Combining Association and Haplotype Studies Towards the Improvement of Fruit Quality in Tomato

Jiantao Zhao

▶ **To cite this version:**

Jiantao Zhao. Combining Association and Haplotype Studies Towards the Improvement of Fruit Quality in Tomato. Agricultural sciences. Université d'Avignon, 2019. English. NNT : 2019AVIG0712 . tel-02790163

HAL Id: tel-02790163
https://hal.inrae.fr/tel-02790163

Submitted on 12 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INRA Centre de recherche PACA
UR 1052, Génétique et Amélioration des Fruits et Légumes

THÈSE

Présentée à Avignon Université pour obtenir le grade de **Docteur en Sciences**

Spécialité : Science Agronomiques

Soutenue le 10 Décembre 2019

Combining Association and Haplotype Studies Towards the Improvement of Fruit Quality in Tomato

par

Jiantao Zhao

Dr. **Charles Eric Durel** (DR, HDR, INRA, Angers)	Rapporteur
Dr. **Laurence Moreau** (DR, HDR, INRA; Gif/Yvette)	Rapporteur
Dr. **Dominique This** (MAF, Montpellier Supagro)	Examinatrice
Dr. **Vincent Segura** (CR, INRA, Montpellier)	Examinateur
Dr. **Mathilde Causse** (DR, HDR, INRA Avignon)	*Directrice de thèse*
Dr. **Christopher Sauvage** (Syngenta)	*Co-encadrant*

Ecole Doctorale :

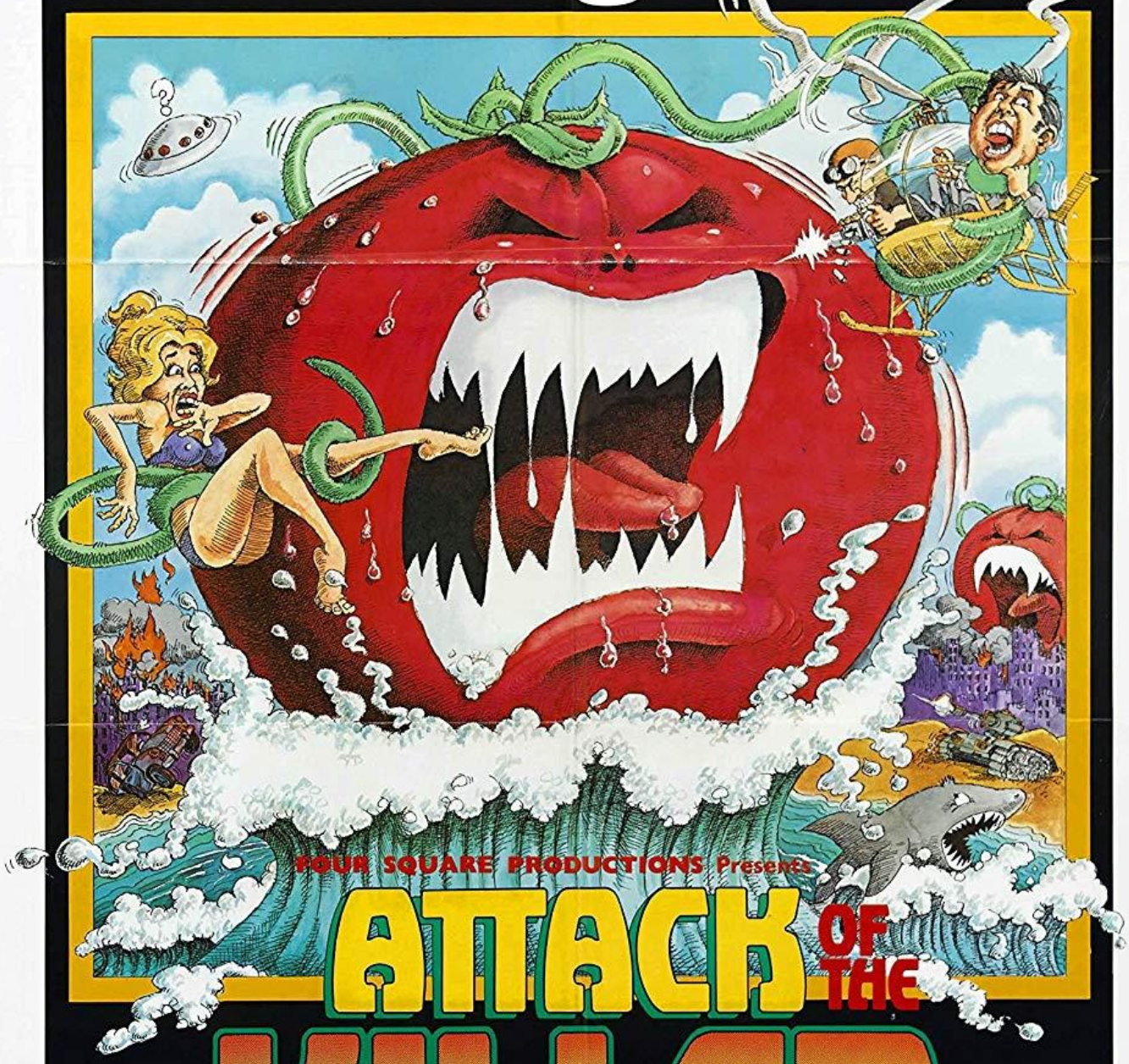
Sciences et Agrosciences (ED536), Avignon

Laboratoires d'accueil :

INRA-UR1052, Génétique et Amélioration des Fruits et Légumes (GAFL), Avignon

Financement : Chinese Scholarship Council (CSC) Scholarship & INRA

"Aaargh!..."



FOUR SQUARE PRODUCTIONS Presents

ATTACK OF THE KILLER TOMATOES

A New Musical-Comedy-Horror Show

Starring **DAVID MILLER • GEORGE WILSON • SHARON TAYLOR • JACK RILEY**

Produced by **STEVE PEACE & JOHN DE BELLO** • Written by **COSTA DILLON, STEVE PEACE & JOHN DE BELLO**

Directed by **JOHN DE BELLO** • Music by **GORDON GOODWIN & PAUL SUNDFOR** • Cinematography by **JOHN K. CULLEY**

AN MGM ENTERTAINMENT RELEASE

PG PARENTAL GUIDANCE SUGGESTED
SOME MATERIAL MAY NOT BE SUITABLE FOR PRE-TEENAGERS

Color by CFI

Acknowledgements

First of all, I would like to thank my home country **People's Republic of China** (**中华人们共和国**, Zhōnghuá rénmen gònghéguó) and the government leaded by the **Chinese Communist Party** (**中国共产党**) for saving my ancestors from sufferings of the World War I and World War II, which had took away almost everything we had ever had in 5000 years, except for hopes. I really hope the world will always be in peace and there will be no wars anymore in the world forever and everyone on this planet works together to make this ordinary world a real paradise. Thanks for the **Chinese Scholarship Council** (CSC, **国家留学基金委**) scholarship for funding me my expenses of the three PhD studies in France. This funding guarantees me to explore the unknown scientific world.

I also want to thank my former university – **Northwest A&F University** (Northwest Agricultural and Forestry University, **西北农林科技大学**) and my supervisor Prof. **Zhirong Zou** (**邹志荣**) and the three guarantors, including Dr. **Jing Zhang** (**张静**), Dr. **Yanfei Cao** (**曹晏飞**) and Dr. **Kai Cao** (**曹凯**) for supporting my applications of the CSC scholarship.

Thank **God** for the world peace and creating tomato – such a wonderful species.

I really want to offer special thanks to my supervisor **Mathilde CAUSSE** (马蒂·考斯) for providing this wonderful PhD offer and this precious opportunity to come to INRA, GAFL. I still clearly remembered the moment I first stepped on the land of France at the airport of Marseille three years ago. The first glance of Mathilde makes me so comfortable and relaxed, which really matters a lot for a boy that comes from a small farming village, let alone going abroad. During the three years, she is always so nice and caring like a mom to me, even though she always tells that she is not. There is an old-saying in China that "**One day teacher, a long-long mother/father**" (**一日为师，终生为父/母**, yī rì wéi shī, zhōng shēn wéi fù), which is the philosophy we still respect deeply in the mindset of all Chinese. As you may know, most of Chinese

are too shy to express their emotions and feelings. Instead, we respect and memorize all those good memories and peoples that love us and us beloved deeply in our mind and soul.

I also want to thank my co-supervisor **Christopher Sauvage** (**苏科瑞**, Sū kēruì) who supervised and leaded me to kingdom of tomato regularly. He is more than a supervisor, and also a smart and talented science uncle and a humor and funny big brother. He is someone that can teach you seriously in doing research and also enjoy a colorful life at the same time. There is one proverb **亦师亦友** (Yì shī yì yǒu), which is exactly the example of the role Christopher plays during my thesis. I really miss him so much since he moved to Syngenta (先正达). Even now, every time I come across to his office, I still cannot stop have a look into his office and I only see our sweetie **Estelle** (**爱美丽**, Ài měilì).

Thanks to **Bitton Frédérique** (**周伯通**, Zhōu bótōng), our team bioinfomatician and engineer, who is a real magician and master of bioinformatics. Under her trainings, I gradually master the primary secrets of these magic languages including R, Linux, python, etc. It's her that shows me how to use these magic powers to explore the kingdom of science more logically, efficiently and powerfully. I also thank to **Guillaume Jean Bauchet** (**傅杰英**, Fù jiéyīng), **Jacques Lagnel** (**卢杰**, Lújié) and **Emmanuel Le-Calonnec** (**罗英**, Luō yīng) for their assistance in maintaining the server and local computer and bioinformatics supports.

I also want to thank my African brother **Isidore Diouf** (**杜思和**, Dù sīhé). He is such a nice guy and like a real brother to me. He helps me a lot not only in regular studies but also daily life. I also want to thank **Stephanie Arnoux** (**奥思凡**, Ào sīfán), **Mariem Omrani** (**欧玛丽**, Ōu mǎlì), **Mariem Nsibi** (**李思琪**, Lǐ sīqí), **Zoe Terret** (**屠悠悠**, Tú yōuyōu), **Anna Bastet** (**巴爱玛**, Bā àimǎ), **Aimeric Agaoua** (**爱中华**, Ài zhōnghuá), **Pierre Sadon** (**张三丰**, Zhāngsānfēng), Hussein Kanso (**吴亦凡**, Wú yìfán) … They are so nice to me during the past three years.

Renaud Duboscq (**杜贺龙**, Dù hèlóng), who shared the same office room, is such a wonderful person to work with. He is funny, very supportive and also a big fun of sports, especially basketball and football. It is such a wonderful experience to find a friend with the same hobby and passion. Everything I am in a very bad mood and stacked with my work, he will always be there to find some Chinese music, some games and some funny jobs to cheer me up…

Many thanks to **Huang Sanwen** (**黄三文**) from CAAS, China, **Harry Klee** (**任我行**), **Denise Tieman** (**田丹妮**) from University of Florida, **Guillaume Bauchet** (**周杰伦**) from BTI, Cornell and **Zhao Jinghua** (**赵京华**) from University of Cambridge for their contributions to my thesis. Deeply thanks to **Fei Zhangjun** (**费张君**), from BTI, Cornell for providing me the postdoc offer. Many thanks to **Zheng Yi** (**郑轶**) and **Wu Shan** (**吴珊**?) and **Wu Yaoyao** (**吴瑶瑶**) for their assistance and sharing of the useful information about Cornell.

Thanks to the administrative stuff from GAFL, such as **Sébastien**, **Evelyne** and **Astrid** for helping with all the registrations, ordre de mission, hotel reservations, tickets… Thanks a lot for your caring. I also thanks to all the team members in GAFL **Rebecca**, **Alexandre**, **Quilot**, **Carretero** … for the good memories we have together.

Thanks to my other Chinese friends in France, including Jin Xiuliang (**金秀良**), Diao Wanying (**刁万英**), Lan Weijie (**兰维杰**), Li Linyuan (**李林源**), Zhang Xu (**张旭**), Wang Xueqiu (**王雪秋**), Yu Jiahao (**余佳浩**), Liu Xuwei (**刘旭伟**), Huang Jing (**黄晶**), Liu Shouyang (**刘守阳**), Jiang Jingyi (**蒋静怡**), Xue Zeyun (**薛泽云**), Zhang Siqi (**张思琪**), Chen Shen (**陈申**), Lu Zijie (**陆子杰**), Lu Zijun (**陆子俊**)…. With you guys I enjoyed a lot of wonderful off-work times, playing cards, basketball, hot spicy Chinese food, travelling, etc.

At last, I want to express my deepest thanks to my family members, my father **Zhao Huaishen** (**赵怀申**), my mom **Cui Xiuhua** (**崔绣花**), my cute lovely younger sister **Zhao Shuhui** (**赵淑慧**) and my beloved girlfriend **Xu Yao** (**徐瑶**). Thanks my grandparent, my uncles, aunts and cousins. They are always caring about me, supporting me, no matter what happens. I love you guys. In particular, words are too week to express how I am grateful to my sweetie girlfriend. It's her that makes me believe in miracles and paradise. I really hope she can marry me someday and that day comes soon! I love you (我爱你, Wǒ àinǐ).

THÈSE

Combinaison d'études d'associations et d'haplotypes pour l'amélioration de la qualité des fruits chez la tomate

THESE

Combining association and haplotype studies towards the improvement of fruit quality in tomato

Jiantao Zhao, 2019

Résumé

Les consommateurs se plaignent de la qualité gustative des tomates depuis des décennies. Celle-ci est influencée principalement par les sucres, les acides et un ensemble de divers composés volatils. L'amélioration de la saveur de la tomate reste l'un des principaux défis à relever pour améliorer la qualité de la tomate et l'acceptabilité des consommateurs pour l'amélioration moderne des tomates. Le but principal de cette thèse était de disséquer le contrôle génétique de la saveur de la tomate en utilisant des SNP à haute densité et un ensemble divers de traits liés à la saveur, notamment les sucres, les acides, les acides aminés et les composés volatils. Dans la première partie, j'ai effectué plusieurs analyses basées sur l'exploration des haplotypes dans une collection d'accessions. Plusieurs approches ont été utilisées et comparées pour identifier les régions génomiques en cours de sélection. Les modèles bayésiens de génétique d'association basés sur les haplotypes et une partie des SNP ont identifié 108 associations significatives pour 26 caractères. Parmi ces associations, certains gènes candidats prometteurs ont été identifiés. Certains avantages de l'utilisation des haplotypes ont également été présentés. Dans la deuxième partie, j'ai réalisé une méta-analyse d'études d'association pangénomique à l'aide de trois panels d'associations de tomates. J'ai démontré l'efficacité de l'imputation des génotypes pour augmenter la couverture de SNP à l'échelle du génome. Des méta-analyses de modèles à effets fixes et à effets aléatoires (pour les SNP présentant une hétérogénéité $I^2 > 25$) ont été effectuées afin de contrôler l'hétérogénéité croisée des études. Au total, 305 locus significatifs ont été identifiés, dont 211 nouveaux. Parmi ceux-ci, 24 locus ont présenté des cis-eQTL lors d'une précédente étude d'association à l'échelle du transcriptome de fruits. L'analyse d'enrichissement pour toutes les associations a montré que jusqu'à 10 processus biologiques étaient enrichis de manière significative et que tous étaient étroitement impliqués dans les métabolites liés aux arômes. Une liste de gènes candidats prometteurs a été fournie, qui pourraient présenter un grand intérêt pour la validation fonctionnelle. J'ai également démontré la possibilité d'augmenter de manière significative le contenu en composés volatils qui contribuent de manière positive aux préférences des consommateurs tout en réduisant les volatils désagréables, en sélectionnant les combinaisons d'allèles pertinentes. Globalement, cette thèse augmente les connaissances du contrôle génétique du goût de la tomate, ce qui devrait contribuer à son amélioration.

Abstract

Consumers have been complaining about tomato flavor for decades. Tomato taste is mainly influenced by sugars, acids and a diverse set of volatiles. Improving tomato flavor remains one of the main challenges for improving tomato sensory quality and consumer acceptability in modern tomato breeding. The main purpose of this thesis was to decipher the genetic and evolutionary control of tomato flavor by using high density SNPs and a diverse set of flavor-related metabolites, including sugars, acids, amino acids and volatiles. In the first part, I performed multiple haplotype-based analyses on a tomato core collection. Several approaches were used and compared to identify the genomic regions under selection. Haplotype and SNP-based Bayesian models identified 108 significant associations for 26 traits. Among these associations, some promising candidate genes were identified. I also compared marker local haplotype sharing (mLHS) with LD in determining the candidate regions. In addition, some general benefits of using haplotypes were also provided as general discussions. In the second part, I pioneered in introducing meta-analysis of genome-wide association studies using three tomato association panels. I demonstrated the efficiency of genotype imputation in increasing the genome-wide SNP coverage. Both fixed-effect and random-effect models (for those SNPs with heterogeneity $I^2 > 25$) of meta-analysis were performed in order to control cross-study heterogeneity. A total of 305 significant loci were identified and 211 of which were new. Among them, 24 loci exhibited cis-eQTLs in a previous transcriptome-wide association study in fruit tissue. Enrichment analysis for all associations showed that up to 10 biological processes were significantly enriched and all of which were closely involved in flavor-related metabolites. A list of promising candidate genes was provided, which could be of great interest for functional validation. I also demonstrated the possibility to significantly increase the content of volatiles that positively contribute to consumer preferences while reducing unpleasant volatiles, by selection of the relevant allele combinations. Taken together, this thesis provides a comprehensive knowledge of the genetic control of tomato flavor, which will promote its improvement.

Résumé substantiel de la thèse en Français

Les consommateurs se plaignent de la dégradation de la saveur des tomates modernes depuis plusieurs décennies. Cependant, l'amélioration de la qualité sensorielle globale présente des difficultés pour plusieurs raisons: 1) le goût est moins important comparé au rendement, aux résistances aux maladies et à l'adaptation aux conditions de croissance qui intéressent les producteurs; 2) la qualité sensorielle est principalement déterminée par un ensemble d'attributs décrivant les propriétés externes du fruit (taille, couleur, fermeté) et internes (saveurs, arômes, textures), ce qui est complexe et difficile à sélectionner et mesurer simplement et largement influencé par l'environnement au sens large; 3) du point de vue métabolique, la saveur est principalement due aux teneurs en sucres et en acides organiques ainsi qu'à leur rapport ainsi qu'à la composition en arômes volatils, dérivés de multiples voies de biosynthèse. Cependant nos connaissance de leur déterminisme génétique est assez limitée et seuls quelques gènes régulateurs ont été fonctionnellement clonés; 4) les métabolites contribuant positivement à la saveur de la tomate, en particulier les sucres présentent généralement une corrélation négative avec le poids du fruit et donc une saveur de tomate globalement améliorée pourrait réduire le rendement, ce qui n'est pas souhaitable pour les producteurs; 5) différents consommateurs peuvent avoir des préférences différentes et il n'existe pas de cultivar de tomate répondant à toutes les préférences.

Dans cette thèse, nous nous sommes concentré sur le contrôle génétique des teneurs en métabolites liés à la qualité, incluant les sucres, les acides organiques, les acides aminés et divers composés volatils. Parmi tous les facteurs ayant un impact sur la saveur globale, la modification de ces métabolites pourrait avoir des effets majeurs directs sur la perception globale de la saveur de la tomate. Avec le développement rapide des méthodes de génotypage, telles que les puces SNP et le séquençage de nouvelle génération (NGS) et des méthodes de phénotypage, telles que la chromatographie en phase gazeuse couplée à la spectrométrie de masse (GC-MS), la chromatographie en phase liquide couplée à ma spectrométrie de masse (LC-MS), le génotypage et le phénotypage d'un large panel d'accessions de tomate est maintenant possible, ce qui offre de nouvelles opportunités pour disséquer le contrôle génétique des métabolites impliqués dans la qualité (dans cette thèse, nous nous concentrons uniquement sur les métabolites liés à la saveur et aux arômes).

Au début de cette thèse, deux collections de tomates (notées panels S et B) avaient été étudiées par le laboratoire INRA du GAFL et étaient phénotypées et génotypées, toutes deux avec un accent particulier sur les métabolites liés à la qualité gustative. Dans le même temps, une autre collection de tomates avait été caractérisée par un autre groupe (également axé sur les métabolites liés aux saveurs et à à l'arôme) et des génotypes produits par NGS (panel T). Toutes les données génotypiques et phénotypiques étaient déjà librement disponibles. Les études individuelles d'association pangénomique (GWAS) basées sur ces trois panels ont révélé une forte hétérogénéité entre les études en termes de nombre et de position des loci significativement associés, bien que certaines associations aient également été détectées dans plusieurs études. Ces résultats démontrent l'une des principales limitations de l'utilisation de la GWAS dans l'identification des loci à effets génétiques modérés à faibles.

Les études génétiques et génomiques chez l'humain sont toujours pionnières dans la production de nouveaux modèles et approches statistiques, car davantage de ressources sont disponibles. Avec des milliers de génomes humains séquencés avec une profondeur de séquence élevée, l'imputation du génotypage a été introduite pour réduire le coût du génotypage (en génotypant les individus avec des puces de SNP au lieu de NGS) et maintenir la couverture génomique (la couverture génomique peut être considérablement accrue après imputation). Ceci a été à la base de la génétique des populations humaines moderne. L'une des principales applications de l'imputation est son intégration dans la GWAS. Dans les études modernes de GWAS humaine, la taille de l'échantillon de GWAS a atteint plusieurs milliers, voire plusieurs millions d'individus, ce qui offre de nouvelles opportunités pour de nouveaux modèles statistiques. Parmi ces modèles, la méta-analyse de GWAS utilisant uniquement les résultats résumés d'études individuelles fournit une excellente opportunité pour intégrer les résultats de GWAS de différents panels. Cette approche s'est révélée pussiante en termes statistiques en 1) confirmant des associations significatives déjà identifiées, 2) en identifiant de nouvelles associations significatives, 3) en traitant une hétérogénéité croisée. Cependant, à notre connaissance, la méta-analyse de GWAS a rarement été introduites dans les études sur les plantes. Par conséquent, une partie majeure de la thèse porte sur la manière d'appliquer la méta-analyse de GWAS à l'aide des trois panels de GWAS disponibles.

Plusieurs modèles statistiques ont été développés afin d'accroître l'efficacité de la GWAS dans l'identification des associations significatives, telles que le modèle mixte multi-locus (MLMM) et le modèle mixte multi-traits (MTMM), qui a été largement appliqué pour ces panels qui sont génotypés avec des puces SNP (portant quelques milliers de SNP). Cependant, la couverture génomique limitée des marqueurs rend difficile 1) l'identification des loci à effet génétique modéré à faible et des régions où le LD est faible et où les marqueurs ne sont pas nombreux; 2) l'identification des gènes candidats par cartographie fine locale.

Les haplotypes sont les combinaisons particulières d'allèles observées sur une région d'un chromosome dans une population donnée. Les blocs haplotypiques sont les régions dans lesquelles il existe peu de traces de recombinaison historique, et dans lesquelles seuls quelques haplotypes communs sont observés. Le génotypage de seulement quelques SNP soigneusement choisis pourrait fournir suffisamment d'informations pour identifier les haplotypes communs. Les allèles d'un même bloc haplotypique ont plus de chances d'être hérités ensemble, tout en partageant la même fréquence d'allèle mineur (MAF). Les analyses basées sur les haplotypes examinent des groupes de SNP plutôt que des SNP individuels et améliorent la puissance de détection statistique pour de nombreux aspects, y compris l'identification des signaux de sélection et la GWAS. Par conséquent, dans la deuxième partie de cette thèse, nous allons nous intéresser à l'introduction d'haplotypes sous plusieurs aspects afin d'obtenir un aperçu global des avantages de l'utilisation des haplotypes pour l'identification de régions sous sélection et la GWAS.

Nous avons donc organisé cette thèse en cinq chapitres. Le chapitre 1 fournit une analyse bibliographique des sujets couverts dans cette thèse. Nous avons d'abord présenté les principaux défis, priorités et objectifs de sélection de la qualité de la tomate. Nous nous sommes principalement concentrés sur deux objectifs, la productivité et la qualité des fruits tant au niveau nutritionnel que sensoriel. Nous avons ensuite présenté les ressources génétiques disponibles au niveau international, y compris l'origine des tomates et les ressources génétiques des apparentées sauvages. Nous avons ensuite présenté les principales ressources génomiques de la tomate disponibles dans le monde, en analysant l'historique du projet de séquencage du génome et de toutes les ressources génomiques de la tomate générées par séquence. Nous avons ensuite introduit les analyses de la diversité génétique des ressources

essentielles pour plusieurs applications. Nous avons ensuite fourni des informations détaillées sur la manière de détecter les empreintes de sélection au niveau génomique, notamment 1) pourquoi il est important de détecter les empreintes de sélection, 2) comment détecter les balayages sélectifs, 3) les applications récentes et leurs limites, avec un objectif central sur la tomate et 4) de nouvelles possibilités de détecter ces empreintes grâce aux nouveaux modèles statistiques. Nous avons ensuite résumé les approches de marquage et leurs applications dans l'identification des gènes / QTL, sous plusieurs aspects: 1) l'évolution des marqueurs moléculaires; 2) les marqueurs SNP et les approches associées pour générer des SNP denses, avec un accent particulier sur les puces de SNP, le reséquençage et l'imputation du génotypage; 3) les populations spécifiques pour disséquer les déterminants de phénotypes; 4) les principaux résultats de la cartographie de QTL et les gènes clonés; 5) la GWAS. Nous avons ensuite fourni une revue détaillée sur la méta-analyse de GWAS,incluant 1) les avantages de la méta-analyse de GWAS; 2) les modèles statistiques utilisés pour l'effectuer et 3) certains problèmes et perspectives de cette approche. Nous avons ensuite fourni une introduction détaillée sur les haplotypes. Nous avons finalement introduit la sélection génomique, notamment: 1) les principes fondamentaux de la sélection/ prédiction génomique; 2) les modèles de prédiction génomique les plus appliqués; 3) les facteurs influençant la précision de la prédiction; 4) ses applications à la tomate. Enfin, nous avons présenté les questions scientifiques et le plan de cette thèse.

Le **chapitre 2**présente le résumé global des matériels et méthodes utilisés dans la thèse. Globalement, cette thèse porte sur trois panels GWAS, génotypés et phénotypés avec un ensemble diversifié de traits liés à la saveur. Ils comprennent le panel S (Sauvage et al., 2014), le panel B (Bauchet et al., 2017) et le panel T (Tieman et al., 2017).

Le **chapitre 3** est axé sur plusieurs analyses basées sur les haplotypes et nous avons démontré que l'utilisation d'haplotypes fournissait de nouvelles informations génétiques et évolutives sur le poids et la composition des fruits de la tomate. Ce chapitre est un projet d'article mettant l'accent sur la combinaison de la génétique des populations et de la génétique quantitative afin d'approfondir nos connaissances sur le contrôle génétique du poids et de la composition des fruits de tomate. Nous avons cherché à déchiffrer les empreintes moléculaires de la sélection, à identifier les

associations haplotype-trait, à fournir une description du paysage haplotypique sous les associations marqueur-trait et à comparer le partage d'haplotype local avec les estimations de déséquilibre de liaison afin d'affiner la recherche de gènes candidats. Nous avons également testé les avantages de l'utilisation des haplotypes pour améliorer la prédiction génomique et mis l'accent sur les promesses de ce type d'approche à des fins de sélection.

Tout d'abord, nous avons détecté un total de 784 blocs haplotypiques dans une collection de 163 accessions. La taille moyenne des blocs d'haplotype était de 58,085 kb. En utilisant le « score d'haplotype intégré » (iHS), nous avons identifié 24 balayages sélectifs positifs, dont neuf ne se chevauchaient pas avec des balayages détectés au niveau de la domestication ou de l'amélioration. Les modèles bayésiens basés sur les haplotypes et les SNP ont identifié 108 associations significatives pour 26 caractères, ce qui est supérieur aux études précédentes. Parmi les associations, 77 étaient situées dans des zones soumises à pressions sélectives. Nous avons montré que le « partage d'haplotype local de marqueurs » (mLHS) constituait une alternative au modèle de décroissance du déséquilibre de liaison pour définir les intervalles de confiance autour des associations pour rechercher des gènes candidats et pu proposer quelques gènes candidats. Le schéma de décomposition des haplotypes locaux et la longueur des haplotypes au sein de différents groupes d'accessions ont fourni de nouvelles informations sur l'histoire démographique des loci associés. Nous démontrons ainsi le pouvoir d'utiliser les haplotypes pour des études évolutives et génétiques, fournissant de nouvelles informations sur l'amélioration de la qualité de la tomate et l'historique de la sélection.

Le **chapitre 4** est un article publié dans Nature Communications (DOI: 10.1038 / s41467-019-09462-w). Dans cet article, nous avons expliqué en détail comment effectuer une méta-analyse d'études d'association pangénomique en utilisant les résultats résumés de trois panels. Avant la méta-analyse, nous avons effectué une imputation de génotype pour les panels B et S, qui ont été génotypés avec des puces de SNP, afin d'augmenter la couverture des génomes. Nous avons ensuite utilisé EMMAX pour les tests d'association pour les panels S et B individuellement. Nous avons ensuite effectué la méta-analyse à l'aide d'un modèle à effets fixes en utilisant le logiciel METAL. Pour les SNP présentant une hétérogénéité (I2> 25), nous avons ensuite utilisé le modèle à effets aléatoires proposé dans METASOFT. La méta-

analyse a identifié un total de 305 locus significatifs, dont 211 nouveaux. Parmi ceux-ci, 24 locus ont présenté des cis-eQTL lors d'une précédente étude d'association à l'échelle du transcriptome sur des tissus de fruits. L'analyse d'enrichissement pour toutes les associations a montré que jusqu'à 10 processus biologiques étaient enrichis de manière significative et qu'ils étaient tous étroitement impliqués dans les métabolites liés à la flaveur (en termes de sucres, d'acides organiques, d'acides aminés et de composés volatils). Une liste de gènes candidats prometteurs a été fournie, ce qui pourrait présenter un intérêt pour la validation fonctionnelle. A partir des associations, nous avons démontré que la sélection lors de la domestication et de l'amélioration a eu un impact sur la teneur en citrate et en malate de fruits, alors que la teneur en sucres a été soumise à une sélection moins stringente. Nous suggérons qu'il est possible d'augmenter de manière significative le contenu en composés volatils qui contribuent positivement aux préférences des consommateurs tout en réduisant les volatils désagréables, en sélectionnant les combinaisons d'allèles pertinentes.

Le **chapitre 5** présente les perspectives et conclusions de la thèse.. En résumé, dans cette thèse, nous avons conçu et mis en œuvre des approches innovantes en génomique qui nous ont permis d'approfondir notre compréhension du contrôle génétique de la qualité de la tomate. Ces résultats nous conduisent à proposer plusieurs questions de recherche pour l'avenir, notamment: 1) comment équilibrer les volatiles positifs / négatifs 2) comment lever les difficultés rencontrées pour identifier de nouvelles associations significatives; 3) comment tirer davantage parti de l'imputation par génotype ; 4) comment approfondir nos connaissances sur l'histoire démographique de la tomate ; 5) comment intégrer les haplotypes dans de véritables pratiques de sélection ; 6) comment calculer l'héritabilité sur la base des données GWAS résumées et enfin 7) comment intégrer tous les résultats de cette thèse aux autres connaissances évolutives, génétiques, génomiques, métaboliques et transcriptomiques disponibles afin d'améliorer le goût général de la tomate.

Substantial Summary of the Thesis in English

Consumers have been complaining about the deteriorated flavor of modern tomato cultivars over decades. However, improving the overall sensory quality is challenging for breeders because of several reasons: 1) flavor is less important then yield, disease resistances or adaptation to growth conditions for growers; 2) sensory quality is determined by many attributes, describing external (size, color, firmness) and internal (flavor, aroma, texture) properties, which can not be assessed by a simple measurement; 3) from the metabolic perspective, flavor is mostly due to sugars and organic acids and to their ratio and also to the composition in volatile aromas, which are derived from multiple synthesis pathways, but our knowledge of their genetic control is quite limited and only a few regulatory genes have been functionally cloned; 4) the metabolites positively contributing to flavor are usually negatively correlated to fruit weight, especially sugars; 5) A high significantly enhanced overall tomato flavor might reduce yield, which is unwanted for growers; 6) different people might have different preferences and there is not a single tomato cultivar meeting all the preferences.

In this thesis, we focused on the genetic control of flavor-related metabolites, including sugars, acids, amino acids and volatiles. Among all the factors impacting the overall flavor, modification of these metabolites could have a direct major effect on the overall tomato flavor. In order to do so, we need to know the genetic architecture of how these metabolites are controlled and regulated. With the fast development of genotyping methods, such as SNP arrays and next-generation sequencing (NGS) and phenotyping methods, such as gas chromatography coupled to mass spectrometry (GC-MS), liquid chromatography coupled to mass spectrometry (LC-MS) and high performance liquid chromatography (HPLC), genotyping and phenotyping a large panel of tomato accessions are nowadays possible, which provides great opportunities to dissect the genetic control of the metabolites that are of interests (in this thesis we only focus on flavor-related central metabolites).

At the beginning of this thesis, the INRA laboratory had studied two diverse tomato collections (named panel S and panel B) that were deeply phenotyped and genotyped, both with a key focus on flavor-related metabolites. At the same time, another group published a tomato collection with deep phenotypes (also focused on flavor-related

metabolites) and genotypes (NGS), which herein was referred as panel T. All the genotypic and phenotypic data were freely available. However, genome-wide association studies (GWAS) based on each of the three panels revealed strong cross-study heterogeneity (non-random variance across different studies) in terms of the number and position of the significantly associated loci, though some same associations were also detected across studies. These results demonstrate one of the main limitations of using GWAS in identifying the loci with moderate to weak genetic effects.

Human genetic and genomic studies are always pioneer in new statistical models and approaches, as more resources are available, in terms of intellectual advantages, funding, samples, etc. After thousands of human genomes with high sequence depth availability, genotyping imputation was introduced to greatly reduce the genotyping cost (genotyped with SNP arrays instead of NGS) and maintain the genomic coverage (the genomic coverage can be significantly increased after imputation), which has been the foundation of modern human population genetics. One of the main applications of imputation is its integration into GWAS. In modern human genomic studies, the sample size of GWAS has reached several thousands to millions, which provides opportunities to introduce new statistical models assuming that the population size is sufficiently large. Among these models, meta-analysis of GWAS using only the summary results of individual study provide great opportunity to integrate the GWAS results from different panels. This approach demonstrates great statistical powers in 1) confirming already identified significant associations, 2) identifying new significant associations, 3) handling cross-study heterogeneity. However, to our best knowledge, meta-analysis of GWAS has been rarely introduced into major crops. Therefore, one major part of the thesis is focusing on the application of GWAS meta-analysis using the three available GWAS panels. We think this strategy will also be helpful and insightful for other crop breeding and improvement.

From the more practical perspective, due to low genetic diversity and high linkage disequilibrium (LD), especially of modern large-fruit tomatoes compared to cherry tomatoes and the closest wild species, several thousands of SNPs will be quite effective in real practices of tomato breeding (due to the strong linkage disequilibrium of markers).

Several statistical models have been developed in order to increase the efficiency of GWAS in identifying significant associations, such as multi-locus mixed model (MLMM) and multi-trait mixed model (MTMM), which have been applied to the panels genotyped with SNP arrays (thousands of high-quality SNPs will be achieved). However, the limited genomic coverage of markers makes it challenging in 1) identifying those loci with moderate to low genetic effect and those regions where LD is weak and the markers are only a few; 2) narrowing down the candidate genes by regional fine-mapping.

Haplotypes are the particular combinations of alleles observed in a chromosome region in a given population. Haplotype blocks are the regions where there is little evidence for historical recombination and within which only a few common haplotypes are observed. Genotyping only a few, carefully chosen tag-SNPs will provide enough information to identify the most common haplotypes. Alleles within the same haplotype block are more likely to be inherited together, while sharing similar minor allele frequency (MAF). Haplotype-based analyses examine groups of SNPs rather than individual SNPs and enhance the statistical detection power for many aspects, including identifying signals of recent positive selection and GWAS. Therefore, in the second major part of this thesis, we analysed the benefits of using haplotypes for multiple aspects, such as identifying genomic regions under selection and haplotype-based GWAS. We think these analyses will open a new window to maximize the genetic gains from the available resources, especially for those panels that are only genotyped with SNP arrays.

Therefore, we organized this thesis into five chapters. **Chapter 1** provides a general bibliographic introduction about the topics studied in this thesis. We first introduced the main challenges, priorities and breeding objectives of tomato quality (productivity and fruit quality at both nutritional and sensory levels). We then introduced the main genetic resources available at the international level, including the origin of tomatoes and its wild relatives and related genetic resources and how to generate new genetic resources. We then introduced the main genomic resources of tomato, the history of the tomato genome sequence project and all the tomato genomic resources that have been generated by sequencing. We then introduced the genetic diversity analyses of the worldwide resources, which is essential for several applications and studies. We next provided detail introductions about how to detect the selective footprints at the

genomic level, showing 1) why it is important to detect selective footprints, 2) how to detect selective sweeps, 3) recent applications and limitations in crops, with a central focus on tomato and 4) new opportunities in detecting these footprints with the benefits of new statistical models. We then summarized the achievements of molecular markers and their applications in identifying genes/QTLs, which was explained in details from several aspects, evolution of molecular markers, SNP markers and related approaches to generate dense SNPs, with a central focus on SNP arrays, resequencing and genotyping imputation, specific populations to dissect phenotype determinants, main achievements of trait mapping using linkage mapping and genes that have been cloned and GWAS. We then provided detailed introductions about meta-analysis of GWAS, as this is quite promising and efficient but with few applications in major crops. We present the benefits of meta-analysis of GWAS, the statistical models in performing meta-analysis of GWAS and some future issues and prospects of this approach. We then introduced haplotype concept and its benefits for genetic studies. We also introduced genomic selection, due to its increasing interests, including the principle of genomic selection/genomic prediction, the most applied genomic prediction models, the factors influencing the prediction accuracy and its applications in tomato.

Chapter 2 provides the global summary of the materials and methods used in the thesis. Overall, this thesis mainly focused on three GWAS panels, which have been both genotyped and phenotyped with a diverse set of flavor-related traits. They include panel S (Sauvage et al., 2014), panel B (Bauchet et al., 2017) and panel T (Tieman et al., 2017).

Chapter 3 is focused on multiple haplotype-based analyses and we demonstrate that using haplotypes provides new genetic and evolutionary insights into tomato fruit weight and composition. This chapter is a draft manuscript focusing on the combination of population and quantitative genetics applied to haplotypes to deepen our knowledge of marker-trait associations for fruit weight and composition in tomato. We aimed at deciphering the molecular footprints of selection, identifying haplotype – trait associations, providing a description of the haplotype landscape under marker – trait associations and comparing marker local haplotype sharing with linkage disequilibrium estimates to narrow down the search for candidate genes. We also

tested the benefits of using haplotypes in improving the genomic prediction as general discussion and put more emphasis on the promise of this type of approach for breeding purposes.

In this chapter, we performed and compared single SNP and multiple haplotype-based association analyses for tomato fruit weight and metabolite contents. First, using 6000 SNPs we detected a total of 784 haplotype blocks in a collection of 163 tomato accessions. The average size of haplotype blocks was 58.085 kb. By using integrated haplotype score (iHS), we identified 24 positive selective sweeps, among which, nine were non-overlapping with either domestication or improvement sweeps. Haplotype and SNP-based Bayesian models identified 108 significant associations for 26 traits, which outperformed previous studies. Among the associations, 77 were located within selective sweeps. Marker local haplotype sharing (mLHS) provided an alternative to linkage disequilibrium decay pattern to define confidence intervals around the associations to seek for candidate genes. Local haplotype decaying pattern and the length of haplotypes within different groups of accessions provided new insights on the demographic history of the associated loci. We thus demonstrate the power of using haplotypes for evolutionary and genetic studies, providing novel insights into tomato quality improvement and breeding history.

Chapter 4 is an article that has been published in Nature Communications (DOI: 10.1038/s41467-019-09462-w). In this paper, we demonstrate in details how to perform meta-analysis of genome-wide association studies by using the summary results from different panels. These results can be helpful in deepening our understandings on the genetic control of tomato flavor. Before meta-analysis, we performed genotype imputation for panels B and S, which were genotyped with SNP arrays, to increase the SNP coverage. We then used the EMMAX package for providing association test statistics for panels S and B, respectively. We first performed the meta-analysis using a fixed effect model using METAL software. For those SNPs with heterogeneity ($I^2 > 25$), we then performed the random effect model in METASOFT software. Meta-analysis identified a total of 305 significant loci, among which 211 were new. Among these, 24 loci exhibited cis-eQTLs in a previous transcriptome-wide association study in fruit tissue, which suggested that a polymorphism in the region of the gene impacts its expression. Enrichment analysis for all associations showed that up to 10 biological processes were significantly

enriched and all of which were closely involved in flavor-related metabolites (in terms of sugars, organic acids, amino acids, and volatiles). A list of promising candidate genes was provided, which could be of great interest for functional validation. We demonstrate that fruit citrate and malate contents have been impacted by selection during domestication and improvement, while sugar content has undergone less stringent selection. We suggest that it may be possible to significantly increase the content of volatiles that positively contribute to consumer preferences while reducing unpleasant volatiles, by selection of the relevant allele combinations.

Chapter 5 provides prospects and conclusions of the thesis. In summary, in this thesis, we have designed and performed innovative genomics approaches in order to deepen our understanding of the genetic control of tomato quality. Following the conclusions, we indicated several aspects that might provide new research windows in the future, including 1) how to balance those positive/negative volatiles; 2) challenges in identifying new significant genotype-phenotype associations; 3) how to gain more from genotype imputation; 4) how to deepen our knowledge about tomato demographic history; 5) how to integrate haplotypes into real tomato breeding practices; 6) how to calculate the heritability based on summary GWAS data and 7) how to integrate all the achievements in this thesis with other available related evolutionary, genetic, genomic, metabolic and transcriptomic knowledge to improve the overall tomato flavor.

Scientific communications

Articles

Zhao J, Sauvage C, Zhao J, Bitton F, Bauchet G, Liu D, Huang S, Tieman DM, Klee HJ, Causse M (2019). Meta-analysis of genome-wide association studies provides insights into genetic control of tomato flavor. **Nat Communications** 10: 1534 (**displayed in Chapter 4**)

Book Chapter

Causse M, **Zhao J**, Diouf I, Wang J, Lefebvre V, Caromel B, Génard M, Bertin N (2019). *Genomic designing for climate smart tomato*. In Genomic Designing of Climate Smart Crops, Springer Nature (**accepted for publication, displayed in Appendix 3**)

Congresses

Zhao J, Sauvage C, Bitton F, Causse M (2019). What can we gain from the analysis of recent positive selection signature and multiple haplotype-based analyses in tomato? The XVI Soanaceae Conference, Jerusalem, Israel (**Oral communication**)

Zhao J, C Sauvage and M Causse (2019) What can we gain from multiple haplotype-based analyses. Journées Jeunes Chercheurs BAP, Avignon, France (**Oral communication**)

Zhao J, C Sauvage and M Causse (2018) Genomic selection for fruit quality and yield in tomato. XIX EUCARPIA Meeting of the Tomato Working Group, Naples, Italy (**Poster**)

Zhao J, C Sauvage and M Causse (2018) Deciphering the genetic background of fruit quality traits through a meta-GWAS strategy. Solanaceae Conference, Chiang Mai, Thailand (**Oral communication, presented by Sauvage**)

Zhao J, C Sauvage and M Causse (2018) Meta-analysis of Genome-wide Association Provides New Insights into Tomato Favor. Journées Jeunes Chercheurs BAP, Paris, France (**Oral communication**)

Zhao J, C Sauvage and M Causse (2017) Genomic selection for fruit quality and yield in tomato. Journées Jeunes Chercheurs BAP, Bordeaux, France (**Poster**)

Table of Contents

Chapter 1

Chapter 1 General introduction

This chapter aims to provide from general to detailed explanations about the genetic challenges of tomato quality breeding, where centrally flavor-related traits were mainly focused, including fruit weight, sugars, acids, amino acids and a series of volatiles. In order to dissect the underlying mechanisms of these important quality traits, and improve tomato flavor, we will herein provide comprehensive reviews on what we have in terms of genetic resources, what breeding challenges of we meet, what we have known about tomato genetics, such as molecular markers and genes/QTLs, how to detect the selective events, how to identify candidate genes associated with fruit quality via linkage mapping and association mapping, how to integrate all these knowledge in accelerating tomato breeding in real breeding practices, such as via genomic prediction. Though not all these aspects will be investigated in thesis, at least not with same efforts and attention, we think these overviews together will be helpful to guide researchers and breeders to breed tomato cultivars.

1.1 Challenges, priorities and breeding objectives of tomato quality

Tomato is the first vegetable consumed worldwide after potato. It has become an important food in many countries, especially in those regions where micronutrients and vitamins are still limited in the diet, such as Asia and Africa. Nowadays, there are two main types of tomato varieties produced, processing tomatoes and fresh market tomatoes. Tomato crop faces several challenges, which impacts its breeding objectives. Breeders will orientate their main breeding objectives according to the wide diversity of growth conditions and final use as fresh or processed. These objectives can be classified in (1) productivity, (2) fruit quality at both nutritional and sensory levels and (3) adaptation to growth conditions in terms of response to biotic and abiotic stresses. However, biotic and abiotic stresses are not the main purpose of this thesis and we will not provide detailed introductions for this aspect.

1.1.1 Productivity

From 1988 to 2017, the tomato world production regularly grew from 64 MT (million tons) to 182 MT. Since 1995, China increased its production and became the first producer, and since then, its production increased up to 60 MT (**Figure 1.1**) covering almost 4,800,000 ha. This growth is mainly due to an increase in production area, especially with the fast

development of Chinese solar greenhouses in China, and to improvement in productivity and variety breeding.

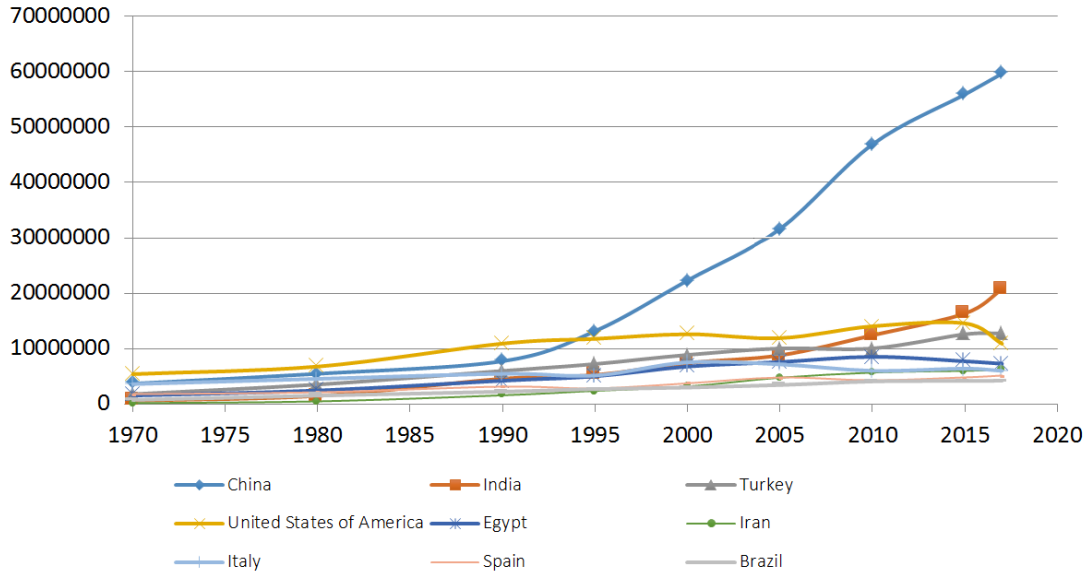


Figure 1.1 Evolution of tomato production over years in the 9 main producing countries

With an average yield of 37 T/ha, compared to 16 t/ha in 1961, yield has increased over years but large differences remain according to countries and growth conditions. In south Europe greenhouses, the average yield is 50-80 T/ha, while it may be more than 400 T/ha in the Netherland and Belgium, with a crop lasting up to 11 months. Expressed per square meter, the average yield is 3.7 kg/m^2, reaching 50 kg/m^2 in the Netherland, while it is 5.6 in China where most of the production is in open field although modern Chinese solar greenhouses are developed (Cao et al., 2019). Advancement in related fields, such as environmental control technologies, engineering, artificial intelligence, also promote the overall production of tomato.

Yield is strongly dependent on cultivars and growth conditions. It results from fruit number and fruit weight. Fruit weight is one of the main breeding target during the long-term domestication and breeding history (Lin et al., 2014). Cultivars for fresh market are classified based on their fruit size and shape from the cherry tomato (less than 20 g) to beef tomato (fruit weight higher than 200 g). The potential size depends on cell number established in pre-anthesis stage, but final fruit size mainly depends on the rate and duration of cell enlargement (Ho, 1996). Seed number and competition among fruits also affect the final fruit size (Bertin et al., 2002; Bertin et al., 2003). Seed and fruit are highly sensitive to biotic and abiotic

stresses, which often lead to seed and fruit abortion (Ruan et al., 2012). Fruit number is controlled by the truss architecture but the increase in flower number often leads to abortion (Soyk et al., 2017a). Fruit shape varies from flat to long or ovate fruit and is also determined at the carpel development stage. Mutations in four genes explain most of the tomato fruit shape (Rodríguez et al., 2011).

1.1.2 Fruit quality

1.1.2.1 Nutritional quality

Tomato consumption has been shown to reduce the risks of certain cancers and cardiovascular diseases (Giovannucci, 1999). Its nutritional value is related to fruit composition in primary and secondary metabolites (**Table 1.1**), but is mostly due to its content in lycopene and carotene (Bramley, 2000; Bramley, 2002). Lycopene is responsible of the red fruit color but also acts as a dietary antioxidant. Tomato also constitutes an important source of vitamin C. In spite of considerable efforts in developing cultivars with higher content in carotenoids, or in vitamin C, none has reached a commercial importance, in part because of a negative relation between yield and these traits (Klee, 2010).

In addition to these well-known vitamins and antioxidants, other compounds in tomato fruit with antioxidant properties include chlorogenic acid, rutin, plastoquinones, tocopherol, and xanthophylls. Tomatoes also contribute but to a lesser extent in carbohydrates, fiber, flavor compounds, minerals, protein, fats and glycoalkaloids to the diet (Davies and Hobson, 1981). Exhaustive metabolome studies have completed the composition of tomato in both primary and secondary metabolites and shown the wide diversity present among tomato accessions and their wild relatives (Tikunov, 2005; Schauer et al., 2006; Tikunov et al., 2013; Tieman et al., 2017; Zhu et al., 2018). Considerable genetic variation exists in tomato for micronutrients with antioxidant activity or other health conferring properties (Hanson et al., 2004; Schauer et al., 2005). A number of these micronutrients, particularly carotenoids, have long been major objectives of breeding programs because of their contribution to the quality of fresh and processed tomato products. Increased recognition of their health promoting properties has stimulated new research to identify loci that influence their concentration in tomato.

Table 1.1 Average tomato fruit nutritional value and composition

Nutrient	Unit	Value per 100 g	Cherry (17 g)	Large (182 g)	Medium (123 g)	Small (91 g)
Water	g	94.52	16.07	172.03	116.26	86.01
Energy	kcal	18	3	33	22	16
Protein	g	0.88	0.15	1.6	1.08	0.8
Total lipid (fat)	g	0.2	0.03	0.36	0.25	0.18
Carbohydrate, by difference	g	3.89	0.66	7.08	4.78	3.54
Fiber, total dietary	g	1.2	0.2	2.2	1.5	1.1
Sugars, total	g	2.63	0.45	4.79	3.23	2.39
Minerals						
Calcium, Ca	mg	10	2	18	12	9
Iron, Fe	mg	0.27	0.05	0.49	0.33	0.25
Magnesium, Mg	mg	11	2	20	14	10
Phosphorus, P	mg	24	4	44	30	22
Potassium, K	mg	237	40	431	292	216
Sodium, Na	mg	5	1	9	6	5
Zinc, Zn	mg	0.17	0.03	0.31	0.21	0.15
Vitamins						
Vitamin C, total	mg	13.7	2.3	24.9	16.9	12.5
Thiamin	mg	0.037	0.006	0.067	0.046	0.034
Riboflavin	mg	0.019	0.003	0.035	0.023	0.017
Niacin	mg	0.594	0.101	1.081	0.731	0.541
Vitamin B6	mg	0.08	0.014	0.146	0.098	0.073
Folate, DFE	µg	15	3	27	18	14
Vitamin B12	µg	0	0	0	0	0
Vitamin A, RAE	µg	42	7	76	52	38
Vitamin A, IU	IU	833	142	1516	1025	758
Vitamin E	mg	0.54	0.09	0.98	0.66	0.49
Vitemin K	µg	7.9	1.3	14.4	9.7	7.2
Lipids						
Fatty acids, total saturated	g	0.028	0.005	0.051	0.034	0.025
Fatty acids, monounsaturated	g	0.031	0.005	0.056	0.038	0.028
Fatty acids, polyunsaturated	g	0.083	0.014	0.151	0.102	0.076

(adapted from USDA: https://www.usda.gov/)

Vitamin A and vitamin C are the principal vitamins in tomato fruit, and potassium the main mineral. Tomatoes also provide moderate levels of folate in the diet and lesser amounts of vitamin E and several water-soluble vitamins. β-carotene is a pro-vitamin A carotenoid. Carotene biosynthesis in tomato has been deciphered and many genes and mutations identified (Ronen et al., 1999). More than 20 genes that influence the type, amount, or distribution of fruit carotenoids have been characterized in tomato (Labate et al., 2007).

Vitamin C pathway in plants has been deciphered by Smirnoff and Wheeler, (2000). The variation in ascorbic acid content may depend on varieties and growth conditions (Gest et al., 2013) and a few QTL controlling the variation of Vitamin C have been identified (Stevens et al., 2007). The biosynthetic pathway of folate is also well characterized and the genes involved identified (Almeida et al., 2011). One of the major QTL controlling its variation has been shown to be due to epigenetic variation (Quadrana et al., 2014).

Glycoalkaloids and their toxic effects are commonly associated with Solanaceous species. Tomato accumulates the glycoalkaloids α-tomatine and dehydrotomatine which are less toxic than those present in potato (Madhavi and Salunkhe, 1998; Milner et al., 2011). Several genes controlling their variations have been identified (Cárdenas et al., 2016; Zhu et al., 2018).

Flavonoids comprise a large group of secondary plant metabolites and include anthocyanins, flavonols, flavones, catechins, and flavonones (Harborne and Williams, 2000). Numerous efforts have focused on manipulation of transgene expression to enhance fruit flavonoids (Muir et al., 2001; Colliver et al., 2002; Bovy et al., 2002). Willits et al. (2005) identified a wild accession that expressed structural genes of the anthocyanin biosynthetic pathway in the fruit peel and fruit flesh. Introgression of the *S. pennellii* accession into tomato produced progeny that accumulated high levels of quercetin in fruit flesh and peel. The mutation responsible for the lack of accumulation of yellow color flavonoid in the pink tomato has been identified (Adato et al., 2009; Ballester et al., 2016). Phenolic acids form a diverse group. Hydroxycinnamic acid esters of caffeic acid predominate in Solanaceous species and chlorogenic acid is the most abundant (Mølgaard and Ravn, 1988). Rousseaux et al. (2005) noted large environmental interactions for fruit antioxidants and identified several QTL for total phenolic concentration in fruit of *S. pennellii* introgression lines.

Tomato mineral composition is greatly influenced by plant nutrition, and as a result, has been characterized in the context of mineral deficiency and the effect of these conditions on plant

health. There is significant genotypic variation for mineral content in tomato fruit. Potassium, together with nitrate and phosphorous, constitutes approximately 93% of the total inorganic fruit constituents (Davies and Hobson, 1981). High throughput metabolic profiling allowed getting insight on the whole metabolic changes in tomato fruits during fruit development or in various genotypes (Overy et al., 2004; Schauer et al., 2005; Baxter et al., 2007).

1.1.2.2 Sensory quality

Fresh-market tomato breeders improved yield, disease resistances, adaptation to growth conditions, fruit aspect, but have lacked clear targets for improving organoleptic fruit quality. Consumers have complained about tomato taste for years (Bruhn et al., 1991). Nevertheless improving sensory fruit quality is complex as it is determined by a set of attributes, describing external (size, color, firmness) and internal (flavor, aroma, texture) properties.

Organoleptic quality is often described as a combination of taste, aroma and smell, appearance and texture (**Figure 1.2**). Flavor is mostly due to sugars and organic acids (Stevens et al., 1977), to their ratio (Stevens et al., 1979; Bucheli et al., 1999), and to the composition in volatile aromas (Klee and Tieman, 2013). Sweetness and acidity are related to sugars and acids content (Malundo et al., 1995). Sweetness seems to be more influenced by the content in fructose than in glucose, while acidity is mostly due to the citric acid, present in higher content than malic acid in mature fruits (Stevens et al., 1977). Depending on the studies, acidity is more related to the fruit pH or to the titratable acidity (Baldwin et al., 1998; Auerswald et al., 1999). Both sugars and acids contribute to the sweetness and to the overall aroma intensity (Baldwin et al., 1998).

Texture traits are more difficult to relate to physical measures or to fruit composition, although firmness in mouth is partly related to instrumental measure of fruit firmness (Causse et al., 2002) and mealiness was found related to the texture parameters of the pericarp (Verkerke et al., 1998). Several studies intended to identify the most important characteristics for consumer preferences (Causse et al., 2010).

Processing tomato has specific quality attributes. The self-pruning mutation (*sp*), characteristic of all the processing varieties, controls the determinate growth habit of tomato plants. Processing cultivars associate the *sp* mutation with concentrated flowering, fruit firmness and resistance of mature fruits to over-ripening, allowing a unique mechanical harvest. The *sp* gene was cloned (Pnueli et al., 1998). This mutation does not only affect plant

architecture, but also modulates the expression of genes controlling fruit weight and composition (Stevens, 1986; Fridman et al., 2002; Quinet et al., 2011). This gene belongs to a gene family which is composed of at least six genes (Carmel-Goren et al., 2003). Recently, this gene was also shown to be responsible for the loss of day-length-sensitive flowering (Soyk et al., 2017a). The jointless mutations, provided by the *j* and *j2* genes, are also useful to processing tomato production. The *j2* mutation has been discovered in a *S. cheesmaniae* accession, and has no abscission zone in fruit pedicel allowing harvest without calyx and pedicel during vine pick-up (Mao et al., 2000; Budiman et al., 2004).

Although production of high quality fruits is dependent on environmental factors (light and climate) and cultural practices (irrigation, nutrition), a large range of genetic variation has been shown, which could be used for breeding tomato quality as earlier reviewed (Davies and Hobson, 1981; Stevens, 1986; Dorais and Papadopoulos, 2001). Preferences of consumers faced to genetic variability have rarely been studied. Causse et al. (2003) showed the importance of flavor and secondarily of texture traits in consumer appreciation. Cherry tomatoes have been identified as a source of flavor (Hobson and Bedford, 1989), with fruits rich in acids and sugars. On the contrary, long shelf life cultivars have been described as generally less tasty than traditional ones (Jones, 1986), with lower volatile content (Baldwin et al., 1991). Furthermore quality has a subjective component and there is not a unique expectation (Causse et al., 2010).

Wild relatives of *S. lycopersicum* may be interesting for improving fruit composition. Mutations of enzymes involved in the carbon metabolism were found in *S. chmielewskii and in S. habrochaites*, leading to particular sugar compositions: The *sucr* mutation in an invertase gene, in S. *chmielewskii*, provides fruits with sucrose instead of glucose and fructose (Chetelat et al., 1995). In *S. habrochaites*, an allele of the ADP glucose pyrophosphorylase enzyme was identified as much more efficient than the allele of the cultivated species, leading to an increase in the final sugar content of the fruit (Schaffer et al., 2000). Another locus *Fgr* modulates the fructose-glucose ratio in mature fruit, a *S. habrochaites* allele yielding higher ratio (Levin et al., 2000). The gene responsible is a sugar transporter of the SWEET family (Shammai et al., 2018). A gene *Lin5* encoding an apoplastic invertase has been shown to be a QTL modulating sugar partitioning, the allele of *S. pennellii* leading to higher sugar concentrations than the *S. lycopersicum* one (Fridman et al., 2000). Wild tomato species may also provide original aromas, either favorable to tomato quality

(Kamal et al., 2001) or unfavorable (Tadmor et al., 2002). Several genes responsible for the variation of aroma production in tomato have been cloned (**Table 1.2**) (Klee, 2010; Bauchet et al., 2017b; Zhu et al., 2018).

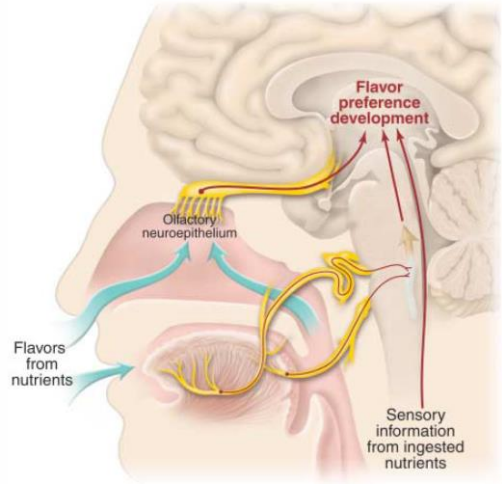


Figure 1.2 Taste and olfactory sensory stimulation are integrated with a variety of sensory inputs including visual, tactile, and nutrient-sensing from gastrointestinal tract to generate the flavor preference and aversions (adapted from Goff and Klee, 2006)

Many efforts for improving fruit quality have failed because of the complex correlations between the various components or between yield or fruit weight and fruit quality components. The correlation between fruit weight and sugar content is frequently negative (Causse et al., 2001), but may be positive in other samples (Grandillo and Tanksley, 1996a). Correlations were also detected between fruit size and antioxidant composition (Hanson et al., 2004). Answering to the demand of producers and retailers of fresh-market tomatoes, breeders have considerably improved external aspect and shelf life of tomato fruit. This improvement was obtained either by the use of the ripening mutations or by the cumulative effect of several genes improving fruit firmness. Several mutations affecting fruit ripening are known, *rin* (ripening inhibitor) the most widely used, *nor* (non ripening), and *alc* (alcobaca). Long shelf life cultivars have invaded the tomato market in the 90's, but consumers have criticized their flavor (Jones, 1986; McGlasson et al., 1987). The corresponding genes have been identified and extensively studied (Vrebalov et al., 2002; Ito et al., 2017; Wang et al., 2019). The impact of the enzymes involved in cell wall modifications during ripening on fruit firmness and shelf life has been extensively studied and modifications of polygalacturonase or pectin methyl esterase activity were proposed to increase fruit shelf life and texture properties (Hobson and Grierson, 1993).

More than 400 volatiles have been identified in tomato fruit (Petró-Turza, 1986), a few of them contributing to the particular aroma of tomato fruit (Baldwin et al., 2000; Tieman et al., 2017). Among the centrally involved metabolites, a set of volatiles play an important role in the overall liking of tomato (**Table 1.3**). Among the most important volatiles, those with a relative high concentration might play a less important role compared to those with lower concentration but with high odor threshold (**Table 1.3**). For example, hexanal has the highest concentration, which is about 27 folds higher than 6-methyl-5-hepten-2-one. However, the odor threshold of 6-methyl-5-hepten-2-one is 400 times higher than hexanal.

Among the essential volatiles, some are common to different fruits, such as blueberry, tomato and strawberry (**Table 1.4**). However, their contributions to consumer preferences might differ and even have completely different effects. The volatiles 1-nitro-2-phenylethane, 1-nitro-3-phenylethane, 1-penten-3-one, 2-phenylethanol, 2-isobutylthiazole, E-2-heptenal, Z-4-decenal, 2,5-dimethyl-4-hydroxy-3(2H)-furanone, 6-methyl-5-hepten-2-ol, phenylacetaldehyde, isovaleric acid, isovaleronitrile and E-3-hexen-1-ol only positively contribute to consumer preferences of tomato, but have no effects on blueberry and strawberry. 6-Methyl-5-hepten-2-one, which is derived from lycopene, has a positive contribution of consumer preference in tomato, but with a negative effect on consumer preference in strawberry and no significant effect in blueberry (**Table 1.4**).

Table 1.2 Cloned genes involved in tomato aroma (adapted from Rothan et al., 2019).

Trait	ITAG gene model	Gene_ID	Function	Chr
Ascorbate	Solyc09g009390	MDHAR	Monodehydroascorbate reductase	9
Benzaldehyde	Solyc08g079270	SlODO1	MYB transcription factor	8
BRIX	Solyc09g010080	Lin5	Invertase	9
Fructose to glucose ratio	Solyc04g064610	FGR	SWEET transporter	4
Geranial	Solyc01g087250	LECCD1A	Carotenoid cleavage dioxygenase	1
Geranial	Solyc01g087260	LECCD1B	Carotenoid cleavage dioxygenase	1
Glucose	Solyc01g079790	AGPL3	ADP-glucose pyrophosphorylase	1
Guaiacol	Solyc09g089580/90	NSGT1	Glycosyltransferase	9
Guaiacol	Solyc10g005060	CTOMT1	Catechol-O-methyltransferase	10
Hexanol & Z-3-hexenol	Solyc09g025210	ADH2	Alcohol dehydrogenase	9
Hexanol and 1-phenylethanol from hexanal and phenylacethaldehyde	Solyc12g056600	SlscADH1	Short chain alcohol dehydrogenase	12
Malic acid	Solyc06g072910 Solyc06g072920	AlMT9	Aluminum-activated malate transporter	6
Methyl salicylate	Solyc09g091550	SlSAMT	Methyl transferase	9
Phenylacetaldehyde	Solyc01g008530; Solyc01g008550	LePAR1;LePAR2	Phenylacetaldehyde reductases	1
Phenylacetaldehyde	Solyc04g064490	SlGT	Glycosyltransferase	4
Phenylacetaldehyde; Phenethylalcohol	Solyc08g068600	LeAADC2	Amino acid decarboxylases	8
Phenylacetaldehyde; Phenethylalcohol	Solyc08g068680	LeAADC1	Amino acid decarboxylases	8

Table 1.3 Volatile compounds and their precursors in two varieties of tomato. Shown are volatile chemicals positively contributing to tomato flavor, as adapted from (Goff and Klee, 2006).

Volatile	Precursor	Concentration (nl/g FW/hour *cerasiforme*)	Concentration (nl/g FW/hour Flora-Dade)	Odor threshold (ppb)
Cis-3-hexenal	Fatty acid	16.28	5.25	0.25
β-ionone	Carotenoid	0.03	0.02	0.007
Hexanal	Fatty acid	27.21	17.15	5
β-Damascenone	Carotenoid	ND	ND	0.002
1-Penten-3-one	Fatty acid	0.21	0.03	1
2-Methylbutanal	Isoleucine	0.75	0.25	1
3-Methylbutanal	Leucine	0.67	0.18	0.2
Trans-2-Hexenal	Fatty acid	0.7	0.26	17
Isobutylthiazole	Unknown	0.32	0.8	3.5
1-Nitro-2-phenylethane	Phenylalanine	0.018	0.013	13
Trans-2-Heptenal	Fatty acid	0.16	0.13	13
Phenylacetaldehyde	Phenylalanine	0.06	0.09	4
6-Methyl-5-hepten-2-one	Carotenoid	0.99	1.84	2000
Cis-3-hexenol	Fatty acid	19.83	13.29	70
2-Phenylethanol	Phenylalanine	0.21	0.32	750
3-Methylbutanol	Leucine	3.83	1.23	120
Methyl salicylate	Phenylalanine	0.08	0.04	40

After the long-term domestication and improvement of tomato, the majority of the essential flavor-associated chemicals have been significantly reduced in modern tomato. For example, the concentration of methional has been significantly reduced by up to approximately 65%. 2-Isobutylthiazole, isovaleric acid and 6-methyl-5-hepten-2-one, which all positively contribute to consumer preference of tomato, have experienced significant decrease in modern tomato.

However, the main synthesis pathways of the essential volatiles in tomato fruits are not that complex and can be mainly subdivided into several pathways, including fatty acids pathway, carotenoid pathway, amino acid pathway and a few others (Goff and Klee, 2006; Klee and Tieman, 2018). These pathways are centrally correlated to several essential primary metabolites (**Figure 1.3**). However, only a few genes involving the synthesis have been cloned and our knowledge on the regulatory mechanisms is still quite limited. Besides, the volatiles are quite sensitive to environmental factors, cultivars, developing stages, measurements, etc. Many of them have a moderate to low heritability (Tieman et al., 2017; Bauchet et al., 2017b). It thus still remains a main breeding challenge to significantly improve multiple volatiles with positive contributions and meanwhile reduce those volatiles with negative contributions (Klee, 2010; Klee and Tieman, 2013; Klee and Tieman, 2018).

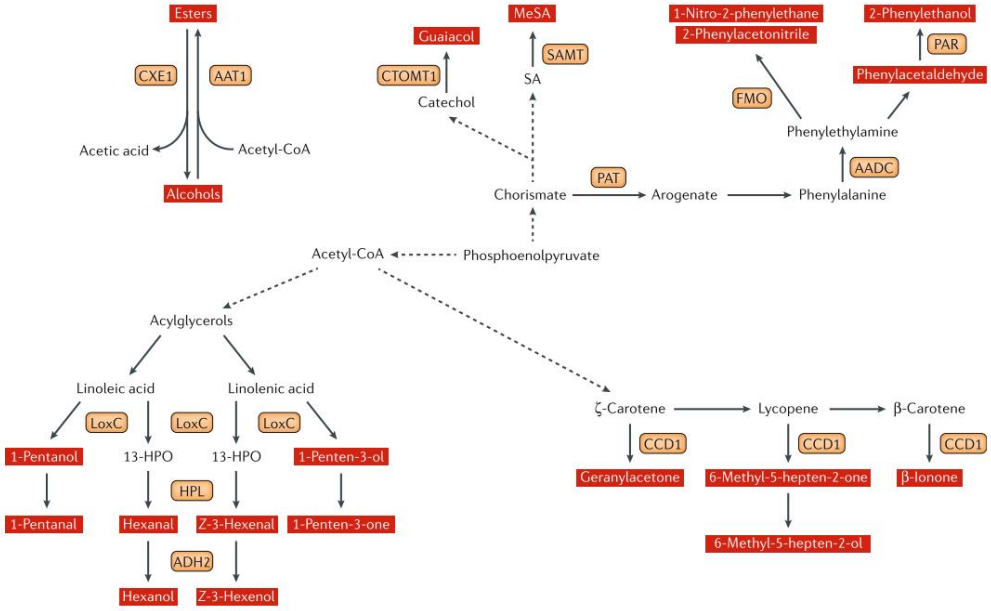


Figure 1.3 Synthesis pathways for tomato flavor volatiles. Solid lines indicate a validated step in a pathway with the responsible enzyme indicated in an orange box. Volatiles are indicated in red. Nonvolatile intermediates are indicated in black. Steps in which the responsible enzyme has not been defined are indicated with dashed lines (adapted from Klee and Tieman, 2018).

Table 1.4 Flavor chemicals and attributes significantly correlated with consumer preferences (adapted from Klee and Tieman, 2018)

Flavor chemicals	Blueberry	Tomato	Strawberry
Flavor intensity	Positive	Positive	Positive
Sweetness	Positive	Positive	Positive
Sugars			
Fructose	Positive	Positive	Positive
Glucose	Positive	Positive	Positive
Sucrose	Positive	NR	Positive
Volatiles			
Hexanol	NSS	NSS	Positive
1-Methylbutylbutyrate	NR	NR	Positive
1,8-Cineole	Negative	NR	NSS
2-Nonanone	Negative	NR	NSS
1-Nitro-2-phenylethane	NR	Positive	NR
1-Nitro-3-phenylethane	NR	Positive	NR
1-Penten-3-ol	NSS	NSS	Negative
1-Penten-3-one	NR	Positive	NR
2,5-Dimethyl-4-methoxy-3(2 H)-furanone	NR	NR	Negative
2,5-Dimethyl-4-hydroxy-3(2 H)-furanone	NR	Positive	NR
2-Ethyl-hexan-1-ol	NR	NR	Positive
2-Heptanone	NSS	NR	Positive
2-Hexanone	NR	NR	Positive
2-Isobutylthiazole	NR	Positive	NR
2-Phenylethanol	NR	Positive	NR
3-Methyl-1-butanol	Positive	NSS	NR
3-Pentanone	NR	Positive	Positive
6-Methyl-5-hepten-2-ol	NR	Positive	NR
3-Ethyloctane	NR	NR	Positive
6-Methyl-2-heptanone	Negative	NR	NR
6-Methyl-5-hepten-2-one	NSS	Positive	Negative
Benzaldehyde	NR	Positive	NR
Benzyl cyanide	NR	Positive	NR
Butyl acetate	NR	Negative	Positive
Butyl butyrate	NR	NR	Positive
Z-2-pentenal	NR	NR	Positive
Z-2-penten-1-ol	Negative	NSS	NSS
Z-4-decenal	NR	Positive	NR
Z-Linalool oxide	NSS	NR	Positive
Decyl butyrate	NR	NR	Positive
Ethyl butyrate	NR	NR	Positive
Ethyl decanoate	NR	NR	Negative
Ethyl propionate	Positive	NR	NSS
Eugenol	NR	Negative	NR
γ-Decalactone	NR	NR	Positive
γ-Dodecalactone	NR	NR	Negative
Heptaldehyde	NSS	Negative	Positive
Hexyl acetate	NSS	Negative	Positive
Hexyl butyrate	NSS	NR	Positive
Isoamyl acetate	Positive	NSS	NSS
Isobutyl acetate	NR	Negative	NR
Isopentyl butyrate	NR	NR	Positive
Isopropyl butyrate	NR	NR	Positive
Methyl anthranilate	NR	NR	Negative
Isovaleraldehyde	Positive	NSS	NR
Isovaleric acid	NR	Positive	NR
Isovaleronitrile	NR	Positive	NR
Linalool	Negative	NR	NSS
Methyl butyrate	NR	NR	Positive
Nerolidol	NR	NR	Negative
Methyl salicylate	Negative	NSS	NR
Nonyl aldehyde	NSS	Positive	Positive
Nonyl 2-methylpropanoate	NR	NR	Positive
Pentyl butyrate	NR	NR	Negative
Phenylacetaldehyde	NSS	Positive	NR
Prenyl acetate	NSS	Negative	NSS
Salicylaldehyde	NR	Negative	NR
S-Methyl thiobutyrate	NR	NR	Positive
E-2-Heptenal	NR	Positive	NR
E-2-Decenal	NR	NR	Positive
E-2-Hexenal	Negative	NSS	Negative
E-2-Hexenyl butyrate	NR	NR	Positive
E-2-Octenal	NR	NR	Positive
E-2-Pentenal	Negative	Positive	Positive
E-3-Hexen-1-ol	NR	Positive	NR

NR, not reported as being present; NSS, present but not significantly correlated with liking

1.2 Genetic resources

Genetic resources for food and agriculture are keys to global food security and nutrition (FAO, 2015). In crop production, maintaining genetic diversity is an essential strategy not only to breed new varieties, to identify candidate genes of target traits, to dissect the evolutionary history, but also to reduce the effects of biotic and abiotic stresses, etc. In this section, we will mainly focus on introducing the genetic resources that are available, which will provide us what genetic resources we have been collected and maintained.

1.2.1 Origin of tomato and its wild relatives

Tomato belongs to the large and diverse *Solanaceae* family also called Nightshades, which includes more than three thousand species. Among them, major crops arose from Old world (eggplant from Asia) and New world (pepper, potato, tobacco, tomato from South America). The *Lycopersicon* clade (**Table 1.5**) contains the domesticated tomato (*Solanum lycopersicum*) and 12 close wild relative species which can be crossed with cultivated tomato (Peralta et al., 2005). Charles Rick and colleagues started the first prospections and studies on the tomato wild relatives in the 40's.

Tomato clade species are originated from the Andean region, including Peru, Bolivia, Ecuador, Colombia and Chile. Their growing environments range from sea level to 3,300 m altitude, from arid to rainy climate and from Andean Highlands to the coast of Galapagos Islands**.** Their habitats are often narrow and isolated valleys and they were adapted to many climates and different soil types. The large range of ecological conditions contributed to the diversity of the wild species. This broad variation is also expressed at the morphological, physiological, sexual and molecular levels (Peralta et al., 2005).

The domestication of tomato is due to a divergence from S. *pimpinellifolium* that occurred several thousand years ago. It probably happened in two steps, first in Peru, leading to S. *lycopersicum cerasiforme* accessions, then in Mexico, leading to large fruit accessions (reviewed in Bauchet and Causse, 2012) as confirmed by molecular analyses (Blanca et al., 2012; Lin et al., 2014; Blanca et al., 2015). Only a few tomato seeds were brought back from Mexico to Europe, leading, after domestication, to a new genetic bottleneck. The tomato cultivation first slowly spread in southern Europe and it is only after the Second World War that its intentional selection started and that it was spread over the world.

Table 1.5 Tomatoes and their wild relative species of the Lycopersicon section according to Peralta et al. 2008 'Lycopersicon group' correspond to the red- and orange-fruited species). For further details of crossability and other biological parameters of wild tomatoes see Grandillo et al. (2011).

Species	**Distribution**	**Habitat;(elevational range**	**Section according to Peralta et al. (2008)**
Solanum lycopersicum L.	Globally cultivated domesticate	Cultivated; sea level-4000 m	*Lycopersicon* 'Lycopersicon group'
Solanum pimpinellifolium L.	Southwestern Ecuador to northern Chile	Dry slopes, plains and around cultivated fields; sea level-3000 m	*Lycopersicon* 'Lycopersicon group'
Solanum peruvianum L.	Central Peru to northern Chile	Dry coastal deserts and lomas; sea level-3000 m	*Lycopersicon* 'Eriopersicon group'
Solanum cheesmaniae (L.Riley) Fosberg	Galápagos Islands	Dry, open, rocky slopes; sea level-1300 m	*Lycopersicon* 'Lycopersicon group'
Solanum galapagense S.C.Darwin & Peralta	Galápagos Islands	Dry, open, rocky slopes; seashores; sea level-1600 m	*Lycopersicon* 'Lycopersicon group'
Solanum arcanum Peralta	Northern Peru	Dry inter-Andean valleys and in coastal lomas (seasonal fog-drenched habitats); 100-4000 m	*Lycopersicon 'Arcanum group'*
Solanum chmielewskii	Southern Peru and northern Bolivia	Dry inter-Andean valleys, usually on open, rocky slopes; often on roadcuts; 1200-3000 m	*Lycopersicon 'Arcanum group'*
Solanum neorickii D.M.Spooner, G.J.Anderson & R.K.Jansen	Southern Ecuador to southern Peru	Dry inter-Andean valleys; 500-3500 m	*Lycopersicon 'Arcanum group'*
Solanum chilense (Dunal)Reiche	Coastal Chile and southern Peru	Dry, open, rocky slopes; sea level-4000 m	*Lycopersicon 'Eriopersicon group'*
Solanum corneliomulleri J.F.Macbr.	Southern Peru (Lima southwards)	Dry, rocky slopes; 20-4500 m (low elevation populations associated with landslides in southern Peru)	*Lycopersicon 'Eriopersicon group'*
Solanum habrochaites S.Knapp & D.M.Spooner	Andean Ecuador and Peru	Montane forests, dry slopes and occasionally coastal lomas; 10-4100 m	*Lycopersicon 'Eriopersicon group'*
Solanum huaylasense Peralta	Río Santa river drainage, north-central Peru	Dry, open, rocky slopes; 950-3300 m	*Lycopersicon 'Eriopersicon group'*
Solanum pennellii Correll	Northern Peru to northern Chile	Dry slopes and washes, usually in flat areas; sea level-4100 m	*Lycopersicon 'Neolycopersicon group'*

1.2.2 Genetic resources as sources for adaptation

There are more than 83,000 tomato accessions stored in different seed banks worldwide (FAO, 2015). These seed banks include the Tomato Genetic Resources Center (TGRC) in Davis, USA (https://tgrc.ucdavis.edu/), the United States Department of Agriculture (USDA) in Geneva, USA (https://www.ars.usda.gov/), the World Vegetable Center in Taiwan, China (https://avrdc.org/), the Centre for Genetic Resources, in the Netherlands (https://www.wur.nl/en/Research-Results/Statutory-research-tasks/Centre-for-Genetic-Resources-the-Netherlands-1.htm) and others. These seed banks maintain most of the genetic diversity of tomatoes.

Thanks to the pioneer work of Charles Rick, the Tomato Genetics Resource Center of the University of California, in Davis, maintains the largest collection of wild relative accessions that he prospected during his life. This collection has been an important source of diversity for breeding tomato and for gene discovery. For instance, there is a collection of 46 *Solanum pennellii* that is only found in Peru, and is particularly adapted to dry conditions (**Figure 1.4**).

1.2.3 Natural and induced mutants

Natural genetic diversity is the main source for adaptation and crop breeding. Natural mutations appeared in cultivated accessions or were introduced from wild relative species, which provide a great source of genetic diversity for many traits, including disease resistance genes and quality trait-related genes (Bauchet and Causse, 2012; Bauchet et al., 2017a; Rothan et al., 2019). However, the number of cloned genes with detailed functional validations is still limited (Rothan et al., 2019). Some biotechnology tools such as TILLING (Targeting Induced Local Lesions in Genomes; Comai and Henikoff, 2006) provide collections of mutants in a specific accession, accelerating functional genomic research and the discovery of interesting alleles at a given locus (Menda et al., 2004; Baldet et al., 2007; Okabe et al., 2011; Mazzucato et al., 2015; Gauffier et al., 2016). This technology typically uses chemical mutagens such as ethyl methanesulfonate (EMS) to generate several base mutations in the genome. There are several TILLING collections worldwide for tomato, such as the UCD Genome Center TILLING laboratory, University of California, USA (http://tilling.ucdavis.edu/index.php/TomatoTilling); The Microtom collection (Okabe et al., 2011); TOMATOMA database, Japan (http://tomatoma.nbrp.jp/); Repository of Tomato Genomics Resources (RTGR), University of Hyderabad, India

(https://www.uohyd.ac.in/images/index.html); The Genes That Make Tomatoes (http://zamir.sgn.cornell.edu/mutants/index.html); the Tilling Platform of Tomato, INRA, France (http://www-urgv.versailles.inra.fr/tilling/tomato.htm) (Minoïa et al., 2010); LycoTILL database, Metapontum Agrobios, Italy (http://www.agrobios.it/tilling/) (Minoia et al., 2010) and others.

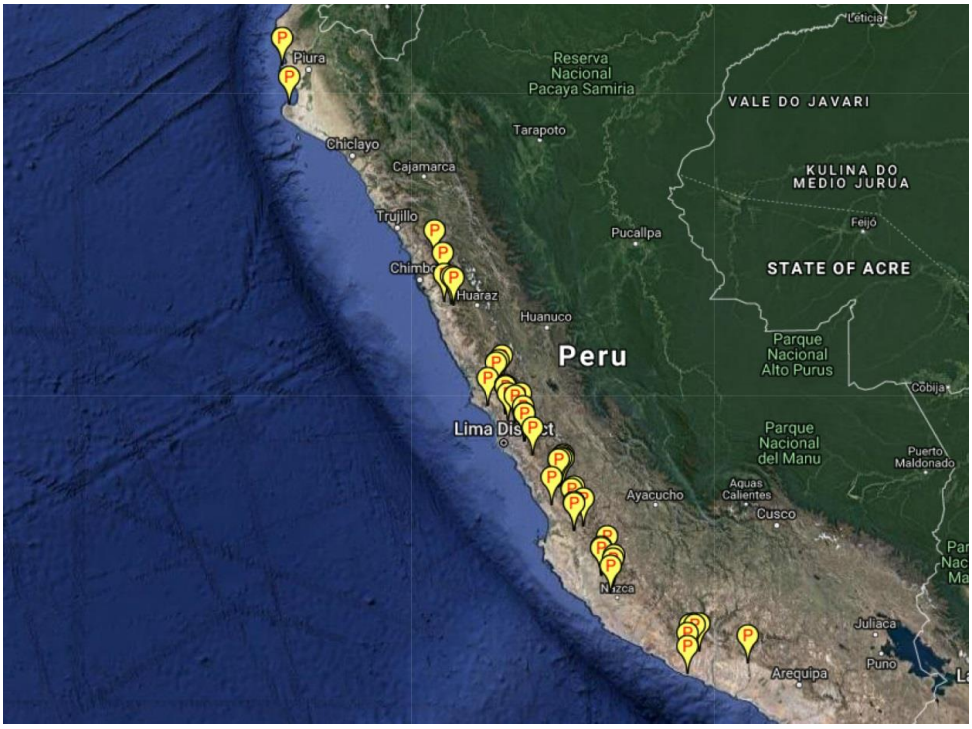


Figure 1.4 Geographical locations of wild tomato species *Solanum pennellii*. Data were collected from Tomato Genetics Resource Center, University of California, Davis (https://tgrc.ucdavis.edu/Data/Acc/Wildspecies.aspx).

1.3 Genomic resources

Genomic information greatly promoted our understanding of the genetic architecture and evolutionary history of modern tomato. In this section will focus on the genomic resources that are available. We believe these resources will greatly promote the genetic, genomic studies of tomato. The tomato genome sequencing project was initiated as part of the International Solanaceae Project (SOL), which was launched on November 3, 2003 at Washington, USA and gathered a consortium of scientists of 10 countries including China, France, Spain, Italy, USA, UK, the Netherlands, Japan, Korea and India (Mueller *et al.*, 2005). The main reason why tomato was first chosen as the reference genome for the Solanaceae was due to its high level of macro and micro-synteny among the species, its relatively small genome (900 Mb) and its economic importance. This project was first started with conventional sequencing technologies, such as Sanger sequencing. In order to reduce the cost of producing a high-quality reference, BAC-by-BAC sequencing strategy based on saturated genetic markers was used to select seed BACs within the gene-rich parts of the tomato genome for sequencing. However, this process was quite slow and became a serious obstacle, which was then greatly accelerated by next-generation sequencing (Pietrella and Giuliano, 2016).

The first tomato genome sequence was published in 2012 for the inbred tomato cultivar 'Heinz 1706' (*S. lycopersicum*) together with a draft of its closest wild species *S. pimpinellifolium* (accession LA1589) (The Tomato Genome Consortium, 2012). In the tomato genome, recombination, genes and transcripts are substantially located in the euchromatin regions compared to the heterochromatin regions, whereas chloroplast insertions and conserved microRNA genes were more evenly distributed throughout the genome (The Tomato Genome Consortium, 2012).

The tomato genome was highly syntenic with other Solanaceae species, such as pepper, eggplant, potato and *Nicotiana*. Tomato had fewer high-copy, full-length long terminal repeat retrotransposons with older insertion ages compared to *Arabidopsis* and Sorghum. Genome annotation showed that there were a total 34,727 protein-coding genes and 30,855 of them were supported by RNA sequencing data. Chromosomal organization of genes, transcripts, repeats and sRNAs were very similar between tomato and potato. Among all the protein-coding genes, 8615 genes were common to tomato, potato, *Arabidopsis*, rice and grape. A total of 96 conserved sRNAs were predicted in tomato, which could be further divided into

34 families, 10 of which being highly conserved in plants. The potato genome showed only 8% divergence from tomato, with nine large and several smaller inversions (The Tomato Genome Consortium, 2012). The *Solanum* lineage has experienced one ancient and one more recent consecutive genome triplications. The genome information provides a basic understanding of the genetic bottlenecks that narrowed tomato genetic diversity (The Tomato Genome Consortium, 2012).

Since the first published version, the sequence has been completed with new sequencing methods, corrected and re-annotated using new sequence data and new RNAseq data and the genome version today is SL4.0 while the annotation is ITAG4.0 (http: solgenomics.net) (**Table 1.6**).

Table 1.6 Highlights of the current tomato genome version SL4.0 and annotation ITAG4.0, adapted from https://solgenomics.net/organism/Solanum_lycopersicum/genome

Build highlights SL4.0	**Annotation highlights ITAG4.0**
Only 44Kb of N's (unknown bases) compared to 81.7Mb in SL3.0	34,075 protein coding genes in ITAG4.0
Only 152 unplaced contigs in Chr 00 compared to 4,374 in SL3.0	Functional descriptions assigned to 29,532 genes
Better annotation of repeat regions in SL4.0	ITAG4.0 has 4,794 novel genes
80X Pacbio coverage with RSII and Sequel (13kb read N50)	29,281 genes preserved from ITAG2.3
Canu assembly (N50 5.5 Mb) and Hi-C scaffolding (12 chromosomes and unplaced contigs)	21,962 of 29,281 preserved genes have been updated
Validated with Bionano optical maps and 10X linked reads	Most of the updated genes have extensions in the 5' and 3' UTRs

1.4 Genetic diversity analyses

Genetic diversity refers to the total genetic characteristics of a species with a population. According to the United Nations Food and Agriculture Organization, 75% of all crop genetic diversity has been lost since the previous century, and one third of the remaining is expected to become extinct by 2050. For instance, according to the National Center for Genetic Resources Preservation, only a few of the commercial seeds that are common in 1903 were still reserved at the preservation center, including tomato and cucumber (**Figure 1.5**). Conserving and managing the genetic diversity of crops is imperative to adapt to changing variables, especially pest and environmental stressors, in order to feed the increasing global population.

Tomato, one of the most important vegetables, belongs to the Solanaceae family. The wild relatives of cultivated tomatoes are native to western South America with diverse habitats (Blanca et al., 2012). Molecular genetic markers play an important role in the modern breeding (Ramstein et al., 2019). They also provided a new vision of tomato genetic diversity, compared with isozyme markers (Bauchet and Causse, 2012). Overall, modern cultivated tomato accessions present a lower polymorphism level compared to wild species (*S. pimpinellifolium*), as shown by different types of markers, such as RFLP (Miller and Tanksley, 1990), AFLP (Suliman-Pollatschek et al., 2002; Park et al., 2004; Van Berloo et al., 2008; Zuriaga et al., 2009), RAPD (Grandillo and Tanksley, 1996b; Archak et al., 2002; Tam et al., 2005; Carelli et al., 2006; El-hady et al., 2010; Meng et al., 2010; Length, 2011), SSR (Suliman-Pollatschek et al., 2002; Jatoi et al., 2008; Mazzucato et al., 2008; Albrecht et al., 2010; Meng et al., 2010; Sim et al., 2010; Zhou et al., 2015), ISSR (Vargas-Ponce et al., 2011; Shahlaei et al., 2014) and SNPs (Blanca et al., 2012; Sim et al., 2012a; Lin et al., 2014; The 100 Tomato Genome Sequencing Consortium, 2014). The phylogenetic studies consistently revealed that *S. pimpinellifolium*, cherry (*S. l. cerasiforme*) and the cultivated tomato (*S. lycopersicum*) are consistently clustered together, compared with other wild relatives. The phylogeny of *Solanum* sect. *Lycopersicon* with other closely related Solanum species is provided in **Figure 1.6**.

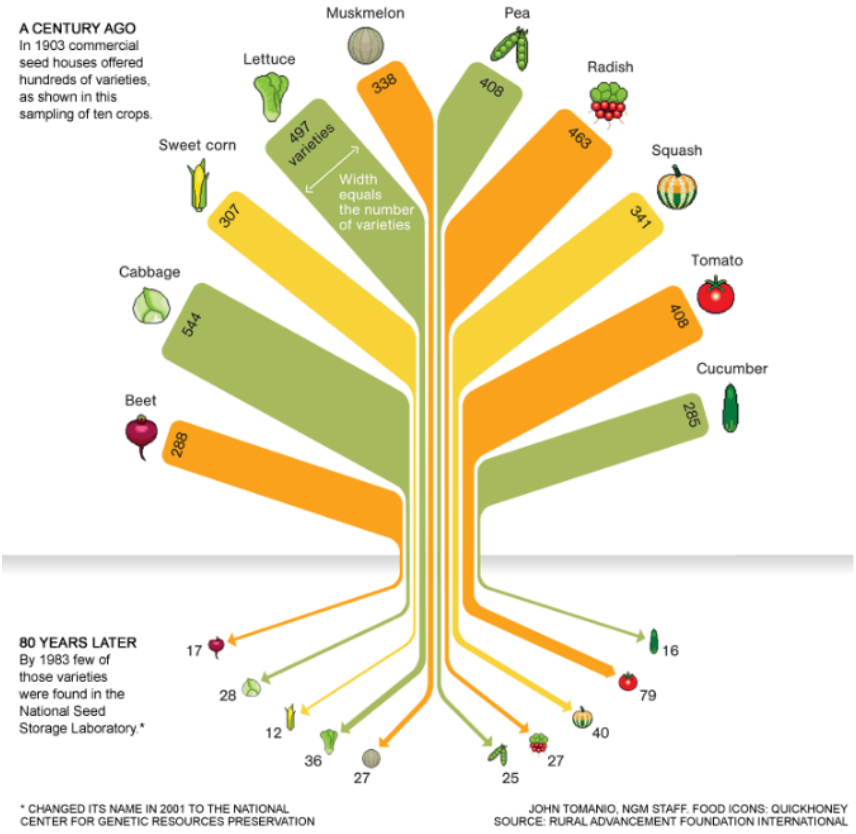


Figure 1.5 Examples of genetic diversity losses in the past century based on the data from the National Center for Genetic Resources Preservation (adapted from https://medium.com/thenextnorm/importance-of-genetic-diversity-in-agriculture-b9f88f5fda55).

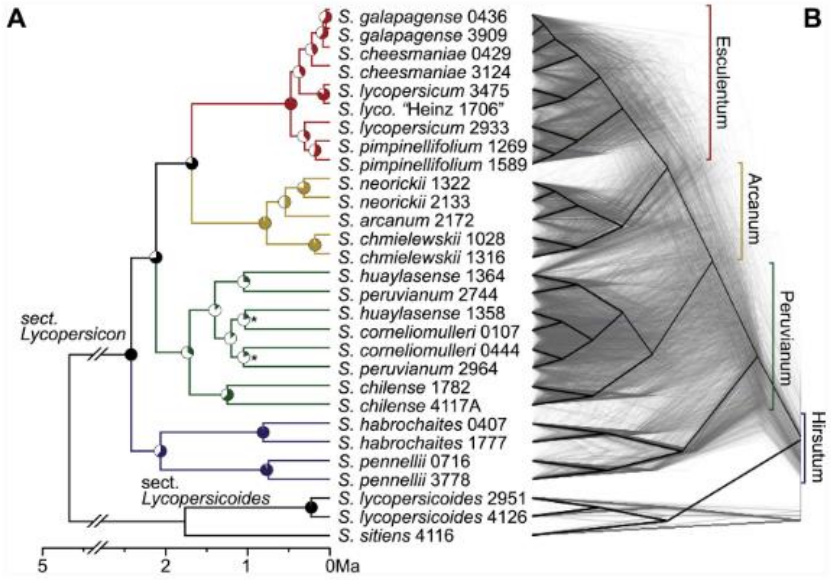


Figure 1.6 The phylogeny of *S. Lycopersicon*. (A) A whole-transcriptome concatenated molecular clock phylogeny with section *Lycopersicoides* as the outgroup. (B) A "cloudogram" of 2,745 trees (grey) inferred from nonoverlapping 100-kb genomic windows (adapted from Pease et al., 2016).

1.4.1 Use of genetic diversity to decipher the origin of tomato

Cherry tomatoes, genetically regarded as the admixture of *S. pimpinellifolium* and *S. lycopersicum* (Ranc et al., 2008; Blanca et al., 2012; Lin et al., 2014). Ranc et al. (2008) revealed that the groups of cherry tomatoes could be further divided into two groups, one being the admixture of *S. pimpinellifolium* and *S. lycopersicum* and the remaining were more genetically close to *S. lycopersicum*. In addition, cherry and wild tomato accessions inhabited strikingly different ecological and climatic regions and a clear relationship was found between the population structure and a geographic map based on the climatic classification (**Figure 1.7**). In addition, principal component analysis revealed that the main *S. pimpinellifolium* group also showed a clear substructure that was consistent segregation within the coastal and montane accessions (**Figure 1.8**). Population structure also revealed that the substructure of *S. pimpinellifolium* and cherry groups was linked to geography, which was consistent with PCA (Blanca et al., 2012). Within the groups of cherry tomatoes, a prominent geographic division was also observed between Ecuador, northern Peru, southern Peru, Mesoamerica and non-American. In contrast, the traditional varieties globally were quite homogeneous (Blanca et al., 2012).

1.4.2 Use of genetic diversity to study the genome dynamic

Linkage disequilibrium (LD) refers to the non-random association of alleles at two or more loci in a general population. In general, the LD in cultivated tomato accessions was larger than that of wild species, which could be up to about 20 Mbp, while cherry tomatoes ranged in between (Mazzucato et al., 2008; Van Berloo et al., 2008; Sim et al., 2010; Ranc et al., 2012; Xu et al., 2013; Sauvage et al., 2014; Zhang et al., 2016; Bauchet et al., 2017a).

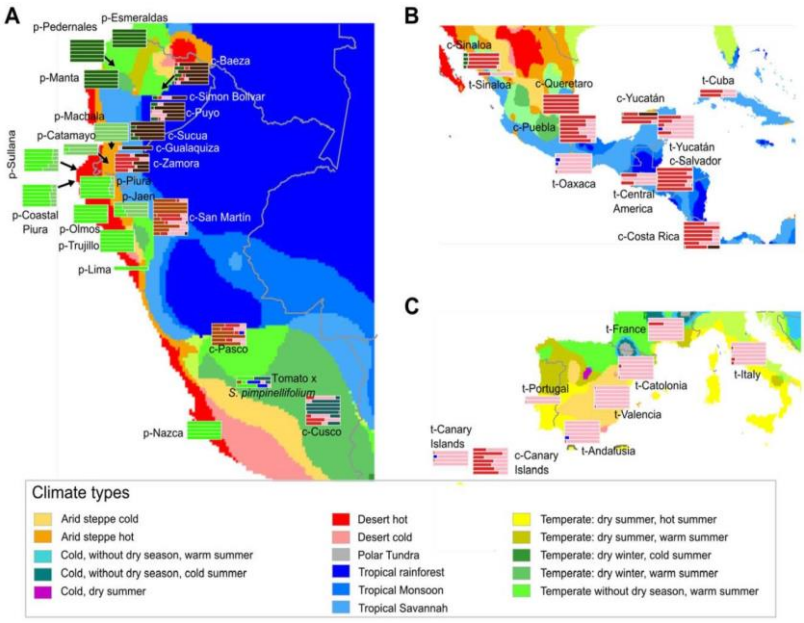


Figure 1.7 Geographical distributions of the population structure revealed by SOLCAP SNPs. Different colored bars represent the proportion of the population structure. The ancestries calculated by the Structure analysis are clustered by geographical group and represented at the corresponding geographical location. The different colors of the geographical background correspond to the climatic classification (adapted from Blanca et al., (2012).

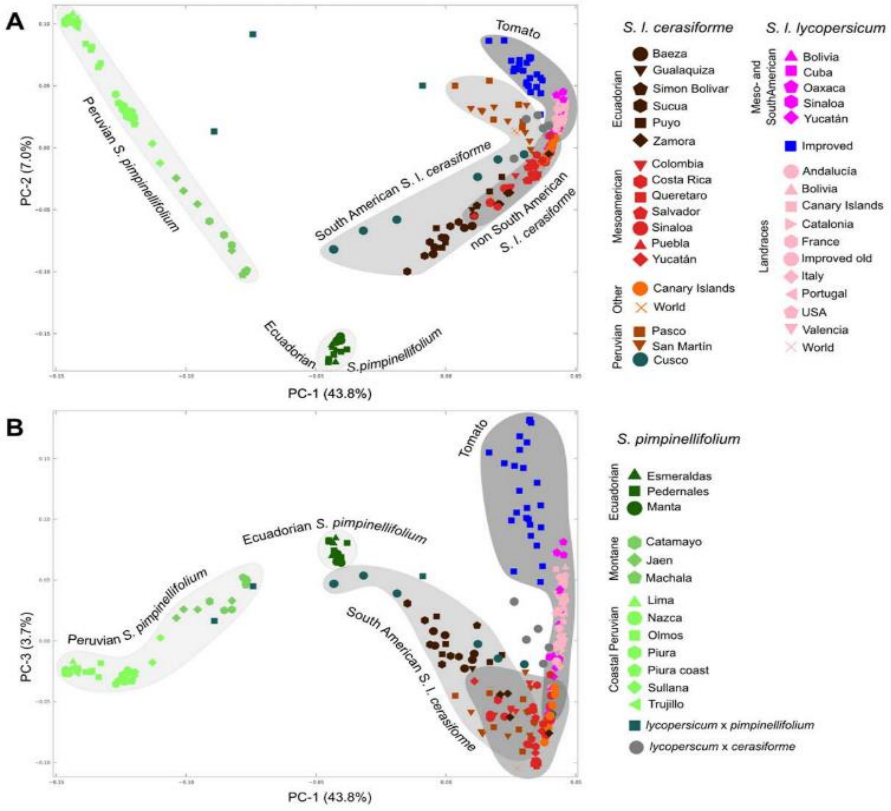


Figure 1.8 PCA analysis of all samples. In panel A the projection along the first and second principal components of the PCA analysis carried out with the SNP genotypes is represented (adapted from Blanca et al., (2012).

Admixture of cherry tomatoes with modern cultivars and wild species could help reduce the large LD and overcome the low resolution of association mapping of modern tomato cultivars (Ranc et al., 2012). The LD based on pairwise r^2 using both the local weighted scatterplot smoothing (LOESS) and non-linear regression (NLR) was different between different representative tomato sub-populations, including processing, fresh market and vintage tomato collections. The largest LD was found on chromosome 3 (11.8 cM for NLR) and the shortest distance on chromosome 12 (1.9 cM for LOESS) (**Table 1.7**). By taking the populations structure into account in estimating the LD (r_s^2), the LD of S.L. (large tomatoes) on every chromosome was always the largest compared to S.C. (cherry) and S.P (the closest wild species) and the shortest LD was always observed in S.P for all 12 chromosomes (**Table 1.8**). All these results revealed a similar trend that the genetic diversity of tomato has been consistently reduced during the long-term domestication and improvement. However, introgression of disease resistance genes into modern breeding had a big influence on the LD (Bauchet et al., 2017a) (**Table 1.9**).

In the tomato genome, there are more genes at the beginning and ending of the chromosome than the center of chromosomes (The Tomato Genome Consortium, 2012). The recombination patterns for all the chromosomes were in general similar where most of the recombination events were occurred at the beginning and the endings (**Figure 1.9**). Those regions with fewer recombinations usually have a relatively larger LD, as recombination events break down the strong linkage.

1.4.3 Use of genetic diversity for breeding purposes

The strong LD can be beneficial in reducing the minimum number of SNPs to cover the whole genome, especially for genome-wide association studies (GWAS). However, at the same time, it will also reduce the resolution and cause challenges in regional fine-mapping to identify the causal variants. Cherry tomatoes serve as the phenotypic and genotypic mosaic between large domesticated tomatoes wild species and are helpful to bridge the gaps between low genetic diversity and high morphological diversity of modern cultivated tomato accessions and wild species. Re-introducing those genes that have been lost during domestication and improvement, such as via introgression and genome editing could be helpful to break down the relatively large LD and also to increase the genetic diversity of modern tomatoes.

Table 1.7 Chromosome by chromosome linkage disequilibrium (LD) analysis within three representative sub-populations of cultivated tomatoes (adapted from Sim et al., 2012).

	LD decay (cM)											
	95th percentile method[1]						Fixed method (r^2 = 0.2)					
	Processing		Fresh market		Vintage		Processing		Fresh market		Vintage	
Chr	LOESS[2]	NLR[3]	LOESS	NLR	LOESS	NLR	LOESS	NLR	LOESS	NLR	LOESS	NLR
1	6.6	4.1	9.3	7.7	9.9	9.4	7.2	5.3	7.7	3.6	7.2	3.8
2	5.4	14.2	4.1	10.2	5.8	6.6	6.0	18.7	3.3	4.6	3.0	2.7
3	6.3	4.4	7.2	12.1	8.6	11.8	6.9	5.7	6.1	5.4	6.1	4.7
4	4.6	6.9	3.8	7.4	5.8	9.9	5.2	9.0	2.9	3.5	4.4	4.1
5	3.5	3.1	6.1	7.7	3.8	4.7	3.9	4.1	4.3	3.5	2.5	1.9
6	7.1	35.2	6.0	20.1	5.8	6.0	7.4	40.6	5.0	9.7	4.1	2.8
7	4.0	4.0	6.0	8.3	4.9	2.8	5.0	6.4	4.4	3.8	4.1	1.1
8	7.7	27.9	5.7	5.7	6.7	6.3	7.7	36.7	3.3	2.8	4.4	2.5
9	3.5	5.0	9.0	9.5	7.8	6.5	4.7	6.6	6.9	4.7	4.1	2.7
10	5.2	2.2	5.8	17.7	5.8	3.6	6.0	2.8	4.2	8.0	3.6	1.4
11	0.8	0.8	47.6	38.0	6.0	n.d.	0.8	1.1	19.7	17.2	4.9	41.3
12	6.7	5.5	4.6	4.1	1.9	1.1	6.7	6.7	2.2	1.9	1.6	0.5
Average	5.0	9.8	10.1	13.1	6.4	6.8	5.5	12.5	6.2	6.1	4.4	6.3

[1]The estimate of the 95th percentile baseline r2 value in each germplasm group was 0.23 in the processing varieties, 0.12 in the fresh market varieties, and 0.11 in the vintage varieties.
[2]Logally weighted scatterplot smoothing [66].
[3]Non-linear regression [67]. For NLR, the expected r2 was calculated using the model of Hill and Weir (1988).
n.d. = not determined.

Table 1.8 Intra chromosomal LD in each group (adapted from Sauvage et al., 2014).

Mean Pairwise r_s^2	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	K11	K12	All K
S.L	0.5508	0.5988	0.5895	0.5570	0.6029	0.6235	0.5416	0.5397	0.5231	0.5539	0.5938	0.5389	0.5678
S.C	0.5391	0.5318	0.5394	0.5191	0.5500	0.5530	0.5320	0.5337	0.5204	0.5315	0.5619	0.5117	0.5353
S.P	0.3323	0.3239	0.2884	0.3872	0.3557	0.2604	0.3478	0.3338	0.3431	0.3923	0.2917	0.3968	0.3378

Table 1.9 Linkage disequilibrium: mean LD (non-linear regression threshold r^2 = 0.1) values according to genetic groups and chromosomes (adapted from Bauchet et al., 2017).

Genetic pool	All chr	Chr01	Chr02	Chr03	Chr04	Chr05	Chr06	Chr07	Chr08	Chr09	Chr10	Chr11	Chr12
S. lycopersicum	0.192	0.184	0.210	0.242	0.157	0.197	0.230	0.187	0.185	0.187	0.151	0.183	0.193
Modern admixture	0.170	0.156	0.162	0.128	0.163	0.202	0.187	0.172	0.131	0.187	0.174	0.156	0.219
Old admixture	0.166	0.168	0.181	0.122	0.147	0.323	0.168	0.115	0.113	0.151	0.116	0.220	0.166
S. pimpinellifolium	0.095	0.073	0.083	0.074	0.085	0.119	0.079	0.076	0.135	0.061	0.066	0.061	0.232
Whole set	0.190	0.176	0.206	0.153	0.211	0.228	0.133	0.190	0.163	0.179	0.231	0.214	0.191

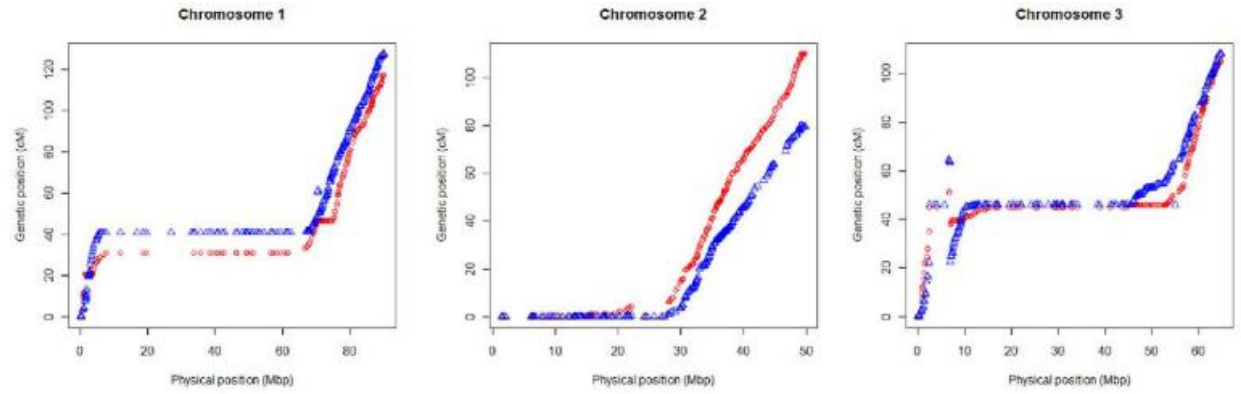


Figure 1.9 Relationship between genetic and physical positions within chromosome 1-3. The genetic positions of SNP markers are indicated by red circles for the EXPEN 2012 population and blue triangles for the EXPIM 2012 population (adapted from Sim et al., 2012; relationship for each chromosome is available within the paper).

1.5 Selection footprints at the genome level

In this section, we will overview the impacts of nature selections (positive or negative) and the most advanced approaches to detect selective signals. Since the most commonly applied models in detecting selective footprints in major crops are quite limited, we also provide detailed explanations about the recently developed statistical models based on humans, especially those composed of multiple signals. We think these most-advanced models will be interesting and helpful in promoting our understandings on what, where and when selection has happened in the genomes of crops, especially in tomato. This knowledge is useful to understand domestication and improvement of tomato, which in turn, will assist tomato quality breeding.

Natural selection tends to increase the frequency of beneficial allele in a population over time (**Figure 1.10**). Those individuals harnessing beneficial traits have higher fitness. In the genomic era, selection refers to any non-random, differential propagation of an allele as a consequence of its phenotypic effect (Vitti et al., 2013). Identifying candidate variants under selection not only demonstrates evolution and shed light on species history but also represent biologically meaningful insights (Vitti et al., 2013). For instance, the first adaptive trait studied in humans was the disorder of red blood cells that is distributed in regions where malaria was endemic (Haldane, 2006). It was further proved that the sickle cell mutation in the *Hemoglobin-B* gene (*HBB*) was responsible for the selection for malaria resistance (Allison, 1954). At the early stage of evolutionary genetics, examples of natural selection were mainly elucidated in adaptive traits using a forward genetic approach, such as for lactase persistence and skin pigmentation in humans (Tishkoff et al., 2007) or armored plates in stickleback fish (Jones et al., 2012). Nowadays, advancements in genomic technology make it possible to use both the forward and backward genetic approaches.

Figure 1.10 Genetic signatures of positive selection (adapted from Scheinfeldt and Tishkoff, 2013)

1.5.1 Different types of selection signals

Selection can be briefly divided into positive (favored selection) and negative selection (disfavored selection or purifying selection). Random mutations are more likely to be deleterious than beneficial and are more likely to be removed from the gene pool before they can achieve detectable frequency within the population (Vitti et al., 2013). Background selection is a form of negative selection of the ongoing of removal of deleterious mutations. Balancing selection is the selection in which multiple alleles are maintained at an appreciable frequency within the gene pool. Stabilizing selection is the selection when intermediate phenotypic values are favored by balancing selection of codominant alleles or by positive selection of alleles underlying intermediate phenotypes (Vitti et al., 2013). Among these types of selections, much research in recent years has focused on detection of positive selection signals. From the practical perspective, positive selection leaves a more conspicuous footprint on the genome compared to negative selection and balancing selection. In contrast, negative selection is mainly observed in highly conserved regions and balancing selection's effect is often subtle (Vitti et al., 2013).

1.5.2 Approaches to detect selection

Methods to detect selection signals can be divided into macro-evolutionary level and micro-evolutionary level. Macro-evolutionary level selections are typically detected by comparing the homologous sequences among related taxa (**Figure 1.11**). Methods for macro-evolutionary selections can be further divided into gene-based methods and other rate-based methods, such as Hudson-Kreitman-Aguadé (HKA) test (Hudson et al., 1987; Wright and Charlesworth, 2004) and identification of accelerated regions (Pollard et al., 2006; Shapiro and Alm, 2008; Lindblad-Toh et al., 2011).

At the micro-evolutionary level, positive selection causes a beneficial allele to high prevalence or fixation, thereby causing a population-wide reduction in genetic diversity (**Figure 1.12**). Recent positive selection can be found with three different signals: high levels of allele differentiation between populations, high frequency of the derived allele and long haplotypes (Karlsson et al., 2014). Main methods to detect selection signals fall broadly into four categories: frequency-based methods (such as Tajima's D and derivatives, Fay & Wu's H), linkage disequilibrium-based methods (such as LRH, iHS, XP-EHH and IBD), population differentiation-based methods (such as LKT, LSBL and hapFLK) and composite methods (such as CLR, XP-CLR) (Vitti et al., 2013). An overview of common approaches for detecting selection is provided in **Table 1.10**.

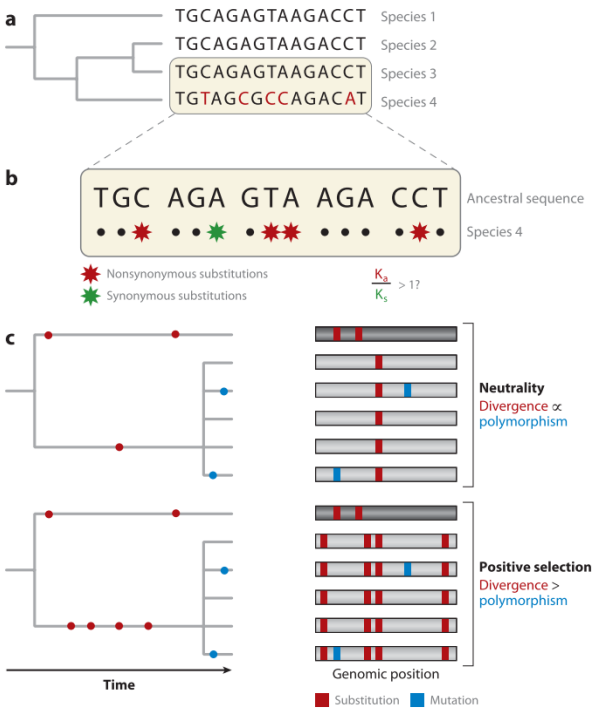


Figure 1.11 Methods for detecting selection at the macro-evolutionary level (adapted from Vitti et al. 2013)

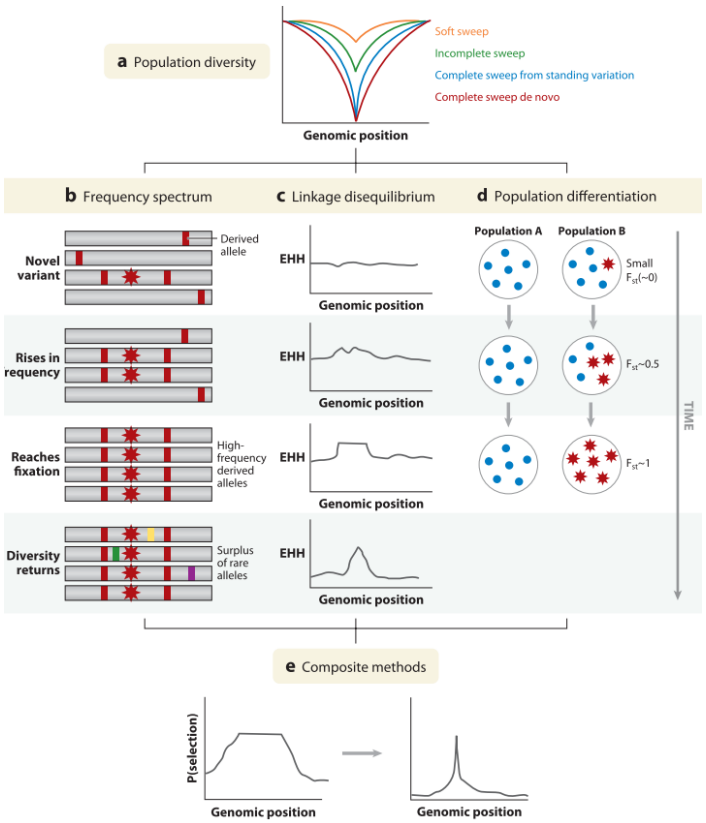


Figure 1.12 Methods for detecting selective sweeps at the micro-evolutionary level (adapted from Vitti et al. 2013).

Table 1.10 An overview of common approaches for detecting selection (adapted from Vitti et al. 2013)

	Approach	Intuition	Representative tests	References
Methods for macroevolution	Gene-based methods	Synonymous substitutions are (assumed to be) selectively neutral. Thus, they tell us about the background rate of evolution. If the rate of nonsynonymous substitution differs significantly, it is suggestive of selection.	K_a/K_s (also referred to as d_N/d_S or ω)	(43, 60)
			McDonald-Kreitman test (MKT)	(27, 78)
	Other rate-based methods	Levels of polymorphism and divergence should be correlated (because both are primarily functions of the mutation rate) unless selection causes one to exceed the other.	Hudson-Kreitman-Aguadé (HKA) test MKT	(59, 135)
		Regions that undergo accelerated change in one lineage but are conserved in related lineages are probable candidates for selection.	Identification of accelerated regions	(14, 77, 100, 102, 116)
Methods for microevolution	Frequency-based methods	In a selective sweep, a genetic variant reaches high prevalence together with nearby linked variants (high-frequency derived alleles). From this homogenous background, new alleles arise but are initially at low frequency (surplus of rare alleles).	Ewens-Watterson test	(30, 133)
			Tajima's D and derivatives	(38, 39, 122, 123)
			Fay & Wu's H	(33)
	Linkage disequilibrium–based methods	Selective sweeps bring a genetic region to high prevalence in a population, including the causal variant and its neighbors. The associations between these alleles define a haplotype, which persists in the population until recombination breaks these associations down.	Long-range haplotype (LRH) test	(111, 141)
			Long-range haplotype similarity test	(52)
			Integrated haplotype score (iHS)	(131)
			Cross-population extended haplotype homozygosity (XP-EHH)	(113)
			Linkage disequilibrium decay (LDD)	(132)
			Identity-by-descent (IBD) analyses	(15, 50)
	Population differentiation–based methods	Selection acting on an allele in one population but not in another creates a marked difference in the frequency of that allele between the two populations. This effect of differentiation stands out against the differentiation between populations with respect to neutral (i.e., nonselected) alleles.	Lewontin-Krakauer test (LKT)	(11, 31, 73, 129)
			Locus-specific branch length (LSBL)	(117)
			hapFLK	(32)
	Composite methods	Combining test scores for multiple sites across a contiguous region can reduce the rate of false positives.	Composite likelihood ratio (CLR)	(67, 68, 87, 89)
			Cross-population composite likelihood ratio (XP-CLR)	(22)
		Combining multiple independent tests at one site can improve resolution and distinguish causal variants. Different tests can provide complementary information.	DH test	(138, 139)
			Composite of multiple signals (CMS)	(44, 45)

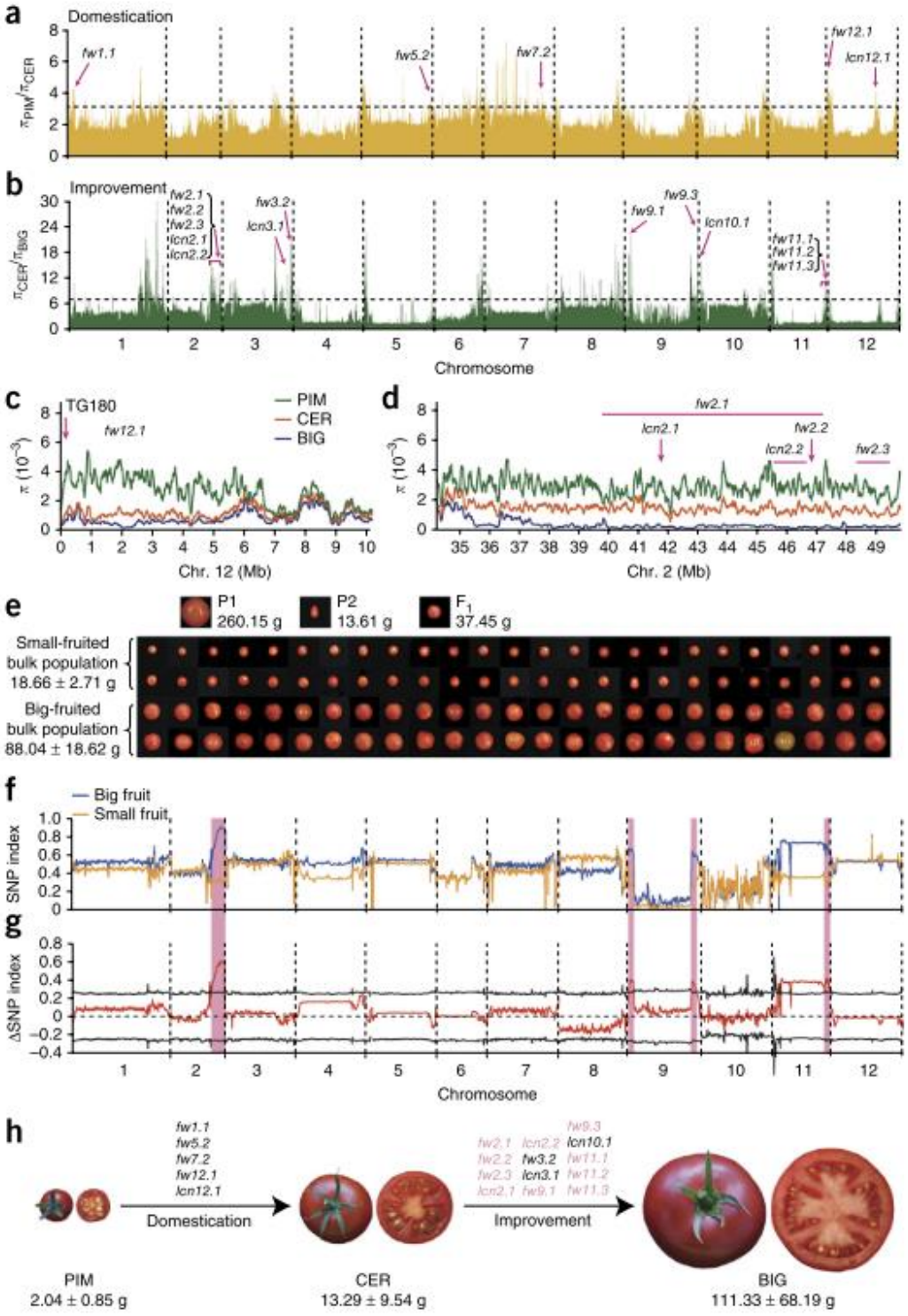


Figure 1.13 Evolution of fruit mass during domestication and improvement (adapted from Lin et al. 2014). (**a,b**) Domestication and improvement sweeps. (**c,d**) Distribution of nucleotide diversity of the PIM (green), CER (orange) and BIG (blue) lines within the domestication sweeps harboring *fw12.1* (**c**) and within the improvement sweep harboring five fruit mass QTLs on chr2 (**e**). (**e-g**) Verification of the improvement sweeps related to fruit mass. (**h**) Schematic of the two-step evolution of tomato fruit size. QTLs that were putatively selected during domestication and improvement are listed, and those in pink were verified in this study.

1.5.3 Selective sweeps in tomato

In tomato, the selection signals were mainly detected by analysis of nucleotide diversity (π) and Fst (Städler et al., 2012; Lin et al., 2014; Tieman et al., 2017; Zhu et al., 2018). The nucleotide diversity of wild *S. pimpinellifolium* (PIM) group was substantially higher than that of the *S. lycopersicum cerasiforme* (CER) and *S. lycopersicum* (BIG) groups. By comparing the nucleotide diversity of PIM and CER (π_{PIM}/π_{CER}; domestication sweeps) and CER and BIG (π_{PIM}/π_{BIG}; improvement sweeps), a total of 186 domestication sweeps and 133 improvement sweeps were identified, covering 8.3% (64.6 Mb) and 7.0% (54.5 Mb) of the tomato genome (Lin et al., 2014). Notably, 21% of the domestication sweeps overlapped with improvement sweeps, indicating that some of the domestication loci might have undergone a second round of selection for further improvement of fruit weight. The enlargement of tomato fruit mass was well explained by several major QTLs located within the domestication and improvement sweeps (**Figure 1.13**). Among these, there was a major improvement sweep on chromosome 2, where five major fruit weight QTLs were located, including two cloned QTL *fw2.2* and *lcn2.1*. These results demonstrated that nucleotide diversity (π) is a good parameter to dissect the domestication and improvement sweeps.

Lin et al. (2014) also investigated the divergence in big-fruit tomatoes based on population differentiation statistic (F_{ST}) for 122 modern processing accessions and 144 BIG accessions and identified a non-random distribution of highly divergent sites, especially on chromosome 5 (**Figure 1.14**). Three SSC QTLs (Tanksley et al., 1996) and one fruit firmness QTL (Xu et al., 2013) were previously reported on chromosome 5. These results indicated that selection for higher SSC and better fruit firmness likely hitchhiked almost of the entire chromosome 5.

In a recent tomato pan-genome study, Gao et al., (2019) showed that genomes of wild accessions carried significantly more genes than those of CER accessions, and that genomes of the BIG group had the lowest number of genes, indicating a general trend of gene loss during tomato domestication and improvement. By treating genes with higher frequencies in CER than PIM and BIG than CER as possible favorable genes, a total of 120 favorable and 1213 unfavorable genes were identified during domestication and 12 favorable and 665 unfavorable genes were identified during improvement stage. Enrichment analysis showed that defense response was the most enriched group of unfavorable genes during both stages, especially those genes related to cell wall thickening. A rare promoter allele in the promoter region of *TomLoxC* demonstrated a very good example of strong negative selection during

both domestication and improvement, which was essential for many C5 and C6 volatiles in tomato fruit (Chen et al., 2004; Shen et al., 2014).

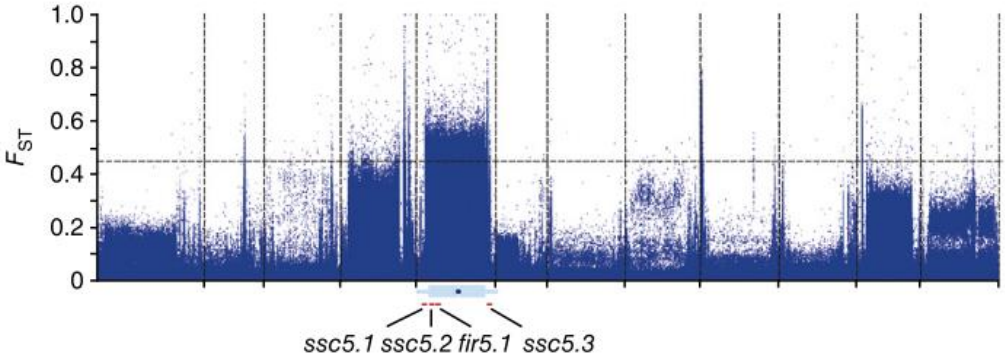


Figure 1.14 A major genomic signature of modern processing tomatoes and three causative variants for pink fruit (adapted from Lin et al. 2014)

1.5.4 Methods to detect selection signals and its application challenges

Several methods were proposed to detect multiple selection signals. Natural selection leaves a number of footprints on the genome, and each single test only detects a slightly different signal. Grossman et al., (2010) developed CMS (composite of multiple signals) method to combine tests for multiple selection signal detection, which could increase the resolution by up to 100-fold both in simulations and real data (**Figure 1.15**). Alachiotis and Pavlidis, (2018) recently proposed another program, called RAiSD (raised accuracy in sweep detection) that composed allele diversity, site frequency spectrum and the linkage disequilibrium (LD) in the region of a sweep and was mainly designed to detect hard selective sweeps (the classical selective sweep model in which a new advantageous mutation arises, and spreads quickly to fixation due to natural selection) (Maynard Smith and Haigh, 1974; Vitti et al., 2013).

Akbari et al., (2018) developed iSAFE (integrated selection of allele favored by evolution) to identify the favored mutation in a positive sweep. iSAFE outperformed CMS in improving the rankings, while CMS had excellent performance in localizing the favored mutations (**Figure 1.16**). In addition, iSAFE scores are normalized and the results from different populations are comparable. However, one limitation of iSAFE is its deteriorated performance in identifying the favored mutation when it was fixed or near fixation (Akbari et al., 2018). However, CMS requires a control population as well as a demographic model, in addition to the target population under selection (Grossman et al., 2010) and, a high depth genotyping is required when applying iSAFE.

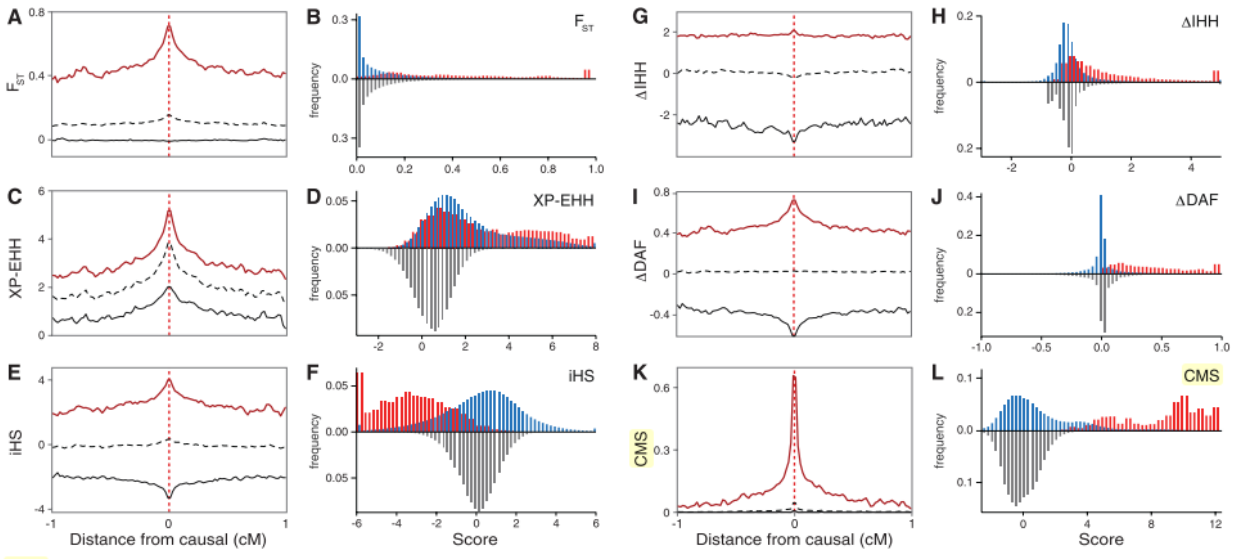


Figure 1.15 Comparison of different models in identifying selective signals (adapted from Grossman et al. 2010).

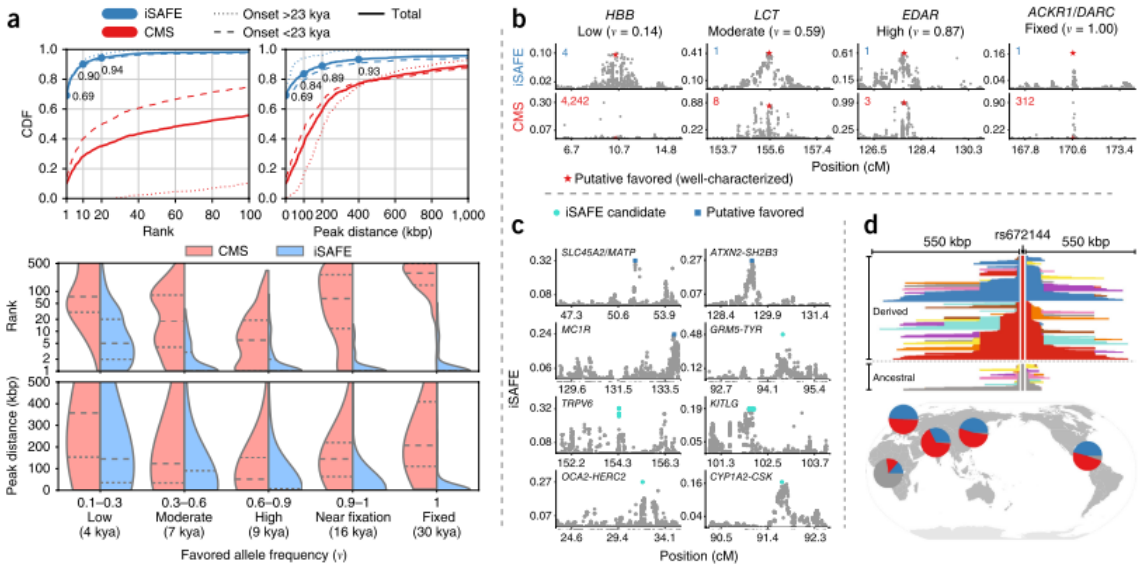
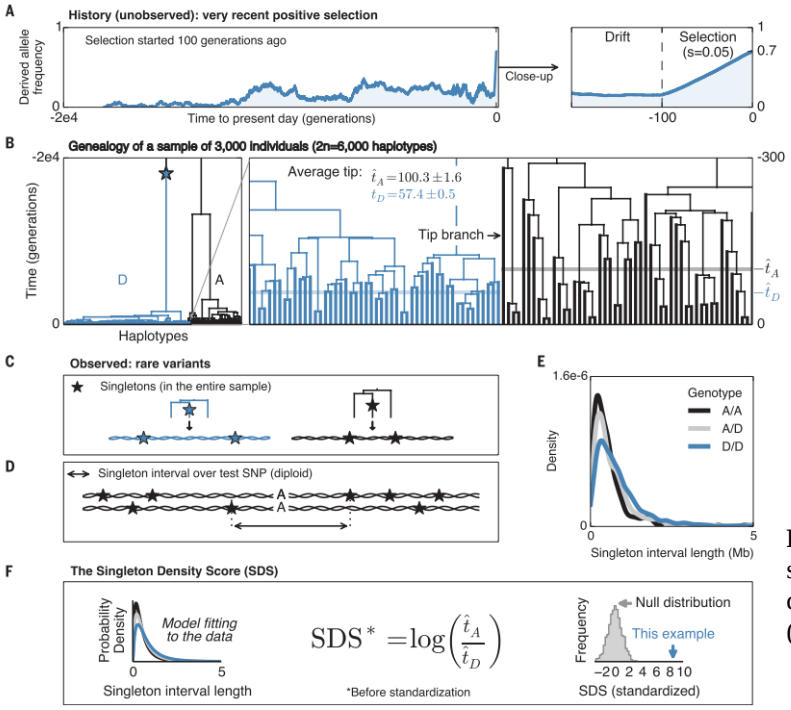


Figure 1.16 iSAFE performance. (a) The cumulative distribution function (CDF) of the favored mutation rank (top left) and peak distance (top right) for iSAFE and CMS scores. Bottom: rank and peak distance distributions of the favored mutation as a function of favored allele frequency (ν) in the target population (EUR). In the bottom plot, the dashed (dotted) line represents the median (quartiles). (b) iSAFE and CMS scores of four well-characterized selective sweeps. The rank of the putative favored mutation in the 5-Mbp region is shown in the top left corner in each plot. (c) iSAFE scores for regions under selection. Top-ranked iSAFE candidates that match reported favored mutations ("putative favored") or are newly suggested by iSAFE ("iSAFE candidate") are indicated. All datasets consisted of a 5-Mbp window around the selected region, unless one side reached the telomere or centromere. (d) The GRM5–TYR region. The mutation rs672144 was ranked first by iSAFE and is very well separated from other mutations in the surrounding 5 Mbp, in all non-African populations, with high confidence (adapted from Akbari et al., 2018).

Field et al. (2016) introduced SDS (singleton density score) to infer very recent selective sweeps in human genome by comparing the ancestral and derived haplotypes (**Figure 1.17**). SDS was more powerful in detecting selection signals within 100 generations compared to iHS (integrated haplotype score). However, iHS was always more powerful than SDS after 100 generations of selection (**Figure 1.18**). It could be interesting to apply SDS in tomato, especially when the studied samples mainly consist of large-fruit tomatoes, which are still undergoing human selections for quality improvement.

Zeng et al., (2018) recently proposed a Bayesian mixed linear model (BayesS) that could distinguish negative selections ($S < 0$) from positive selections ($S > 0$) when the trait-associated variants have pleiotropic effect (**Figure 1.19**).

Tomato has undergone long-term selection during domestication and improvement processes, during which, fruit weight and biotic/abiotic resistance were among the major breeding targets. However, some important quality traits, such as tomato flavor, sugar contents have been strongly deteriorated in modern large-fruit tomatoes, compared to the wild cherry tomatoes. These results indicated that tomato has undergone both positive and negative selection during the domestication and improvement stages. It is thus of great interest to distinguish negative selection from positive selection sweeps for both the domestication and improvement steps. However, most of these composite models usually require substantially large populations with high quality of in-depth genotyping, which could limit its potential applications in tomato at present stage.



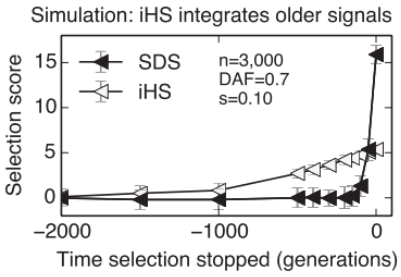


Figure 1.17 Illustration of the SDS method (adapted from Field et al. 2016)

Figure 1.18 Comparison of the selection score of SDS and iHS with different selection generations (adapted from Field et al. 2016)

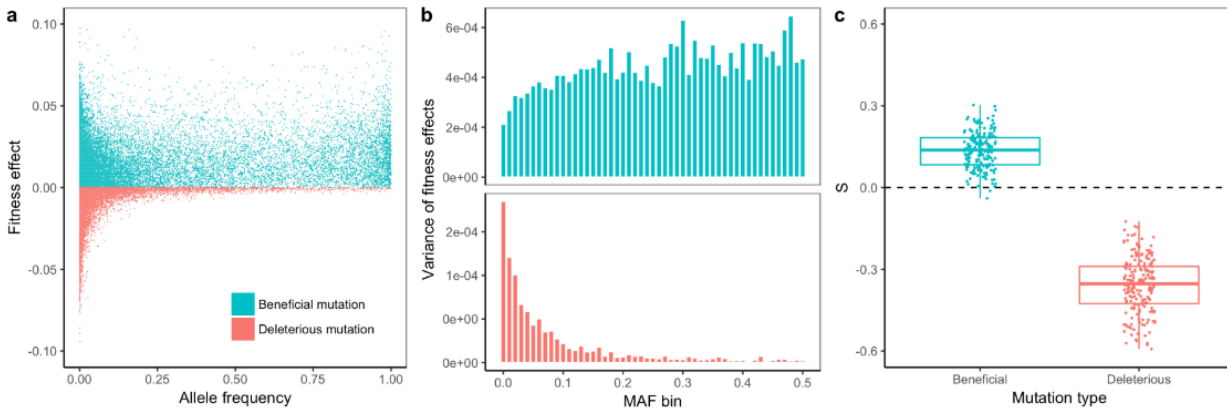


Figure 1.19 Forward simulations with mutations of direct effect on fitness to distinguish negative selections (S <0) from positive selections (S > 0) (adopted from Zeng et al. 2018)

1.6 Molecular markers and gene/QTL mapping

In this section, we will introduce the genetic markers, with a central focus on SNPs and how to develop large datasets of SNPs via SNP arrays and next-generation sequencing (NGS). We will also introduce the landmarks of gene/QTL mapping and also the populations used dating back last century. These achievements lay the foundation of modern tomato breeding and are still important quantitative studies of tomato.

1.6.1 Evolution of molecular markers

Tomato has been used for genetic studies and mutation mapping of interesting traits even before the discovery of molecular markers (Butler, 1952). Genes of interest were first mapped thanks to pairs of near isogenic lines differing only in the region of the interesting gene (Philouze, 1991; Laterrot, 1996). Nevertheless, until the 1980s, the location of mutations of interest on genetic maps was not precise. The first isozyme markers were limited in number and rapidly replaced by restriction fragment length polymorphism (RFLP) markers. The first high-density genetic map based on RFLP markers was constructed (Tanksley et al., 1992). With more than 1000 loci, spread on the 12 chromosomes, it allowed the localization of several mutations and genes of interest. Then, PCR based markers, including RAPD, AFLP and microsatellites, were used, but remained limited in polymorphism level and distribution across the genome. Following the identification of PCR markers linked to the gene of interest, specific PCR markers were set up, simplifying the genotyping step for breeders. Nevertheless, PCR markers such as RAPD or AFLP map in majority close to the centromeres, reducing their potential efficiency for gene mapping in tomato (Grandillo and Tanksley, 1996a; Haanstra et al., 1999; Saliba-Colombani et al., 2001).

1.6.2 SNP markers

1.6.2.1 SNP discovery

Single nucleotide polymorphisms (SNPs) are the most abundant molecular markers for major crops. SNPs can be detected in any region of the genome, including coding sequences or non-coding sequences of genes, as well as the intergenic regions. Only the non-synonymous SNPs in the coding regions of genes change the amino acid sequences of proteins. However, SNPs in the non-coding region are also likely to affect gene expression through different mechanisms (Farashi et al., 2019). Millions of SNPs can be directly generated via

genotyping-by-sequencing (GBS) or resequencing of a few lines (Catchen et al., 2011). Next-generation sequencing-based technologies have also accelerated the identification and isolation of genes associated with agronomic traits in major crops (Nguyen et al., 2018). There are many GBS methods available, including at least 13 reduced-representation sequencing (RRS) approaches and at least four whole-genome resequencing (WGR) approaches (Scheben et al., 2017). Among them, RNA sequencing and exome sequencing based on transcriptome sequences is an important alternative RRS approach (Haseneyer et al., 2011; Scheben et al., 2017). The sequenced data can be used for expression analysis and also does not require prior genomic sequence information (Wang et al., 2010a).

Since the availability of the reference tomato genome, whole-genome resequencing of different tomato accessions has directly generated millions of SNPs, covering the whole tomato genome (Bolger et al., 2014; Lin et al., 2014; Menda et al., 2014; The 100 Tomato Genome Sequencing Consortium, 2014; Tieman et al., 2017; Ye et al., 2017; Zhu et al., 2018). The number of SNPs in the wild tomato species compared to the reference cultivated genome exceeds 10 million, which are 20-folds higher than that in most of the domesticated accessions

1.6.2.2 SNP arrays

SNP array is another popular and cost-effective genotyping approach. Several arrays have been developed in tomato, such as those produced by the Solanaceae Coordinated Agricultural Project (SolCAP) (Hamilton et al., 2012; Sim et al., 2012a), the Centre of Biosystems Genomics (CBSG) consortium (Víquez-Zamora et al., 2013) or by the Diversity Arrays Technology (DArTseq) (Pailles et al., 2017). However, RNA-seq based SNP arrays, such as SolCAP and ddRAD-Seq (Arafa et al., 2017), have some major limitations: Gene expression is dependent on tissue and time, multiple biases are introduced by library preparation during RNA fragmentation (Wang et al., 2009) and SNP density is low in coding regions (Scheben et al., 2017). In tomato, these SNP arrays have been widely used to genotype different tomato collections (Sim et al., 2012a; Víquez-Zamora et al., 2013; Ruggieri et al., 2014; Sauvage et al., 2014; Blanca et al., 2015; Bauchet et al., 2017a; Bauchet et al., 2017b; Pailles et al., 2017; Albert et al., 2016b).

1.6.2.3 Resequenced tomato accessions

Next generation sequencing technologies made it possible to sequence genomes at large scales (Goodwin et al., 2016). Soon after the availability of the reference tomato genome, the genome of the stress-tolerant wild tomato species *S. pennellii* was published (Bolger et al., 2014). This species is characterized by extreme drought tolerance and unusual morphology. Many stress-related candidate genes were mapped in this wild species. Large gene expression differences were observed between *S. lycopersicum* cv. M82 and *S. pennellii* (LA716) due to polymorphisms at the promoter and/or coding sequence levels. This wild species and others were further re-sequenced and assembled using long read sequencing platforms complemented with Illumina sequencing (Usadel et al., 2017).

After the genome of *S. pennellii*, a panel of diversified tomato accessions and related wild species were sequenced (The 100 Tomato Genome Sequencing Consortium, 2014)(The 100 Tomato Genome Sequencing Consortium, 2014)(The 100 Tomato Genome Sequencing Consortium, 2014)(The 100 Tomato Genome Sequencing Consortium, 2014). The allogamous self-incompatible wild species have the highest level of heterozygosity, which was low for the autogamous self-compatible species (The 100 Tomato Genome Sequencing Consortium, 2014). Almost at the same time, a comprehensive genomic analysis based on resequencing 360 tomato accessions elucidated the history of tomato breeding (Lin et al., 2014). This study showed that domestication and improvement of tomato mainly involved two independent sets of QTLs leading to fruit size increase. Five major QTLs (*fw1.1*, *fw5.2*, *fw7.2*, *fw12.1* and *lcn12.1*) contributed to the enlargement of tomato fruit during domestication process. Then, up to 13 major QTLs (*fw2.1*, *fw2.2*, *fw2.3*, *lcn2.1*, *lcn2.2*, *fw3.2*, *lcn3.1*, *fw9.1*, *fw9.3*, *lc10.1*, *fw11.1*, *fw11.2* and *fw11.3*) contributed to the second improvement of tomato fruit size. This study also detected several independent mutations in a major gene *SlMYB12* that changed modern red tomato to pink tomato appreciated in Asia. This study also illustrated the linkage drag associated with wild introgressions (Lin et al., 2014).

Since then, low-depth resequencing or genotyping-by-sequencing has become a common practice and is widely applied in many tomato collections. Up to now, around 900 tomato accessions have been re-sequenced, with the sequence depth ranging from low to high (The Tomato Genome Consortium, 2012; Causse et al., 2013; Bolger et al., 2014; Lin et al., 2014; The 100 Tomato Genome Sequencing Consortium, 2014; Tieman et al., 2017; Ye et al., 2017;

Tranchida-Lombardo et al., 2018). These genomic resources are freely available (https://solgenomics.net) and will greatly facilitate modern breeding of new tomato cultivars.

In a recent pan-genome study comparing the genomes of 725 phylogenetically and geographically representative tomato accessions, a total of 4,873 genes were newly discovered compared to the reference genome (Gao et al., 2019a). Among these, 272 were potential contaminations and were removed from the 'Heinz 1706' reference genome. Substantial gene loss and intensive negative selection of genes and promoters were detected during tomato domestication and improvement. During tomato domestication, a total of 120 favorable (genes with higher frequencies in CER than PIM, or in BIG than CER) and 1213 unfavorable genes (those genes with lower frequencies) were identified, whereas 12 favorable and 665 unfavorable genes were identified during improvement process. Disease resistance genes were especially lost or negatively selected. Gene enrichment indicated that defense response was the most enriched group of unfavorable genes during both domestication and improvement. No significantly enriched gene families were found in favorable genes during improvement. A rare allele in the *TomLoxC* promoter was found under selected during domestication. Taken together with other findings, this pan-genome study provides useful knowledge for further biological discovery and breeding (Gao et al., 2019a).

Recently, the research groups of Michael C. Schatz at Johns Hopkins University and Zachary B. Lippman at Cold Spring Harbor Laboratory released the reference genome of 13 diverse tomato accessions, each with their own independent versioning. These new reference accessions include Brandywine, M82, Floradade, EA00371, EA00990, PAS014479, BGV006775, BGV006865, BGV007989, BGV007931, PI303731, PI169588 and LYC1410 (https://solgenomics.net/projects/tomato13/).

1.6.3 Specific populations to dissect phenotype determinants

Rapidly, molecular breeding strategies were set up and implemented to try to "pyramid" genes and QTL of interest for agronomical traits, notably using Advanced Backcross QTL method (AB-QTL) (Grandillo and Tanksley, 1996a). Using this approach with a *S. lycopersicum x S. pimpinellifolium* progeny, in which agronomical favorable QTL alleles were detected, Grandillo et al. (1996) showed how a wild species could contribute to improve cultivated tomato(Grandillo et al., 1996)(Grandillo et al., 1996)(Grandillo et al., 1996)(Grandillo et al., 1996). Introgression Lines (ILs) derived from interspecific crosses

allowed to dissect the effect of chromosome fragments from a donor (usually from a wild relative) introgressed into a recurrent elite line. ILs offer the possibility to evaluate the agronomic performance of a specific set of QTL (Paran et al., 1995). ILs were used as a base for fine mapping and positional cloning of several genes and QTL of interest. The first IL library was developed between *S. pennellii* and *S. lycopersicum* (Eshed and Zamir, 1995; Zamir, 2001). QTL mapping power was increased compared to biallelic QTL mapping population, and was again improved by the constitution of sub-IL set with smaller introgressed fragments. This population was used in identifying QTLs for fruit traits (Causse et al., 2004); anti-oxidants (Rousseaux et al., 2005), vitamin C (Stevens et al., 2007) and volatile aromas (Tadmor et al., 2002). The introgression of a QTL identified in these IL has allowed plant breeders to boost the content of soluble solids (SSC) in commercial varieties and largely increased tomato yield in California (Fridman et al., 2004). Complementary genetic resources are now available, including a new backcrossed inbred line (BIL) population generated by repeated backcrosses, followed by selfing (Ofner et al., 2016). This BIL population could be used in combination with ILs for fine-mapping QTLs previously identified and to pinpoint strong candidate genes (Fulop et al., 2016). The creation of systematic sub-ILs carrying smaller introgressions, further facilitated the identification of candidate genes (Alseekh et al., 2013). These sub-ILs are available to the scientific community and have been used to map loci affecting fruit chemical composition (Alseekh et al., 2015; Liu et al., 2016a). Such exotic libraries were also designed with other species, involving *S. pimpinellifolium* (Doganlar et al., 2003), *S. habrochaites* (Monforte and Tanksley, 2000; Finkers et al., 2007) and *S. lycopersicoides* (Canady et al., 2005).

Introgression lines were also used to dissect the genetic basis of heterosis (Eshed and Zamir, 1995). Heterosis refers to phenomenon where hybrids between distant varieties or crosses between related species exhibit greater values than both parents (Birchler et al., 2010). Heterosis involves genome–wide dominance complementation and inheritance model such as locus-specific over-dominance (Lippman and Zamir, 2007). Heterotic QTL for several traits were identified in tomato ILs (Semel et al., 2006a). A unique QTL was shown to display at the heterozygous level improved harvest index, earliness and metabolite content (sugars and amino acids) in processing tomatoes (Gur et al., 2010; 2011). Furthermore, a natural mutation in the SFT gene, involved in flowering (Shalit et al., 2009), was shown to correspond to a single over-dominant gene increasing yield in hybrids of processing tomato (Krieger et al., 2010).

1.6.4 Achievements of trait mapping

The construction of genetic maps of molecular markers Tanksley et al. (1992) permitted the dissection of quantitative traits into QTL (Quantitative Trait Loci) since the pioneer work of Paterson et al., (1988). This strategy also opened the way to investigate physical mapping and molecular cloning of genetic factors underlying quantitative traits (Paterson et al., 1991). The first gene cloned by positional cloning was the *Pto* gene, conferring resistance to *pseudomonas syringae* (Martin *et al.* 1993). Since then, several interspecific progenies with each wild relative species were studied. Due to the low genetic diversity within the cultivated compartment (Miller and Tanksley 1990), most of the mapping populations were based on interspecific crosses between a cultivar and a related wild species from the lycopersicon group (as reviewed by Labate et al. (2007), Foolad (2007) and Grandillo et al. (2011) or from lycopersicoides (Pertuzé et al., 2003) and juglandifolia group (Albrecht et al., 2010). However, maps based on intraspecific crosses have proved their interest notably for fruit quality aspects (Saliba-Colombani et al., 2001). All those populations allowed the discovery and characterization of a myriad of major genes (Rothan et al., 2019) and QTLs involved in various traits (Grandillo and Tanksley, 1996b; Tanksley et al., 1996; Fulton et al., 1997; Bernacchi et al., 1998; Chen et al., 1999; Grandillo et al., 1999; Frary et al., 2000; Monforte and Tanksley, 2000; Causse et al., 2001; Saliba-Colombani et al., 2001; Causse et al., 2002; Doganlar et al., 2003; Frary et al., 2004; Schauer et al., 2006; Baldet et al., 2007; Jiménez-Gómez et al., 2007; Cagas et al., 2008; Kazmi et al., 2012; Haggard et al., 2013; Alseekh et al., 2015; Pascual et al., 2015; Ballester et al., 2016; Rambla et al., 2016; Kimbara et al., 2018).

The main results of QTL studies can be summarized as follows:

- QTLs are detected in every case, sometimes with strong effects. A few QTLs explaining a large part of the phenotypic variation, acting together with minor QTLs, are frequently detected. Most of the QTLs act in an additive manner, but a few dominant and even over-dominant QTLs were detected (Paterson et al., 1988; DeVicente and Tanksley, 1993).

- QTLs can be separated in two types: QTLs stable over the environments, years or types of progeny, and QTLs more specific of one condition (Paterson et al., 1991).

- Some regions involved in the variation of a trait are found in progenies derived from different accessions of a species, or from different species (Fulton et al., 1997; Bernacchi et al., 1998; Chen et al., 1999; Grandillo et al., 1999; Fulton, 2002).

- The dissection of complex traits in relevant components and the QTL mapping of these components allowed the genetic bases of the variability of complex traits to be understood. For example, a map of QTLs controlling several attributes of organoleptic quality in fresh-market tomato revealed relations between QTLs for sensory attributes and chemical components of the fruit (Causse et al., 2002). The analysis of biochemical composition of a trait is also important.

- Fine mapping experiments allowed to precisely map the QTLs in a chromosome region and to verify the existence of several QTLs linked in the same region (Paterson et al., 1990; Frary et al., 2003; Lecomte et al., 2004a). For example, by reducing the size of an introgressed fragments from *S. pennellii*, Eshed and Zamir (1995) identified three linked QTLs controlling fruit weight on a single chromosome arm. Fine mapping is also an important step for cloning QTLs, as first shown by the successes in cloning QTLs controlling fruit weight (Alpert and Tanksley, 1996; Frary et al., 2000), fruit shape (Tanksley, 2004) and soluble solid content (Fridman et al., 2000; Fridman et al., 2004).

- Wild species, in spite of their low characteristics in comparison to cultivars, can carry alleles, which may contribute to the improvement of most of the agronomic traits (DeVicente and Tanksley, 1993).

1.6.5 QTL discovery towards cloning of candidate genes

Tomato is probably one of the crops with the largest number of single mutations used for its breeding (as reviewed by Grandillo and Cammareri, 2018, and Rothan et al., 2019). Before the SNP discovery, due to the limited genetic diversity of domesticated tomato accessions, the populations used for linkage mapping have been generated by crosses between a cultivated and a close wild tomato species (Foolad, 2007; Foolad and Panthee, 2012). Since the development of molecular markers, these segregating populations have become an effective and efficient tool to construct high density genetic linkage maps (Tanksley et al., 1992), allowing the detection of Quantitative Trait Loci (QTLs). By using different linkage populations and multiple molecular markers, including RFLP, SSR and SNPs, hundreds of

QTLs have been reported, for different agronomical, morphological, and quality related traits (Grandillo and Tanksley, 1996b; Tanksley et al., 1996; Fulton et al., 1997; Bernacchi et al., 1998; Chen et al., 1999; Grandillo et al., 1999; Fulton et al., 2000; Monforte and Tanksley, 2000; Saliba-Colombani et al., 2001; Causse et al., 2002; Doganlar et al., 2003; van der Knaap and Tanksley, 2003; Fridman et al., 2004; Baldet et al., 2007; Foolad, 2007; Jiménez-Gómez et al., 2007; Cagas et al., 2008; Dal Cin et al., 2009; Sim et al., 2010; Ashrafi et al., 2012; Haggard et al., 2013; Kinkade and Foolad, 2013).

However, among the detected QTLs, only a few have been cloned and functionally validated (Bauchet and Causse, 2012; Rothan et al., 2019). The first gene cloned by positional cloning in tomato was the *Pto* gene, conferring resistance to *Pseudomonas syringae* races, with the assistance of RFLP markers (Martin et al., 1993). Based on the same RFLP map, *Fen*, another member of this gene family, was also soon reported (Martin et al., 1994). From then on, different resistance genes were identified and cloned based on RFLP markers, such as *Cf-2*, a leucine-rich repeat protein conferring resistance to *Cladosopum fulvum* strains (Dixon et al., 1996); *Prf*, another resistance gene to *Pseudomonas syringae* pv. tomato (Pst) strains (Salmeron et al., 1996); *Ve* conferring Verticilium wilt resistance, encoding surface-like receptors (Kawchuk et al., 2001) and others.

Some important genes/QTL involved in developmental processes were also identified and cloned with the assistance of molecular markers. Among them, *fw2.2*, a major QTL controlling tomato fruit weight, was one of the first examples (Frary et al., 2000). It alters tomato fruit size likely by expression regulation rather than sequence and structure variation of the encoded protein (Nesbitt and Tanksley, 2002). Recently, some other major QTLs were functionally validated, such *fw3.2* (corresponding to a cytochrome P450 gene) (Chakrabarti et al., 2013) and *fw11.2* (corresponding to a cell size regulator) (Mu et al., 2017). Some major QTLs related to fruit shape were also reported, such as *OVATE*, a negative regulatory gene causing pear-shaped tomato fruits (Liu et al., 2002); SUN, a retrotransposon-mediated gene (Xiao et al., 2008); locule number *fas* (Huang and van der Knaap, 2011) and *lc* (Munos et al., 2011). Other cloned genes related to tomato development are summarized in a recent review paper (Rothan et al., 2019).

Tomato fruits are rich in diverse nutrients and health-promoting compounds, such as sugars, organic acids, amino acids and volatiles (Goff and Klee, 2006; Klee, 2013). However, breeding tomatoes with high nutrition and strong flavor still remain a major breeding

challenge (Tieman et al., 2012; Klee and Tieman, 2013; Klee and Tieman, 2018). *Lin5*, a major QTL modifying sugar content in tomato fruit, was cloned about 20 year ago (Fridman et al., 2000). In various genetic backgrounds and environments, the wild-species allele increased glucose and fructose contents compared to the cultivated allele (Fridman et al., 2000). In addition, this gene shared a similar expression pattern in tomato, potato and Arabidopsis (Fridman and Zamir, 2003). Recently a *SWEET* protein, a plasma membrane-localized glucose efflux transporter, was shown to play a role in the ratio of glucose and fructose accumulation (Shammai et al., 2018). A balanced content of sugars and organic acids is crucial for consumer preference (Tieman et al., 2017). Recently, a major QTL regulating malate content was cloned, corresponding to an *Aluminium Activated Malate Transporter 9* (*Sl-ALMT9*) (Ye et al., 2017). Though only a few QTLs regulating sugars and organic acids have been functionally validated, this knowledge is important for understanding the regulation mechanisms. Several genes involved in the variation of volatile production were also characterized (Tieman et al., 2006; Klee, 2010; Tikunov et al., 2013; Shen et al., 2014; Bauchet et al., 2017b; Klee and Tieman, 2018).

1.6.6 New genomic and biological resources for QTL and candidate genes identification

Lin et al., (2014) demonstrated the benefits of whole-genome resequencing of two extreme bulk populations derived from an F_2 population of tomato, where many fruit weight QTLs were identified. Whole-genome-sequencing of bulked F_2 plants with contrasted phenotypes offers the opportunity to identify the SNPs that are putatively related to the target phenotypes via aligning the sequenced data to the reference genome (Garcia et al., 2016). This approach has been also efficient in identifying mutations, especially generated by EMS (Garcia et al., 2016).

However, the genetic diversity of linkage populations (parental mapping population) is limited to the two parental accessions used for crossing. In order to overcome this limitation, multi-parent advanced generation intercross (MAGIC) populations offer an alternative, which has been generated for different species, such as Arabidopsis (Kover et al., 2009), rice (Bandillo et al., 2013), wheat (Huang et al., 2012b; Mackay et al., 2014), faba bean (Sallam and Martsch, 2015), sorghum (Ongom and Ejeta, 2017) and tomato (Pascual et al., 2015). The first tomato MAGIC population was developed by crossing eight re-sequenced tomato lines and there was no obvious population structure in this population. The linkage map was 87% larger than those derived from bi-parental populations and some major fruit quality

QTLs were identified by using this approach (Pascual et al., 2015). Recently, this MAGIC population was also used for identifying QTLs under water deficit and salinity stresses and many stress-specific QTLs were identified (Diouf et al., 2018). The MAGIC population harnesses the benefits of mapping populations and GWAS populations and also overcomes some of the limitations of these two populations. However, generating the MAGIC population is more complex than two parental mapping populations and many potential recombinations from the MAGIC population is also removed, due to the population size, which limits the efficiencies and applications.

1.7 Genotype imputation

1.7.1 Principle and interest

The main idea of genotype imputation is to compare the missing genotypes with the reference panel and impute the missing alleles with probabilities (**Figure 1.20**). The main benefit of imputation is to greatly increase the genome coverage of SNPs without additional sequencing or genotyping efforts, once a high quality reference panel is available. High quality and high genome coverage of molecular markers is essential for most of marker-based genetic analyses, such as evolutionary analyses, population genetics, genomic selection (GS) and genome-wide association studies (GWAS). Imputation accurately assigns genotypes at untyped markers, improving genome coverage, facilitating comparison and combination of studies that use different marker panels, increasing power to detect genetic association, and guiding fine-mapping (Marchini and Howie, 2010; Das et al., 2016). Genotype imputation can boost the power up to 10% percent (Spencer et al., 2009), and be further used for fine-mapping (Marchini et al., 2007) and meta-analysis of GWAS (Marchini and Howie, 2010).

When a large diverse reference panel is available, SNP density can be significantly increased by genotype imputation (Guan and Stephens, 2008; Halperin and Stephan, 2009; Iwata and Jannink, 2010; Marchini and Howie, 2010; Pasaniuc et al., 2012; Das et al., 2016; Browning and Browning, 2016; Wang et al., 2018). In human and model plant species, there are some very good reference panels suitable for genotype imputation, such as the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) and the UK10K Project in humans (The UK10K Consortium, 2015; Danecek et al., 2015), the 3000 Rice Genome Project (The 3000 rice genomes project, 2014; McCouch et al., 2016) and the 1001 Genomes Consortium in *Arabidopsis thaliana* (The 1001 Genomes Consortium, 2016). These reference imputation panels greatly benefited researchers in 1) increasing genome-wide SNP coverage, 2) GWAS and meta-analysis of GWAS, 3) regional fine-mapping, 4) investigating the missing heritability and others.

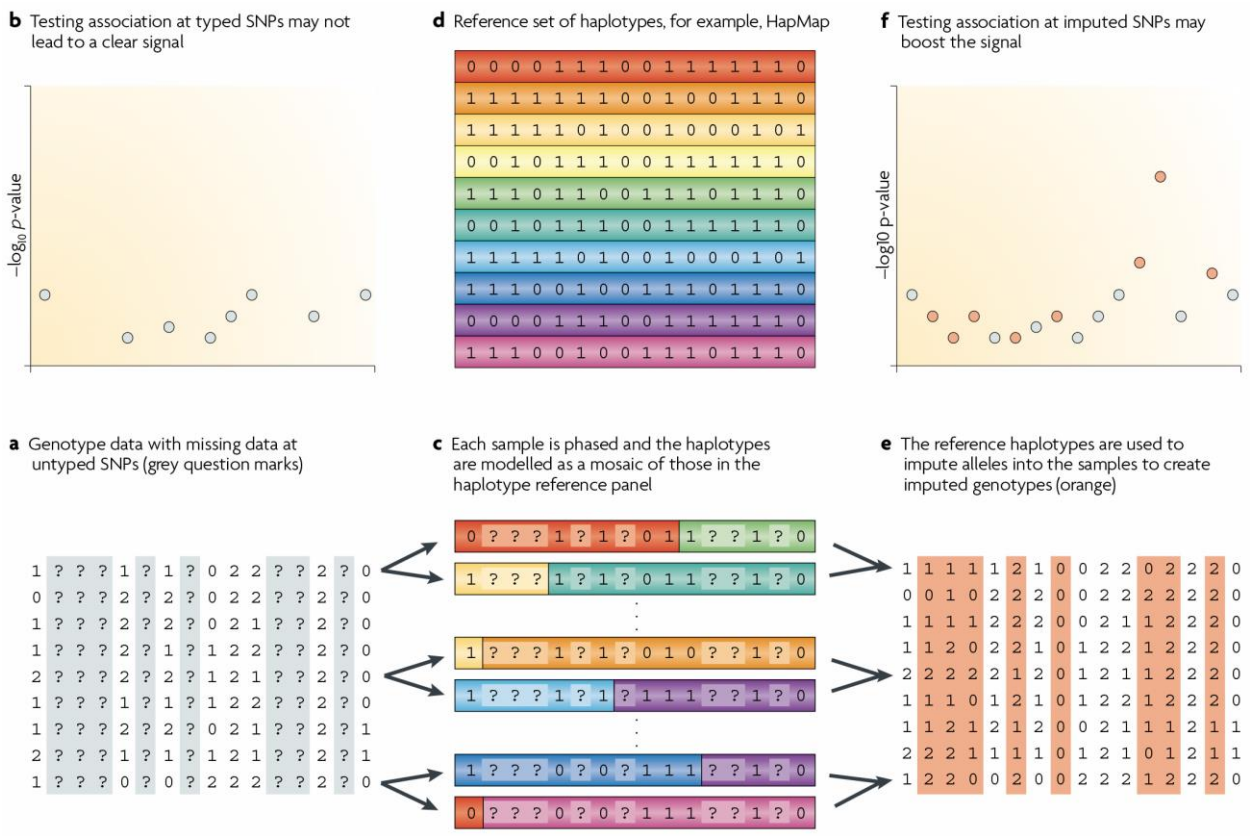


Figure 1.20 How genotype imputation works. The raw data consist of a set of genotyped SNPs that has a large number of SNPs without any genotype data (part a). Testing for association at just these SNPs may not lead to a significant association (part b). Imputation attempts to predict these missing genotypes. Algorithms differ in their details but all essentially involve phasing each individual in the study at the typed SNPs. The figure highlights three phased individuals (part c). These haplotypes are compared to the dense haplotypes in the reference panel (part d). Strand alignment between data sets must be done before this comparison takes place. The phased study haplotypes have been coloured according to which reference haplotypes they match. This highlights the idea implicit in most phasing and imputation models that the haplotypes of a given individual are modelled as a mosaic of haplotypes of other individuals. Missing genotypes in the study sample are then imputed using those matching haplotypes in the reference set (part e). In real examples, the genotypes are imputed with uncertainty and a probability distribution over all three possible genotypes is produced. It is necessary to take account of this uncertainty in any downstream analysis of the imputed data. Testing these imputed SNPs can lead to more significant associations (part f) and a more detailed view of associated regions (adapted from Marchini and Howie, 2010).

1.7.2 Comparison of imputation programs and software

There are several genotyping imputation software available, such as IMPUTE (Marchini et al., 2007), PLINK (Purcell et al., 2007), BIMBAM (Servin and Stephens, 2007), BEAGLE (Browning and Browning, 2008), FIMPUTE (Sargolzaei et al., 2014), MACH-admix (Liu et al., 2013). Each of them has their own pros and cons (see **Table 1.11** for details). For example, PLINK and Beagle are computationally more efficient because they focus on genotypes for a relatively small number of neighboring markers when imputing each missing genotype. IMPUTE, MaCH and fastPHASE are computationally more intensive but provide a better estimate of missing genotypes because they take into account all available markers when imputing each missing genotype (Porcu et al., 2013).

While it has been shown that the imputation accuracy does not appear to be substantially affected by a GWAS QC (quality control) step, this observation is only valid for common variants and may not be generalized to the imputation of low frequency (1-5% MAF) and rare variants (<1% MAF) (Southam et al., 2011). Detailed explanations of the key important information about these methods have been summarized by Marchini and Howie (2010). The imputation accuracy of SNP with rare alleles (MAF ≤ 0.05) is important since rare alleles may account for a large portion of the genetic variation that is not explained by common alleles (Manolio et al., 2009; Makowsky et al., 2011; Gibson, 2012; VanRaden et al., 2013; Ma et al., 2013).

Nazzicari et al., (2016) compared the performance of four general imputation methods (K-nearest neighbors, Random Forest, singular value decomposition, and mean value) and two genotype-specific methods (‘‘Beagle’’ and FILLIN) was tested on GBS data from alfalfa (*Medicago sativa* L., autotetraploid, heterozygous, without reference genome) and rice (*Oryza sativa* L., diploid, 100 % homozygous, with reference genome). Beagle was the best performing method, both for accuracy and time-wise, in rice. In alfalfa, KNNI and RFI gave the highest accuracies, but KNNI was much faster. Hickey et al. (2012) used IMPUTE2 for imputation and found that the accuracy of imputation was high even when only 8774 SNP constitute the low-density platform. The correlation between the true and imputed genotypes was 0.87. However, there was a dramatic reduction in the accuracy of imputation when the low-density platform had fewer than 8774 genotypes.

Table 1.11 Comparison of imputation methods (adapted from Marchini and Howie, 2010).

Properties	Imputation method				
	IMPUTE v1	**IMPUTE v2.2**	**MACH v1.0.16**	**fastPHASE v1.4.0 BIMBAM v0.99**	**BEAGLE v3.2**
Reference panels					
Can use a haplotype reference panel?	Yes	Yes	Yes	Yes	Yes
Can use a genotyped reference panel?	No	Yes	Yes	Yes	Yes
Can two haplotype or genotype reference panels be used in the same run?	No	Yes	No	No	No
Reference panels available in correct format	HapMap2 HapMap3 1KGP pilot data	HapMap2 HapMap3 1KGP pilot data	HapMap2 HapMap3 1KGP pilot data	HapMap2	No
Study samples					
Can take genotypes specified with uncertainty?	No	Yes	No	No	Yes
Can accommodate trios and related samples?	No	No	No	No	Trios and duos
Can impute into a study sample of autosomal haplotypes?	Yes	Yes	No	No	Yes
Can impute on the X chromosome?	Yes	Yes	No	No	Yes
Program options and features					
Does phasing as well as imputation?	No	Yes	Yes	Yes	Yes
Can impute sporadic missing genotypes?	No	Yes	Yes	Yes	Yes
Has internal performance assessment?	Yes	Yes	Yes	No	No
Can impute only in a specified interval?	Yes	Yes	No	No	No
Can handle strand alignment between data sets?	Yes	Yes	Yes	No	No
SNP and sample inclusion and exclusion options?	Yes	Yes	No	Yes	Yes
Joint model for imputation and association testing?	No	No	No	No	No
Operating system requirements	Linux, Solaris, Windows, Mac	Linux, Solaris, Windows, Mac	Linux, Windows, Mac	BIMBAM (source code + Windows) fastPHASE (Linux, Solaris, Windows, Mac)	Java executable
Computational performance					
Assessment 1*	43m (1000 Mb)	75m (180 Mb)	105m (80 Mb)	855m (16 Mb)	56m (3100 Mb)
Assessment 2‡	---	48m (115m)	---	157m (211m)	104m (234m)
Error rates§					
Rows correspond to the Scenario A, Scenario B (restricted) and Scenario B (full) data sets	5.42%	5.16%	5.46%	5.92%	6.33%
	---	3.4% (0.86%)	---	5.33% (1.32%)	3.46% (0.93%)
	---	3.4% (0.86%)	---	---	4.01% (1.04%)
Output files					
Genotype posteriors produced?	Yes	Yes	Yes	Yes	Yes
Information measures?	Yes	Yes	Yes	No	Yes
Easiest use of output files to test association	Feed files directly into SNPTEST. Test based on genotype posteriors, dosages or thresholded genotypes	Feed files directly into SNPTEST. Test based on genotype posteriors, dosages or thresholded genotypes	Genotype dosage files can be fed into MACH2DAT or MACH2QTL	BIMBAM can produce file formats used by BIMBAM. fastPHASE out files need to be processed	Best-guess phased haplotypes can be tested in BEAGLE. Processing required to use genotype posteriors or dosage

1.7.3 Measurement of imputation accuracy

Imputation accuracy is a key parameter to evaluate the efficiency of genotype imputation of different programs. In order to calculate the imputation accuracy, correct allele rate (CR) and correlation coefficient r^2 between true and imputed genotypes and imputation error rates are frequently evaluated. Based on the nature of both measures and results reported in the literature, imputation accuracy appears to be a more useful measure of the correctness of imputation than imputation error rates, because imputation accuracy does not depend on minor allele frequency (MAF), whereas imputation error rate depends on MAF. Imputation accuracy depends on the ability of identifying the correct haplotype of a SNP, but many other factors have been identified as well, including the number of genotyped immediate ancestors, the number of individuals with genotypes at the high-density panel, the SNP density on the low- and high-density panel, the MAF of the imputed SNP and whether imputed SNP are located at the end of a chromosome or not.

There are several different ways to compare true and imputed genotypes. Marchini et al. (2007) used imputation certainty as the measurement of imputation quality. In some studies the percentage of incorrectly imputed alleles or genotypes is reported, and termed allelic or genotype imputation error rate (Zhang and Druet, 2010). Other studies report the percentage of correctly imputed genotypes, and call this imputation accuracy (Weigel et al., 2010), while other refer to the (squared) correlation between true and imputed genotypes as imputation accuracy (Druet and Georges, 2010; Calus et al., 2011; Mulder et al., 2012). Other measures that have been developed or suggested include the imputation quality score (Lin et al., 2010) and those that are derived internally in imputation algorithms. The imputation packages MaCH and Beagle compute a measure that attempts to predict the imputation R^2 value based on the posterior distribution of the Gibbs sampler, without having any information of the true genotypes. It was suggested that the correlation between true and imputed genotypes being independent from the allele frequency at the imputed locus, it may therefore be a measure with more desirable properties than allelic imputation error rates (Browning and Browning, 2008; Hickey et al., 2012b).

1.7.4 Factors affecting imputation accuracy

The most important factor for imputation success in livestock is the number of genotyped immediate ancestors (Hickey et al., 2011; Huang et al., 2012a). When there are no or

few immediate ancestors with genotypes, the total number of animals at the imputed density becomes important, that is, having too few animals with genotypes at the imputed SNP density yields poor imputation results (Hayes et al., 2012; Wang et al., 2012a). Conversely, the impact of having only a small number of animals available at the imputed SNP density on imputation accuracy may be limited if those animals are close relatives, for example, immediate ancestors, of the imputed individuals (Gualdrón Duarte et al., 2013).

Other factors include the SNP density on the low and high density panel (Mulder et al., 2012), the MAF of the imputed SNP (van Binsbergen et al., 2014) and whether imputed SNP are located at the end of a chromosome or not (Badke et al., 2013; Cleveland and Hickey, 2013; Wellmann et al., 2013). For very low-density SNP panels (e.g. 384 SNP), the impact of the LD between imputed SNP and SNP on the low-density panel can also be reduced considerably if the individuals genotyped at the imputed density are close relatives of the imputed individuals (Wang et al., 2013; Hickey and Kranis, 2013; Wellmann et al., 2013).

When using software that does not explicitly utilise pedigree information, other important factors affecting imputation accuracy include the number of individuals with genotypes at the imputed density (Zhang and Druet, 2010), and the relationship between imputed individuals and individuals genotyped at high density (Hickey et al., 2012a).

1.8 Genome-wide association studies

1.8.1 General introduction

Association study aims at detecting associations between genetic variants and targeted phenotypes in a given population. It was first used in human diseases and then widely applied in other non-human species. It can be briefly subdivided into candidate gene association studies and genome-wide association studies (GWAS) (Hirschhorn and Daly, 2005). GWAS has been first applied to detect variants underlying complex traits and common diseases in humans (Hirschhorn and Daly, 2005; Visscher et al., 2017; Sud et al., 2017; Schaid et al., 2018; Tam et al., 2019). In addition, various direct and indirect methods have been developed to test for associations (**Figure 1.21**). Nowadays, GWAS have become popular to detect candidate genes in humans population studies (**Figure 1.22**).

1.8.2 Models used in genome-wide association studies

With the increasing of population size, number of markers and complexity of cases, newer models and tools were developed for GWAS in order to reduce computational demand and also increase statistical power (Gupta et al., 2014; Gupta et al., 2019). The main models developed recently are listed in **Table 1.12**. Based on the differences of models, these approaches can be generally subdivided into six groups: (1) Single-locus, single-trait (SLST) mixed model; (2) Multi-locus and multi-trait mixed model (MLMM and MTMM); (3) Joint-linkage association mapping (JLAM); (4) Use of diverse panels for GWAS; (5) Epistasis and Q × E interactions; (6) Bayesian methods for GWAS, as detailed reviewed by Gupta et al., (2019).

1.8.2.1 Single-locus and single-trait association model

Most of the GWAS were first performed with a single-locus and a single-trait association model. Multi-locus and multi-trait association models were mainly developed in the last 5-10 years. The first mixed association model was proposed by Yu et al., (2006). These mixed models can be classified into two groups: exact methods and approximate methods. The exact methods, such as GEMMA and Fast-LMM, estimate each marker effect and are comparatively slower (Zhou and Stephens, 2012; Lipka et al., 2015). In contrast, the approximate methods, such as EMMAX and GRAMMAR do not need to estimate population parameters for each marker and are computationally fast. In most cases, these two approaches are asymptotically equivalent and the choice of a method depends on the dataset, computational speed and the level of user-friendliness (Gupta et al., 2014; Eu-ahsunthornwattana et al., 2014; Gupta et al., 2019).

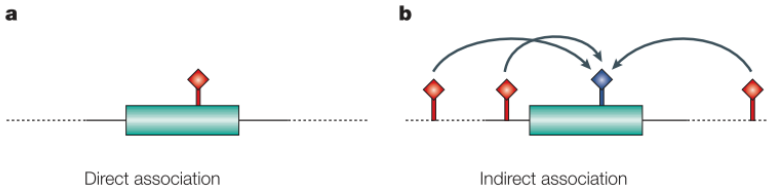


Figure 1.21 Testing SNPs for association by direct and indirect methods, as adapted from Hirschhorn and Daly, (2005).

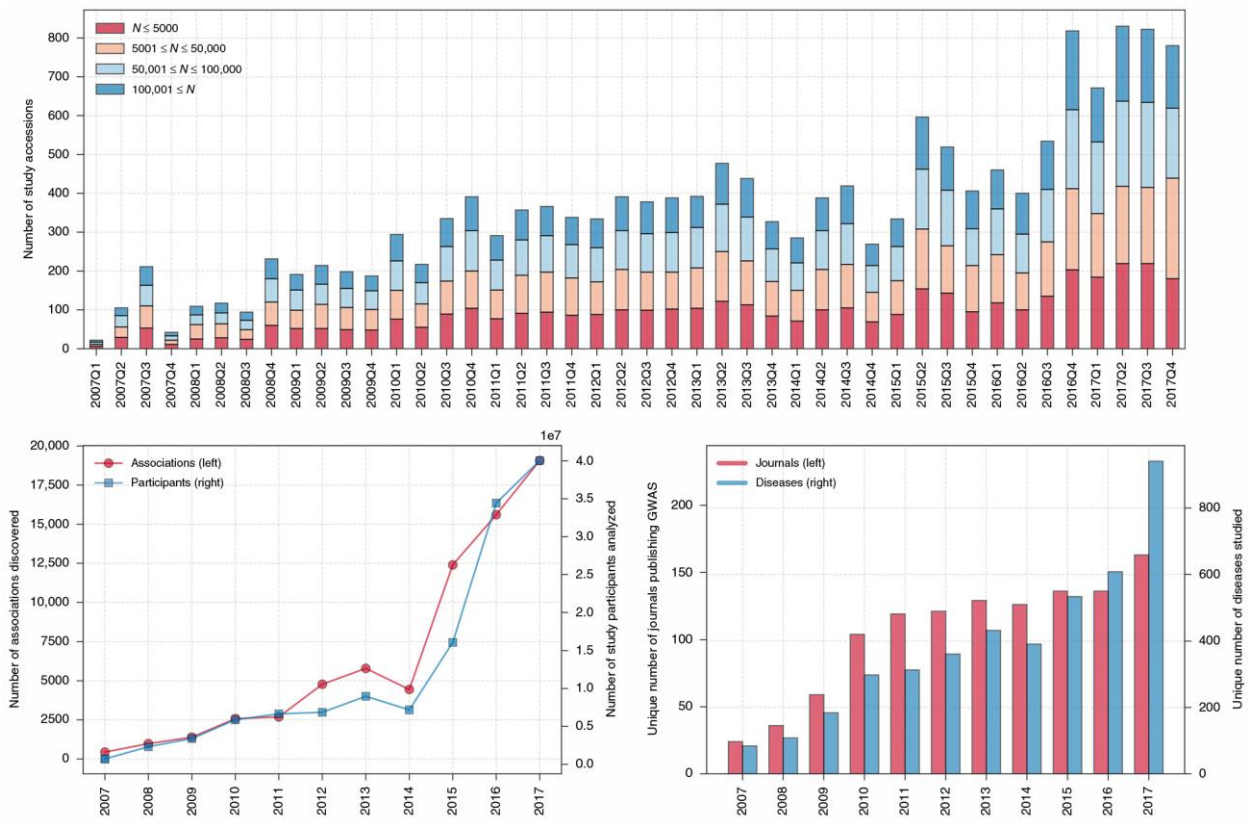


Figure 1.22 The growth of GWAS, 2007 – 2017. The upper panel shows the number of study accessions published per quarter over time colored according to sample size to show the growth of larger (100,001 ⩽N) GWAS. The lower left panel shows the strong positive correlation between the number of associations found and the number of participants used in GWAS over time. The lower right panel shows the growth in the number of unique traits examined as well as the number of unique journals publishing GWAS over time. 2007–2017 is selected since only 10 entries occurred before 2007. Each panel contains full calendar years only. Source: NHGRI-EBI GWAS Catalog (adapted from Mills and Rahal, 2019).

1.8.2.2 Multi-locus and multi-trait association model

Single-locus and single-marker association models have some limitations, such as multiple testing, background genotype effects and pleiotropic effects (Akey et al., 2001; Korte et al., 2012; Segura et al., 2012; Buzdugan et al., 2016). Multi-trait and multi-locus mixed associations provide new opportunities. The multi-locus models include: (1) Bayesian-inspired penalized maximum likelihood approach (Hoggart et al., 2008), (2) penalized logistic regression approach (Ayers and Cordell, 2010), (3) elastic-net approach (Cho et al., 2010), (5) empirical Bayes approach (Lü et al., 2011), (6) multi-locus mixed model, MLMM (Segura et al., 2012) and (7) random-SNP-effect model (Wang et al., 2016b). Several approaches are proposed to handle the "large p, small n" problem, such as efficient exact variance component test (ExactVCTests) (Zhou et al., 2016). Compared to single linear model, models taking into account different co-factors could greatly reduce false positives and have become the most commonly used in GWAS (**Figure 1.23**).

In contrast to multi-locus mixed model (MLMM), multi-trait mixed model (MTMM) considers the trait-trait interactions and is helpful in identifying regions controlling more than one trait (Korte et al., 2012). However, MTMM can only be used for two traits, which was recently improved to multi-trait associations by using the matrix-variate linear mixed model (mvLMM) (Zhou and Stephens, 2014; Furlotte and Eskin, 2015). Multiple-trait interactions can also be managed by principal component analysis (PCA) and then analysis of each pseudo-PC separately (Gao et al., 2014). As GWAS become more and more popular and common, multi-trait meta-analysis of genome-wide associations using summary statistics will become much more promising (Turley et al., 2018). It is nowadays even possible to model multi-locus multi-trait at the same time (Lippert et al., 2014; Kim et al., 2016; Zhan et al., 2017).

Table 1.12 Different mixed model approaches proposed over the years for GWAS in crop plants along with their features (adapted from Gupta et al., 2019).

No.	Approach	Features	Reference
1	Mixed linear model (MLM)	Takes care of multiple levels of relatedness; effectively controls population structure and type I and type II error rate	(Yu et al., 2006)
2	Genome-wide rapid association using mixed model and regression (GRAMMAR)	An approximate method which first estimates the residuals adjusted for family effects and then treats these as phenotypes along with genotyping data for analysis using rapid least- squares method; reduce computation time for each individual SNP	(Aulchenko et al., 2007)
3	Efficient mixed-model association (EMMA)	An exact method that accounts for population structure and genetic relatedness with substantially increased computational speed and reliability of the result.	(Kang et al., 2008)
4	Efficient mixed-model association eXpedited (EMMAX)	An approximate method in which VCA is not repeated for each marker, as each marker is assumed to explain only a small fraction of phenotypic effect; instead, heritability estimated from the null model is used for all markers; can perform AM using vast amount of data in a short time	(Kang et al., 2010)
5	Compressed mixed linear model (CMLM)	Clusters the individuals into fewer groups based on the kinship among the individuals; the kinship between pairs of groups is replaced by the kinship between pairs of individuals; reduce the computation demand substantially	(Zhang et al., 2010)
6	Population parameters previously determined (P3D)	A complementary approach to CMLM; eliminates the need of estimating population parameters (such as VCs); computationally fast	(Zhang et al., 2010)
7	Factored spectrally transformed linear mixed models (FaST-LMM)	An exact method with improvement over MLM approach brought out by use of a low-rank relatedness matrix (matrix based on a few thousand markers instead of all markers); reduces computation time considerably	(Lippert et al., 2011)
8	Multi-locus mixed model (MLMM)	An improvement over MLM; can effectively control for population structure and false discovery rate in GWAS; takes into account the background genotypes	(Segura et al., 2012)
9	Multi-trait mixed model (MTMM)	Performs GWAS of correlated phenotypes using the principle of MLM; takes into account both, within-trait and between-trait VCs simultaneously of multiple traits	(Korte et al., 2012)
10	GRAMMAR-Gamma	A VC-based two-step approximate method; an improvement over GRAMMAR; reduces computational demand and provides correct estimates of SNP effects; suitable for using genotyping data based on whole-genome resequencing with large sample size	(Svishcheva et al., 2012)
11	GEMMA	An efficient-exact method; faster than EMMA; yield accurate p values even in the presence of strong population structure, and even when the marker effect is large; suitable for studies with large association panels	(Zhou and Stephens, 2012)
12	Linear mixed model Lasso (LMM-Lasso)	Combines multivariate analysis and corrects for population structure (combination of MLM and Lasso regression); can partition the total phenotypic variance into different components, like the one caused due to individual SNP effects as well as that caused by population structure	(Rakitsch et al., 2013)
13	Selecting CONnected explanatory SNPs (SConES)	An efficient multi-locus method for discovering sets of loci which are associated with a phenotype while being connected in an underlying network; computationally fast	(Azencott et al., 2013)
14	Low rank linear mixed model (LRLMM)	Take into account the effective degrees of freedom for interpreting model complexity of the LRLMM along with principal components (for controlling population structure) and kinship	(Hoffman, 2013)
15	Bayesian sparse linear	A combination of MLM and sparse regression models	(Zhou et al.,

	mixed model (BSLMM)		2013)
16	Settlement of MLM under progressively exclusive relationship (SUPER)	An improvement over FaST-LMM; extracts a subset of SNPs and uses them in FaST-LMM; increased statistical power	(Wang et al., 2014)
17	Genetic analysis incorporating Pleiotropy and Annotation (GPA)	Enables joint analysis of multiple GWA data sets and the annotation information	(Chung et al., 2014)
18	Enriched CMLM (ECMLM)	An improvement over CMLM with increased statistical power; calculates kinship using several different algorithms and uses this information during analysis	(Li et al., 2014)
19	Principal components-Select (PC-Select)	A hybrid approach that includes the PCs of the genotype matrix as fixed effects in FaSTLMM Select method	(Tucker et al., 2014)
20	Multivariate Linear mixed models (mvLMM)	Uses computationally-efficient algorithm for fitting mvLMMs with one covariance component (in addition to the residual error term), and for performing the LR test for GWAS; improvement over GEMMA	(Zhou and Stephens, 2014)
21	BOLT-LKMM	Based on Bayesian mixed-model association; increased computational power	(Loh et al., 2015)
22	Random-SNP-effect MLM (RMLM)	SNP-effects are treated as random; the threshold p value for significance tests are calculated based on a modified Bonferroni correction	(Wang et al., 2016b)
23	Multi-locus RMLM (MRMLM)	A multi-locus model that includes markers selected from the RMLM with less stringent selection criterion ; multiple test correction is not required	(Wang et al., 2016b)
24	Fixed and random model Circulating Probability Unification (FarmCPU)	Combines both, the fixed effect and random effect models in analysis and improves statistical power with reduced computing time	(Liu et al., 2016b)
25	Penalized multitrait mixed modeling approach	Accommodate both types of correlations; i.e., between subjects and traits during analysis	(Liu et al., 2016c)
26	pLARmEB (polygenic-background-control-based least angle regression plus empirical Bayes)	Integrates least angle regression with empirical Bayes and ca perform multilocus GWAS; it is more powerful in detection of QTN and its effect; has less false positive rate and require less computing time than Bayesian hierarchical generalized linear model	(Zhang et al., 2017)
27	LFMM 2	A R based fast gene-environment association model, and outperforms other approaches based on principal component or surrogate variable analysis	(Caye et al., 2019)
28	hapQTL	A Bayesian model that relies on a hidden Markov model to infer ancestral haplotypes and their loadings at each marker for each individual.	(Xu and Guan, 2014)

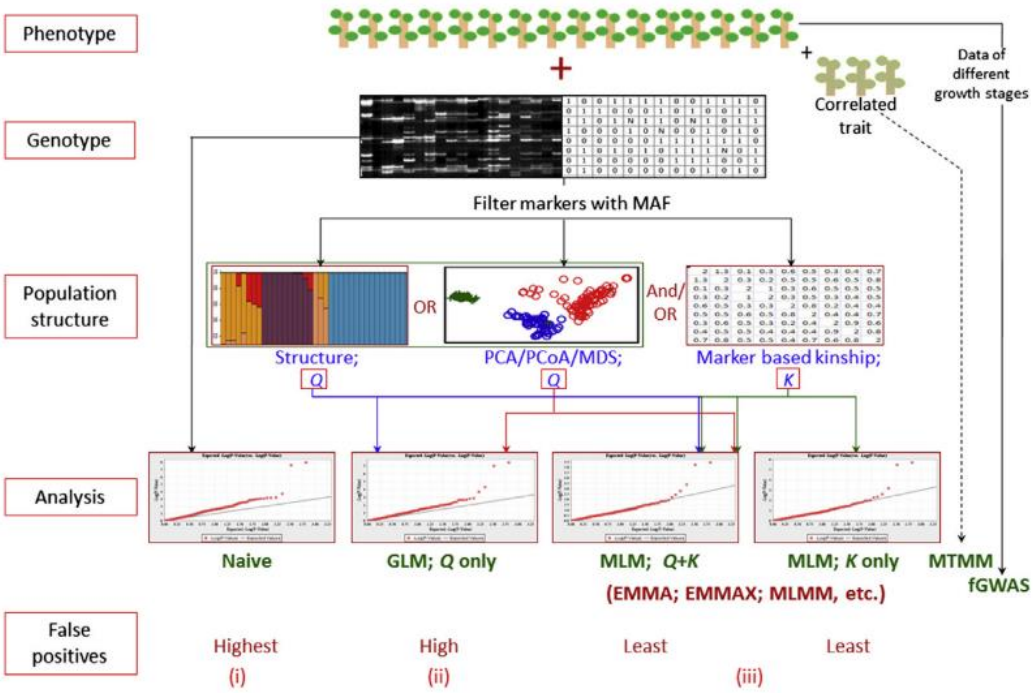


Figure 1.23 Schematic representations of various steps involved in association analysis. On the extreme left are shown the various steps involved in association mapping including phenotyping, genotyping, study of population structure, association analysis, and the rate of false positives (adapted from Gupta et al., 2014).

1.8.2.3 Joint-linkage association mapping (JLAM)

Both linkage mapping and association mapping have merits and limits. Joint linkage-association mapping (JLAM) was proposed to harness the benefits and overcome the limitations of these two approaches (Wu and Zeng, 2001). Among the earliest examples, a multi-parental maize population was developed to test the potential benefits of nested association mapping (NAM). Later on, many new multi-parental mapping populations were developed, such as multi-parent advanced generation intercross (MAGIC), multi-line Cross Inbred Lines (MCILs), Recombinant Inbred Advanced Intercross Lines (RIAILs), recombinant inbred chromosome substitution lines (RICSLs) (Gupta et al., 2019). Notably, the MAGIC population has been successfully applied in mapping the ecologically and evolutionarily relevant traits in Arabidopsis (Kover and Mott, 2012) and other agronomical traits in major crops, such as rice (Bandillo et al., 2013), wheat (Huang et al., 2012b), cotton (Islam et al., 2016) and tomato (Pascual et al., 2015).

1.8.2.4 Epistasis and Q × E interactions

Some models were also developed to handle the epistasis (QTL × QTL) and Q × E interactions (reviewed by Wei et al., 2014; Upton et al., 2016; Gupta et al., 2019). Epistasis has both functional (the variant effect at one locus depends on the variant at another locus) and statistical effects (variance attributed to the interactions between variants, apart from their independent effects) (Wei et al., 2014). The multi-trait mixed model (MTMM) provides new opportunities to dissect G × E interactions (Korte et al., 2012). Recently, Lü et al. (2011) proposed an epistatic association mapping (EAM) approach in soybean, which could estimate all the main-effect quantitative trait loci (QTLs), environmental effects, Q × E interactions and QTL × QTL interactions by empirical Bayes approach. Saïdou et al. (2014) proposed another mixed linear model which estimated the effects of SNP by environment interaction, ancestry by environment interaction, SNP by ancestry interaction and three way interactions. Diouf et al. (2018) used a tomato MAGIC population to investigate the genotype × environment (G × E) interactions. They found significant G × E interactions for five of the seven traits over 2 years and 15 QTLs revealed G × E interactions and 35 QTLs were treatment specific.

1.8.2.5 Bayesian association models

Bayesian association models provide new opportunities for GWAS. Compared to frequentist approaches, Bayesian approaches can deal with the problems of multiple testing and rare marker alleles and also increase the computation speed (Fernando and Garrick, 2013). Bayesian approaches can also be efficient for fine mapping of candidate genes (Schaid et al., 2018) but also in meta-analysis (Ashby, 2006; Stephens and Balding, 2009). Notably, Xu and Guan (2014) proposed Bayesian association model and demonstrated the benefits of using haplotypes in identifying associations. With the fast development of machine learning and artificial intelligence, Bayesian approaches should become more popular in GWAS (Gupta et al., 2019).

1.8.3 Landmarks of genome-wide association studies in tomato

After the demonstration of GWAS power to analyze human diseases (Klein et al., 2005), it was quickly adopted in major crops (Brachi et al., 2011; Luo, 2015; Liu and Yan, 2019). In tomato, the first reported association study was performed to identify the SNPs associated with the fruit weight QTL *fw2.2*. However, the authors did not find any positive associated SNP in a small collection of 39 cherry tomato accessions (Nesbitt and Tanksley, 2002).

The high degree of LD in tomato genome, especially within modern large fruit tomato collections, is beneficial in terms of the minimum number of molecular markers needed to cover the whole genome. Before the availability of large SNP number, molecular markers such as SSRs were popular for GWAS. For example, Xu et al. (2013) performed an association mapping on 188 tomato accessions with 121 polymorphic SNPs and 22 SSRs. They successfully identified 132 significant associations for six quality traits. Zhang et al., (2016) genotyped 174 tomato accessions including 123 cherry tomato and 51 heirlooms with 182 SSRs and performed GWAS for fruit quality traits. A total of 111 significant associations were identified for 10 traits and many previously identified major QTLs were located in/near regions of the significant associated markers. The authors further extended the phenotypes to volatiles (Zhang et al., 2015), as well as sugars and organic acids (Zhao et al., 2016).

With the availability of the reference tomato genome (The Tomato Genome Consortium, 2012), millions of SNPs became available and allowed the identification of causative polymorphisms. For instance, the causative gene *SlMYB12* conferring pink tomato fruit color was identified in a GWAS using 231 sequenced tomato accessions (Lin et al., 2014). Several

mutations were further identified in the protein structure of SlMYB12 and the authors identified three recessive alleles of this gene controlling pink tomato color (Lin et al., 2014).

However, whole-genome-sequencing is still quite expensive, especially at a large population scale, which greatly limits its wide applications. SNP arrays were thus developed to overcome this limit (Hamilton et al., 2012; Sim et al., 2012a). Sauvage et al., (2014) genotyped 163 tomato accessions composed of large-fruit, cherry and wild tomato accessions with the SolCAP array, generating a total of 5995 high quality SNPs. Then they performed GWAS using a multi-locus mixed model (MLMM; Segura et al., 2012) for 36 metabolites that were highly correlated across two years of experiment and identified 44 candidate loci associated for different fruit metabolite contents (Sauvage et al., 2014). Among the candidate loci, they identified a gene with unknown function on chromosome 6 that was strongly associated with malate content. This association was further identified in different GWAS and meta-analysis of GWAS based on different populations (Tieman et al., 2017; Bauchet et al., 2017b; Ye et al., 2017) and was further validated as an *Al-Activated Malate Transporter 9* (*Sl-ALMT9*) (Ye et al., 2017).

Bauchet et al., (2017b) genotyped 300 tomato accessions with both the SolCAP and CBSG arrays, generating a total of 11,012 high quality SNPs, which were used for GWAS using both MLMM and multi-trait mixed model (MTMM) (Korte et al., 2012). A total of 79 significant associations were identified for the content in 13 primary and 19 secondary metabolites in tomato fruits. Among these, two associations involving fruit acidity and phenylpropanoid content were particularly investigated (Bauchet et al., 2017b). The same population was also characterized for agronomic traits and many QTLs were identified, such as *fw2.2* and *fw3.2*. Within this panel, the authors also demonstrated that admixed accessions shared different haplotype patterns compared to domesticated and wild tomatoes (Bauchet et al., 2017a). GWAS for similar quality traits were also performed in other collections (Ruggieri et al., 2014; Zhang et al., 2016).

With the fast development of whole-genome-sequencing technology and the reduction of price per genome, it is possible to sequence hundreds of diverse tomato collections. For instance, Tieman et al. (2017) sequenced 231 new accessions and combined these data with 245 previously sequenced genomes, generating a total of 476 genome sequences. These data were then used for GWAS for diverse flavor-related metabolites, including 27 volatiles, total soluble solids, glucose, fructose, citric acid, and malic acid. A total of 251 significant

associations were detected for 20 traits. Two loci were significantly associated with both glucose and fructose, corresponding to two major QTL *Lin5* and *SSC11.1*. By combining with selection analysis, it was further shown that the negative correlation between sugar content and fruit weight was likely caused by the loss of high-sugar alleles during domestication and improvement of ever-larger tomato fruits (Tieman et al., 2017). In addition, some candidate genes involved in tomato volatile contents were also identified, such as Solyc09g089580 for guaiacol and methylsalicylate. By combining the three significant associated loci for geranylacetone and 6-methyl-5-hepten-2-one, it was shown that the allelic combinations conferring favorable aromas were progressively lost during domestication and breeding (Tieman et al., 2017).

1.8.4 Limitations and challenges of genome-wide association studies

1.8.4.1 GWAS findings published to date represent only the tip of the iceberg

Nowadays, GWAS is best demonstrated in human studies, regardless of statistical models, population size and complexity of targeted traits, etc. However, the best achievements of GWAS findings in humans still only represent the tip of the iceberg (**Figure 1.24**). For example, human obesity arises from different factors and their interactions, including genetic predisposition, demographic factors, medical conditions, lifestyles and environmental exposures (McAllister et al., 2009; Locke et al., 2015; Pigeyre et al., 2016; Reddon et al., 2016). Though several waves of GWAS have been extensively applied to investigate human obesity, only few of the candidate genes have been validated and our knowledge about the underlying genetic regulation networks is still limited (Thorleifsson et al., 2009; Meyre et al., 2009; Sandholt et al., 2012; Wheeler et al., 2013). For tomato, though fruit weight has undergone two main evolutionary stages of domestication and improvement, there were at least 5 and 13 major QTLs located within domestication and improvement sweeps, respectively (Lin et al., 2014). In addition, the number of QTLs for fruit weight was much larger than that. However, only a few of them (fw2.2, fw3.2 and fw11.1) have been functionally characterized (Rothan et al., 2019) and our knowledge about the genetic control of fruit weight and other important traits of tomato is still quite limited.

1.8.4.2 Limitations of multiple testing burden of statistical models

A genome-wide significant threshold is required in GWAS to cutoff the significant associations, which is usually based on a Bonferroni correction to maintain a genome-wide false-positive rate at 5% (Bonferroni, 1936). There are some other correction methods available, such as Bonferroni step-down or Holm correction (Holm, 1979), Westfall and Young permutation (Westfall and Young, 1993), False discovery rate correction (FDR

correction) (Benjamini and Hochberg, 1995), q value (Storey, 2002) and step-up adaptive correction (Benjamini et al., 2006) (**Figure 1.25**). Among all these methods available, Bonferroni correction is the most stringent, which makes many true association unable to be detected. In contrast, the FDR correction and the q value are the less stringent (Qian and Huang, 2005).

There are several other strategies to reduce the number of multiple tests, including gene-based (Gamazon et al., 2015; Hägg et al., 2015; Savage et al., 2018; Sewda et al., 2019) or pathway-based association tests (Wang et al., 2007; Liu et al., 2010; De Las Fuentes et al., 2012; Cirillo et al., 2017), haplotype-based association studies (Clark, 2004; Xu and Guan, 2014; Wang et al., 2016a; N'Diaye et al., 2017; Daware et al., 2017; Maldonado et al., 2019), combining linkage mapping and GWAS (Johansson et al., 2010; Londin et al., 2013; Talukder et al., 2019; Gao et al., 2019b), genes specifically expressed in an important tissue (Du et al., 2011; Yang et al., 2018), genes with differential expression patterns (Yang et al., 2016; Zhou et al., 2019; Kuan et al., 2019) or SNPs harbouring evolutionary signatures (Grossman et al., 2010; Tam et al., 2019).

Sample size is an important factor in GWAS to map a genetic variant, especially for those rare variants (Altshuler et al., 2008). Including more samples in the studied panel can be one strategy to overcome this multiple testing limitations in GWAS. However, there are some additional problems and challenges to do so. (1) Population structure; if more samples are included from different genetic backgrounds, with the increasing of subgroups, the overall structure will be complex and could lead to statistical challenges. (2) Genotyping challenges; though with the fast development of NGS, the unit cost of genome sequencing is gradually reduced, sequencing a large GWAS panel still remains a main challenge for most of the crops. (3) Imputation: when using genotyping imputation, a high quality reference panel is required. Besides, the post-imputation quality control steps and parameters will also influence the imputation quality, which in turn will impact the results of GWAS. (4) Phenotyping challenges. Even though the genome sequencing is finally manageable at large population scale, phenotyping becomes another challenge, especially for those sample collected from distinct environmental backgrounds, which could lead to strong heterogeneity.

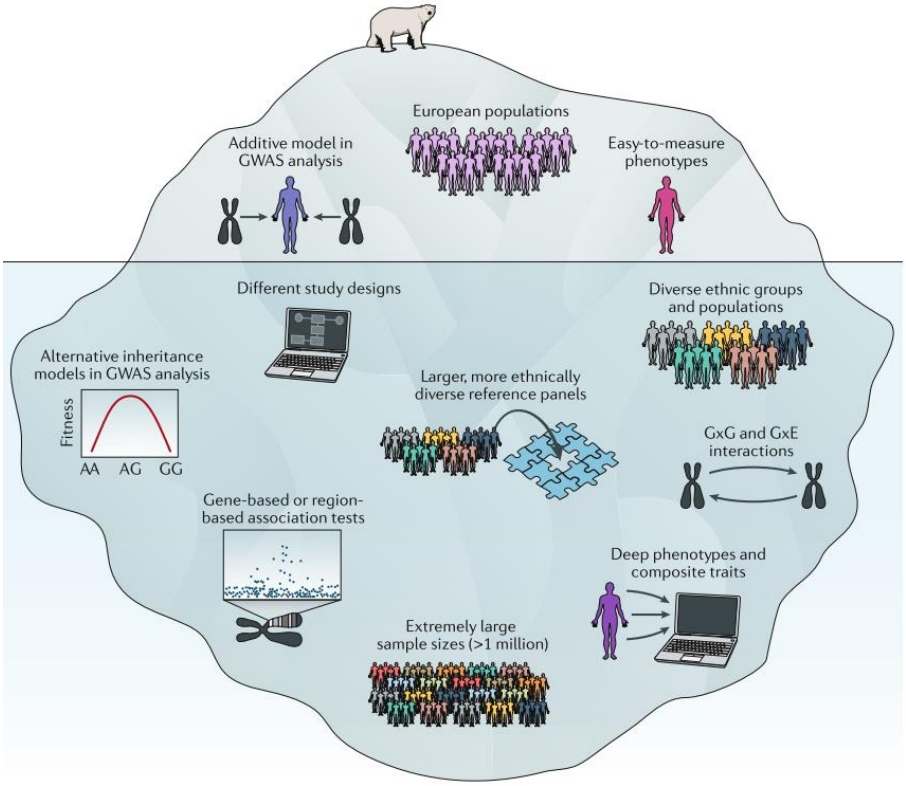


Figure 1.24 GWAs performed to date represent the tip of the iceberg. The discoveries that can be made using genome-wide association studies (GWAS) are represented by an iceberg. The portion of the iceberg above water represents the discoveries that have been made by GWAS to date, using easy- to-measure phenotypes, predominantly European populations, and an additive genetic model. Most of the iceberg is submerged under water. The submerged portion represents the vast number of discoveries that can potentially be made by expanding the current paradigm of GWAS to include a wider range of phenotypes, substantially larger sample sizes, more diverse populations and ethnic groups, and different study designs and analyses. G×G, gene-gene; G×E, gene-environment (adapted from Tam et al., 2019).

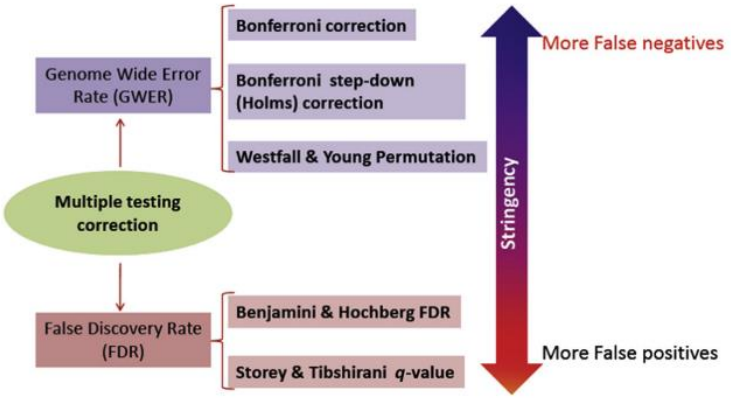


Figure 1.25 A comparison of methods used for corrections recommended to overcoming the multiple testing problem [genome-wide error rate (GWER) and false discovery rate (FDR)]. Stringency of the results of association mapping involving false negatives and false positives differs in different approaches of GWER and FDR in the order in which they are shown (on the extreme right, the upward arrow indicates the direction of more false negatives and the downward arrow indicates direction of more false positives) (adapted from Gupta et al., 2014).

Since GWAS are usually performed under an additive model, which is quite restrictive, new association models handling more complex models could further improve the statistical power to detect new associations. For example, apart from additive model, there are some other association models handling specific effects, such as recessive (Wood et al., 2016), dominant (Meyre et al., 2009; Lopes et al., 2014; Chen et al., 2015), over-dominant (Semel et al., 2006b; Wermter et al., 2008), multiplicative (Joo et al., 2009), parent-of-origin-specific (phenotypic effects of an allele depends on whether it is inherited from the mother or the father) (Lawson et al., 2013; Hoggart et al., 2014) and X-linked inheritance models (Tukiainen et al., 2014).

1.8.4.3 Missing heritability

Missing heritability is another main challenge in GWAS as only a small to modest fraction of the missing heritability is usually explained. It has been revealed that the heritability of important tomato quality traits ranged from low to high. However, taking tomato as an example, for many important quality traits, only a small proportion of heritability has been explained by the significant associated loci detected in GWAS (Sauvage et al., 2014; Tieman et al., 2017; Bauchet et al., 2017b). Understanding this genetic variation is important for the improvement of tomato quality. In human diseases, it is important to better prevent, diagnose and treat diseases (Manolio et al., 2009). There are several factors accounting for the missing heritability: (1) small population size and marker density; (2) large number of unidentified variants with small effect; (3) less common and rare allele variants; (4) gene-gene and gene-environment interactions (de Vlaming et al., 2017)(de Vlaming et al., 2017)(de Vlaming et al., 2017)(de Vlaming et al., 2017) (see reviews in Manolio et al., 2009; Eichler et al., 2010; Yang et al., 2017; Tam et al., 2019).

The size of samples analyzed and the number of markers genotyped is important in GWAS, and the statistical power will increase with larger population size and more markers (Fan and Song, 2016). Including more samples from different genetic backgrounds will diversify the total genetic diversity and alleles with low to moderate genetic effect might be able to be detected. More markers will also be helpful, and genotyping imputation provides an efficient and cost-effective approach to greatly increase the density of markers, when a large reference panel genotyped with dense markers is available (Marchini and Howie, 2010; Das et al., 2016; Fan and Song, 2016). For example, Hysi et al. (2018) performed a genome-wide association meta-analysis of European hair color based on genotyping imputation and showed that all the

identified significant associations explained substantially more heritability compared with previous estimates.

The heritability of human adult height is frequently quoted of approximately 80% based on family or twin studies (Silventoinen et al., 2003; Macgregor et al., 2006). However, GWAS using about 63,000 individuals only explains less than 20% of the total heritability (Visscher, 2008; Wood et al., 2014). The remaining missing heritability can be due to the incomplete linkage disequilibrium between causal variants and SNPs. In fact, a large proportion of the heritability is not missing but undetected due to the small effects that are unable to pass the stringent significance tests (Yang et al., 2010; Yang et al., 2017). For some traits, such as human schizophrenia, a large number of unidentified common variants with small effect is expected to explain the vast majority of genetic effects (International Schizophrenia Consortium et al., 2009).

In the statistical models usually used in GWAS practices, less common alleles (MAF < 0.05) and rare alleles (MAF < 0.01) are usually removed from association tests. However, some of them are expected to have an intermediate to high effect (**Figure 1.26**). These rare and low-frequency SNPs also explain a proportion of the missing heritability (Cirulli and Goldstein, 2010; Dickson et al., 2010; Gibson, 2012; Marouli et al., 2017). Detecting their effect is also important, because both theoretical and empirical evidence suggest that variants with strong phenotypic effects are more likely to be deleterious (Kryukov et al., 2007; Tennessen et al., 2012). In human, the vast of majority of coding variation is rare (MAF < 0.05), accounting for up to 86% of total single-nucleotide variants (SNVs) (Tennessen et al., 2012). Also, many rare human disorders are due to rare alleles with large phenotypic effects (Gibson, 2012). From the evolutionary point of view, the deleterious selection is more generally referred as background selection (Vitti et al., 2013). In order to test the effect of less common and rare SNPs, a substantially large population is required. In a recent study focusing on human height, the authors successfully identified 83 (32 of which were rare variants) height-associated coding variants with lower MAF (ranging from 10 to 4.8 %), with effects ten times greater than the average effects of common variants. Besides, these variants overlapped genes involved in monogenic growth disorders (Marouli et al., 2017) (**Figure 1.27**).

Though these research strategies dealing with rare and low frequency variants and structural variants are proposed for human genetic studies, many of them can be adapted for investigating the genetic effects of rare and low frequency variants in major crops. Apart from the aforementioned factors, gene-gene and gene-environment interactions also account for a certain degree of missing heritability (Frazer et al., 2009; Aschard et al., 2012; de

Vlaming et al., 2017), though inclusion of these interaction effects in risk-prediction models is unlikely to dramatically improve the discrimination ability of these prediction models (Aschard et al., 2012).

1.8.4.4 GWAS do not necessarily pinpoint causal variants and genes

Linkage disequilibrium (LD) is a double-edged sword in GWAS: it facilitates the initial identification of candidate regions but makes it difficult to target the causal variant(s) (Altshuler et al., 2008). In many cases of human diseases and traits, the vast majority of associations fall outside coding regions (Hindorff et al., 2009; Schork et al., 2013; Mahajan et al., 2018; Timpson et al., 2018). The majority of significant associated SNPs in tomato also fall outside coding regions, even though some candidate genes are identified in the nearby regions (Sauvage et al., 2014; Tieman et al., 2017; Bauchet et al., 2017b). Therefore, substantial additional steps are required to identify the causal variants, which are important to deepen our understanding on the genetic control and regulations of the targeted phenotypes. These approaches include regional fine-mapping (Gaulton et al., 2015; Huang et al., 2017; Mahajan et al., 2018; Schaid et al., 2018; Westra et al., 2018; Dadaev et al., 2018), transcriptional analysis (Lonsdale et al., 2013; Zouine et al., 2017; Shinozaki et al., 2018), functional validation (Bauchet et al., 2017b; Ye et al., 2017; Peng et al., 2017; Zhu et al., 2018; Du et al., 2018; Gao et al., 2019a) and evolutionary genetic analyses (Ye et al., 2017; Gao et al., 2019a) or combination of these approaches.

Despite the difficulties in interpreting the GWAS results, several progresses have been made in bridging associations to function and cause, which can be briefly divided into the following aspects:

- Improving the density of SNP arrays (Víquez-Zamora et al., 2013; Delaneau et al., 2014).
- Custom genotyping arrays targeting particular candidate regions (Ghoussaini et al., 2014; Onengut-Gumuscu et al., 2015).
- Developing a high quality reference panel for genotyping imputation (The 1000 Genomes Project Consortium et al., 2015; Browning and Browning, 2016; Wang et al., 2018).
- Using multi populations for fine-mapping (Li and Keating, 2014; Asimit et al., 2016; Mägi et al., 2017; Liu et al., 2019).
- Availability of datasets of other regulatory elements (The ENCODE Project Consortium, 2012; Zhong et al., 2013; Lonsdale et al., 2013; Andersson et al., 2014; Ward and Kellis, 2016; Zouine et al., 2017; Shinozaki et al., 2018).
- Developing new fine-mapping models, such as Bayesian models (Schaid et al., 2018).

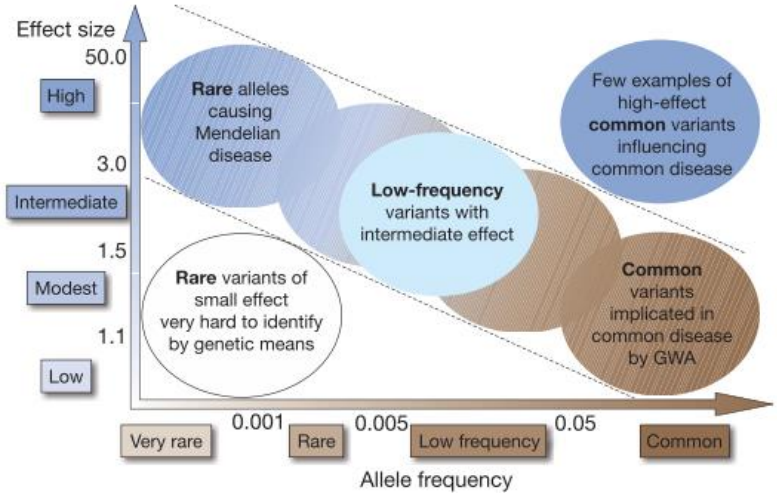


Figure 1.26 Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines (adapted from McCarthy et al., 2008; Manolio et al., 2009)

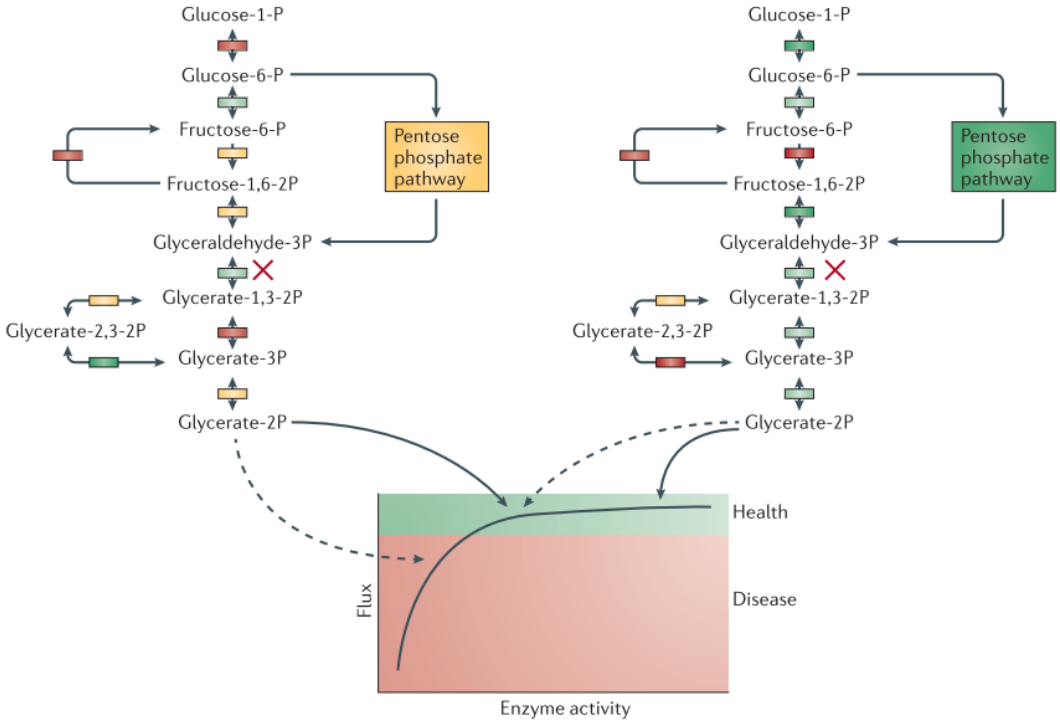


Figure 1.27 Joint effects of rare and common variants. A straightforward reconciliation of the effects of rare and common variants supposes that pervasive common variation influences the expression and activity of genes in pathways, establishing the background liability to disease that is then further modified by rare variants with larger effects. In this hypothetical example of central metabolism, standing variation results in some individuals having lower flux than others (left versus right; colored boxes imply enzyme activity differences from low activity (red shading) to high activity (green shading)), but according to standard biochemical theory, systems evolve such that most variation is accommodated within the healthy range. (adapted from Gibson, 2012).

1.8.4.5 Limitations of GWAS in detecting epistasis

Epistasis and Q × E interactions could also account for a degree of the missing heritability, which remains a challenge to detect, though it has been commonly observed in many species (Causse et al., 2007; Lehner, 2011; Mackay, 2014; Wei et al., 2014; Buchner and Nadeau, 2015; Upton et al., 2016; Soyk et al., 2017b). Statistical power should be increased (such as by using very large sample sizes) and methodological challenges should be solved for better handling the epistasis effects, such as multi-trait mixed association model (MTMM) (Korte et al., 2012), Bayesian methods (Fernando and Garrick, 2013; Xu and Guan, 2014) and artificial intelligence (Wei et al., 2014; Tam et al., 2019).

1.8.4.6 Limitations specific to SNP array-based GWAS

SNPs arrays have been shown to be effective and efficient in GWAS, and thus gained increasing popularity. When a high quality reference panel is available, genotyping imputation will greatly bridge the gaps between arrays and whole-genome sequencing (WGS). GWAS using SNP arrays and WGS has both advantages and disadvantages (**Table 1.13**). However, those limitations are less urgent or important compared to the benefits of using SNP arrays in GWAS for most of the crops, especially for those species with less importance, but of great scientific interests. Even for important crops, such as tomato and apple, whole-genome sequencing of hundreds of accessions is still quite expensive. When a core reference panel is once available, first genotyping with SNP arrays and then using imputation will be more practical for many good benefits, such as improving statistical powers in identifying new causal variants, handling less common and rare variants, missing heritability, etc. (**Figure 1.28**).

1.8.5 Post-GWAS studies

The main purpose of GWAS is to deepen our knowledge about the genetic architecture and the biological bases of trait variation by identifying the causal variants. However, due to statistical limitations, genome coverage, population size, linkage disequilibrium and other technical limitations, for many cases, it is challenging to directly identify the candidate genes, especially for those causal variants with small genetic effects and those located in regions with strong LD. Post-GWAS analysis is crucial to narrow down the candidate variants and then validate their biological functions.

With the fast implementation of GWAS, more and more significant loci associated with different targeted phenotypes in different species are available. The advancements of methodology and statistical tools make it possible to move forward to the post-GWAS era

Table 1.13 GWAS using SNP arrays versus WGS. Genetic variants can be genotyped using numerous technologies, including genome- wide single- nucleotide polymorphism (SNP) arrays (combined with statistical imputation of unobserved genotypes from population reference panels) and whole- genome sequencing (WGS). SNP arrays are the most widely used genotyping technology in GWAS, primarily owing to their lower costs, and performing WGS in very large sample sizes is currently cost- prohibitive. although the switch to WGS is likely to be inevitable with declining sequencing costs, the choice to use SNP arrays or WGS in GWAS should be made taking into consideration other factors (adapted from Tam et al., 2019).

Factor	SNP arrays	WGS
Cost	Relatively inexpensive (~US$40 per sample)	Expensive (>US$1,000 per sample)
Reliability	Reliable, highly accurate technology	Less mature and less accurate technology
Genomic coverage	• Mainly restricted to common and low- frequency variants, although imputation of rare variants is increasingly accurate (ultra- rare variants, however, can never be identified) • Biased towards variants discovered in well- studied or sequenced populations	From low- frequency, common variants to nearly all genetic variation in the genome, depending on the depth of sequencing
GWAS analysis	Well- established analytical pipeline and tools for data analysis	• Higher computational costs and greater analytical complexity • Eventually, larger multiple testing burden when conducting single- variant tests
Other considerations	Custom genotyping arrays can be extremely cost-effective	• As all variation is ascertained, fine-mapping is easier • Greater costs to store, process, analyse and interpret the resulting data
Suitable research objectives	• Analysing known or candidate associations in large cohorts • Detecting low- frequency, common variant associations in extremely large sample sizes	• Detecting and fine- mapping rare variants • Detecting ultra- rare risk variants when it becomes economically viable to perform WGS at a very large scale

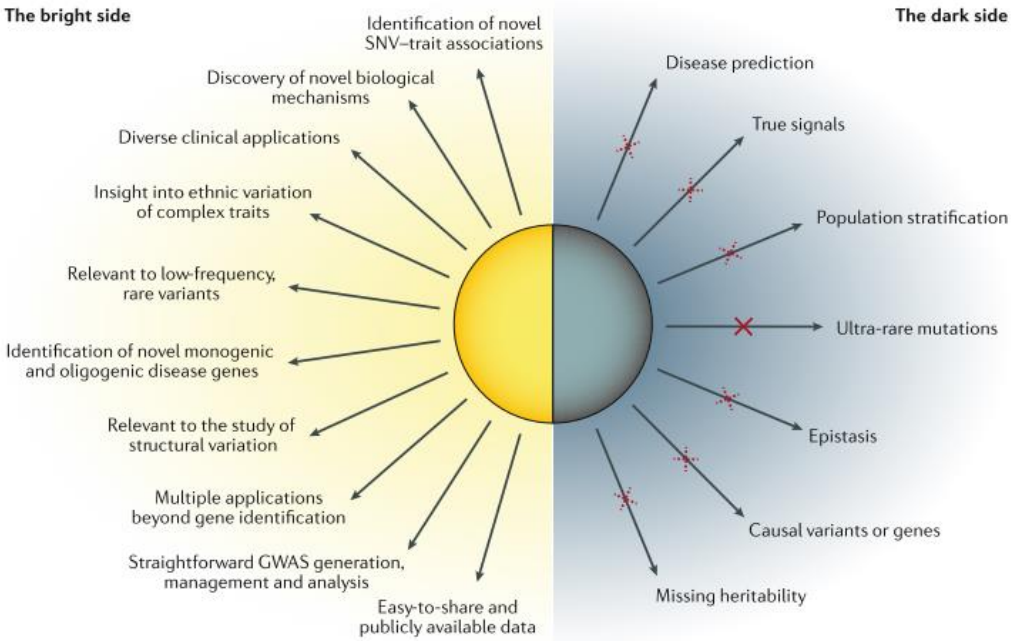


Figure 1.28 Benefits and limitations of GWAs using SNP arrays. A visual depiction of the current benefits (the bright side) and limitations (the dark side) of GWAS. The solid X indicates a permanent limitation. The dotted Xs represent limitations that have the potential to be overcome, at least to some extent, in the future (adapted from Tam et al., 2019).

from several aspects: (1) Genome-Wide Complex Trait Joint Analysis (GCTA-COJO); (2) multiple-based GWAS; (3) meta-analysis of GWAS; (4) rare alleles/variant analysis; (5) use of associated markers in the coding versus non-coding regions; (6) candidate genes/alleles identification (such as via fine mapping, localization success rate approach and conditional analysis) and annotation; (7) other non-phenotypic analyses, such as RNA-seq, eQTLs, DNA methylation and mQTL, metabolite analysis (See detailed reviews in Gupta et al., 2014; Gupta et al., 2019; Chen et al., 2019) (**Figure 1.29**). In real breeding programs, not all of the significant associated will be used and prioritization of the GWAS signals will be needed, which can be achieved in several ways and additional knowledge: (1) meta-analysis of GWAS (Evangelou and Ioannidis, 2013; Gupta et al., 2019); (2) pathway-based analysis GWAS (Wang et al., 2010b; Akula et al., 2011; Lipka et al., 2013; Richter et al., 2016; Costanzo et al., 2019); (3) methylation analysis (Lister et al., 2008; Zhong et al., 2013; Gardiner et al., 2015); (4) non-coding region analysis (such as eQTL, miRNA and lncRNA analyses) (Gong et al., 2015; Zhu et al., 2016; Engreitz et al., 2016; Zhu et al., 2018; Schaid et al., 2018); (5) integration of multi-omic analyses, such as transcriptome-wide association studies (TWAS) (Gusev et al., 2016; Kremling et al., 2018; Mancuso et al., 2019; Wainberg et al., 2019) and metabolite-based GWAS (mGWAS) (Luo, 2015; Alseekh and Fernie, 2018; Fernie and Gutierrez-Marcos, 2019; Chen et al., 2019); (6) haplotype-based analyses (Lorenz et al., 2010; Hamblin and Jannink, 2011; Hao et al., 2012; Lu et al., 2012).

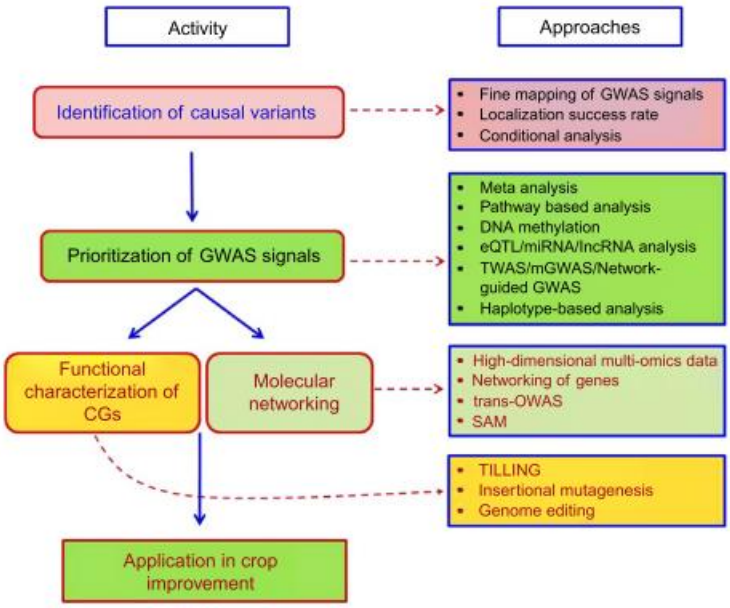


Figure 1.29 A flow chart showing the activities, which can be carried out in the post-GWAS era (adapted from Gupta et al., 2019).

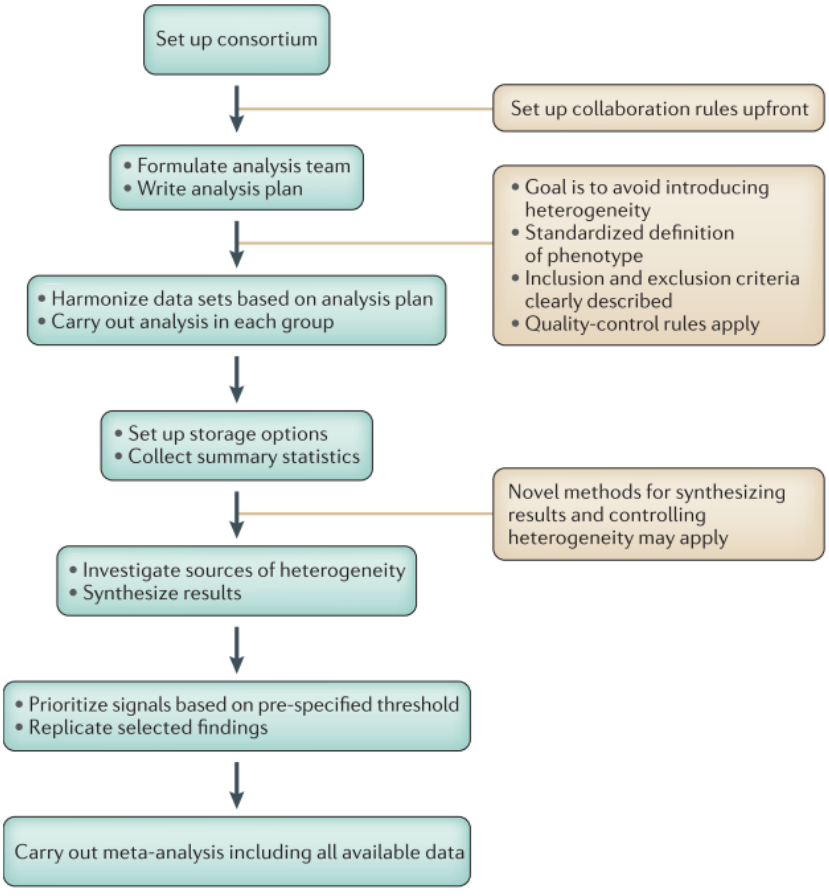


Figure 1.30 Stages in a meta-analysis. A typical plan for a meta-analysis of genome-wide and next-generation sequence data (adapted from Evangelou and Ioannidis, 2013).

1.9 Meta-analysis of genome-wide association studies

1.9.1 General introduction

Meta-analysis is the quantitative synthesis of research results focusing on the same research question and investigating their diversity across different studies (Ioannidis et al., 2007; Gurevitch et al., 2018). Though this idea has been proposed for over a century (Pearson, 1904), it only began to affect scientific research to a large extent till the 1970s (Smith and Glass, 1977; Glass, 2015). For example, Smith and Glass (1977) performed a meta-analysis of nearly 400 controlled evaluations of psychotherapy and the results provided convincing evidence of the efficacy of psychotherapy. Since then, meta-analysis has grown to a major academic industry in the past 40 years, especially in the field of medicine (Glass, 2015). The main purpose of meta-analysis of GWAS is either replication analyses of previously reported associations or discovering new significant associations (Gupta et al., 2019).

However, meta-analysis is only applied to GWAS since 20 years and has become a popular approach for discovering genetic variants associated with targeted traits (Begum et al., 2012; Evangelou and Ioannidis, 2013; Panagiotou et al., 2013; Wijmenga and Zhernakova, 2018; Gurevitch et al., 2018). There are several advantages of using meta-analyses in GWAS:

- **Single GWAS may only explain a small proportion of heritability for most traits.** Though many common significant associated loci can be identified in a single GWAS, for many traits, these associations only explain a limited proportion of the total heritability, such as human adult height (Visscher, 2008; Yang et al., 2010; Wood et al., 2014; Yang et al., 2017). A similar phenomenon is also observed in other species, such as tomato, where only a limited number of associations are identified, and only moderate level of heritability is explained, depending on the traits (Sauvage et al., 2014; Tieman et al., 2017; Bauchet et al., 2017b).
- **Single GWAS is limited in identifying large number of alleles with minor genetic effect.** In order to identify more common alleles with small effects, substantially large sample sizes are required (Ioannidis et al., 2006; Moonesinghe et al., 2008; Chapman et al., 2011).
- **Single GWAS cannot handle cross-study heterogeneity.** The significant associated loci from different GWAS might differ in terms of both numbers and locations, and the difference is usually referred to heterogeneity (non-random variance across

studies). The heterogeneity can be caused by several factors, including phenotyping measurements, populations from different ancestry, population stratification, G × E interactions, linkage disequilibrium, genotyping platforms and genotyping imputation, etc. (see detailed explanations in Begum et al., 2012; Evangelou and Ioannidis, 2013; Panagiotou et al., 2013; Gupta et al., 2019).

- **Meta-analysis of GWAS versus mega-analysis of GWAS.** When performing meta-analysis, only summary result data from individual GWAS are needed. If the individual-level data from all panels are also available and the cross-study heterogeneity is properly managed, it is also possible to first combine the phenotypic and genotypic data and then perform a single GWAS, which is technically referred to mega-analysis of GWAS. To do so, phenotypic data should be properly managed (the difference between different panels could reach several folds) and quality control is also required after combining genotypic data. Both simulations and real dataset have shown that meta-analysis has a similar statistical performance compared to mega-analysis (Lin and Zeng, 2010; Panagiotou et al., 2013). Therefore, even when all the individual-level data is available, it is not necessary to re-analyze the raw data again, as only summary results are needed for meta-analysis.

There are several important stages to perform a well-designed GWAS meta-analysis and several reviews are available (de Bakker et al., 2008; Zeggini and Ioannidis, 2009; Thompson et al., 2011; Evangelou and Ioannidis, 2013). Briefly, the performance of GWAS meta-analysis include four main steps (**Figure 1.30**)

1.9.2 Statistical models

Depending on the hypothesis of genetic effects, meta-analysis was first performed based on two general models: fixed-effect model (the genetic effects of markers are the same across studies) and random-effect model (the genetic effects of markers are different across studies), which idea has been proposed back in the late 1930s (Yates and Cochran, 1938) and then formalized and generalized later (Cochran, 1954). A detailed statistical explanation of the simplest GWAS meta-analysis as well as the measurement of heterogeneity is demonstrated in **Figure 1.31**.

Apart from the simplest *P*-value meta-analysis, there are some other more complex models, including fixed effects, random effects, Bayesian approach, multivariate approaches and other extensions (**Table 1.14**). Different models have their distinct advantages and disadvantages and have been integrated into different software for applications (**Table 1.15**). In real examples, fixed-effect model is the most commonly used approach and METAL and

R packages (such as Metafor, rmeta and CATMAP) are the two most popular software used for these analyses (**Figure 1.32**).

1.9.3 Prospects of meta-analysis of genome-wide association studies

For important crops, such as rice, maize and wheat, where the genome sequencing is usually conducted by international consortia and data centralised, it should be possible to apply meta-analysis (Gupta et al., 2019). Better data management and newly developed computing technologies, such as cloud-based platform easyGWAS (Grimm et al., 2017), will accelerate its application.

In addition, there are some other interesting prospects of GWAS meta-analysis based on human genetics, which could also be interesting and applicable for crops (**Table 1.16**) (adapted from Panagiotou et al., 2013).

1.10 Potential benefits of using haplotypes

Our understanding of the genetic architecture of agronomical traits is guided by technical (i.e. sequencing), analytical (i.e. statistics) and theoretical advances (i.e. population and quantitative genetics). Up to now, the vast majority of marker trait-associations was revealed using QTL and GWAS mapping. The later approach relies on the linkage disequilibrium (LD) between the gene(s) that control the variance of the trait and a single molecular marker. While being successful for detecting loci of large effect, it remains limited to decipher the additional medium to low effect loci. In addition, a strong knowledge of the structure of LD is required, particularly the distance to which LD extends and how much it varies from one chromosomal region to another in the population under study. Switching from single marker to multiple markers has benefited to the discovery of LD 'blocks', namely haplotypes, carrying the (un)favorable alleles to select for.

The simplest genome-wide association study (GWAS) meta-analysis approach is to combine *P* values using Fisher's method. The formula for the statistic is

$$X^2 = -2\sum_{i=1}^{k}\log(P_i)$$

where P_i is the *P* value for the i^{th} study, and *k* is the number of studies in the meta-analysis. Under the null hypothesis, X^2 follows a χ^2 distribution with $2k$ degrees of freedom. The *Z* scores meta-analysis can be implemented using the equation

$$Z = \frac{\sum_i Z_i w_i}{\sqrt{\sum_i w_i^2}}$$

where w_i is the square root of sample size of the i^{th} study and

$$Z_i = \Phi^{-1}\left(1-\frac{P_i}{2}\right)^*(\text{effect direction for study } i)$$

where Φ is the standard normal cumulative distribution function. For fixed effects models, inverse variance weighting is widely used. The weighted average of the effect sizes can be calculated as

$$\hat{\theta}_F = \frac{\sum_i w_i \hat{\theta}_i}{\sum_i w_i}$$

and the variance is

$$\text{var}(\hat{\theta}_F) = \frac{1}{\sum_i w_i}$$

where $\hat{\theta}_i$ is the i^{th} study normalized effect (for example, logarithm of odds ratio or β-coefficient for a logistic regression for a binary phenotype or mean difference or standardized mean difference for a continuous phenotype), and w_i is the reciprocal of the estimated variance of the effect study. The random effects model incorporates the between-study variance of heterogeneity, and therefore the weight for the random effects model is calculated as

$$w_i^R = \frac{1}{\left(\frac{1}{w_i} + \hat{\tau}^2\right)}$$

where

$$\tau^2 = \frac{(Q-(k-1))}{\left(\sum_i w_i - \left(\frac{\sum_i w_i^2}{\sum_i w_i}\right)\right)}$$

and *Q* is Cochran's *Q* statistic, which is given by

$$Q = \sum_i w_i(\hat{\theta}_i - \hat{\theta}_F).$$

Another popular heterogeneity metric, I^2, is given by

$$I^2 = \frac{100^*(Q-(k-1))}{Q}.$$

The multivariate meta-analysis approaches are based on the calculation of a variance–covariance matrix for the correlated phenotypes or the single-nucleotide polymorphisms in linkage disequilibrium that will allow the calculation of the marginal effects. In cross-phenotype meta-analysis, the developed statistic measures the likelihood of the null hypothesis, given the data. The test is asymptotically distributed as

$$\chi^2_{df=1}.$$

Figure 1.31 Statistical properties of common GWAS meta-analysis approaches (adapted from Evangelou and Ioannidis, 2013).

Table 1.14 Summary of methods for meta-analysis of genome-wide data (adapted from Evangelou and Ioannidis, 2013).

Method	Description	Advantages	Disadvantages	Main software used
P value meta-analysis	Simplest meta-analytical approach	Allows meta-analysis when effects are not available	Direction of effect is not always available; inability to provide effect sizes; difficulties in interpretation	METAL, GWAMA, R packages
Fixed effects	Synthesis of effect sizes. Between-study variance is assumed to be zero	Effects readily available through specialized software	Results may be biased if a large amount of heterogeneity exists	METAL, GWAMA, R packages
Random effects	Synthesis of effect sizes. Assumes that the individual studies estimate different effects	Generalizability of results	Power deserts in discovery efforts; may yield spuriously large summary effect estimates when there are selection biases	GWAMA, R packages
Bayesian approach	Incorporates prior assessment of the genetic effects	Most direct method for interpretation of results as posterior probabilities given the observed data	Methodologically challenging; GWAS-tailored routine software not available; subjective prior information used	R packages
Multivariate approaches	Incorporates the possible correlation between outcomes or genetic variants	Increased power can identify variants that conventional meta-analysis do not reveal using the same data sets	Computationally intensive; software not available for all analyses; some may require individual-level data	GCTA for multi-locus approaches
Other extensions	A set of different approaches that allows for the identification of multiple variants across different diseases	Summary results of previous meta-analyses can be used	May need additional exploratory analyses for the identification of variants; prone to systematic biases	Software developed by the authors of the proposed methodologies

GCTA, genome-wide complex trait analysis; GWAS, genome-wide association study.

Table 1.15 Comparison of meta-analysis software package (adapted from Evangelou and Ioannidis, 2013).

	METAL	GWAMA	MetABEL	PLINK	R packages
Ability to process files from GWAS analysis tools; software used	No	Yes; SNPTEST, PLINK	Yes; ABEL	Yes; PLINK	No
Fixed effects implemented?	Yes	Yes	Yes	Yes	Yes
Random effects implemented?	No	Yes	No	No	Yes
Heterogeneity metrics generated	Q, I^2	Q, I^2	Q, I^2	Q, I^2	Q, I^2
Graphical illustration of meta-analysis results	No	Manhattan and QQ plots	Forest plots	No	Yes

GWAS, genome-wide association study.

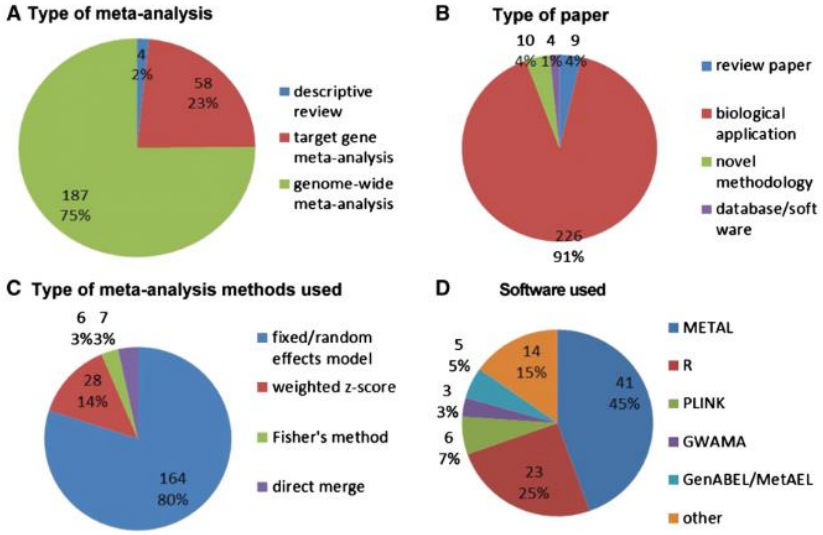


Figure 1.32 Summary of GWAS meta-analysis review: (A) type of meta-analysis; (B) type of paper; (C) type of meta-analysis method; (D) software used (adapted from Begum et al., 2012).

Table 1.16 Future issues related to meta-analysis of GWAS.

1. Meta-analysis of GWA data is expected to become even more common in the future as individual teams worldwide create more consortia and share data.
2. Given the widespread role of consortia, mega-analysis of individual GWA data could increase the power to detect subtle effects, especially for rare variants, and could also detect cryptic relatedness, overcoming some of the limitations of GWA meta-analysis.
3. Detection of rare variants will be greatly enhanced by advancements in imputation of these variants and by sequencing of thousands of individuals within the consortial setting.
4. The potential for clinical translation of this information will depend largely and idiosyncratically on the clinical features of specific diseases and on available treatments and preventive measures.
5. Continued meta-analysis of larger samples—a key approach, in addition to imputation and sequencing, for extending the analysis to rarer variants—is likely to continue to yield useful information, although some methodological changes may be necessary.
6. Contributions from individual teams to shared efforts and recognition for collaborative work within a consortial setting are vital for maximizing the potential of future discoveries.

Haplotypes are the particular combinations of alleles observed on a single chromosome, or part of a chromosome in a given population (Gabriel et al., 2002; Belmont et al., 2003). Haplotype blocks are the regions where there is little evidence for historical recombination and within which only a few common haplotypes are observed (Gabriel et al., 2002). Alleles within the same haplotype block are more likely to be inherited together (Farashi et al., 2019). Genotyping only a few, carefully chosen tag-SNPs should provide enough information to identify the common haplotypes (**Figure 1.33**) (Daly et al., 2001; Johnson et al., 2001; Belmont et al., 2003; Hafler and Jager, 2005).

1.10.1 Using haplotypes in identifying selective sweeps

Sabeti et al. (2002) introduced extended haplotype homozygosity (EHH) to detect recent positive selection in human populations by analyzing long-range haplotypes. EHH is defined as the probability of two randomly chosen chromosomes carrying the same core haplotype that are identical by descent (Sabeti et al., 2002)**.** A core haplotype at a locus of interest was first identified and the decay of its association to alleles at different distances from the focal locus was calculated to identify the selective signal. An unusually high EHH frequency within the core haplotypes indicated the presence of a mutation with faster prominence than expected under neutral selection (Sabeti et al., 2002)**.** Significant evidence of selection was observed on different genes. However, other common approaches, including Tajima's D-test (Tajima, 1989), Fu and Li's D-test (Fu and Li, 1993), Fay and Wu's H-test (Fay and Wu, 2000), the Ka/Ks test (Hughes and Nei, 1988), the McDonald and Kreitman test (McDonald and Kreitman, 1991), and the Hudson-Kreitman-Aguadè (HKA) test (Hudson et al., 1987) (Sabeti et al., 2002)**,** could not detect the same significant selection signals**.**

The long-range haplotype (LRH) looks for haplotypes that are extended and common by comparing a haplotype's frequency to its relative EHH at various distances (Vitti et al., 2013). The integrated haplotype score (iHS) compares the area under the curve defined by EHH for the derived and ancestral variant and can capture both extreme EHH for a short distance and the moderate EHH for a longer distance (Vitti et al., 2013). The cross-population extended haplotype homozygosity (XP-EHH) can detect positive selections by comparing the haplotype lengths between populations (Sabeti et al., 2007). XP-EHH is more useful to detect those selective sweeps in near fixation within one population, while iHS is more powerful to detect incomplete or ongoing selections (**Figure 1.34**).

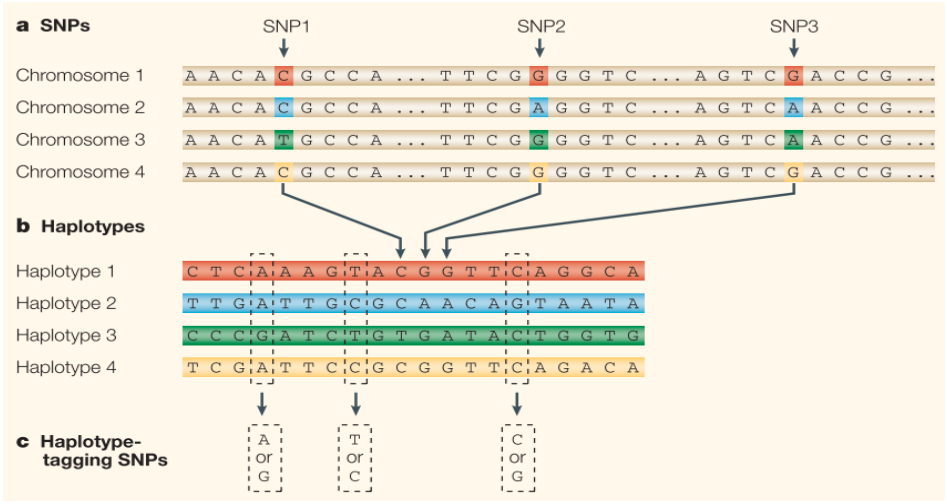


Figure 1.33 SNPs, haplotypes and haplotype-tagging SNPs (adapted from Hafler and De Jager 2005).

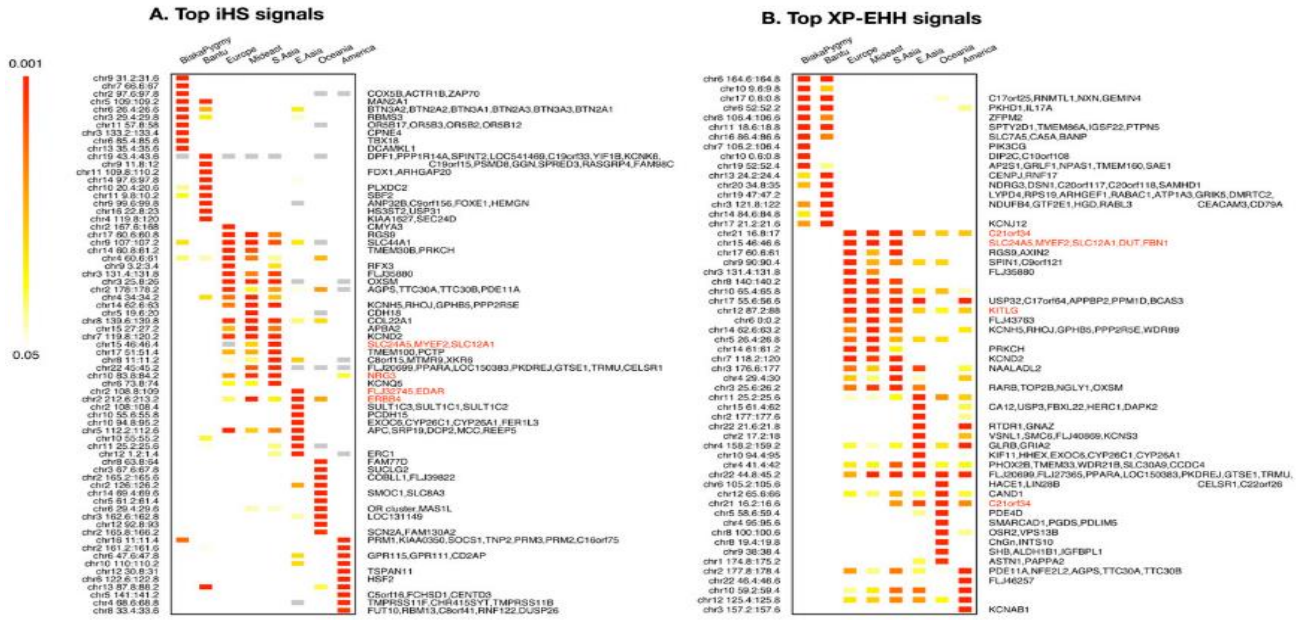


Figure 1.34 Top 10 iHS (A) and XP-EHH (B) signals by population cluster (adapted from Pickrell et al., 2009).

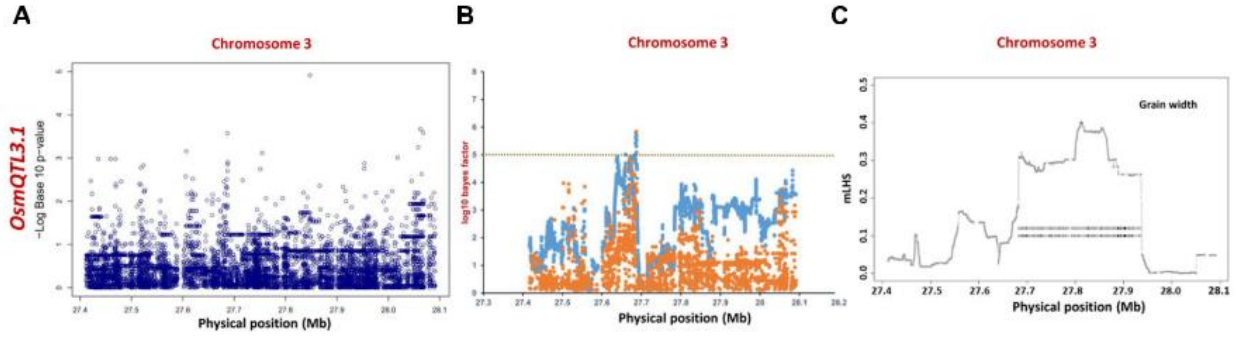


Figure 1.35 Manhattan plots showing significant trait associations identified using 365 *indica* accessions. (A) Individual SNP-based regional association analysis, (B) haplotype-based regional association analysis and (C) trait-associated LD blocks calculated for each of the associated haplotype. (adapted from Daware et al., 2017).

1.10.2 Using haplotypes in genome-wide associations

Haplotype-based analyses examine groups of SNPs rather than individual SNPs and enhance the statistical detection power for GWAS (Khatkar et al., 2007; Xu and Guan, 2014; Negro et al., 2019). However, to my best knowledge, there is no example of haplotype-based associations in tomato. It will thus be interesting to validate whether it could be helpful to apply haplotype-based associations in tomato.

1.10.3 Using haplotypes as an alternative of linkage disequilibrium

One crucial step in association study is trying to find the promising candidate genes for the targeted phenotypes for either validating the candidate genes (i.e. through knockout) or developing molecular markers for breeding purposes. In tomato, linkage disequilibrium (LD) was frequently adopted to choose the window size to search for candidate genes at a given threshold, such as $R^2 > 0.3$ chosen after resampling (Albert et al, 2016), $R^2 > 0.7$ (Bauchet *et al.*, 2017), or $R^2 > 0.8$ (Tieman et al., 2017) chosen more or less empirically. Even within the window size at a high threshold, the LD between the focal SNP and close SNPs does not decay gradually as many SNPs in strong and weak LD could appear in the same region(Zhao et al., 2019), which makes it difficult to choose the optimal threshold to look for candidate genes. In contrast, mLHS between nearby SNPs and the focal SNP decrease more gradually on both sides (Xu and Guan, 2014). For example, Daware et al. (2017) applied mLHS for identifying metaQTLs associated with grain size in rice, and they successfully identified several major QTLs. In addition, they showed that mLHS was helpful in choosing the candidate LD block where a dramatic decrease of mLHS was observed (**Figure 1.35**)

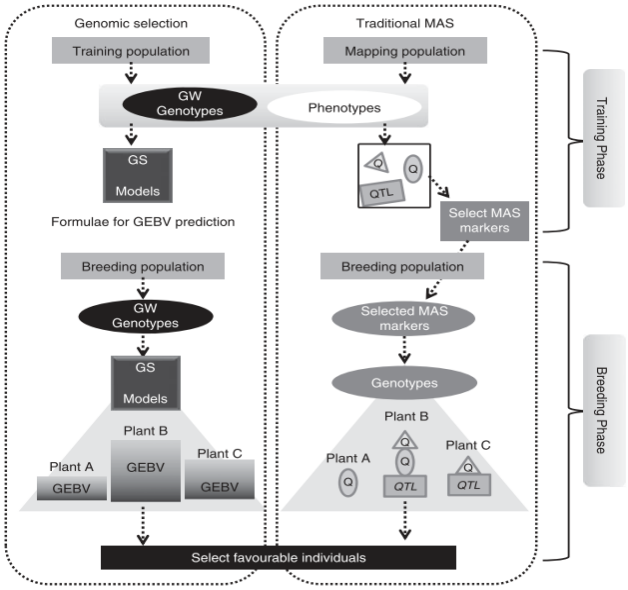


Figure 1.36 Schemes of genomic selection (GS) (left) and traditional MAS for the selection of quantitative traits (right). Both GS and traditional MAS contained training and breeding phases. In the training phase, quantitative trait loci (QTLs) are identified in traditional MAS to produce formulae for genomic estimated breeding value (GEBV) prediction, i.e. GS models. In the breeding phase, favorable individuals are selected based on the genotypes of the selected markers in MAS, whereas GEBVs are used for selection in GS (adapted from Nakaya and Isobe, 2012).

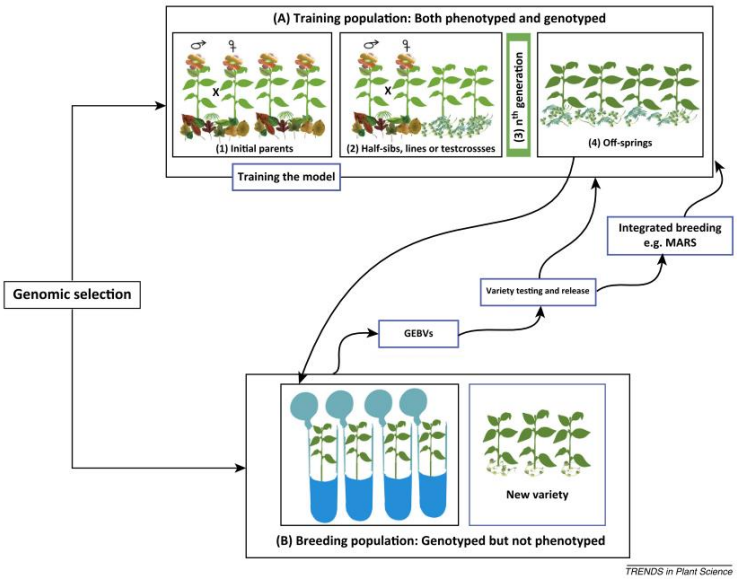


Figure 1.37 Genomic selection (GS) steps and applications in plant breeding. (A) The training population (TP) is the population set being phenotyped and genotyped. The initial parents used to produce the next generation (1) by crossing selected parents, half-siblings, lines, or test crosses are included in (2), and continue until the nth generation (3) that delivers offspring to be used as a validation set to train the model against the training sets in the TP (4). (B) Breeding populations (BP) are only genotyped but not phenotyped. These can also comprise new varieties introduced as BP but related to the TP. The breeding lines with highest genomic estimated breeding values (GEBVs) are selected and this will routinely continue as a turn cycle of GS to the TP. The selected candidates with high GEBVs can be integrated with other breeding schemes, such as marker-assisted recurrent selection (MARS) to introgress the required agro-morphological traits(s) to well-adapted crop species (adapted from Desta and Ortiz, 2014).

1.11 The application of Genomic selection in crops: towards phenotypic prediction

1.11.1 Principle of genomic prediction

The fast development of low-cost molecular markers, especially SNPs, has greatly promoted their applications in crop breeding. Genome-assisted breeding can be roughly classified into two categories. The first class includes marker-assisted selection (MAS) (Ribaut and Hoisington, 1998; Xu and Crouch, 2008; Gupta et al., 2010) and marker-assisted recurrent selection (MARS) (Ribaut and Hoisington, 1998; Xu and Crouch, 2008; Gupta et al., 2010; Foolad and Panthee, 2012). The second class is genomic selection (GS), which has been proposed first for animal breeding almost 20 years ago (Meuwissen et al., 2001). Its basic principles rely on the fact that many traits are controlled by a large number of QTL with low effect. Both linkage mapping and GWAS have limitations in identifying and quantifying these small effects and also rare QTL alleles or associations that are highly susceptible to environmental conditions (Crossa et al., 2017). In MAS and MARS practices, markers that are significantly associated with targeted phenotypes are used as indicators for introgression. In contrast, GS uses the genomic estimated breeding values (GEBVs) calculated by genomic prediction (GP) to guide the selection of promising candidate individuals (**Figure 1.36**). GS uses the genotypic and phenotypic data in a training population to predict the GEBVs of individuals in a testing population that have been only genotyped (**Figure 1.37**). GS takes into account the genetic effect of all molecular markers by summing the total marker effects of GEBV (Heffner et al., 2009) and is expected to address small effect genes that cannot be captured by traditional MAS or MARS (Hayes et al., 2009; Nakaya and Isobe, 2012). The main advantages of GS include notably cost reduction and time saving compared to phenotype-based selection (Crossa et al., 2017). At the same time, there are some limitations of GS, including 1) high cost of genotyping the training and testing populations, 2) low to moderate prediction accuracy, 3) challenges in handling G × E interactions, 4) difficulty in implementing in real breeding programs, etc.

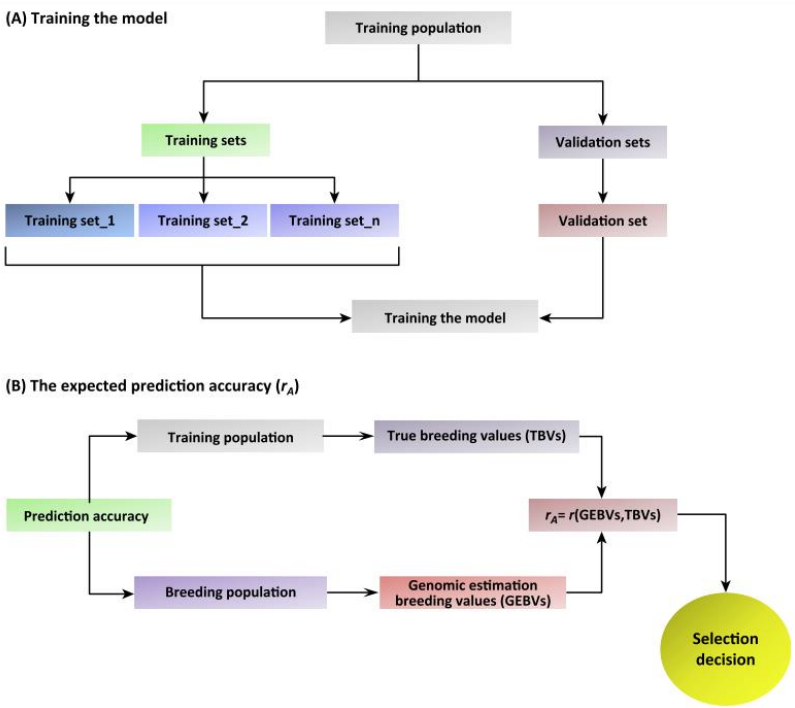


Figure 1.38 Training the model and prediction accuracy. (A) Training the model: different groups of the training population (TP) are represented as training sets to correlate against the validation set. The training sets along with validation set are used for cross-validation with K-folds to train the prediction models. (B) Prediction accuracy: the selected prediction model(s) are used to estimate the expected prediction accuracy or genomic estimated breeding values (GEBVs) in the selected candidates of the target species. (adapted from Desta and Ortiz, 2014).

Table 1.17 Main features of genome-wide prediction models (adapted from Desta and Ortiz, 2014).

Model acronym[a]	Features
RR-BLUP	Assumes that all markers have equal variances with small but non-zero effect
	Applies homogeneous shrinkage of predictors towards zero, but allows for markers to have uneven effects
	Computed from a realized-relation matrix based on markers
	Some QTL are in LD to marker loci, whereas others are not
LASSO	Combines both shrinkage and variable selection methods
	RR-BLUP does not use variable selection, but outsmarts LASSO when there is multicolinearity between the predictors
EN	Double regularization using ℓ_1 and ℓ_2 penalty norms combines the merited features of these norms to confront the challenge of high-dimensional data
BRR	Induces homogeneous shrinkage of all marker effects towards zero and yields a Gaussian distribution of marker effects
	Similar to RR-BLUP, there is a problem of QTL linkages to the marker lloci
BL	Applies to both shrinkage and variable selection
	Has an exponential prior on marker variances resulting in a double exponential (DE) distribution
	The DE distribution has a higher mass density at zero and heavier prior tails compared with a Gaussian distribution
Bayes A	Utilizes an inverse chi-square (χ^2) on marker variances yielding a scaled *t*-distribution for marker effects
	Similar to BL and in contrast to BRR, it shrinks tiny marker effects towards zero and larger values survive
	Has a higher peak of mass density zero compared with the DE distribution
Bayes B	Similar to Bayes A, uses an inverse χ^2 resulting in a scaled *t*-distribution
	Unlike Bayes A, utilizes both shrinkage and variable selection methods
	When $\pi = 0$, then it is similar to Bayes A
Bayes C	Applies both shrinkage and variable selection methods
	Characterized by a Gaussian distribution
	Bayes B and Bayes C consist of point of mass at zero in their slab priors
Bayes Cπ	A modified variant of Bayes B
	Used to alleviate the shortcomings of Bayes A and Bayes B
	Unlike Bayes B, π is not fixed, but estimated from the data
RKHS	Based on genetic distance and a kernel function with a smoothing parameter to regulate the distribution of QTL effects
	Effective for detecting nonadditive gene effects
RF	Uses the regression model rooted in bootstrapping sample observations
	Takes the average of all tree nodes to find the best prediction model
	Captures the interactions between markers

[a]EN, elastic net; RF, random forest; RHKS, reproducing kernels Hilbert spaces regression.

1.11.2 Factors influencing the prediction accuracy

The accuracy of genomic prediction (GP) is usually measured during a cross validation experiment by comparing the predicted GEBVs with observed true breeding values (**Figure 1.38**) using the level of correlation between these estimates. Several factors influence the accuracy of GP, including the size, structure and genetic diversity of the training population, trait heritability, the number and distribution of molecular markers, linkage disequilibrium, prediction model and number of QTLs (Isidro et al., 2015; Spindel et al., 2015; Duangjit et al., 2016; Kooke et al., 2016; Yamamoto et al., 2016; Boison et al., 2017; Crossa et al., 2017; Yamamoto et al., 2017; Minamikawa et al., 2017; Müller et al., 2017; Crain et al., 2018; Edwards et al., 2019; Sun et al., 2019; Mangin et al., 2019). In order to improve the prediction accuracy, complex GS models were developed in order to handle different factors, such as the multi-trait and multi-environment G × E interactions (Montesinos-López et al., 2016; Fernandes et al., 2018). To date, a large number of GS models are available and the prediction accuracy vary according to traits and conditions (Heslot et al., 2012; Jonas and de Koning, 2013; Yamamoto et al., 2016; Yamamoto et al., 2017).

1.11.3 Genomic prediction models and applications in tomato

Many genomic prediction models have been proposed to improve the prediction accuracy (**Table 1.17**). Broadly, these prediction models can be categorized as parametric regressions and non-parametric regressions. The former can be further categorized as penalized approach and Bayesian approach (**Figure 1.39**).

The first GP in tomato was on a simulation-based breeding design and phenotypic prediction, where a theoretical method was proposed to apply GS to actual breeding of simultaneous improvement of yield and flavor (Yamamoto et al., 2016). Briefly, 96 big-fruited tomato varieties were selected and 20 agronomic traits were measured, including yield, quality, physiological disorder of fruit and others, with the broad-sense heritability ranging from 0.10 to 1.00. Seven GP models were compared, including five linear methods, Ridge regression (RR) (Endelman, 2011), Bayesian Lasso (BL) (Park and Casella, 2008), extended Bayesian Lasso (EBL) (Mutshinda and Sillanpää, 2010), weighted Bayesian shrinkage regression (wBSR) (Hayashi and Iwata, 2010), and Bayes C (Habier et al., 2011), and two nonlinear methods,

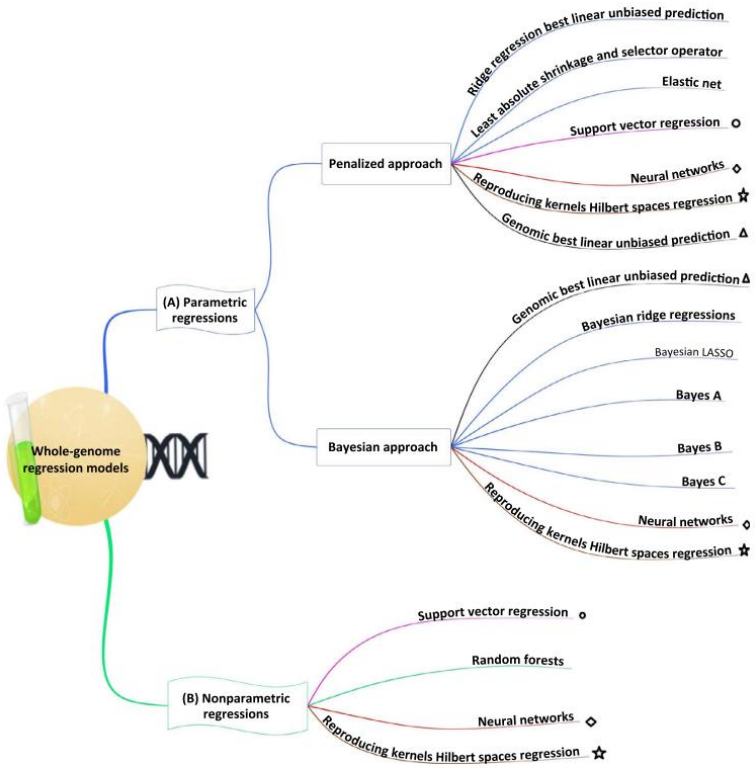


Figure 1.39 Classification of whole-genome regression models. Broadly, these models are categorized as parametric regressions (A) (in blue) and nonparametric regressions (B) (in green). Models that are indicated by multiline colors are additionally tagged with symbols for further identification because they are classified in different whole- genome regressions (as adapted from Desta and Ortiz, 2014).

Table 1.18 Accuracy of genomic estimated breeding values (GEBVs) in traits evaluated in this study. Accuracy was evaluated as a Pearson's correlation coefficient between phenotypic values and GEBVs from leave-one-out cross validation. Bold italics indicate the highest value in the same trait. RR, Ridge regression; BL, Bayesian Lasso; EBL, Extended Bayesian Lasso; wBSR, Weighted Bayesian shrinkage regression; RKHS, Reproducing kernel Hilbert space regression; RF, Random forest (adapted from Yamamoto et al., 2016).

Trait	RR	BL	EBL	wBSR	Bayes C	RKHS	RF
Percentage of fruit set	0.238	0.244	0.220	***0.319***	0.290	0.256	0.207
Total fruit weight	0.590	0.591	0.576	0.602	***0.606***	0.599	0.472
Average fruit weight	0.461	0.424	***0.461***	0.450	0.450	0.455	0.302
Percentage of marketable fruits	0.199	0.206	0.133	0.238	***0.256***	0.225	0.017
Total marketable fruit weight	0.403	0.381	0.377	0.400	***0.414***	0.413	0.118
Average marketable fruit weight	0.437	0.387	***0.482***	0.429	0.420	0.408	0.221
Soluble solids content	0.772	0.768	***0.807***	0.779	0.778	0.771	0.679
Pericarp colour	0.482	0.371	0.456	0.387	0.465	***0.514***	0.498
Style scar	***0.513***	0.493	0.496	0.505	0.511	0.495	0.508
Percentage of blossom-end rot fruits	−0.064	0.113	0.015	0.030	0.153	***0.206***	0.012
Percentage of irregular-shaped fruits	0.454	0.439	0.406	0.448	***0.465***	0.427	0.413
Percentage of cracked fruits	−0.018	0.034	−0.240	0.095	0.022	0.117	***0.422***
Percentage of small fruits	−0.048	0.131	−0.030	0.018	***0.156***	−0.103	−0.049
Leaf length	0.361	0.366	0.244	0.307	***0.384***	0.346	0.365
Leaf width	0.282	0.307	0.302	0.328	***0.335***	0.305	0.213
Stem width	0.336	***0.353***	0.337	0.340	0.347	0.342	0.258
Height to the first truss	0.397	0.409	0.381	***0.415***	0.399	0.355	0.394
Number of flowers	0.332	0.367	0.350	0.323	***0.376***	0.331	0.343
Days to flowering	0.576	0.584	0.581	0.580	0.591	***0.703***	0.653
Number of leaves under the first truss	0.285	0.275	0.212	0.311	0.304	***0.326***	0.276

reproducing kernel Hilbert space regression (RKHS) (Gianola and Kaam, 2008) and random forest (RF) (Breiman, 2001). The highest prediction accuracy for different traits varied and the accuracy of Bayes C was highest for up to eight traits, ranking the best among all models (**Table 1.18**). These results demonstrated the potential benefits of using Bayesian models in increasing the accuracy of genomic prediction. Some individuals with high GEBV of total fruit weight and soluble solid contents were selected as parents to simulate later generations. Then, simulations demonstrated that after five generations, the simulated GEBVs were comparable with parental varieties. Breeding target traits could also have impacts on some non-target traits. In particular, simultaneous selection for yield and flavor resulted in morphological changes, such as the increase in plant height. These results demonstrated the benefits of simulations for real breeding design.

Yamamoto et al., (2017) then used a population of big-fruited F1s to construct the GS models to assess its potential for the improvement of total fruit weight and soluble solid content in a practical experiment. By testing six GS models and 10-fold cross-validation, the prediction accuracy for soluble solid content was higher than for total fruit weight. GBLUP and BL had significantly higher predictability compared to other models for soluble solid content. In contrast, RKHS and RF had significantly higher predictability compared to other linear models for total fruit weight. The authors further developed four progenies to predict trait segregations and demonstrated that all individuals in the four progenies were genetically distinct from each other but intermediate between their parental varieties. However, the genetic diversity within each population was much lower compared to the training population.

Duangjit et al., (2016) investigated the impact of some key factors on the efficiency of GP in tomato, including the size of training population, the number and density of SNPs and individual relatedness. Based on the analysis of 163 tomato accessions, the optimal size of the training population was 122. The prediction accuracy also increased with the increase of marker density and number, but weakly. Individual relatedness also influenced the prediction accuracy, and predictions were better in closer individual relatedness. Based on this population, only about 2300 SNPs distributed every 0.1 cM were efficient to reach similar prediction accuracy, compared with using all SNPs. However, there were some limitations in this study: 1) it only tested the ridge regression best linear unbiased prediction (rrBLUP) statistical model (Endelman, 2011); 2) the number of SNPs was relatively small and the genomic coverage in certain genomic regions was quite limited(Zhao et al., 2019)(Zhao et al.,

2019); 3) Population structure existed and the number of wild accessions was quite small compared to cherry and large-fruited tomato accessions.

1.11.4 Bayesian models in genomic prediction

Bayesian models, especially BayesC, have demonstrated their advantages in improving the prediction accuracy (Yamamoto et al., 2016). Pérez et al. (2014) reported an BGLR (Bayesian generalized linear regression) statistical R package for genome-wide regression and prediction analyses, which was an extension of BLR package (Bayesian linear regression) (Pérez et al., 2010). BLR can only handle continuous outcomes, while BGLR extends BLR by allowing regressions for binary and censored outcomes (Pérez et al., 2014). This package includes several prior densities for regression coefficient, such as flat (fixed effect), gaussian (Bayesian ridge regression, BRR), scaled-t (BayesA), double exponential (BL), Gaussian mixture (BayesC) and scaled-t mixture (BayesB). These Bayesian models can also be further extended to handle G × E interactions (Cuevas et al., 2017).

1.11.5 Haplotype-based genomic prediction

Most of the GS models rely on marker-based information and are unable to exploit local epistatic interactions among markers. Molecular markers can also be combined into haplotypes by combining linkage disequilibrium and linkage analysis to improve prediction accuracy (Clark, 2004; Calus et al., 2008; Jiang et al., 2018), which has been recently shown especially in animals (Calus et al., 2008; Cuyabano et al., 2014; Cuyabano et al., 2015a; Cuyabano et al., 2015b; Hess et al., 2017; Karimi et al., 2018). Haplotype-based genome-wide prediction models make it possible to exploit local epistatic effects inside haplotype blocks (Wang et al., 2012b; de Los Campos et al., 2013; He et al., 2016; Jiang et al., 2018). The benefits of haplotype-based GS remain to be investigated in major crops (Jiang et al., 2018).

Simulations and analyses of cattle data showed that haplotype-based genomic prediction further improve the prediction accuracy (Calus et al., 2008; Villumsen et al., 2009; Cuyabano et al., 2014; Cuyabano et al., 2015a; Hess et al., 2017; Jiang et al., 2018; Karimi et al., 2018). Villumsen et al. (2009) showed that with the increase of the length of haplotypes, prediction accuracy was first increased gradually and then decreased gradually. Cuyabano et al. (2014) demonstrated that LD-based haplotype blocks increased the prediction accuracy compared with the commonly used individual SNP approach in the Nordic Holstein population. In

addition, the prediction accuracy of Bayesian model was highest for the milk protein in every scenario compared with best linear unbiased prediction (BLUP). These results showed the potentials of using haplotype-based Bayesian models in increasing the prediction accuracy. Hess et al. (2017) tested the potential benefits of fixed-length of haplotypes in improving the prediction accuracy in an admixed dairy cattle population. The main results showed that genomic predictions were more accurate with short haplotypes than those with longer haplotypes and no significant differences were observed between different Bayesian models (BayesA, BayesB and BayesN). Jiang et al. (2018) extended haplotype-based genomic prediction (HGBLUP) model to exploit local epistatic interactions among markers. Applying the HGBLUP model to empirical data sets using a mouse panel revealed higher prediction accuracies than for marker-based models. In contrast, only a small subset of the traits analyzed in crop populations showed such a benefit.

1.12 General conclusion and scientific plans

In summary, nowadays we have entered a new breeding era where new statistical models (such as Bayesian models and meta-analysis) and the availability of diverse multi-omic data (genomes, phenomes, metabolomes and transcriptomes) together makes it possible (1) to understand what has happened in the past breeding history, especially from the evolutionary perspective, (2) to promote our understanding on the genetic architecture of important traits and (3) to effectively design and select ‘the ideal crop’ for the targeted populations and meet the global increasing requirements of food and nutrition security.

In this thesis, I mainly focused on haplotype-based analyses and genome-wide association meta-analysis using three GWAS panels for the analysis of tomato fruit quality. For the haplotype-based analyses, I calculated the haplotype blocks within wild species, cherry and large tomato collections. I also used integrated haplotype score (iHS), a haplotype-based approach to detect selection sweeps in tomato genome. I then compared haplotype-based association model with multi-locus and single-locus mixed model in identifying association. For the significant associations, I then checked the haplotypes in the nearby regions of the focal SNP to check whether there are distinct haplotypes within different tomato subgroups. I also compared the bifurcation diagram of haplotypes around the peak SNPs to see if there are distinct differences between the reference alleles and alternative alleles. I tested marker local haplotype score (mLHS) to determine the candidate regions around the associations. Together with gene annotations and transcriptome data, I tested the potential benefits of haplotype

study in tomato, to assess what we can gain in dissecting the genetic control of tomato flavor-related traits. Additionally, we also tested the possibilities of using haplotype to improve the genomic prediction accuracy, on a few traits. However, this part is not the main focus of this thesis and I will only give an example of its applications.

In the second part, I introduced genotyping imputation-driven meta-analysis of genome-wide association studies of three available GWAS panels, which were all focused on tomato flavor-related traits. To do so, I first applied genotyping imputations for the two panels that were genotyped with SNP arrays to increase the genome-wide coverage. A detailed quality control was performed and the imputation quality also cross-checked by comparing the genotyped and imputed data. I then re-run the GWAS using the imputed genotypes following the same association model. Once the summary GWAS results were available, I performed both fixed-effect and random-effected meta-analysis in order to handle heterogeneity. Once significant associations were detected, I analyzed the most significant regions studying gene annotation, transcriptome analysis, genes under selections, etc. I also tested the possibilities of using summary GWAS data to assess the heritability. However, though many novel candidate genes are expected to be detected, due to the limited time and resources, functional validation, such as gene editing, was not performed in this thesis.

In summary, I wanted to apply the most recently published approaches (in terms of haplotypes and meta-analysis) in deepening our understandings on the genetic control of tomato quality traits. We expect the achievement of this thesis will promote the improvement and breeding of tomato, especially with a focused attention on breeding tomato cultivars with an overall enhanced flavor.

References

Adato, A., Mandel, T., Mintz-Oron, S., Venger, I., Levy, D., Yativ, M., Domínguez, E., Wang, Z., De Vos, R. C. H., Jetter, R., et al. (2009). Fruit-surface flavonoid accumulation in tomato is controlled by a *SLMYB12*-regulated transcriptional network. *PLoS Genet.* **5**:e1000777.

Akbari, A., Vitti, J. J., Iranmehr, A., Bakhtiari, M., Sabeti, P. C., Mirarab, S., and Bafna, V. (2018). Identifying the favored mutation in a positive selective sweep. *Nat. Methods* **15**:279–282.

Akey, J., Jin, L., and Xiong, M. (2001). Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur. J. Hum. Genet.* **9**:291–300.

Akula, N., Baranova, A., Seto, D., Solka, J., Nalls, M. A., Singleton, A., Ferrucci, L., Tanaka, T., Bandinelli, S., Cho, Y. S., et al. (2011). A Network-Based Approach to Prioritize Results from Genome-Wide Association Studies. *PLoS One* **6**:e24220.

Alachiotis, N., and Pavlidis, P. (2018). RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Commun. Biol.* **1**:79.

Albrecht, E., Escobar, M., and Chetelat, R. T. (2010). Genetic diversity and population structure in the tomato-like nightshades Solanum lycopersicoides and S. sitiens. *Ann. Bot.* **105**:535–554.

Allison, A. C. (1954). Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection. *Br. Med. J.* **1**:290.

Almeida, J., Quadrana, L., Asís, R., Setta, N., de Godoy, F., Bermúdez, L., Otaiza, S. N., Corrêa da Silva, J. V., Fernie, A. R., Carrari, F., et al. (2011). Genetic dissection of vitamin E biosynthesis in tomato. *J. Exp. Bot.* **62**:3781–3798.

Alpert, K. B., and Tanksley, S. D. (1996). High-resolution mapping and isolation of a yeast artificial chromosome contig containing *fw2.2*: a major fruit weight quantitative trait locus in tomato. *Proc. Natl. Acad. Sci. U. S. A.* **93**:15503–7.

Alseekh, S., and Fernie, A. R. (2018). Metabolomics 20 years on: what have we learned and what hurdles remain? *Plant J.* **94**:933–942.

Alseekh, S., Ofner, I., Pleban, T., Tripodi, P., Di Dato, F., Cammareri, M., Mohammad, A., Grandillo, S., Fernie, A. R., and Zamir, D. (2013). Resolution by recombination: Breaking up *Solanum pennellii* introgressions. *Trends Plant Sci.* **18**:536–538.

Alseekh, S., Tohge, T., Wendenberg, R., Scossa, F., Omranian, N., Li, J., Kleessen, S., Giavalisco, P., Pleban, T., Mueller-Roeber, B., et al. (2015). Identification and Mode of Inheritance of Quantitative Trait Loci for Secondary Metabolite Abundance in Tomato. *Plant Cell* **27**:485–512.

Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. *Science (80-.).* **322**:881–888.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature* **507**:455–461.

Arafa, R. A., Rakha, M. T., Soliman, N. E. K., Moussa, O. M., Kamel, S. M., and Shirasawa, K. (2017). Rapid identification of candidate genes for resistance to tomato late blight disease using next-generation sequencing technologies. *PLoS One* **12**:e0189951.

Archak, S., Karihaloo, J. L., and Jain, A. (2002). RAPD markers reveal narrowing genetic base of Indian tomato cultivars. *Curr. Sci.* **82**:1139–1143.

Aschard, H., Chen, J., Cornelis, M. C., Chibnik, L. B., Karlson, E. W., and Kraft, P. (2012). Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases. *Am. J. Hum. Genet.* **90**:962–72.

Ashby, D. (2006). Bayesian statistics in medicine: a 25 year review. *Stat. Med.* **25**:3589–3631.

Ashrafi, H., Kinkade, M. P., Merk, H. L., and Foolad, M. R. (2012). Identification of novel quantitative trait loci for increased lycopene content and other fruit quality traits in a tomato recombinant inbred line population. *Mol. Breed.* **30**:549–567.

Asimit, J. L., Hatzikotoulas, K., McCarthy, M., Morris, A. P., and Zeggini, E. (2016). Trans-ethnic study design approaches for fine-mapping. *Eur. J. Hum. Genet.* **24**:1330–1336.

Auerswald, H., Schwarz, D., Kornelson, C., Krumbein, A., and Brückner, B. (1999). Sensory analysis, sugar and acid content of tomato at different EC values of the nutrient solution. *Sci. Hortic. (Amsterdam).* **82**:227–242.

Aulchenko, Y. S., de Koning, D.-J., and Haley, C. (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**:577–85.

Ayers, K. L., and Cordell, H. J. (2010). SNP Selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet. Epidemiol.* **34**:879–891.

Azencott, C.-A., Grimm, D., Sugiyama, M., Kawahara, Y., and Borgwardt, K. M. (2013). Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics* **29**:i171–i179.

Badke, Y. M., Bates, R. O., Ernst, C. W., Schwab, C., Fix, J., Van Tassell, C. P., and Steibel, J. P. (2013). Methods of tagSNP selection and other variables affecting imputation accuracy in swine. *BMC Genet.* **14**:8.

Baldet, P., Stevens, R., Causse, M., Duffe, P., Buret, M., Rothan, C., Garchery, C., Duffé, P., Carchery, C., Baldet, P., et al. (2007). Candidate Genes and Quantitative Trait Loci Affecting Fruit Ascorbic Acid Content in Three Tomato Populations. *Plant Physiol.* **143**:1943–1953.

Baldwin, E. A., Nisperos-Carriedo, M. O., Baker, R., and Scott, J. W. (1991). Quantitative analysis of flavor parameters in six Florida tomato cultivars (*Lycopersicon esculentum* Mill). *J. Agric. Food Chem.* **39**:1135–1140.

Baldwin, E. A., Scott, J. W., Einstein, M. A., Malundo, T. M. M., Carr, B. T., Shewfelt, R. L., and Tandon, K. S. (1998). Relationship between sensory and instrumental analysis for tomato flavor. *J. Am. Soc. Hortic. Sci.* **123**:906–915.

Baldwin, E., Scott, J., Shewmaker, C., and Schuch, W. (2000). Flavor trivia and tomato aroma: biochemistry and possible mechanisms for control of important aroma components. *Hortscience* **35**:1013–1022.

Ballester, A.-R., Bovy, A. G., Viquez-Zamora, M., Tikunov, Y., Grandillo, S., de Vos, R., de Maagd, R. A., van Heusden, S., and Molthoff, J. (2016). Identification of Loci Affecting Accumulation of Secondary Metabolites in Tomato Fruit of a *Solanum lycopersicum* × *Solanum chmielewskii* Introgression Line Population. *Front. Plant Sci.* **7**:1428.

Bandillo, N., Raghavan, C., Muyco, P., Sevilla, M. A. L., Lobina, I. T., Dilla-Ermita, C., Tung, C.-W., McCouch, S., Thomson, M., Mauleon, R., et al. (2013). Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. *Rice* **6**:11.

Bauchet, G., and Causse, M. (2012). Genetic diversity in tomato (*Solanum lycopersicum*) and its wild relatives. In *Genetic Diversity in Plants*, p. InTech.

Bauchet, G., Grenier, S., Samson, N., Bonnet, J., Grivet, L., and Causse, M. (2017a). Use of modern tomato breeding germplasm for deciphering the genetic control of agronomical traits by Genome Wide Association study. *Theor. Appl. Genet.* **130**:875–889.

Bauchet, G., Grenier, S., Samson, N., Segura, V., Kende, A., Beekwilder, J., Cankar, K., Gallois, J.-L., Gricourt, J., Bonnet, J., et al. (2017b). Identification of major loci and genomic regions controlling acid and volatile content in tomato fruit: implications for flavor improvement. *New Phytol.* **215**:624–641.

Baxter, C. J., Liu, J. L., Fernie, A. R., and Sweetlove, L. J. (2007). Determination of metabolic fluxes in a non-steady-state system. *Phytochemistry* **68**:2313–2319.

Begum, F., Ghosh, D., Tseng, G. C., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Res.* **40**:3777–3784.

Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'Ang, L. Y., Huang, W., Liu, B., Shen, Y., Tam, P. K. H., et al. (2003). The international HapMap project. *Nature* **426**:789–796.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**:289–300.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**:491–507.

Bernacchi, D., Beck-Bunn, T., Eshed, Y., Lopez, J., Petiard, V., Uhlig, J., Zamir, D., and Tanksley, S. (1998). Advanced backcross QTL analysis in tomato. I. Identification of QTLs for traits of agronomic importance from *Lycopersicon hirsutum*. *Theor. Appl. Genet.* **97**:381–397.

Bertin, N., Gautier, H., and Roche, C. (2002). Number of cells in tomato fruit depending on fruit position and source-sink balance during plant development. *Plant Growth Regul.* **36**:105–112.

Bertin, N., Borel, C., Brunel, B., Cheniclet, C., and Causse, M. (2003). Do genetic make-up and growth manipulation affect tomato fruit size by cell number, or cell size and DNA endoreduplication? *Ann. Bot.* **92**:415–424.

Birchler, J. A., Yao, H., Chudalayandi, S., Vaiman, D., and Veitia, R. A. (2010). Heterosis. *Plant Cell* **22**:2105–2112.

Blanca, J., Cañizares, J., Cordero, L., Pascual, L., Diez, M. J., and Nuez, F. (2012). Variation Revealed by SNP Genotyping and Morphology Provides Insight into the Origin of the Tomato. *PLoS One* **7**:e48198.

Blanca, J., Montero-Pau, J., Sauvage, C., Bauchet, G., Illa, E., Díez, M. J., Francis, D., Causse, M., van der Knaap, E., and Cañizares, J. (2015). Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. *BMC Genomics* **16**:257.

Boison, S. A., Utsunomiya, A. T. H., Santos, D. J. A., Neves, H. H. R., Carvalheiro, R., Mészáros, G., Utsunomiya, Y. T., do Carmo, A. S., Verneque, R. S., Machado, M. A., et al. (2017). Accuracy of genomic predictions in Gyr (*Bos indicus*) dairy cattle. *J. Dairy Sci.* **100**:5479–5490.

Bolger, A., Scossa, F., Bolger, M. E., Lanz, C., Maumus, F., Tohge, T., Quesneville, H., Alseekh, S., Sørensen, I., Lichtenstein, G., et al. (2014). The genome of the stress-tolerant wild tomato species *Solanum pennellii*. *Nat. Genet.* **46**:1034–1038.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubbl. del R Ist. Super. di Sci. Econ. e Commer. di Firenze* **8**:3–62.

Bovy, A., de Vos, R., Kemper, M., Schijlen, E., Pertejo, M. A., Muir, S., Collins, G., Robinson, S., Verhoeyen, M., Hughes, S., et al. (2002). Engineering secondary metabolism in maize cells by ectopic expression of transcription factors. *Plant Cell* **14**:2509–2526.

Brachi, B., Morris, G. P., and Borevitz, J. O. (2011). *Genome-wide association studies in plants: The missing heritability is in the field*. BioMed Central.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**:2633–2635.

Bramley, P. M. (2000). Is lycopene beneficial to human health? *Phytochemistry* **54**:233–236.

Bramley, P. M. (2002). Regulation of carotenoid formation during tomato fruit ripening and development. *J. Exp. Bot.*

53:2107–2113.

Breiman, L. (2001). Random Forests. *Mach. Learn.* **45**:5–32.

Browning, B. L., and Browning, S. R. (2008). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**:210–223.

Browning, B. L., and Browning, S. R. (2016). Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* **98**:116–126.

Bruhn, C. M., Feldman, N., Garlitz, C., Harwood, J., Ivans, E., Marshall, M., Riley, A., Thurber, D., and Williamson, E. (1991). Consumer Perceptions of Quality: Apricots, Cantaloupes, Peaches, Pears, Strawberries, and Tomatoes. *J. Food Qual.* **14**:187–195.

Bucheli, P., Voirol, E., De La Torre, R., López, J., Rytz, A., Tanksley, S. D., and Pétiard, V. (1999). Definition of nonvolatile markers for flavor of tomato (*Lycopersicon esculentum* Mill.) as tools in selection and breeding. *J. Agric. Food Chem.* **47**:659–664.

Buchner, D. A., and Nadeau, J. H. (2015). Contrasting genetic architectures in different mouse reference populations used for studying complex traits. *Genome Res.* **25**:775–91.

Budiman, M. A., Chang, S.-B., Lee, S., Yang, T. J., Zhang, H.-B., de Jong, H., and Wing, R. A. (2004). Localization of *jointless-2* gene in the centromeric region of tomato chromosome 12 based on high resolution genetic and physical mapping. *Theor. Appl. Genet.* **108**:190–196.

Butler, L. (1952). The linkage map of the tomato. *J. Hered.* **43**:25–36.

Buzdugan, L., Kalisch, M., Navarro, A., Schunk, D., Fehr, E., and Bühlmann, P. (2016). Assessing statistical significance in multivariable genome wide association analysis. *Bioinformatics* **32**:1990–2000.

Cagas, C. C., Lee, O. N., Nemoto, K., and Sugiyama, N. (2008). Quantitative trait loci controlling flowering time and related traits in a *Solanum lycopersicum* × *S. pimpinellifolium* cross. *Sci. Hortic. (Amsterdam).* **116**:144–151.

Calus, M. P. L., Meuwissen, T. H. E., Roos, A. P. W. de, and Veerkamp, R. F. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics* **178**:553–561.

Calus, M. P. L., Veerkamp, R. F., and Mulder, H. A. (2011). Imputation of missing single nucleotide polymorphism genotypes using a multivariate mixed model framework. *J. Anim. Sci.* **89**:2042–2049.

Canady, M. A., Meglic, V., and Chetelat, R. T. (2005). A library of *Solanum lycopersicoides* introgression lines in cultivated tomato. *Genome* **48**:685–697.

Cao, K., Xu, H., Zhang, R., Xu, D., Yan, L., Sun, Y., Xia, L., Zhao, J., Zou, Z., and Bao, E. (2019). Renewable and sustainable strategies for improving the thermal environment of Chinese solar greenhouses. *Energy Build.* **202**:109414.

Cárdenas, P. D., Sonawane, P. D., Pollier, J., Vanden Bossche, R., Dewangan, V., Weithorn, E., Tal, L., Meir, S., Rogachev, I., Malitsky, S., et al. (2016). GAME9 regulates the biosynthesis of steroidal alkaloids and upstream isoprenoids in the plant mevalonate pathway. *Nat. Commun.* **7**:10654.

Carelli, B. P., Gerald, L. T. S., Grazziotin, F. G., and Echeverrigaray, S. (2006). Genetic diversity among Brazilian cultivars and landraces of tomato *Lycopersicon esculentum* Mill. revealed by RAPD markers. *Genet. Resour. Crop Evol.* **53**:395–400.

Carmel-Goren, L., Liu, Y. S., Lifschitz, E., and Zamir, D. (2003). The *SELF-PRUNING* gene family in tomato. *Plant Mol. Biol.* **52**:1215–1222.

Catchen, J. M., Boone, J. Q., Davey, J. W., Hohenlohe, P. A., Etter, P. D., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12**:499–510.

Causse, M., Saliba-Colombani, V., Lesschaeve, I., and Buret, M. (2001). Genetic analysis of organoleptic quality in fresh market tomato. 2. Mapping QTLs for sensory attributes. *Theor. Appl. Genet.* **102**:273–283.

Causse, M., Saliba-Colombani, V., Lecomte, L., Duffé, P., Rousselle, P., and Buret, M. (2002). QTL analysis of fruit quality in fresh market tomato: a few chromosome regions control the variation of sensory and instrumental traits. *J. Exp. Bot.* **53**:2089–2098.

Causse, M., Buret, M., Robini, K., and Verschave, P. (2003). Inheritance of Nutritional and Sensory Quality Traits in Fresh Market Tomato and Relation to Consumer Preferences. *J. Food Sci.* **68**:2342–2350.

Causse, M., Duffe, P., Gomez, M. C., Buret, M., Damidaux, R., Zamir, D., Gur, A., Chevalier, C., Lemaire-Chamley, M., and Rothan, C. (2004). A genetic map of candidate genes and QTLs involved in tomato fruit size and composition. *J. Exp. Bot.* **55**:1671–1685.

Causse, M., Chaïb, J., Lecomte, L., Buret, M., and Hospital, F. (2007). Both additivity and epistasis control the genetic variation for fruit quality traits in tomato. *Theor. Appl. Genet.* **115**:429–442.

Causse, M., Friguet, C., Coiret, C., LéPicier, M., Navez, B., Lee, M., Holthuysen, N., Sinesio, F., Moneta, E., and Grandillo, S. (2010). Consumer Preferences for Fresh Tomato at the European Scale: A Common Segmentation on Taste and Firmness. *J. Food Sci.* **75**:S531–S541.

Caye, K., Jumentier, B., Lepeule, J., and François, O. (2019). LFMM 2: Fast and accurate inference of gene-environment associations in genome-wide studies. *Mol. Biol. Evol.* **36**:852–860.

Chakrabarti, M., Zhang, N., Sauvage, C., Muños, S., Blanca, J., Cañizares, J., Diez, M. J., Schneider, R., Mazourek, M., McClead, J., et al. (2013). A cytochrome P450 regulates a domestication trait in cultivated tomato. *Proc. Natl. Acad. Sci. U. S. A.* **110**:17125–30.

Chapman, K., Ferreira, T., Morris, A., Asimit, J., and Zeggini, E. (2011). Defining the power limits of genome-wide association scan meta-analyses. *Genet. Epidemiol.* **35**:781–9.

Chen, F. Q., Foolad, M. R., Hyman, J., St. Clair, D. A., and Beelaman, R. B. (1999). Mapping of QTLs for lycopene and other fruit traits in a *Lycopersicon esculentum* × *L. pimpinellifolium* cross and comparison of QTLs across tomato species. *Mol. Breed.* **5**:283–299.

Chen, G., Hackett, R., Walker, D., Taylor, A., Lin, Z., and Grierson, D. (2004). Identification of a specific isoform of tomato lipoxygenase (TomloxC) involved in the generation of fatty acid-derived flavor compounds. *Plant Physiol.* **136**:2641–51.

Chen, X., Kuja-Halkola, R., Rahman, I., Arpegård, J., Viktorin, A., Karlsson, R., Hägg, S., Svensson, P., Pedersen, N. L., and Magnusson, P. K. E. (2015). Dominant Genetic Variation and Missing Heritability for Human Complex Traits: Insights from Twin versus Genome-wide Common SNP Models. *Am. J. Hum. Genet.* **97**:708–714.

Chen, H.-H., Petty, L. E., Bush, W., Naj, A. C., and Below, J. E. (2019). GWAS and Beyond: Using Omics Approaches to Interpret SNP Associations. *Curr. Genet. Med. Rep.* **7**:30–40.

Chetelat, R. T., DeVerna, J. W., and Bennett, A. B. (1995). Introgression into tomato (*Lycopersicon esculentum*) of the *L. chmielewskii* sucrose accumulator gene (*sucr*) controlling fruit sugar composition. *Theor. Appl. Genet.* **91**:327–333.

Cho, S., Kim, K., Kim, Y. J., Lee, J.-K., Cho, Y. S., Lee, J.-Y., Han, B.-G., Kim, H., Ott, J., and Park, T. (2010). Joint Identification of Multiple Genetic Variants via Elastic-Net Variable Selection in a Genome-Wide Association Analysis. *Ann. Hum. Genet.* **74**:416–428.

Chung, D., Yang, C., Li, C., Gelernter, J., and Zhao, H. (2014). GPA: A Statistical Approach to Prioritizing GWAS Results by Integrating Pleiotropy and Annotation. *PLoS Genet.* **10**:e1004787.

Cirillo, E., Parnell, L. D., and Evelo, C. T. (2017). A Review of Pathway-Based Analysis Tools That Visualize Genetic Variants. *Front. Genet.* **8**:174.

Cirulli, E. T., and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11**:415–425.

Clark, A. G. (2004). The role of haplotypes in candidate gene studies. *Genet. Epidemiol.* **27**:321–333.

Cleveland, M. A., and Hickey, J. M. (2013). Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation1. *J. Anim. Sci.* **91**:3583–3592.

Cochran, W. G. (1954). The Combination of Estimates from Different Experiments. *Biometrics* **10**:101.

Colliver, S., Bovy, A., Collins, G., Muir, S., Robinson, S., De Vos, C. H. R., and Verhoeyen, M. E. (2002). Improving the nutritional content of tomatoes through reprogramming their flavonoid biosynthetic pathway. *Phytochem. Rev.* **1**:113–123.

Comai, L., and Henikoff, S. (2006). TILLING: Practical single-nucleotide mutation discovery. *Plant J.* **45**:684–694.

Costanzo, M., Kuzmin, E., van Leeuwen, J., Mair, B., Moffat, J., Boone, C., and Andrews, B. (2019). Global Genetic Networks and the Genotype-to-Phenotype Relationship. *Cell* **177**:85–100.

Crain, J., Mondal, S., Rutkoski, J., Singh, R. P., and Poland, J. (2018). Combining High-Throughput Phenotyping and Genomic Information to Increase Prediction and Selection Accuracy in Wheat Breeding. *Plant Genome* **11**:0.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* **22**:961–975.

Cuevas, J., Crossa, J., Montesinos-López, O. A., Burgueño, J., Perez-Roudriguez, P., and De Los Campos, G. (2017). Bayesian genomic prediction with genotype x environment interaction kernel models. *G3 Genes|Genomes|Genetics* **7**:41–53.

Cuyabano, B. C., Su, G., and Lund, M. S. (2014). Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. *BMC Genomics* **15**.

Cuyabano, B. C. D., Su, G., Rosa, G. J. M., Lund, M. S., and Gianola, D. (2015a). Bootstrap study of genome-enabled prediction reliabilities using haplotype blocks across Nordic Red cattle breeds. *J. Dairy Sci.* **98**:7351–7363.

Cuyabano, B. C. D., Su, G., and Lund, M. S. (2015b). Selection of haplotype variables from a high-density marker map for genomic prediction. *Genet. Sel. Evol.* **47**:61.

Dadaev, T., Saunders, E. J., Newcombe, P. J., Anokian, E., Leongamornlert, D. A., Brook, M. N., Cieza-Borrella, C., Mijuskovic, M., Wakerell, S., Olama, A. A. Al, et al. (2018). Fine-mapping of prostate cancer susceptibility loci in a large meta-analysis identifies candidate causal variants. *Nat. Commun.* **9**:2256.

Dal Cin, V., Kevany, B., Fei, Z., and Klee, H. J. (2009). Identification of *Solanum habrochaites* loci that quantitatively influence tomato fruit ripening-associated ethylene emissions. *Theor. Appl. Genet.* **119**:1183–1192.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nat. Genet.* **29**:229–232.

Danecek, P., Huang, J., Min, J. L., Timpson, N. J., Trabetti, E., Richards, J. B., Durbin, R., Howie, B., Gambaro, G., Zheng, H.-F., et al. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**:8111.

Dani, Z. (2001). Improving plant breeding with exotic genetic libraries. *Nat. Rev. Genet.* **2**:3–9.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. *Nat. Genet.* **48**:1284–1287.

Davies, J. N., and Hobson, G. E. (1981). The constituents of tomato fruit — the influence of environment, nutrition, and genotype. *Crit. Rev. Food Sci. Nutr.* **15**:205–280.

Daware, A. V., Srivastava, R., Singh, A. K., Parida, S. K., and Tyagi, A. K. (2017). Regional Association Analysis of MetaQTLs Delineates Candidate Grain Size Genes in Rice. *Front. Plant Sci.* **8**:807.

de Bakker, P. I. W., Ferreira, M. A. R., Jia, X., Neale, B. M., Raychaudhuri, S., and Voight, B. F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**:122–128.

De Las Fuentes, L., Yang, W., Dávila-Román, V. G., and Gu, C. C. (2012). Pathway-based genome-wide association analysis of coronary heart disease identifies biologically important gene sets. *Eur. J. Hum. Genet.* **20**:1168–1173.

de Los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., Calus, M. P. L., Kirst, M., Huber, D., and Peter, G. F. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **193**:327–45.

de Vlaming, R., Okbay, A., Rietveld, C. A., Johannesson, M., Magnusson, P. K. E., Uitterlinden, A. G., van Rooij, F. J. A., Hofman, A., Groenen, P. J. F., Thurik, A. R., et al. (2017). Meta-GWAS Accuracy and Power (MetaGAP) Calculator Shows that Hiding Heritability Is Partially Due to Imperfect Genetic Correlations across Studies. *PLoS Genet.* **13**:1–23.

DeVicente, M. C., and Tanksley, S. D. (1993). QTL analysis of transgressive segregation in an interspecific tomato cross. *Genetics* **134**:585–596.

Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D. B. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**:e1000294.

Diouf, I. A., Derivot, L., Bitton, F., Pascual, L., and Causse, M. (2018). Water Deficit and Salinity Stress Reveal Many Specific QTL for Plant Growth and Fruit Quality Traits in Tomato. *Front. Plant Sci.* **9**:279.

Dixon, M. S., Jones, D. A., Keddie, J. S., Thomas, C. M., Harrison, K., and Jones, J. D. . (1996). The Tomato *Cf-2* Disease Resistance Locus Comprises Two Functional Genes Encoding Leucine-Rich Repeat Proteins. *Cell* **84**:451–459.

Doganlar, S., Frary, A., Ku, H.-M., and Tanksley, S. D. (2003). Mapping quantitative trait loci in inbred backcross lines of *Lycopersicon pimpinellifolium* (LA1589). *Genome* **45**:1189–1202.

Dorais, M., and Papadopoulos, A. P. (2001). Greenhouse Tomato Fruit Quality. In *Horticulture Reviews*, p. 349. Wiley & Sons.

Druet, T., and Georges, M. (2010). A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics* **184**:789–798.

Du, H., Vimaleswaran, K. S., Ängquist, L., Hansen, R. D., van der A, D. L., Holst, C., Tjønneland, A., Overvad, K., Jakobsen, M. U., Boeing, H., et al. (2011). Genetic Polymorphisms in the Hypothalamic Pathway in Relation to Subsequent Weight Change – The DiOGenes Study. *PLoS One* **6**:e17436.

Du, X., Huang, G., He, S., Yang, Z., Sun, G., Ma, X., Li, N., Zhang, X., Sun, J., Liu, M., et al. (2018). Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nat. Genet.* **50**:1–7.

Duangjit, J., Causse, M., and Sauvage, C. (2016). Efficiency of genomic selection for tomato fruit quality. *Mol. Breed.* **36**:36:29.

Edwards, S. M. K., Buntjer, J. B., Jackson, R., Bentley, A. R., Lage, J., Byrne, E., Burt, C., Jack, P., Berry, S., Flatman, E., et al. (2019). The effects of training population design on genomic prediction accuracy in wheat. *Theor. Appl. Genet.* Advance Access published 2019, doi:10.1007/s00122-019-03327-y.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**:446–450.

El-hady, E. a a A., Haiba, A. a a, El-hamid, N. R. A., Rizkalla, A. a, and Phylogenetic, A. a R. (2010). Phylogenetic Diversity and Relationships of Some Tomato Varieties by Electrophoretic Protein and RAPD analysis. *J. Am. Sci.* **6**:434–441.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome J.* **4**:250.

Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., McDonel, P. E., Guttman, M., and Lander, E. S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**:452–455.

Eshed, Y., and Zamir, D. (1995). An Introgression Line Population of *Lycopersicon pennellii* in the Cultivated Tomato Enables the Identification and Fine Mapping of Yield-Associated QTL. *Genetics* **141**:1147–1162.

Eu-ahsunthornwattana, J., Miller, E. N., Fakiola, M., Jeronimo, S. M. B., Blackwell, J. M., Cordell, H. J., and Cordell, H. J. (2014). Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data. *PLoS Genet.* **10**:e1004445.

Evangelou, E., and Ioannidis, J. P. A. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **14**:379–389.

Fan, Y., and Song, Y.-Q. (2016). Finding the Missing Heritability of Genome-wide Association Study Using Genotype Imputation. *Matters* Advance Access published 2016, doi:10.1111/j.1440-1789.2005.00628.x.

FAO (2015). *Coping with climate change – the roles of genetic resources for food and agriculture*.

Farashi, S., Kryza, T., Clements, J., and Batra, J. (2019). Post-GWAS in prostate cancer: from genetic association to biological contribution. *Nat. Rev. Cancer* **19**:46–59.

Fay, J. C., and Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**:1405–1413.

Fernandes, S. B., Dias, K. O. G., Ferreira, D. F., and Brown, P. J. (2018). Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. *Theor. Appl. Genet.* **131**:747–755.

Fernando, R. L., and Garrick, D. (2013). Bayesian Methods Applied to GWAS. In *Genome-Wide Association Studies and Genomic Prediction*, pp. 237–274. Berlin: Humana Press, Totowa, NJ.

Fernie, A. R., and Gutierrez-Marcos, J. (2019). From genome to phenome: genome-wide association studies and other approaches that bridge the genotype to phenotype gap. *Plant J.* **97**:5–7.

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M. I., et al. (2016). Detection of human adaptation during the past 2000 years. *Science (80-.).* **354**:760–764.

Finkers, R., Van Heusden, A. W., Meijer-Dekens, F., Van Kan, J. A. L., Maris, P., and Lindhout, P. (2007). The construction of a *Solanum habrochaites* LYC4 introgression line population and the identification of QTLs for resistance to *Botrytis cinerea*. *Theor. Appl. Genet.* **114**:1071–1080.

Foolad, M. R. (2007). Genome mapping and molecular breeding of tomato. *Int. J. Plant Genomics* **2007**:64358.

Foolad, M. R., and Panthee, D. R. (2012). Marker-Assisted Selection in Tomato Breeding. *CRC. Crit. Rev. Plant Sci.* **31**:93–123.

Frary, A., Nesbitt, T. C., Frary, A., Grandillo, S., Van Der Knaap, E., Cong, B., Liu, J., Meller, J., Elber, R., Alpert, K. B., et al. (2000). *fw2.2*: A quantitative trait locus key to the evolution of tomato fruit size. *Science (80-.).* **289**:85–88.

Frary, A., Doganlar, S., Daunay, M. C., and Tanksley, S. D. (2003). QTL analysis of morphological traits in eggplant and implications for conservation of gene function during evolution of solanaceous species. *Theor. Appl. Genet.* **107**:359–370.

Frary, A., Fulton, T. M., Zamir, D., and Tanksley, S. D. (2004). Advanced backcross QTL analysis of a *Lycopersicon esculentum* × *L. pennellii* cross and identification of possible orthologs in the Solanaceae. *Theor. Appl. Genet.* **108**:485–496.

Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**:241–251.

Fridman, E., and Zamir, D. (2003). Functional divergence of a syntenic invertase gene family in tomato, potato, and Arabidopsis. *Plant Physiol.* **131**:603–9.

Fridman, E., Pleban, T., and Zamir, D. (2000). A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. *Proc. Natl. Acad. Sci.* **97**:4718–4723.

Fridman, E., Liu, Y. S., Carmel_Goren, L., Gur, A., Shoresh, M., Pleban, T., Eshed, Y., and Zamir, D. (2002). Two tightly linked QTLs modify tomato sugar content via different physiological pathways. *Mol. Genet. Genomics* **266**:821–826.

Fridman, E., Carrari, F., Liu, Y. S., Fernie, A. R., and Zamir, D. (2004). Zooming in on a quantitative trait for tomato yield using interspecific introgressions. *Science (80-.).* **305**:1786–1789.

Fu, Y. X., and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**:693–709.

Fulop, D., Ranjan, A., Ofner, I., Covington, M. F., Chitwood, D. H., West, D., Ichihashi, Y., Headland, L., Zamir, D., Maloof, J. N., et al. (2016). A New Advanced Backcross Tomato Population Enables High Resolution Leaf QTL Mapping and Gene Identification. *G3 Genes|Genomes|Genetics* **6**:3169–3184.

Fulton, T. M. (2002). Identification, Analysis, and Utilization of Conserved Ortholog Set Markers for Comparative Genomics in Higher Plants. *Plant Cell Online* **14**:1457–1467.

Fulton, T. M., Beck-Bunn, T., Emmatty, D., Eshed, Y., Lopez, J., Petiard, V., Uhlig, J., Zamir, D., and Tanksley, S. D. (1997). QTL analysis of an advanced backcross of *Lycopersicon peruvianum* to the cultivated tomato and comparisons with QTLs found in other wild species. *Theor. Appl. Genet.* **95**:881–894.

Fulton, T. M., Grandillo, S., Beck-Bunn, T., Fridman, E., Frampton, A., Lopez, J., Petiard, V., Uhlig, J., Zamir, D., and Tanksley, S. D. (2000). Advanced backcross QTL analysis of a *Lycopersicon esculentum* × *Lycopersicon parviflorum* cross. *Theor. Appl. Genet.* **100**:1025–1042.

Furlotte, N. A., and Eskin, E. (2015). Efficient Multiple-Trait Association and Estimation of Genetic Correlation Using the Matrix-Variate Linear Mixed Model. *Genetics* **200**:59–68.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002). The structure of haplotype blocks in the human genome. *Science (80-.).* **296**:2225–2229.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V, Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**:1091–1098.

Gao, H., Zhang, T., Wu, Y., Wu, Y., Jiang, L., Zhan, J., Li, J., and Yang, R. (2014). Multiple-trait genome-wide association study based on principal component analysis for residual covariance matrix. *Heredity (Edinb).* **113**:526–532.

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., Burzynski-Chang, E. A., Fish, T. L., Stromberg, K. A., Sacks, G. L., et al. (2019a). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat.*

Genet. **51**:1044–1051.

Gao, J., Wang, S., Zhou, Z., Wang, S., Dong, C., Mu, C., Song, Y., Ma, P., Li, C., Wang, Z., et al. (2019b). Linkage mapping and genome-wide association reveal candidate genes conferring thermotolerance of seed-set in maize. *J. Exp. Bot.* **70**:4849–4864.

Garcia, V., Bres, C., Just, D., Fernandez, L., Tai, F. W. J., Mauxion, J. P., Le Paslier, M. C., Bérard, A., Brunel, D., Aoki, K., et al. (2016). Rapid identification of causal mutations in tomato EMS populations via mapping-by-sequencing. *Nat. Protoc.* **11**:2401–2418.

Gardiner, L.-J., Quinton-Tulloch, M., Olohan, L., Price, J., Hall, N., and Hall, A. (2015). A genome-wide survey of DNA methylation in hexaploid wheat. *Genome Biol.* **16**:273.

Gauffier, C., Lebaron, C., Moretti, A., Constant, C., Moquet, F., Bonnet, G., Caranta, C., and Gallois, J.-L. (2016). A TILLING approach to generate broad-spectrum resistance to potyviruses in tomato is hampered by *eIF4E* gene redundancy. *Plant J.* **85**:717–729.

Gaulton, K. J., Ferreira, T., Lee, Y., Raimondo, A., Mägi, R., Reschen, M. E., Mahajan, A., Locke, A., William Rayner, N., Robertson, N., et al. (2015). Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* **47**:1415–1425.

Gest, N., Gautier, H., and Stevens, R. (2013). Ascorbate as seen through plant evolution: the rise of a successful molecule? *J. Exp. Bot.* **64**:33–53.

Ghoussaini, M., Edwards, S. L., Michailidou, K., Nord, S., Cowper-Sal·lari, R., Desai, K., Kar, S., Hillman, K. M., Kaufmann, S., Glubb, D. M., et al. (2014). Evidence that breast cancer risk at the 2q35 locus is mediated through *IGFBP5* regulation. *Nat. Commun.* **5**:4999.

Gianola, D., and Kaam, J. B. C. H. M. van (2008). Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits. *Genetics* **178**:2289.

Gibson, G. (2012). Rare and common variants: Twenty arguments. *Nat. Rev. Genet.* **13**:135–145.

Giovannucci, E. (1999). Tomatoes, Tomato-Based Products, Lycopene, and Cancer: Review of the Epidemiologic Literature. *J. Natl. Cancer Inst.* **91**:317–331.

Glass, G. V. (2015). Meta-analysis at middle age: a personal history. *Res. Synth. Methods* **6**:221–231.

Goff, S. A., and Klee, H. J. (2006). Plant volatile compounds: Sensory cues for health and nutritional value? *Science (80-.).* **311**:815–819.

Gong, J., Liu, C., Liu, W., Wu, Y., Ma, Z., Chen, H., and Guo, A.-Y. (2015). An update of miRNASNP database for better SNP selection by GWAS data, miRNA expression and online tools. *Database* **2015**.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**:333–351.

Grandillo, S., and Tanksley, S. D. (1996a). QTL analysis of horticultural traits differentiating the cultivated tomato from the closely related species *Lycopersicon pimpinellifolium*. *Theor. Appl. Genet.* **92**:935–951.

Grandillo, S., and Tanksley, S. D. (1996b). Genetic analysis of RFLPs, GATA microsatellites and RAPDs in a cross between *L. esculentum* and *L. pimpinellifolium*. *Theor. Appl. Genet.* **92**:957–965.

Grandillo, S., Ku, H. M., and Tanksley, S. D. (1996). Characterization of *fs8.1*, a major QTL influencing fruit shape in tomato. *Mol. Breed.* **2**:251–260.

Grandillo, S., Ku, H. M., and Tanksley, S. D. (1999). Identifying the loci responsible for natural variation in fruit size and shape in tomato. *Theor. Appl. Genet.* **99**:978–987.

Grandillo, S., Chetelat, R., Knapp, S., Spooner, D., Peralta, I., Cammareri, M., Perez, O., Termolino, P., Tripodi, P., Chiusano, M. L., et al. (2011). *Solanum* sect. *Lycopersicon*. In *Wild Crop Relatives: Genomic and Breeding Resources*, pp. 129–215. Berlin, Heidelberg: Springer Berlin Heidelberg.

Grimm, D. G., Roqueiro, D., Salomé, P. A., Kleeberger, S., Greshake, B., Zhu, W., Liu, C., Lippert, C., Stegle, O., Schölkopf, B., et al. (2017). easyGWAS: A Cloud-Based Platform for Comparing the Results of Genome-Wide Association Studies. *Plant Cell* **29**:5–19.

Grossman, S. R., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., et al. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science (80-.).* **327**:883–886.

Gualdrón Duarte, J. L., Bates, R. O., Ernst, C. W., Raney, N. E., Cantet, R. J., and Steibel, J. P. (2013). Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. *BMC Genet.* **14**:38.

Guan, Y., and Stephens, M. (2008). Practical issues in imputation-based association mapping. *PLoS Genet.* **4**.

Gupta, P. K., Kumar, J., Mir, R. R., and Kumar, A. (2010). Marker-Assisted Selection as a Component of Conventional Plant Breeding. In *Plant Breeding Reviews*, pp. 145–217. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Gupta, P. K., Kulwal, P. L., and Jaiswal, V. (2014). Association mapping in crop plants: Opportunities and challenges. In *Advances in Genetics*, pp. 109–147. Elsevier.

Gupta, P. K., Kulwal, P. L., and Jaiswal, V. (2019). Association mapping in plants in the post-GWAS genomics era. In *Advances in Genetics*, pp. 75–154.

Gur, A., Osorio, S., Fridman, E., Fernie, A. R., and Zamir, D. (2010). *hi2-1*, A QTL which improves harvest index, earliness and alters metabolite accumulation of processing tomatoes. *Theor. Appl. Genet.* **121**:1587–1599.

Gur, A., Semel, Y., Osorio, S., Friedmann, M., Seekh, S., Ghareeb, B., Mohammad, A., Pleban, T., Gera, G., Fernie, A. R., et al. (2011). Yield quantitative trait loci from wild tomato are predominately expressed by the shoot. *Theor. Appl. Genet.* **122**:405–420.

Gurevitch, J., Koricheva, J., Nakagawa, S., and Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature* **555**:175–182.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., Jansen, R., De Geus, E. J. C., Boomsma, D. I., Wright, F. A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**:245–252.

Haanstra, J. P. W., Wye, C., Verbakel, H., Meijer-Dekens, F., Van Den Berg, P., Odinot, P., Van Heusden, A. W., Tanksley, S., Lindhout, P., and Peleman, J. (1999). An integrated high-density RFLP-AFLP map of tomato based on two *Lycopersicon esculentum* x *L. pennellii* F2 populations. *Theor. Appl. Genet.* **99**:254–271.

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* **12**:186.

Hafler, D. A., and Jager, P. L. De (2005). Applying a new generation of genetic maps to understand human inflammatory disease. *Nat. Rev. Immunol.* **5**:83–91.

Hägg, S., Ganna, A., Laan, S. W. Van Der, Esko, T., Pers, T. H., Locke, A. E., Berndt, S. I., Justice, A. E., Kahali, B., Siemelink, M. A., et al. (2015). Gene-based meta-analysis of genome-wide association studies implicates new loci involved in obesity. *Hum. Mol. Genet.* **24**:6849.

Haggard, J. E., Johnson, E. B., and St. Clair, D. A. (2013). Linkage Relationships Among Multiple QTL for Horticultural Traits and Late Blight (*P. infestans*) Resistance on Chromosome 5 Introgressed from Wild Tomato *Solanum habrochaites*. *G3 Genes|Genomes|Genetics* **3**:2131–2146.

Haldane, J. B. S. (2006). Disease and Evolution. In *Malaria: Genetic and Evolutionary Aspects*, pp. 175–187. Boston, MA: Springer US.

Halperin, E., and Stephan, D. A. (2009). SNP imputation in association studies. *Nat. Biotechnol.* **27**:349–351.

Hamblin, M. T., and Jannink, J.-L. (2011). Factors Affecting the Power of Haplotype Markers in Association Studies. *Plant Genome J.* **4**:145.

Hamilton, J. P., Sim, S.-C., Stoffel, K., Van Deynze, A., Buell, C. R., and Francis, D. M. (2012). Single nucleotide polymorphism discovery in cultivated tomato via sequencing by synthesis. *Plant Genome J.* **5**:17.

Hanson, P. M., Yang, R., Wu, J., Chen, J., Ledesma, D., Tsou, S. C. S., and Lee, T.-C. (2004). Variation for Antioxidant Activity and Antioxidants in Tomato. *J. Am. Soc. Hortic. Sci.* **129**:704–711.

Hao, D., Cheng, H., Yin, Z., Cui, S., Zhang, D., Wang, H., and Yu, D. (2012). Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (*Glycine max*) landraces across multiple environments. *Theor. Appl. Genet.* **124**:447–458.

Harborne, J. B., and Williams, C. A. (2000). Advances in flavonoid research since 1992. *Phytochemistry* **55**:481–504.

Haseneyer, G., Schmutzer, T., Seidel, M., Zhou, R., Mascher, M., Schön, C. C., Taudien, S., Scholz, U., Stein, N., Mayer, K. F. X., et al. (2011). From RNA-seq to large-scale genotyping - genomics resources for rye (*Secale cereale* L.). *BMC Plant Biol.* **11**:131.

Hayashi, T., and Iwata, H. (2010). EM algorithm for Bayesian estimation of genomic breeding values. *BMC Genet.* **11**:3.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). *Invited review*: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.* **92**:433–443.

Hayes, B. J., Bowman, P. J., Daetwyler, H. D., Kijas, J. W., and van der Werf, J. H. J. (2012). Accuracy of genotype imputation in sheep breeds. *Anim. Genet.* **43**:72–80.

He, S., Schulthess, A. W., Mirdita, V., Zhao, Y., Korzun, V., Bothe, R., Ebmeyer, E., Reif, J. C., and Jiang, Y. (2016). Genomic selection in a commercial winter wheat population. *Theor. Appl. Genet.* **129**:641–651.

Heffner, E. L., Sorrells, M. E., and Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Sci.* **49**:1–12.

Heslot, N., Yang, H. P., Sorrells, M. E., and Jannink, J. L. (2012). Genomic selection in plant breeding: A comparison of models. *Crop Sci.* **52**:146–160.

Hess, M., Druet, T., Hess, A., and Garrick, D. (2017). Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. *Genet. Sel. Evol.* **49**:54.

Hickey, J. M., and Kranis, A. (2013). Extending long-range phasing and haplotype library imputation methods to impute genotypes on sex chromosomes. *Genet. Sel. Evol.* **45**:10.

Hickey, J. M., Kinghorn, B. P., Tier, B., Wilson, J. F., Dunstan, N., and van der Werf, J. H. (2011). A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet. Sel. Evol.* **43**:12.

Hickey, J. M., Crossa, J., Babu, R., and de los Campos, G. (2012a). Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. *Crop Sci.* **52**:654.

Hickey, J. M., Kinghorn, B. P., Tier, B., van der Werf, J. H., and Cleveland, M. A. (2012b). A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. *Genet. Sel. Evol.* **44**:9.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106**:9362–7.

Hirschhorn, J. N., and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**:95–108.

Ho, L. C. (1996). The mechanism of assimilate partitioning and carbohydrate compartmentation in fruit in relation to the quality and yield of tomato. *J. Exp. Bot.* **47**:1239–1243.

Hobson, G. E., and Bedford, L. (1989). The composition of cherry tomatoes and its relation to consumer acceptability. *J. Hortic. Sci.* **64**:321–329.

Hobson, G., and Grierson, D. (1993). Tomato. In *Biochemistry of Fruit Ripening*, pp. 405–442. Dordrecht: Springer Netherlands.

Hoffman, G. E. (2013). Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. *PLoS One* **8**:e75707.

Hoggart, C. J., Whittaker, J. C., De Iorio, M., and Balding, D. J. (2008). Simultaneous Analysis of All SNPs in Genome-Wide and Re-Sequencing Association Studies. *PLoS Genet.* **4**:e1000130.

Hoggart, C. J., Venturini, G., Mangino, M., Gomez, F., Ascari, G., Zhao, J. H., Teumer, A., Winkler, T. W., Tšernikova, N., Luan, J., et al. (2014). Novel approach identifies SNPs in *SLC2A10* and *KCNK9* with evidence for parent-of-origin effect on body mass index. *PLoS Genet.* **10**:e1004508.

Holm, S. (1979). A simple rejective test procedure. *Scand. J. Stat.* **6**:65–70.

Huang, Z., and van der Knaap, E. (2011). Tomato fruit weight *11.3* maps close to fasciated on the bottom of chromosome 11. *Theor. Appl. Genet.* **123**:465–474.

Huang, Y., Hickey, J. M., Cleveland, M. A., and Maltecca, C. (2012a). Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. *Genet. Sel. Evol.* **44**:25.

Huang, B. E., George, A. W., Forrest, K. L., Kilian, A., Hayden, M. J., Morell, M. K., and Cavanagh, C. R. (2012b). A multiparent advanced generation inter-cross population for genetic analysis in wheat. *Plant Biotechnol. J.* **10**:826–839.

Huang, H., Fang, M., Jostins, L., Umićević Mirkov, M., Boucher, G., Anderson, C. A., Andersen, V., Cleynen, I., Cortes, A., Crins, F., et al. (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**:173–178.

Hudson, R. R., Kreitman, M., and Aguadé, M. (1987). A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics* **116**:153–159.

Hughes, A. L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**:167–170.

Hysi, P. G., Valdes, A. M., Liu, F., Furlotte, N. A., Evans, D. M., Bataille, V., Visconti, A., Hemani, G., McMahon, G., Ring, S. M., et al. (2018). Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability. *Nat. Genet.* Advance Access published April 16, 2018, doi:10.1038/s41588-018-0100-5.

International Schizophrenia Consortium, S., Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**:748–52.

Ioannidis, J. P. A., Trikalinos, T. A., and Khoury, M. J. (2006). Implications of Small Effect Sizes of Individual Genetic Variants on the Design and Interpretation of Genetic Association Studies of Complex Diseases. *Am. J. Epidemiol.* **164**:609–614.

Ioannidis, J. P. A., Patsopoulos, N. A., and Evangelou, E. (2007). Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS One* **2**.

Isidro, J., Jannink, J.-L., Akdemir, D., Poland, J., Heslot, N., and Sorrells, M. E. (2015). Training set optimization under population structure in genomic selection. *Theor. Appl. Genet.* **128**:145–158.

Islam, M. S., Thyssen, G. N., Jenkins, J. N., Zeng, L., Delhom, C. D., McCarty, J. C., Deng, D. D., Hinchliffe, D. J., Jones, D. C., and Fang, D. D. (2016). A MAGIC population-based genome-wide association study reveals functional association of *GhRBB1_A07* gene with superior fiber quality in cotton. *BMC Genomics* **17**:903.

Ito, Y., Nishizawa-Yokoi, A., Endo, M., Mikami, M., Shima, Y., Nakamura, N., Kotake-Nara, E., Kawasaki, S., and Toki, S. (2017). Re-evaluation of the rin mutation and the role of RIN in the induction of tomato ripening. *Nat. Plants* **3**:866–874.

Iwata, H., and Jannink, J. L. (2010). Marker genotype imputation in a low-marker-density panel with a high-marker-density reference panel: Accuracy evaluation in barley breeding lines. *Crop Sci.* **50**:1269–1278.

Jatoi, S. A., Fujimura, T., Yamanaka, S., Watanabe, J., Watanabe, K. N., and Watanabe, K. N. (2008). Potential loss of unique genetic diversity in tomato landraces by genetic colonization of modern cultivars at a non-center of origin. *Plant Breed.* **127**:189–196.

Jiang, Y., Schmidt, R. H., and Reif, J. C. (2018). Haplotype-based genome-wide prediction models exploit local epistatic interactions among markers. *G3 Genes|Genomes|Genetics* **8**:g3.300548.2017.

Jiménez-Gómez, J. M., Alonso-Blanco, C., Borja, A., Anastasio, G., Angosto, T., Lozano, R., and Martínez-Zapater, J. M. (2007). Quantitative genetic analysis of flowering time in tomato. *Genome* **50**:303–315.

Johansson, Å., Marroni, F., Hayward, C., Franklin, C. S., Kirichenko, A. V., Jonasson, I., Hicks, A. A., Vitart, V., Isaacs, A., Axenovich, T., et al. (2010). Linkage and Genome-wide Association Analysis of Obesity-related Phenotypes: Association of Weight With the MGAT1 Gene. *Obesity* **18**:803–808.

Johnson, G. C. L., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., et al. (2001). Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29**:233–237.

Jonas, E., and de Koning, D.-J. (2013). Does genomic selection have a future in plant breeding? *Trends Biotechnol.* **31**:497–504.

Jones, J. B. (1986). Survival of *Xanthomonas campestris* pv. *vesicatoria* in Florida on Tomato Crop Residue, Weeds, Seeds, and Volunteer Tomato Plants. *Phytopathology* **76**:430.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., et al. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**:55–61.

Joo, J., Kwak, M., Ahn, K., and Zheng, G. (2009). A Robust Genome-Wide Scan Statistic of the Wellcome Trust Case-Control Consortium. *Biometrics* **65**:1115–1122.

Kamal, H. M., Takashina, T., Egashira, H., Satoh, H., and Imanishi, S. (2001). Introduction of aromatic fragrance into cultivated tomato from the "peruvianum complex." *Plant Breed.* **120**:179–181.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckermann, D., Daly, M. J., and Eskin, E. (2008). Statistical framework for phylogenomic analysis of gene family expression profiles. *Genetics* **178**:1709–1723.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**:348–354.

Karimi, Z., Sargolzaei, M., Robinson, J. A. B., and Schenkel, F. S. (2018). Assessing haplotype-based models for genomic evaluation in holstein cattle. *Can. J. Anim. Sci.* **98**:750–759.

Karlsson, E. K., Kwiatkowski, D. P., and Sabeti, P. C. (2014). Natural selection and infectious disease in human

populations. *Nat. Rev. Genet.* **15**:379–393.

Kawchuk, L. M., Hachey, J., Lynch, D. R., Kulcsar, F., Van Rooijen, G., Waterer, D. R., Robertson, A., Kokko, E., Byers, R., Howard, R. J., et al. (2001). Tomato Ve disease resistance genes encode cell surface-like receptors. *Proc. Natl. Acad. Sci. U. S. A.* **98**:6511–6515.

Kazmi, R. H., Khan, N., Willems, L. A. J., Van Heusden, A. W., Ligterink, W., and Hilhorst, H. W. M. (2012). Complex genetics controls natural variation among seed quality phenotypes in a recombinant inbred population of an interspecific cross between *Solanum lycopersicum* × *Solanum pimpinellifolium*. *Plant, Cell Environ.* **35**:929–951.

Khatkar, M. S., Zenger, K. R., Hobbs, M., Hawken, R. J., Cavanagh, J. A. L., Barris, W., McClintock, A. E., McClintock, S., Thomson, P. C., Tier, B., et al. (2007). A Primary Assembly of a Bovine Haplotype Block Map Based on a 15,036-Single-Nucleotide Polymorphism Panel Genotyped in Holstein–Friesian Cattle. *Genetics* **176**:763–772.

Kim, J., Zhang, Y., Pan, W., and Alzheimer's Disease Neuroimaging Initiative (2016). Powerful and Adaptive Testing for Multi-trait and Multi-SNP Associations with GWAS and Sequencing Data. *Genetics* **203**:715–31.

Kimbara, J., Ohyama, A., Chikano, H., Ito, H., Hosoi, K., Negoro, S., Miyatake, K., Yamaguchi, H., Nunome, T., Fukuoka, H., et al. (2018). QTL mapping of fruit nutritional and flavor components in tomato (*Solanum lycopersicum*) using genome-wide SSR markers and recombinant inbred lines (RILs) from an intra-specific cross. *Euphytica* **214**:210.

Kinkade, M. P., and Foolad, M. R. (2013). Validation and fine mapping of *lyc12.1*, a QTL for increased tomato fruit lycopene content. *Theor. Appl. Genet.* **126**:2163–2175.

Klee, H. J. (2010). Improving the flavor of fresh fruits: genomics, biochemistry, and biotechnology. *New Phytol.* **187**:44–56.

Klee, H. J. (2013). Purple tomatoes: Longer lasting, less disease, and better for you. *Curr. Biol.* **23**:R520–R521.

Klee, H. J., and Tieman, D. M. (2013). Genetic challenges of flavor improvement in tomato. *Trends Genet.* **29**:257–262.

Klee, H. J., and Tieman, D. M. (2018). The genetics of fruit flavour preferences. *Nat. Rev. Genet.* **19**:347–356.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., Paul SanGiovanni, J., Mane, S. M., Mayne, S. T., et al. (2005). Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science (80-.).* **308**:385–389.

Kooke, R., Kruijer, W., Bours, R., Becker, F., Kuhn, A., van de Geest, H., Buntjer, J., Doeswijk, T., Guerra, J., Bouwmeester, H., et al. (2016). Genome-Wide Association Mapping and Genomic Prediction Elucidate the Genetic Architecture of Morphological Traits in Arabidopsis. *Plant Physiol.* **170**:2187–2203.

Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q., and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* **44**:1066–1071.

Kover, P. X., and Mott, R. (2012). Mapping the genetic basis of ecologically and evolutionarily relevant traits in *Arabidopsis thaliana*. *Curr. Opin. Plant Biol.* **15**:212–217.

Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., Purugganan, M. D., Durrant, C., and Mott, R. (2009). A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in *Arabidopsis thaliana*. *PLoS Genet.* **5**:e1000551.

Kremling, K., Diepenbrock, C., Gore, M., Buckler, E., and Bandillo, N. (2018). Transcriptome-wide association supplements genome-wide association in Zea mays. *bioRxiv* Advance Access published July 6, 2018, doi:10.1101/363242.

Krieger, U., Lippman, Z. B., and Zamir, D. (2010). The flowering gene *SINGLE FLOWER TRUSS* drives heterosis for yield in tomato. *Nat. Genet.* **42**:459–463.

Kryukov, G. V., Pennacchio, L. A., and Sunyaev, S. R. (2007). Most Rare Missense Alleles Are Deleterious in Humans: Implications for Complex Disease and Association Studies. *Am. J. Hum. Genet.* **80**:727.

Kuan, P.-F., Yang, X., Clouston, S., Ren, X., Kotov, R., Waszczuk, M., Singh, P. K., Glenn, S. T., Gomez, E. C., Wang, J., et al. (2019). Cell type-specific gene expression patterns associated with posttraumatic stress disorder in World Trade Center responders. *Transl. Psychiatry* **9**:1.

Labate, J. A., Grandillo, S., Fulton, T., Muños, S., Caicedo, A. L., Peralta, I., Ji, Y., Chetelat, R. T., Scott, J. W., Gonzalo, M. J., et al. (2007). Tomato. In *Vegetables* (ed. Kole, C.), pp. 1–125. Berlin, Heidelberg: Springer Berlin Heidelberg.

Laterrot, H. (1996). Twenty-one near isogenic lines in Moneymaker type with different genes for disease resistances. *Rep Tomato Genet Coop* **46**:34.

Lawson, H. A., Cheverud, J. M., and Wolf, J. B. (2013). Genomic imprinting and parent-of-origin effects on complex traits. *Nat. Rev. Genet.* **14**:609–617.

Lecomte, L., Duffé, P., Buret, M., Servin, B., Hospital, F., and Causse, M. (2004). Marker-assisted introgression of five QTLs controlling fruit quality traits into three tomato lines revealed interactions between QTLs and genetic backgrounds. *Theor. Appl. Genet.* **109**:658–668.

Lehner, B. (2011). Molecular mechanisms of epistasis within and between genes. *Trends Genet.* **27**:323–331.

Length, F. (2011). Genetic diversity in 14 tomato (*Lycopersicon esculentum* Mill .) varieties in Nigerian markets by RAPD-PCR technique. *J. Biotechnol.* **10**:4961–4967.

Levin, I., Gilboa, N., Yeselson, E., Shen, S., and Schaffer, A. A. (2000). *Fgr*, a major locus that modulates the fructose to glucose ratio in mature tomato fruits. *Theor. Appl. Genet.* **100**:256–262.

Li, Y. R., and Keating, B. J. (2014). Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Med.* **6**:91.

Li, M., Liu, X., Bradbury, P., Yu, J., Zhang, Y.-M., Todhunter, R. J., Buckler, E. S., and Zhang, Z. (2014). Enrichment of statistical power for genome-wide association studies. *BMC Biol.* **12**:73.

Lin, D. Y., and Zeng, D. (2010). Meta-analysis of genome-wide association studies: No efficiency gain in using individual participant data. *Genet. Epidemiol.* **34**:60–66.

Lin, P., Hartz, S. M., Zhang, Z., Saccone, S. F., Wang, J., Tischfield, J. A., Edenberg, H. J., Kramer, J. R., M.Goate, A., Bierut, L. J., et al. (2010). A New Statistic to Evaluate Imputation Reliability. *PLoS One* **5**:e9697.

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S., Wang, X., et al. (2014). Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.* **46**:1220–1226.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**:476–482.

Lipka, A. E., Gore, M. A., Magallanes-Lundback, M., Mesberg, A., Lin, H., Tiede, T., Chen, C., Buell, C. R., Buckler, E. S., Rocheford, T., et al. (2013). Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain. *G3 Genes, Genomes, Genet.* **3**:1287–99.

Lipka, A. E., Kandianis, C. B., Hudson, M. E., Yu, J., Drnevich, J., Bradbury, P. J., and Gore, M. A. (2015). From association to prediction: Statistical methods for the dissection and selection of complex traits in plants. *Curr. Opin. Plant Biol.* **24**:110–118.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**:833–835.

Lippert, C., Casale, F. P., Rakitsch, B., and Stegle, O. (2014). LIMIX: genetic analysis of multiple traits. *bioRxiv* Advance Access published May 22, 2014, doi:10.1101/003905.

Lippman, Z. B., and Zamir, D. (2007). Heterosis: revisiting the magic. *Trends Genet.* **23**:60–66.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008). Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* **133**:523–536.

Liu, H. J., and Yan, J. (2019). Crop genome-wide association study: a harvest of biological relevance. *Plant J.* **97**:8–18.

Liu, J., Van Eck, J., Cong, B., and Tanksley, S. D. (2002). A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proc. Natl. Acad. Sci.* **99**:13302–13306.

Liu, Y. J., Guo, Y. F., Zhang, L. S., Pei, Y. F., Yu, N., Yu, P., Papasian, C. J., and Deng, H. W. (2010). Biological pathway-based genome-wide association analysis identified the vasoactive intestinal peptide (VIP) pathway important for obesity. *Obesity* **18**:2339–2346.

Liu, E. Y., Li, M., Wang, W., and Li, Y. (2013). MaCH-Admix: Genotype Imputation for Admixed Populations. *Genet. Epidemiol.* **37**:25–37.

Liu, Z., Alseekh, S., Brotman, Y., Zheng, Y., Fei, Z., Tieman, D. M., Giovannoni, J. J., Fernie, A. R., and Klee, H. J. (2016a). Identification of a *Solanum pennellii* Chromosome 4 Fruit Flavor and Nutritional Quality-Associated Metabolite QTL. *Front. Plant Sci.* **7**:1–15.

Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016b). Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLOS Genet.* **12**:e1005767.

Liu, J., Yang, C., Shi, X., Li, C., Huang, J., Zhao, H., and Ma, S. (2016c). Analyzing Association Mapping in Pedigree-Based GWAS Using a Penalized Multitrait Mixed Model. *Genet. Epidemiol.* **40**:382–393.

Liu, H. Y., Alyass, A., Abadi, A., Peralta-Romero, J., Suarez, F., Gomez-Zamudio, J., Audirac, A., Parra, E. J., Cruz, M., and Meyre, D. (2019). Fine-mapping of 98 obesity loci in Mexican children. *Int. J. Obes.* **43**:23–32.

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**:197–206.

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**:284–290.

Londin, E., Yadav, P., Surrey, S., Kricka, L. J., and Fortina, P. (2013). Use of Linkage Analysis, Genome-Wide Association Studies, and Next-Generation Sequencing in the Identification of Disease-Causing Mutations. In *Methods in molecular biology (Clifton, N.J.)*, pp. 127–146.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**:580–585.

Lopes, M. S., Bastiaansen, J. W. M., Harlizius, B., Knol, E. F., and Bovenhuis, H. (2014). A Genome-Wide Association Study Reveals Dominance Effects on Number of Teats in Pigs. *PLoS One* **9**:e105867.

Lorenz, A. J., Hamblin, M. T., and Jannink, J.-L. (2010). Performance of Single Nucleotide Polymorphisms versus Haplotypes for Genome-Wide Association Analysis in Barley. *PLoS One* **5**:e14079.

Lu, Y., Xu, J., Yuan, Z., Hao, Z., Xie, C., Li, X., Shah, T., Lan, H., Zhang, S., Rong, T., et al. (2012). Comparative LD mapping using single SNPs and haplotypes identifies QTL for plant height and biomass as secondary traits of drought tolerance in maize. *Mol. Breed.* **30**:407–418.

Lü, H.-Y., Liu, X.-F., Wei, S.-P., and Zhang, Y.-M. (2011). Epistatic Association Mapping in Homozygous Crop Cultivars. *PLoS One* **6**:e17773.

Luo, J. (2015). Metabolite-based genome-wide association studies in plants. *Curr. Opin. Plant Biol.* **24**:31–38.

Ma, P., Brøndum, R. F., Zhang, Q., Lund, M. S., and Su, G. (2013). Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. *J. Dairy Sci.* **96**:4666–4677.

Macgregor, S., Cornes, B. K., Martin, N. G., and Visscher, P. M. (2006). Bias, precision and heritability of self-reported and clinically measured height in Australian twins. *Hum. Genet.* **120**:571–580.

Mackay, T. F. C. (2014). Epistasis and quantitative traits: Using model organisms to study gene-gene interactions. *Nat. Rev. Genet.* **15**:22–33.

Mackay, I. J., Bansept-Basler, P., Barber, T., Bentley, A. R., Cockram, J., Gosman, N., Greenland, A. J., Horsnell, R., Howells, R., O'Sullivan, D. M., et al. (2014). An Eight-Parent Multiparent Advanced Generation Inter-Cross Population for Winter-Sown Wheat: Creation, Properties, and Validation. *G3 Genes|Genomes|Genetics* **4**:1603–1610.

Madhavi, D. L., and Salunkhe, D. K. (1998). Tomato. In *Handbook of vegetable science and technology : production, composition, storage, and processing / edited by D.K. Salunkhe, S.S. Kadam. - Version details - Trove*, p. New York.

Mägi, R., Horikoshi, M., Sofer, T., Mahajan, A., Kitajima, H., Franceschini, N., McCarthy, M. I., COGENT-Kidney Consortium, T2D-GENES Consortium, C.-K., Morris, A. P., and Morris, A. P. (2017). Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum. Mol. Genet.* **26**:3639–3650.

Mahajan, A., Wessel, J., Willems, S. M., Zhao, W., Robertson, N. R., Chu, A. Y., Gan, W., Kitajima, H., Taliun, D., Rayner, N. W., et al. (2018). Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* **50**:559–571.

Makowsky, R., Pajewski, N. M., Klimentidis, Y. C., Vazquez, A. I., Duarte, C. W., Allison, D. B., and de los Campos, G. (2011). Beyond Missing Heritability: Prediction of Complex Traits. *PLoS Genet.* **7**:e1002051.

Maldonado, C., Mora, F., Scapim, C. A., and Coan, M. (2019). Genome-wide haplotype-based association analysis of key traits of plant lodging and architecture of maize identifies major determinants for leaf angle: *HAPla4*. *PLoS One* Advance Access published 2019, doi:10.1371/journal.pone.0212925.

Malundo, T. M. M., Shewfelt, R. L., and Scott, J. W. (1995). Flavor quality of fresh tomato (*Lycopersicon esculentum* Mill.) as affected by sugar and acid levels. *Postharvest Biol. Technol.* **6**:103–110.

Mancuso, N., Freund, M. K., Johnson, R., Shi, H., Kichaev, G., Gusev, A., and Pasaniuc, B. (2019). Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet.* **51**:675–682.

Mangin, B., Rincent, R., Rabier, C. E., Moreau, L., and Goudemand-Dugue, E. (2019). Training set optimization of genomic prediction by means of EthAcc. *PLoS One* **14**:e0205629.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461**:747–753.

Mao, L., Begum, D., Chuang, H., Budiman, M. A., Szymkowiak, E. J., Irish, E. E., and Wing, R. A. (2000). *JOINTLESS* is a MADS-box gene controlling tomato flower abscissionzone development. *Nature* **406**:910–913.

Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**:499–511.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**:906–913.

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K. S., Wood, A. R., Kjaer, T. R., Fine, R. S., Lu, Y., Schurmann, C., Highland, H. M., et al. (2017). Rare and low-frequency coding variants alter human adult height. *Nature* **542**:186–190.

Martin, G. B., Brommonschenkel, S. H., Chunwongse, J., Frary, A., Ganal, M. W., Spivey, R., Wu, T., Earle, E. D., Tanksley, S. D., Sipvey, R., et al. (1993). Map-based cloning of a protein kinase gene conferring disease resistance in tomato. *Science (80-.).* **262**:1432–1436.

Martin, G. B., Frary, A., U, T. W., Brommonschenkel, S., Chunwongse, J., Earle, E. D., and Tanksley, S. D. (1994). A Member of the Tomato Pto Gene Family Confers Sensitivity to Fenthion Resulting in Rapid Cell Death. *Plant Cell* **6**:1543–1552.

Maynard Smith, J., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res. (Camb).* **23**:23–35.

Mazzucato, A., Papa, R., Bitocchi, E., Mosconi, P., Nanni, L., Negri, V., Picarella, M. E., Siligato, F., Soressi, G. P., Tiranti, B., et al. (2008). Genetic diversity, structure and marker-trait associations in a collection of Italian tomato (*Solanum lycopersicum* L.) landraces. *Theor. Appl. Genet.* **116**:657–669.

Mazzucato, A., Cellini, F., Bouzayen, M., Zouine, M., Mila, I., Minoia, S., Petrozza, A., Picarella, M. E., Ruiu, F., and Carriero, F. (2015). A TILLING allele of the tomato *Aux/IAA9* gene offers new insights into fruit set mechanisms and perspectives for breeding seedless tomatoes. *Mol. Breed.* **35**:22.

McAllister, E. J., Dhurandhar, N. V., Keith, S. W., Aronne, L. J., Barger, J., Baskin, M., Benca, R. M., Biggio, J., Boggiano, M. M., Eisenmann, J. C., et al. (2009). Ten putative contributors to the obesity epidemic. *Crit. Rev. Food Sci. Nutr.* **49**:868–913.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**:356–369.

McCouch, S. R., Wright, M. H., Tung, C.-W., Maron, L. G., McNally, K. L., Fitzgerald, M., Singh, N., DeClerck, G., Agosto-Perez, F., Korniliev, P., et al. (2016). Open access resources for genome-wide association mapping in rice. *Nat. Commun.* **7**:10532.

McDonald, J. H., and Kreitman, M. (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**:652–654.

Menda, N., Semel, Y., Peled, D., Eshed, Y., and Zamir, D. (2004). In silico screening of a saturated mutation library of tomato. *Plant J.* **38**:861–872.

Menda, N., Strickler, S. R., Edwards, J. D., Bombarely, A., Dunham, D. M., Martin, G. B., Mejia, L., Hutton, S. F., Havey, M. J., Maxwell, D. P., et al. (2014). Analysis of wild-species introgressions in tomato inbreds uncovers ancestral origins. *BMC Plant Biol.* **14**:287.

Meng, F. J., Xu, X. Y., Huang, F. L., and Li, J. F. (2010). Analysis of genetic diversity in cultivated and wild tomato varieties in Chinese market by RAPD and SSR. *Agric. Sci. China* **9**:1430–1437.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**:1819–1829.

Meyre, D., Delplanque, J., Chèvre, J.-C., Lecoeur, C., Lobbens, S., Gallina, S., Durand, E., Vatin, V., Degraeve, F., Proença, C., et al. (2009). Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat. Genet.* **41**:157–159.

Miller, J. C., and Tanksley, S. D. (1990). RFLP analysis of phylogenetic relationships and genetic variation in the genus Lycopersicon. *Theor. Appl. Genet.* **80**:437–448.

Mills, M. C., and Rahal, C. (2019). A scientometric review of genome-wide association studies. *Commun. Biol.* **2**:9.

Milner, S. E., Brunton, N. P., Jones, P. W., O' Brien, N. M., Collins, S. G., and Maguire, A. R. (2011). Bioactivities of Glycoalkaloids and Their Aglycones from Solanum Species. *J. Agric. Food Chem.* **59**:3454–3484.

Minamikawa, M. F., Nonaka, K., Kaminuma, E., Kajiya-Kanegae, H., Onogi, A., Goto, S., Yoshioka, T., Imai, A., Hamada, H., Hayashi, T., et al. (2017). Genome-wide association study and genomic prediction in citrus: Potential of genomics-assisted breeding for fruit quality traits. *Sci. Rep.* **7**:4721.

Minoia, S., Cellini, F., Bendahmane, A., D'Onofrio, O., Petrozza, A., Carriero, F., Piron, F., Mosca, G., and Sozio, G. (2010). A new mutant genetic resource for tomato crop improvement by TILLING technology. *BMC Res. Notes* **3**:39.

Minoïa, S., Bendahmane, A., Piron, F., Salgues, A., Moretti, A., Caranta, C., Piednoir, E., Nicolaï, M., and Zamir, D. (2010). An Induced Mutation in Tomato eIF4E Leads to Immunity to Two Potyviruses. *PLoS One* **5**:e11313.

Mølgaard, P., and Ravn, H. (1988). Evolutionary aspects of caffeoyl ester distribution In Dicotyledons. *Phytochemistry* **27**:2411–2421.

Monforte, A. J., and Tanksley, S. D. (2000). Fine mapping of a quantitative trait locus (QTL) from *Lycopersicon hirsutum* chromosome 1 affecting fruit characteristics and agronomic traits: breaking linkage among QTLs affecting different traits and dissection of heterosis for yield. *Theor. Appl. Genet.* **100**:471–479.

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Toledo, F. H., Pérez-Hernández, O., Eskridge, K. M., and Rutkoski, J. (2016). A Genomic Bayesian Multi-trait and Multi-environment Model. *G3 Genes|Genomes|Genetics* **6**:2725–2744.

Moonesinghe, R., Khoury, M. J., Liu, T., and Ioannidis, J. P. A. (2008). Required sample size and nonreplicability thresholds for heterogeneous genetic associations. *Proc. Natl. Acad. Sci. U. S. A.* **105**:617–22.

Mu, Q., Huang, Z., Chakrabarti, M., Illa-Berenguer, E., Liu, X., Wang, Y., Ramos, A., and van der Knaap, E. (2017). Fruit weight is controlled by Cell Size Regulator encoding a novel protein that is expressed in maturing tomato fruits. *PLOS Genet.* **13**:e1006930.

Mueller, L. A., Tanskley, S. D., Giovannoni, J. J., van Eck, J., Stack, S., Choi, D., Kim, B. D., Chen, M., Cheng, Z., Li, C., et al. (2005). The tomato sequencing project, the first cornerstone of the International Solanaceae Project (SOL). *Comp. Funct. Genomics* **6**:153–158.

Muir, S. R., Collins, G. J., Robinson, S., Hughes, S., Bovy, A., Ric De Vos, C. H., van Tunen, A. J., and Verhoeyen, M. E. (2001). Overexpression of petunia chalcone isomerase in tomato results in fruit containing increased levels of flavonols. *Nat. Biotechnol.* **19**:470–474.

Mulder, H. A., Calus, M. P. L., Druet, T., and Schrooten, C. (2012). Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. *J. Dairy Sci.* **95**:876–889.

Müller, B. S. F., Neves, L. G., de Almeida Filho, J. E., Resende, M. F. R., Muñoz, P. R., dos Santos, P. E. T., Filho, E. P., Kirst, M., and Grattapaglia, D. (2017). Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of Eucalyptus. *BMC Genomics* **18**:524.

Mutshinda, C. M., and Sillanpää, M. J. (2010). Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. *Genetics* **186**:1067–75.

N'Diaye, A., Haile, J. K., Cory, A. T., Clarke, F. R., Clarke, J. M., Knox, R. E., and Pozniak, C. J. (2017). Single marker and haplotype-based association analysis of semolina and pasta colour in elite durum wheat breeding lines using a high-density consensus map. *PLoS One* **12**:1–24.

Nakaya, A., and Isobe, S. N. (2012). Will genomic selection be a practical method for plant breeding? *Ann. Bot.* **110**:1303–1316.

Nazzicari, N., Biscarini, F., Cozzi, P., Brummer, E. C., and Annicchiarico, P. (2016). Marker imputation efficiency for genotyping-by-sequencing data in rice (*Oryza sativa*) and alfalfa (*Medicago sativa*). *Mol. Breed.* **36**:69.

Negro, S. S., Millet, E. J., Madur, D., Bauland, C., Combes, V., Welcker, C., Tardieu, F., Charcosset, A., and Nicolas, S. D. (2019). Genotyping-by-sequencing and SNP-arrays are complementary for detecting quantitative trait loci by tagging different haplotypes in association studies. *BMC Plant Biol.* **19**:318.

Nesbitt, T. C., and Tanksley, S. D. (2002). Comparative Sequencing in the Genus Lycopersicon: Implications for the Evolution of Fruit Size in the Domestication of Cultivated Tomatoes. *Genetics* **162**:365–379.

Nguyen, K. Le, Grondin, A., Courtois, B., and Gantet, P. (2018). Next-Generation Sequencing Accelerates Crop Gene Discovery. *Trends Plant Sci.* **24**:263–274.

Ofner, I., Lashbrooke, J., Pleban, T., Aharoni, A., and Zamir, D. (2016). *Solanum pennellii* backcross inbred lines (BILs) link small genomic bins with tomato traits. *Plant J.* **87**:151–160.

Okabe, Y., Asamizu, E., Saito, T., Matsukura, C., Ariizumi, T., Brès, C., Rothan, C., Mizoguchi, T., and Ezura, H. (2011). Tomato TILLING Technology: Development of a Reverse Genetics Tool for the Efficient Isolation of Mutants from Micro-Tom Mutant Libraries. *Plant Cell Physiol.* **52**:1994–2005.

Onengut-Gumuscu, S., Chen, W.-M., Burren, O., Cooper, N. J., Quinlan, A. R., Mychaleckyj, J. C., Farber, E., Bonnie, J. K., Szpak, M., Schofield, E., et al. (2015). Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47**:381–386.

Ongom, P. O., and Ejeta, G. (2017). Mating Design and Genetic Structure of a Multi-Parent Advanced Generation Intercross (MAGIC) Population of Sorghum (*Sorghum bicolor* (L.) Moench). *G3 Genes|Genomes|Genetics* **8**:331–341.

Overy, S. A., Walker, H. J., Malone, S., Howard, T. P., Baxter, C. J., Sweetlove, L. J., Hill, S. A., and Quick, W. P. (2004). Application of metabolite profiling to the identification of traits in a population of tomato introgression lines. *J. Exp. Bot.* **56**:287–296.

Pailles, Y., Ho, S., Pires, I. S., Tester, M., Negrão, S., and Schmöckel, S. M. (2017). Genetic Diversity and Population Structure of Two Tomato Species from the Galapagos Islands. *Front. Plant Sci.* **8**:138.

Panagiotou, O. A., Willer, C. J., Hirschhorn, J. N., and Ioannidis, J. P. A. (2013). The Power of Meta-Analysis in Genome-Wide Association Studies. *Annu. Rev. Genomics Hum. Genet.* **14**:441–465.

Paran, I., Goldman, I., Tanksley, S. D., and Zamir, D. (1995). Recombinant inbred lines for genetic mapping in tomato. *Theor. Appl. Genet.* **90**:542–548.

Park, T., and Casella, G. (2008). The Bayesian Lasso. *J. Am. Stat. Assoc.* **103**:681–686.

Park, Y. H., West, M. A. ., and St. Clair, D. A. (2004). Evaluation of AFLPs for germplasm fingerprinting and assessment of genetic diversity in cultivars of tomato (Lycopersicon esculentum L.). *Genome* **47**:510–518.

Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B. M., Daly, M. J., Sklar, P., et al. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.* **44**:631–635.

Pascual, L., Desplat, N., Huang, B. E., Desgroux, A., Bruguier, L., Bouchet, J. P., Le, Q. H., Chauchard, B., Verschave, P., and Causse, M. (2015). Potential of a tomato MAGIC population to decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. *Plant Biotechnol. J.* **13**:565–577.

Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E., and Tanksley, S. D. (1988). Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* **335**:721–726.

Paterson, A. H., DeVerna, J. W., Lanini, B., and Tanksley, S. D. (1990). Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes, in an interspecies cross of tomato. *Genetics* **124**:735–742.

Paterson, A. H., Damon, S., Hewitt, J. D., Zamir, D., Rabinowitch, H. D., Loncoln, S. E., Lander, E. S., and Tanksley, S. D. (1991). Mendelian factors underlying quantitative traits in tomato: Comparison across species, generations, and environments. *Genetics* **127**:181–197.

Pearson, K. (1904). Report on Certain Enteric Fever Inoculation Statistics. *Br. Med. J.* **2**:1243–6.

Pease, J. B., Haak, D. C., Hahn, M. W., and Moyle, L. C. (2016). Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. *PLoS Biol.* **14**:1–24.

Peng, M., Shahzad, R., Gul, A., Subthain, H., Shen, S., Lei, L., Zheng, Z., Zhou, J., Lu, D., Wang, S., et al. (2017). Differentially evolved glucosyltransferases determine natural variation of rice flavone accumulation and UV-tolerance. *Nat. Commun.* **8**:1975.

Peralta, I. E., Knapp, S., and Spooner, D. M. (2005). New Species of Wild Tomatoes (*Solanum* Section *Lycopersicon*: Solanaceae) from Northern Peru. *Syst. Bot.* **30**:424–434.

Pérez, P., de los Campos, G., Crossa, J., and Gianola, D. (2010). Genomic-Enabled Prediction Based on Molecular Markers and Pedigree Using the Bayesian Linear Regression Package in R. *Plant Genome J.* **3**:106.

Pérez, P., de los Campos, G., and Goddard, M. E. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **198**:483–95.

Pertuzé, R. A., Ji, Y., and Chetelat, R. T. (2003). Comparative linkage map of the *Solanum lycopersicoides* and *S. sitiens* genomes and their differentiation from tomato. *Genome* **45**:1003–1012.

Petró-Turza, M. (1986). Flavor of tomato and tomato products. *Food Rev. Int.* **2**:309–351.

Philouze, J. (1991). Description of isogenic lines, except for one, or two, monogenically controlled morphological traits in tomato, *Lycopersicon esculentum* Mill. *Euphytica* **56**:121–131.

Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Srinivasan, B. S., Barsh, G. S., Myers, R. M., Feldman, M. W., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* **19**:826–837.

Pietrella, M., and Giuliano, G. (2016). The sequencing: How it was done and whiat is produced. In *The tomato genome* (ed. Causse, M.), Giovannoni, J.), Bouzayen, M.), and Zouine, M.), pp. 39–73. Springer, Berlin, Heidelberg.

Pigeyre, M., Yazdi, F. T., Kaur, Y., and Meyre, D. (2016). Recent progress in genetics, epigenetics and metagenomics unveils the pathophysiology of human obesity. *Clin. Sci.* **130**:943–986.

Pnueli, L., Carmel-Goren, L., Hareven, D., Gutfinger, T., Alvarrez, H., Ganal, M., Zanir, D., and Lifschitz, E. (1998). The *SELF-PRUNING* gene of tomato regulates vegetative to reproductive switching of sympodial meristems and is the ortholog of *CEN* and *TFL1*. *Development* **125**:1979–1989.

Pollard, K. S., Salama, S. R., King, B., Kern, A. D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J. S., Bejerano, G., Baertsch, R., et al. (2006). Forces Shaping the Fastest Evolving Regions in the Human Genome. *PLoS Genet.* **2**:e168.

Porcu, E., Sanna, S., Fuchsberger, C., and Fritsche, L. (2013). Genotype Imputation in Genome-Wide Association Studies. In *Current Protocols in Human Genetics*, pp. 1.25.1-1.25.14. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**:559–575.

Qian, H.-R., and Huang, S. (2005). Comparison of false discovery rate methods in identifying genes with differential expression. *Genomics* **86**:495–503.

Quadrana, L., Almeida, J., Asís, R., Duffy, T., Dominguez, P. G., Bermúdez, L., Conti, G., Corrêa da Silva, J. V.,

Peralta, I. E., Colot, V., et al. (2014). Natural occurring epialleles determine vitamin E accumulation in tomato fruits. *Nat. Commun.* **5**:4027.

Quinet, M., Kinet, J.-M., and Lutts, S. (2011). Flowering response of the uniflora:blind:self-pruning and jointless:uniflora:self-pruning tomato (*Solanum lycopersicum*) triple mutants. *Physiol. Plant.* **141**:166–176.

Rakitsch, B., Lippert, C., Stegle, O., and Borgwardt, K. (2013). A Lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics* **29**:206–214.

Rambla, J. L., Medina, A., Fernández-del-Carmen, A., Barrantes, W., Grandillo, S., Cammareri, M., López-Casado, G., Rodrigo, G., Alonso, A., García-Martínez, S., et al. (2016). Identification, introgression, and validation of fruit volatile QTLs from a red-fruited wild tomato species. *J. Exp. Bot.* **68**:erw455.

Ramstein, G. P., Jensen, S. E., and Buckler, E. S. (2019). Breaking the curse of dimensionality to identify causal variants in Breeding 4. *Theor. Appl. Genet.* **132**:559–567.

Ranc, N., Mũos, S., Santoni, S., and Causse, M. (2008). A clarified position for *solanum lycopersicum* var. *cerasiforme* in the evolutionary history of tomatoes (solanaceae). *BMC Plant Biol.* **8**:130.

Ranc, N., Muños, S., Xu, J., Le Paslier, M.-C., Chauveau, A., Bounon, R., Rolland, S., Bouchet, J.-P., Brunel, D., and Causse, M. (2012). Genome-Wide Association Mapping in Tomato (*Solanum lycopersicum*) Is Possible Using Genome Admixture of Solanum lycopersicum var. cerasiforme. *G3 Genes|Genomes|Genetics* **2**:853–864.

Reddon, H., Gueant, J.-L., and Meyre, D. (2016). The importance of gene-environment interactions in human obesity. *Clin. Sci.* **130**:1571–1597.

Ribaut, J.-M., and Hoisington, D. (1998). Marker-assisted selection: new tools and strategies. *Trends Plant Sci.* **3**:236–239.

Richter, A., Schaff, C., Zhang, Z., Lipka, A. E., Tian, F., Köllner, T. G., Schnee, C., Preiß, S., Irmisch, S., Jander, G., et al. (2016). Characterization of Biosynthetic Pathways for the Production of the Volatile Homoterpenes DMNT and TMTT in *Zea mays*. *Plant Cell* **28**:2651–2665.

Rodríguez-Leal, D., Lemmon, Z. H., Man, J., Bartlett, M. E., and Lippman, Z. B. (2017). Engineering quantitative trait variation for crop improvement by genome editing. *Cell* **171**:470-480.e8.

Rodríguez, G. R., Muños, S., Anderson, C., Sim, S.-C., Michel, A., Causse, M., Gardener, B. B. M., Francis, D., and van der Knaap, E. (2011). Distribution of *SUN*, *OVATE*, *LC*, and *FAS* in the tomato germplasm and the relationship to fruit shape diversity. *Plant Physiol.* **156**:275–85.

Ronen, G., Cohen, M., Zamir, D., and Hirschberg, J. (1999). Regulation of carotenoid biosynthesis during tomato fruit development: expression of the gene for lycopene epsilon-cyclase is down-regulated during ripening and is elevated in the mutantDelta. *Plant J.* **17**:341–351.

Rothan, C., Diouf, I., and Causse, M. (2019). Trait discovery and editing in tomato. *Plant J.* **97**:73–90.

Rousseaux, M. C., Jones, C. M., Adams, D., Chetelat, R., Bennett, A., and Powell, A. (2005). QTL analysis of fruit antioxidants in tomato using *Lycopersicon pennellii* introgression lines. *Theor. Appl. Genet.* **111**:1396–1408.

Ruan, Y.-L., Patrick, J. W., Bouzayen, M., Osorio, S., and Fernie, A. R. (2012). Molecular regulation of seed and fruit set. *Trends Plant Sci.* **17**:656–665.

Ruggieri, V., Francese, G., Sacco, A., Alessandro, A. D., Rigano, M. M., Parisi, M., Milone, M., Cardi, T., Mennella, G., and Barone, A. (2014). An association mapping approach to identify favourable alleles for tomato fruit quality breeding. *BMC Plant Biol.* **14**:1–15.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**:832–837.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**:913–918.

Saïdou, A.-A., Thuillet, A.-C., Couderc, M., Mariac, C., and Vigouroux, Y. (2014). Association studies including genotype by environment interactions: prospects and limits. *BMC Genet.* **15**:3.

Saliba-Colombani, V., Causse, M., Langlois, D., Philouze, J., and Buret, M. (2001). Genetic analysis of organoleptic quality in fresh market tomato. 1. Mapping QTLs for physical and chemical traits. *Theor. Appl. Genet.* **102**:259–272.

Sallam, A., and Martsch, R. (2015). Association mapping for frost tolerance using multi-parent advanced generation inter-cross (MAGIC) population in faba bean (*Vicia faba* L.). *Genetica* **143**:501–514.

Salmeron, J. M., Oldroyd, G. E. ., Rommens, C. M. ., Scofield, S. R., Kim, H.-S., Lavelle, D. T., Dahlbeck, D., and Staskawicz, B. J. (1996). Tomato *Prf* Is a Member of the Leucine-Rich Repeat Class of Plant Disease Resistance Genes and Lies Embedded within the *Pto* Kinase Gene Cluster. *Cell* **86**:123–133.

Sandholt, C. H., Hansen, T., and Pedersen, O. (2012). Beyond the fourth wave of genome-wide obesity association studies. *Nutr. Diabetes* **2**:e37–e37.

Sargolzaei, M., Chesnais, J. P., and Schenkel, F. S. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* **15**:478.

Sauvage, C., Segura, V., Bauchet, G., Stevens, R., Do, P. T., Nikoloski, Z., Fernie, A. R., and Causse, M. (2014). Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant Physiol.* **165**:1120–1132.

Savage, J. E., Jansen, P. R., Stringer, S., Watanabe, K., Bryois, J., de Leeuw, C. A., Nagel, M., Awasthi, S., Barr, P. B., Coleman, J. R. I., et al. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**:912–919.

Schaffer, A. A., Levin, I., Oguz, I., Petreikov, M., Cincarevsky, F., Yeselson, Y., Shen, S., Gilboa, N., and Bar, M. (2000). ADPglucose pyrophosphorylase activity and starch accumulation in immature tomato fruit: the effect of a *Lycopersicon hirsutum*-derived introgression encoding for the large subunit. *Plant Sci.* **152**:135–144.

Schaid, D. J., Chen, W., and Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**:491–504.

Schauer, N., Zamir, D., and Fernie, A. R. (2005). Metabolic profiling of leaves and fruit of wild species tomato: A survey of the *Solanum lycopersicum* complex. *J. Exp. Bot.* **56**:297–307.

Schauer, N., Semel, Y., Roessner, U., Gur, A., Balbo, I., Carrari, F., Pleban, T., Perez-Melis, A., Bruedigam, C., Kopka, J., et al. (2006). Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nat. Biotechnol.* **24**:447–454.

Scheben, A., Batley, J., and Edwards, D. (2017). Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant Biotechnol. J.* **15**:149–161.

Scheinfeldt, L. B., and Tishkoff, S. A. (2013). Recent human adaptation: genomic approaches, interpretation and insights. *Nat. Rev. Genet.* **14**:692–702.

Schork, A. J., Thompson, W. K., Pham, P., Torkamani, A., Roddey, J. C., Sullivan, P. F., Kelsoe, J. R., O'Donovan, M. C., Furberg, H., Schork, N. J., et al. (2013). All SNPs Are Not Created Equal: Genome-Wide Association Studies Reveal a Consistent Pattern of Enrichment among Functionally Annotated SNPs. *PLoS Genet.* **9**:e1003449.

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* **44**:825–830.

Semel, Y., Nissenbaum, J., Menda, N., Zinder, M., Krieger, U., Issman, N., Pleban, T., Lippman, Z., Gur, A., and

Zamir, D. (2006a). Overdominant quantitative trait loci for yield and fitness in tomato. *Proc. Natl. Acad. Sci.* **103**:12981–12986.

Semel, Y., Nissenbaum, J., Menda, N., Zinder, M., Krieger, U., Issman, N., Pleban, T., Lippman, Z., Gur, A., and Zamir, D. (2006b). Overdominant quantitative trait loci for yield and fitness in tomato. *Proc. Natl. Acad. Sci.* **103**:12981–12986.

Servin, B., and Stephens, M. (2007). Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genet.* **3**:1296–1308.

Sewda, A., Agopian, A. J., Goldmuntz, E., Hakonarson, H., Morrow, B. E., Taylor, D., Mitchell, L. E., and Consortium, on behalf of the P. C. G. (2019). Gene-based genome-wide association studies and meta-analyses of conotruncal heart defects. *PLoS One* **14**:e0219926.

Shahlaei, A., Torabi, S., and Khosroshahli, M. (2014). Efficiacy of SCoT and ISSR marekers in assesment of tomato (*Lycopersicum esculentum* Mill.) genetic diversity. *Int. J. Biosci.* **5**:14–22.

Shalit, A., Rozman, A., Goldshmidt, A., Alvarez, J. P., Bowman, J. L., Eshed, Y., and Lifschitz, E. (2009). The flowering hormone florigen functions as a general systemic regulator of growth and termination. *Proc. Natl. Acad. Sci.* **106**:8392–8397.

Shammai, A., Petreikov, M., Yeselson, Y., Faigenboim, A., Moy-Komemi, M., Cohen, S., Cohen, D., Besaulov, E., Efrati, A., Houminer, N., et al. (2018). Natural genetic variation for expression of a SWEET transporter among wild species of *Solanum lycopersicum* (tomato) determines the hexose composition of ripening tomato fruit. *Plant J.* **96**:343–357.

Shapiro, B. J., and Alm, E. J. (2008). Comparing Patterns of Natural Selection across Species Using Selective Signatures. *PLoS Genet.* **4**:e23.

Shen, J., Tieman, D., Jones, J. B., Taylor, M. G., Schmelz, E., Huffaker, A., Bies, D., Chen, K., and Klee, H. J. (2014). A 13-lipoxygenase, TomloxC, is essential for synthesis of C5 flavour volatiles in tomato. *J. Exp. Bot.* **65**:419–428.

Shinozaki, Y., Nicolas, P., Fernandez-Pozo, N., Ma, Q., Evanich, D. J., Shi, Y., Xu, Y., Zheng, Y., Snyder, S. I., Martin, L. B. B., et al. (2018). High-resolution spatiotemporal transcriptome mapping of tomato fruit development and ripening. *Nat. Commun.* **9**:364.

Silventoinen, K., Sammalisto, S., Perola, M., Boomsma, D. I., Cornes, B. K., Davis, C., Dunkel, L., de Lange, M., Harris, J. R., Hjelmborg, J. V. B., et al. (2003). Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight Countries. *Twin Res.* **6**:399–408.

Sim, S.-C., Robbins, M. D., Van Deynze, A., Agee, M., and Francis, D. M. (2010). Population structure and genetic differentiation associated with breeding history and selection in tomato (*Solanum lycopersicum* L.). *Heredity (Edinb).* **106**:927–935.

Sim, S.-C., Durstewitz, G., Plieske, J., Wieseke, R., Ganal, M. W., van Deynze, A., Hamilton, J. P., Buell, C. R., Causse, M., Wijeratne, S., et al. (2012a). Development of a large snp genotyping array and generation of high-density genetic maps in tomato. *PLoS One* **7**.

Sim, S.-C., Van Deynze, A., Stoffel, K., Douches, D. S., Zarka, D., Ganal, M. W., Chetelat, R. T., Hutton, S. F., Scott, J. W., Gardner, R. G., et al. (2012b). High-Density SNP Genotyping of Tomato (*Solanum lycopersicum* L.) Reveals Patterns of Genetic Variation Due to Breeding. *PLoS One* **7**:e45520.

Smirnoff, N., and Wheeler, G. L. (2000). Ascorbic Acid in Plants: Biosynthesis and Function. *CRC. Crit. Rev. Plant Sci.* **19**:267–290.

Smith, M. L., and Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *Am. Psychol.* **32**:752–760.

Southam, L., Panoutsopoulou, K., Rayner, N. W., Chapman, K., Durrant, C., Ferreira, T., Arden, N., Carr, A.,

Deloukas, P., Doherty, M., et al. (2011). The effect of genome-wide association scan quality control on imputation outcome for common variants. *Eur. J. Hum. Genet.* **19**:610–614.

Soyk, S., Müller, N. A., Park, S. J., Schmalenbach, I., Jiang, K., Hayama, R., Zhang, L., Van Eck, J., Jiménez-Gómez, J. M., and Lippman, Z. B. (2017a). Variation in the flowering gene *SELF PRUNING 5G* promotes day-neutrality and early yield in tomato. *Nat. Genet.* **49**:162–168.

Soyk, S., Lemmon, Z. H., Oved, M., Fisher, J., Liberatore, K. L., Park, S. J., Goren, A., Jiang, K., Ramos, A., van der Knaap, E., et al. (2017b). Bypassing Negative Epistasis on Yield in Tomato Imposed by a Domestication Gene. *Cell* **169**:1142-1155.e12.

Spencer, C. C. A., Su, Z., Donnelly, P., and Marchini, J. (2009). Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet.* **5**.

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., Atlin, G., Jannink, J. L., and McCouch, S. R. (2015). Genomic Selection and Association Mapping in Rice (*Oryza sativa*): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Line. *PLoS Genet.* **11**:1–25.

Städler, T., Florez-Rueda, A. M., and Paris, M. (2012). Testing for "Snowballing" hybrid incompatibilities in *Solanum*: Impact of ancestral polymorphism and divergence estimates. *Mol. Biol. Evol.* **29**:31–34.

Stephens, M., and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10**:681–690.

Stevens, M. A. (1986). Inheritance of Tomato Fruit Quality Components. *Plant Breed. Rev.* **4**:273–311.

Stevens, M. A., Kader, A. A., and Albright-Holton, M. (1977). Intercultivar variation in composition of locular and pericarp portions of fresh market tomatoes. *J. Am. Soc. Hortic. Sci.* **102**:689–692.

Stevens, M., Kader, A. a., and Albright-Holton, M. (1979). Potential fot increasing tomato flavor via increased sugar and acid content. *J. Amer. Soc. Hort. Sci.* **104**:40–42.

Stevens, R., Buret, M., Duffe, P., Garchery, C., Baldet, P., Rothan, C., and Causse, M. (2007). Candidate Genes and Quantitative Trait Loci Affecting Fruit Ascorbic Acid Content in Three Tomato Populations. *Plant Physiol.* **143**:1943–1953.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **64**:479–498.

Sud, A., Kinnersley, B., and Houlston, R. S. (2017). Genome-wide association studies of cancer: Current insights and future perspectives. *Nat. Rev. Cancer* **17**:692–704.

Suliman-Pollatschek, S., Kashkush, K., Shats, H., Hillel, J., and Lavi, U. (2002). Generation and mapping of AFLP, SSRs and SNPs in Lycopersicon esculentum. *Cell. Mol. Biol. Lett.* **7**:583–597.

Sun, J., Poland, J. A., Mondal, S., Crossa, J., Juliana, P., Singh, R. P., Rutkoski, J. E., Jannink, J.-L., Crespo-Herrera, L., Velu, G., et al. (2019). High-throughput phenotyping platforms enhance genomic selection for wheat grain yield across populations and cycles in early stage. *Theor. Appl. Genet.* Advance Access published February 18, 2019, doi:10.1007/s00122-019-03309-0.

Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M., and Aulchenko, Y. S. (2012). Rapid variance components–based method for whole-genome association analysis. *Nat. Genet.* **44**:1166–1170.

Tadmor, Y., Fridman, E., Gur, A., Larkov, O., Lastochkin, E., Ravid, U., Zamir, D., and Lewinsohn, E. (2002). Identification of malodorous, a wild species allele affecting tomato aroma that was selected against during domestication. *J. Agric. Food Chem.* **50**:2005–2009.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**:585–

595.

Talukder, Z. I., Ma, G., Hulke, B. S., Jan, C.-C., and Qi, L. (2019). Linkage Mapping and Genome-Wide Association Studies of the Rf Gene Cluster in Sunflower (*Helianthus annuus* L.) and Their Distribution in World Sunflower Collections. *Front. Genet.* **10**:216.

Tam, S. M., Mhiri, C., Vogelaar, A., Kerkveld, M., Pearce, S. R., and Grandbastien, M. A. (2005). Comparative analyses of genetic diversities within tomato and pepper collections detected by retrotransposon-based SSAP, AFLP and SSR. *Theor. Appl. Genet.* **110**:819–831.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**:467–484.

Tanksley, S. D. (2004). The Genetic , developmental, and molecular bases of fruit size in tomato and shape variation. *Plant Cell* **16**:181–190.

Tanksley, S. D., Ganal, M. W., Prince, J. P., De Vicente, M. C., Bonierbale, M. W., Broun, P., Fulton, T. M., Giovannoni, J. J., Grandillo, S., Martin, G. B., et al. (1992). High density molecular linkage maps of the tomato and potato genomes. *Genetics* **132**:1141–1160.

Tanksley, S. D., Grandillo, S., Fulton, T. M., Zamir, D., Eshed, Y., Petiard, V., Lopez, J., and Beck-Bunn, T. (1996). Advanced backcross QTL analysis in a cross between an elite processing line of tomato and its wild relative *L. pimpinellifolium*. *Theor. Appl. Genet.* **92**:213–224.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science (80-.).* **337**:64–69.

The 100 Tomato Genome Sequencing Consortium (2014). Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. *Plant J.* **80**:136–148.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**:56–65.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* **526**:68–74.

The 1000 Genomes Project Consortium, Durbin, R. M., Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* **467**:1061–1073.

The 1000 Genomes Project Consortium, Gibbs, R. A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., et al. (2015). A global reference for human genetic variation. *Nature* **526**:68–74.

The 1001 Genomes Consortium (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. *Cell* **166**:481–491.

The 3000 rice genomes project (2014). The 3,000 rice genomes project. *Gigascience* **3**:7.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**:57–74.

The Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**:635–641.

The UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. *Nature* **526**:82–89.

Thompson, J. R., Attia, J., and Minelli, C. (2011). The meta-analysis of genome-wide association studies. *Brief. Bioinform.*

12:259–269.

Thorleifsson, G., Walters, G. B., Gudbjartsson, D. F., Steinthorsdottir, V., Sulem, P., Helgadottir, A., Styrkarsdottir, U., Gretarsdottir, S., Thorlacius, S., Jonsdottir, I., et al. (2009). Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat. Genet.* **41**:18–24.

Tieman, D., Taylor, M., Schauer, N., Fernie, A. R., Hanson, A. D., and Klee, H. J. (2006). Tomato aromatic amino acid decarboxylases participate in synthesis of the flavor volatiles 2-phenylethanol and 2-phenylacetaldehyde. *Proc. Natl. Acad. Sci. U. S. A.* **103**:8287–92.

Tieman, D., Bliss, P., McIntyre, L. M. M., Blandon-Ubeda, A., Bies, D., Odabasi, A. Z. Z., Rodríguez, G. R. R., Van Der Knaap, E., Taylor, M. G. G., Goulet, C., et al. (2012). The chemical interactions underlying tomato flavor preferences. *Curr. Biol.* **22**:1035–1039.

Tieman, D., Zhu, G., Resende, M. F. R., Lin, T., Nguyen, C., Bies, D., Rambla, J. L., Beltran, K. S. O., Taylor, M., Zhang, B., et al. (2017). A chemical genetic roadmap to improved tomato flavor. *Science (80-.).* **355**:391–394.

Tikunov, Y. (2005). A Novel Approach for Nontargeted Data Analysis for Metabolomics. Large-Scale Profiling of Tomato Fruit Volatiles. *Plant Physiol.* **139**:1125–1137.

Tikunov, Y. M., Molthoff, J., de Vos, R. C. H., Beekwilder, J., van Houwelingen, A., van der Hooft, J. J. J., Nijenhuis-de Vries, M., Labrie, C. W., Verkerke, W., van de Geest, H., et al. (2013). NON-SMOKY GLYCOSYLTRANSFERASE1 Prevents the Release of Smoky Aroma from Tomato Fruit. *Plant Cell* **25**:3067–3078.

Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J., and Richards, J. B. (2018). Genetic architecture: The shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19**:110–124.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39**:31–40.

Tranchida-Lombardo, V., Aiese Cigliano, R., Anzar, I., Landi, S., Palombieri, S., Colantuono, C., Bostan, H., Termolino, P., Aversano, R., Batelli, G., et al. (2018). Whole-genome re-sequencing of two Italian tomato landraces reveals sequence variations in genes associated with stress tolerance, fruit quality and long shelf-life traits. *DNA Res.* **25**:149–160.

Tucker, G., Price, A. L., and Berger, B. (2014). Improving the Power of GWAS and Avoiding Confounding from Population Stratification with PC-Select. *Genetics* **197**:1045–1049.

Tukiainen, T., Pirinen, M., Sarin, A.-P., Ladenvall, C., Kettunen, J., Lehtimäki, T., Lokki, M.-L., Perola, M., Sinisalo, J., Vlachopoulou, E., et al. (2014). Chromosome X-Wide Association Study Identifies Loci for Fasting Insulin and Height and Evidence for Incomplete Dosage Compensation. *PLoS Genet.* **10**:e1004127.

Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., Nguyen-Viet, T. A., Wedow, R., Zacher, M., Furlotte, N. A., et al. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**:229–237.

Upton, A., Trelles, O., Cornejo-García, J. A., and Perkins, J. R. (2016). Review: High-performance computing to detect epistasis in genome scale data sets. *Brief. Bioinform.* **17**:368–379.

Usadel, B., Chetelat, R., Koren, S., Maumus, F., Fernie, A. R., Aury, J.-M., Maß, J., Schmidt, M. H.-W., Denton, A. K., Wormit, A., et al. (2017). De Novo Assembly of a New *Solanum pennellii* Accession Using Nanopore Sequencing. *Plant Cell* **29**:2336–2348.

Van Berloo, R., Zhu, A., Ursem, R., Verbakel, H., Gort, G., and van Eeuwijk, F. A. (2008). Diversity and linkage disequilibrium analysis within a selected set of cultivated tomatoes. *Theor. Appl. Genet.* **117**:89–101.

van Binsbergen, R., Bink, M. C., Calus, M. P., van Eeuwijk, F. A., Hayes, B. J., Hulsegge, I., and Veerkamp, R. F.

(2014). Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genet. Sel. Evol.* **46**:41.

van der Knaap, E., and Tanksley, S. D. (2003). The making of a bell pepper-shaped tomato fruit: identification of loci controlling fruit morphology in Yellow Stuffer tomato. *Theor. Appl. Genet.* **107**:139–147.

VanRaden, P. M., Null, D. J., Sargolzaei, M., Wiggans, G. R., Tooker, M. E., Cole, J. B., Sonstegard, T. S., Connor, E. E., Winters, M., van Kaam, J. B. C. H. M., et al. (2013). Genomic imputation and evaluation using high-density Holstein genotypes. *J. Dairy Sci.* **96**:668–678.

Vargas-Ponce, O., Pérez-Álvarez, L. F., Zamora-Tavares, P., and Rodríguez, A. (2011). Assessing Genetic Diversity in Mexican Husk Tomato Species. *Plant Mol. Biol. Report.* **29**:733–738.

Verkerke, W., Janse, J., and Kersten, M. (1998). Instrumental Measurement and Modelling of Tomato Fruit Taste. *Acta Hortic.* **456**:199–206.

Villumsen, T. M., Janss, L., and Lund, M. S. (2009). The importance of haplotype length and heritability using genomic selection in dairy cattle. *J. Anim. Breed. Genet.* **126**:3–13.

Víquez-Zamora, M., Vosman, B., van de Geest, H., Bovy, A., Visser, R. G., Finkers, R., and van Heusden, A. W. (2013). Tomato breeding in the genomics era: Insights from a SNP array. *BMC Genomics* **14**:354.

Visscher, P. M. (2008). Sizing up human height variation. *Nat. Genet.* **40**:489–490.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**:5–22.

Vitti, J. J., Grossman, S. R., and Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annu. Rev. Genet.* **47**:97–120.

Vrebalov, J., Ruezinsky, D., Padmanabhan, V., White, R., Medrano, D., Drake, R., Schuch, W., and Giovannoni, J. (2002). A MADS-box gene necessary for fruit ripening at the tomato *Ripening-inhibitor* (Rin) locus. *Science (80-.).* **296**:343–346.

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**:592–599.

Wang, K., Li, M., and Bucan, M. (2007). Pathway-Based Approaches for Analysis of Genomewide Association Studies. *Am. J. Hum. Genet.* **81**:1278–1283.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**:57–63.

Wang, K., Li, M., and Hakonarson, H. (2010a). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**:e164.

Wang, K., Li, M., and Hakonarson, H. (2010b). Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* **11**:843–854.

Wang, H., Woodward, B., Bauck, S., and Rekaya, R. (2012a). Imputation of missing SNP genotypes using low density panels. *Livest. Sci.* **146**:80–83.

Wang, D., Salah El-Basyoni, I., Stephen Baenziger, P., Crossa, J., Eskridge, K. M., and Dweikat, I. (2012b). Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. *Heredity (Edinb).* **109**:313–9.

Wang, C., Habier, D., Peiris, B. L., Wolc, A., Kranis, A., Watson, K. A., Avendano, S., Garrick, D. J., Fernando, R. L., Lamont, S. J., et al. (2013). Accuracy of genomic prediction using an evenly spaced, low-density single nucleotide

polymorphism panel in broiler chickens. *Poult. Sci.* **92**:1712–1723.

Wang, Q., Tian, F., Pan, Y., Buckler, E. S., and Zhang, Z. (2014). A SUPER Powerful Method for Genome Wide Association Study. *PLoS One* **9**:e107684.

Wang, S., Zhao, J. H., An, P., Guo, X., Jensen, R. A., Marten, J., Huffman, J. E., Meidtner, K., Boeing, H., Campbell, A., et al. (2016a). General Framework for Meta-Analysis of Haplotype Association Tests. *Genet. Epidemiol.* **40**:244–252.

Wang, S.-B., Feng, J.-Y., Ren, W.-L., Huang, B., Zhou, L., Wen, Y.-J., Zhang, J., Dunwell, J. M., Xu, S., and Zhang, Y.-M. (2016b). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* **6**:19444.

Wang, D. R., Agosto-Pérez, F. J., Chebotarov, D., Shi, Y., Marchini, J., Fitzgerald, M., McNally, K. L., Alexandrov, N., and McCouch, S. R. (2018). An imputation platform to enhance integration of rice genetic resources. *Nat. Commun.* **9**:3519.

Wang, S., Alseekh, S., Fernie, A. R., and Luo, J. (2019). The Structure and Function of Major Plant Metabolite Modifications. *Mol. Plant* **12**:899–919.

Ward, L. D., and Kellis, M. (2016). HaploReg v4: Systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* **44**:D877–D881.

Wei, W.-H., Hemani, G., and Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nat. Rev. Genet.* **15**:722–733.

Weigel, K. A., Van Tassell, C. P., O'Connell, J. R., VanRaden, P. M., and Wiggans, G. R. (2010). Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. *J. Dairy Sci.* **93**:2229–2238.

Wellmann, R., Preuß, S., Tholen, E., Heinkel, J., Wimmers, K., and Bennewitz, J. (2013). Genomic selection using low density marker panels with application to a sire line in pigs. *Genet. Sel. Evol.* **45**:28.

Wermter, A.-K., Scherag, A., Meyre, D., Reichwald, K., Durand, E., Nguyen, T. T., Koberwitz, K., Lichtner, P., Meitinger, T., Schäfer, H., et al. (2008). Preferential reciprocal transfer of paternal/maternal *DLK1* alleles to obese children: first evidence of polar overdominance in humans. *Eur. J. Hum. Genet.* **16**:1126–1134.

Westfall, P. H., and Young, S. S. (1993). *Resampling-based multiple testing : examples and methods for P-value adjustment*. 1st Editio. Wiley.

Westra, H.-J., Martínez-Bonet, M., Onengut-Gumuscu, S., Lee, A., Luo, Y., Teslovich, N., Worthington, J., Martin, J., Huizinga, T., Klareskog, L., et al. (2018). Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes. *Nat. Genet.* **50**:1.

Wheeler, E., Huang, N., Bochukova, E. G., Keogh, J. M., Lindsay, S., Garg, S., Henning, E., Blackburn, H., Loos, R. J. F., Wareham, N. J., et al. (2013). Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. *Nat. Genet.* **45**:513–517.

Wijmenga, C., and Zhernakova, A. (2018). The importance of cohort studies in the post-GWAS era. *Nat. Genet.* **50**:1–7.

Willits, M. G., Kramer, C. M., Prata, R. T. N., De Luca, V., Potter, B. G., Stephens, J. C., and Graser, G. (2005). Utilization of the genetic resources of wild species to create a nontransgenic high flavonoid tomato. *J. Agric. Food Chem.* **53**:1231–1236.

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**:1173–1186.

Wood, A. R., Tyrrell, J., Beaumont, R., Jones, S. E., Tuke, M. A., Ruth, K. S., Yaghootkar, H., Freathy, R. M., Murray, A., Frayling, T. M., et al. (2016). Variants in the *FTO* and *CDKAL1* loci have recessive effects on risk of obesity and type 2 diabetes, respectively. *Diabetologia* **59**:1214–1221.

Wright, S. I., and Charlesworth, B. (2004). The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. *Genetics* **168**:1071–6.

Wu, R., and Zeng, Z. B. (2001). Joint linkage and linkage disequilibrium mapping in natural populations. *Genetics* **157**:899–909.

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J., and Knaap, E. van der (2008). A Retrotransposon-Mediated Gene Duplication Underlies Morphological Variation of Tomato Fruit. *Science (80-.).* **319**:1527–1530.

Xu, Y., and Crouch, J. H. (2008). Marker-Assisted Selection in Plant Breeding: From Publications to Practice. *Crop Sci.* **48**:391.

Xu, H., and Guan, Y. (2014). Detecting local haplotype sharing and haplotype association. *Genetics* **197**:823–838.

Xu, J., Ranc, N., Muños, S., Rolland, S., Bouchet, J.-P. P., Desplat, N., Le Paslier, M.-C. C., Liang, Y., Brunel, D., and Causse, M. (2013). Phenotypic diversity and association mapping for fruit quality traits in cultivated tomato and related species. *Theor. Appl. Genet.* **126**:567–581.

Yamamoto, E., Matsunaga, H., Onogi, A., Kajiya-Kanegae, H., Minamikawa, M., Suzuki, A., Shirasawa, K., Hirakawa, H., Nunome, T., Yamaguchi, H., et al. (2016). A simulation-based breeding design that uses whole-genome prediction in tomato. *Sci. Rep.* **6**:19454.

Yamamoto, E., Matsunaga, H., Onogi, A., Ohyama, A., Miyatake, K., Yamaguchi, H., Nunome, T., Iwata, H., and Fukuoka, H. (2017). Efficiency of genomic selection for breeding population design and phenotype prediction in tomato. *Heredity (Edinb).* **118**:202–209.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**:565–569.

Yang, J., Liu, D., Wang, X., Ji, C., Cheng, F., Liu, B., Hu, Z., Chen, S., Pental, D., Ju, Y., et al. (2016). The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. *Nat. Genet.* **48**:1225–1232.

Yang, J., Zeng, J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2017). Concepts, estimation and interpretation of SNP-based heritability. *Nat. Genet.* **49**:1304–1310.

Yang, R. Y., Quan, J., Sodaei, R., Aguet, F., Segrè, A. V., Allen, J. A., Lanz, T. A., Reinhart, V., Crawford, M., Hasson, S., et al. (2018). A systematic survey of human tissue-specific gene expression and splicing reveals new opportunities for therapeutic target identification and evaluation. *bioRxiv* Advance Access published April 30, 2018, doi:10.1101/311563.

Yates, F., and Cochran, W. G. (1938). The analysis of groups of experiments. *J. Agric. Sci.* **28**:556–580.

Ye, J., Wang, X., Hu, T., Zhang, F., Wang, B., Li, C., Yang, T., Li, H., Lu, Y., Giovannoni, J. J., et al. (2017). An inDel in the promoter of *Al-ACTIVATED MALATE TRANSPORTER9* selected during tomato domestication determines fruit malate contents and aluminum tolerance. *Plant Cell* **29**:2249–2268.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**:203–208.

Zeggini, E., and Ioannidis, J. P. A. (2009). Meta-analysis in genome-wide association studies. *Pharmacogenomics* **10**:191–201.

Zeng, J., De Vlaming, R., Wu, Y., Robinson, M. R., Lloyd-Jones, L. R., Yengo, L., Yap, C. X., Xue, A., Sidorenko, J., McRae, A. F., et al. (2018). Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**:746–753.

Zhan, X., Zhao, N., Plantinga, A., Thornton, T. A., Conneely, K. N., Epstein, M. P., and Wu, M. C. (2017). Powerful Genetic Association Analysis for Common or Rare Variants with High-Dimensional Structured Traits. *Genetics* **206**:1779–1790.

Zhang, Z., and Druet, T. (2010). Marker imputation with low-density marker panels in Dutch Holstein cattle. *J. Dairy Sci.* **93**:5487–5494.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**:355–360.

Zhang, J., Zhao, J., Xu, Y., Liang, J., Chang, P., Yan, F., Li, M., Liang, Y., and Zou, Z. (2015). Genome-Wide Association Mapping for Tomato Volatiles Positively Contributing to Tomato Flavor. *Front. Plant Sci.* **6**:1042.

Zhang, J., Zhao, J., Liang, Y., and Zou, Z. (2016). Genome-wide association-mapping for fruit quality traits in tomato. *Euphytica* **207**:439–451.

Zhang, J., Feng, J.-Y., Ni, Y.-L., Wen, Y.-J., Niu, Y., Tamba, C. L., Yue, C., Song, Q., and Zhang, Y.-M. (2017). pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity (Edinb).* **118**:517–524.

Zhao, J., Xu, Y., Ding, Q., Huang, X., Zhang, Y., Zou, Z., Li, M., Cui, L., and Zhang, J. (2016). Association Mapping of Main Tomato Fruit Sugars and Organic Acids. *Front. Plant Sci.* **7**:1–11.

Zhao, J., Sauvage, C., Zhao, J., Bitton, F., Bauchet, G., Liu, D., Huang, S., Tieman, D. M., Klee, H. J., and Causse, M. (2019). Meta-analysis of genome-wide association studies provides insights into genetic control of tomato flavor. *Nat. Commun.* **10**:1534.

Zhong, S., Fei, Z., Chen, Y., Zheng, Y., Huang, M., Vrebalov, J., McQuinn, R., Gapper, N., Liu, B., Xiang, J., et al. (2013). Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. *Nat. Biotechnol.* **31**:154–159.

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**:821–824.

Zhou, X., and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* **11**:407–409.

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genet.* **9**:e1003264.

Zhou, R., Wu, Z., Cao, X., and Jiang, F. (2015). Genetic diversity of cultivated and wild tomatoes revealed by morphological traits and SSR markers. *Genet. Mol. Res.* **14**:13868–13879.

Zhou, J. J., Hu, T., Qiao, D., Cho, M. H., and Zhou, H. (2016). Boosting Gene Mapping Power and Efficiency with Efficient Exact Variance Component Tests of Single Nucleotide Polymorphism Sets. *Genetics* **204**:921–931.

Zhou, P., Hirsch, C. N., Briggs, S. P., and Springer, N. M. (2019). Dynamic Patterns of Gene Expression Additivity and Regulatory Variation throughout Maize Development. *Mol. Plant* **12**:410–425.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**:481–487.

Zhu, G., Wang, S., Huang, Z., Zhang, S., Liao, Q., Zhang, C., Lin, T., Qin, M., Peng, M., Yang, C., et al. (2018). Rewiring of the fruit metabolome in tomato breeding. *Cell* **172**:249-261.e12.

Zouine, M., Maza, E., Djari, A., Lauvernier, M., Frasse, P., Smouni, A., Pirrello, J., and Bouzayen, M. (2017). TomExpress, a unified tomato RNA-Seq platform for visualization of expression data, clustering and correlation networks. *Plant J.* **92**:727–735.

Zuriaga, E., Blanca, J., and Nuez, F. (2009). Classification and phylogenetic relationships in *Solanum* section *Lycopersicon* based on AFLP and two nuclear gene sequences. *Genet. Resour. Crop Evol.* **56**:663–678.

Chapter 2

Chapter 2 Summary of Materials and Methods

This chapter provides brief summaries of the materials and methods used in this thesis. Detailed information will not be provided here but in the two following chapters.

2.1 Summary of materials

Overall, this thesis will mainly focus on three GWAS panels, which have been both genotyped and phenotyped with a diverse set of flavor-related traits. They include panel S (Sauvage et al., 2014), panel B (Bauchet et al., 2017) and panel T (Tieman et al., 2017). Summary of the three panels and traits is provided in **Table 2.1**. Detailed explanations about field experiment, genotyping, phenotyping, quality control and GWAS are available in the corresponding articles.

Table 2.1 Summary of three GWAS panels used in this thesis.

GWAS panel	Panel S	Panel B	Panel T
Panel code	S	B	T
Population size	163	300	402
Phenotype replications	2007 and 2008	2011 and 2012	Florida and Israel
Genotype method	SOLCAP arrays	SOLCAP and CBSG arrays	Whole-genome sequencing
Genotyped SNPs	5,995	9,013	2,014,488
MAF	0.037 < MAF < 0.45	MAF > 0.01	MAF > 0.05
Population structure	K=2	K=6	K=5
GWAS model	MLMM	MLMM	EMMAX
Sugars and acids 4	Citrate & Malate	Citrate & Malate	Citrate & Malate
	Fructose	Fructose	Fructose
	Glucose	Glucose	Glucose
Amino acids 10	Asparagine	Asparagine	
	Aspartate	Aspartic acid	
	Glutamine	Glutamine	
	… 7 others	… 7 others	
Volatiles 17		(E)-2-heptenal	(E)-2-heptenal
		(E)-2-hexenal	(E)-2-hexenal
		(Z)-3-hexenal	(Z)-3-hexenal
		… 14 others	… 14 others

2.2 Multi-haplotype based analyses

In order to test the potential benefits of using haplotypes compared to single markers, we chose panel S as an example for detailed investigations. We applied haplotypes to several main aspects: 1) haplotypes/haplotype block evaluations within subgroups of panel S, 2) using integrated haplotype score (iHS) to detect selective sweeps and compare them with allelic diversity (π) derived sweeps, 3) compare haplotype-based mixed association model (hapQTL) (Xu and Guan, 2014) with multi-locus mixed model (MLMM) (Segura et al., 2012) and single-locus mixed model (EMMAX) (Kang et al., 2010), 4) use marker local haplotype sharing (mLHS) to choose the candidate block and compare it with linkage disequilibrium, compare the ancestral and derived haplotypes of the focal significantly associated SNPs and following functional analyses, such as gene annotation and transcriptome analyses (**Figure 2.1**).

Though genomic selection is not the central focus of this thesis, we still think it is interesting and quite helpful to apply haplotypes to improve the prediction accuracy. We thus tested its efficiency. All these knowledge together will tell us 1) where has been selected in tomato genome? Will haplotype improve the statistical power in identifying new associations? Whether these associations are overlapped with the selective sweeps? Are there some promising candidate genes with a close functional annotation, etc. We hope these analyses will bridge us from past to the future of tomato breeding (see details in **Chapter 3**).

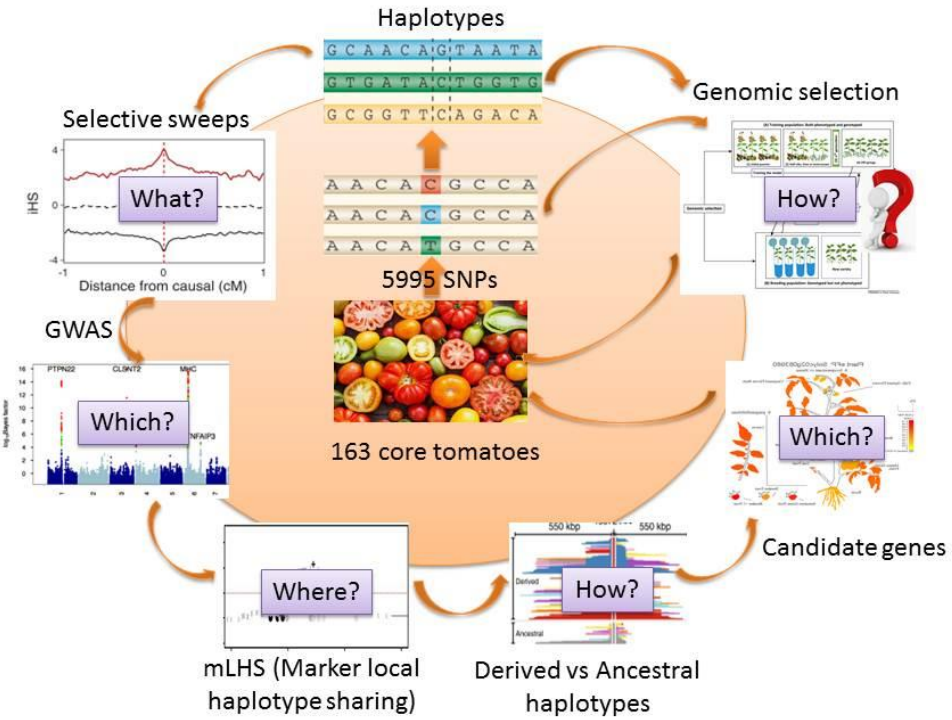


Figure 2.1 Global illustration of technical plan of multi-haplotype based analyses using panel S.

2.3 Meta-analysis of GWAS

The three GWAS panels were used for GWAS meta-analysis following the steps illustrated **Figure 2.2**. A total of 788 tomato accessions and 2,316,117 SNPs from the three GWAS panels were used for the final meta-analysis. We have imputed genotypes in panels S and B using IMPUTE2 (Howie et al., 2009) and demonstrated that it could increase the SNP density to 30-50 folds more, with high quality controls (detailed methods in Chapter 4). In order to avoid the heterogeneity caused by association model, we re-run the GWAS following the same model (EMMAX) for those flavor-related traits that were measured in at least two panels. We then first performed the fixed-effect meta-analysis model for all SNPs using the software METAL (Willer et al., 2010). For those SNPs with heterogeneity, we then performed the random-effect meta-analysis model proposed in METASOFT software (Han and Eskin, 2011). Once significant associations were detected, we also screened for promising candidate genes as examples to provide clues why modern tomatoes have a deteriorated flavor and how to improve the overall quality of tomato (**see details in Chapter 4**).

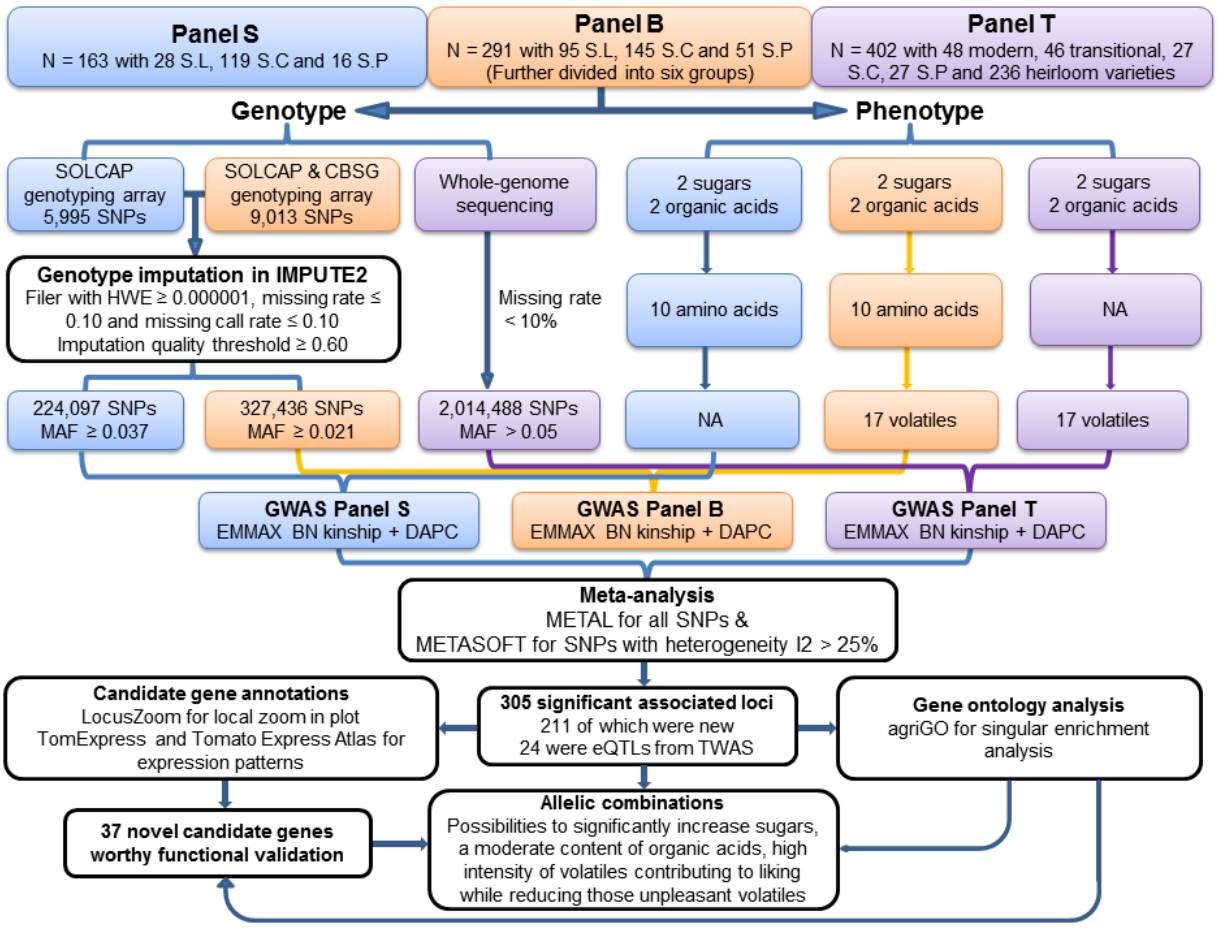


Figure 2.3 Global strategy of meta-analysis of genome-wide association studies from three GWAS panels.

References

Bauchet G, Grenier S, Samson N, Segura V, Kende A, Beekwilder J, Cankar K, Gallois J-L, Gricourt J, Bonnet J, et al (2017) Identification of major loci and genomic regions controlling acid and volatile content in tomato fruit: implications for flavor improvement. New Phytol **215**: 624–641

Han B, Eskin E (2011) Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am J Hum Genet **88**: 586–598

Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet **5**: e1000529

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet **42**: 348–354

Sauvage C, Segura V, Bauchet G, Stevens R, Do PT, Nikoloski Z, Fernie AR, Causse M (2014) Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. Plant Physiol **165**: 1120–1132

Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, Nordborg M (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet **44**: 825–830

Tieman D, Zhu G, Resende MFR, Lin T, Nguyen C, Bies D, Rambla JL, Beltran KSO, Taylor M, Zhang B, et al (2017) A chemical genetic roadmap to improved tomato flavor. Science (80-) **355**: 391–394

Willer CJ, Li Y, Abecasis GR (2010) METAL: Fast and efficient meta-analysis of genomewide association scans. Bioinformatics **26**: 2190–2191

Xu H, Guan Y (2014) Detecting local haplotype sharing and haplotype association. Genetics **197**: 823–838

Chapter 3

Chapter 3

Multiple haplotype-based analyses provide genetic and evolutionary insights into tomato fruit weight and composition

This chapter is a draft manuscript with a central focus on the combination of population and quantitative genetics to deepen our knowledge of marker-trait associations for fruit weight and composition in tomato. We aimed at deciphering the molecular footprints of selection, identifying haplotype – trait associations, providing a description of the haplotype landscape under marker – trait associations and comparing marker local haplotype sharing with linkage disequilibrium estimates to narrow down the search for candidate genes. We also tested the benefits of using haplotypes in improving the genomic prediction as general discussion and put more emphasis on the promise of this type of approach for breeding purposes.

Multiple haplotype-based analyses provide genetic and evolutionary insights into tomato fruit weight and composition

Jiantao Zhao[1], Christopher Sauvage[1,2], Frédérique Bitton[1], Mathilde Causse[1]

INRA, UR1052, Génétique et Amélioration des Fruits et Légumes, Domaine Saint Maurice, 67 Allée des Chênes CS 60094 – 84143 Montfavet Cedex, France

***Corresponding author**

Mathilde Causse: mathilde.causse@inra.fr

[1]INRA, UR1052, Centre de Recherche PACA, Génétique et Amélioration des Fruits et Légumes, Domaine Saint Maurice, 67 Allée des Chênes CS 60094 – 84140
Montfavet Cedex, France

[2]Present address: Syngenta SAS France, 1228 Chemin de l'Hobit, Saint Sauveur 31790, France

Running title: Multiple haplotype-based analyses of tomato quality traits

Abstract

Improving fruit quality through its metabolic composition is a main challenge for tomato breeders. For a better understanding of the demographic history of the loci controlling tomato quality traits and deepening the knowledge of the genetic architecture of these traits, an innovative approach was applied based on multiple haplotype-based analyses. In this study, we performed and compared single SNP and multiple haplotype-based association analyses focusing on tomato fruit weight and metabolite contents. First, we detected a total of 784 haplotype blocks in a collection of 163 tomato accessions. The average size of haplotype blocks was 58.085 kb. By using integrated haplotype score (iHS), we identified 24 positive selective sweeps, among which, nine were non-overlapping with either domestication or improvement sweep. Haplotype and SNP-based Bayesian models identified 108 significant associations for 26 traits, which outperformed previous studies. Among the associations, 77 were located within selective sweeps. Marker local haplotype sharing (mLHS) provided an alternative to linkage disequilibrium decay pattern to define confidence intervals around the associations to seek for candidate genes. Local haplotype decaying pattern and the length of haplotypes within different groups of accessions provided new insights on the demographic history of the associated loci. We thus demonstrate the power of using haplotypes for evolutionary and genetic studies, providing novel insights into tomato quality improvement and breeding history.

Introduction

Our understanding of the genetic architecture of agronomical traits is guided by the combination of technical (i.e. sequencing), analytical (i.e. statistics) and theoretical advances (i.e. population and quantitative genetics). Up to now, the vast majority of marker trait-associations were revealed using QTL and GWAS mapping. The later approach relies on the linkage disequilibrium (LD) between the gene(s) that control the variation of the trait and a single molecular marker. While being successful for detecting loci of large effect, it remains limited to decipher the additional medium to low effect loci to track down the missing heritability of traits of interest (Brachi et al., 2011; Eichler et al., 2010). In addition, a strong knowledge of the structure of LD is required, particularly the distance to which LD extends and how much it varies from one chromosomal region to another in the population under study as it mostly drives towards the successful identification of a candidate gene. Also,

defining the window size around a significant SNP to look for candidate genes still remains challenging, especially when LD is high over a relatively large region (Schaid et al., 2018). To do so, the LD R^2 is usually used as the threshold (usually ranging from 0.3 to 0.8, depending on stringency levels; Xu et al., 2013; Sauvage et al., 2014; Tieman et al., 2017; Bauchet et al., 2017).

Haplotypes are the particular combinations of alleles observed on a region of a chromosome in a given population (Gabriel et al., 2002; Belmont et al., 2003). Haplotype blocks are the regions where there is little evidence for historical recombination for example and within which only a few common haplotypes are observed (Gabriel et al., 2002). Genotyping only a few, carefully chosen tag-SNPs will provide enough information to identify the common haplotypes (Daly et al., 2001; Johnson et al., 2001; Belmont et al., 2003). Alleles within the same haplotype block are more likely to be inherited together (Farashi et al., 2019) while sharing similar minor allele frequency (MAF). Haplotype-based analyses examine groups of SNPs rather than individual SNPs and enhance the statistical detection power for many aspects, including identifying signals of recent positive selection (Sabeti et al., 2002) and genome-wide association studies (GWAS, Khatkar et al., 2007; Gawenda et al., 2015; Maldonado et al., 2019).

The nature and diversity of haplotypes also witnesses of the forces that shaped genetic variation in a genome and the consequences for breeding notably. Among the four major evolutionary forces, selection refers to any non-random, differential propagation of an allele. Positive selection causes a beneficial allele and hitchhicked variants to sweep to high prevalence (soft-sweep) or even reach to fixation (hard sweep) within a population, thereby producing a population-wide reduction of genetic diversity (Vitti et al., 2013; Hermisson and Pennings, 2017). Identifying genomic regions with unusually high local haplotype homozygosity represents a powerful strategy to identify natural or artificial recent selection events (Gautier et al., 2017). Among the haplotype detection methods, the integrated haplotype score (iHS) compares the area under the curve defined by extended haplotype homozygosity (EHH) for the derived and ancestral allele (Voight et al., 2006). Thus, iHS provides helpful insights into the genome-wide distributions of very recent selective events in favor of alleles that have not yet reached fixation (Voight et al., 2006).

Tomato is the most consumed vegetable worldwide (http://www.fao.org/faostat). Cultivated tomato (*Solanum lycopersicum* L.) has experienced severe bottlenecks during its

domestication and breeding history, resulting in a narrow genetic diversity (Bauchet and Causse, 2012; Lin et al., 2014; Gao et al., 2019). Tomatoes, among the intensively bred fruit crops, are widely viewed as lacking flavor (Klee and Tieman, 2018). Many important metabolites in tomato are flavor-related, but often time-consuming to quantify (Klee and Tieman, 2013). High-throughput genomic approaches thus provide new opportunities for tomato quality improvements (Tieman et al., 2017; Zhu et al., 2018; Zhu et al., 2019; Zhao et al., 2019).

In this study we aimed at benefiting from the haplotype landscape of the tomato genome to (1) deepen our knowledge about the recent breeding history, (2) test the potential of haplotypes to detect new candidate regions for QTL and predict phenotypes in genomic selection context. To do so, we first calculated in a large set of wild and cultivated accessions the haplotype blocks between all accessions and across population subgroups. We then identified the recent positive selective sweeps, which were compared with domestication and improvement sweeps previously reported in the literature (Lin et al., 2014). We then compared associations based on haplotype or SNP-based Bayes model with single-marker based association model (EMMAX; Kang et al., 2010) and multi-locus mixed model (MLMM; Segura et al., 2012). We showed that marker local haplotype sharing (mLHS) provided an alternative to linkage disequilibrium to choose the window size for detecting candidate genes. Our multiple haplotype-based analyses demonstrated the potentials of using haplotypes for several aspects.

RESULTS

Haplotype block estimation

To investigate the haplotype landscape in the tomato genome, we first determined the haplotype blocks within all accessions and within each of the three subgroups composed of 116 *S. l. cerasiforme* accessions (cherry tomato, CER), 31 *S. lycopersicum* (large-fruit tomato, BIG) and 16 *S. pimpinellifolium* (the closest wild species, PIM). Within all the 163 tomato accessions, we detected a total of 784 haplotype blocks (**Table S1**). Within the three subgroups, we detected 704, 259 and 134 haplotype blocks for CER, BIG and PIM groups, respectively (**Table S2-S4**). We observed significant positive correlation between the number of accessions and haplotype blocks (R^2 = 0.99, P = 0.0497). The average size of haplotype

blocks was 58.085 kb. The genome-wide distribution pattern of the size of haplotype blocks was similar in different groups (**Figure 1A**). Across chromosomes, the global distribution of the size of haplotype blocks was similar, though many differences were also observed (**Figure S1**). Within the haplotype blocks, there were 5.2 SNPs on average, ranging from 2 to 46 SNPs (**Figure 1B**). The distribution patterns of SNPs were similar between the different groups (**Figure S2**).

Haplotype blocks were more likely to be located at both ends of the chromosome arms. The most frequenct haplotype blocks within the genetic subgroups were overlapping with SNP marker distribution along chromosomes and were more numerous at each end of chromosomes' arms. in terms of their positions. (**Figure 1C**).

Identification of positive selective sweeps using integrated haplotype score

By using integrated haplotype score (iHS), we identified a total of 24 positive selective sweeps (PSS01 – PSS24) (**Table S5**). We then calculated the linkage disequilibrium of the peak SNPs for each positive selective sweep and used $R^2 = 0.5$ as the threshold to define the window size, which ranged from 12 kb to 18.7 Mb, with an average size of 3.43 Mb (**Table S5**). Genes identified within the PSS were listed in **Table S6.**

In order to classify whether the positive selective sweeps were caused during the domestication or the improvement stage, we calculated the nucleotide diversity (π) in PIM, CER and BIG, and identified 132 domestication sweeps (DS001-DS132, $\pi_{PIM/CER} > 3.43$) (**Table S7**) and 93 improvement sweeps (IS001-IS093, $\pi_{CER/BIG} > 6.16$) (**Table S8**). All the genes within the domestication/improvement sweeps were listed in **Table S9** and **Table S10**. Among these, 13 domestication sweeps and 10 improvement sweeps overlapped with 8 and 6 positive selective sweeps, respectively. Among the 24 PSS, 9 were not overlapping with either domesticated or improvement sweeps. There were 5967, 3455 and 5700 genes located within PSS, DS and IS, respectively. Among these, 871 were overlapping across PSS and DS and 1270 were overlapping across PSS and IS (**Figure S3A**). In addition, PSS, DS and IS covered a total of genomic size of 75.208; 43.959 and 52.539 Mb, respectively, accounting for 30.46%, 20.38% and 23.43% of the tomato genome (**Figure S3B**).

Marker local haplotype sharing provided an alternative to linkage disequilibrium

For each of the associations detected in hapQTL, we calculated the marker local haplotype sharing (mLHS) for the peak associated SNPs and used mLHS = 0.2 as the threshold to define the window size to search for candidate genes (**Figure S4-S29**). The average window size around the associations was 1,370,418 bp, ranging from 51,916 bp to 25,667,882 bp. In order to compare mLHS with linkage disequilibrium (LD), we focused on the three associations detected for glutarate2oxo content as example and chose the window size defined by mLHS for comparison (**Figure 2**). The mLHS revealed a similar decaying pattern on both sides of the peak SNP on chr6 and chr11 (**Figure 2B, 2D**). However, the mLHS decayed more rapidly on the upstream side of the peak SNP on chr10 (**Figure 2C**). Though the LD near the association on chr11 was larger than on chr6, the LD pattern on both sides of the peak SNP was similar for both associations (**Figure 2E, 2G**). However, given the same window sizes based on mLHS, the LD reached different R^2 threshold for the associations on chr6 and chr11, approximately 0.25 and 0.6, respectively. For the association on chr10, $R^2 = 0.25$ gave a similar window size compared to mLHS = 0.20 (**Figure 2F**).

Haplotype-based association model outperformed multi-locus association model

In order to test the efficiency of haplotypes in identifying associations between traits and markers, we performed the regional association mapping using the SNP/haplotype-based Bayes factor model in hapQTL (Xu and Guan, 2014). A total of 108 significant associations were detected for 26 traits (**Table S11**). In order to validate the benefits of haplotypes in identifying associations, we also compared haplotype-based association model in hapQTL with multi-locus mixed model (MLMM) (Segura et al., 2012) as well as single marker mixed model in EMMAX (Kang et al., 2010). Taking the population structure and kinship as cofactors, EMMAX detected a total of 8 significant associations for 6 traits (**Table S12**). Among these, citrate and malate were both significantly associated at Chr06:44996740, which corresponded to the *Al-Activated Malate Transporter 9* (*Sl-ALMT9*) (Sauvage et al., 2014; Ye et al., 2017) and were used as positive control our the approach.

In addition to previously analyzed metabolites (Sauvage et al., 2014), we further analyzed fruit weight (fw) and detected a total of 9 significant associations for fw (**Table**

S13). Among all associations detected in hapQTL, 39 (36.1% of all associations) were detected in both the haplotype- and SNP-based Bayes models (**Figure 3A**). Haplotype-based Bayes approach outperformed SNP-based Bayes approach in terms of the number of significant associations, with 97 compared to 50, respectively (**Figure 3A**). We then compared the number of significant associations between these three association models. The largest number of significant associations was detected in hapQTL, with a total of 108 significant associations, followed by MLMM, with a total of 53 significant associations (**Figure 3B**). Among these, there were four significant loci that were detected by all three association methods on three traits, including malate (Chr02:22214295 and Chr06:44919354), citrate (Chr06:44996740) and aspartate (Chr04:60724790). In addition, a total of 30 loci were co-detected between MLMM and hapQTL. These results showed that when the number of SNPs was limited (~6000), single-marker-based association model (EMMAX) had the lowest performance, compared to multi-marker (MLMM) and haplotype/SNP-based Bayes model (hapQTL).

Associations of fruit weight and diverse metabolites were located within selective sweeps

In order to validate whether fruit weight and metabolite associations were in regions under selection, we checked the overlap between haplotype-based associations and selective sweeps, including PSS, DS and IS. Among all the associations detected, 77 (71.3% of all associations) were located within any selective sweeps type (**Figure 3C**). For the 108 associations identified in hapQTL, 22, 20 and 33 were located within PSS, DS and IS, respectively (**Table S11**). Among these, 7 were overlapped between PSS and IS, while no overlaps were observed between PSS and DS (**Table S11**).We also listed the candidate genes identified using hapQTL with previously identified domestication and improvement sweeps using another population (Lin et al., 2014). In details, 19 of these associations were within the domestication sweeps and 15 of them were within the improvement sweeps. Apart from this, there were 6 positive selective sweeps that overlapped with improvement sweeps (**Table S11**). In addition, among all the associations detected, 19 of them were detected as significant cis-eQTLs in a previous study (Zhu et al., 2018) and 10 of these cis-eQTLs were within either a positive selective sweeps or domestication/improvement sweeps (Lin et al., 2014).

The eight significant associations detected in EMMAX were located either in DS or IS. However, none of them were located within the 24 PSS (**Table S12**). For the 53

significant associations detected in MLMM, 40 were located within DS or IS (**Table S13**). For those associations, 25 were located within IS. Among these associations, only two were within PSS and the first one was for fruit weight (Chr04:59,964,407). This association was also located within DS042-DS043 and IS048-IS049, and could mainly be due to the large LD in the nearby region, which covered a region of approximately 5 Mb. The other association located within PSS was for Erythritol (Chr02:41,981,476), which was also overlapping with IS019.

Fruit weight was notably improved by allele fixation

Fruit weight is one of the most important traits selected during the long-term domestication and improvement processes of tomato breeding. For fruit weight, hapQTL identified 23 associations, while MLMM identified 9 associations and no significant associations were identified using EMMAX (**Figure 4A-C**). Among the 23 associations detected with hapQTL, 16 overlapped with selective sweeps (11 of which were located within improvement sweeps). In particular, four associations were located within positive selective sweeps (PSS11, PSS12, PSS16 and PSS21) (**Figure 4A-D**). In particular, the most significant association detected using MLMM (**Figure 4B**), which was also significant in hapQTL (**Figure 4C**), was located within the strongest positive selective sweep PSS16 identified using iHS (**Figure 4D**). In addition, the association on chr5 identified in hapQTL was within the second strongest positive selective sweep PSS12.

Together with another association on chr6, which was located between DS066 and IS061, we took these three associations as examples for further illustration of the results. For these three loci, the extended haplotype homozygosity revealed that haplotypes carrying the allele A extended differently than those carrying alternative alleles, especially the strongest association on chr7 (**Figure 4E**). In addition, high significant differences of fruit weight were observed between allele A and B, as well as between the three main groups of accessions (**Figure 4F**). We then evaluated their combined effects on fruit weight. We found that the wild species group (PIM) was dominated with one allele combination, while the large fruit cultivars were dominated with the other allele combination (**Figure 4G**). Highly significant differences of fruit weight were observed between different allele combinations (**Figure 4H**).

Multiple associations for metabolomic traits experienced positive selective or domestication/improvement sweeps

Among the haplotype-based associations identified for metabolite contents, 61 of which were located within the positive selective sweeps (**Figure 5A-D**). For example, apart from the two associations on chr2 and chr6 previously identified for malate content, we identified a new association on chr3, which corresponded to a candidate gene (the most promising candidate gene near the peak SNP) annotated as a *UDP-glucose dehydrogenase* (Solyc03g115380) and was located within PSS10. This gene was previously reported as carrying a major cis-eQTL (Zhu et al., 2018). The association identified for fructose content on chr5 was located on a candidate gene annotated as *ATP synthase F1 delta subunit* (Solyc05g050500) and was located within PSS13. This gene was also previously detected as a cis-eQTL (Zhu et al., 2018).

The association for proline content on the chr9, was located in a region with two candidate genes that were annotated as *NADH dehydrogenase* (Solyc09g064450) and *aromatic L-amino acid decarboxylase* (Solyc09g064430). This region was located within PSS21. The candidate genes for the three other associations for this trait detected on chr2:34,220,988 (*2-oxoglutarate-dependent dioxygenase*, Solyc02g062500), chr5: 1664103 (*bifunctional N-succinyldiaminopimelate-aminotransferase/acetylornithine transaminase protein*, Solyc05g007060) and chr8: 928474 (*UDP-glucose salicylic acid glucosyltransferase*, Solyc08g006330) were all reported as cis-eQTLs (Zhu et al., 2018). The candidate gene *UDP-glucose salicylic acid glucosyltransferase* was located within PSS18 (**Table S11**).

Haplotypes carrying both alleles extended differently around the peak association of proline (**Figure 5E-G**). For the association of malate, highly significant difference of malate content was observed between both alleles.

Haplotype based-QTL identified two loci co-associated with fructose, glucose and sucrose

Among all the associations identified using hapQTL (both in haplotype- and SNP-based Bayes model), we found two associations on chr2 and chr6 that were significantly co-associated with fructose, glucose and sucrose (**Figure 6A-C**). The mLHS patterns on chr2 and chr6 decayed within a window size of ~350 kb and ~560 kb, respectively (**Figure 6D,**

6E). The association on chr2 was located within PSS07, which was reported as an improvement sweep (IS030) (Lin et al., 2014). The peak SNP was located within the same haplotype. However, the haplotype length was quite different: it was largest in CER tomato, followed by BIG tomato and was shortest in PIM tomato (**Figure 6F**). In contrast, the peak SNP of the association on chr6 was located between two haplotypes and did not overlap with any selective sweeps (**Figure 6G**).

For the association on chr2, 45 candidate genes were described within this region . Among them, the most promising candidate gene was *wuschel* gene (Solyc02g083950), which was included in the same haplotype. For the association on chr6, within this region, there were 77 candidate genes. Among these, the most promising candidate gene was a *Solute carrier family facilitated glucose transporter* gene (Solyc06g066600), which had a particular strong expression level in fruits compared to other tissues (**Figure S30**). For both loci, significant difference between alleles and sugar contents were observed for all three sugars (**Figure S31, S32**) as well as their total content (**Figure 6H, 6I**). We also compared the sugar content between the three groups and found significant difference in the content of fructose, glucose, sucrose (**Figure S31, S32**) and their total content between genetic groups (**Figure 6H, 6I**).

Discussion

Benefits of haplotypes in identifying new associations

Compared to whole genome sequencing (WGS), SNP array is still a reliable, highly accurate and relatively cheap technology for GWAS, especially in very large sample sizes (Tam et al., 2019). However, when using SNP arrays, there might be many large genomic gaps in the genome and the linkage disequilibrium (LD) in different regions might also greatly differ (Sim et al., 2012; Víquez-Zamora et al., 2013; Zhao et al., 2019). Also, SNP arrays are limited in identifying ultra-rare mutations, epistasis, causal variants and missing heritability (Tam et al., 2019) due to ascertainment bias.

Though different association models are available (Gupta et al., 2019), such as EMMAX (Kang et al., 2010) and MLMM (Segura et al., 2012), haplotype association mapping takes into account not only allelic heterogeneity, but also possible statistical interactions among markers (epistasis), which is more powerful than single marker and

multiple marker analysis (Guan and Stephens, 2011; Xu and Guan, 2014). In this study, we identified more associations using haplotype-based Bayes model, compared to MLMM and EMMAX, which demonstrated the potential benefit of haplotypes in identifying new associations. In human genetics, the threshold of hapQTL is usually set at 10^{-6}, which was comparable to the genome-wide threshold (10^{-8}) in a typical human GWAS study (Xu and Guan, 2014). We thus thought the suggestive threshold based on the effective number of SNPs was appropriate for comparing the number of significant associations. In addition, the overlap between the most significant associations of MLMM, EMMAX and hapQTL also provided additional support for this threshold.

Marker local haplotype sharing provided an alternative to linkage disequilibrium for interval definition

One crucial step in association study is trying to find the promising candidate genes for the targeted phenotypes for either validate the candidate genes (i.e. functional study) or develop molecular markers for breeding purposes. In tomato, linkage disequilibrium (LD) was frequently adopted to choose the window size to search for candidate genes at a given threshold, such as $R^2 > 0.3$ (Albert et al, 2016), $R^2 > 0.5$ (Zhao et al., 2019), $R^2 > 0.7$ (Bauchet *et al.*, 2017), or $R^2 > 0.8$ (Tieman et al., 2017). Even within the window size at a high threshold, the LD between the peak SNP and close SNPs does not decay gradually as many SNPs in strong and weak LD could appear in the same region (Zhao et al., 2019), which makes it difficult to choose the optimal threshold to look for candidate genes. In contrast, mLHS between nearby SNPs and the peak SNP decreased gradually on both sides (Xu and Guan, 2014). Our results showed that at the same mLHS threshold, where dramatic decreases of mLHS were observed for the majority of associations, the corresponding R^2 based on LD varied in different associations (**Figure 2**). Thus, compared to the wave patterns of LD, mLHS provided a good alternative, although conservative, to choose the window size for screening for candidate genes.

Metabolite composition and fruit weight were improved by mutation fixation

During the tomato breeding history, flavor has not been the priority compared to yield, disease resistance and postharvest shelf life (Klee and Tieman, 2013; Klee and Tieman, 2018). However, due to reduced genetic diversity and large LD, selection of the main breeding

targets during tomato domestication and improvement might leave long-term direct or indirect effects on diverse flavor-related metabolites and volatiles. In this study, only 24 positive selective sweeps were identified, which together accounted for 8.36% of the tomato genome size. In contrast, though more domestication and improvement sweeps were identified (132 DS and 93 IS, respectively), they covered a smaller genomic size compared to PSS, with 4.88% and 5.83% of the tomato genome, respectively, which was lower than a previous study (DS and IS accounted for 8.3% and 7.0% of the genome, respectively) where 360 accessions were resequenced (Lin et al., 2014). This could be mainly explained by the limited genomic coverage of using SNP arrays and the size of the panel we analyzed.

In this study, the majority of associations for metabolites were located within selective sweeps of PSS, DS or IS. These results demonstrated that some genes with major effects might probably have undergone positive artificial selection, due to their major phenotypic effects. This is the case for instance for *Al-Activated Malate Transporter 9* (Sl-ALMT9), the major QTL responsible for variation in malate accumulation in fruit, which has been identified in different GWAS panels (Sauvage et al., 2014; Tieman et al., 2017; Bauchet et al., 2017; Ye et al., 2017; Zhao et al., 2019). This gene has been selected during domestication stage (Ye et al., 2017) by a gradual increase in the frequency of the haplotypes carrying the beneficious allele. In this study, we found it also located within a domestication sweep (DS069). In addition, other hplotype based associations with several genes controlling soluble solid content, citrate, glutarate2oxo and phenylalanine were also located within the same selective sweep. These results demonstrated that Sl-ALMT9 might not only directly regulate malate accumulation in tomato fruit, but also influence other closely related metabolites, especially citrate. *Lin5* (Solyc09g010080), a major QTL regulating soluble solid content in tomato fruit (Fridman et al., 2000), was located within both the domestication sweep DS149 and positive selective sweep PSS20, indicating strong selection experienced by this gene. The locus *fw3.2* (Solyc03g114940), a major fruit weight QTL, was located within PSS10 (Chakrabarti et al., 2013) supporting previous results that showed lower level of genetic diversity at this locus. *Uniform ripening* (u), which encodes a Golden 2-like (GLK) transcription factor, and contributed to the reduction of sugar content in modern tomato was located within PSS21 (Powell et al., 2012).

The haplotype landscape we defined revealed selective sweeps that occurred during selective events, a typical footprint of human-driven selective process. In wheat, many drought and heat stress tolerance related genes detected from GWAS were also located within

some selective sweeps (Li et al., 2019). For example, for the two associations that were co-associated with fructose, glucose and sucrose, the first association was located within one major haplotype. The length of the haplotype, which was located within PSS07 and a previously reported improvement sweep (IS030) (Lin et al., 2014), differed in BIG, CER and PIM, indicating that this haplotype has been strongly selected. In contrast, for these sugars, the peak SNP of association on chr6 was located between two major haplotypes and no selective sweeps were identified in the nearby region.

Promising candidate genes within the associations identified by hapQTL

Identifying the promising candidate genes controlling traits of interest is one of the major outputs in GWAS, notably when following a top-down approach. However, choosing the candidate genes to focus on remains challenging, especially when the LD near the peak SNP is large. However, for most of the cases, the most promising candidate gene was quite close to the peak SNP notably for the associations involving *Sl-ALMT9* and *Lin5* (Bauchet *et al.*, 2017; Sauvage *et al.*, 2014; Tieman *et al.*, 2017; Ye *et al.*, 2017; Zhao *et al.*, 2019; Zhu *et al.*, 2018).

Promising candidate genes involving important metabolites and volatiles using SNP arrays have already been identified (Bauchet *et al.*, 2017; Sauvage *et al.*, 2014; Zhao *et al.*, 2019). But all these studies, showed limits (ie. sampling size, marker density, part of the variance explained). In this study, we identified additional candidate genes. We identified a glucose transporter (Solyc06g066600) associated with sugar contents. Two other sugar transporters (Solyc08g081090 and Solyc11g062360) were associated with Soluble Solid Content. Another association for this trait was found with a gene corresponding to the *phosphoenolpyruvate carboxylase* (Solyc04g006970) gene which is highly expressed in fruit and shows variable expression in fruit. Furthermore, a candidate gene for fructose on chr5 was annotated as *ATP synthase F1 delta subunit* (Solyc05g050500). This gene was located within PSS13 and was also previously detected as carrying a cis-eQTL suggesting that its expression is regulated by a polymorphism in or close to the gene (Zhu et al., 2018). Among the associations identified for proline, one candidate on chr2 was a *proline dehydrogenase* (Solyc02g089620), which was directly involved in the dehydrogenase of proline. The candidate gene on chr3 was an *amino acid transporter* (Solyc03g117350), whose function is

also close to the trait. Fine mapping of the candidate regions and further functional validation is needed to definitively validate these candidate genes (Gupta et al., 2019).

Conclusion

Haplotype blocks are the results of demographic history of tomato through its domestication and breeding stages. Then selecting the optimal haplotype blocks carrying the positive alleles could provide new opportunities in accelerating tomato breeding. We identified a few novel candidate genes. Their functional validation will provide new genetic and evolutionary insights into tomato quality.

Experimental procedures

Materials

The studied panel consists of 163 tomato accessions derived from a core collection previously described (Xu et al., 2013; Sauvage et al, 2014). Briefly, among these, there were 116 *S.* l. *cerasiforme* accessions (CER, cherry tomato), 31 *S. lycopersicum* (BIG, large-fruit tomato) and 16 *S. pimpinellifolium* (PIM, the closest wild species). Plants were grown in a plastic greenhouse during summers of 2007 and 2008, in Avignon, France. Pericarp tissues from five fruits at the ripe stages were collected and stored at -80℃ before metabolic profiling. Genomic DNA was isolated from 100 mg frozen leaves (see Sauvage et al., 2014 for additional details).

Genotyping and quality control

All accessions were genotyped with the SOLCAP SNP array (Hamilton et al., 2012; Sim et al., 2012). SNPs with genotyping call rate lower than 90% were removed. The remaining SNPs were then filtered with minor allele frequency (MAF, $0.037 < MAF < 0.45$), generating a total of 5,995 high quality SNPs, as explained in Sauvage et al. (2014).

Haplotype block estimation

We first estimated the haplotype blocks within all accessions using Plink (Purcell et al., 2007) following the default procedure in Haploview. We then estimated the haplotype blocks within each group (BIG, CER and PIM), following the same parameters. The

graphical representation of the Genome-wide distribution of haplotypes was generated using ShinyCircos (Yu et al., 2018).

Identification of positive selective sweeps using iHS

We used integrated haplotype score (iHS) as implemented in rehh 2.0 R package (Gautier et al., 2017) to identify positive selective sweeps. The genotypic data was first phased using SHAPEIT v2 (Delaneau et al., 2014) with default settings (the number of effective population size was set at 2000) (Zhao et al., 2019). For those SNPs passing the significant threshold (-log10 [1-2|Φ_{iHS}-0.5|] > 2), we calculated the linkage disequilibrium (r^2) around the peak SNP with the following parameters: --ld --ld-window-kb 100000 --ld-window 1000 --r2 --ld-window-r2 0.5 --maf 0.037 in PLINK 1.9 (https://www.cog-genomics.org/plink2). We used r^2 = 0.5 as the threshold to group those SNPs passing the significant iHS threshold as one selective sweep (sweeps closer 100 kb were also combined as one).

Identification of domestication and improvement sweeps

In order to check whether the positive selective sweeps were caused during domestication or improvement stages, two key stages during tomato evolution, we calculated the nucleotide diversity (π) between three subgroup using a 100 kb window with a step size of 10 kb in PIM, CER and BIG separately, using vcftools (Danecek et al., 2011). We then scanned the ratios of genetic diversity between PIM and CER (PIM/CER) for domestication sweeps and between CER and BIG (CER/BIG) for improvement sweeps. We selected windows with the top 5% of ratios as the domestication and improvement sweeps, respectively (3.43 and 6.16 for domestication and improvement, respectively). All sweeps with the windows closer than 100 kb were merged into a single selected region. Comparing with iHS, we could assess which domestication and improvement sweeps were positive selective sweeps.

Phenotyping

Briefly, ten fruits per accession were measured for fruit weight. The metabolite profiles were measured as detailed in Sauvage et al. (2014), including sugars, sugar alcohols, organic acids and amino acids. Only phenotypes with a high correlation over two years were retained for further analyses (Xu et al., 2013; Sauvage et al., 2014).

Single marker genomewide association

We performed association analysis using the efficient mixed-model association expedited software (EMMAX) (Kang et al., 2010), which is a single-marker based association model. The BN kinship matrix and the first five discriminant axes of principal components (DAPC, six in total) were added as cofactors. The BN kinship matrix was calculated in EMMAX with the default command: emmax-kin -v -h -d 10. The optimal number of clusters was determined by Bayesian Information Criteria (BIC) with minor increase or decrease (K = 6). All PCs and all discriminant functions were retained to find the optimal number of clusters (Tieman et al., 2017). Genome-wide significant threshold was determined in Genetic type 1 Error Calculator (GEC) (Li et al., 2012). The genome-wide suggestive and significant threshold were set to 4.10×10^{-4} and 2.05×10^{-5}, respectively.

Regional haplotype-based association (hapQTL)

We performed regional haplotype-based association using hapQTL (Xu and Guan, 2014). The first five discriminant principal components were added as covariates, as required in EMMAX. The number of EM runs was set at 10 to avoid uncertainty in LD inference. The number of upper clusters was set at 3. We defined the genome-wide significant threshold in hapQTL lower than the threshold of typical GWAS as suggested in Xu and Guan (2014). So we used 3.387 ($-\log_{10}$(the suggestive p-value 4.10×10^{-4})) as the Bayes factor threshold, which was comparable with the significant threshold in EMMAX. Significant associations in the same strong LD block or haplotype regions were treated as a unique association and the peak SNP was retained.

Marker local haplotype sharing (mLHS) was calculated based on 10 independent EM runs with the same parameter, which could be used to define the LD block around the peak SNP (Xu and Guan, 2014). We found that the threshold of 0.25 (2.5/10) was too stringent for most loci that they could not cover most of the top ancestral haplotypes. So, we adopted 0.20 as the LD block threshold. Gene annotations were done according to the tomato genome annotation version 2.40. For those SNPs passing the significant threshold, we then calculated the LD (r^2 = 0.5) for each marker to group those closely linked SNPs as one selective sweep. Those sweeps within 100 kb were also grouped as one single sweep.

Genome-wide association via MLMM

In order to compare the efficiency of hapQTL in identifying associations with other models, we then compared our results with association analysis using multi-locus mixed-model (MLMM) (Segura et al., 2012) as described in Sauvage et al., (2014). Population structure was estimated in Structure v2.3.3 (Pritchard et al., 2000) and kinship matrix estimated in SPAGeDi (Hardy and Vekemans, 2002). We then compared the number of significantly associated regions obtained by the three methods.

Candidate gene expression patterns and visualization of local haplotype structure

In order to provide supporting evidence whether some candidate genes were functionally related to the analyzed phenotypic traits, we screened the genome annotation (version 2.40) and for those candidate genes of particular interests, we linked their annotation to their expression level in tomato fruits at different developing stages. Data were retrieved from the Tomato Expression Atlas database (http://tea.solgenomics.net/expression_viewer/input) (Fernandez-Pozo et al., 2017; Shinozaki et al., 2018). Candidate gene expression levels at different tissues and developing stages were also checked using ePlant (Waese et al., 2017). In order to see whether there was major difference between the ancestral and derived haplotypes (in order to avoid this uncertainty, we used allele A and B instead of ancestral and derived allele, as the true ancestral and derived alleles of the SNPs are unknown). Visualization of local haplotype structure around a peak SNP near the candidate gene was performed using the bifurcation.diagram() function in rehh 2.0 R package (Gautier et al., 2017).

Supplementary files

Supplementary Tables and Figures are provided in Appendix 1.

Acknowledgements

Jiantao Zhao was founded by Chinese Scholarship Council (CSC) scholarship (No. 201606300007). We thank Yongtao Guan from Baylor College of Medicine, USA for suggestions and his detailed explanations on the threshold of hapQTL and the example data and R codes he provided to calculate mLHS. We thank Anurag Daware and Akhilesh Tyagi from National Institute of Plant Genome Research, New Delhi, India for sharing the R codes and example data in rice to calculate mLHS. We are also grateful to Jacques Lagnel from

UR1052 GAFL INRA, France for the bioinformatic assistance he provided to perform the analyses described in the present manuscript.

Conflict of interest

The authors declare that they have no conflict of interests.

Author contributions

Study design/conception: M.C., J-T.Z., C.S.; supervision: C.S, M.C.; data collection and analysis: J-T.Z.; data interpretation: J-T.Z., F.B., C.S., M.C.; first draft of the manuscript: J-T.Z.; critical revisions of the manuscript: all co-authors.

References

Bauchet, G., and Causse, M. (2012). Genetic diversity in tomato (*Solanum lycopersicum*) and its wild relatives. In *Genetic Diversity in Plants*, p. InTech.

Bauchet, G., Grenier, S., Samson, N., Segura, V., Kende, A., Beekwilder, J., Cankar, K., Gallois, J.-L., Gricourt, J., Bonnet, J., et al. (2017). Identification of major loci and genomic regions controlling acid and volatile content in tomato fruit: implications for flavor improvement. *New Phytol.* **215**:624–641.

Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'Ang, L. Y., Huang, W., Liu, B., Shen, Y., Tam, P. K. H., et al. (2003). The international HapMap project. *Nature* **426**:789–796.

Calus, M. P. L., Meuwissen, T. H. E., Roos, A. P. W. de, and Veerkamp, R. F. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics* **178**:553–561.

Chakrabarti, M., Zhang, N., Sauvage, C., Muños, S., Blanca, J., Cañizares, J., Diez, M. J., Schneider, R., Mazourek, M., McClead, J., et al. (2013). A cytochrome P450 regulates a domestication trait in cultivated tomato. *Proc. Natl. Acad. Sci. U. S. A.* **110**:17125–30.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* **22**:961–975.

Cuyabano, B. C., Su, G., and Lund, M. S. (2014). Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. *BMC Genomics* **15**.

Cuyabano, B. C. D., Su, G., Rosa, G. J. M., Lund, M. S., and Gianola, D. (2015a). Bootstrap study of genome-enabled prediction reliabilities using haplotype blocks across Nordic Red cattle breeds. *J. Dairy Sci.* **98**:7351–7363.

Cuyabano, B. C., Su, G., and Lund, M. S. (2015b). Selection of haplotype variables from a high-density marker map for genomic prediction. *Genet. Sel. Evol.* **47**:61.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nat. Genet.* **29**:229–232.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., et al. (2011). The variant call format and VCFtools. *Bioinformatics* **27**:2156–2158.

Delaneau, O., Marchini, J., Consortium, T. 1000 G. P., McVean, G. A., Donnelly, P., Lunter, G., Marchini, J. L., Myers, S., Gupta-Hinch, A., Iqbal, Z., et al. (2014). Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat. Commun.* **5**:3934.

Desta, Z. A., and Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci.* **19**:592–601.

Duangjit, J., Causse, M., and Sauvage, C. (2016). Efficiency of genomic selection for tomato fruit quality.

Mol. Breed. **36**:36:29.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome J.* **4**:250.

Farashi, S., Kryza, T., Clements, J., and Batra, J. (2019). Post-GWAS in prostate cancer: from genetic association to biological contribution. *Nat. Rev. Cancer* **19**:46–59.

Fernandez-Pozo, N., Zheng, Y., Snyder, S. I., Nicolas, P., Shinozaki, Y., Fei, Z., Catala, C., Giovannoni, J. J., Rose, J. K. C., and Mueller, L. A. (2017). The tomato expression atlas. *Bioinformatics* **33**:2397–2398.

Fridman, E., Pleban, T., and Zamir, D. (2000). A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. *Proc. Natl. Acad. Sci. U. S. A.* **97**:4718–23.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002). The structure of haplotype blocks in the human genome. *Science (80-.).* **296**:2225–2229.

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., Burzynski-Chang, E. A., Fish, T. L., Stromberg, K. A., Sacks, G. L., et al. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* **51**:1044–1051.

Gautier, M., Klassmann, A., and Vitalis, R. (2017). rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Mol. Ecol. Resour.* **17**:78–90.

Gawenda, I., Thorwarth, P., Günther, T., Ordon, F., and Schmid, K. J. (2015). Genome-wide association studies in elite varieties of German winter barley using single-marker and haplotype-based methods. *Plant Breed.* **134**:28–39.

Guan, Y., and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5**:1780–1815.

Gupta, P. K., Kulwal, P. L., and Jaiswal, V. (2019). Association mapping in plants in the post-GWAS genomics era. In *Advances in Genetics*, pp. 75–154.

Hamilton, J. P., Sim, S.-C., Stoffel, K., Van Deynze, A., Buell, C. R., and Francis, D. M. (2012). Single nucleotide polymorphism discovery in cultivated tomato via sequencing by synthesis. *Plant Genome J.* **5**:17.

Hardy, O. J., and Vekemans, X. (2002). SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol. Ecol. Notes* **2**:618–620.

Hermisson, J., and Pennings, P. S. (2017). Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods Ecol. Evol.* **8**:700–716.

Hess, M., Druet, T., Hess, A., and Garrick, D. (2017). Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. *Genet. Sel. Evol.* **49**:54.

Hickey, J. M., Chiurugwi, T., Mackay, I., and Powell, W. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nat. Genet.* **49**:1297–1303.

Hickey, L. T., N. Hafeez, A., Robinson, H., Jackson, S. A., Leal-Bertioli, S. C. M., Tester, M., Gao, C., Godwin, I. D., Hayes, B. J., and Wulff, B. B. H. (2019). Breeding crops to feed 10 billion. *Nat. Biotechnol.* **37**:744–754.

Jiang, Y., Schmidt, R. H., and Reif, J. C. (2018). Haplotype-based genome-wide prediction models exploit local epistatic interactions among markers. *G3 Genes|Genomes|Genetics* **8**:g3.300548.2017.

Johnson, G. C. L., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., et al. (2001). Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29**:233–237.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**:348–354.

Karimi, Z., Sargolzaei, M., Robinson, J. A. B., and Schenkel, F. S. (2018). Assessing haplotype-based models for genomic evaluation in holstein cattle. *Can. J. Anim. Sci.* **98**:750–759.

Khatkar, M. S., Zenger, K. R., Hobbs, M., Hawken, R. J., Cavanagh, J. A. L., Barris, W., McClintock, A. E., McClintock, S., Thomson, P. C., Tier, B., et al. (2007). A Primary Assembly of a Bovine Haplotype Block Map Based on a 15,036-Single-Nucleotide Polymorphism Panel Genotyped in Holstein–Friesian Cattle. *Genetics* **176**:763–772.

Klee, H. J., and Tieman, D. M. (2013). Genetic challenges of flavor improvement in tomato. *Trends Genet.* **29**:257–262.

Klee, H. J., and Tieman, D. M. (2018). The genetics of fruit flavour preferences. *Nat. Rev. Genet.* **19**:347–356.

Li, M.-X. X., Yeung, J. M. Y., Cherny, S. S., and Sham, P. C. (2012). Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* **131**:747–756.

Li, T., Yang, X., Yu, Y., Si, X., Zhai, X., Zhang, H., Dong, W., Gao, C., and Xu, C. (2018). Domestication

of wild tomato is accelerated by genome editing. *Nat. Biotechnol.* **36**:1160–1163.

Li, L., Mao, X., Wang, J., Chang, X., Reynolds, M., and Jing, R. (2019). Genetic dissection of drought and heat-responsive agronomic traits in wheat. *Plant. Cell Environ.* **42**:pce.13577.

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S., Wang, X., et al. (2014). Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.* **46**:1220–1226.

Liu, H. J., and Yan, J. (2019). Crop genome-wide association study: a harvest of biological relevance. *Plant J.* **97**:8–18.

Maldonado, C., Mora, F., Scapim, C. A., and Coan, M. (2019). Genome-wide haplotype-based association analysis of key traits of plant lodging and architecture of maize identifies major determinants for leaf angle: *HAPla4*. *PLoS One* Advance Access published 2019, doi:10.1371/journal.pone.0212925.

Meuwissen, T. H. E., Odegard, J., Andersen-Ranberg, I., and Grindflek, E. (2014). On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. *Genet. Sel. Evol.* **46**.

Millet, E. J., Kruijer, W., Coupel-Ledru, A., Alvarez Prado, S., Cabrera-Bosquet, L., Lacube, S., Charcosset, A., Welcker, C., van Eeuwijk, F., and Tardieu, F. (2019). Genomic prediction of maize yield across European environmental conditions. *Nat. Genet.* **51**:952–956.

Pérez, P., de los Campos, G., and Goddard, M. E. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **198**:483–95.

Powell, A. L. T., Nguyen, C. V., Hill, T., Cheng, K. L. L., Figueroa-Balderas, R., Aktas, H., Ashrafi, H., Pons, C., Fernández-Muñoz, R., Vicente, A., et al. (2012). Uniform ripening encodes a *Golden 2-like* transcription factor regulating tomato fruit chloroplast development. *Science (80-.).* **336**:1711–1715.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**:945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**:559–575.

Rodríguez-Leal, D., Lemmon, Z. H., Man, J., Bartlett, M. E., and Lippman, Z. B. (2017). Engineering quantitative trait variation for crop improvement by genome editing. *Cell* **171**:470-480.e8.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**:832–837.

Sauvage, C., Segura, V., Bauchet, G., Stevens, R., Do, P. T., Nikoloski, Z., Fernie, A. R., and Causse, M. (2014). Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant Physiol.* **165**:1120–1132.

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* **44**:825–830.

Shinozaki, Y., Nicolas, P., Fernandez-Pozo, N., Ma, Q., Evanich, D. J., Shi, Y., Xu, Y., Zheng, Y., Snyder, S. I., Martin, L. B. B., et al. (2018). High-resolution spatiotemporal transcriptome mapping of tomato fruit development and ripening. *Nat. Commun.* **9**:364.

Sim, S.-C., Durstewitz, G., Plieske, J., Wieseke, R., Ganal, M. W., van Deynze, A., Hamilton, J. P., Buell, C. R., Causse, M., Wijeratne, S., et al. (2012). Development of a large snp genotyping array and generation of high-density genetic maps in tomato. *PLoS One* **7**.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**:467–484.

Tieman, D., Zhu, G., Resende, M. F. R., Lin, T., Nguyen, C., Bies, D., Rambla, J. L., Beltran, K. S. O., Taylor, M., Zhang, B., et al. (2017). A chemical genetic roadmap to improved tomato flavor. *Science (80-.).* **355**:391–394.

Villumsen, T. M., Janss, L., and Lund, M. S. (2009). The importance of haplotype length and heritability using genomic selection in dairy cattle. *J. Anim. Breed. Genet.* **126**:3–13.

Víquez-Zamora, M., Vosman, B., van de Geest, H., Bovy, A., Visser, R. G. F., Finkers, R., and van Heusden, A. W. (2013). Tomato breeding in the genomics era: Insights from a SNP array. *BMC Genomics* **14**:354.

Vitti, J. J., Grossman, S. R., and Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annu. Rev. Genet.* **47**:97–120.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol.* **4**:0446–0458.

Waese, J., Fan, J., Pasha, A., Yu, H., Fucile, G., Shi, R., Cumming, M., Kelley, L. A., Sternberg, M. J., Krishnakumar, V., et al. (2017). ePlant: visualizing and exploring multiple levels of data for hypothesis generation in plant biology. *Plant Cell* **29**:1806–1821.

Xu, H., and Guan, Y. (2014). Detecting local haplotype sharing and haplotype association. *Genetics* **197**:823–

838.

Xu, J., Ranc, N., Muños, S., Rolland, S., Bouchet, J.-P. P., Desplat, N., Le Paslier, M.-C. C., Liang, Y., Brunel, D., and Causse, M. (2013). Phenotypic diversity and association mapping for fruit quality traits in cultivated tomato and related species. *Theor. Appl. Genet.* **126**:567–581.

Yamamoto, E., Matsunaga, H., Onogi, A., Kajiya-Kanegae, H., Minamikawa, M., Suzuki, A., Shirasawa, K., Hirakawa, H., Nunome, T., Yamaguchi, H., et al. (2016). A simulation-based breeding design that uses whole-genome prediction in tomato. *Sci. Rep.* **6**:19454.

Yamamoto, E., Matsunaga, H., Onogi, A., Ohyama, A., Miyatake, K., Yamaguchi, H., Nunome, T., Iwata, H., and Fukuoka, H. (2017). Efficiency of genomic selection for breeding population design and phenotype prediction in tomato. *Heredity (Edinb).* **118**:202–209.

Ye, J., Wang, X., Hu, T., Zhang, F., Wang, B., Li, C., Yang, T., Li, H., Lu, Y., Giovannoni, J. J., et al. (2017). An inDel in the promoter of *Al-ACTIVATED MALATE TRANSPORTER9* selected during tomato domestication determines fruit malate contents and aluminum tolerance. *Plant Cell* **29**:2249–2268.

Yu, Y., Ouyang, Y., and Yao, W. (2018). shinyCircos: an R/Shiny application for interactive creation of Circos plot. *Bioinformatics* **34**:1229–1231.

Zhao, J., Sauvage, C., Zhao, J., Bitton, F., Bauchet, G., Liu, D., Huang, S., Tieman, D. M., Klee, H. J., and Causse, M. (2019). Meta-analysis of genome-wide association studies provides insights into genetic control of tomato flavor. *Nat. Commun.* **10**:1534.

Zhu, G., Wang, S., Huang, Z., Zhang, S., Liao, Q., Zhang, C., Lin, T., Qin, M., Peng, M., Yang, C., et al. (2018). Rewiring of the fruit metabolome in tomato breeding. *Cell* **172**:249-261.e12.

Zhu, G., Gou, J., Klee, H., and Huang, S. (2019). Next-gen approaches to flavor-related metabolism. *Annu. Rev. Plant Biol.* **70**:annurev-arplant-050718-100353.

Zsögön, A., Čermák, T., Naves, E. R., Notini, M. M., Edel, K. H., Weinl, S., Freschi, L., Voytas, D. F., Kudla, J., and Peres, L. E. P. (2018). De novo domestication of wild tomato using genome editing. *Nat. Biotechnol.* **36**:1211–1216.

Figure legends

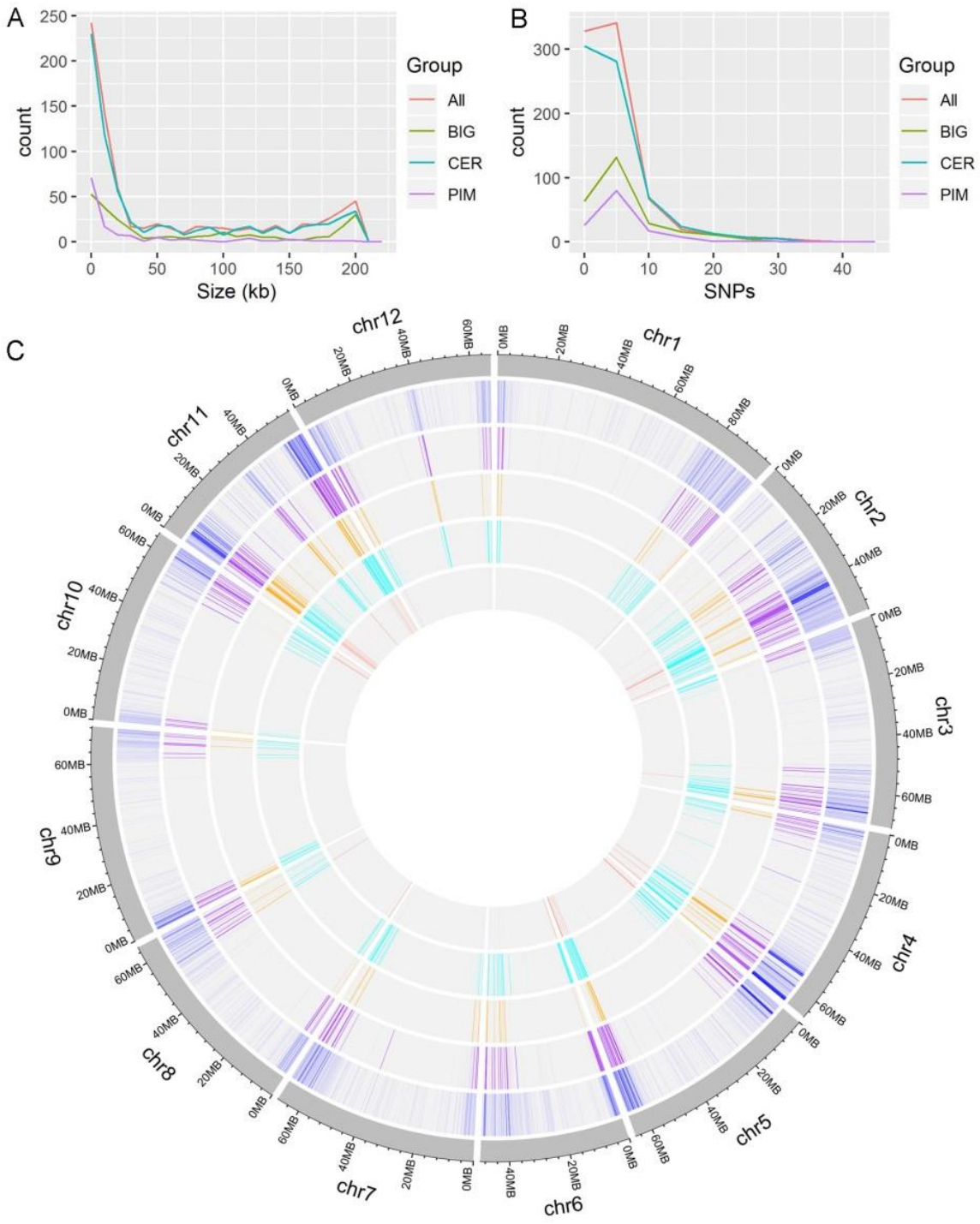


Figure 1. Haplotype block distributions within the 163 tomato accessions that were genotyped with 5995 SNPs. (**A**) Distribution patterns of the size of haplotype blocks within all accessions and subgroups. All, 163 tomato accessions; BIG, 31 large-fruit tomato accessions; CER, 116 cherry tomato accessions; PIM, 16 wild tomato species. (**B**) Distribution patterns of the number of SNPs within all accessions and subgroups. (**C**) Genome-wide distribution of SNPs and haplotype blocks. From outside to inside: distribution of 5995 SNPs; haplotype blocks within 163 tomato accessions; haplotype blocks within 31 large-fruit tomato accessions; haplotype blocks within 116 cherry tomato accessions; haplotype blocks within 16 wild tomato species.

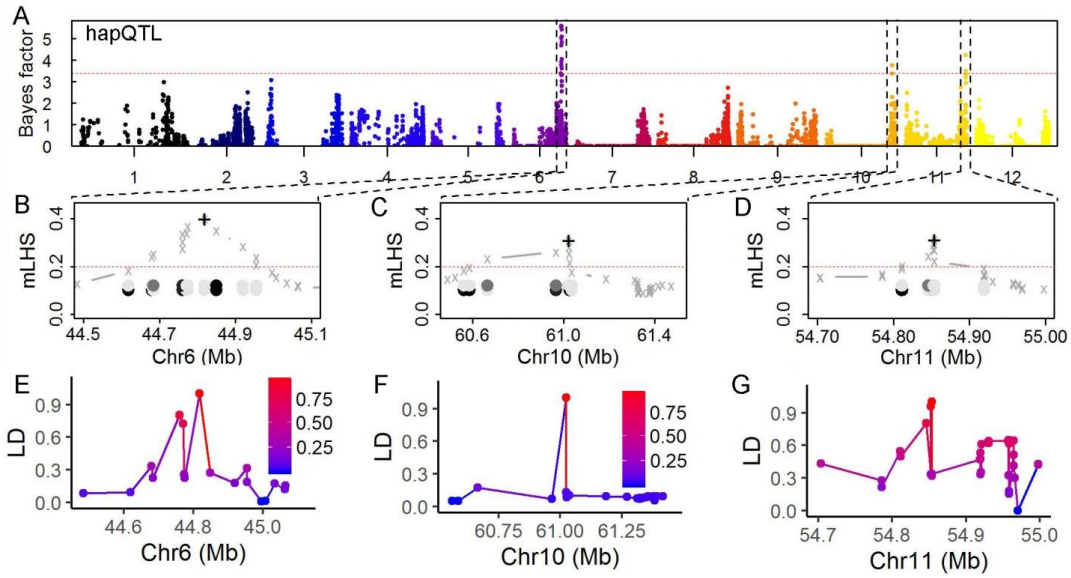


Figure 2. Comparison between marker local haplotype sharing (mLHS) and linkage disequilibrium (LD) based on the three associations detected for glutarate2oxo content. (**A**) Manhattan plot of genome-wide associations for fruit weight using haplotype- and SNP-based Bayes model (hapQTL). (**B-D**) Maker local haplotype sharing of the peak SNPs for the associations detected on chr6 (**B**), chr10 (**C**) and chr11 (**D**), respectively. The threshold was set at 0.20. (**E-G**) Linkage disequilibrium distribution patterns of the peak SNPs for the associations detected on chr6 (**E**), chr10 (**F**) and chr11 (**G**), respectively.

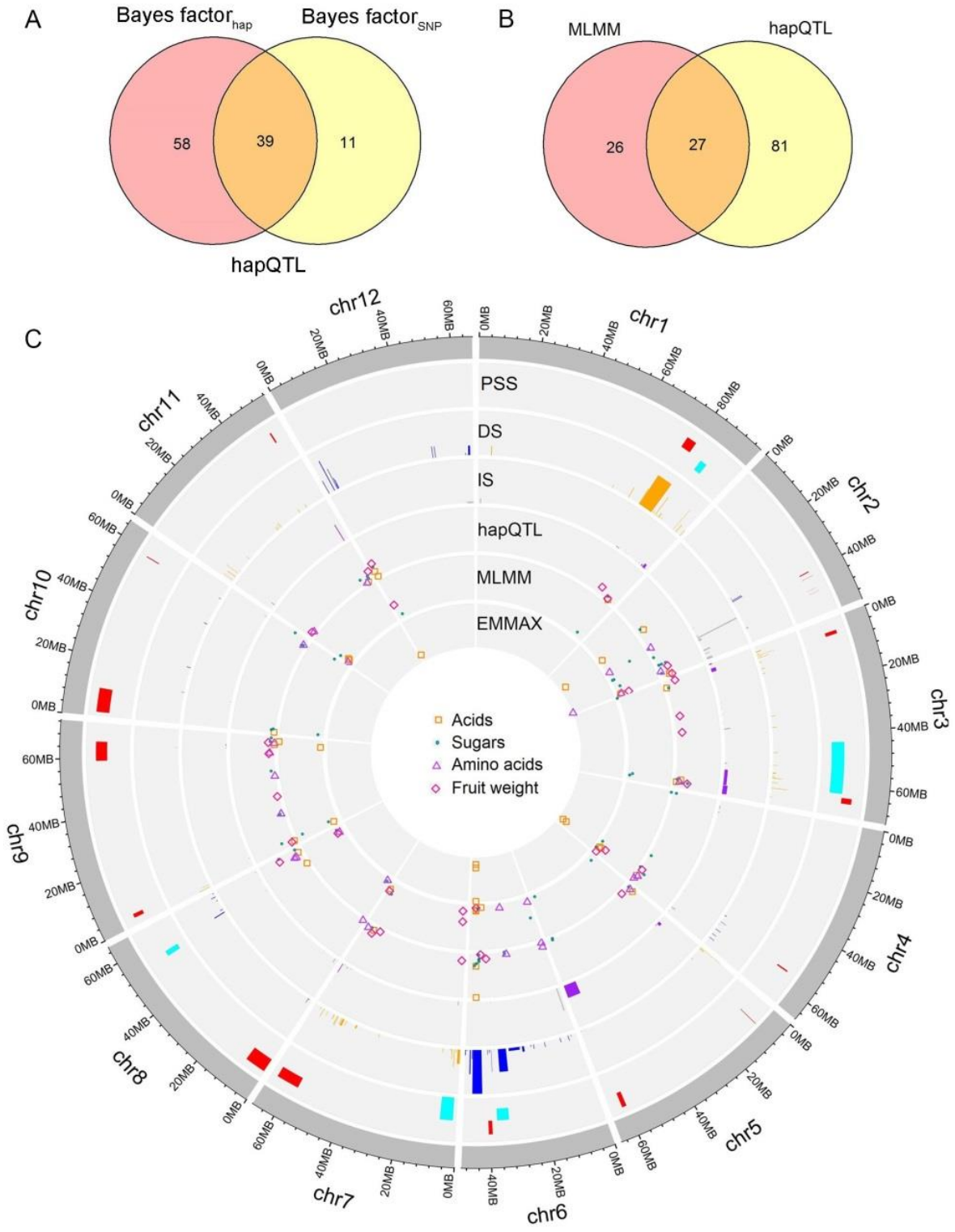


Figure 3. Association model comparison and genome-wide distributions of selective sweeps and significant associations. (**A**) Venn diagram of haplotype- and SNP-based Bayes associations using hapQTL. (**B**) Venn diagram of EMMAX, MLMM and hapQTL. EMMAX, efficient mixed-model association expedited; MLMM, multi-locus mixed-model. (**C**) Genome-wide distributions of selective sweeps and significant associations. PSS, positive selective sweeps; DS, domestication sweeps; IS, improvement sweeps. Traits were subdivided into four main groups as acids, sugars, amino acids and fruit weight and indicated by different colored shapes.

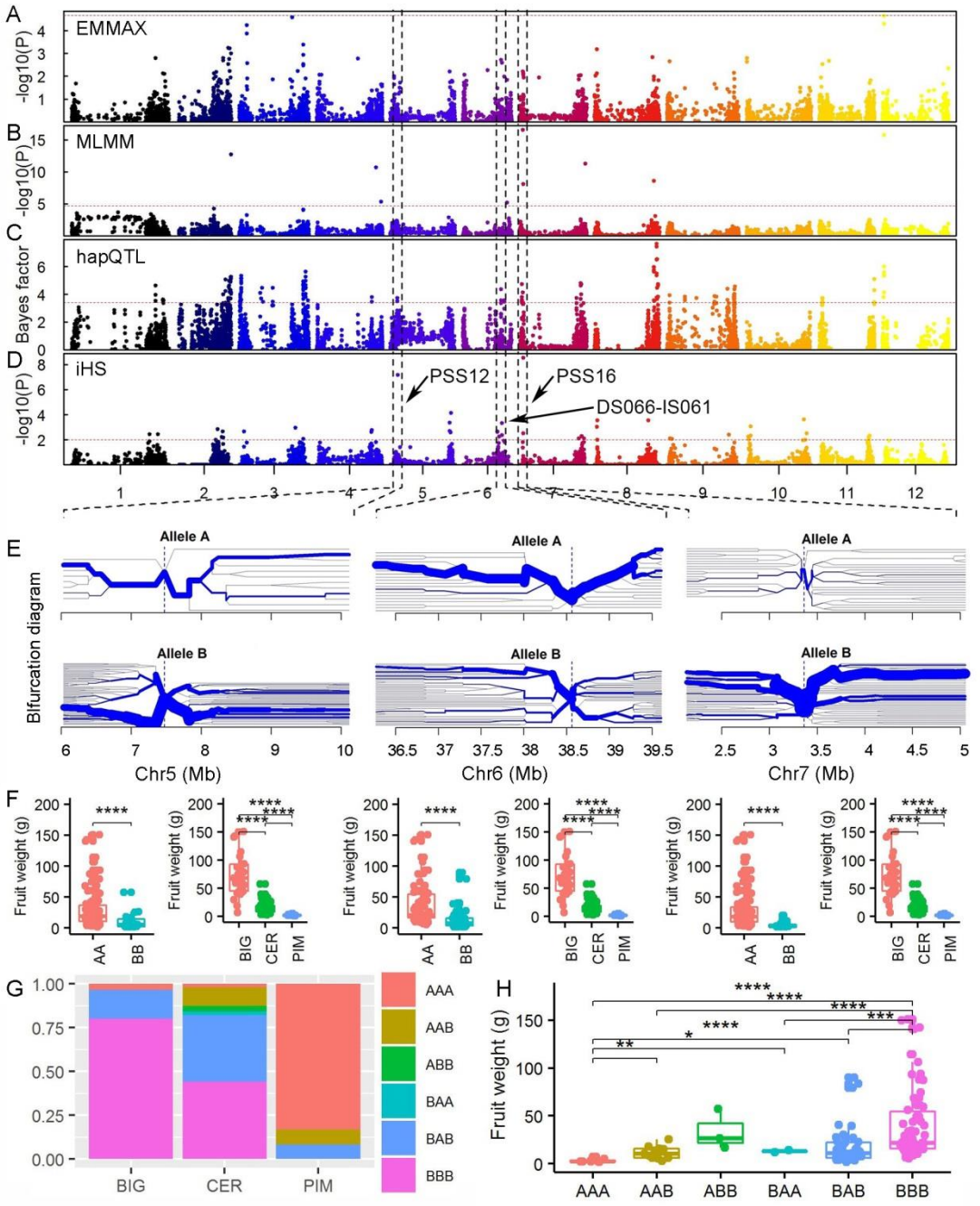


Figure 4. Identification of associations for fruit weight that were within positive selective sweeps and their impacts on fruit weight. (**A**) Manhattan plot of genome-wide associations for fruit weight using efficient mixed-model association expedited model (EMMAX). (**B**) Manhattan plot of genome-wide associations for fruit weight using multi-locus mixed model (MLMM). (**C**) Manhattan plot of genome-wide associations for fruit weight using haplotype- and SNP-based Bayes model (hapQTL). (**D**) Manhattan plot of genome-wide distribution of positive selective sweeps identified using integrated haplotype score (iHS). (**E**) Bifurcation diagram for the extended haplotypes starting from the allele A and B of the peak SNPs on chr5, chr6 and chr7, respectively. (**F**) Comparisons of the fruit weight between the allele A and B in different tomato groups of the peak SNPs on chr5, chr6 and chr7, respectively. BIG, big-fruit tomato (*S. lycopersicum*), CER, cherry tomato (*S. lycopersicum* var *cerasiforme*) and PIM, the closest wild species (*S. pimpinellifolium*). (**G**) Allele combination effects between the three groups. (**H**) Effects of different allele combinations on fruit weight. ns, not significant; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

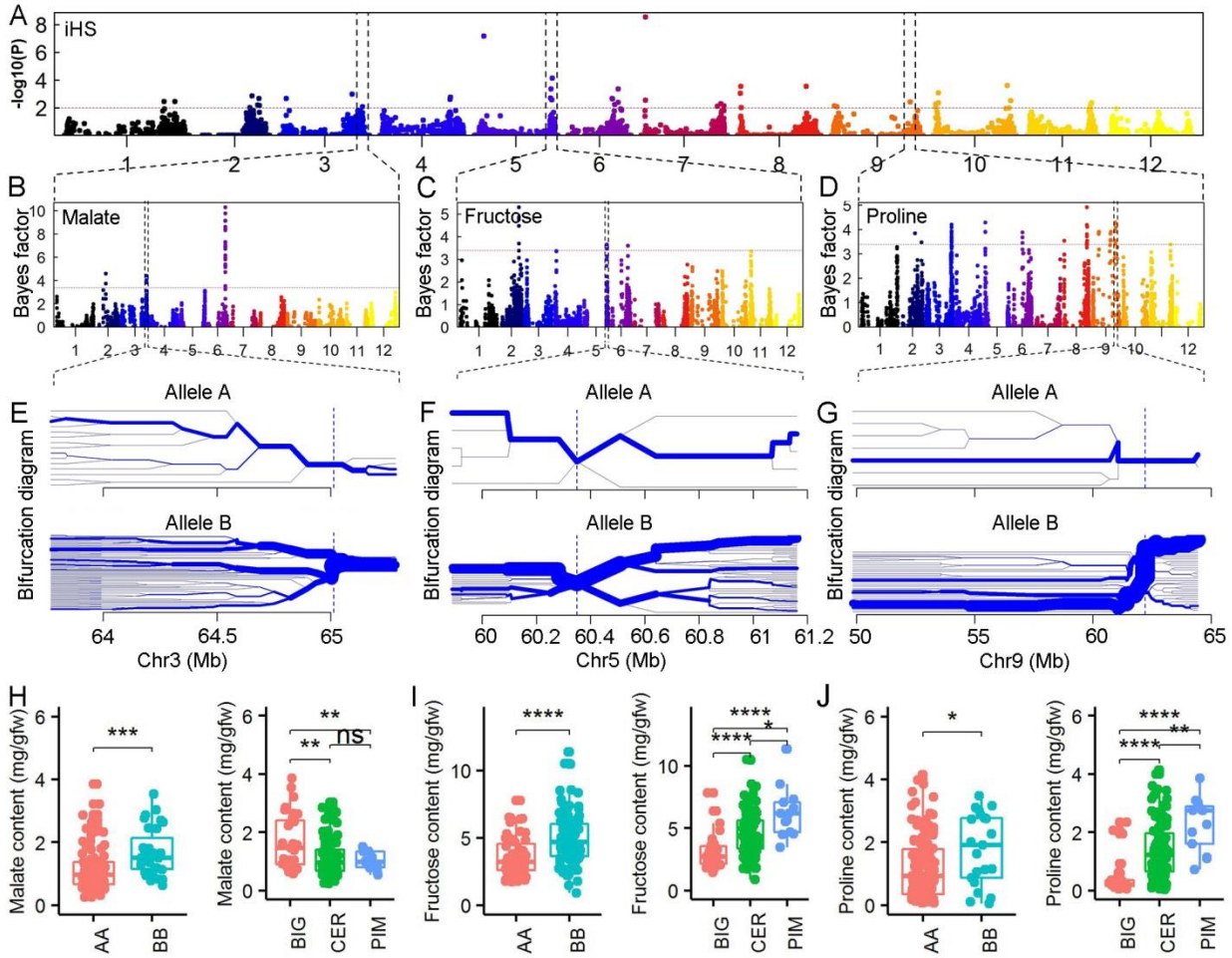


Figure 5. Identification of associations for malate, fructose and proline content that were within positive selective sweeps. (**A**) Manhattan plot of genome-wide distribution of positive selective sweeps identified using integrated haplotype score (iHS). (**B**) Manhattan plot of genome-wide associations for malate content using haplotype- and SNP-based Bayes model (hapQTL). (**C**) Manhattan plot of genome-wide associations for fructose content using haplotype- and SNP-based Bayes model (hapQTL). (**D**) Manhattan plot of genome-wide associations for proline content using haplotype- and SNP-based Bayes model (hapQTL). (**E-G**) Bifurcation diagram for the extended haplotypes starting from allele A and allele B of the peak SNPs for malate, fructose and proline, respectively. (**H-J**) Comparisons between allele A and allele B and different tomato groups of the peak SNPs for malate, fructose and proline content, respectively. BIG, big-fruit tomato (*S. lycopersicum*), CER, cherry tomato (*S. lycopersicum* var *cerasiforme*) and PIM, the closest wild species (*S. pimpinellifolium*). ns, not significant; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

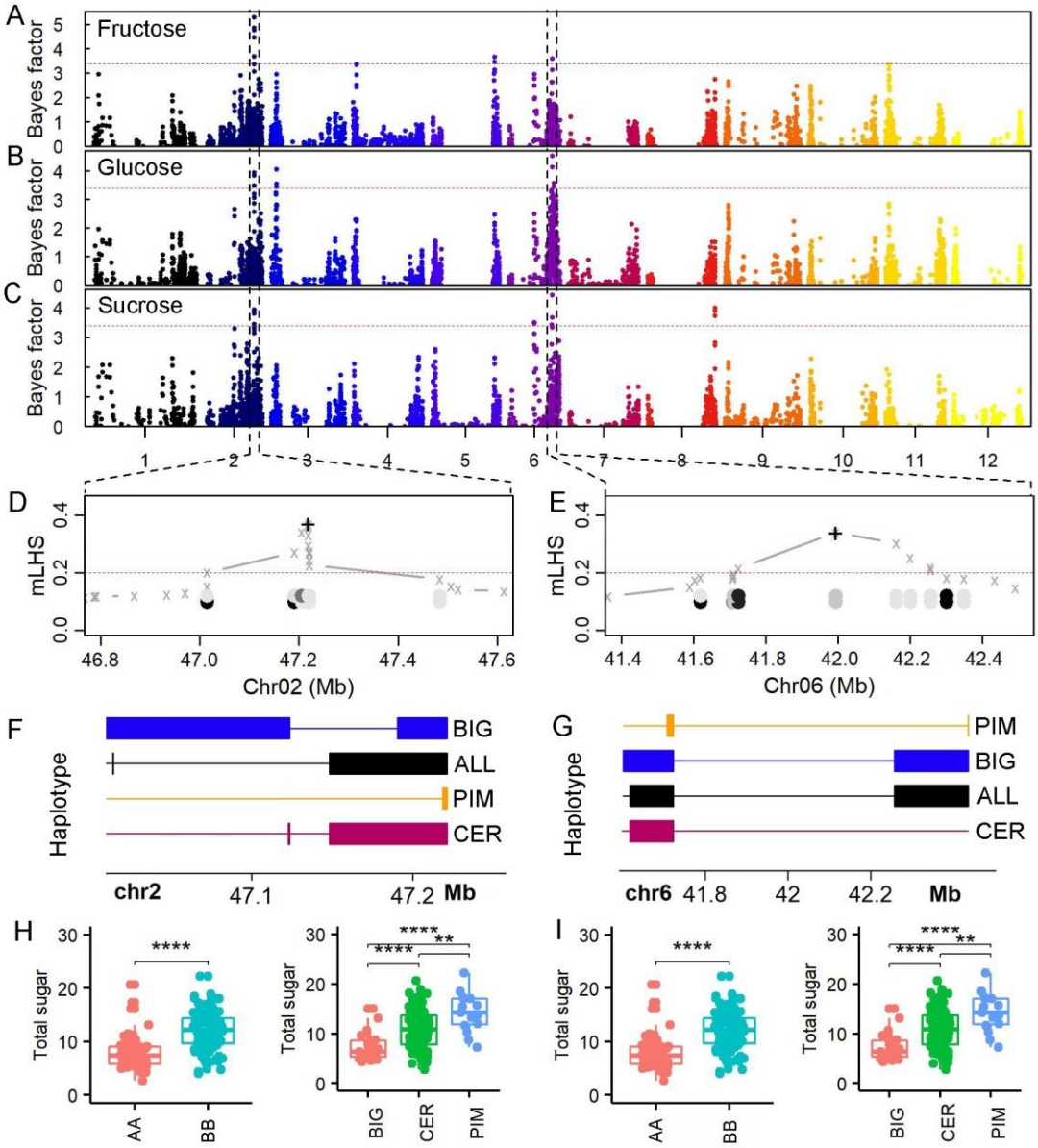


Figure 6. Identification of new candidate genes that was significantly co-associated with fructose, glucose and sucrose using hapQTL. (**A-C**) Regional Manhattan plot of the significant association for fructose (**A**), glucose (**B**) and sucrose (**C**). (**D,E**) Marker local haplotype sharing of the peak SNP that were significantly co-associated with fructose, glucose and sucrose on chr2 (**D**) and chr6 (**E**), respectively. (**F,G**) Haplotype distributions of the target regions between all and three subgroups for the associations detected on chr2 (**F**) and chr6 (**G**), respectively. (**H, I**) Comparison of the total sugar content of fructose, glucose and sucrose between allele A and B as well as three subgroups for the association detected on chr2 (**H**) and chr6 (**I**), respectively. BIG, big-fruit tomato (*S. lycopersicum*); CER, cherry tomato (*S. lycopersicum* var *cerasiforme*); PIM, the closest wild species (*S. pimpinellifolium*). ns, not significant; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

Supplementary results and general discussions

Topic 1: Haplotype-based genomic prediction

To sustain genetic gains and accelerate the breeding cycles of crops, genomic selection (GS) (Crossa et al., 2017) is progressively adopted. It has great potentials to improve selection efficiency, reduce phenotyping costs and orientate breeding schemes (Hickey et al., 2017). It is for instance possible to accurately predict the maize yield in a wide range of environmental scenarios (Millet et al., 2019). Combining GS with other breeding technologies could greatly reduce generation intervals and include the precise locations of causative mutations (Hickey et al., 2019). Single SNPs were first used for genomic prediction while haplotypes were recently adopted to increase prediction accuracy. Simulations and analyses of cattle data showed that haplotype-based genomic prediction further improves the prediction accuracy (Calus et al., 2008; Villumsen et al., 2009; Cuyabano et al., 2014; Cuyabano et al., 2015a; Hess et al., 2017; Jiang et al., 2018; Karimi et al., 2018). However, the application of genomic selection in tomato is limited to simulations (Yamamoto et al., 2016), or cross validations experiments (Duangjit et al., 2016; Yamamoto et al., 2017) without any public report of its application.

In this chapter, we have demonstrated the benefits of using haplotypes from several aspects. In order to test if haplotypes could also be used to predict the phenotypic values using genomic prediction, we also tested the potential benefits of using haplotypes in improving the genomic prediction accuracy. To do so, we first converted the haplotypes within each haplotype block to pseudo SNPs. Briefly, if the unique haplotype appears in one accession, we treated it as allele A; if not, we treated it as alternative allele B. Rare pseudo SNPs with low minor allele frequency were removed. We used the same MAF threshold (0.037) as in Sauvage et al., (2014). The remaining pseudo SNPs were then combined with the SNPs that were not located within any haplotype blocks, which was used to re-running all the models aforementioned following the same parameters. After filtering minor allele frequency lower than 0.037 (Sauvage et al., 2014), a total of 2496 pseudo-SNPs remained. They were then combined with those SNPs located outside haplotype blocks, generating a total of 4650 SNPs.

In order to compare the prediction accuracy, we selected 10 traits as examples based on their biological significance and wide range of heritability, including fruit weight (FW), brix, fructose (Fru), sucrose (Suc), ascorbic acid (ASA), malate, citrate, asparagine, proline (Pro) and lysine (Lys). We compared different genomic prediction models: genomic best linear

unbiased prediction (GBLUP) with multiple Bayes genomic prediction models, including BayesA, BayesB, BayesC, BL and BRR using BGLR package (Pérez et al., 2014). Then, to test the predictive accuracy of each tested models, we implemented a cross validation approach as following: we randomly selected 75% of the accessions as the training population (TP), trained the model and predicted the phenotypic values within the remaining accessions or test population, with 100 replications.

In overall, we found the haplotype-based prediction accuracy was increased compared to single SNP-based prediction (**Figure 1**). For fruit weight and ASA, the prediction accuracy of haplotype-based models was always higher than single SNP-based models. For those traits that are likely controlled by a few major QTLs, such as malate and citrate, the prediction of BayesB outperformed all the other models, which was further improved by using haplotypes. In summary, haplotypes are also helpful in improving the genomic prediction accuracy, though the highest prediction accuracy varied, depending on the traits and models used.

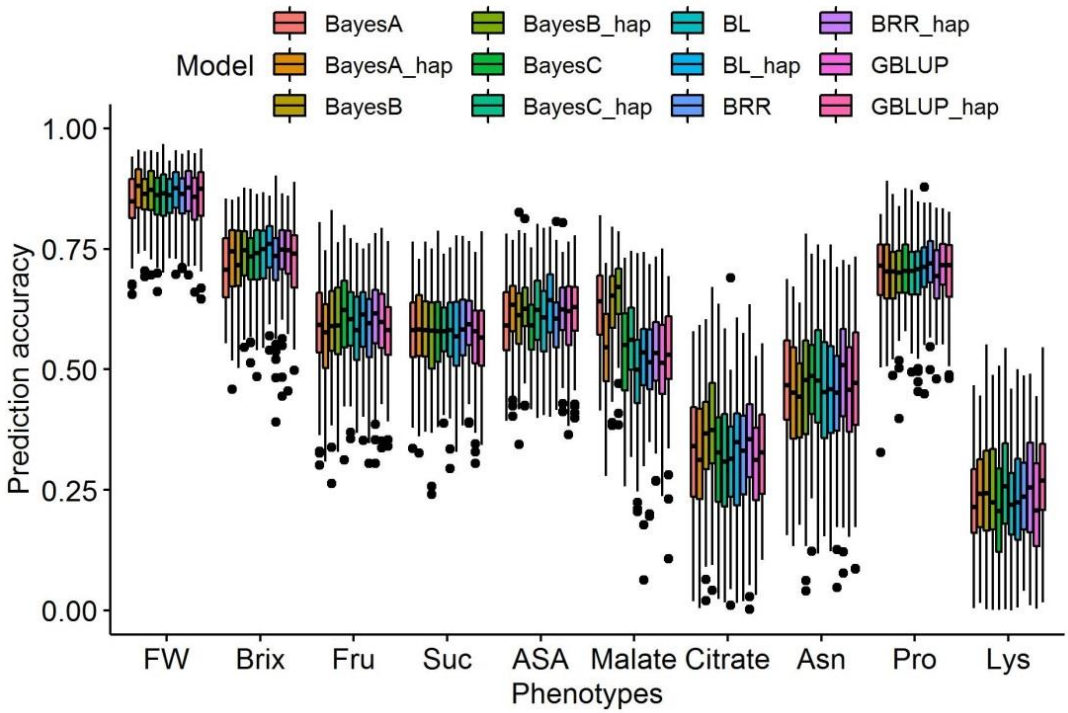


Figure 1 Comparison of genomic prediction accuracy using different models based on 100 replicates. FW, fruit weight; Fru, fructose; Suc, sucrose; ASA, ascorbic acid; Asn, asparagine; Pro, proline; Lys, lysine. GBLUP, genomic best linear unbiased prediction; BL, Bayesian LASSO; BRR, Bayesian Ridge Regression. hap indicated that the prediction was based on haplotypes.

Topic 2: Multi-haplotype mixed model versus multi-locus mixed model

In order to further demonstrate the benefits of using haplotypes, we herein compared the efficiency of multi-haplotype mixed association model (MHMM) versus multi-locus mixed model (MLMM) in identifying significant associations, taking brix and malate content traits measured on panel S as examples. To do so, we first converted the haplotype genotypes within each haplotype block into corresponding pseudo SNPs. For each unique haplotype, if it appeared in one individual, we coded it as 1; if it did not appear in the other individual, we coded it as 0. We then repeated this for the remaining unique haplotypes and for all the haplotype blocks detected. We then combined pseudo SNPs (filtered with MAF > 0.037) with the remaining SNPs that were not located within haplotype blocks. We found multi-haplotype mixed model (MHMM) had a similar performance in identifying associations for both brix and malate (**Figure 2**). For Brix, MHMM and MLMM detected 8 and 9 associations, respectively. For malate, MHMM and MLMM detected 5 and 2 associations, respectively. New associations were identified, which supported the potential benefits of using haplotypes. However, there were some associations detected in MLMM that could not be found in MHMM. This can be caused by the quality control of pseudo SNPs with MAF. As many less common haplotypes were removed, which could cause those alleles with major effects but less common removed for further analyses.

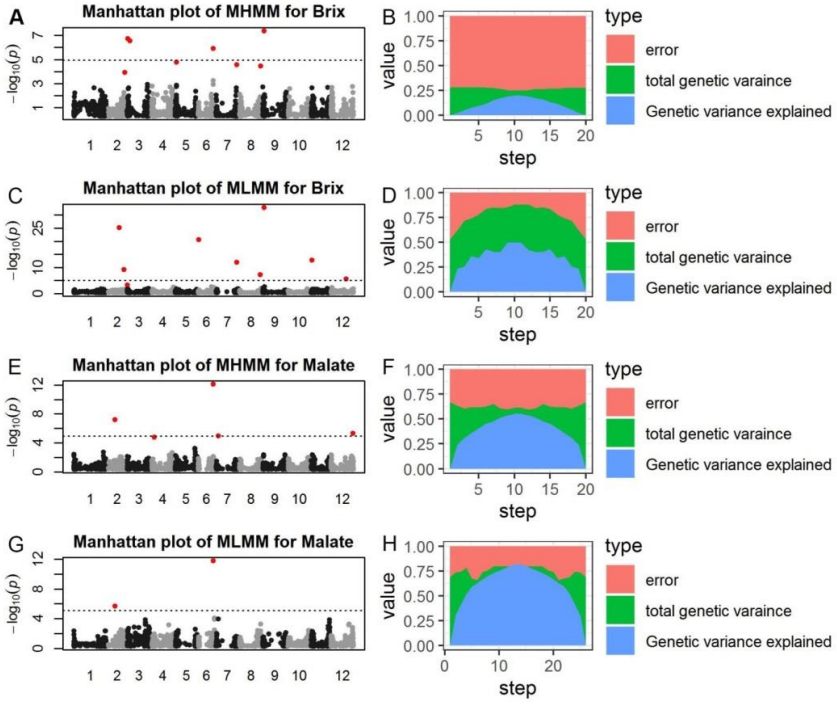


Figure 2 Comparision of MHMM and MLMM for brix and malate.

Topic 3: Composite of multiple selective signals

Identifying the selective footprints in the genome will provide useful information on which genomic regions have been selected during domestication and improvement, which in turn could help tomato breeding of quality improvement.We took both panels S and T as an example to test RAiSD in tomato (**Figure 3**). Though many strong selective sweeps were identified, most of the signals were located in the middle of the chromosomes. In some cases, no functional genes were located within the hard selective sweeps detected, even when it covered up to 3 Mbs. INotably, panel T was genotyped with next-generation sequencing and the genomic coverage of SNPs were much denser compared to panel S (about 340-folds denser). In sum, the results using RAiSD were quite difficult to interpret from an evolutionary and biological point of view and we abandoned this analysis.

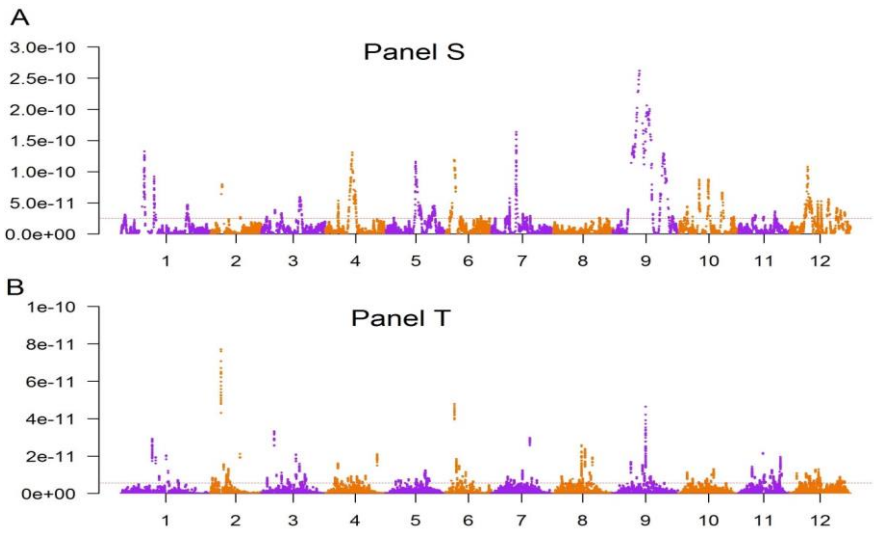


Figure 3 RAiSD (raised accuracy in sweep detection) of panel S and panel T that composited multiple selective signals.

Chapter 4

Chapter 4

Meta-analysis of genome-wide association studies provides insights into genetic control of tomato flavor

This chapter is an article that has been published in Nature Communications (DOI: 10.1038/s41467-019-09462-w). In this paper, we demonstrated in details about how to perform meta-analysis of genome-wide association studies by using the summary results from different panels. We also presented some very interesting results we found in the meta-analysis, which can be quite helpful in deepening our understandings on the genetic control of tomato flavor. The full online pdf version of this paper is available at **Appendix 3**.

Meta-analysis of genome-wide association studies provides insights into genetic control of tomato flavor

Jiantao Zhao[1], Christopher Sauvage[1,6], Jinghua Zhao[2,7], Frédérique Bitton[1], Guillaume Bauchet[1,8], Dan Liu[3], Sanwen Huang[3,4], Denise M. Tieman[5], Harry J. Klee[5], and Mathilde Causse[1*]

[1]INRA, UR1052, Génétique et Amélioration des Fruits et Légumes, Domaine Saint Maurice, 67 Allée des Chênes CS 60094 – 84143 Montfavet Cedex, France

[2]MRC Epidemiology Unit & Institute of Metabolic Science, University of Cambridge, Addrenbrooke's Hospital, Box 285, Hills Road, Cambridge CB2 0QQ, UK

[3]Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong 518124, China

[4]Key Laboratory of Biology and Genetic Improvement of Horticultural Crops of the Ministry of Agriculture, Sino-Dutch Joint Laboratory of Horticultural Genomics, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[5]Horticultural Sciences, Plant Innovation Center, University of Florida, Post Office Box 110690, Gainesville, FL 32611, USA

[6]Present address: Syngenta, 12 Chemin de l'Hobit, Saint Sauveur 31790, France

[7]Present address: Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Wort's Causeway, Cambridge, CB1 8RN, UK

[8]Present address: Boyce Thompson Institute, Cornell University, 533 Tower Rd, Ithaca, NY 14853, USA

*Corresponding author: mathilde.causse@inra.fr

Abstract

Tomato flavor has changed over the course of long-term domestication and intensive breeding. To understand the genetic control of flavor, we report the meta-analysis of genome-wide association studies (GWAS) using 775 tomato accessions and 2,316,117 SNPs from three GWAS panels. We discover 305 significant associations for the contents of sugars, acids, amino acids, and flavor-related volatiles. We demonstrate that fruit citrate and malate contents have been impacted by selection during domestication and improvement, while sugar content has undergone less stringent selection. We suggest that it may be possible to significantly increase volatiles that positively contribute to consumer preferences while reducing unpleasant volatiles, by selection of the relevant allele combinations. Our results provide genetic insights into the influence of human selection on tomato flavor and demonstrate the benefits obtained from meta-analysis.

Introduction

The deterioration of tomato flavor has been a source of complaint from consumers for decades[1]. During long-term domestication and breeding history, flavor has not been a priority, in contrast to yield, disease resistance, and postharvest shelf life[1,2]. However, flavor is one of the most important traits for improving tomato sensory quality and consumer acceptability[3]. Flavor is centrally influenced by sugars, acids, amino acids and a diverse set of volatiles[4–6]. Most of these compounds are quantitatively inherited as shown by many QTL studies but only a few QTLs have been positionally cloned[7]. Genome-wide association studies (GWAS) have detected many significant associated loci for tomato flavor related traits[6,8–12] However, reducing a QTL to a causative gene is difficult and only a few candidate genes have been functionally validated[7]. The underlying genetic control of tomato flavor is still incomplete and remains an important breeding target.

Meta-analysis of genome-wide associations is powerful in dissecting complex human diseases[13,14]. A recent meta-analysis in cattle stature also demonstrated its power in non-human species[15]. However, to the best of our knowledge, no GWAS meta-analysis has been reported in major crops, despite the increasing number of GWAS studies in major crops, such as rice. To date, the genomes of over 500 tomato accessions have been fully sequenced[6,12,16–19], making it possible to perform genotype imputation[20,21] and subsequent meta-analysis of GWAS using summary data[14] to decipher the polygenic architecture of agronomic traits. In this study, we perform a meta-GWAS on 775 tomato accessions and 2,316,117 SNPs and discover 305 significant associations for diverse flavor-related traits. Our results provide genetic insights into tomato flavor.

Results

Meta-analysis

Here we report the first meta-analysis of GWAS in tomato using results of three publicly available GWAS panels: 163 tomato accessions from panel S[8], 291 accessions from panel B[11] and 402 accessions from panel T[6] (Fig. 1). We analyzed a large set of tomato flavor-related quality chemicals, including sugars, organic acids, amino acids and volatiles measured in each of these panels.

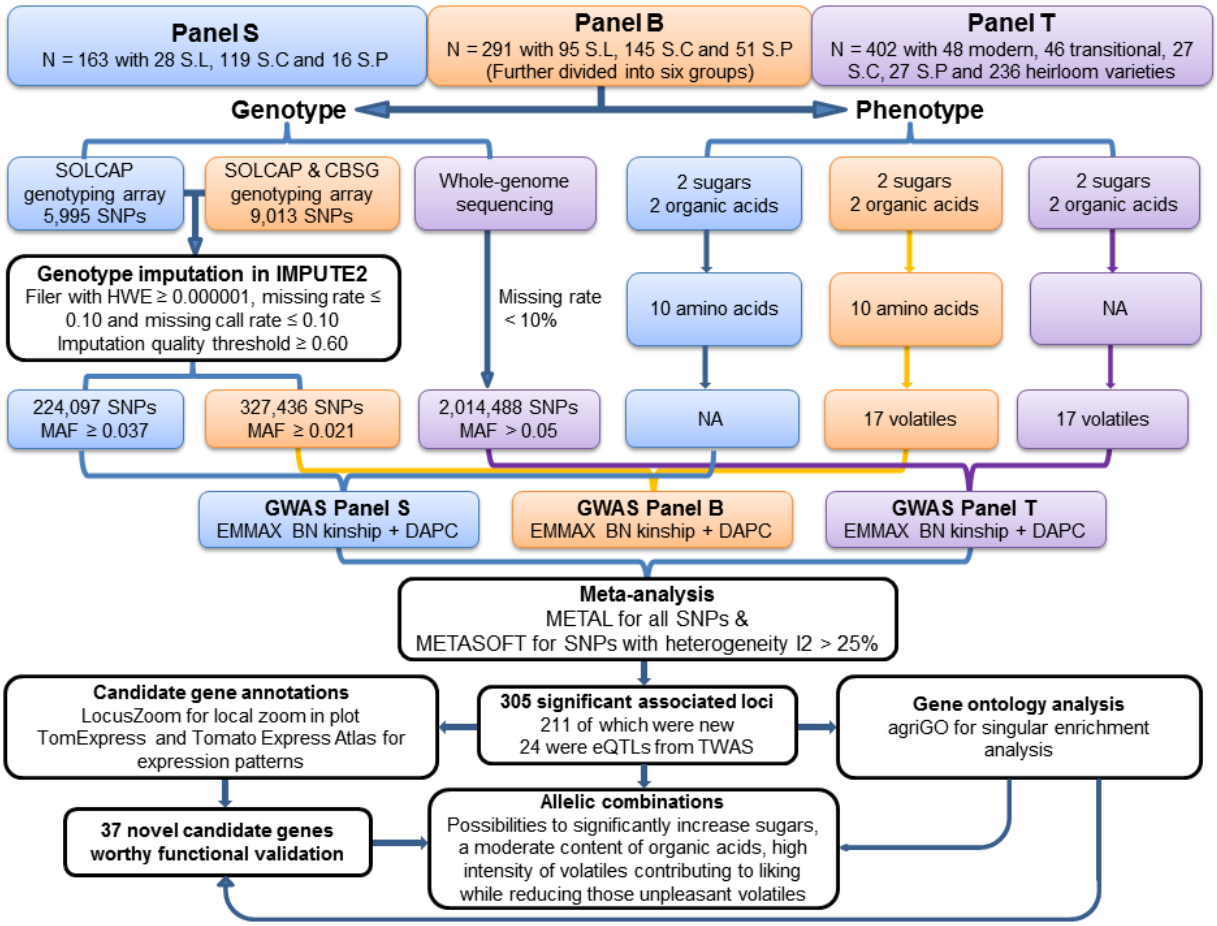


Fig. 1. Overview of study design. N, the number of individuals; S.L, *S. lycopersicum*; S.C, *S. lycopersicum* var *cerasiforme*; S.P, *S. pimpinellifolium*; Genotyping arrays: SOLCAP, Solanaceae Coordinated Agricultural Project; CBSG, Centre of Biosystems Genomics consortium; HWE, Hardy-Weinberg equilibrium; MAF, minor allele frequency; GWAS, genome-wide association study; EMMAX, Efficient Mixed-Model Association eXpedited; DAPC, Discriminant Analysis of Principal Components; eQTL, expression quantitative trait locus; TWAS, transcriptome-wide association study.

First, we used IMPUTE2 software[22] to increase the genome-wide SNP densities of panel S[8] and panel B[11], which were genotyped using SNP arrays (Online methods). After quality control (Supplementary Figures 1-3, Supplementary Tables 1-2, Supplementary Data 1-3), a total of 209,152 and 252,414 SNPs was retained for panel S and B, respectively. Imputation greatly increased the density of genomic coverage (Supplementary Figure 4) and revealed a similar genetic population structure compared with genotyped SNPs for both panels (Supplementary Figures 5-12 and Supplementary Data 4-5). We used the Efficient Mixed-Model Association eXpedited (EMMAX) software for association tests for panel S and B[23], as reported for panel T[6] (Online methods, Supplementary Figure 13). After imputation, we observed a similar or slight statistical increase in terms of the significance and the number of associated loci compared with MLMM[24] (Supplementary Figures 14-44) and no genomic

inflation ($\lambda < 1$) was detected for most (83.3%) of the traits (Supplementary Data 6). For panel T, which was characterized by 2,040,403 SNPs, the association tests had also been performed using EMMAX[6].

By combining the three separate studies, a total of 775 unique tomato accessions were used for the final meta-analysis of 31 flavor-related traits (2 sugars, 2 organic acids, 10 amino acids and 17 flavor-related volatiles). We performed the meta-analysis with two software: METAL[25] using a fixed effect model and METASOFT[26] for those SNPs where heterogeneity occurred ($I^2 > 25$) using a random effect model. Manhattan plots and quantile–quantile (Q-Q) plots for all traits are shown in Supplementary Figures 45-75. Meta-analysis identified a total of 305 significant loci ($P < 4 \times 10^{-7}$ for sugars, acids and volatiles; $P < 2.99 \times 10^{-6}$ for amino acids), among which 211 were new (Supplementary Data 7). A total of 87 strong effect meta-QTLs were identified with high probability ($P < 10^{-9}$). Most of these loci passed the suggestive thresholds in at least one panel (Figure S14-75). Among the identified loci, 35 had a moderate to strong heterogeneity ($I^2 > 25$). We generated a local SQLite dataset for tomato (Online methods) and provided the LocusZoom plots for all the genome-wide significant associated loci (Supplementary Figures 76-123). Among the 305 loci, 24 loci exhibited cis-eQTLs in a previous transcriptome-wide association study[12] in fruit tissue (Supplementary Data 7). Among the 211 associated loci, we identified 37 promising candidate genes (7 with significant cis-eQTLs[12]) with functional annotations related to the pathways of flavor chemicals (Table 1).

We performed a singular enrichment analysis for all associations using agriGO[27] (http://bioinfo.cau.edu.cn/agriGO/index.php). Up to ten biological processes were significantly enriched ($P < 0.005$) (Supplementary Data 8). All these enriched processes or groups were closely involved in flavor-related metabolites (in terms of sugars, organic acids, amino acids and volatiles), such as UDP-glycosyltransferase activity, transferase activity, oxidoreductase activity and carbohydrate metabolic processes.

Previously reported flavor-related loci in the three panels were all strongly associated in the meta-analysis at a higher significance level, such as *Lin5* (Solyc09g010080, fructose, $P = 6.16 \times 10^{-10}$; glucose, $P = 4.30 \times 10^{-10}$), *TFM6* (Solyc06g072920, malate, $P = 2.26 \times 10^{-37}$) and *Phytoene synthase 1* (Solyc03g031860, geranyl acetone, $P = 6.73 \times 10^{-26}$)[6,28]. In meta-analysis of GWAS, heterogeneity represents the genetic variations observed across combined studies[13]. In this study, strong heterogeneity occurred even for those loci with major effects,

Table 1. Summary of 37 candidate genes associated with main flavor-related traits in tomato fruit*

Trait	Chr	BP	Ref	Alt	*P*	I^2	Locus name	Candidate gene
Citrate	1	1749084	c	g	$3.62 \times 10^{-}$	0	Solyc01g007090	Aluminum-activated malate
Citrate	2	47904426	a	g	$4.30 \times 10^{-}$	97.9	Solyc02g084820	Glycosyl transferase group 1
Citrate	3	52998165	a	c	$1.84 \times 10^{-}$	0	Solyc03g083090	Glycogen synthase
Citrate	6	44955568	a	c	$7.46 \times 10^{-}$	98.4	Solyc06g072920	Aluminum-activated malate
Citrate	7	63601724	t	g	$4.70 \times 10^{-}$	0	Solyc07g055840	Citrate synthase
Fructose	1	3327330	a	g	$6.37 \times 10^{-}$	0	Solyc01g009150	Glycosyl hydrolase
Fructose	5	63485334	c	g	$4.68 \times 10^{-}$	0	Solyc05g053400[a]	Glucosyltransferase
Fructose	7	63757414	a	c	$4.28 \times 10^{-}$	0	Solyc07g055840	Citrate synthase
Fructose	8	64470216	a	g	$2.33 \times 10^{-}$	96.2	Solyc08g081420	Glycosyltransferase-like protein
Fructose	10	422707	a	t	$6.27 \times 10^{-}$	0	Solyc10g005510[a]	Glyceraldehyde-3-phosphate
Fructose	10	65465775	t	c	$6.84 \times 10^{-}$	0	Solyc10g086720	Fructose-1 6-bisphosphatase class 1
Glucose	1	1998383	a	g	$2.36 \times 10^{-}$	0	Solyc01g007910	Succinyl-CoA ligase
Glucose	2	43844073	t	c	$2.87 \times 10^{-}$	96.7	Solyc02g079220	Solute carrier family facilitated
Glucose	4	911809	a	g	$6.62 \times 10^{-}$	0	Solyc04g007160	Alpha-glucosidase
Glucose	8	58158082	a	g	$4.99 \times 10^{-}$	0	Solyc08g069060	Beta-1 3-galactosyltransferase 6
Glucose	10	332069	t	g	$1.20 \times 10^{-}$	0	Solyc10g005510[a]	Glyceraldehyde-3-phosphate
Malate	1	2650772	t	c	$2.08 \times 10^{-}$	0	Solyc01g008550	Cinnamoyl CoA reductase-like
Malate	9	72364359	a	t	$1.34 \times 10^{-}$	0	Solyc09g098590	Sucrose synthase
Malate	11	55879120	a	c	$7.14 \times 10^{-}$	0	Solyc11g072700	Glycosyltransferase-like protein
Malate	12	1824226	t	g	$1.75 \times 10^{-}$	0	Solyc12g008430	Malic enzyme
Asparagine	2	54365596	a	g	$3.72 \times 10^{-}$	94	Solyc02g093550[a]	Methyltransferase type 11
Asparagine	5	62468569	a	g	$8.92 \times 10^{-}$	0	Solyc05g052170	Acetyltransferase GNAT family
Asparagine	12	64463407	t	c	$1.13 \times 10^{-}$	0	Solyc12g089350	GDSL esterase/lipase
Aspartate	8	60307917	t	c	$6.35 \times 10^{-}$	0	Solyc08g076350	Abhydrolase domain-containing
Aspartate	11	4008385	t	g	$7.24 \times 10^{-}$	0	Solyc11g010960	Alcohol dehydrogenase
Aspartate	12	37536492	a	t	$9.16 \times 10^{-}$	0	Solyc12g044940[a]	Short-chain
Phenylalanin	11	4002767	t	c	$9.57 \times 10^{-}$	0	Solyc11g010960	Alcohol dehydrogenase
Proline	3	66798980	t	g	$2.39 \times 10^{-}$	0	Solyc03g117770[a]	Serine incorporator 1
Serine	3	69913055	a	g	$3.06 \times 10^{-}$	0	Solyc03g121910	Threonine synthase
Geranyl	2	40883244	a	g	$6.00 \times 10^{-}$	0	Solyc07g049670	Alcohol acetyltransferase
Hexenal	1	1083181	c	g	$1.45 \times 10^{-}$	0	Solyc01g006540	Lipoxygenase
Methyl	9	69293875	a	g	$2.34 \times 10^{-}$	0	Solyc09g089580	1-aminocyclopropane-1-carboxylate
1-penten-3-	5	3036212	a	g	$7.07 \times 10^{-}$	0	Solyc05g008800*	Lipid phosphate phosphatase 3
2-methyl-1-	6	37782796	a	g	$5.50 \times 10^{-}$	0	Solyc06g059850	3-methyl-2-oxobutanoate
6-methyl-5-	3	3212583	t	c	$6.76 \times 10^{-}$	0	Solyc03g025720	Long-chain-fatty-acid--CoA ligase
6-methyl-5-	4	60345897	a	t	$3.00 \times 10^{-}$	0	Solyc04g074360	UDP-glucuronosyltransferase
6-methyl-5-	10	61007386	a	g	$9.28 \times 10^{-}$	0	Solyc10g079470	L-galactono--lactone dehydrogenase

* A total of 305 loci for main tomato flavor-related quality traits were identified by meta-analysis of 775 tomato accessions and 2,316,117 SNPs. For each association, associated traits, chromosome (Chr), reference allele (Ref), alternative allele (Alt), the marker-trait association *P* value (*P*), heterogeneity I square (I^2), locus name (International Tomato Annotation Group 2.4) and candidate genes are shown. All SNP positions were aligned on the tomato reference genome version 2.50. The *P-value* is reported from the random-effect model in performed using the inverse variance-weighted fixed-effect model in METAL[25]. For those SNPs where heterogeneity occurs ($I^2 > 25$, indicating moderate heterogeneity), we used the Han and Eskin random-effects model (RE2) implemented in METASOFT[26]. We also treated those candidate genes as new if previous GWAS did not report them though the association might be significant.

** Significant cis expression quantitative trait loci (cis-eQTLs) from a previous transcriptome-wide association study (TWAS)[12] mainly based on panel T.

such as *Lin5* (fructose, I^2 = 95.6, $P = 1.05 \times 10^{-10}$; glucose, I^2 = 95.3, $P = 5.85 \times 10^{-10}$). This could be due to population structure, linkage disequilibrium, phenotyping platforms, G × E interactions, etc[13]. We then focused on loci in regions showing low LD, where one or a few candidate genes could be identified and regions with medium LD but with candidate genes near the peak SNPs.

Meta-analysis for sugar content

We looked into six candidate genes that were significantly associated both with fructose and glucose. In addition to *Lin5* and *SSC11.1*, we found four loci from the meta-analysis that were significantly associated both with fructose (Fig. 2a) and glucose content (Fig. 2b). These associations are in strong linkage disequilibrium with four candidate genes: *alpha-L-fucosidase 1* (*FUCA*; chr3: 1,506,106; fructose, $P = 3.39 \times 10^{-8}$; glucose, $P = 1.46 \times 10^{-8}$), *fatty acid elongase 3-ketoacyl-CoA synthase* (*KCS*; chr5: 3,403,706, fructose, $P = 2.57 \times 10^{-8}$; chr5: 3,406,424, glucose, $P = 1.49 \times 10^{-8}$), *glucosyltransferase* (*GTF*; chr5: 63,485,334; fructose, $P = 4.68 \times 10^{-10}$; glucose, $P = 8.36 \times 10^{-10}$) and *glyceraldehyde-3-phosphate dehydrogenase* (*GAPDH*; chr10:422,707, fructose, $P = 6.27 \times 10^{-10}$; chr10:332,069, glucose, $P = 1.20 \times 10^{-9}$). Notably, near the region of *FUCA* (up to ten genes), there are two candidate genes (Solyc03g006870, *phosphoglucomutase* and Solyc03g006860, *fructokinase*) which are also promising candidate genes for association with fructose and glucose content. Notably, *GTF* ($P = 7.55 \times 10^{-34}$) and GAPDH ($P = 7.84 \times 10^{-17}$) also showed significant cis-eQTL in a related transcriptome-wide association study[12].

Interestingly, all these loci, except *Lin5* (which falls in the domestication sweep DW149[19]), were not associated with any domestication[19] or improvement sweeps[19]. We compared the frequencies of different combinations of alleles of these candidate genes in relation to sugar content in wild, transitional, heirloom and modern accessions (more detailed explanations about group definition in Online Methods). All modern, heirloom and transitional accessions lost most of the diversity of allele combinations that is present in the wild species group (Fig. 2c). The sugar content of heirloom+transitional (heir_trans) and heirloom+modern (heir_mod) groups were both significantly lower than that of the wild species (Fig. 2d). Fruit sugar content increased gradually as the number of alternative alleles increased (Fig. 2e). We observed significant positive correlations between the number of alternative alleles within allele combinations and sugar content (Fig. 2f). In addition, total sugar content (glucose + fructose) of all alternative allele combinations was significantly higher ($P = 0.024$) than that

of all reference allele combinations (Fig. 2g). Together, these results provide insights into possibilities for tomato sugar improvement.

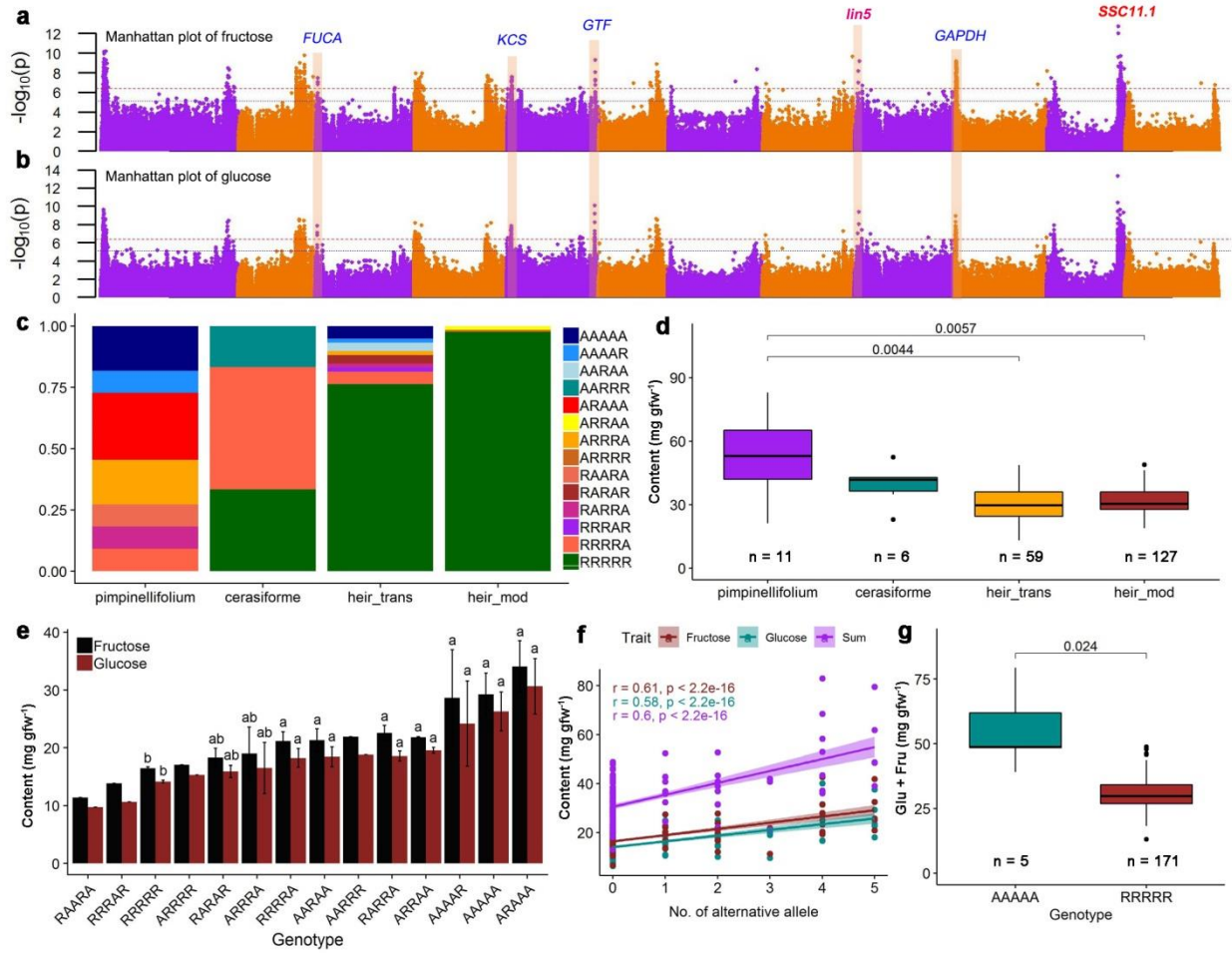


Fig. 2. Combinations of fructose and glucose alleles for the improvement of tomato sugar content. Only alleles that were significantly associated both with fructose and glucose were analyzed. **a**, **b**, Manhattan plot for meta-analysis of genome-wide association analysis of fructose (**a**) and glucose (**b**) content. Candidates and previously identified genes were labeled in blue and red, respectively. FUCA, *alpha-L-fucosidase 1*; KCS, f*atty acid elongase 3-ketoacyl-CoA synthase*; GTF, *glucosyltransferase*; GADPH, *glyceraldehyde-3-phosphate dehydrogenase*. **c**, Allele distribution of fructose/glucose content at positions: chr3:1,506,106, chr5:3,403,706, chr5:63,485,334, chr9:3,477,979 and chr10:422,707 that were both significantly associated with fructose and glucose in *S. lycopersicum* var *cerasiforme* (cerasiforme), heirloom + transitional (heir_trans), heir + modern (heir_mod) and the closest wild species *S. pimpinellifolium* (pimpinellifolium) tomato accessions (see detailed information about groups in online methods). **d**, Comparison of sugar content (fructose + glucose) between different tomato types in cerasiforme, heir_trans, heir_mod and pimpinellifolium tomato accessions. **e**, Mean (±SE) content of fructose (black) and glucose (brown) at different allele combinations in cerasiforme, heir_trans, heir_mod and pimpinellifolium tomato accessions. Significant t-test *P* values are also provided. **f**, Correlation between the number of alternative alleles and sugar content. Fructose, glucose and the sum of fructose + glucose were colored in brown4, cyan4 and purple. **g**, Comparison of sugar content (fructose + glucose) between all alternative and reference allele combinations at position chr3: 1,506,106, chr5: 3,403,706, chr5: 63,485,334, chr9: 3,477,979 and chr10: 422,707. Center line and limits of box were the mean and interquartile ranges. Error bars represent the maximum and minimum values. Whiskers indicate variability outside the upper and lower quartiles. Significant t-test *P* values are also provided. Source data of Figure 2c-g are provided in a Source Data file.

Meta-analysis for organic acids

The meta-analysis also provided several candidate genes for tomato fruit acid content. A strong association ($P = 2.26 \times 10^{-37}$ was detected for malate at an aluminum-activated malate transporter-like gene on chromosome 6, which has been reported to have a major effect on malate content[6,8,11], and was further validated as *Al-Activated Malate Transporter 9* (*Sl-ALMT9*)[28]. We found a strong significant association for citrate (chr6: 44,955,568, $P = 7.46 \times 10^{-27}$), which was 1.54kb away from *Sl-ALTM9* (Supplementary Figure 45 and Table 1). We also identified a significant association with another aluminum-activated malate transporter on chromosome 1 (chr1:1,749,084, $P = 3.62 \times 10^{-13}$; Supplementary Figure 45 and Table 1). The strong linkage with both citrate and malate indicated that *Al-Activated Malate Transporter* also plays an important role in regulating citrate content in tomato fruit.

Candidate genes directly involved in the biosynthesis of citrate and malate were also identified. For example, we identified an association with citrate on chromosome 7, 150kb away from a gene coding a citrate synthase (Solyc07g055840, $P = 4.70 \times 10^{-12}$). This candidate gene was also significantly associated with fructose ($P = 4.28 \times 10^{-09}$). For malate content, we found one association on chromosome 12 (chr12: 1,824,226, $P = 1.75 \times 10^{-19}$) close (36kb) to a gene coding a malic enzyme (Solyc12g008430, four genes away from the peak SNP). We then took six candidate genes to analyze the relationships between different allele combinations and citrate and malate content, respectively (Fig. 3). The six candidate genes for citrate were *AIMT* (*Aluminum-activated malate transporter*, chr1: 1,749,084, $P = 3.62 \times 10^{-13}$), *GTF* (*Glycosyl transferase group 1*, chr2: 47,904,426, $P = 4.30 \times 10^{-13}$), *GS* (*Glycogen synthase*, chr3: 52,998,165, $P = 1.84 \times 10^{-15}$), *AIMT* (*Aluminum-activated malate transporter*, chr6: 44,955,568, $P = 7.46 \times 10^{-27}$), *CS* (*Citrate synthase*, chr7: 63,601,724, $P = 4.70 \times 10^{-12}$) and *Rubisco* (*Ribulose-1 5-bisphosphate carboxylase/oxygenase activase 1*, chr10: 65,378,714, $P = 5.35 \times 10^{-09}$). The six candidate genes for malate were *GTF* (*UDP-glucosyltransferase*, chr2: 48,509,791, $P = 3.47 \times 10^{-28}$), *PDHB* (*Pyruvate dehydrogenase E1 component subunit beta*, chr4: 2,156,747, $P = 4.45 \times 10^{-17}$), *AIMT* (*Aluminum-activated malate transporter*, chr6: 44,999,916, $P = 2.26 \times 10^{-37}$), *SS* (*Sucrose synthase*, chr9: 72,364,359, $P = 1.34 \times 10^{-15}$), *ME* (*Malic enzyme*, chr12: 1,824,226, $P = 1.75 \times 10^{-19}$) and *GAPB* (*Glyceraldehyde-3-phosphate dehydrogenase B*, chr12: 64,816,056, $P = 5.99 \times 10^{-16}$).

Among the selected candidates, *GTF* on chromosome 2 and *AIMT* on chromosome 6 were associated with both citrate and malate (Fig. 3a, 3b). Both *GTF* and *GS* are located within improvement sweeps (IS031 and IS044, respectively)[19] and domestication sweeps (DS050 and DS175)[19] were observed for malate on *PDHB* and *ME*. For citrate and malate, the modern tomato accessions presented very different allele combinations than those in wild species and cherry tomatoes (Fig. 3c, 3d). In comparison, the total number of allele combinations for malate was approximately three times that of citrate. The citrate content was significantly different between some allele combinations (Fig. 3e). With the increase in the total number of alternative alleles in different allele combinations, the citrate content first increased gradually, with a peak at n=2, and then steadily decreased (Fig. 3f). The malate content also showed a wide range of variation among alleles (Fig. 3g and Supplementary Data 9). We observed a weak but significant (P = 0.02) positive linear correlation (r = 0.16) between the number of alternative alleles and malate content (Fig. 3h).

These results demonstrated that citrate content was more influenced by improvement sweeps while malate was more influenced by domestication sweeps in the long-term breeding history. In addition, citrate has much less allele diversity than malate and a distinct pattern of relationships between the number of alternative alleles and its content.

Meta-analysis for amino acids and volatiles

Many candidate genes associated with amino acid and volatile contents were identified. For example, we found a significant association for serine on chromosome 3 ($P = 3.06 \times 10^{-14}$) (Supplementary Figure 57 and Table 1), which was only significant in panel B ($P = 2.13 \times 10^{-9}$) (Supplementary Figure 26). The candidate gene is annotated as a threonine synthase, an enzyme involved in the serine biosynthesis pathway. For proline, we found one associated locus (Solyc03g117770, $P = 2.39 \times 10^{-9}$), which was also reported as a significant eQTL ($P = 1.04 \times 10^{-35}$)[12]. This gene is a serine incorporator, and directly regulates serine content. One locus corresponding to GDSL esterase/lipase (Solyc12g089350) was also significantly associated with four amino acids (asparagine, GABA, glutamine and threonine). For hexanal, we found the strongest association corresponding to the lipoxygenase gene *LoxC* (Solyc01g006540, $P = 1.45 \times 10^{-10}$), which encodes an enzyme that is essential for synthesis of C6 and C5 fatty acid-derived volatiles[29,30]. This candidate gene was also significantly associated with (Z)-3-hexen-1-ol ($P = 3.94 \times 10^{-07}$). For 2-methyl-1-butanol, the strongest association corresponded to a 3-methyl-2-oxobutanoate dehydrogenase gene (Solyc06g059850, $P = 5.50 \times 10^{-09}$), an enzyme associated with branched chain amino acid metabolism.

We then looked at the possibility that significantly increasing the overall intensity of volatiles contributed to consumer liking as well as significantly reducing the overall content of unpleasant volatiles by combining the strongest loci associated with the contents of six volatiles (Fig. 4). The four volatiles positively contributing to liking included geranyl acetone (chr3: 4,328,514, $P = 6.73 \times 10^{-26}$), hexanal (chr1: 1,083,181, $P = 1.45 \times 10^{-10}$), phenylacetaldehyde (chr4: 55,635,636, $P = 5.59 \times 10^{-22}$) and 6-methyl-5-hepten-2-one (chr3: 3,212,583, $P = 6.76 \times 10^{-26}$). The two unpleasant (or negative) volatiles were guaiacol (chr9: 69,299,940, $P = 5.90 \times 10^{-18}$) and methyl salicylate (chr9: 69,293,875, $P = 2.34 \times 10^{-19}$) (Fig. 4a-4f). Modern and heirloom+transitional accessions had the lowest allele diversity, especially compared with *S. pimpinellifolium* and cherry tomato accessions (*S. l. cerasiforme*). Interestingly, we also found that cherry tomatoes had the greatest diversity of allele combinations and some of them only appeared in this group (Fig. 4g). The highest total content of the four positive volatiles was observed in allele combinations of cherry tomato accessions, which were significantly higher than the allele combinations of all modern tomato accessions (Fig. 4h). In contrast, modern accessions have, on average, a significantly higher content of unpleasant volatiles, compared with the cherry accessions (Fig. 4i). These results revealed the combinations of alleles that have the potential to significantly enhance the total contents of volatiles associated with consumer liking.

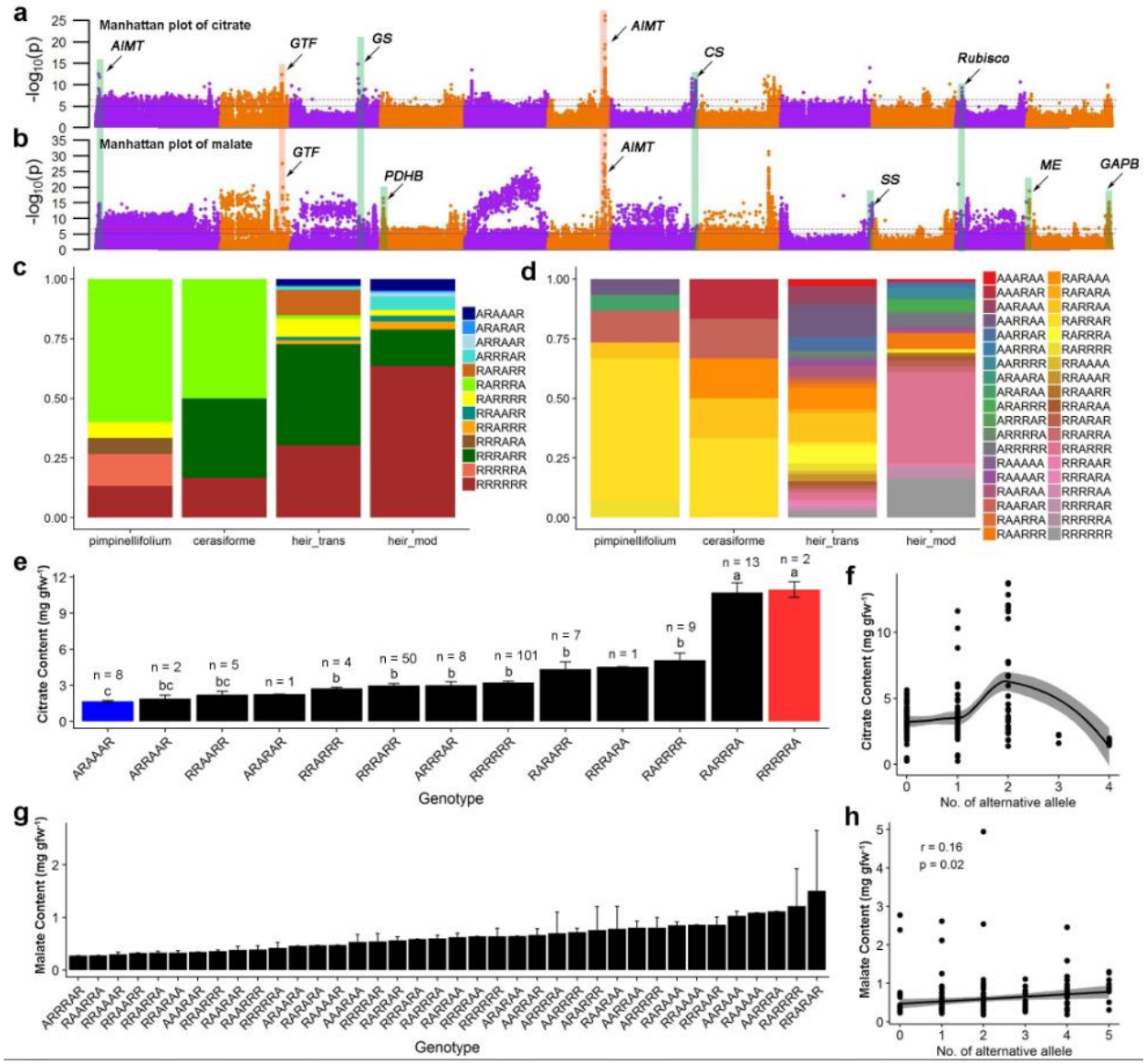


Fig. 3. Combinations of citrate and malate alleles for the improvement of tomato organic acid content. a, **b**, Manhattan plot for meta-analysis of genome-wide association analysis of citrate (**a**) and malate (**b**) content. AIMT, *Aluminum-activated malate transporter*; GTF, *Glycosyl transferase group 1*; GS, *Glycogen synthase*; AIMT, *Aluminum-activated malate transporter*; CS, *Citrate synthase*; Rubisco, *Ribulose-1 5-bisphosphate carboxylase/oxygenase activase 1*; PDHB, *Pyruvate dehydrogenase E1 component subunit beta*; SS, *Sucrose synthase*; ME, *Malic enzyme*; GAPB, *Glyceraldehyde-3-phosphate dehydrogenase B*. **c**, Allele distribution of citrate content at positions: chr1:1749084, chr2: 47,904,426, chr3: 52,998,165, chr6: 44,955,568, chr7: 63,601,724 and chr10: 65,378,714 in cerasiforme, heir_trans, heir_mod and pimpinellifolium tomato accessions. **d**, Allele distribution of malate content at positions: chr2: 48,509,791, chr4: 2,156,747, chr6: 44,999,916, chr9: 72,364,359, chr12: 1,824,226 and chr12: 64,816,056 in cerasiforme, heir_trans, heir_mod and pimpinellifolium tomato accessions. **e**, Mean (±SE, standard error) content of citrate content at different allele combinations in cerasiforme, heir_trans, heir_mod and pimpinellifolium tomato accessions. **f**, Correlation between the number of alternative alleles and citrate content. **g**, Mean (±SE) content of malate content at different allele combinations in cerasiforme, heir_trans, heir_mod and pimpinellifolium tomato accessions. **h**, Correlations between the number of alternative alleles and malate content. Source data of Figure 3c-h are provided as a Source Data file.

Discussion

With the development of next-generation sequencing technology, GWAS has become a classical genetic approach to identify QTLs and causal genes in crops[31]. We herein demonstrate the potential of meta-analysis of GWAS following the detailed protocols first proposed in human genetics[32,33], which can be easily applied in other crops. Meta-analysis of GWAS is used when pooling raw data of separate panels (mega-analysis) is not possible. It has been shown both theoretically and numerically that meta-analysis is statistically as efficient as mega-analysis[34,35]. Even when possible, it is thus not necessary to re-analyze the raw data to perform meta-analysis. Only summary data (beta, standard error and p-values of associations at each SNP) from each panel is needed and should be provided with each GWAS result. For mega-analysis, genotypes and phenotypes from all panels should be first combined and then analyzed, which requires proper management of phenotypic structure (data coming from different studies with different plant growth conditions, different harvesting and sampling procedures, different metabolic analysis protocols etc.) and genotypic structure (such as population structure and kinship). Compared to mega-analysis, meta-analysis can assess the heterogeneity (consistency) of studies, which can be caused by many factors, such as phenotypic structure, genetic structure, linkage disequilibrium, imputation accuracies or G×E interactions[13,34].

Flavor remains a major breeding challenge in tomato[1,6]. Here, we used imputation-driven meta-analysis of genome-wide association studies to greatly increase the number of SNPs linked to chemicals associated with flavor. Among the 305 significantly associated loci, 41% of the SNPs had a low frequency (MAF < 0.1). Very low-frequency (0.01 < MAF < 0.05) SNPs were also detected (3 significant associated loci) (Supplementary Figure 124). These results demonstrated that a sufficiently large sample size is needed to uncover these low-frequency and less common variants and to account for missing heritability[36–38]. Although hundreds of tomato genome sequences have been published[6,12,16–19], a high sequence depth reference panel is needed, such as the 1000 Genomes Project[39] in humans or the 1,135 Arabidopsis genomes[40] in Arabidopsis, to perform genotype imputation[20,21], heritability estimation[36,41–43] and meta-analysis[13,14] with higher accuracy. Also, an imputation server could greatly enhance the integration of genetic resources[44].

In this study, we identified 37 promising candidate genes with functional annotations consistent with their involvement in biosynthesis of flavor chemicals. With the advancement

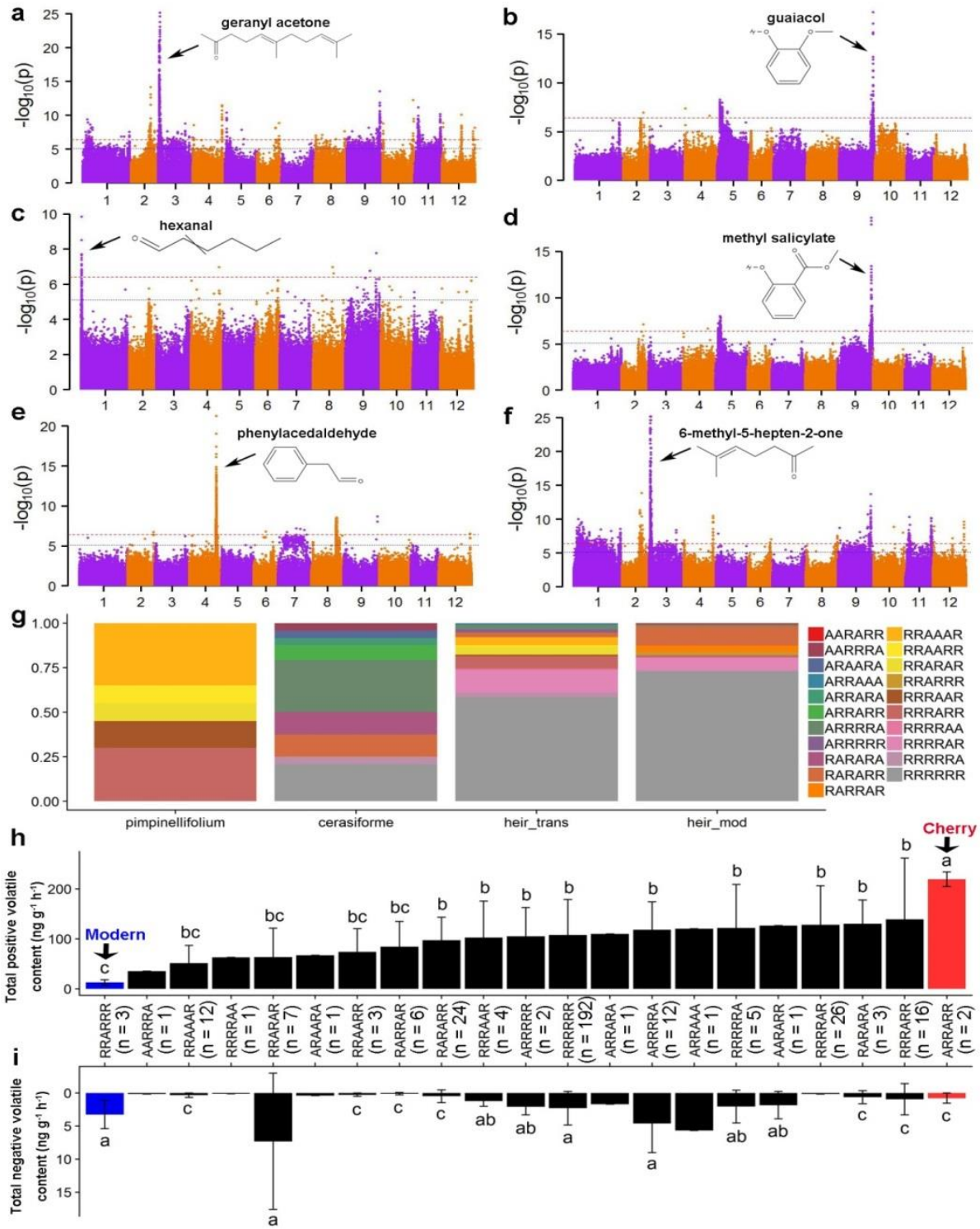


Fig. 4. Combinations of six volatile alleles for the improvement of tomato volatile content. a-f, Manhattan plot for meta-analysis of genome-wide association analysis of geranyl acetone (**a**), guaiacol (**b**), hexanal (**c**), methyl salicylate (**d**), phenylacetaldehyde (**e**) and 6-methyl-5-hepten-2-one (**f**) content. **g**, Allele distribution of six volatiles content at positions: chr3: 4,328,514 (geranyl acetone), chr9: 69,299,940 (guaiacol), chr1: 1,083,181 (hexanal), chr9: 69,293,875 (methyl salicylate), chr4: 55,635,636 (phenylacetaldehyde) and chr3: 3,212,583 (6-methyl-5-hepten-2-one) in cerasiforme, heir_trans, heir_mod and pimpinellifolium tomato accessions. **h, i**, Mean (±SE, standard error) content of total content of the four positive volatiles (geranyl acetone, hexanal, phenylacetaldehyde and 6-methyl-5-hepten-2-one) (**h**) and two unpleasant volatiles (lower panel, guaiacol and methyl salicylate) (**i**) at different allele combinations in cerasiforme, heir_trans, heir_mod and pimpinellifolium tomato accessions. Source data of Figure 4g-i are provided as a Source Data file.

of genome editing technologies, their functional analysis could greatly promote our knowledge of the genetic architecture of tomato flavor, provide fully linked markers for breeding and ensure consumer satisfaction[45–48]. It is also possible now to introduce desirable traits into wild stress-tolerant tomato accessions by genome editing [49,50]. However, tomato flavor can only be significantly improved when multiple genes are modified.

Many consumers are more attracted by small and medium size tomatoes with superior taste[51], as higher sugar content is usually associated with smaller fruit size[6]. In the meta-analysis, we found that modern cultivars have lost the majority of high-sugar alleles that were present in transitional, cherry tomato varieties and wild species. All these loci did not seem to have been influenced by any domestication or improvement sweeps, with the exception of *Lin5*, but some were loosely linked to fruit weight QTLs due to large LD in tomato. These results reflect the fact that sugar content has not been a breeding priority, in contrast to fruit size, yield, biotic and abiotic resistances[1,6]. Strong positive correlations between the number of alternative alleles and sugar content provide clues on how to select higher sugar content tomato cultivars. However, sugar content can only be significantly improved when almost all the alternative alleles are selected, and will probably be accompanied by reduced fruit size[6] except if precise recombination or genetic modifications limits the linkage drag effect.

Malate and citrate are the main organic acids in most ripe fruits[52]. In tomato, citrate has a stronger impact on consumer preferences. In this study, candidate genes potentially impacting both citrate and malate contents were identified. We also demonstrated that citrate has been more influenced by improvement sweeps and malate by domestication sweeps. These results show that citrate was probably selected for improving tomato flavor.

Flavor-related volatiles are strongly influenced by the environment[53,54]. Nevertheless this meta-analysis illustrates that it should be possible to significantly enhance the content of favorable aromas via replacement of undesirable alleles. However, unlike sugars, the undesirable alleles should be carefully chosen[6]. Cherry tomato varieties have been introduced to the market since the 1990s. Their genomes are an admixture of those of big-fruited tomatoes and *S. pimpinellifolium* species[19,55] and may still contain a large number of favorable alleles. Thus they may serve as the most promising allele reservoir for breeding of high-flavor tomatoes.

In conclusion, we performed the first meta-analysis of genome-wide association analyses in a major vegetable and identified numerous loci involved in tomato flavor that were not identified in the three independent studies. A strong positive correlation between allele combinations and sugar content provides clues for breeding for higher sugar content. Modern cultivars have lost most of the allelic diversity for sugars, acids and volatiles that is present within the species. Significant improvements should be achieved by replacing undesirable alleles. Taken together, our meta-analysis provides genetic insights into the genetic control of tomato flavor and gives a roadmap for flavor improvement.

Methods

Three GWAS panels

The meta-GWASs approach is based on three different GWAS panels already published and genotyped using different technologies. Our approach consisted in imputing SNP data for panels S[8] and B[11] from a reference panel, then conducting separate GWAS using the same mixed linear model (MLM) as described in[6] and collecting the summary statistics to run a meta-GWAS.

Panel S consists of 163 accessions[8], including 28 *S. lycopersicum* (large tomato), 119 *S. lycopersicum* var *cerasiforme* (cherry tomato) and 16 *S. pimpinellifolium* (closest wild species). This panel was genotyped using the Solanaceae Coordinated Agricultural Project (SOLCAP) genotyping array[56,57], generating 5,995 high quality SNPs. The minimal success genotyping rate per accession was fixed at 90%. The minor allele frequency of SNPs ranged from 0.037 to 0.45. Tomato accessions in Panel S were grown in Avignon, France, following a randomized complete block design, in a greenhouse during the summers of 2007 and 2008[8,58].

Panel B consists of 300 accessions with 62 *S. pimpinellifolium*, 48 *S. lycopersicum* and 190 *S.* l. *cerasiforme* accessions[11]. This panel was genotyped both with the SOLCAP[56,57] and CBSG arrays[59]. After quality control, 9,013 SNPs (minor allele frequency, MAF > 0.1) and 291 accessions were kept. Accessions in Panel B were grown in Agadir, Morocco, France, under

passive greenhouse irrigated conditions in 2011 and 2012[11]. Each trial followed a randomized complete block design, with three and two blocks, in 2011 and 2012, respectively.

Panel T consists of 402 tomato accessions from two separate panels[6]. Panel T was genotyped by whole genome resequencing technology, generating a number of 2,014,488 SNPs passing quality control (MAF > 0.05, missing rate < 10%). This panel includes five tomato types, including modern (51), transitional (50), cherry (27), heirloom (243) and wild species (27)[6].

Phenotypes

A total of 31 flavor-related quality traits in tomato were analyzed for meta-analysis, including two sugars (fructose and glucose), two organic acids (citrate and malate), 10 amino acids and 17 flavor-related volatiles. The 10 amino acids were asparagine, aspartate, GABA, glutamine, lysine, methionine, phenylalanine, proline, serine and threonine. The 17 volatiles were (E)-2-heptenal (E2HEP), (E)-2-hexenal (E2HEX), (E)-2-pentenal (E2PEN), (E,E)-2,4-decadienal (EE24D), (Z)-3-hexen-1-ol (Z3H1X), (Z)-3-hexenal (Z3HEX), 1-octen-3-one (X1O3ON), 1-penten-3-one (X1P3ON), 2-methyl-1-butanol (X2M1BU), 3-methyl-1-butanol (X3M1BU), 6-methyl-5-hepten-2-one (X6MHON), beta-ionone (BIONO), geranylacetone (GRACE), guaiacol (GUAIA), hexanal (XEXAN), phenylacetaldehyde (PHEAC) and methylsalicylate (METHY).

Sugars and organic acids were measured in all three panels. Amino acids were measured both in panel S and B, while flavor-related volatiles were measured both in panel B and T. Briefly, fructose and glucose in panel S were measured using the micro-method. Citrate and malate were measured by gas chromatography-mass spectrometry (GC-MS)[8]. Data distribution was tested using the Shapiro-Wilk test and data with a non-normal distribution were Log_{10} transformed. In panel B, these metabolites were measured within the Product Metabolism and Analytical Sciences Endogenous Metabolite Profiling Platform at Syngenta Jealott's Hill International Research Center, Bracknell, UK. Fructose and glucose were analyzed by high pH ion-exchange chromatography. Citrate and malate were analyzed using electrospray ionization-liquid chromatography (ESI-LC-MS/MS). Fructose and malate were transformed using the Boxcox method. Citrate was transformed using the Log_{10} method. In panel T, citrate and malate were measured using the citrate and malate analysis kits (R-Biopharm, Marshall, MI), according to the manufacturer's instructions[60]. Measurements of amino acids and volatiles in panel S was measured using GC-MS by comparing with a database of authentic standards. Small organic acids and amino acids in panel B were analyzed using

electrospray ionization-liquid chromatography (ESI-LC-MS/MS).Volatiles in panel T were first captured by headspace solid phase micro extraction (HS-SPME) coupled GC-MS.

Reference panel for SNP imputation

A reference panel was selected from the 360 re-sequenced tomato accessions [19] to perform SNP imputation in panels S and B. Among this panel, only accessions with genome coverage ⩾ 90% and mean sequencing depth ⩾ 4.0 were kept. Wild tomato species were also removed, generating a total reference set of 221 accessions genotyped with 3,809,156 SNPs (Table S1).

Recombination map

A high-density recombination map is required for imputation and computing genomic partitions. However, the available tomato genetic maps EXPIM 2012 and EXPEN 2012[57] have a limited genomic coverage (~3500 mapped SNPs). In order to use a much denser genetic map, we developed a Python script to infer the corresponding genetic positions of the 3,809,156 SNPs in the reference panel. Before calculating the recombination rate, we first compared the physical *vs* genetic distribution patterns for each chromosome (Fig. S1). Comparing with EXPIM 2012, this newly built genetic map had the same distribution pattern (Fig. S1). This comparison indicated the inferred genetic positions were accurate and were then used for estimating the recombination rate, as required for imputation. Minor adjustments were also done for some SNPs in order to follow an overall increasing positional order. Extreme recombination rate values were also removed (> 2000 cM/Mb).

Genotype imputation

One unphased reference panel from IMPUTE2 (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#home)[22] was adopted for imputation of panel S and B independently. The 221 filtered sequenced accessions passing quality control were used as the reference panel. The newly built recombination map was used instead of EXPIM 2012. The whole genome was then divided into genomic intervals of 5 Mb for imputation and the effective size of population (*Ne*) was set at 2000.

Quality control

After imputation, the minimum MAF for panel S and B was set at 0.037 and 0.021, respectively, according to the formula: [Number of chromosomes/(2 × Number of

individuals)][61]. After combining all the imputed data, basic statistic summaries were obtained in QCTOOL v2 (http://www.well.ox.ac.uk/~gav/qctool_v2/) with the following command: ./qctool -g GWAS.gen -snp-stats. We then filtered all imputed SNPs with Hardy-Weinberg equilibrium (HWE) $\geqslant$ 0.000001, MAF$\geqslant$ 0.037 (0.021 for panel B), missing rate $\leqslant$ 0.10 and missing call rate $\leqslant$ 0.10. After these primary control steps, a total of 224,097 and 327,436 SNPs were retained for panel S and B, respectively.

In order to determine the optimal threshold of imputation quality (Info criteria), we compared the imputed and sequenced genotype data of the nine overlapping accessions in panel S that have been genotyped by SNP arrays and whole-genome sequencing. If the maximum of the three probabilities at a locus was higher than 0.9, we treated it as a certainty. This was done by converting the imputed data to ped/map format via GTOOL (http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html). We then compared the imputed and genotyped values of the nine accessions (Fig. S2). Total numbers of corrected SNPs at different MAF and Info thresholds were obtained to validate the optimal threshold of MAF and Info. The average value of Info was 0.882 (with no filtering of MAF). With the increase of Info, the number of correctly genotyped SNPs increased from less than 200 to about 50,000 for panel S (Fig. S2a, Table S2). On average, 51.45% of the SNPs have been correctly imputed for all Info values. There was no significant difference between the numbers of corrected imputed SNPs for different Info values of the three tomato groups (Fig. S2b). The majority of imputed SNPs had a MAF value ranging from 0.037 to 0.25, with a mean value of 0.172 ± 0.103 (with no filtering of Info). The percentage of successfully genotyped SNPs averaged at 57.3% and a higher percentage of corrected imputed SNPs decreased gradually with the increase of MAF (Fig. S2c). Similarly, no significant difference was found between the numbers of corrected imputed SNPs for different MAF values of three tomato genetic groups (Fig. S2d). Details of the number and percentage of corrected imputed SNPs at different MAF bins among the nine accessions are listed in Table S3. We than compared the relationship between MAF and Info. The average value of Info was 0.912 for all values of MAF (Fig. S2e). We found that the lowest mean value of Info (0.622) was observed on less common SNPs (0.037 < MAF < 0.05) (Fig. S2e, Table S4). However, this value is still higher than the proper imputation quality threshold (0.4) in common quality control of meta-analysis of genome-wide association studies[33]. So, we decided to set the Info threshold at 0.60 as the threshold of high imputation quality.

After filtering with imputation quality threshold (Info) ≥ 0.60, total of 209,152 and 252,414 SNPs were retained for panel S and B, respectively. The mean Info value at different MAF values for panel S and B were 0.929 and 0.922, respectively (Table S5). The lowest mean value of Info at different MAF value was 0.810 and 0.783, respectively (Fig. S2f, Fig. S3). These SNPs offered a much denser genomic coverage for both panel S and B (35-fold and 28-fold, respectively) (Fig. S4). Only some large genomic gaps still remained where there were few genotyped SNPs over a long genomic region (Fig. S4). These results indicated that all the retained SNPs had a high imputation quality and were used for further analyses.

Linkage disequilibrium analysis

For population structure and kinship analyses, only independent SNPs ($r^2 < 0.2$) were used. This was done in PLINK (https://www.cog-genomics.org/plink2) with: --indep-pairwise 50 5 0.2 (windows, step, r2) –maf 0.05, generating a total of 3,602 and 4,294 independent SNPs for panel S and B, respectively.

Principal component analysis

In order to compare the genetic structure revealed before and after imputation, we performed a principal component analysis (PCA) for panels S and B, using all genotyped SNPs and independent imputed SNPs ($r^2 < 0.2$) in PLINK: --pca. Principal component analysis showed that genotype imputation did not lead to significant differences in genetic group composition and pairwise individual distances, for all three accession classes of panel S (S.C., S.L., S.P.) (Fig. S5a-c). For the first principal component (PC1), there were strong positive correlations (0.93, 0.82, 0.93 for S.C, S.L. and S.P. respectively) between genotyped and imputed SNPs (only imputed SNPs) (Fig. S5d). By combining genotyped and imputed SNPs together (hereafter called ‘All’ dataset), a similar strong positive correlation (0.94, 0.82, 0.94 for S.C, S.L. and S.P. respectively) was also found (Fig. S5e). Correlation between imputed and all SNPs was also strong for all tomato classes (Fig. S5f). For the panel B, a previous study revealed a population structure composed of six groups[62]. After imputation, we found they had a similar distribution pattern (Fig. S6). PC1 between genotyped SNPs and all (genotyped and imputed) SNPs had a strong positive correlation (higher than 0.7 for all six groups) (Fig. S6c). In contrast, the second principal component (PC2) had strong negative correlations for all six groups (lower than -0.6 for all six groups) (Fig. S6d).

Population structure

In a previous study, the population structure of panel S was evaluated by Structure v2.3.4[63] (https://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/html/structure.html). So we first compared the structure following the same parameters, with 1×10^6 burn-in period and 5×10^6 MCMC steps. Based on the Evanno method[63], the optimal number of ancestral populations was two. Only minor population assignment differences were found for both subpopulations, compared with structure from genotyped SNPs (Figure S7).

We further used discriminant analysis of principal components (DAPC)[64] (http://adegenet.r-forge.r-project.org/files/tutorial-dapc.pdf) using the independent 3,602 and 4,294 SNPs ($r^2 < 0.2$) to infer the optimal population structure for panels S and B. This method partitioned the variance within and among groups without assumptions on LD or Hardy-Weinberg equilibrium[65], which has shown a better performance in clustering individuals[11]. The optimal number of clusters was determined by Bayesian Information Criteria (BIC) with a minor increase or decrease. All PCs and all discriminant functions were retained to find the optimal number of clusters. In the following DAPC analyses, all discriminant functions and the first 50 PCs were retained in order to achieve 80% of cumulative variance for both panel S and B.

For panel S, the optimal number of clusters was six (Fig. S8) and DAPC revealed a clear structure of all the accessions (Fig. S9). For panel B, the optimal number of cluster was six, which was the same as that revealed by using genotyped SNPs (Fig. S10). Membership of each cluster was also quite similar (Fig. S11), compared with that of genotyped SNPs (Fig. S12). Detailed information of the membership of each cluster revealed by all independent SNPs for panels S and B is listed in Table S6 and S7**,** respectively. These results indicated that imputation did not cause significant differences in the genetic structure for both panels S and B. For panel T, the optimal number of clusters was five from DAPC with the first 20 PCs retained and a cross validation run of 100 times[6].

Genome-wide association analysis

Though SNPTEST v2.5.4 (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html#introduction) can use the imputed data from IMPUTE2 to detect associations directly, it cannot however handle too many cofactors in the model. For accessions from each panel used in this study, there is strong genetic structure. We first took one trait (malate) in panel S as an example to choose the optimal association software to perform the association tests.

In order to add kinship as a cofactor in SNPTEST, we performed a principal component analysis of the kinship calculated in SPAGeDi (http://ebe.ulb.ac.be/ebe/SPAGeDi.html) and structure in Structure v2.3.4. We then added the first 20 PCs as cofactors in the frequentist association test model in SNPTEST. In the next step, we used EMMAX (http://genetics.cs.ucla.edu/emmax/index.html) with the BN kinship matrix and DAPC results to conduct association analyses. For BN kinship calculation, the default command was used: emmax-kin -v -h -d 10. A uniform threshold (P=1/n, n is the effective number of independent SNPs) was used as the genome-wide significance threshold for all three panels. The effective number of independent SNPs was calculated in Genetic type 1 Error Calculator (GEC)[66] (http://grass.cgs.hku.hk/gec/download.php). The suggestive p-value for the 224,097 SNPs of panel S was 9.63×10^{-5} and the significant *p-value* was 4.82×10^{-6}. For the 327,436 SNPs of panel B, the suggestive and significant p-value was 5.99×10^{-5} and 2.99×10^{-6}, respectively.

After comparing the association results for malate of panel S, we found the strongest p-value in SNPTEST was still quite low, compared with other approaches (Fig. S13). Results from MLMM (https://github.com/Gregor-Mendel-Institute/MultLocMixMod) and EMMAX were quite similar. So, in the following analyses, we only used SNPTEST to compute summary statistics, not for finding associations. For MLMM, this model adds the marker as co-factor using a window of 10. If too many markers are in full LD, the genetic variance calculation may be biased[24]. So, we used EMMAX for association analyses for all traits with the BN kinship matrix and DAPC results as covariance.

Meta-analysis

A total of 788 tomato accessions and 2,316,117 SNPs from three GWAS panels were used for the final meta-analysis. Since each panel was stratified and a small number of individuals overlapped between panels (38 between panel B and S, 18 between panel S and T, 17 between panel B and T), genomic inflation factor (λ) was corrected before meta-analysis using GenABEL[61] (http://www.genabel.org/packages/GenABEL) in R. Genomic inflation can be caused by population structure, cryptic relatedness, genotyping errors, sample size, LD, trait heritability, number of causal variants and other technical artefacts[67]. Though no adjustment is necessary when λ is lower or equal to one, we still corrected the standard errors of beta coefficients by applying the formula SE × $\sqrt{\lambda}$ in general for each individual studies to get the chi-squares to its optimal values[68].

METAL[25] (fixed-effect model) (https://genome.sph.umich.edu/wiki/METAL_Documentation) and METASOFT[26] (random-effect model) (http://genetics.cs.ucla.edu/meta/) are two most commonly used meta-analysis software[13]. Meta-analysis was first performed using the inverse variance-weighted fixed-effect model in METAL[25]. The genome-wide significant p-value for meta-analysis was set as 4.0×10^{-7}, except for SNPs that only appeared between panel S and B (the significant p-value was set at 2.99×10^{-6}). For those SNPs where heterogeneity occurs ($I^2 > 25$, indicating moderate heterogeneity), we used the Han and Eskin random-effects model (RE2) in METASOFT[26]. This model assumes no heterogeneity under the null hypothesis and offers greater power under heterogeneity, compared with conventional random-effect models[26].

Local SQLite database for LocusZoom

In order to obtain a regional zoom plot of the candidate SNPs in LocusZoom[69] (https://genome.sph.umich.edu/wiki/LocusZoom_Standalone), a local SQLite database of tomato was required. We thus created a custom SQLite database in LocusZoom with the following steps. SNP positions in the 221 accessions of the reference panel were inserted by: dbmeister.py --db my_database.db --snp_pos my_snp_pos_file. For the gene information, we first downloaded the gene annotation file from Solgenomics (ftp://ftp.solgenomics.net/genomes/Solanum_lycopersicum/annotation/ITAG2.4_release/). We then converted it to genePred file format by gff3ToGenePred (http://hgdownload.cse.ucsc.edu/admin/exe/). Gene names were replaced with short codes instead of providing full names to avoid long names and overlapping. We then inserted the gene information by the following command line: dbmeister.py --db my_database.db --refflat my_refflat_file. For the recombination file, we used the recombination map previously inferred and inserted the data into our database by: dbmeister.py --db my_database.db --snp_set my_snpset_file. We used the 221 reference panel to calculate the linkage disequilibrium (LD) in PLINK by the following parameter: --ld-snp my.snp --ld-window-kb 100000 --ld-window 1000 --r2 --ld-window-r2 0 (windows, step, r2).

LD in candidate gene regions

In order to define the window size of the candidate genes, we first calculated the LD around the significant associated SNP with the window size of 5 Mb in PLINK with the following command line: --ld-window-kb 500000 --ld-window 1000 --r2 --ld-window-r2 0 (windows, step, r2). We then chose LD higher than 0.5 as the threshold of LD decay for the candidate

gene region sizes. Within the regions, we chose the candidate genes based on both the distance of the peak SNP as well as the closest genes with known functions related to the trait. If no gene fell in the candidate regions, we provided the closest gene. We further crosschecked the candidate gene expression patterns using the Tomato Expression Atlas[70] (http://tea.solgenomics.net/expression_viewer/input).

Group re-definition of panel T

The relationship between allele combinations and flavor-related metabolites (sugars, organic acids and volatiles) was only based on panel T. For the accessions in panel T, they were previously defined as five clusters, namely *S. lycopersicum* var *cerasiforme*, heirloom, transitional, modern and the closest wild species *S. pimpinellifolium* tomato accessions[6]. However, there were up to 11 accessions with duplicated individual IDs (Supplementary Data 10) and we cross-checked these duplicated lines and only kept one. In addition, some accessions in the group of heirloom, modern and transitional were labeled inappropriately based on the DAPC analysis. In order to correct for this, we generated the principal component analysis (PCA) based on independent SNPs (LD = 0.1) (Supplementary Figure 125). Based on PCA, some heirloom accessions are mixed with modern accessions and were labeled as heir_mod (heirloom and modern). For the remaining heirloom accessions, they were combined with transitional accessions and labeled as heir_trans (heirloom and transitional) (Supplementary Figure 126). The accessions of panel T were thus re-defined as four clusters, namely *S. lycopersicum* var *cerasiforme*, (cerasiforme, 26 members), heirloom and modern (heir_mod, 196 members), heirloom and transitional (heir_trans, 138 members), and *S. pimpinellifolium* (27 members) (Supplementary Data 10-11). These redefined groups were then used for allelic combination analyses. Statistical tests were only performed for those allele combinations with at least two observations (either labeled with letters or with p-values).

Data availability

Data supporting the findings of this work are available within the paper and its Supplementary Information files. All new meta-analysis data associated with the paper are available in a repository [https://doi.org/10.15454/TWFDYW]. The source data underlying Figs 2c-g, 3c-h and 4g-i and Supplementary Figs 5a-f, 6a-d and 124-126 are provided as a Source Data file. Additional datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgements

Jiantao Zhao was funded by a Chinese Scholarship Council (CSC) scholarship. We thank Guangtao Zhu from Huang's group in helping by providing the original GWAS results of panel T and discussions about the results. We thank Qi Wu from the University of Cambridge for detailed theoretical explanations about linkage disequilibrium and population genetics. We thank David Francis from Ohio State University for the positive discussions and cross-checking the misclassification of the accessions in panel T. We thank Rebecca Stevens for the English language editing.

Author contributions

Study design/conception: M.C, J-T.Z., C.S.; Supervision: C.S, M.C.; Data collection and analysis: J-T.Z., F.B. J-H.Z., D.L., G.B., S.H., D.T., H.K.; Data interpretation: J-T.Z., F.B., C.S., M.C., D.T., H.K.; First draft of the manuscript: J-T.Z.; Critical revisions of the manuscript: all co-authors.

Competing interests

The authors declare no competing of interests.

Supplementary materials

Supplementary Tables 1-2 and **supplementary Figures 1-13,124-126** are available at the end of the thesis in **Appendix 2**. **Supplementary Figures 14-123** and **supplementary Data1-5,9-10** are available on line: https://www.nature.com/articles/s41467-019-09462-w.

References

1. Klee, H. J. & Tieman, D. M. The genetics of fruit flavour preferences. *Nat. Rev.Genet.***19,** 347–356 (2018).
2. Tieman, D. *et al.* The chemical interactions underlying tomato flavor preferences. *Curr. Biol.* **22,** 1035–1039 (2012).
3. Causse, M. *et al.* Consumer preferences for fresh tomato at the European scale: A common segmentation on taste and firmness. *J. Food Sci.* **75,** S531–S541 (2010).
4. Baldwin, E. A., Scott, J. W., Shewmaker, C. K. & Schuch, W. Flavor trivia and tomato aroma: Biochemistry and possible mechanisms for control of important aroma components. *HortScience* **35,** 1013–1022 (2000).
5. Goff, S. A. & Klee, H. J. Plant volatile compounds: Sensory cues for health and nutritional value? *Science (80-.).* **311,** 815–819 (2006).
6. Tieman, D. *et al.* A chemical genetic roadmap to improved tomato flavor. *Science (80-.).* **355,** 391–394 (2017).
7. Rothan, C., Diouf, I. & Causse, M. Trait discovery and editing in tomato. *Plant J.* **97,** 73–90 (2019).
8. Sauvage, C. *et al.* Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant Physiol.* **165,** 1120–1132 (2014).
9. Zhang, J. *et al.* Genome-wide association mapping for tomato volatiles positively contributing to tomato flavor. *Front. Plant Sci.* **6,** 1042 (2015).
10. Zhao, J. *et al.* Association mapping of main tomato fruit sugars and organic acids. *Front. Plant Sci.* **7,** 1–11 (2016).
11. Bauchet, G. *et al.* Identification of major loci and genomic regions controlling acid and volatile content in tomato fruit: implications for flavor improvement. *New Phytol.* **215,** 624–641 (2017).
12. Zhu, G. *et al.* Rewiring of the fruit metabolome in tomato breeding. *Cell* **172,** 249–261.e12 (2018).
13. Evangelou, E. & Ioannidis, J. P. A. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **14,** 379–389 (2013).
14. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18,** 117–127 (2017).
15. Bouwman, A. C. *et al.* Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nat. Genet.* **50,** 362–367 (2018).
16. Sato, S. *et al.* The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485,** 635–641 (2012).
17. Aflitos, S. *et al.* Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. *Plant J.* **80,** 136–148 (2014).
18. Bolger, A. *et al.* The genome of the stress-tolerant wild tomato species *Solanum pennellii*. *Nat. Genet.* **46,** 1034–1038 (2014).
19. Lin, T. *et al.* Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.* **46,** 1220–1226 (2014).
20. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11,** 499–511 (2010).
21. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48,** 1284–1287 (2016).
22. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).
23. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42,** 348–354 (2010).
24. Segura, V. *et al.* An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* **44,** 825–830 (2012).
25. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–2191 (2010).
26. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88,** 586–598 (2011).
27. Tian, T. *et al.* agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* **45,** W122–W129 (2017).
28. Wang, B. *et al.* An InDel in the promoter of Al-ACTIVATED MALATE TRANSPORTER9 selected during tomato domestication determines fruit malate contents and aluminum tolerance . *Plant Cell* **29,** 2249–2268 (2017).
29. Chen, G. *et al.* Identification of a specific isoform of tomato lipoxygenase (TomloxC) involved in the generation of fatty acid-derived flavor compounds. *Plant Physiol.* **136,** 2641–2651 (2004).

30. Shen, J. *et al.* A 13-lipoxygenase, TomloxC, is essential for synthesis of C5 flavour volatiles in tomato. *J. Exp. Bot.* **65,** 419–428 (2014).
31. Liu, H. J. & Yan, J. Crop genome-wide association study: a harvest of biological relevance. *Plant J.* **97,** 8–18 (2019).
32. Turner, S. *et al.* in *Current Protocols in Human Genetics* **Chapter 1,** Unit1.19 (NIH Public Access, 2011).
33. Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9,** 1192–1212 (2014).
34. Panagiotou, O. A., Willer, C. J., Hirschhorn, J. N. & Ioannidis, J. P. A. The power of meta-analysis in genome-wide association studies. *Annu. Rev. Genomics Hum. Genet.* **14,** 441–465 (2013).
35. Lin, D. & Zeng, D. Meta-analysis of genome-wide association studies: No efficiency gain in using individual participant data. *Genet. Epidemiol.* **34,** 60–66 (2010).
36. Yang, J., Zeng, J., Goddard, M. E., Wray, N. R. & Visscher, P. M. Concepts, estimation and interpretation of SNP-based heritability. *Nat. Genet.* **49,** 1304–1310 (2017).
37. Gibson, G. Rare and common variants: Twenty arguments. *Nat. Rev. Genet.* **13,** 135–145 (2012).
38. Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* **542,** 186–190 (2017).
39. Gibbs, R. A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
40. Alonso-Blanco, C. *et al.* 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166,** 481–491 (2016).
41. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
42. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11,** 446–450 (2010).
43. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S. & Balding, D. J. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49,** 986–992 (2017).
44. Wang, D. R. *et al.* An imputation platform to enhance integration of rice genetic resources. *Nat. Commun.* **9,** 3519 (2018).
45. Gao, C. The future of CRISPR technologies in agriculture. *Nat. Rev. Mol. Cell Biol.* **19,** 275–276 (2018).
46. Rodríguez-Leal, D., Lemmon, Z. H., Man, J., Bartlett, M. E. & Lippman, Z. B. Engineering quantitative trait variation for crop improvement by genome editing. *Cell* **171,** 470–480.e8 (2017).
47. Huang, S., Weigel, D., Beachy, R. N. & Li, J. A proposed regulatory framework for genome-edited crops. *Nat. Genet.* **48,** 109–111 (2016).
48. Yin, K., Gao, C. & Qiu, J.-L. Progress and prospects in plant genome editing. *Nat. Plants* **3,** 17107 (2017).
49. Zsögön, A. *et al.* De novo domestication of wild tomato using genome editing. *Nat. Biotechnol.* **36,** 1211–1216 (2018).
50. Gao, C. *et al.* Domestication of wild tomato is accelerated by genome editing. *Nat. Biotechnol.* **36,** 1160–1163 (2018).
51. Oltman, A. E., Jervis, S. M. & Drake, M. A. Consumer attitudes and preferences for fresh market tomatoes. *J. Food Sci.* **79,** S2091–S2097 (2014).
52. Etienne, A., Génard, M., Lobit, P., Mbeguié-A-Mbéguié, D. & Bugaud, C. What controls fleshy fruit acidity? A review of malate and citrate accumulation in fruit cells. *J. Exp. Bot.* **64,** 1451–1469 (2013).
53. Cebolla-Cornejo, J. *et al.* Evaluation of genotype and environment effects on taste and aroma flavor components of Spanish fresh tomato varieties. *J. Agric. Food Chem.* **59,** 2440–2450 (2011).
54. Karppinen, K., Zoratti, L., Nguyenquynh, N., Häggman, H. & Jaakola, L. On the Developmental and environmental regulation of secondary metabolism in *Vaccinium* spp. berries. *Front. Plant Sci.* **7,** 655 (2016).
55. Blanca, J. *et al.* Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. *BMC Genomics* **16,** 257 (2015).
56. Hamilton, J. P. *et al.* Single nucleotide polymorphism discovery in cultivated tomato via sequencing by synthesis. *Plant Genome J.* **5,** 17 (2012).
57. Sim, S. C. *et al.* Development of a large snp genotyping array and generation of high-density genetic maps in tomato. *PLoS One* **7,** (2012).
58. Xu, J. *et al.* Phenotypic diversity and association mapping for fruit quality traits in cultivated tomato and related species. *Theor. Appl. Genet.* **126,** 567–581 (2013).
59. Víquez-Zamora, M. *et al.* Tomato breeding in the genomics era: Insights from a SNP array. *BMC Genomics* **14,** 354 (2013).
60. Tieman, D. M. *et al.* Identification of loci affecting flavour volatile emissions in tomato fruits. *J. Exp. Bot.* **57,** 887–896 (2006).
61. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: An R library for genome-wide

association analysis. *Bioinformatics* **23,** 1294–1296 (2007).
62. Bauchet, G. *et al.* Use of modern tomato breeding germplasm for deciphering the genetic control of agronomical traits by Genome Wide Association study. *Theor. Appl. Genet.* **130,** 875–889 (2017).
63. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol. Ecol.* **14,** 2611–2620 (2005).
64. Jombart, T. *et al.* Package ' adegenet '. *Bioinforma. Appl. Note* **24,** 1403–1405 (2008).
65. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genet.* **11,** 94 (2010).
66. Li, M. X., Yeung, J. M. Y., Cherny, S. S. & Sham, P. C. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* **131,** 747–756 (2012).
67. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19,** 807–812 (2011).
68. de Bakker, P. I. W. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17,** 122–128 (2008).
69. Pruim, R. J. *et al.* LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics* **27,** 2336–2337 (2011).
70. Fernandez-Pozo, N. *et al.* The tomato expression atlas. *Bioinformatics* **33,** 2397–2398 (2017).

Chapter 5

Chapter 5 Conclusion and Prospects

5.1 Conclusion

In this thesis, we have designed and performed innovative genomics approaches in order to deepen our understanding of the genetic control of tomato quality. In more details we developed and applied multiple haplotype-based analyses in order to benefit from footprints of selection that occurred during tomato domestication and modern breeding phases and also pioneered in introducing a meta-analysis of genome-wide association studies in tomato, the first application in a crop plant. We focused on flavor-related traits, including sugars, organic acids, amino acids and volatiles, most of the main targets for better tasting tomato.

In the multiple haplotype-based study, we defined and used haplotypes to detect the footprints of human induced selection, detected significant associations between haplotypes and traits of agronomical interests and studied haplotype structures near the associated peak SNPs. In the meta-analysis of genome-wide association studies, we benefited from genotyping imputation tools to greatly increase the genome-wide SNP coverage from SNP arrays, and performed fixed-effect and random-effect meta-analysis model to control cross-study heterogeneity. We successfully identified more than 200 new significant loci associated with a diverse set of important flavor-related sugars, organic acids, amino acids and volatiles compared to already known loci. We also brought additional knowledge regarding why modern large-fruit tomato collections have a deteriorated overall flavor, compared to cherry tomatoes and its closest wild species, *S. pimpinellifolium*. In addition, we also provided genetic clues about future tracks to improve the overall tomato flavor, especially facing the negative balance of fruit weight versus sugars and "positive volatiles" versus "negative volatiles".

5.2 Challenges and prospects

In the new breeding era, apart from continuing focusing on fruit yield and biotic and abiotic resistances/tolerances, tomato flavor is an important breeding target with an ever-increasing demand, especially from the consumers' perspective. Tomato flavor is a quite complex breeding target, which is due to the combined interactions of consumer perception, appearance, smell, taste, texture, temperature, past experience (Goff and Klee, 2006), and

fruit structure and composition of many important flavor-related sugars, organic acids, amino acids and volatiles (Tieman et al., 2012; Tieman et al., 2017; Zhu et al., 2018; Klee and Tieman, 2018; Zhao et al., 2019).

However, flavor improvement of tomato still remains an important genetic challenge of modern tomato breeding (Klee, 2010; Klee and Tieman, 2013; Klee and Tieman, 2018) for several reasons: (1) the complexity of assessing flavor. Flavor is influenced by several factors, and it is not possible to find a simple measurement to represent the impacts of all these factors; (2) significant negative correlations between flavor-related sugars and yield (Tieman et al., 2017); (3) pleasant volatiles versus unpleasant volatiles. Among all the volatiles impacting consumer preference, their concentrations, odor threshold and contributions vary, which makes it difficult to balance them; (4) low genetic diversity and large linkage disequilibrium (LD), which makes it easier to identify significant associations via GWAS but more challenging to identify the causal variant; (5) limited understanding on the biosynthesis and regulation pathways of these important flavor-related metabolites, especially volatiles (Klee and Tieman, 2013; Klee and Tieman, 2018).

Next-generation sequencing (NGS) technologies provide new opportunities to dissect the genetic architecture of tomato flavor. To date, hundreds of tomato genomes have been re-sequenced, though the sequence depths differ (The Tomato Genome Consortium, 2012; Bolger et al., 2014; Lin et al., 2014; Tieman et al., 2017; Zhu et al., 2018). SNP arrays provide a cost-effective and efficient alternative approach to generate thousands of SNPs (Hamilton et al., 2012; Sim et al., 2012; Víquez-Zamora et al., 2013), which have been successfully applied in identifying novel causal variants associated with different flavor-related traits (Sauvage et al., 2014; Bauchet et al., 2017b).

5.2.1 How to balance those positive/negative volatiles?

Even though tomato can produce over 400 volatiles, only about 30 play an important role in impacting the tomato flavor (Goff and Klee, 2006; Tieman et al., 2017; Klee and Tieman, 2018). However, the influence of volatiles is not only related to their content, but also to their odor threshold. Besides, based on the contributions to the overall liking of consumers, volatiles can be briefly divided into two groups: pleasant volatiles (positive volatiles) and unpleasant volatiles (negative volatiles). Volatiles sharing a similar structure are usually

derived from the same or close pathways, such as carotenoid, fatty acid, or phenolic pathways. In this case, the content of these volatiles are often correlated together (positively or negatively). Besides, the positive and negative volatiles are also often correlated, making it a real breeding challenge to achieve a balance between both types of volatiles. In addition, the complexity of the genetic control of volatiles makes it another big challenge to improve tomato flavor as some may have a low heritability. Nevertheless the meta-analysis (chapter 4) allowed us to identify some of the major loci to select for improving the volatile content of modern tomatoes as the beneficial alleles were lost during selection.

5.2.2 Challenges in identifying new significant genotype-phenotype associations

Identifying new causal variants or linked markers still remains one major genetic challenge of crop improvement, notably in tomato. In the third chapter, we demonstrated that haplotype-based association model (hapQTL) outperformed multi-locus mixed model (MLMM) and single-locus mixed model (EMMAX) for the identification of new associations. However, haplotype-based approach did not always outperform multi-locus mixed model, indicating that the interest of using haplotypes also depends on the phenotypes and SNPs used. Large LD decay makes it possible to identify associations using only hundreds to thousands of SNPs in this crop, instead of using GBS. However, the unevenly distributed patterns of markers as well as LD make it difficult to cover the whole genome evenly and identify new causal variants with moderate to low genetic effects and explain the missing heritability, which is especially true when using the single-marker association model. In this case, developing a new SNP array with much denser genome coverage would help to overcome these limitations.

There are some other important factors influencing the results of phenotype-genotype associations via GWAS, such as population structure and G × E interactions. For example, fruit weight is strongly correlated with population structure. Association studies focusing on fruit weight using different study panels might detect different associations (Xu et al., 2013; Ruggieri et al., 2014; Zhang et al., 2016; Bauchet et al., 2017a). Similar phenomena were also observed in linkage mapping. In addition, G × E interaction is another important factor influencing the identification of associations, especially for those traits with moderate to low heritability.

5.2.3 How to gain more from genotype imputation?

A high quality reference panel is important for genotype imputation (Marchini and Howie, 2010; Das et al., 2016). Recently, a published imputation server for the imputation of rice has been made available (Wang et al., 2018b). In this study, we have demonstrated the great benefits of genotype imputation in increasing the density of SNPs (the density of SNPs was increased to 30-50 folds, depending on the panels and SNP arrays), which can be further used for imputation-driven GWAS and meta-analysis of GWAS (Zhao et al., 2019). For example, among all 307 significant associations detected in the meta-analysis, 249 were derived from imputation. Though several hundred tomato accessions have been sequenced, the sequence quality varied depending on the consortium and materials. In order to achieve the potential benefits of imputation, a new international imputation consortium would be quite helpful in collecting, sharing and genotyping a core collection of over 500 tomato accessions, with approximately 100 wild, 200 cherry and 200 large-fruit tomatoes (which will allow to perform imputation for cases where only one type of tomatoes are used). These datasets should be deposited at public website, such as Sol Genomics Network, with free access for research purposes. Notably, in the recent pan-genome study, it has been shown that even with genome sequencing, up to about 5000 genes were still absent from the reference genome, including genes of important functions, such as *TomLoxC* (Gao et al., 2019). Higher sequence coverage of a larger core collection will be important to develop the reference imputation panel. If possible, taking those genes that are absent from the reference genome will also be helpful to improve the quality of the reference imputation panel.

Besides, the imputation quality is also influenced by several factors, such as the composition of samples and population size and structure (Schurz et al., 2019), genetic similarity (Roshyara and Scholz, 2015), marker density (Mulder et al., 2012) and MAF (minor allele frequency) (van Binsbergen et al., 2014). In addition, it is important to evaluate the effects of the size and composition (wild, cherry and big-fruit tomatoes) of the reference panel. Notably, even for the group of wild tomato (*S. pimpinellifolium*), accessions could be further divided into three single-ancestry subpopulations and four mixed-ancestry subpopulations (Lin et al., 2019). Compared to the wild tomato species, the domesticated large-fruit tomatoes have lost a lot of genetic diversity (Lin et al., 2014). A similar population structure between reference panel and studied panel will increase the overall imputation accuracy. In most GWAS cases in tomato, the studied populations usually consist of large-fruit and cherry tomatoes, with or without wild species (*S. pimpinellifolium*). If a large high-quality reference panel is available,

it will be interesting to check whether it is necessary to separate both the studies and reference panels and then impute each sub-group separately to see whether the overall imputation quality will be improved, compared to a single imputation of all accessions together.

Once a core imputation reference panel will be available, it will be interesting to test whether it is necessary to develop a new SNP array for genotyping imputation. Though there are some SNP arrays already available, such as SolCAP (Hamilton et al., 2012; Sim et al., 2012) and CBSG (Víquez-Zamora et al., 2013), imputation could only achieve about 1% of the total SNPs genotyped with GBS, based on our analyses for panel S and B (Zhao et al., 2019). Even when combining both SolCAP and CBSG SNP array, there are still many large genomic gaps uncovered (Bauchet et al., 2017a; Bauchet et al., 2017b). Though many gaps occured near the centromere regions, where there are only a few genes, gaps were also observed in other regions with many genes. In addition, SolCAP array has been shown to be less appropriate for genotyping wild tomatoes (Lin et al., 2019). This could be overcome by the use of the RADseq technique that targets the vicinity of restriction enzyme cutting sites and could provide an alternative to balance of cost-effectiveness and marker density, especially for *S. pimpinellifolium* (Chen et al., 2014; Bhakta et al., 2015; Lin et al., 2019). RADseq could increase marker density in regions with high recombination frequency and reduce marker density in regions with lower recombination frequency (Lin et al., 2019). This knowledge could be useful to design a new SNP array in order to achieve a higher imputation quality or design different SNP arrays for each subgroup. For example, the Axiom arrays (Thermos Fisher scientific) are optimally suited for applications involving 500 to 500,000 markers (https://www.snpexpert.com/Our-DNA-services/Genotyping-Arrays), which would be quite helpful. Also, SNP arrays do not target copy number variants (CNVs), which is also important for their diverse biological functions (Klopocki and Mundlos, 2011; Girirajan et al., 2011). Developing SNP arrays focusing on regions linked to a specific association could also be helpful combined with imputation, for regional fine-mapping (Schaid et al., 2018). However, reference panels used for imputation and imputation errors have to be carefully managed (Chundru et al., 2019). How to handle these problems or integrate these effects in developing the reference panel will be challenging. In addition, once both the core imputation reference panel and the SNP arrays are available, it will be very beneficial to develop a public imputation server for efficient applications. A similar public imputation server in rice (Wang et al., 2018b) is already available, which could be helpful for the future applications in tomato.

5.2.4 How to gain knowledge about tomato demographic history?

In the genomic era, especially with the benefits of molecular markers, it is possible to look at genome dynamic across time with a high resolution, especially studying which regions have undergone selection (whether it is positive or negative) and the consequences onto the genome structure. Identifying these regions is important because they provide information about the role of natural or human selection in shaping the modern crops which have been domesticated, and, in turn, could help us designing new crops to meet the increasing demand for high quality agriculture products. Nowadays, modern breeding of major crops has several major challenges, including maintaining and managing genetic diversity in breeding programs, increasing allelic diversity to adapt plant to biotic and abiotic stresses, climate change, etc. In contrast, the closest wild relatives usually harness many candidate genes that could help promote the performance of modern crops, such as resistance genes and others (Rothan et al., 2019). For example, modern tomatoes have a bad overall flavor compared to cherry tomatoes and wild species. During long-term domestication and improvement of tomato, the flavor has never been a major breeding target up to recent time, compared to yield and biotic/abiotic stress resistance where large improvement were achieved (Klee and Tieman, 2013). *TomLoxC*, a gene involved in C5 volatile biosynthesis, contributes to the desirable tomato flavor. However, this gene has been strongly negatively selected during both domestication and improvement processes (Gao et al., 2019). Identifying the regions under selection could thus help researchers and breeders identifying the promising candidate genes that can be translated into modern tomatoes.

Selection footprints can be integrated into breeding strategies from two complementary aspects. The first aspect starts from detecting selection footprints and then investigating the candidate genes within the sweeps to see whether there are some genes with important biological functions and major influence on the phenotypes and of potential breeding values. On the other hand, we can start from the target traits from a diversity panel. We can genotype the panel and then perform the GWAS and detect the selective sweeps to see if there are some overlaps between sweeps and significantly associated loci. If so, this information could guide us where to find the preferable allele to improve the targeted traits, such as via genomic editing and introgressions.

Recent positive selection can be found with three different signals: high levels of allele differentiation between populations, high frequency of the derived allele and long haplotypes

(Karlsson et al., 2014). Main methods to detect selection signals fall broadly into four categories: frequency-based methods (such as Tajima's D and derivatives, Fay & Wu's H), linkage disequilibrium-based methods (such as LRH, iHS, XP-EHH and IBD), population differentiation-based methods (such as LKT, LSBL and hapFLK) and composite methods (such as CLR, XP-CLR) (Vitti et al., 2013). However, identification of one of the footprint such as selective sweeps in major crops is only limited to certain types of selective signals, such as π and F_{st} (Verde et al., 2013; Jia et al., 2013; Qi et al., 2013; Meyer et al., 2016; Li et al., 2017; Du et al., 2018). In tomato, the selective signal studies were mainly focused on allele diversity (π) and F_{st} (Lin et al., 2014; Tieman et al., 2017; Zhu et al., 2018). For example, Plassais et al. (2019) used the cross-population composite likelihood ratio (XP-CLR) and cross-population extended haplotype homozygosity (XP-EHH) to identify selective regions in multiple dog breeds and several selective signatures were consistently significant across populations. In addition, some of the signatures were associated with important phenotypes. Wang et al. (2018a) identified a gene for green seed coat in soybean via GWAS, *G*, which only exists in wild soybeans and was significantly reduced to 4% in cultivars (**Figure 5.1a**). F_{ST}, nucleotide diversity (π), and cross-population composite likelihood ratio (XP-CLR) all showed that this gene was located within a strong selective sweep, where XP-CLR had the best performance in locating this gene (**Figure 5.1b**).

Grossman et al., (2010) developed CMS (composite of multiple signals) method to combine tests for multiple selection signals, which could increase the resolution by up to 100-fold both in simulations and real data. Alachiotis and Pavlidis, (2018) recently proposed another program RAiSD (raised accuracy in sweep detection) that composed allele diversity, site frequency spectrum and the linkage disequilibrium (LD) in the region of a sweep and was mainly designed to detect hard selective sweeps. Akbari et al., (2018) developed iSAFE (integrated selection of allele favored by evolution) to identify the favored mutation in a positive sweep. Field et al. (2016) introduced SDS (singleton density score) to infer very recent selective sweeps in human genome by comparing the ancestral and derived haplotypes. All these approaches provide new opportunities to identify the selective regions.

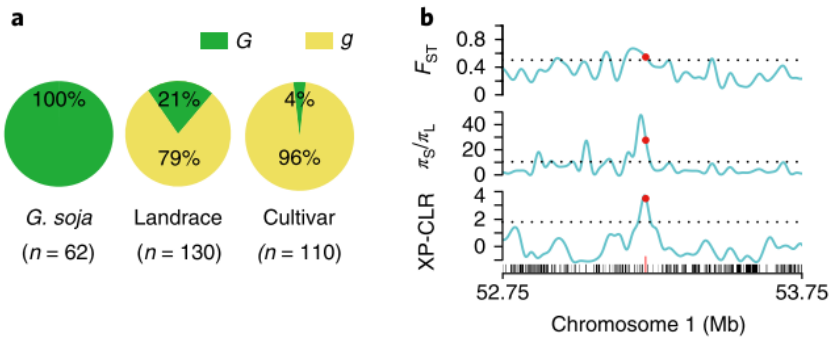


Figure 5.1 *G* is a domestication gene that contributes to soybean dormancy. a, Genotype frequency distribution of SNP1128991. **b**, F_{ST}, π, and XP-CLR values between *G. soja* (S) and the landrace (L) across the 1-Mb genomic region of the G locus. The dashed horizontal line indicates the genome-wide threshold (top 5% of the genome) of the selection signals. The bottom line indicates annotated genes in this region. The red line and dot denote the G gene—*Glyma.01G198500* (adapted from Wang et al., 2018a)

As previously mentioned, tomato has undergone long-term selection during domestication and improvement processes, during which fruit weight and biotic/abiotic resistances were among the major breeding targets. However, some important quality traits, such as tomato flavor, sugar contents have been strongly deteriorated in modern large-fruit tomatoes, compared to the wild cherry tomatoes. These results indicated that tomato has undergone both positive and negative selections during the domestication and improvement stages. Zeng et al., (2018) recently proposed a Bayesian mixed linear model (BayesS) that could distinguish negative selections ($S < 0$) from positive selections ($S > 0$) when the trait-associated variants have pleiotropic effect (**Figure 5.2**). It is important to distinguish negative selection from positive selection sweeps for all the domestication and improvement sweeps. However, the Bayesian model usually requires substantially large populations with high quality of in-depth genotyping, which could limit its potential applications in tomato at present stage.

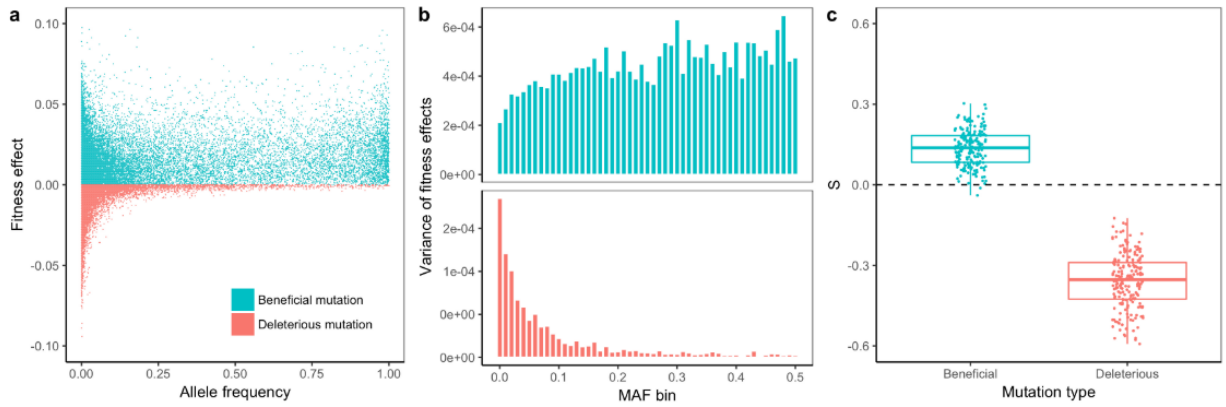


Figure 5.2 Forward simulations with mutations of direct effect on fitness to distinguish negative selections (S <0) from positive selections (S > 0). Two sets of simulations under this scenario were performed. The authors first simulated a 10Mb segment with 95% of new mutations being neutral and 5% being deleterious (or beneficial in the second set of simulations) with their effects on fitness sampled from an exponential distribution with mean 0.01 and a negative (or positive in the second set of simulations) sign. Shown are the results after 5,000 generations of selection with a constant population size of 1,000, across 200 simulation replicates for each type of selection. (a) The joint distribution of the fitness effects and allele frequencies at the causal variants. (b) The relationship between the variance of fitness effects and MAF. (c) The estimate of S based on the fitness effects in each simulation replicate. The band inside the box is the median, the bottom and top of the box are the first and third quartiles, respectively (Q1 and Q3), and the lower and upper whiskers are Q1 – 1.5 IQR and Q3 + 1.5 IQR, respectively, where IQR = Q3 – Q1 (adapted from Zeng et al., 2018).

5.2.5 How to integrate haplotypes into real tomato breeding practices?

In this thesis, we applied several haplotype-based analyses on tomato and obtained interesting results. It is still challenging to integrate all these results into real breeding practices to improve tomato quality.

- **How to choose the SNPs to define haplotypes?**

The first challenge is to define the haplotypes in a given population. To do so, high quality and density genome-wide SNPs are needed. However, how many SNPs are needed to detect the majority of most important haplotypes or how to track and use the evolution of the haplotype landscape across cycles of selection? Genomic analyses have revealed that coding genes in the tomato genome are not evenly distributed, as well as the LD patterns (The Tomato Genome Consortium, 2012). So, a well-designed SNP dataset will be crucial to define the haplotypes. The information obtained from re-sequenced accessions might be used to identify such patterns. From a more practical and cost-effective perspective, using imputation will be a good approach to greatly increase the genome-wide SNP coverage. Once thousands to millions of SNPs will be available, it will be interesting to select a core haplotype-based Tag SNPs. The core Tag SNPs should achieve a balance between harnessing

most of the haplotype diversity and the number of SNPs, in order to keep the efficiency and also a high cost-effectiveness in real practices.

- **How to choose the best haplotype combinations for breeding?**

The combination and the number of optimal haplotypes might differ depending on the genetic architecture of the target phenotype to be improved. If more than one phenotype is targeted, it might remain a big challenge to choose an optimal combination. Instead, a balanced combination of haplotypes might be more practicable, especially when the targeted phenotypes have a strong negative correlation, such as sugar content and fruit weight. In the other cases, if two traits are positively correlated, it might be easier to choose the optimal haplotype combinations to improve both traits. For the negative correlation between fruit weight and sugar content, Tieman et al (2017) suggested to increase the content of some volatiles which may increase the sweet perception without any change in sugars, as phenylacetaldehyde or phenylethanol. For these two compounds, a strong association was observed by Bauchet et al (2017) and Tieman et al (2017). In the nearby region of this associations, no major fruit weight QTLs were detected, which indicates the possibility to significantly enhance the relative contents of these two volatiles with no major impacts of fruit weight.

- **How to identify the candidate causal variants from haplotypes?**

Even when it is possible to theoretically select the optimal haplotype combinations to improve the targeted phenotypes, it still remains a major challenge to integrate them in practice. The first trying can be converting haplotypes into pseudo SNP-like markers to detect the medium to low effect associations (Meuwissen et al., 2014; Jiang et al., 2018; Karimi et al., 2018). Though in this way we might be able to identify new associations and improve the genomic prediction accuracy, we need to provide a likely physical position in order to generate the Manhattan plot in GWAS. The other limits could due to the quality control of removing less common and rare haplotypes, which could remove some useful information. Finding the most promising candidate causal variant will be very helpful in deepening our understanding on the influence of causal variants on the nearby haplotypes. For a fast and quick application of haplotypes, we can identify the candidate genes within a relatively large region. However, for many cases, the haplotype block carries more than one gene and extends to a region up to several Mbs, which makes it impossible to identify the causal mutation and clone the candidate haplotype blocks via PCR. In such case, it will be

interesting to cross two individuals with distinct haplotype lengths in the candidate haplotype block or develop a fine-mapping SNP array especially focusing on the candidate region. Cross-checking of the haplotypes focusing on the same candidate genes from multiple samples or different populations will also help in narrowing down the candidate causal haplotype regions. Also, combining the association results from different populations could also help narrowing down the list of candidate causal variant-associated haplotypes (**Figure 5.3**). However, this will be time-consuming, and restricted to the most relevant associations.

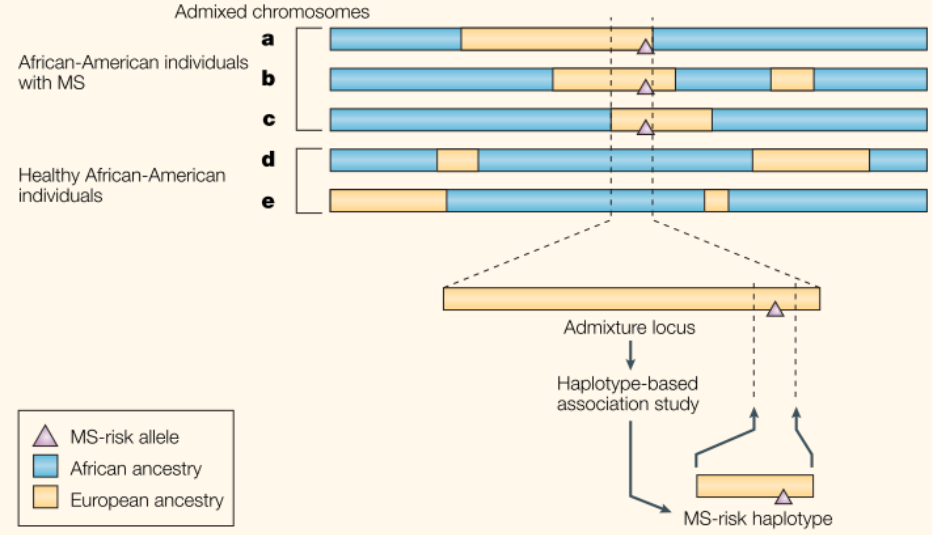


Figure 5.3 Outline of an admixture study of multiple sclerosis. Five idealized African-American chromosomes from different individuals are shown (a–e). The first three (a–c) are from patients with multiple sclerosis (MS), and the other two are from healthy control individuals (d,e). The chromosomal segments of African ancestry are shown in blue, and those of European ancestry are shown in yellow. The pink triangle depicts the position of a risk allele of European ancestry that confers susceptibility to MS. Each chromosome has a different proportion of European ancestry within the chromosomal region being examined. When the location of these European segments is compared, one smaller segment has increased European frequency relative to any other segment among the chromosomes of affected individuals but not among those of healthy control individuals. This is the admixture locus. The next phase of the analysis then relies on fine-mapping techniques, such as identifying all haplotype blocks within the admixture locus and testing each haplotype within those blocks for association with MS. This analysis will yield a disease-risk haplotype that contains the disease-risk allele and will be followed by an exhaustive assessment of all genetic variation within the risk haplotype to determine which allele is the risk allele (adapted from Hafler and Jager, 2005).

➢ How to use haplotypes in marker-assisted selection and genomic selection?

Depending on the targeted phenotypes, some markers can also be possibly selected from the tag SNPs derived from haplotypes, and be used for marker-assisted selection (MAS) or genomic prediction. In tomato, Duangjit et al., (2016) has demonstrated that about 2300 independent SNPs that were distributed evenly in the genome were enough for trait prediction and there was no significant additional gains when increasing the number of markers. This is consistent with our results and indicates that selecting the tag SNPs from

haplotypes will have potential benefits in achieving similar or even increasing prediction accuracy with fewer markers.

Nowadays, it is possible to take the G × E interactions into account in the prediction models (Jiang and Reif, 2015; Lado et al., 2016; Cuevas et al., 2017; Jiang et al., 2018; Millet et al., 2019). Similarly, we can also take the G × E interactions in the haplotype-based prediction models. This can be achieved by treating the pseudo SNPs as regular SNPs and adapt all the steps that are based on single SNP-based models. Also, we can take the significant associations as co-factors in the prediction models, which will be helpful in either investigating the effects of the significant associations or improving the prediction accuracy (Yamamoto et al., 2016).

5.2.6 How to calculate the heritability based on summary GWAS data?

Finding the missing heritability still remains a main challenge in human genetics and also crop plants (Manolio et al., 2009; Eichler et al., 2010; Brachi et al., 2011; Speed et al., 2017; Yang et al., 2017). While it is straight forward to estimate the heritability of a trait from single GWAS experiment, doing so from meta-analysis outputs is not trivial. Nowadays, some methods have been developed to calculate the heritability based on the summary data from GWAS. For example, Yang et al., (2011) developed GCTA (Genome-wide complex trait analysis) to estimate the proportion of phenotypic variance explained by all markers, which has been further extended for many other analyses. Finucane et al. (2015) developed LDSC (LD score) that can partition heritability by functional annotation using GWA summary statistics. Shi et al. (2016) developed HESS (heritability estimation from summary statistics) to estimate and visualize the local SNP-heritability. Speed and Balding, (2019) recently developed SumHer for estimating SNP-based heritability, which outperformed LDSC by allowing the user to specify the heritability model to calculate the heritability. However, most of these approaches require a large population size and high density of SNPs, which is still difficult to apply in tomato, even after genotype imputation. However, with the fast development of NGS, it should become soon possible to apply these new approaches to investigate the missing heritability of major tomato quality traits. While genotyping and analyzing large scale dataset becomes routine, the main bottleneck comes from our ability to measure phenotype on such a large scale with a higher degree of accuracy.

5.2.7 Including new GWAS datasets for meta-analysis of GWAS

In chapter 4, we have demonstrated the benefits of meta-analysis of GWAS in identifying new potential causal variants. With the increasing popularity of GWAS, new GWAS datasets are expected to be available, and should be included in a new meta-GWAS analysis . It will be quite helpful to develop an open-public service (FAIR, findable, available, accessible and reusable, https://www.go-fair.org/fair-principles/) to 1) store all the summary GWAS datasets for each panel, 2) submit new GWAS datasets, 3) perform new meta-analysis, 4) generate plots for some important results, such as Manhattan plot, Q-Q plot, cross-study heterogeneity, forest plot, Z-M plot and 5) integrate these results with other available datasets, such as Sol Genomic Network, Tomato Express Atlas, TomExpress and others.

It will be interesting to develop new programs specific to single-haplotype-based and multi-haplotype-based mixed models and apply them to major crops, including tomato. If the program can directly take regular SNPs as genotypic inputs into the haplotype-based association models, then it will also be interesting to perform the meta-analysis of GWAS based on haplotypes. However, to do so, the raw genotypes and phenotypes from each GWAS panels should be available, and re-analyse the haplotype-based associations, in order to avoid heterogeneity caused by association models, which could be challenging in some cases of data sharing and computation.

5.2.8 How to integrate these achievements to improve tomato flavor?

The breeding success of improving tomato quality will strongly depend on the main breeding targets to follow and on available technologies. If the main breeding purpose for fresh market tomatoes is to enhance overall flavor, then multiple flavor-related metabolites should be targeted, such as sugars, acids and volatiles. To do so, a deep understanding of the metabolic pathways and regulations is needed and more genes regulating these pathways should be identified and selected or modified in order to achieve a balanced overall flavor. Besides, consumers from different backgrounds might have different preferences. For example, in China, pink tomatoes are much more popular than red tomatoes, which was mainly due to a mutation in a major gene, *myb12*. However, up to 122 metabolites are significantly altered between red and pink tomatoes (Zhu et al., 2018). Using appropriate methodology, though it is possible to statistically model the flavor chemical composition of an average ‘ideal’ fruit averaged over the sampled consumer population, this ideal fruit may not be the most liked by every individual (Klee and Tieman, 2018). A diversity of proposed tastes and texture is

needed. New technologies apart from traditional genetic tools for breeding could help promote the fast breeding of tomato, such as genome engineering and genome editing. For example, gene editing targeting at *SELF-PRUNING 5G* (*SP5G*) resulted in a quick burst of flower and demonstrated the power to rapidly improve tomato yield (Soyk et al., 2017). Genome editing technologies such as CRISPR-Cas9 can introduce desirable traits, such as traits associated morphology, flower, fruit production and ascorbic acid synthesis, into stress-tolerant wild tomato accessions and also retain the disease resistance and salt tolerance (Li et al., 2018). These results demonstrate the great potentials of genomic editing in the new breeding stage 4 (ideotype-based selection and transformation) (Wallace et al., 2018; Ramstein et al., 2019). In addition, directed evolution-genome editing (DE-GE) could also be very useful in improving tomato yield and important metabolites (**Figure 5.4**). Unfortunately, these technologies will be difficult to be applied in Europe as they are considered as genetically modified organisms (GMOs) and also due to the rapid turnover of tomato varieties. However, this will not be a major limit for those international breeding companies and those research institutes outside of Europe. Another limit comes from the introgression of mutations or resistance genes from wild species, which may introduce some unfavorable effects for other quality related traits. For example, most modern cultivars contain the *uniform* (*u*) mutation, which turns wild tomatoes with a green shoulder to more uniformly red (which are more attractive). However, this mutation reduces the contents of chloroplasts, carotenoids and soluble solids, all of which contribute to the tomato flavor (Powell et al., 2012; Klee and Tieman, 2013). Finally, even if breeders create tomato varieties with good flavor, the whole chain will have to be modified if the new varieties require more strict growth and post-harvest conditions and management.

5.3 Final conclusion

Overall, there are still several aspects that are of great interest and promising in the future of tomato genetic and genomic studies, though they might seem challenging at the current status. These prospects, together with what we have shown, should benefit the modern breeding of tomato to feed the increasing global population with better, nutritious and health-promoting tomatoes.

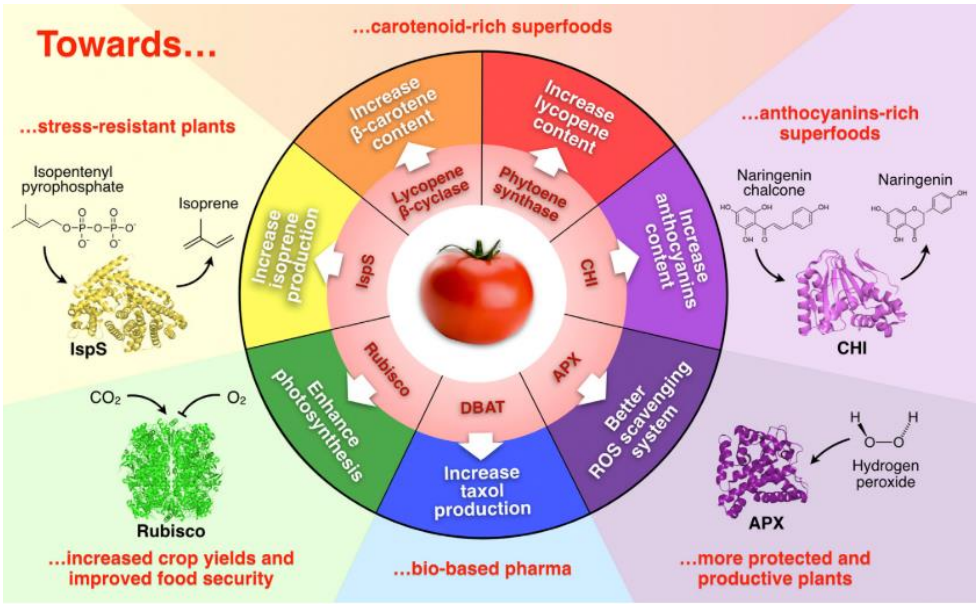


Figure 5.4 Plant Pathways Where Directed Evolution-Genome Editing (DE-GE) May Be Usefully Applied. The figure summarizes examples of how DE-GE could help to increase the yield of useful secondary metabolites or to improve primary metabolism in plants, as detailed in the text. Working on lycopene b-cyclase, phytoene synthase, and chalcone isomerase (CHI) may help to increase plant production of b-carotene, lycopene, and anthocyanins, respectively. Increasing the catalytic activity of isoprene synthase (IspS) and 10-deacetylbaccatin III-10-b-O- acetyltransferase (DBAT) may trigger isoprene and taxol production, respectively. Enhancing the affinity/resistance of ascorbate peroxidase (APX) to hydrogen peroxide will increase plant resilience to stress, and Rubisco with improved carboxylation properties is predicted to increase agricultural yield (adapted from Gionfriddo et al., 2019).

References

Akbari, A., Vitti, J. J., Iranmehr, A., Bakhtiari, M., Sabeti, P. C., Mirarab, S., and Bafna, V. (2018). Identifying the favored mutation in a positive selective sweep. *Nat. Methods* **15**:279–282.

Alachiotis, N., and Pavlidis, P. (2018). RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Commun. Biol.* **1**:79.

Bauchet, G., Grenier, S., Samson, N., Bonnet, J., Grivet, L., and Causse, M. (2017a). Use of modern tomato breeding germplasm for deciphering the genetic control of agronomical traits by Genome Wide Association study. *Theor. Appl. Genet.* **130**:875–889.

Bauchet, G., Grenier, S., Samson, N., Segura, V., Kende, A., Beekwilder, J., Cankar, K., Gallois, J.-L., Gricourt, J., Bonnet, J., et al. (2017b). Identification of major loci and genomic regions controlling acid and volatile content in tomato fruit: implications for flavor improvement. *New Phytol.* **215**:624–641.

Bhakta, M. S., Jones, V. A., and Vallejos, C. E. (2015). Punctuated distribution of recombination hotspots and demarcation of pericentromeric regions in *Phaseolus vulgaris* L. *PLoS One* **10**:e0116822.

Bolger, A., Scossa, F., Bolger, M. E., Lanz, C., Maumus, F., Tohge, T., Quesneville, H., Alseekh, S., Sørensen, I., Lichtenstein, G., et al. (2014). The genome of the stress-tolerant wild tomato species *Solanum pennellii*. *Nat. Genet.* **46**:1034–1038.

Brachi, B., Morris, G. P., and Borevitz, J. O. (2011). *Genome-wide association studies in plants: The missing heritability is in the field*. BioMed Central.

Chen, A.-L., Liu, C.-Y., Chen, C.-H., Wang, J.-F., Liao, Y.-C., Chang, C.-H., Tsai, M.-H., Hwu, K.-K., and Chen, K.-Y. (2014). Reassessment of QTLs for Late Blight Resistance in the Tomato Accession L3708 Using a Restriction Site Associated DNA (RAD) Linkage Map and Highly Aggressive Isolates of *Phytophthora infestans*. *PLoS One* **9**:e96417.

Chundru, V. K., Marioni, R. E., Prendergast, J. G. D., Vallerga, C. L., Lin, T., Beveridge, A. J., Gratten, J., Hume, D.

A., Deary, I. J., Wray, N. R., et al. (2019). Examining the Impact of Imputation Errors on Fine-Mapping Using DNA Methylation QTL as a Model Trait. *Genetics* **212**:577–586.

Cuevas, J., Crossa, J., Montesinos-López, O. A., Burgueño, J., Perez-Roudriguez, P., and De Los Campos, G. (2017). Bayesian genomic prediction with genotype x environment interaction kernel models. *G3 Genes|Genomes|Genetics* **7**:41–53.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. *Nat. Genet.* **48**:1284–1287.

Du, X., Huang, G., He, S., Yang, Z., Sun, G., Ma, X., Li, N., Zhang, X., Sun, J., Liu, M., et al. (2018). Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nat. Genet.* **50**:1–7.

Duangjit, J., Causse, M., and Sauvage, C. (2016). Efficiency of genomic selection for tomato fruit quality. *Mol. Breed.* **36**:36:29.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**:446–450.

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M. I., et al. (2016). Detection of human adaptation during the past 2000 years. *Science (80-.).* **354**:760–764.

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R. R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**:1228–1235.

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., Burzynski-Chang, E. A., Fish, T. L., Stromberg, K. A., Sacks, G. L., et al. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* **51**:1044–1051.

Gionfriddo, M., De Gara, L., and Loreto, F. (2019). Directed Evolution of Plant Processes: Towards a Green (r)Evolution? *Trends Plant Sci.* Advance Access published October 2019, doi:10.1016/j.tplants.2019.08.004.

Girirajan, S., Campbell, C. D., and Eichler, E. E. (2011). Human Copy Number Variation and Complex Genetic Disease. *Annu. Rev. Genet.* **45**:203–226.

Goff, S. A., and Klee, H. J. (2006). Plant volatile compounds: Sensory cues for health and nutritional value? *Science (80-.).* **311**:815–819.

Grossman, S. R., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., et al. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science (80-.).* **327**:883–886.

Hafler, D. A., and Jager, P. L. De (2005). Applying a new generation of genetic maps to understand human inflammatory disease. *Nat. Rev. Immunol.* **5**:83–91.

Hamilton, J. P., Sim, S.-C., Stoffel, K., Van Deynze, A., Buell, C. R., and Francis, D. M. (2012). Single nucleotide polymorphism discovery in cultivated tomato via sequencing by synthesis. *Plant Genome J.* **5**:17.

Jia, G., Huang, X., Zhi, H., Zhao, Y., Zhao, Q., Li, W., Chai, Y., Yang, L., Liu, K., Lu, H., et al. (2013). A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). *Nat. Genet.* **45**:957–961.

Jiang, Y., and Reif, J. C. (2015). Modeling Epistasis in Genomic Selection. *Genetics* **201**:759–68.

Jiang, Y., Schmidt, R. H., and Reif, J. C. (2018). Haplotype-based genome-wide prediction models exploit local epistatic interactions among markers. *G3 Genes|Genomes|Genetics* **8**:g3.300548.2017.

Karimi, Z., Sargolzaei, M., Robinson, J. A. B., and Schenkel, F. S. (2018). Assessing haplotype-based models for genomic evaluation in holstein cattle. *Can. J. Anim. Sci.* **98**:750–759.

Karlsson, E. K., Kwiatkowski, D. P., and Sabeti, P. C. (2014). Natural selection and infectious disease in human populations. *Nat. Rev. Genet.* **15**:379–393.

Klee, H. J. (2010). Improving the flavor of fresh fruits: genomics, biochemistry, and biotechnology. *New Phytol.* **187**:44–56.

Klee, H. J., and Tieman, D. M. (2013). Genetic challenges of flavor improvement in tomato. *Trends Genet.* **29**:257–262.

Klee, H. J., and Tieman, D. M. (2018). The genetics of fruit flavour preferences. *Nat. Rev. Genet.* **19**:347–356.

Klopocki, E., and Mundlos, S. (2011). Copy-Number Variations, Noncoding Sequences, and Human Phenotypes. *Annu. Rev. Genomics Hum. Genet.* **12**:53–72.

Lado, B., Barrios, P. G., Quincke, M., Silva, P., and Gutiérrez, L. (2016). Modeling genotype × Environment interaction for genomic selection with unbalanced data from a wheat breeding program. *Crop Sci.* **56**:2165–2179.

Li, L. F., Li, Y. L., Jia, Y., Caicedo, A. L., and Olsen, K. M. (2017). Signatures of adaptation in the weedy rice genome.

Nat. Genet. **49**:811–814.

Li, T., Yang, X., Yu, Y., Si, X., Zhai, X., Zhang, H., Dong, W., Gao, C., and Xu, C. (2018). Domestication of wild tomato is accelerated by genome editing. *Nat. Biotechnol.* **36**:1160–1163.

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S., Wang, X., et al. (2014). Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.* **46**:1220–1226.

Lin, Y.-P., Liu, C.-Y., and Chen, K.-Y. (2019). Assessment of Genetic Differentiation and Linkage Disequilibrium in Solanum pimpinellifolium Using Genome-Wide High-Density SNP Markers. *G3 Genes|Genomes|Genetics* **9**:g3.200862.2018.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461**:747–753.

Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**:499–511.

Meuwissen, T. H. E., Odegard, J., Andersen-Ranberg, I., and Grindflek, E. (2014). On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. *Genet. Sel. Evol.* **46**.

Meyer, R. S., Choi, J. Y., Sanches, M., Plessis, A., Flowers, J. M., Amas, J., Dorph, K., Barretto, A., Gross, B., Fuller, D. Q., et al. (2016). Domestication history and geographical adaptation inferred from a SNP map of African rice. *Nat. Genet.* **48**:1083–1088.

Millet, E. J., Kruijer, W., Coupel-Ledru, A., Alvarez Prado, S., Cabrera-Bosquet, L., Lacube, S., Charcosset, A., Welcker, C., van Eeuwijk, F., and Tardieu, F. (2019). Genomic prediction of maize yield across European environmental conditions. *Nat. Genet.* **51**:952–956.

Mulder, H. A., Calus, M. P. L., Druet, T., and Schrooten, C. (2012). Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. *J. Dairy Sci.* **95**:876–889.

Plassais, J., Kim, J., Davis, B. W., Karyadi, D. M., Hogan, A. N., Harris, A. C., Decker, B., Parker, H. G., and Ostrander, E. A. (2019). Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nat. Commun.* **10**:1489.

Powell, A. L. T., Nguyen, C. V., Hill, T., Cheng, K. L. L., Figueroa-Balderas, R., Aktas, H., Ashrafi, H., Pons, C., Fernández-Muñoz, R., Vicente, A., et al. (2012). Uniform ripening encodes a *Golden 2-like* transcription factor regulating tomato fruit chloroplast development. *Science (80-.).* **336**:1711–1715.

Qi, J., Liu, X., Shen, D., Miao, H., Xie, B., Li, X., Zeng, P., Wang, S., Shang, Y., Gu, X., et al. (2013). A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat. Genet.* **45**:1510–1515.

Ramstein, G. P., Jensen, S. E., and Buckler, E. S. (2019). Breaking the curse of dimensionality to identify causal variants in Breeding 4. *Theor. Appl. Genet.* **132**:559–567.

Roshyara, N. R., and Scholz, M. (2015). Impact of genetic similarity on imputation accuracy. *BMC Genet.* **16**.

Rothan, C., Diouf, I., and Causse, M. (2019). Trait discovery and editing in tomato. *Plant J.* **97**:73–90.

Ruggieri, V., Francese, G., Sacco, A., Alessandro, A. D., Rigano, M. M., Parisi, M., Milone, M., Cardi, T., Mennella, G., and Barone, A. (2014). An association mapping approach to identify favourable alleles for tomato fruit quality breeding. *BMC Plant Biol.* **14**:1–15.

Sauvage, C., Segura, V., Bauchet, G., Stevens, R., Do, P. T., Nikoloski, Z., Fernie, A. R., and Causse, M. (2014). Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant Physiol.* **165**:1120–1132.

Schaid, D. J., Chen, W., and Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**:491–504.

Schurz, H., Müller, S. J., Van Helden, P. D., Tromp, G., Hoal, E. G., Kinnear, C. J., and Möller, M. (2019). Evaluating the accuracy of imputation methods in a five-way admixed population. *Front. Genet.* **10**.

Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am. J. Hum. Genet.* **99**:139–153.

Sim, S.-C., Durstewitz, G., Plieske, J., Wieseke, R., Ganal, M. W., van Deynze, A., Hamilton, J. P., Buell, C. R., Causse, M., Wijeratne, S., et al. (2012). Development of a large snp genotyping array and generation of high-density genetic maps in tomato. *PLoS One* **7**.

Soyk, S., Müller, N. A., Park, S. J., Schmalenbach, I., Jiang, K., Hayama, R., Zhang, L., Van Eck, J., Jiménez-Gómez, J. M., and Lippman, Z. B. (2017). Variation in the flowering gene *SELF PRUNING 5G* promotes day-neutrality and early yield in tomato. *Nat. Genet.* **49**:162–168.

Speed, D., and Balding, D. J. (2019). SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* **51**:277–284.

Speed, D., Cai, N., Johnson, M. R., Nejentsev, S., and Balding, D. J. (2017). Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**:986–992.

The Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**:635–641.

Tieman, D., Bliss, P., McIntyre, L. M. M., Blandon-Ubeda, A., Bies, D., Odabasi, A. Z. Z., Rodríguez, G. R. R., Van Der Knaap, E., Taylor, M. G. G., Goulet, C., et al. (2012). The chemical interactions underlying tomato flavor preferences. *Curr. Biol.* **22**:1035–1039.

Tieman, D., Zhu, G., Resende, M. F. R., Lin, T., Nguyen, C., Bies, D., Rambla, J. L., Beltran, K. S. O., Taylor, M., Zhang, B., et al. (2017). A chemical genetic roadmap to improved tomato flavor. *Science (80-.).* **355**:391–394.

van Binsbergen, R., Bink, M. C., Calus, M. P., van Eeuwijk, F. A., Hayes, B. J., Hulsegge, I., and Veerkamp, R. F. (2014). Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genet. Sel. Evol.* **46**:41.

Verde, I., Abbott, A. G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., Zhebentyayeva, T., Dettori, M. T., Grimwood, J., Cattonaro, F., et al. (2013). The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* **45**:487–494.

Víquez-Zamora, M., Vosman, B., van de Geest, H., Bovy, A., Visser, R. G., Finkers, R., and van Heusden, A. W. (2013). Tomato breeding in the genomics era: Insights from a SNP array. *BMC Genomics* **14**:354.

Vitti, J. J., Grossman, S. R., and Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annu. Rev. Genet.* **47**:97–120.

Wallace, J. G., Rodgers-Melnick, E., and Buckler, E. S. (2018). On the Road to Breeding 4.0: Unraveling the Good, the Bad, and the Boring of Crop Quantitative Genomics. *Annu. Rev. Genet.* **52**:421–444.

Wang, M., Li, W., Fang, C., Xu, F., Liu, Y., Wang, Z., Yang, R., Zhang, M., Liu, S., Lu, S., et al. (2018a). Parallel selection on a dormancy gene during domestication of crops from multiple families. *Nat. Genet.* **50**:1435–1441.

Wang, D. R., Agosto-Pérez, F. J., Chebotarov, D., Shi, Y., Marchini, J., Fitzgerald, M., McNally, K. L., Alexandrov, N., and McCouch, S. R. (2018b). An imputation platform to enhance integration of rice genetic resources. *Nat. Commun.* **9**:3519.

Xu, J., Ranc, N., Muños, S., Rolland, S., Bouchet, J.-P. P., Desplat, N., Le Paslier, M.-C. C., Liang, Y., Brunel, D., and Causse, M. (2013). Phenotypic diversity and association mapping for fruit quality traits in cultivated tomato and related species. *Theor. Appl. Genet.* **126**:567–581.

Yamamoto, E., Matsunaga, H., Onogi, A., Kajiya-Kanegae, H., Minamikawa, M., Suzuki, A., Shirasawa, K., Hirakawa, H., Nunome, T., Yamaguchi, H., et al. (2016). A simulation-based breeding design that uses whole-genome prediction in tomato. *Sci. Rep.* **6**:19454.

Yamamoto, E., Matsunaga, H., Onogi, A., Ohyama, A., Miyatake, K., Yamaguchi, H., Nunome, T., Iwata, H., and Fukuoka, H. (2017). Efficiency of genomic selection for breeding population design and phenotype prediction in tomato. *Heredity (Edinb).* **118**:202–209.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**:76–82.

Yang, J., Zeng, J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2017). Concepts, estimation and interpretation of SNP-based heritability. *Nat. Genet.* **49**:1304–1310.

Zeng, J., De Vlaming, R., Wu, Y., Robinson, M. R., Lloyd-Jones, L. R., Yengo, L., Yap, C. X., Xue, A., Sidorenko, J., McRae, A. F., et al. (2018). Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**:746–753.

Zhang, J., Zhao, J., Liang, Y., and Zou, Z. (2016). Genome-wide association-mapping for fruit quality traits in tomato. *Euphytica* **207**:439–451.

Zhao, J., Sauvage, C., Zhao, J., Bitton, F., Bauchet, G., Liu, D., Huang, S., Tieman, D. M., Klee, H. J., and Causse, M. (2019). Meta-analysis of genome-wide association studies provides insights into genetic control of tomato flavor. *Nat. Commun.* **10**:1534.

Zhu, G., Wang, S., Huang, Z., Zhang, S., Liao, Q., Zhang, C., Lin, T., Qin, M., Peng, M., Yang, C., et al. (2018). Rewiring of the fruit metabolome in tomato breeding. *Cell* **172**:249-261.e12.

Appendix 1

Appendix 1 Supplementary tables and figures related to Chapter 3

Note: in order to reduce the space of too-long information, we only provide the first 10 lines as examples. All data will be available once the manuscript is published.

Table S1. Haplotype blocks estimation based on 163 tomato accessions genotyped with 5995 SNPs (only the first 10 lines were provided).

Chr	BP1	BP2	KB	No. SNPS	SNPS
1	336495	348201	11.707	3	rs01_336495|rs01_338651|rs01_348201
1	534448	541692	7.245	2	rs01_534448|rs01_541692
1	663658	859321	195.664	6	rs01_663658|rs01_669686|rs01_678452|rs01_730154|rs01_853700|rs01_859321
1	1116398	1118868	2.471	2	rs01_1116398|rs01_1118868
1	1508203	1524332	16.13	3	rs01_1508203|rs01_1511380|rs01_1524332
1	2200864	2371872	171.009	5	rs01_2200864|rs01_2255882|rs01_2311631|rs01_2311760|rs01_2371872
1	2441831	2446729	4.899	8	rs01_2441831|rs01_2442427|rs01_2444487|rs01_2445620|rs01_2445949|rs01_2446381|rs01_2446525|rs01_2446729
1	2458075	2461631	3.557	2	rs01_2458075|rs01_2461631
1	2554370	2751369	197	5	rs01_2554370|rs01_2572815|rs01_2574889|rs01_2581713|rs01_2751369
1	5340027	5364660	24.634	3	rs01_5340027|rs01_5346991|rs01_5364660
…					

Table S2. Haplotype blocks estimation based on 28 large-fruited tomato accessions genotyped with 5995 SNPs (only the first 10 lines were provided).

Chr	BP1	BP2	KB	No. SNPS	SNPS
1	259974	303202	43.229	7	rs01_259974|rs01_299550|rs01_299696|rs01_301597|rs01_301603|rs01_303076|rs01_303202
1	663658	853700	190.043	5	rs01_663658|rs01_669686|rs01_678452|rs01_730154|rs01_853700
1	2255882	2371872	115.991	4	rs01_2255882|rs01_2311631|rs01_2311760|rs01_2371872
1	2554370	2581713	27.344	3	rs01_2554370|rs01_2574889|rs01_2581713
1	53643429	53643536	53643429	2	rs01_53643429|rs01_53643536
1	80185036	80186340	80185036	4	rs01_80185036|rs01_80185248|rs01_80185807|rs01_80186340
1	80286023	80485539	80286023	4	rs01_80286023|rs01_80289583|rs01_80485031|rs01_80485539
1	82598325	82785510	82598325	5	rs01_82598325|rs01_82599713|rs01_82600039|rs01_82722469|rs01_82785510
1	83802297	83816618	83802297	2	rs01_83802297|rs01_83816618
1	85134616	85150066	85134616	5	rs01_85134616|rs01_85136490|rs01_85141065|rs01_85144858|rs01_85150066
…					

Table S3. Haplotype blocks estimation based on 119 cherry tomato accessions genotyped with 5995 SNPs (only the first 10 lines were provided).

Chr	BP1	BP2	KB	No. SNPS	SNPS
1	336495	348201	11.707	3	rs01_336495|rs01_338651|rs01_348201
1	534448	541692	7.245	2	rs01_534448|rs01_541692
1	663658	853700	190.043	5	rs01_663658|rs01_669686|rs01_678452|rs01_730154|rs01_853700
1	1508203	1524332	16.13	3	rs01_1508203|rs01_1511380|rs01_1524332
1	2458075	2461631	3.557	2	rs01_2458075|rs01_2461631
1	2507736	2581713	73.978	5	rs01_2507736|rs01_2554370|rs01_2572815|rs01_2574889|rs01_2581713
1	26851884	26852331	0.448	2	rs01_26851884|rs01_26852331
1	35295335	35298571	3.237	2	rs01_35295335|rs01_35298571
1	78355519	78357558	2.04	2	rs01_78355519|rs01_78357558
1	78486259	78500502	14.244	2	rs01_78486259|rs01_78500502
…					

Table S4. Haplotype blocks estimation based on 16 wild tomato accessions genotyped with 5995 SNPs (only the first 10 lines were provided).

Chr	BP1	BP2	KB	No. SNPS	SNPS
1	299696	301603	1.908	3	rs01_299696|rs01_301597|rs01_301603
1	84767805	84768069	0.265	3	rs01_84767805|rs01_84767919|rs01_84768069
2	24677141	24707601	30.46	4	rs02_24677141|rs02_24677229|rs02_24694492|rs02_24707601
2	35214875	35269874	55	4	rs02_35214875|rs02_35214970|rs02_35269758|rs02_35269874
2	36790187	36794605	4.419	5	rs02_36790187|rs02_36790206|rs02_36790680|rs02_36790779|rs02_36794605
2	40332127	40332938	0.812	4	rs02_40332127|rs02_40332187|rs02_40332255|rs02_40332938
2	41412660	41412673	0.014	2	rs02_41412660|rs02_41412673
2	41427320	41427791	0.472	3	rs02_41427320|rs02_41427572|rs02_41427791
2	41787643	41787849	0.207	2	rs02_41787643|rs02_41787849
2	41966641	41967490	0.85	2	rs02_41966641|rs02_41967490
…					

Table S5. List of positive selective sweeps identified by using integrated haplotype score (iHS).

No.	Chr	Position	iHS	P	Start	End	Size
PSS01	1	78579761	2.93	3.41×10^{-03}	76879885	80451423	3571538
PSS02	1	87082281	-2.92	3.53×10^{-03}	85480347	87987834	2507487
PSS03	2	39617885	2.58	9.80×10^{-03}	39515971	39806330	290359
PSS04	2	41431612	3.19	1.40×10^{-03}	41409922	41458782	48859
PSS05	2	45761358	2.75	6.02×10^{-03}	45757485	45785075	27590
PSS06	2	46767771	3.07	2.11×10^{-03}	46765116	46777030	11913
PSS07	2	47204573	2.71	6.64×10^{-03}	47196668	47216586	19916
PSS08	3	4934033	3.08	2.05×10^{-03}	4436828	5656642	1219814
PSS09	3	56892686	-3.28	1.05×10^{-03}	43908591	62589864	18681273
PSS10	3	65012271	2.64	8.28×10^{-03}	64077021	66016878	1939857
PSS11	4	55727428	3.15	1.64×10^{-03}	55315682	55751257	435575
PSS12	5	7473754	5.41	6.42×10^{-08}	7405325	7553523	148198
PSS13	5	61500171	3.97	7.05×10^{-05}	59961746	61500228	1538482
PSS14	6	34732308	-3.07	2.13×10^{-03}	33100482	37372440	4271958
PSS15	6	39539433	3.51	4.47×10^{-04}	39197248	40546914	1349666
PSS16	7	3360966	-5.94	2.84×10^{-09}	1340132	6021290	4681157
PSS17	7	62755438	2.8	5.08×10^{-03}	57514198	65574391	8060193
PSS18	8	2645233	3.64	2.71×10^{-04}	378191	7741110	7362919
PSS19	8	54296663	-3.64	2.74×10^{-04}	53079559	55028674	1949115
PSS20	9	2673326	2.68	7.37×10^{-03}	2543318	3903235	1359917
PSS21	9	62209136	2.9	3.68×10^{-03}	59154740	65830455	6675715
PSS22	10	3991642	3.34	8.31×10^{-04}	1461758	9768513	8306756
PSS23	10	58189616	3.68	2.37×10^{-04}	58002494	58262063	259569
PSS24	11	51352658	2.85	4.38×10^{-03}	51016418	51507052	490635

Table S6. List of Genes within the positive selective sweeps identified by using integrated haplotype score (iHS) (only the first 10 lines were provided).

No.	Chr	Locus name	Gene start	Gene end	Candidates
PSS01	1	Solyc01g067900	76923481	76924353	Rapid alkalinization factor 2
PSS01	1	Solyc01g067910	76928803	76929208	Unknown Protein
PSS01	1	Solyc01g067920	76930648	76932644	Protein FAR1-RELATED SEQUENCE 5
PSS01	1	Solyc01g067930	76941648	76943047	Alpha-1 6-xylosyltransferase
PSS01	1	Solyc01g067940	76942454	76943451	Unknown Protein
PSS01	1	Solyc01g067950	76969993	76970214	Unknown Protein
PSS01	1	Solyc01g067960	77002670	77002966	Unknown Protein
PSS01	1	Solyc01g067970	77009619	77009915	Unknown Protein
PSS01	1	Solyc01g067980	77028046	77031990	Protein phosphatase-2C
PSS01	1	Solyc01g067990	77038390	77039288	Transducin family protein
…					

Table S7. Putative domestication sweeps (only the first lines were provided).

Domestication sweep	Chr	start	end	π_{PIM}	π_{CER}	$\pi_{PIM/CER}$
DS001	1	5630001	5920000	4.84E-06	9.85E-07	4.913705584
DS002	1	61510001	61700000	4.17E-06	1.06E-06	3.933962264
DS003	1	69820001	70010000	4.66E-06	9.07E-07	5.137816979
DS004	1	75250001	75460000	5.08E-06	8.28E-07	6.1352657
DS005	1	75620001	81980000	5.08E-06	5.88E-07	8.639455782
DS006	1	77710001	77810000	5.08E-06	9.07E-07	5.600882029
DS007	1	81790001	81890000	4.98E-06	1.44E-06	3.458333333
DS008	1	82790001	82980000	5.08E-06	1.29E-06	3.937984496
DS009	1	83990001	84180000	5.08E-06	7.49E-07	6.782376502
…						

Table S8. Putative improvement sweeps (only the first lines were provided).

Improvement sweeps	Chr	start	end	π_{BIG}	π_{CER}	$\pi_{CER/BIG}$
IS001	1	1240001	1430000	3.23E-07	5.18E-06	16.05798234
IS002	1	77610001	79660000	6.35E-07	4.42E-06	6.965181969
IS003	1	82990001	83180000	3.23E-07	2.00E-06	6.19999318
IS004	1	83320001	83430000	3.23E-07	2.51E-06	7.780991441
IS005	1	83820001	83920000	3.23E-07	2.00E-06	6.19999318
IS006	1	84710001	84830000	6.35E-07	4.54E-06	7.154281932
IS007	1	90580001	90780000	6.35E-07	4.42E-06	6.965181969
IS008	1	92380001	92520000	1.27E-06	1.01E-05	7.957956775
IS009	1	93660001	93800000	6.35E-07	4.42E-06	6.965181969
…						

Table S9. Genes within the putative domestication sweeps (only the first lines were provided).

No.	Chr	Locus name	Gene start	Gene end	Candidates
DS001	1	Solyc01g010630	5634522	5639332	Ulp1 protease family C-terminal catalytic domain containing protein
DS001	1	Solyc01g010640	5646161	5647113	Uncharacterized membrane protein
DS001	1	Solyc01g010650	5655290	5660901	UDP-galactose transporter 3
DS001	1	Solyc01g010660	5673997	5680626	Receptor-like protein kinase At3g21340
DS001	1	Solyc01g010670	5692071	5693406	Unknown Protein
DS001	1	Solyc01g010680	5698129	5701473	Genomic DNA chromosome 5 P1 clone MRD20
DS001	1	Solyc01g010690	5704672	5706189	Polyvinylalcohol dehydrogenase
DS001	1	Solyc01g010700	5726264	5728807	AKIN gamma
DS001	1	Solyc01g010710	5740284	5746099	Serine carboxypeptidase 1
DS001	1	Solyc01g010720	5747143	5750774	Serine carboxypeptidase K10B2.2
…					

Table S10. Genes within the putative improvement sweeps (only the first lines were provided).

No.	Chr	Locus name	Gene start	Gene end	Candidates
IS001	1	Solyc01g006640	1235698	1244411	AMP-dependent synthetase and ligase
IS001	1	Solyc01g006650	1246050	1248906	Ethylene insensitive 3 class transcription factor
IS001	1	Solyc01g006660	1249071	1255565	Subtilisin-like serine protease
IS001	1	Solyc01g006670	1258330	1259991	UDP-glucosyltransferase HvUGT5876
IS001	1	Solyc01g006680	1260327	1264553	Transcription factor
IS001	1	Solyc01g006690	1267165	1270752	Unknown Protein
IS001	1	Solyc01g006700	1277385	1285298	Downstream neighbor of SON
IS001	1	Solyc01g006710	1286242	1300706	ATP-dependent RNA helicase
IS001	1	Solyc01g006720	1301729	1309554	ABC transporter G family member 22
…					

Table S11. Summary of significant haplotype/SNP-based regional association analysis for fruit weight, sugars, organic acids and amino acids. Cis-eQTLs were highlighted in bold.

Trait	Chr	BP	BF_{hap}	BF_{SNP}	Start	End	Size	No	Sweeps	Locus name	Candidates
ASA	9	2411368	2.621	3.427	2213892	2558851	344959	38	PSS20	Solyc09g009080	Repressor of silencing 1
ASA	9	67045070	3.434	1.249	66929618	67391497	461879	66	DS152*	Solyc09g075350	Pectinesterase
Asparagine	10	61023555	3.689	2.017	60606622	61027927	421305	53	DS161*	Solyc10g079490	Beta-1-3-galactosyl-o-glycosyl-glycoprotein
Asparagine	11	4242564	3.468	1.929	4213379	4265295	51916	9		Solyc11g011180	LRR receptor-like serine/threonine-protein kinase%2C RLP
Aspartate	4	60724790	3.442	3.835	60367116	61131508	764392	83	DS044	**Solyc04g074530**	Alcohol dehydrogenase
Brix	2	47218316	4.502	0.469	47014198	47222747	208549	30	PSS07, IS030*	Solyc02g083980	Endoglucanase 1
Brix	4	993580	4.332	3.767	845831	1166436	320605	37	DS048*	**Solyc04g006970**	Phosphoenolpyruvate carboxylase
Brix	5	60837144	3.726	0.723	60575012	61009554	434542	42	PSS13, IS056	Solyc05g050700	LRR receptor-like serine/threonine-protein kinase%2C RLP
Brix	6	43589609	5.186	4.065	43182556	44112996	930440	119	DS069	Solyc06g071060	Short-chain dehydrogenase/reductase family protein
Brix	8	64194956	7.343	4.897	63846570	64242138	395568	57		Solyc08g081090	Solute carrier family 22 member
Brix	9	3477979	5.498	5.211	3435498	3800385	364887	45	PSS20, DS149*	Solyc09g010080	Beta-fructofuranosidase%2C insoluble isoenzyme 1
Brix	9	71452675	3.753	-0.2	71414324	71539487	125163	18		Solyc09g092330	NAD dependent epimerase/dehydratase
Brix	11	49507731	4.36	1.021	49185016	50173539	988523	63		Solyc11g062360	Sugar transporter superfamily
Brix	12	2329931	4.961	5.254	1898656	2758225	859569	101	DS175*	Solyc12g009040	Long-chain-fatty-acid--CoA ligase 4
Citrate	6	44996740	5.666	3.84	44907792	45064771	156979	34	DS069, IS062	Solyc06g072920	Aluminum-activated malate transporter
DHA	9	69358878	3.9	3.855	68540575	69789519	1248944	143		Solyc09g089680	1-aminocyclopropane-1-carboxylate oxidase-like protein
Erythritol	2	41134765	3.7	2.654	41108585	41223176	114591	14	IS018	Solyc02g071790	Receptor-like serine/threonine kinase
Erythritol	6	42161946	4.831	-0.273	41772852	42203598	430746	94		Solyc06g068040	Polygalacturonase
Erythritol	10	65121066	5.948	4.693	64921523	65303811	382288	59	DS165*	Solyc10g086240	Glucosyltransferase

Fructose	2	47218316	5.296	0.21	47014408	47362878	348470	45	PSS07, IS030*	Solyc02g083980	Endoglucanase
Fructose	5	60641293	3.462	1.909	60282745	60846344	563599	50	PSS13, IS056	**Solyc05g050500**	ATP synthase F1 delta subunit
Fructose	6	41994475	3.139	3.591	41714220	42271906	557686	77		Solyc06g066600	Solute carrier family 2%2C facilitated glucose transporter member 3
Fucose	2	52448991	3.761	3.195	52371581	53094402	722821	96	IS033*	Solyc02g091100	Oxalyl-CoA decarboxylase
Fucose	3	66807096	5.799	6.014	66764367	66885084	120717	19	IS043	Solyc03g117750	Polygalacturonase
Fucose	4	54480764	3.21	3.475	53832505	54572156	739651	41	DS041	**Solyc04g056540**	Riboflavin biosynthesis protein RibD
Fucose	10	2542891	3.552	1.224	2309960	2941178	631218	75	PSS22	Solyc10g008430	Unknown Protein
Fucose	10	61186144	3.429	2.731	61029096	61325347	296251	37	DS161*	Solyc10g079490	Beta-1-3-galactosyl-o-glycosyl-glycoprotein
FW	1	85022895	4.653	3.533	84843243	85145069	301826	39	DS010, DS013*	Solyc01g091410	Acetyl esterase
FW	1	92475884	3.643	2.019	91846139	92644787	798648	101	IS008, IS022*	Solyc01g104040	UDP-glucose glycoprotein glucosyltransferase
FW	2	49223988	5.086	3.234	48853972	49485324	631352	81	IS024, IS032*	Solyc02g086530	Alpha-galactosidase
FW	2	54758149	5.272	1.687	54222217	55088053	865836	123	IS028	Solyc02g094120	Sulfite oxidase
FW	3	1877032	5.341	1.114	1276788	2073938	797150	79	IS029	Solyc03g007310	Abscisic acid receptor PYL8
FW	3	23141786	4.49	-0.494	20701580	28040495	7338915	136	IS032-IS036	Solyc03g058370	UDP-glucosyltransferase family 1 protein
FW	3	33959056	4.488	3.477	33514046	36310383	2796337	59	IS038	Solyc03g063500	NADH-quinone oxidoreductase subunit N
FW	3	67226288	5.651	-0.323	67114250	67330182	215932	30	IS043	Solyc03g118480	Cell division cycle associated 7
FW	4	55774471	3.464	3.8	55727437	55829706	102269	14	PSS11	**Solyc04g064610**	RAG1-activating protein 1 homolog
FW	5	7352686	3.492	3.726	6635487	8176287	1540800	88	PSS12, IS055	**Solyc05g013910**	Unknown Protein
FW	6	38562515	3.566	4.375	38143928	38977488	833560	83	DS066, IS061	Solyc06g060560	Pentatricopeptide repeat-containing protein
FW	6	41994475	3.566	3.477	41741958	42255545	513587	71		Solyc06g066820	Gibberellin 3-beta-hydroxylase
FW	7	2011211	4.737	1.546	1818113	2140134	322021	39	PSS16	Solyc07g007290	Dimethylaniline monooxygenase 5

FW	7	57465601	3.725	1.231	57318431	58119782	801351	55		Solyc07g043550	UDP-glucose 4-epimerase
FW	7	62048905	4.728	4.713	61814899	62201557	386658	39		Solyc07g053630	Response regulator
FW	8	62865804	7.666	2.604	62402595	63319744	917149	121		Solyc08g079250	Lipid transfer protein
FW	9	2210922	3.406	3.953	2143941	2217142	73201	10		**Solyc09g008720**	Ethylene receptor
FW	9	29626075	3.744	1.232	20430388	46098270	25667882	244		Solyc09g042750	Acyl-CoA thioesterase 9
FW	9	61071784	4.421	2.745	60430166	62186388	1756222	94	PSS21	Solyc09g063060	GDSL esterase/lipase At4g28780
FW	9	68371848	4.565	0.251	68171540	68572873	401333	45		Solyc09g082660	Caffeoyl-CoA O-methyltransferase
FW	11	3158623	3.484	3.748	2965055	3248688	283633	36	DS111, IS089	**Solyc11g009010**	Hydrolase alpha/beta fold family protein
FW	11	55958042	4.324	5.106	55707648	56018369	310721	32	IS131*	Solyc11g072700	Glycosyltransferase-like protein
FW	12	1430145	5.615	6.023	1324253	1937156	612903	67	DS175*	Solyc12g006990	Arf GTPase activating protein
GABA	6	1330594	3.63	4.185	1266788	1447768	180980	24		Solyc06g007310	Deoxyribonuclease tatD
Glucose	2	47218316	3.945	0.13	47014383	47322782	308399	41	PSS07, IS030*	**Solyc02g083980**	Endoglucanase 1
Glucose	3	5239104	4.069	2.345	4912145	6052295	1140150	97	DS025	Solyc03g033650	Unknown Protein
Glucose	6	41994475	4.134	4.53	41748343	42244811	496468	67		Solyc06g066600	Solute carrier family 2%2C facilitated glucose transporter member 3
Glutamate	4	60724790	3.189	3.531	60383511	61132407	748896	82	DS044	Solyc04g076090	Glucose-6-phosphate isomerase 2
Glutamate	8	62814480	3.826	1.064	62061884	63145320	1083436	140	IS095*	Solyc08g079440	UDP-D-glucuronate 4-epimerase 2
Glutamate	12	904424	3.543	4.49	715073	1357927	642854	77	DS175*	Solyc12g006410	UDP-glucose 4-epimerase
Glutarate2oxo	6	44818656	5.606	0.547	44635984	44963860	327876	59	DS069, IS062	Solyc06g072670	Short-chain dehydrogenase/reductase
Glutarate2oxo	10	61023555	3.396	3.76	60596099	61029824	433725	55	DS161*	Solyc10g079470	L-galactono-1%2C4-lactone dehydrogenase
Glutarate2oxo	11	54853867	4.236	1.593	54811596	54901085	89489	14		Solyc11g071290	Alcohol dehydrogenase
Inositol1P	2	35173023	2.915	3.613	34431239	35294389	863150	54	DS019*	Solyc02g063180	UDP-N-acetylenolpyruvoylglucosamine reductase
Malate	2	22214295	3.75	4.597	21308389	22427306	1118917	30		Solyc02g021210	Unknown Protein
Malate	3	65147049	4.028	4.413	65049357	65365080	315723	49	PSS10,	**Solyc03g115380**	UDP-glucose dehydrogenase

									IS043		
Malate	6	44919354	11.407	-0.187	44760991	45039057	278066	54	DS069, IS062	Solyc06g072920	Aluminum-activated malate transporter
Maltitol	6	25117534	3.849	3.198	24945058	25732733	787675	45	DS061	Solyc06g035960	4-coumarate-CoA ligase-like protein
Maltitol	9	17881465	3.526	3.824	14521797	28155050	13633253	186		Solyc09g019970	Ubiquitin carboxyl-terminal hydrolase
Maltitol	9	49839317	3.611	3.824	45638641	54538454	8899813	155		Solyc09g057630	Glucan endo-1 3-beta-glucosidase
Nicotinate	1	92644961	3.515	3.273	92419952	93310548	890596	107	IS008, DS019*	Solyc01g104200	Fatty acyl coA reductase
Nicotinate	2	54118621	4.804	-0.197	53929927	54432210	502283	69	IS028	Solyc02g093230	Caffeoyl-CoA O-methyltransferase
Nicotinate	3	4884492	3.426	1.641	4624368	4973889	349521	30	PSS08	Solyc03g032210	Acyl-CoA synthetase/AMP-acid ligase II
Nicotinate	3	66809350	3.474	3.623	66764523	66899361	134838	21	IS043	Solyc03g117750	Polygalacturonase
Nicotinate	5	1860363	4.804	4.879	1439222	2133518	694296	78	DS048, IS053, DS055*	**Solyc05g007260**	Ribulose-phosphate 3-epimerase
Nicotinate	7	61822724	4.059	3.104	61655076	62050829	395753	40		Solyc07g053430	Glucan endo-1 3-beta-glucosidase
Nicotinate	8	53602606	3.459	2.208	53040148	54178942	1138794	61	PSS19, IS075, IS089*	**Solyc08g065490**	Serine hydroxymethyltransferase
Nicotinate	12	1420748	2.758	3.428	1253256	1762413	509157	53	DS175*	Solyc12g007030	Aldehyde dehydrogenase 1
Phenylalanine	7	65807351	4.094	0.892	65614739	65975230	360491	50	IS069	Solyc07g063460	Methyltransferase family protein
Proline	2	34220988	3.847	2.344	33651320	34801684	1150364	79	DS017	**Solyc02g062500**	2-oxoglutarate-dependent dioxygenase
Proline	2	51094633	2.728	3.469	50554695	51388460	833765	116	IS033*	Solyc02g089620	Proline dehydrogenase
Proline	3	66720661	3.931	4.202	66440503	66734401	293898	40	IS043	Solyc03g117350	Amino acid transporter
Proline	5	1664103	4.291	-0.023	1373499	1998377	624878	70	DS048, IS053, DS055*	**Solyc05g007060**	Bifunctional N-succinyldiaminopimelate-aminotransferase/acetylornithine transaminase protein
Proline	6	25117534	3.672	3.881	24950810	25660585	709775	37	DS061	Solyc06g035860	Unknown Protein
Proline	8	928474	2.971	3.538	871320	1177414	306094	43	PSS18	**Solyc08g006330**	UDP-glucose salicylic acid glucosyltransferase

Proline	8	60598551	4.913	0.483	60162211	60821780	659569	70	DS145*	Solyc08g076650	Alpha alpha-trehalose-phosphate synthase
Proline	9	17881465	3.891	3.9	15588107	28030987	12442880	160		Solyc09g019970	Ubiquitin carboxyl-terminal hydrolase
Proline	9	49223167	3.569	-0.315	39774798	53575317	13800519	175		Solyc09g057630	Glucan endo-1 3-beta-glucosidase A6
Proline	9	62209136	4.078	4.247	61415244	62314387	899143	59	PSS21	Solyc09g064450	NADH dehydrogenase
Proline	11	51523371	2.926	3.385	51460553	51754681	294128	21	PSS24	Solyc11g065830	2-oxoglutarate/malate translocator-like protein
Rhamnose	1	92492935	4.219	3.136	92203206	92645216	442010	55	IS008, IS022*	Solyc01g104040	UDP-glucose glycoprotein glucosyltransferase
Rhamnose	2	20852994	4.341	-0.191	20335127	22198488	1863361	30		Solyc02g020980	4-alpha-glucanotransferase
Rhamnose	3	65045517	4.049	3.084	64909434	65063674	154240	18	PSS10, IS042, IS054*	Solyc03g115200	Glucan endo-1 3-beta-glucosidase 1
Rhamnose	3	67080375	4.963	3.704	67019742	67151919	132177	20	IS043	**Solyc03g118120**	Transferase transferring glycosyl groups
Rhamnose	4	1163761	3.392	3.186	875038	1220614	345576	36	DS049*	**Solyc04g007520**	Phosphatidylserine synthase 2
Rhamnose	4	44934889	4.051	1.35	40797863	47615608	6817745	126		Solyc04g049960	Plasma membrane protein 3
Rhamnose	4	56044571	4.883	2.321	55986542	57883396	1896854	23		Solyc04g049960	Endo-1 4-beta-glucanase
Rhamnose	5	558180	4.042	2.292	200737	636135	435398	57	DS046, DS055*	Solyc05g005750	Alpha alpha-trehalose-phosphate synthase
Rhamnose	7	62755438	3.631	2.78	62299667	63029123	729456	93	PSS17, IS067	**Solyc07g054440**	Beta-1%2C3-galactosyl-O-glycosyl-glycoprotein beta-1%2C6-N-acetylglucosaminyltransferase 7
Rhamnose	11	54852419	3.816	1.197	54785657	54854095	68438	10		Solyc11g071290	Alcohol dehydrogenase
Saccharate	10	474132	3.367	3.653	185887	744572	558685	74		**Solyc10g005510**	Glyceraldehyde-3-phosphate dehydrogenase
Saccharate	11	52804397	3.4	2.859	52539907	52843886	303979	36		Solyc11g066820	Cellulose synthase-like C6 glycosyltransferase family 2
Sucrose	2	47218316	3.943	-0.111	47070434	47332914	262480	32	PSS07, IS030*	Solyc02g083980	Endoglucanase
Sucrose	6	25117534	3.503	3.468	24865867	25656439	790572	41	DS061	Solyc06g035960	4-coumarate-CoA ligase-like protein
Sucrose	6	41994475	4.43	3.338	41710556	42257843	547287	77		Solyc06g066600	Solute carrier family 2%2C

											facilitated glucose transporter member 3
Sucrose	8	64447573	3.899	0.201	64359649	64986565	626916	78	DS100, IS079, IS097*	Solyc08g081390	Phosphoglycerate mutase family protein
Threonine	6	1330574	3.402	2.448	1252287	1357763	105476	16		Solyc06g007190	Integrin-linked kinase-associated serine/threonine phosphatase 2C
Tocopherol	10	2199297	3.402	3.886	1985675	2389583	403908	50	PSS22	Solyc10g008020	Methyltransferase
Tyramine	4	62948548	3.468	2.642	62825112	63536688	711576	90	DS045, IS050	Solyc04g078090	Acyl-CoA-binding domain-containing protein 6
Tyramine	8	60405484	4.572	3.941	60007557	60671836	664279	75	IS077, DS145*	Solyc08g076390	Lysine-specific demethylase 5A
Tyramine	9	68174329	3.119	3.454	67851515	68401059	549544	63	IS110*	Solyc09g082460	Homocysteine s-methyltransferase

PSS, positive selective sweeps; DS, domestication sweeps; IS, Improvement sweeps; * Sweeps identified in Lin et al. (2014).

Table S12. Significant associations detected in EMMAX for tomato fruit weight, sugars, organic acids and amino acids.

Trait	Chr	Position	P	Sweeps	Locus name	Candidates
Asparagine	2	54365596	1.81E-05	IS028	Solyc02g093520	Copine-like protein
Aspartate	4	60724790	2.25E-06	DS044	Solyc04g074530	Alcohol dehydrogenase
Citrate	6	44955568	2.00E-07	DS069, IS062	Solyc06g072920	Aluminum-activated malate transporter
Glutamate	4	60724790	4.09E-07	DS044	Solyc04g076090	Glucose-6-phosphate isomerase 2
Glutamate	12	904424	9.62E-07	DS175*	Solyc12g006410	UDP-glucose 4-epimerase
Glutarate2oxo	6	44955568	9.02E-07	DS069, IS062	Solyc06g072670	Short-chain dehydrogenase/reductase
Malate	2	22214295	1.31E-06		Solyc02g021210	Unknown Protein
Malate	6	44955568	1.22E-13	DS069, IS062	Solyc06g072920	Aluminum-activated malate transporter

PSS, positive selective sweeps; DS, domestication sweeps; IS, Improvement sweeps; * Sweeps identified in Lin et al. (2014).

Table S13. Significant associations detected for fruit weight using the MLMM.

Trait	Chr	Position	P	Start	End	Sweeps	Locus name	Candidates
fw	2	54873732	2.03E-13	54599224	54874151	IS028	Solyc02g094300	Uridylyltransferase
fw	4	59964407	1.95E-11	55460761	60524629	PSS11,DS042-DS043, IS048-IS049	Solyc04g073960	Major facilitator superfamily transporter
fw	4	65358078	4.51E-06	63582333	66489674	IS051-IS052	Solyc04g081340	Receptor expression-enhancing protein 3
fw	6	44849440	6.92E-06	42359139	46959297	DS069	Solyc06g072670	Short-chain dehydrogenase/reductase
fw	7	3279240	2.60E-17	2247152	3330664	DS074, IS064	Solyc07g008430	Unknown Protein
fw	7	3745280	8.68E-09	3429976	3815708	DS074, IS065	Solyc07g008760	Tetratricopeptide repeat
fw	7	66672032	5.56E-12	66657641	66668624		Solyc07g064670	Beta-1 3-galactosyltransferase 6
fw	8	60195041	2.45E-09	59963068	60636375	IS077, DS145*	Solyc08g076140	Phosphomevalonate kinase
fw	12	1430145	1.78E-16	1395964	1449699		-	fw12.1
ASA	6	40541416	1.42E-05	38691771	41800372	DS066-DS068	Solyc06g065020.2	Peptide transporter
ASA	7	66663289	2.94E-10	66657641	66668624		Solyc07g064580.2	Conserved gene of unknown function
ASA	9	2411368	1.09E-07	1218234	3251393	DS102	Solyc09g009080.2	Repressor of silencing1
ASA	9	66376508	1.07E-05	64391250	67475096	IS083-IS085, IS108*	Solyc09g074480.1	Gene of unknown function
ASA	11	3393788	4.66E-08	3174682	3434250		Solyc11g010310.1	ATP-dependent RNA helicase
Asn	2	54365596	1.93E-07	54310988	54625509	IS028	Solyc02g093520.2	Copine-like protein
Asp	4	60724790	1.67E-07	60425430	62287804	DS044	Solyc04g074810.2	Basic helix-loop-helix transcription factor
Citrate	6	44955568	1.48E-07	44641317	44955621	DS069, IS062	Solyc06g072930.2	Conserved gene of unknown function
DHA	9	69358878	3.16E-39	65744780	70214577	IS084-IS085, DS154*, IS111*	Solyc09g089560.2	Ubiquitin C-terminal hydrolase family protein
DHA	11	3063738	8.49E-07	2933146	3210568	IS089	SGN-U564017	Pentatricopeptide repeat-containing protein
Ery	2	41981476	1.24E-07	41158136	42148698	PSS04, IS019	Solyc02g076860.2	Pollen allergen Chenopodium a1
Erythritol	10	65121066	5.98E-16	63752582	65232094	DS165*	Solyc10g086220.1	Flavin oxidoreductase/NADH oxidase
Fru	5	60640821	9.31E-07	60387327	60641168	IS056	Solyc05g050500.1	Conserved gene of unknown function
Fru	6	41994475	9.05E-07	41404898	42102847		Solyc06g066810.2	Katanin p60 ATPase-containing subunit
Fuc	3	66807096	2.70E-07	66806258	66836647	IS043	Solyc03g117780.2	UV excision repair protein RAD23

Fuc	4	54480764	1.63E-06	54185403	54545844	DS041	Solyc04g056530.1	Structural constituent of ribosome
GABA	6	1330594	5.53E-08	1329361	1475391	IS058	Solyc06g007310.2	D-type of twin-arginine translocation DNase
Malate	2	22214295	1.28E-06	10570194	28803661		SGN-U565892	Gene of unknown function
Malate	6	44955568	2.48E-08	44641317	44955621	DS069, IS062	Solyc06g072930.2	Conserved gene of unknown function
Nicotinate	2	54873732	3.83E-06	54599224	54874151	IS028	Solyc02g094300.2	Uridylyltransferase PII
Pro	2	34220988	3.71E-06	5408025	38642232		Nonavailable	Conserved gene of unknown function
Pro	6	25117534	3.91E-07	24171480	25429720	DS060-DS061	Solyc06g035870.2	Membrane-associated progesterone receptor component1
Rha	1	92492935	2.61E-08	91808475	95927839	DS014-DS015	SGN-U565850	Embryo-specific
Rha	3	65049140	2.32E-09	65048545	65049261		Solyc03g115250.2	Conserved gene of unknown function
Rha	8	1403227	9.41E-06	1401680	3652087	DS093-DS094, IS072, DS135*	Solyc08g006860.2	Patatin1-Kuras2
Rha	9	3484890	2.10E-10	3479971	3507971		SGN-U565153	Gene of unknown function
SSC	2	35274016	7.79E-26	26923732	35497195	DS017-DS019	Solyc02g063220.2	Man-6-P isomerase
SSC	2	48629882	7.73E-10	47785343	50325174	DS020, IS024	Solyc02g085840.2	UV excision DNA repair protein RAD23
SSC	3	71076	0.0006	0	1971818	IS029	Solyc03g005100.2	CXE carboxylesterase
SSC	6	1748321	2.92E-21	1254	14993414	IS058	Solyc06g007830.1	Auxin signaling F-box1 family protein
SSC	7	62755438	1.22E-12	57514198	63709083	DS088-DS090, IS066-IS067	Solyc07g054440.2	b-1,3-Galactosyl-O-glycosyl-glycoprotein
SSC	8	62311446	5.57E-08	59940364	66571680	IS077	Solyc08g078530.2	Agenet domain-containing protein
SSC	9	3477979	1.34E-33	3477563	3510003	DS149*	Solyc09g010080.2	b-Fructofuranosidase (lin5)
SSC	11	2481288	1.89E-13	2444016	2520530		Solyc11g008250.1	Single-stranded nucleic acid-binding R3H domain protein
SSC	12	5274083	2.41E-06	4907413	9855738	DS125-DS127	Nonavailable	Gene of unknown function
Suc	2	41913145	2.57E-06	40657169	43979649	IS017-IS021	Solyc02g076800.1	Conserved gene of unknown function
Suc	4	61799612	6.01E-05	61685935	61846586		Solyc04g076870.2	Glutamyl-tRNA reductase
Suc	5	4037126	9.51E-09	3934446	4082059	DS049	Solyc05g009820.2	Glycosyltransferase family GT8 protein factor
Thr	2	54365596	3.75E-07	54310988	54625509	IS028	Solyc02g093520.2	Copine-like protein
Threonate	4	60724790	5.73E-06	60425430	62287804	DS044	Solyc04g074810.2	Basic helix-loop-helix transcription factor
Tocopherol	10	2199297	4.35E-07	1626097	2440580		Solyc10g008030.2	Conserved gene of unknown function
Tyramine	8	2587919	1.12E-05	2576623	2658946		Solyc08g008120.2	Conserved gene of unknown function
Tyramine	8	60405484	1.18E-07	59536365	60762142	IS077, DS145*	Solyc08g076390.2	Lys-specific demethylase5A
Tyramine	11	762353	1.54E-06	758724	1306196	DS109	SGN-U275742	Transcription regulator

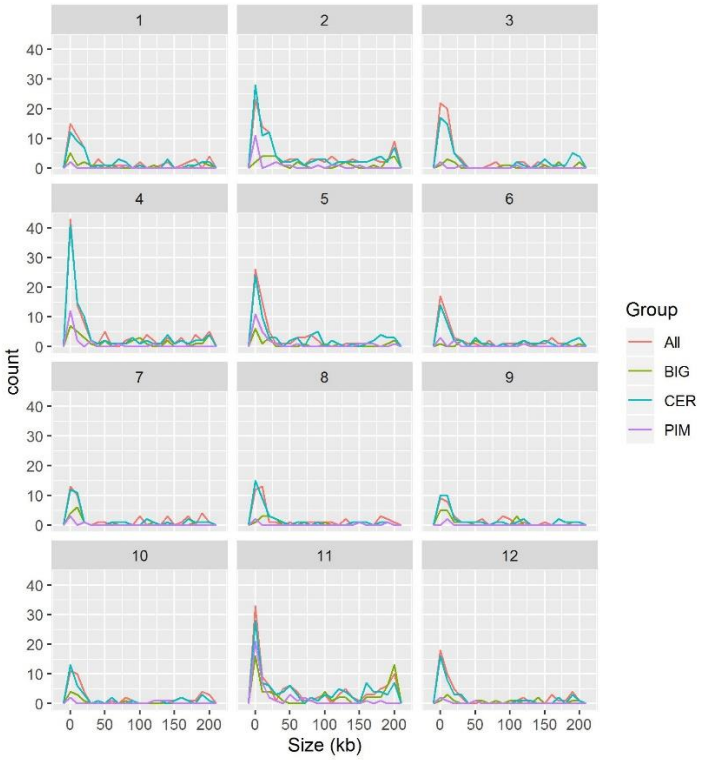


Figure S1. Distribution of the size of haplotype blocks within all accessions and subgroups. All, 163 tomato accessions; BIG, 31 large-fruit tomato accessions; CER, 116 cherry tomato accessions; PIM, 16 wild tomato species.

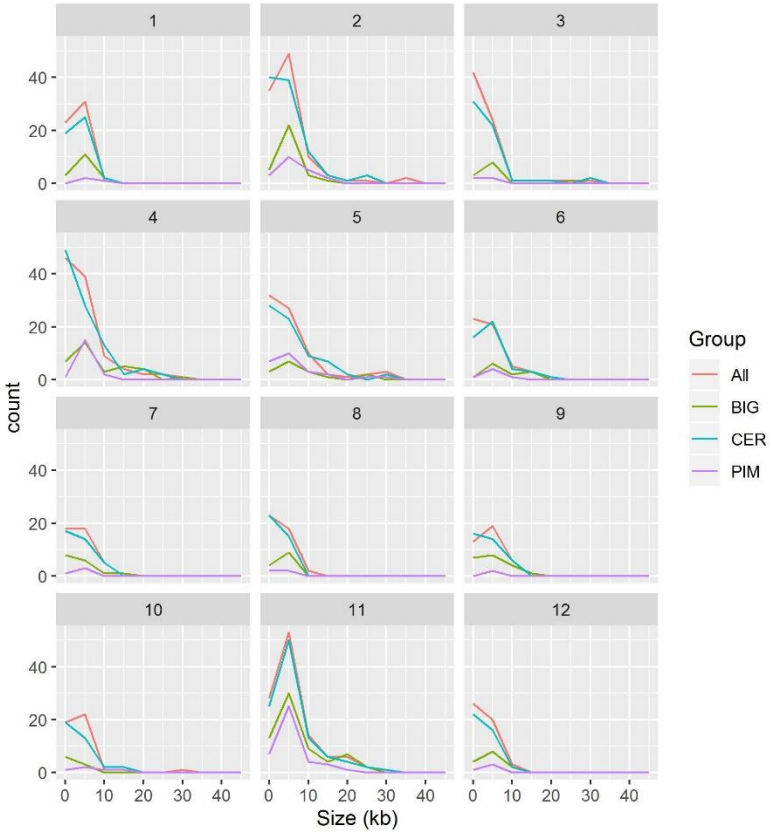


Figure S2. Distribution of the number of SNPs per chromosome within all accessions and subgroups. All, 163 tomato accessions; BIG, 31 large-fruit tomato accessions; CER, 116 cherry tomato accessions; PIM, 16 wild tomato species.

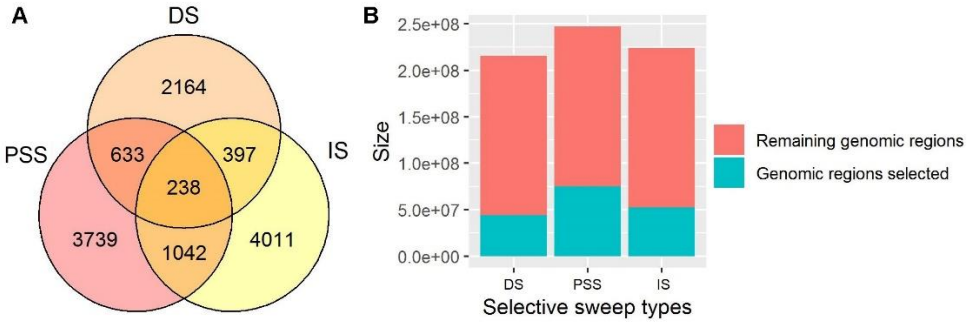


Figure S3. Genomic coverage comparison of positive selective sweeps, domestication and improvement sweeps. (**A**) Venn diagram of the number of genes within positive selective sweeps, domestication and improvement sweeps. PSS, positive selective sweeps; DS, domestication sweeps; IS, improvement sweeps. (**B**) Comparison of the genomic coverage of positive selective sweeps, domestication and improvement sweeps.

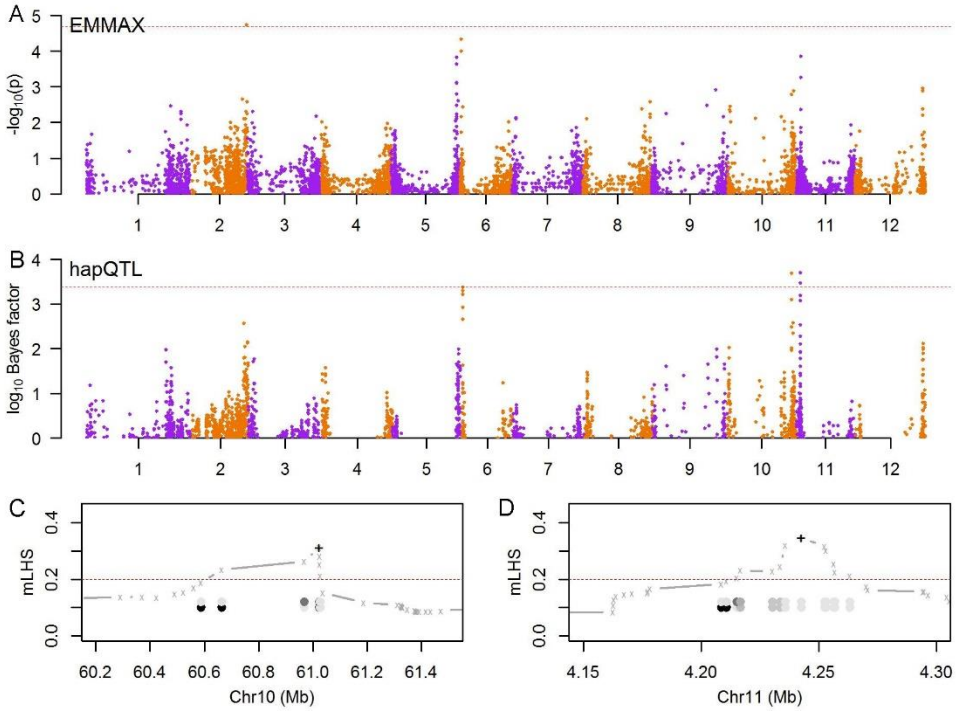


Figure S4. Genome-wide association of ascorbic acid and marker local haplotype sharing (mLHS) of the peak associated SNPs. (**A**) Manhattan plot of genome-wide associations using efficient mixed-model association expedited model (EMMAX). (**B**) Manhattan plot of haplotype- and SNP-based Bayes model using hapQTL. (**C,D**) mLHS patterns of the peak associated SNPs.

Figure S5-S29 were genome-wide association of ascorbic acid and marker local haplotype sharing (mLHS) of the peak associated SNPs for the other traits and were not provided here.

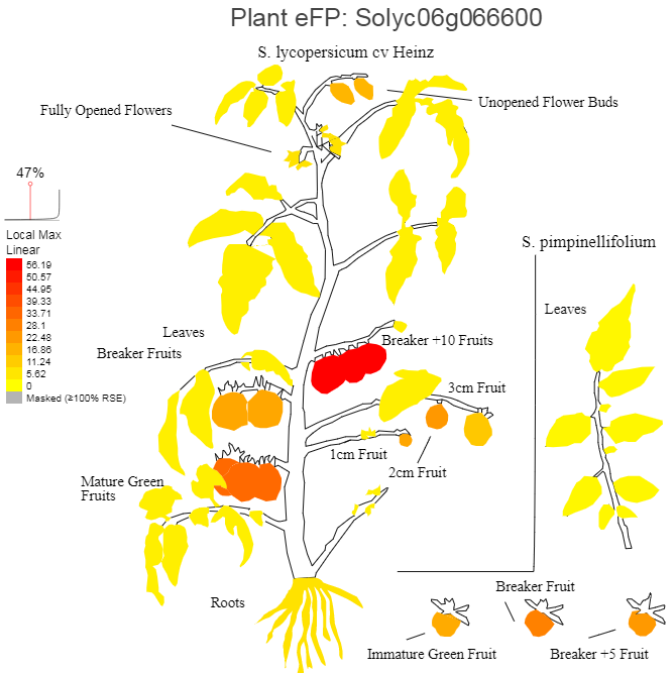


Figure S30. Gene expressions at different tissues and developing stages of the candidate gene (*Solute carrier family 2%2C facilitated glucose transporter member 3*, Solyc06g066600) for the associations detected for fructose, glucose and fructose on chr6.

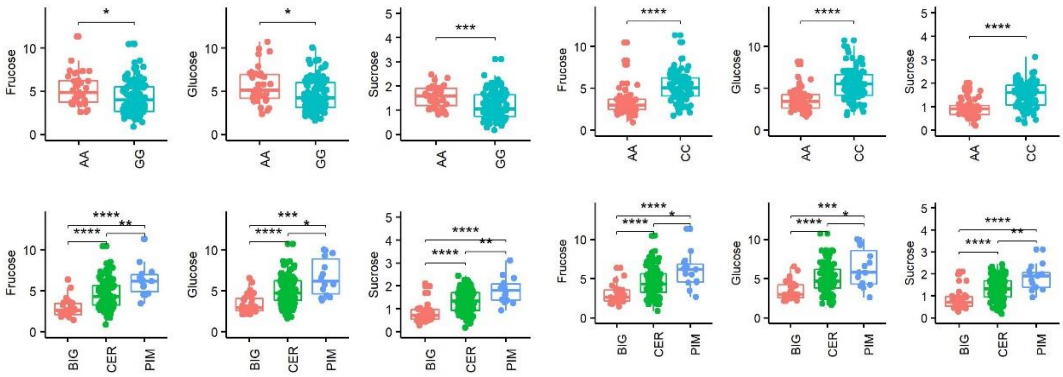


Figure S31. Comparison of the content of fructose, glucose and sucrose between allele A and B as well as three subgroups for the association detected on chr2. BIG, big-fruit tomato (*S. lycopersicum*); CER, cherry tomato (*S. lycopersicum* var *cerasiforme*); PIM, the closest wild species (*S. pimpinellifolium*). ns, not significant; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

Figure S32. Comparison of the content of fructose, glucose and sucrose between allele A and B as well as three subgroups for the association detected on chr6. BIG, big-fruit tomato (*S. lycopersicum*); CER, cherry tomato (*S. lycopersicum* var *cerasiforme*); PIM, the closest wild species (*S. pimpinellifolium*). ns, not significant; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

Appendix 2

Appendix 2 Supplementary tables and figures related to Chapter 4

Supplementary Table 1. Mean and standard error of imputation info.

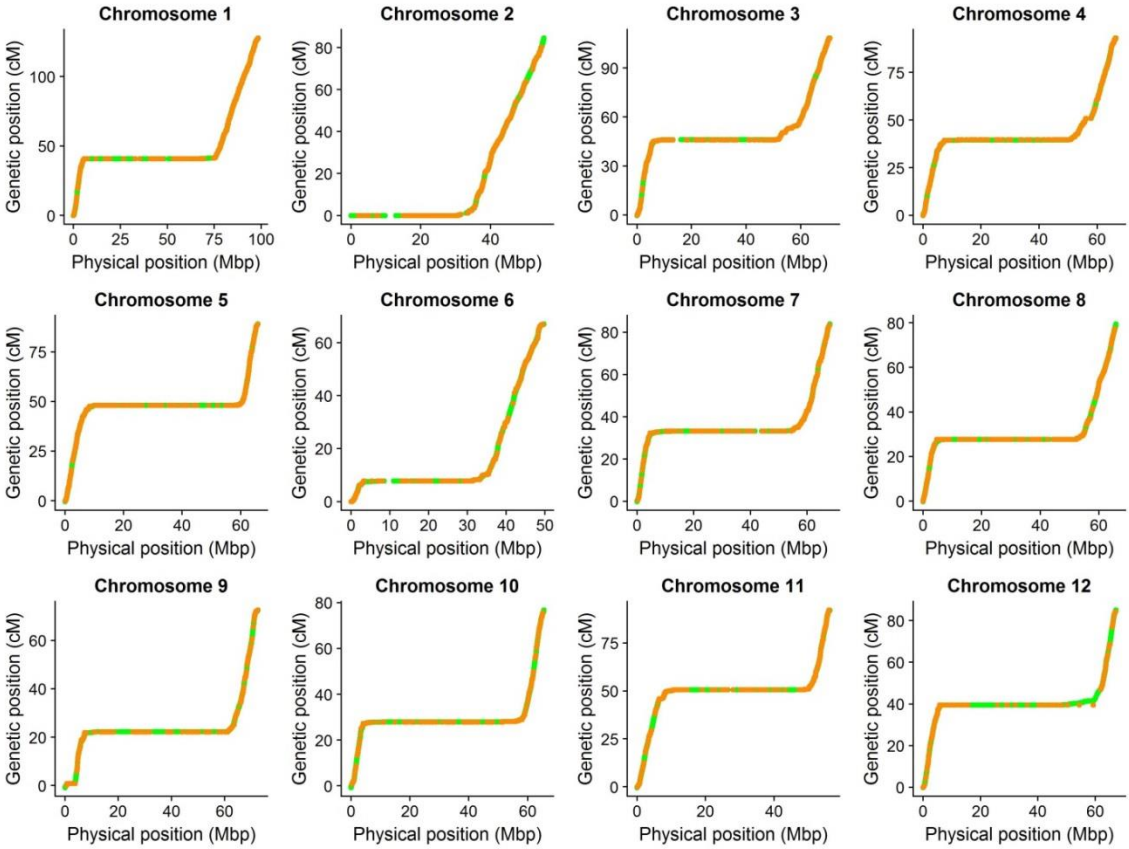
MAF	No. of SNPs	Mean info	SD info
0.050	27724	0.6221	0.2628
0.075	17990	0.7525	0.1965
0.100	9608	0.8639	0.0853
0.125	17185	0.9197	0.0702
0.150	31414	0.9436	0.0536
0.175	29102	0.9418	0.0536
0.200	16478	0.9253	0.0645
0.225	11171	0.9289	0.0527
0.250	15136	0.9659	0.0424
0.275	6106	0.9248	0.0373
0.300	5640	0.9496	0.0358
0.325	11673	0.9628	0.0294
0.350	3152	0.9433	0.0366
0.375	5370	0.9560	0.0264
0.400	2213	0.9535	0.0242
0.425	1810	0.9543	0.0228
0.450	2265	0.9494	0.0247
0.475	1963	0.9401	0.0261
0.500	2108	0.9367	0.0260
Mean	11479.3684	0.9123	0.0616

This analysis was done based on the nine accessions appeared both in the reference and panel S across different MAF bins after filtering with MAF $\geqslant$ 0.037, HWE $\geqslant$ 0.000001, missing $\leqslant$ 0.1 and missing_call $\leqslant$ 0.1 and Info $\geqslant$ 0.60.

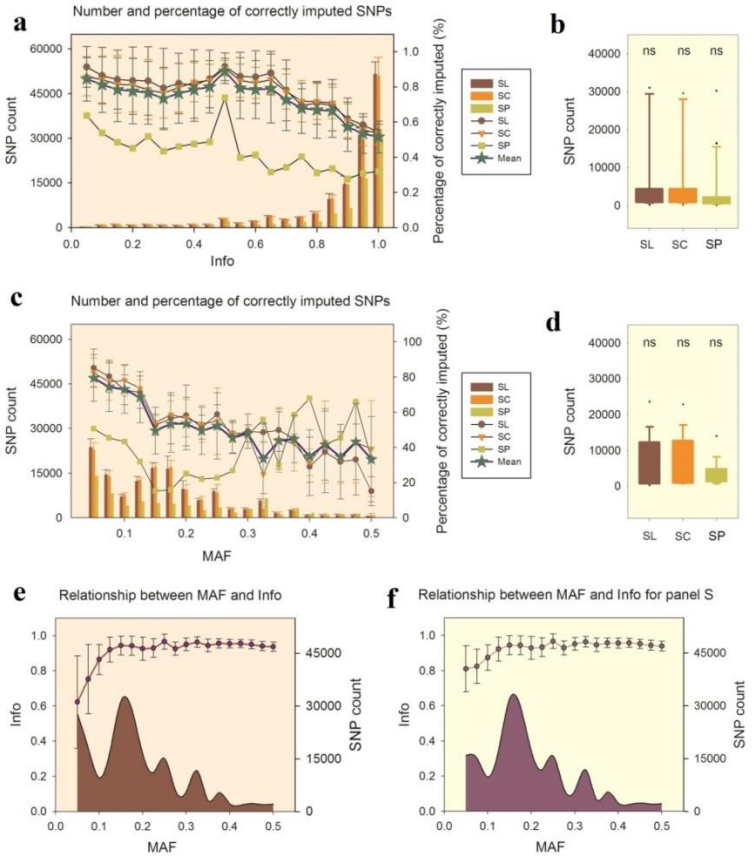
Supplementary Table 2. Mean and standard error of imputation info for Panel S and B.

	Panel S			Panel B		
MAF	No. of SNPs	Mean info	SD info	No. of SNPs	Mean info	SD info
0.050	15855	0.8103	0.1314	30874	0.7834	0.1360
0.075	15150	0.8242	0.0973	19803	0.8512	0.1048
0.100	9762	0.8742	0.0717	9754	0.8669	0.0717
0.125	17547	0.9224	0.0683	6462	0.8706	0.1006
0.150	31929	0.9447	0.0533	9183	0.9349	0.0468
0.175	29897	0.9434	0.0537	9329	0.9223	0.0723
0.200	17278	0.9288	0.0649	20899	0.9694	0.0439
0.225	11905	0.9333	0.0538	28911	0.9646	0.0358
0.250	15806	0.9673	0.0421	18438	0.9387	0.0555
0.275	6491	0.9293	0.0403	10637	0.9481	0.0492
0.300	5885	0.9517	0.0365	6291	0.9263	0.0552
0.325	11855	0.9633	0.0295	20592	0.9741	0.0351
0.350	3328	0.9463	0.0378	6208	0.9451	0.0400
0.375	5527	0.9573	0.0270	7087	0.9601	0.0333
0.400	2341	0.9558	0.0290	7912	0.9561	0.0281
0.425	1948	0.9576	0.0249	8019	0.9535	0.0313
0.450	2378	0.9518	0.0264	11151	0.9672	0.0281
0.475	2082	0.9435	0.0289	14826	0.9736	0.0263
0.500	2188	0.9390	0.0282	6038	0.9667	0.0281
Sum	209152	-	-	252414	-	
Mean	11008.000000	0.9286	0.0497	13284.947368	0.9218	0.0940
Min	1948	0.8103	0.0249	6038	0.7834	0.0263
Max	31929	0.9673	0.1314	30874	0.9741	0.1360

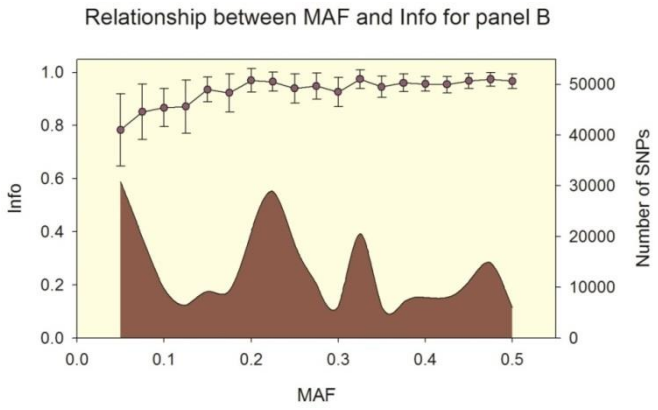
This analysis was done based on all accessions in panel S and B across different MAF bins after filtering with HWE $\geqslant$ 0.000001, missing $\leqslant$ 0.1 and missing_call $\leqslant$ 0.1 and Info $\geqslant$ 0.60. The MAF filtering threshold for Panel S and B was 0.037 and 0.021, respectively.



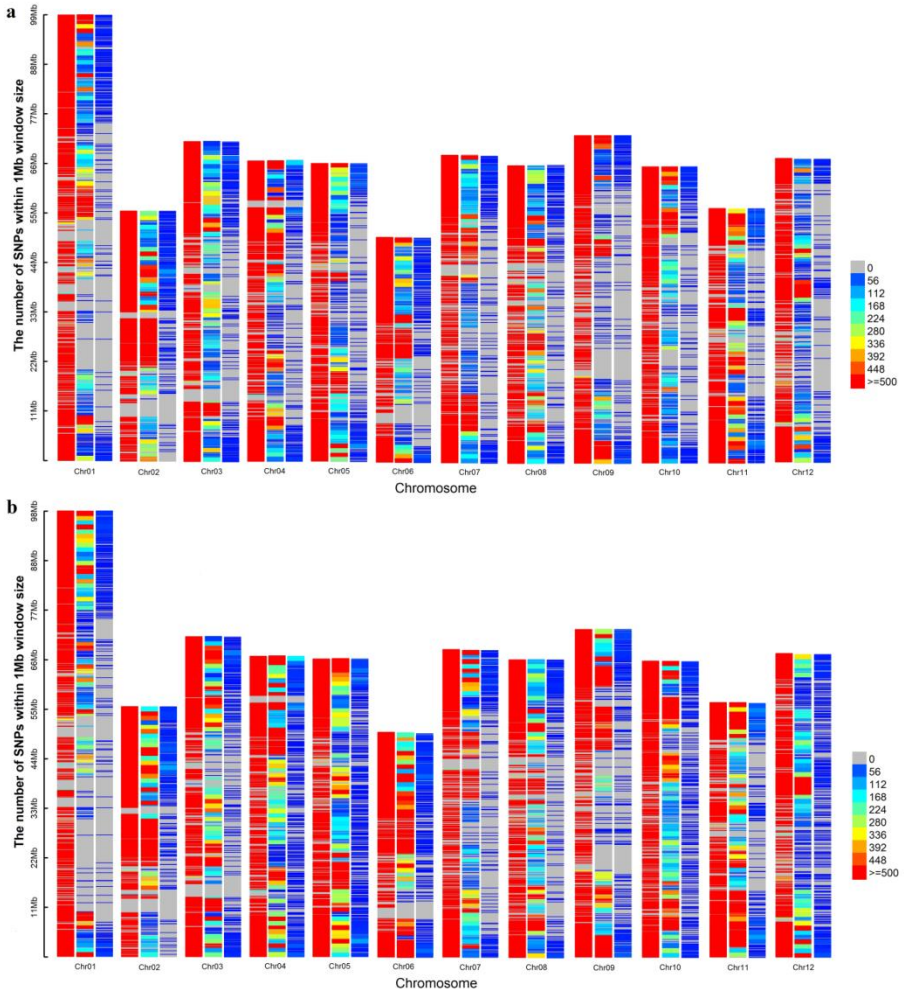
Supplementary Figure 1. Relationship between the inferred genetic and the physical positions on each chromosome. The genetic and physical relationships of SNPs in EXPIM 2012 are indicated by orange circles. The genetic positions of the 3,809,156 SNPs in the reference panel were inferred based on the relationship of the physical and genetic positions of EXPIM 2012 and the corresponding physical position on the reference. The physical position and inferred genetic positions of SNPs in the reference panel are indicated in green circles.



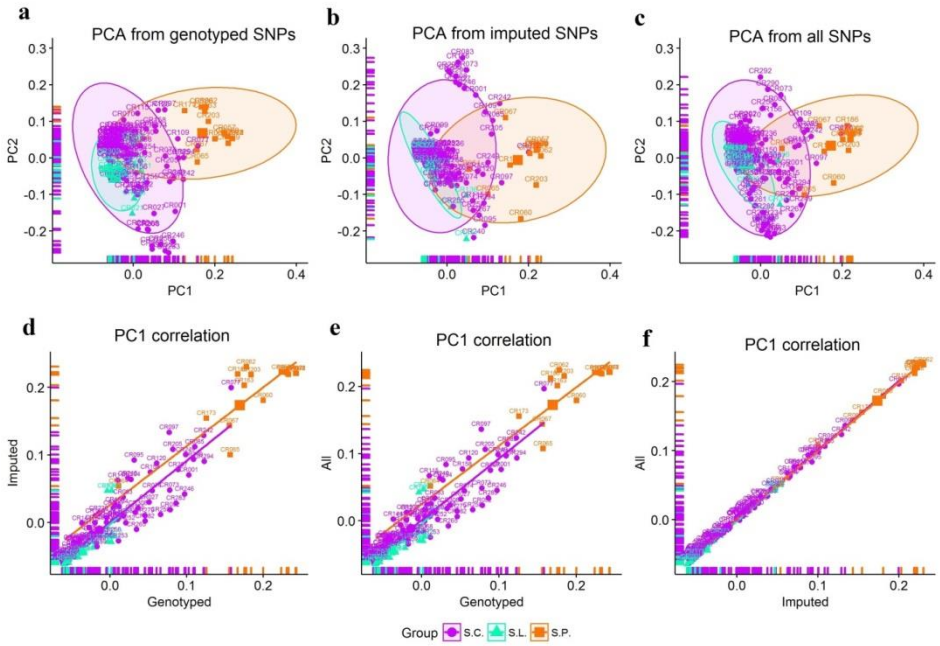
Supplementary Figure 2. Relationship between imputation quality Info and MAF. This relationship was evaluated among different tomato classes (2 *S. lycopersicum*, 6 *S.l.* var *cerasiforme* and 1 *S. pimpinellifolium*) that are in common both in the reference panel and GWAS panel S. Primary quality control steps were done before analysis with Hardy-Weinberg equilibrium (HWE) ≥ 0.000001, MAF ≥ 0.037, missing rate ≤ 0.10 and missing call rate ≤ 0.10. (**a**) The number and percentage of correctly imputed SNPs at different Info values. Each bar represents the number of correctly imputed SNPs. This was done by comparing the maximum of the three probabilities at a locus that was higher than 0.9 with the sequenced genotyping calls. (**b**) Significant t-test of the number of correctly imputed SNPs for three tomato classes at different Info values. (**c**) The number and percentage of correctly imputed SNPs at different MAF values. (**d**) Significant t-test of the number of correctly imputed SNPs for three tomato classes at different MAF values. (**e**) Relationship between MAF and Info for the nine accessions. Smooth line with stars represents the number of SNPs at different MAF values. Line plot with error bars represents the mean and standard error of Info values at different MAF values. (**f**) Relationship between MAF and Info values for all accessions of panel S after genotype imputation quality control. Error bars stand for the stand error.



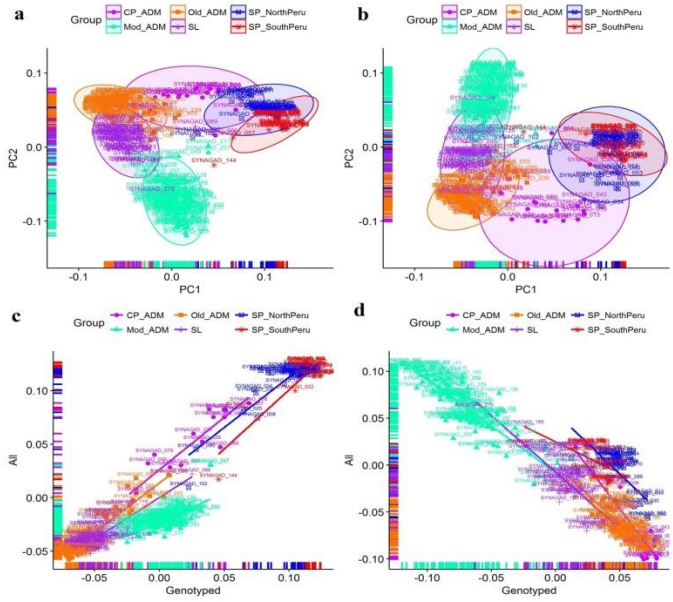
Supplementary Figure 3. Relationship between imputation quality Info criteria and MAF for panel B. Main quality control include Hardy-Weinberg equilibrium (HWE) ≥ 0.000001, MAF ≥ 0.021, missing rate ≤ 0.10 and missing call rate ≤ 0.10. Smooth line represents the number of SNPs at different MAF values. Line plot with error bars represents the mean and standard error of Info values at different MAF values. Error bars stand for the stand error.



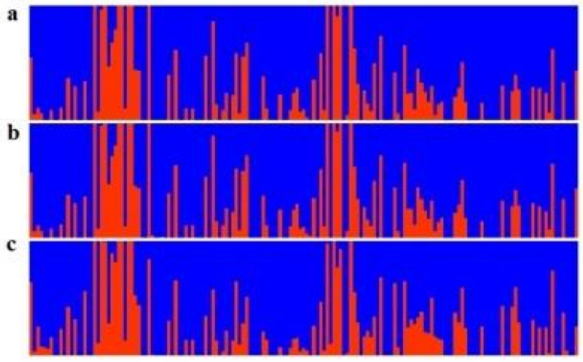
Supplementary Figure 4. Physical positions of SNPs on each. Color legend indicates the density of SNPs on 12 chromosomes. (**a**) Physical positions of SNPs on each chromosome before (right) and after imputation (middle) and the reference panel (left) for panel S. (**b**) Physical positions of SNPs on each chromosome before (right) and after imputation (middle) and the reference panel (left) for panel B.



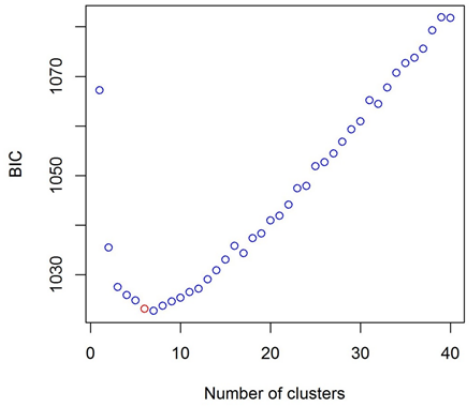
Supplementary Figure 5. Principal component analysis comparison before and after genotype imputation for panel S. (**a**) PCA revealed by genotyped SNPs. (**b**) PCA revealed by independent imputed SNPs ($r^2 \leq 0.2$). (**c**) PCA revealed by all independent genotyped plus imputed SNPs ($r^2 \leq 0.2$). (**d**) Correlation between the first principal component of genotyped and independent imputed SNPs. (**e**) Correlation between the first principal component of genotyped and all independent genotyped plus imputed SNPs. (**f**) Correlation between the first principal component of independent imputed and all independent genotyped plus imputed SNPs. S.C., *Solanum lycopersicum*; S.C., *S. lycopersicum* var *cerasiforme* and S.P., *S. pimpinellifolium*. Source data of Figure S5a-5f are provided as a Source Data file.



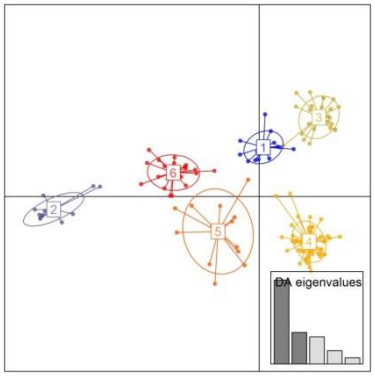
Supplementary Figure 6. Principal component analysis comparison before and after genotype imputation for the six groups previously defined for panel B. (**A**) PCA revealed by genotyped SNPs. (**B**) PCA revealed by all (genotyped and imputed) SNPs ($r^2 \leq 0.2$). (**C**) Correlation between the first principal component of genotyped and all SNPs. (**D**) Correlation between the second principal component of genotyped and all SNPs. Source data of Figure S6a-dare provided as a Source Data file.



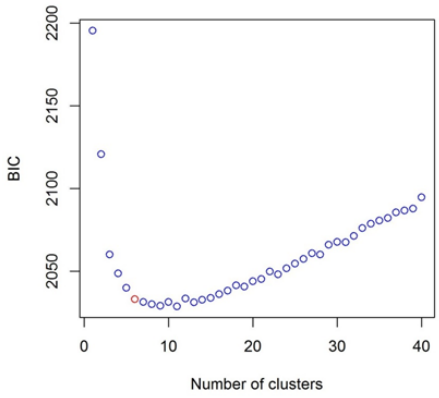
Supplementary Figure 7. Population structure of panel S. Each bar represents one individual. (**a**) Population structure revealed by independent imputed SNPs. (**b**) Population structure revealed by all independent (imputed and genotyped) SNPs. (**c**) Population structure revealed by genotyped SNPs.



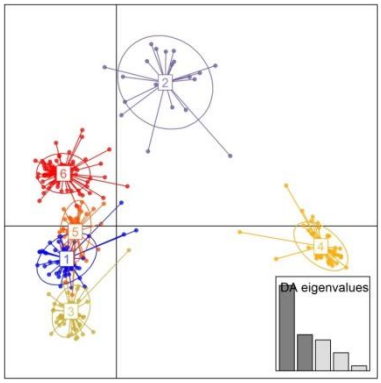
Supplementary Figure 8. Relationship between Bayesian Information Criteria and the number of clusters revealed by all independent SNPs for panel S. Red circle indicates the chosen number of clusters (6).



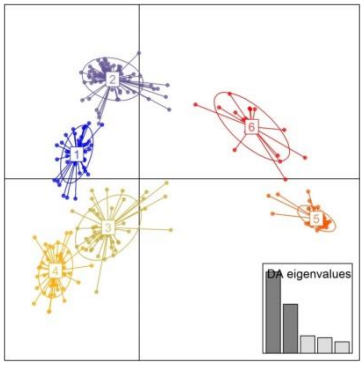
Supplementary Figure 9. Population structure revealed by discriminant analysis of principal components based on all independent SNPs of panel S.



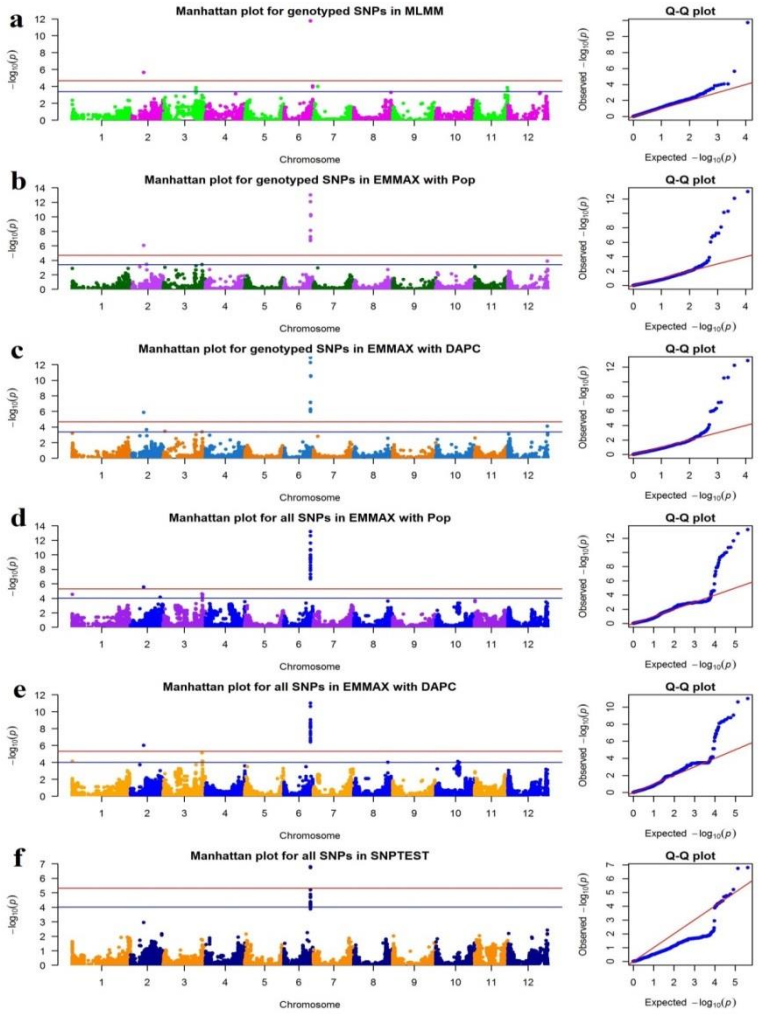
Supplementary Figure 10. Relationship between Bayesian Information Criteria and the number of clusters revealed by all independent SNPs for panel B. Red circle indicates the chosen number of clusters (6).



Supplementary Figure 11. Population structure revealed by discriminant analysis of principal components based on all independent SNPs of panel B.



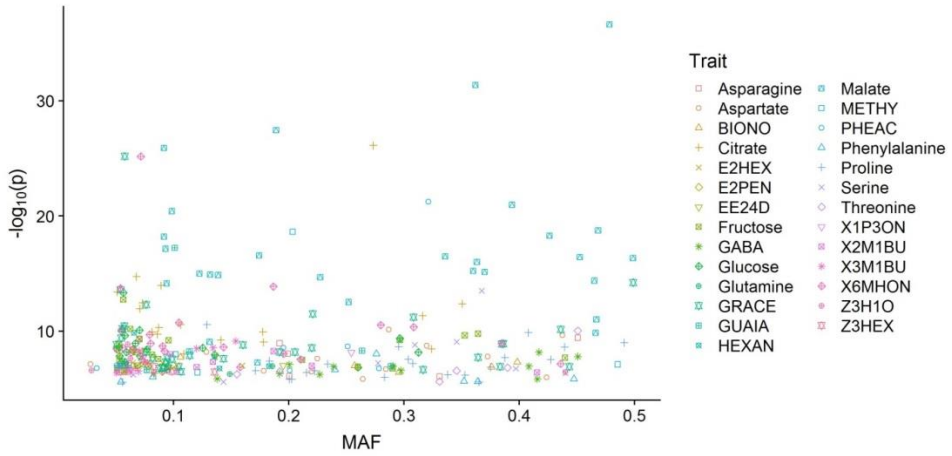
Supplementary Figure 12. Population structure revealed by discriminant analysis of principal components based on all genotyped SNPs of panel B.



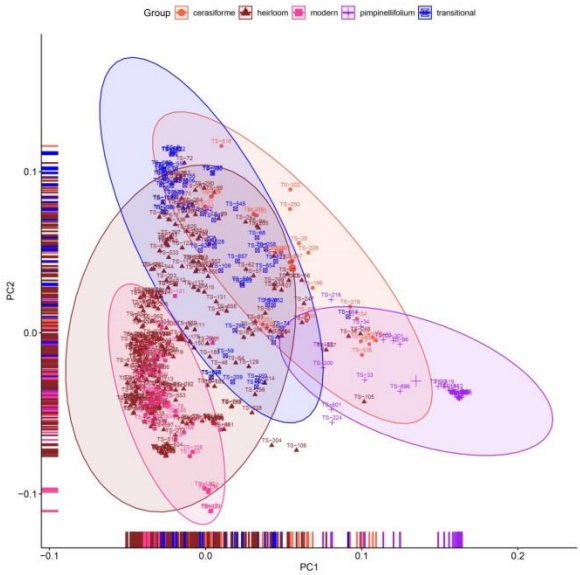
Supplementary Figure 13. Comparison of different GWAS approaches using malate from panel S as an example. (**a**) Manhattan and quantile-quantile (Q-Q) plot for GWAS of genotyped SNPs in multi-locus mixed model (MLMM). (**b**) Manhattan and Q-Q plot for GWAS of genotyped SNPs in efficient mixed-model association expedited (EMMAX) with the cofactor of structure revealed by Structure v2.3.4. (**c**) Manhattan and Q-Q plot for GWAS of genotyped SNPs in EMMAX with the cofactor of structure revealed by discriminant analysis of principal components (DAPC). (**d**) Manhattan and Q-Q plot for GWAS of all (imputed and genotyped) SNPs in EMMAX with the cofactor of structure revealed by Structure v2.3.4. (**e**) Manhattan and Q-Q plot for GWAS of all (imputed and genotyped) SNPs in EMMAX with the cofactor of structure revealed by DAPC. (**f**) Manhattan and Q-Q plot for GWAS of all (imputed and genotyped) SNPs in SNPTEST with the cofactor of the first 20 principal components of kinship and structure.

Supplementary Figures 14-75 are Manhattan and Q-Q plots for all the traits analyzed, which are available on line: https://www.nature.com/articles/s41467-019-09462-w.

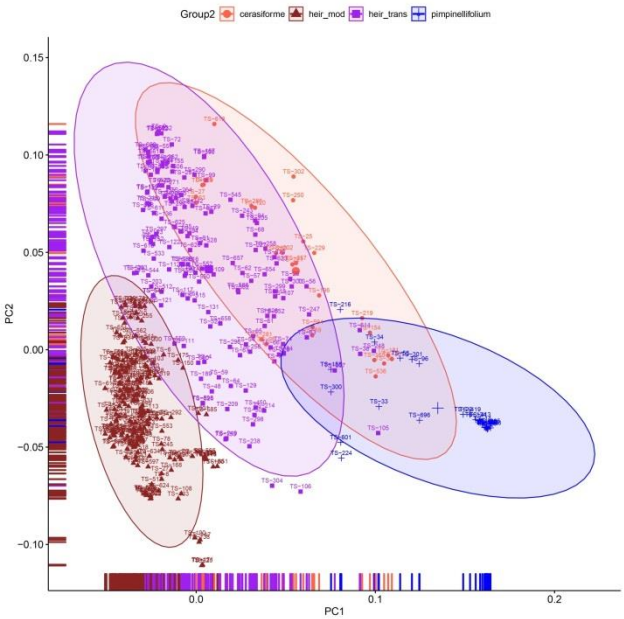
Supplementary Figures 76-123 are the LocusZoom plots of all significant associations identified in the meta0analysis, which are available on line: https://www.nature.com/articles/s41467-019-09462-w.



Supplementary Figure 124. Relationship between minor allele frequency and all significant association *P*-values for different metabolites. Source data are provided as a Source Data file.



Supplementary Figure 125. Principal component analysis of panel T based on independent SNPs. Accessions were previously defined as five clusters based on DAPC analysis. Source data are provided as a Source Data file.



Supplementary Figure 126. Principal component analysis of panel T based on independent SNPs. Accessions were redefined as four main groups as indicated in different colored shapes. Source data are provided as a Source Data file.

Supplementary Data 1-5,9-10 are available on line: https://www.nature.com/articles/s41467-019-09462-w.

Supplementary Data 6. Genomic inflation factors for all traits from three GWAS panels before and after meta-analysis.

Traits	Genomic inflation factor (λ)			
	Panel S	Panel B	Panel T	Meta
Fructose	0.6159032	0.7317135	0.6977581	0.717784
Glucose	0.6744505	0.806336	0.6794763	0.7093804
Malate	0.5707265	0.9136797	0.4129685	0.4664901
Citrate	0.7464719	0.5974109	0.572013	0.6053611
Asparagine	0.7083479	0.7402882	-	0.7941405
Aspartate	1.059803	0.6515481	-	0.8631273
GABA	0.7624631	0.6058975	-	0.7218693
Glutamine	0.7685328	0.8101971	-	-
Lysine	1.176799	1.056351	-	-
Methionine	1.133842	0.8257894	-	-
Phenylalanine	0.6895109	0.7378588	-	0.8525867
Proline	0.6310696	0.679162	-	0.7323248
Serine	0.8970691	0.7190813	-	0.8669158
Threonine	0.7768599	0.8141008	-	0.8614273
E2HEP	-	1.098389	0.9562011	0.9645404
E2HEX	-	0.9546533	0.7569697	0.7877014
E2PEN	-	0.9438439	0.9815885	0.9862116
EE24D	-	0.9679739	0.7956035	0.8194058
Z3H1O	-	0.6943333	0.9398763	0.923969
Z3HEX	-	0.9563462	0.810053	0.836626
X1O3ON	-	1.165792	0.8582511	0.8834625
X1P3ON	-	0.8197511	1.037743	1.017698
X2M1BU	-	0.9711903	0.8148172	0.8353744
X3M1BU	-	1.099926	0.7903689	0.8173063
X6MHON	-	1.048308	0.6081483	0.6498792
BIONO	-	0.6673034	0.8386514	0.8280938
GRACE	-	0.9322994	0.6064222	0.6302904
GUAIA	-	1.213274	0.861091	0.8916016
HEXAN	-	0.9651382	0.8354117	0.8663764
PHEAC	-	0.5993031	0.8780406	0.8583641
METHY	-	1.062173	0.8324349	0.8622192

Supplementary Data 7. Summary of all identified significant associations via meta-analysis for main flavor-related traits in tomato fruit. For each association, associated traits, SNP, chromosome (CHR), position (bp), reference allele (Ref), alternative allele (Alt), minor allele frequency (MAF), beta value of meta,analysis (Beta), standard error of meta analysis (SE), oddor ratio and 95% confidence meta-analysis P value (P), heterogeneity (I2), locus name of the annotated gene (Soly name) and annotated gene (Candidate gene) were provided.

No.	Trait	SNP	CHR	BP	Ref	Alt	Beta	SE	P	I2	Soly name	Candidate gene
1	Citrate	rs01_1749084	1	1749084	c	g	0.111	0.015	3.62E-13	0	Solyc01g007090	Aluminum-activated malate transporter
2	Citrate	**rs01_89465187**	1	8.9E+07	a	g	-0.092	0.014	5.10E-11	0	Solyc01g099150	Lipoxygenase
3	Citrate	rs02_47904426	2	4.8E+07	a	g	-0.048	0.016	4.30E-13	98	Solyc02g084820	Glycosyl transferase group 1
4	Citrate	**rs03_515147**	3	515147	t	g	0.087	0.014	1.89E-10	0	Solyc03g005730	3-isopropylmalate dehydratase large subunit 2
5	Citrate	rs03_52998165	3	5.3E+07	a	c	0.126	0.016	1.84E-15	0	Solyc03g083090	Glycogen synthase
6	Citrate	**rs04_1626376**	4	1626376	a	t	0.075	0.013	9.49E-09	0	Solyc04g007980	1-aminocyclopropane-1-carboxylate oxidase
7	Citrate	**rs04_55737963**	4	5.6E+07	a	g	0.093	0.014	1.14E-10	0	Solyc04g064560	RNA exonuclease
8	Citrate	rs05_5607068	5	5607068	a	g	-0.150	0.020	3.99E-14	0	Solyc05g012350	Hydrolase alpha/beta fold family protein
9	Citrate	**rs05_63225005**	5	6.3E+07	t	c	-0.054	0.009	3.57E-09	0	Solyc05g053120	Glucosyltransferase
10	Citrate	rs06_44955568	6	4.5E+07	a	c	0.066	0.018	7.46E-27	98	Solyc06g072920	Aluminum-activated malate transporter
11	Citrate	rs07_63601724	7	6.4E+07	t	g	0.033	0.005	4.70E-12	0	Solyc07g055840	Citrate synthase
12	Citrate	rs08_55810689	8	5.6E+07	t	c	0.112	0.016	1.15E-12	0	Solyc08g066970	UDP-N-acetylglucosamine/UDP-glucose/GDP-mannose transporter
13	Citrate	rs09_70312902	9	7E+07	t	c	0.105	0.014	1.10E-14	0	Solyc09g090900	3-isopropylmalate dehydratase large subunit 2
14	Citrate	**rs10_40561098**	10	4.1E+07	a	t	0.063	0.010	9.27E-10	0	Solyc10g047340	Hydrolase
15	Citrate	**rs10_61118865**	10	6.1E+07	t	c	-0.118	0.018	7.15E-11	0	Solyc10g079640	Iaa-amino acid hydrolase 9
16	Citrate	**rs10_65378714**	10	6.5E+07	t	g	-0.121	0.021	5.35E-09	0	Solyc10g086580	Ribulose-1 5-bisphosphate carboxylase/oxygenase activase 1
17	Citrate	**rs11_5153986**	11	5153986	a	g	-0.085	0.014	6.16E-10	0	Solyc11g012300	Unknown Protein
18	Citrate	**rs11_51288879**	11	5.1E+07	t	c	-0.133	0.020	2.77E-11	0	Solyc11g065890	1-acyl-sn-glycerol-3-phosphate acyltransferase
19	Citrate	**rs12_3523663**	12	3523663	t	c	-0.028	0.005	1.46E-08	0	Solyc12g010530	Clathrin assembly protein
20	Citrate	**rs12_64042516**	12	6.4E+07	a	g	0.075	0.012	1.16E-10	0	Solyc12g088820	Serine/threonine-protein kinase
21	Fructose	rs01_3327330	1	3327330	a	g	-0.103	0.016	6.37E-11	0	Solyc01g009150	Glycosyl hydrolase
22	Fructose	**rs01_90932104**	1	9.1E+07	a	g	0.114	0.019	3.25E-09	0	Solyc01g101120	Glucan endo-1 3-beta-glucosidase 1
23	Fructose	rs02_47218103	2	4.7E+07	t	c	-0.073	0.011	1.67E-10	0	Solyc02g083990	Calcium-dependent protein kinase CPK1 adapter protein 2-like
24	Fructose	**rs03_1506106**	3	1506106	t	c	-0.095	0.017	3.39E-08	0	Solyc03g006980	Alpha-L-fucosidase 1
25	Fructose	**rs04_5377163**	4	5377163	a	t	0.071	0.013	1.11E-08	0	Solyc04g015200	6-phosphofructokinase

26	Fructose	rs04_52634650	4	5.3E+07	t	g	0.041	0.007	2.07E-08	45	Solyc04g054510	DNA-directed RNA polymerase I subunit RPA12
27	Fructose	**rs05_3403706**	5	3403706	c	g	0.078	0.014	2.57E-08	0	Solyc05g009270	Fatty acid elongase 3-ketoacyl-CoA synthase
28	Fructose	**rs05_63485334**	5	6.3E+07	c	g	0.094	0.015	4.68E-10	0	Solyc05g053400	Glucosyltransferase
29	Fructose	rs06_41390811	6	4.1E+07	a	g	-0.099	0.016	1.30E-09	0	Solyc06g066040	Unknown Protein
30	Fructose	**rs07_2280695**	7	2280695	t	c	0.089	0.017	2.74E-07	0	Solyc07g007620	Processive diacylglycerol glucosyltransferase
31	Fructose	**rs07_48449156**	7	4.8E+07	a	g	0.094	0.017	7.45E-08	0	Solyc07g039600	Unknown Protein
32	Fructose	**rs07_63757414**	7	6.4E+07	a	c	0.095	0.016	4.28E-09	0	Solyc07g055840	Citrate synthase
33	Fructose	**rs08_2487158**	8	2487158	t	c	0.025	0.006	1.28E-07	84	Solyc08g008000	Serine/threonine dehydratase
34	Fructose	**rs08_58158082**	8	5.8E+07	a	g	0.086	0.016	9.75E-08	0	Solyc08g069060	Beta-1 3-galactosyltransferase 6
35	Fructose	**rs08_64470216**	8	6.4E+07	a	g	0.005	0.025	2.33E-10	96	Solyc08g081420	Glycosyltransferase-like protein
36	Fructose	rs09_3477979	9	3477979	a	g	-0.027	0.028	6.16E-10	96	Solyc09g010080	Lin5
37	Fructose	**rs10_422707**	10	422707	a	t	-0.081	0.013	6.27E-10	0	Solyc10g005510	Glyceraldehyde-3-phosphate dehydrogenase
38	Fructose	**rs10_59004825**	10	5.9E+07	t	c	-0.129	0.025	1.59E-07	0	Solyc10g076180	Plant-specific domain TIGR01568 family protein
39	Fructose	**rs10_65465775**	10	6.5E+07	t	c	-0.080	0.014	6.84E-09	0	Solyc10g086720	Fructose-1 6-bisphosphatase class 1
40	Fructose	**rs11_3063738**	11	3063738	a	g	0.092	0.017	3.21E-08	0	Solyc11g008900	Zinc finger CCCH domain-containing protein 66
41	Fructose	rs11_51186147	11	5.1E+07	a	c	-0.119	0.016	1.86E-13	0	Solyc11g065660	SSC11.1
42	Fructose	**rs12_1835753**	12	1835753	c	g	-0.082	0.015	9.68E-08	0	Solyc12g008400	Receptor like kinase
43	Fructose	**rs12_54374393**	12	5.4E+07	a	g	-0.082	0.016	1.12E-07	0	Solyc12g035700	Unknown Protein
44	Fructose	**rs12_64182643**	12	6.4E+07	t	c	-0.033	0.019	1.71E-07	95	Solyc12g089020	Receptor-like protein kinase
45	Glucose	rs01_1998383	1	1998383	a	g	-0.123	0.019	2.36E-10	0	Solyc01g007910	Succinyl-CoA ligase
46	Glucose	**rs01_91816864**	1	9.2E+07	a	g	0.107	0.018	3.92E-09	0	Solyc01g103150	Os03g0731050 protein
47	Glucose	**rs02_43844073**	2	4.4E+07	t	c	-0.085	0.014	2.87E-09	97	Solyc02g079220	Solute carrier family 2%2C facilitated glucose transporter member 8
48	Glucose	**rs03_1506106**	3	1506106	t	c	-0.113	0.020	1.46E-08	0	Solyc03g006980	Alpha-L-fucosidase 1
49	Glucose	**rs04_911809**	4	911809	a	g	-0.120	0.021	6.62E-09	0	Solyc04g007160	Alpha-glucosidase
50	Glucose	**rs04_55433991**	4	5.5E+07	t	c	0.030	0.005	6.96E-09	0	Solyc04g063300	Pentatricopeptide repeat-containing protein
51	Glucose	**rs05_3406424**	5	3406424	t	c	0.099	0.017	1.49E-08	0	Solyc05g009270	Fatty acid elongase 3-ketoacyl-CoA synthase
52	Glucose	**rs05_53075176**	5	5.3E+07	a	c	0.092	0.018	2.27E-07	0	Solyc05g041670	Caffeoyl-CoA O-methyltransferase
53	Glucose	rs05_63485334	5	6.3E+07	c	g	0.115	0.018	8.36E-11	0	Solyc05g053400	Glucosyltransferase
54	Glucose	**rs06_41390811**	6	4.1E+07	a	g	-0.114	0.019	2.51E-09	0	Solyc06g066040	Unknown Protein
55	Glucose	**rs08_2487158**	8	2487158	t	c	0.024	0.005	1.42E-07	69	Solyc08g008000	Serine/threonine dehydratase

56	Glucose	**rs08_58158082**	8	5.8E+07	a	g	0.103	0.019	4.99E-08	0	Solyc08g069060	Beta-1 3-galactosyltransferase 6
57	Glucose	**rs09_3477979**	9	3477979	a	g	-0.028	0.029	4.30E-10	95	Solyc09g010080	Lin5
58	Glucose	**rs10_332069**	10	332069	t	g	-0.118	0.019	1.20E-09	0	Solyc10g005510	Glyceraldehyde-3-phosphate dehydrogenase
59	Glucose	**rs11_5389261**	11	5389261	t	c	-0.066	0.012	1.14E-08	0	Solyc11g012580	Glucan endo-1 3-beta-glucosidase 3
60	Glucose	rs11_51186147	11	5.1E+07	a	c	-0.143	0.019	4.45E-14	0	Solyc11g065660	SSC11.1
61	Glucose	**rs12_3331482**	12	3331482	a	g	-0.085	0.016	9.92E-08	0	Solyc12g010210	Translation initiation factor 6
62	Malate	**rs01_2650772**	1	2650772	t	c	-0.110	0.014	2.08E-15	0	Solyc01g008550	Cinnamoyl CoA reductase-like protein
63	Malate	**rs01_75186951**	1	7.5E+07	t	g	0.112	0.014	4.10E-15	0	Solyc01g066910	PVR3-like protein
64	Malate	**rs01_86203659**	1	8.6E+07	a	g	-0.106	0.013	1.02E-16	0	Solyc01g094800	Chromodomain-helicase-DNA-binding protein 1
65	Malate	**rs01_93076562**	1	9.3E+07	a	t	-0.143	0.018	1.02E-15	0	Solyc01g104640	Helicase
66	Malate	rs02_25428529	2	2.5E+07	a	g	0.182	0.019	3.80E-21	0	Solyc02g030230	UDP-glucose 4-epimerase
67	Malate	rs02_36806140	2	3.7E+07	a	c	0.108	0.012	5.09E-19	0	Solyc02g065650	RING finger protein 13
68	Malate	rs02_48509791	2	4.9E+07	t	c	-0.158	0.014	3.47E-28	0	Solyc02g085660	UDP-glucosyltransferase
69	Malate	rs03_12600405	3	1.3E+07	t	c	0.192	0.022	6.11E-19	0	Solyc03g046430	Unknown Protein
70	Malate	rs03_65842555	3	6.6E+07	t	c	0.177	0.023	7.38E-15	0	Solyc03g116370	Microtubule plus-end binding protein
71	Malate	rs04_2156747	4	2156747	a	g	0.106	0.013	4.45E-17	0	Solyc04g008590	Pyruvate dehydrogenase E1 component subunit beta
73	Malate	rs05_52219368	5	5.2E+07	a	g	-0.212	0.020	1.25E-26	0	Solyc05g041540	Serine carboxypeptidase
74	Malate	**rs06_2927918**	6	2927918	t	c	0.102	0.017	8.55E-10	0	Solyc06g008990	Unknown Protein
75	Malate	rs06_44999916	6	4.5E+07	a	g	-0.147	0.012	2.26E-37	0	Solyc06g072920	Aluminum-activated malate transporter
76	Malate	rs07_38655367	7	3.9E+07	t	g	0.138	0.016	2.68E-17	0	Solyc07g032400	Unknown Protein
77	Malate	**rs07_60030573**	7	6E+07	a	t	0.079	0.012	9.65E-12	0	Solyc07g049670	Alcohol acetyltransferase
78	Malate	**rs08_36375788**	8	3.6E+07	a	c	0.144	0.018	1.20E-15	0	Solyc08g029380	Unknown Protein
79	Malate	rs08_56376169	8	5.6E+07	t	g	-0.146	0.012	4.15E-32	0	Solyc08g067420	Glucose-repressible alcohol dehydrogenase transcriptional effector
80	Malate	rs09_433683	9	433683	a	c	0.105	0.013	3.23E-17	0	Solyc09g005630	Os03g0291800 protein
81	Malate	rs09_49364393	9	4.9E+07	t	g	0.172	0.020	6.94E-18	0	Solyc09g056490	MuDRA transposase-like
82	Malate	rs09_72364359	9	7.2E+07	a	t	-0.135	0.017	1.34E-15	0	Solyc09g098590	Sucrose synthase
83	Malate	**rs10_2557193**	10	2557193	a	g	-0.078	0.012	1.46E-10	0	Solyc11g010480	Threonine endopeptidase
84	Malate	**rs10_37300187**	10	3.7E+07	a	g	0.088	0.016	5.27E-08	0	Solyc10g046790	Unknown Protein
85	Malate	**rs10_59315327**	10	5.9E+07	a	g	-0.103	0.014	2.99E-13	0	Solyc10g076350	Macrophage migration inhibitory factor family protein
86	Malate	rs11_2255630	11	2255630	a	g	-0.121	0.013	1.09E-21	0	Solyc11g008050	Glycogen debranching enzyme
87	Malate	rs11_9681366	11	9681366	a	g	0.106	0.013	3.77E-17	0	Solyc11g018850	Unknown Protein

88	Malate	rs11_55879120	11	5.6E+07	a	c	0.102	0.013	7.14E-16	0	Solyc11g072700	Glycosyltransferase-like protein
89	Malate	rs12_1824226	12	1824226	t	g	-0.119	0.013	1.75E-19	0	Solyc12g008430	Malic enzyme
90	Malate	**rs12_37704315**	12	3.8E+07	a	g	0.106	0.018	5.65E-09	0	Solyc12g044830	Unknown Protein
91	Malate	rs12_64816056	12	6.5E+07	a	c	-0.102	0.013	5.99E-16	0	Solyc12g094640	Glyceraldehyde-3-phosphate dehydrogenase B
92	Asparagine	rs02_54365596	2	5.4E+07	a	g	0.044	0.007	3.72E-10	94	Solyc02g093550	Methyltransferase type 11
93	Asparagine	**rs03_4945916**	3	4945916	t	c	0.117	0.024	7.98E-07	0	Solyc03g033350	Aspartyl protease family protein
94	Asparagine	rs05_62468569	5	6.2E+07	a	g	-0.100	0.017	8.92E-09	0	Solyc05g052170	Acetyltransferase GNAT family protein
95	Asparagine	**rs10_61023555**	10	6.1E+07	t	c	-0.051	0.037	8.61E-07	97	Solyc10g079480	Lycopene beta-cyclase 2
96	Asparagine	rs12_64463407	12	6.4E+07	t	c	0.139	0.023	1.13E-09	0	Solyc12g089350	GDSL esterase/lipase
97	Aspartate	rs03_56191183	3	5.6E+07	a	t	0.160	0.030	7.24E-08	0	Solyc03g095220	Translocase of chloroplast 34
98	Aspartate	**rs03_64586740**	3	6.5E+07	a	t	-0.046	0.007	2.29E-10	0	Solyc03g114690	Wd-40 repeat-containing protein
99	Aspartate	**rs03_70719370**	3	7.1E+07	a	c	-0.070	0.012	2.43E-08	95	Solyc03g124010	Helicase sen1
100	Aspartate	rs04_60700555	4	6.1E+07	t	c	0.052	0.020	1.59E-08	96	Solyc04g074780	Plastid-targeted protein 3
101	Aspartate	**rs05_108012**	5	108012	t	c	-0.034	0.007	1.48E-06	77	Solyc05g005100	Os06g0207500 protein
102	Aspartate	**rs05_64786945**	5	6.5E+07	a	g	-0.062	0.022	2.86E-07	88	Solyc05g055000	Cysteine desulfurase
103	Aspartate	**rs06_37293354**	6	3.7E+07	a	g	-0.043	0.008	1.97E-07	83	Solyc06g054550	Lipase
104	Aspartate	**rs07_63644154**	7	6.4E+07	a	g	-0.025	0.005	1.06E-06	88	Solyc07g055700	Solute carrier family 35 member C2
105	Aspartate	rs08_60307917	8	6E+07	t	c	-0.065	0.011	6.35E-09	0	Solyc08g076350	Abhydrolase domain-containing protein
106	Aspartate	**rs09_2418642**	9	2418642	t	c	0.053	0.010	3.38E-07	0	Solyc09g009080	Repressor of silencing 1
107	Aspartate	rs11_4008385	11	4008385	t	g	0.051	0.008	7.24E-11	0	Solyc11g010960	Alcohol dehydrogenase
108	Aspartate	rs11_55766395	11	5.6E+07	t	c	0.047	0.008	1.86E-09	0	Solyc11g072640	Mitochondrial trans-2-enoyl-CoA reductase
109	Aspartate	**rs12_37536492**	12	3.8E+07	a	t	0.079	0.015	9.16E-08	0	Solyc12g044940	Short-chain dehydrogenase/reductase
110	GABA	**rs01_81646059**	1	8.2E+07	a	t	0.043	0.013	1.21E-07	61	Solyc01g086680	Glutathione S-transferase
111	GABA	**rs02_42305666**	2	4.2E+07	a	g	0.049	0.015	1.37E-06	61	Solyc02g077340	GDSL esterase/lipase At5g45960
112	GABA	rs02_54365596	2	5.4E+07	a	g	0.031	0.005	1.62E-08	94	Solyc02g093550	Methyltransferase type 11
113	GABA	rs03_4945916	3	4945916	t	c	0.156	0.029	8.19E-08	0	Solyc03g033300	Ammonium transporter
114	GABA	**rs03_65004108**	3	6.5E+07	a	t	-0.043	0.067	1.15E-07	97	Solyc03g115200	Glucan endo-1 3-beta-glucosidase
115	GABA	**rs05_60837144**	5	6.1E+07	t	c	0.050	0.010	5.71E-07	0	Solyc05g050700	LRR receptor-like serine/threonine-protein kinase
116	GABA	rs06_1330594	6	1330594	a	c	0.073	0.081	2.69E-07	97	Solyc06g007310	Deoxyribonuclease tatD
117	GABA	**rs09_4768980**	9	4768980	a	t	0.029	0.006	1.45E-06	0	Solyc09g011490	Glutathione S-transferase-like protein
118	GABA	**rs10_61035945**	10	6.1E+07	a	g	0.027	0.005	1.14E-07	0	Solyc10g079480	Lycopene beta-cyclase 2
119	GABA	**rs11_55196715**	11	5.5E+07	t	c	0.037	0.006	6.86E-09	36	Solyc11g071840	Calmodulin binding protein
120	GABA	**rs12_4534284**	12	4534284	a	g	-0.038	0.008	4.69E-07	86	Solyc12g013680	LRR receptor-like serine/threonine-

												protein
121	GABA	**rs12_64463407**	12	6.4E+07	t	c	0.143	0.028	5.15E-07	0	Solyc12g089350	GDSL esterase/lipase
122	Glutamine	**rs02_50640320**	2	5.1E+07	t	c	-0.123	0.025	5.43E-07	0	Solyc02g088630	Glycosyltransferase
123	Glutamine	rs12_64463340	12	6.4E+07	a	t	0.131	0.025	1.08E-07	0	Solyc12g089350	GDSL esterase/lipase
124	Phenylalanine	**rs01_97877551**	1	9.8E+07	t	g	0.036	0.008	2.40E-06	33	Solyc01g111680	Ubiquitin-conjugating enzyme 22
125	Phenylalanine	**rs02_44569659**	2	4.5E+07	a	g	-0.139	0.027	2.23E-07	0	Solyc02g080310	Beta-glucosidase
126	Phenylalanine	**rs03_64488435**	3	6.4E+07	a	t	-0.045	0.009	1.42E-06	0	Solyc03g114500	Enolase
127	Phenylalanine	**rs04_65474317**	4	6.5E+07	a	g	0.241	0.051	2.63E-06	0	Solyc04g081530	Chaperone protein dnaJ
128	Phenylalanine	**rs09_2213892**	9	2213892	c	g	-0.047	0.010	2.31E-06	0	Solyc09g008790	Serine/threonine protein kinase
129	Phenylalanine	**rs11_4002767**	11	4002767	t	c	0.058	0.010	9.57E-09	0	Solyc11g010960	Alcohol dehydrogenase
130	Phenylalanine	**rs12_3388816**	12	3388816	t	c	-0.146	0.030	9.51E-07	0	Solyc12g010330	Response regulator 9
131	Proline	**rs01_9031369**	1	9031369	a	c	0.176	0.034	2.87E-07	0	Solyc01g011420	Unknown Protein
132	Proline	**rs01_66355392**	1	6.6E+07	t	c	0.165	0.034	1.53E-06	0	Solyc01g058380	Coatomer subunit gamma
133	Proline	**rs01_94361411**	1	9.4E+07	a	g	0.051	0.101	2.26E-09	91	Solyc01g106480	Malate dehydrogenase
134	Proline	**rs02_41981476**	2	4.2E+07	a	g	-0.100	0.056	1.35E-07	97	Solyc02g076860	Pollen allergen Phl p 11
135	Proline	rs02_51094633	2	5.1E+07	t	g	-0.256	0.039	2.77E-11	0	Solyc02g089170	Alpha-1 4-glucan-protein synthase
136	Proline	**rs03_52836444**	3	5.3E+07	a	g	-0.595	0.127	2.77E-06	0	Solyc03g082980	Nucleic acid binding protein
137	Proline	rs03_66798980	3	6.7E+07	t	g	-0.204	0.034	2.39E-09	0	Solyc03g117770	Serine incorporator 1
138	Proline	rs04_2268836	4	2268836	t	c	-0.823	0.135	1.00E-09	0	Solyc04g008650	Receptor like kinase
139	Proline	**rs04_59999847**	4	6E+07	t	c	-0.075	0.013	2.87E-08	89	Solyc04g073960	Major facilitator superfamily transporter
140	Proline	**rs05_1860363**	5	1860363	t	g	0.104	0.095	6.02E-07	97	Solyc05g007220	Kelch repeat-containing F-box family protein
141	Proline	**rs05_65206088**	5	6.5E+07	t	g	-0.116	0.023	4.00E-07	90	Solyc05g055630	Palmitoyltransferase PFA4
142	Proline	**rs06_3502385**	6	3502385	t	c	0.079	0.036	6.49E-07	96	Solyc06g009530	Nodal modulator 3
143	Proline	rs06_25117534	6	2.5E+07	a	g	-0.092	0.112	1.37E-10	97	Solyc06g035860	Unknown Protein
144	Proline	**rs06_47277960**	6	4.7E+07	t	c	0.103	0.068	2.78E-07	97	Solyc06g076170	Glucan endo-1 3-beta-glucosidase
145	Proline	**rs07_65315310**	7	6.5E+07	a	c	-0.056	0.012	1.51E-06	0	Solyc07g062620	Serine/threonine-protein kinase receptor
146	Proline	rs08_64272688	8	6.4E+07	t	c	-0.089	0.044	2.07E-08	97	Solyc08g081250	Aminopeptidase
147	Proline	**rs09_3477979**	9	3477979	a	g	0.068	0.012	2.97E-08	74	Solyc09g010090	Beta-fructofuranosidase insoluble isoenzyme
148	Proline	rs09_17881465	9	1.8E+07	t	c	0.116	0.088	4.04E-08	97	Solyc09g019980	Ferric-chelate reductase 1

149	Proline	rs09_49839317	9	5E+07	t	c	-0.100	0.019	7.44E-08	92	Solyc09g057580	Homology to unknown gene
150	Proline	rs09_60202238	9	6E+07	a	g	0.115	0.092	6.42E-08	96	Solyc09g061710	Ribonuclease P protein subunit p25
151	Proline	**rs11_51513717**	11	5.2E+07	a	t	0.147	0.031	2.62E-06	0	Solyc11g065930	Xanthine dehydrogenase/oxidase
152	Proline	**rs12_46135264**	12	4.6E+07	t	c	0.067	0.014	1.02E-06	0	Solyc12g038990	Unknown Protein
153	Serine	**rs02_49223988**	2	4.9E+07	a	g	-0.042	0.013	2.52E-06	93	Solyc02g086560	Cobalamin
154	Serine	rs03_69913055	3	7E+07	a	g	0.063	0.008	3.06E-14	0	Solyc03g121910	Threonine synthase
155	Serine	rs07_64739981	7	6.5E+07	t	g	0.057	0.009	8.35E-10	96	Solyc07g061780	Ubiquitin carboxyl-terminal hydrolase family protein
156	Serine	**rs08_60730948**	8	6.1E+07	t	c	0.041	0.008	1.73E-07	18	Solyc08g076800	Ring H2 finger protein
157	Serine	**rs11_934339**	11	934339	a	g	0.051	0.008	1.52E-09	96	Solyc11g006190	AT-hook motif nuclear localized protein 14
158	Serine	**rs11_55717679**	11	5.6E+07	a	c	-0.037	0.007	4.28E-07	0	Solyc11g072530	V-type proton ATPase subunit a
159	Serine	**rs12_241878**	12	241878	a	c	0.036	0.007	5.89E-08	94	Solyc12g005400	Cyclic nucleotide gated channel
160	Serine	**rs12_64183302**	12	6.4E+07	a	g	0.068	0.014	6.13E-07	0	Solyc12g089020	Receptor-like protein kinase At5g59670
161	Threonine	rs02_54365596	2	5.4E+07	a	g	0.040	0.006	9.18E-11	90	Solyc02g093550	Methyltransferase type 11
162	Threonine	**rs07_64739981**	7	6.5E+07	t	g	0.043	0.008	2.77E-07	94	Solyc07g061780	Ubiquitin carboxyl-terminal hydrolase family protein
163	Threonine	**rs10_61023555**	10	6.1E+07	t	c	-0.032	0.014	2.42E-06	95	Solyc10g079480	Lycopene beta-cyclase 2
164	Threonine	**rs11_6034973**	11	6034973	t	g	-0.042	0.009	5.44E-07	0	Solyc11g013170	Aminotransferase
165	Threonine	**rs11_55671803**	11	5.6E+07	a	c	0.043	0.008	1.46E-07	90	Solyc11g072480	Senescence-associated protein
166	Threonine	**rs12_64463407**	12	6.4E+07	t	c	0.097	0.019	3.33E-07	0	Solyc12g089350	GDSL esterase/lipase
167	BIONO	**rs01_90047419**	1	9E+07	a	g	-0.075	0.015	3.74E-07	0	Solyc01g099940	Pectinesterase
168	BIONO	**rs06_1482454**	6	1482454	a	t	-0.153	0.029	9.99E-08	0	Solyc06g007470	40S ribosomal protein S26
169	BIONO	rs08_62728440	8	6.3E+07	a	g	0.203	0.036	1.87E-08	0	Solyc08g079080	Acid beta-fructofuranosidase
170	BIONO	**rs10_42490535**	10	4.2E+07	t	c	-0.077	0.014	5.08E-08	25	Solyc10g047780	Seed lectin
171	BIONO	**rs11_3559223**	11	3559223	t	c	0.047	0.079	1.29E-07	97	Solyc11g010480	Threonine endopeptidase
172	E2HEX	**rs04_50955974**	4	5.1E+07	a	g	-0.222	0.042	1.78E-07	0	Solyc04g051660	TPR repeat
173	E2HEX	**rs05_64279541**	5	6.4E+07	a	c	-0.428	0.075	1.13E-08	0	Solyc05g054360	Pectinesterase
174	E2HEX	**rs06_48664404**	6	4.9E+07	a	g	0.206	0.039	1.01E-07	0	Solyc06g083120	Unknown Protein
175	E2HEX	**rs08_29549369**	8	3E+07	a	c	-0.492	0.096	3.06E-07	0	Solyc08g022240	Lipase-like protein
176	E2HEX	**rs08_41852822**	8	4.2E+07	a	c	-0.522	0.100	1.99E-07	0	Solyc08g028700	Unknown Protein
177	E2HEX	**rs08_62551973**	8	6.3E+07	a	g	-0.287	0.054	1.20E-07	0	Solyc08g078850	L-lactate dehydrogenase
179	E2HEX	**rs09_70173370**	9	7E+07	a	g	0.411	0.076	5.44E-08	0	Solyc09g090730	Ammonium transporter
180	E2HEX	**rs11_49146698**	11	4.9E+07	a	g	-0.296	0.054	5.12E-08	0	Solyc11g062240	Zinc finger CCCH domain-containing protein 19
181	E2HEX	**rs11_54870921**	11	5.5E+07	t	c	0.230	0.043	9.04E-08	0	Solyc11g071350	Aluminum-activated malate transporter
182	E2HEX	**rs12_64016961**	12	6.4E+07	a	t	-0.352	0.065	5.38E-08	0	Solyc12g088770	Subtilisin-like protease
183	E2PEN	rs09_70173370	9	7E+07	a	g	0.332	0.057	4.39E-09	0	Solyc09g090730	Ammonium transporter

184	EE24D	**rs03_64618191**	3	6.5E+07	a	g	-0.123	0.024	2.89E-07	0	Solyc03g114710	Glucosyltransferase
185	EE24D	**rs07_60351035**	7	6E+07	a	c	0.120	0.023	9.93E-08	0	Solyc07g051840	WRKY transcription factor
186	GRACE	**rs01_9655961**	1	9655961	t	c	-0.619	0.099	4.02E-10	0	Solyc01g011510	Aldehyde dehydrogenase
187	GRACE	**rs01_78675745**	1	7.9E+07	a	g	-0.131	0.025	1.20E-07	0	Solyc01g079600	Lipase
188	GRACE	**rs01_84841566**	1	8.5E+07	t	c	-0.219	0.041	1.21E-07	0	Solyc01g091140	Nitroreductase
189	GRACE	**rs01_95596615**	1	9.6E+07	a	g	0.420	0.075	2.51E-08	0	Solyc01g108230	LAG1 longevity assurance homolog 2
190	GRACE	rs02_40883244	2	4.1E+07	a	g	0.286	0.037	6.00E-15	0	Solyc02g081330	Phytoene synthase 2
191	GRACE	**rs02_48689491**	2	4.9E+07	a	g	0.231	0.041	1.89E-08	0	Solyc02g085880	Cytochrome P450
192	GRACE	rs03_4328514	3	4328514	a	g	0.800	0.076	6.73E-26	0	Solyc03g031860	Phytoene synthase 1
193	GRACE	rs03_19862683	3	2E+07	a	t	0.363	0.065	2.05E-08	0	Solyc03g051630	Unknown Protein
194	GRACE	**rs03_40120552**	3	4E+07	c	g	0.346	0.062	2.70E-08	0	Solyc03g065180	Leucine-rich repeat receptor-like protein kinase
195	GRACE	**rs03_48507215**	3	4.9E+07	t	c	0.348	0.064	4.99E-08	0	Solyc03g077900	Unknown Protein
196	GRACE	**rs03_55675376**	3	5.6E+07	a	g	-0.483	0.080	1.49E-09	0	Solyc03g094010	Glutamate dehydrogenase
197	GRACE	**rs03_63842990**	3	6.4E+07	a	c	0.298	0.056	1.23E-07	0	Solyc03g113790	Mannose-1-phosphate guanyltransferase
198	GRACE	**rs03_70676024**	3	7.1E+07	a	t	-0.344	0.066	1.71E-07	0	Solyc03g123970	Lipid-binding serum glycoprotein
199	GRACE	**rs04_3079421**	4	3079421	t	c	-0.266	0.045	2.85E-09	0	Solyc04g009780	RING finger protein
200	GRACE	rs04_60434091	4	6E+07	a	c	0.334	0.048	3.29E-12	0	Solyc04g074390	UDP-glucuronosyltransferase
201	GRACE	rs05_3171653	5	3171653	a	g	0.474	0.072	3.73E-11	0	Solyc05g008980	Receptor-like protein kinase
202	GRACE	**rs05_10033587**	5	1E+07	a	g	0.221	0.043	2.27E-07	0	Solyc05g015220	Mannosyl-oligosaccharide glucosidase
203	GRACE	**rs05_34292178**	5	3.4E+07	t	c	0.313	0.055	1.35E-08	0	Solyc05g025590	Single-stranded DNA-binding replication protein A large subunit
204	GRACE	**rs05_38579570**	5	3.9E+07	t	g	0.289	0.057	3.33E-07	0	Solyc05g025960	Unknown Protein
205	GRACE	**rs06_37731702**	6	3.8E+07	t	c	-0.257	0.044	6.54E-09	0	Solyc06g059850	3-methyl-2-oxobutanoate dehydrogenase
206	GRACE	**rs06_45007096**	6	4.5E+07	a	t	0.227	0.037	1.20E-09	0	Solyc06g073080	Flavonol synthase/flavanone 3-hydroxylase
207	GRACE	**rs08_2332460**	8	2332460	a	g	-0.417	0.078	8.92E-08	0	Solyc08g007790	Hydroxymethylglutaryl-CoA synthase
208	GRACE	**rs08_60938276**	8	6.1E+07	t	c	0.370	0.066	2.26E-08	0	Solyc08g077000	Ramosa1 C2H2 zinc-finger transcription factor
209	GRACE	**rs09_3790788**	9	3790788	a	c	0.331	0.063	1.43E-07	0	Solyc09g010420	Arginine biosynthesis bifunctional protein
210	GRACE	**rs09_44414362**	9	4.4E+07	c	g	0.402	0.069	6.11E-09	0	Solyc09g055760	T-snare
211	GRACE	**rs09_53512297**	9	5.4E+07	t	c	-0.347	0.066	1.56E-07	0	Solyc09g059170	Anthocyanidin 3-O-glucosyltransferase
212	GRACE	rs09_66583920	9	6.7E+07	a	t	-0.628	0.082	2.56E-14	0	Solyc09g074770	Unknown Protein

213	GRACE	**rs10_2097495**	10	2097495	a	g	0.342	0.057	1.56E-09	0	Solyc10g007930	Cytochrome P450
214	GRACE	rs10_63081709	10	6.3E+07	a	g	-0.457	0.063	4.93E-13	0	Solyc10g083210	Auxin response factor
215	GRACE	rs11_7652084	11	7652084	t	c	-0.286	0.042	6.25E-12	0	Solyc11g016980	Genomic DNA chromosome 5 TAC clone K21L13
216	GRACE	rs11_53251102	11	5.3E+07	c	g	0.493	0.075	6.06E-11	0	Solyc11g068580	Germin-like protein 1
217	GRACE	**rs12_2159033**	12	2159033	a	g	-0.243	0.040	1.32E-09	0	Solyc12g008830	GATA transcription factor 20
218	GRACE	rs12_39603482	12	4E+07	t	c	-0.263	0.040	7.30E-11	0	Solyc12g042850	Unknown Protein
219	GRACE	**rs12_66856936**	12	6.7E+07	c	g	0.405	0.070	6.59E-09	0	Solyc12g099900	GRAS family transcription factor
220	GUAIA	**rs02_42087438**	2	4.2E+07	t	c	-0.294	0.055	1.12E-07	0	Solyc02g077010	Lipase-like
221	GUAIA	rs04_1372745	4	1372745	a	g	0.636	0.116	4.14E-08	0	Solyc04g007690	Auxin efflux carrier
222	GUAIA	**rs04_51282247**	4	5.1E+07	a	g	-0.654	0.127	2.36E-07	0	Solyc04g052940	Unknown Protein
223	GUAIA	rs05_6004337	5	6004337	a	g	0.345	0.059	4.94E-09	0	Solyc05g012800	UDP-N-acetylmuramoylalanine--D-glutamate ligase
224	GUAIA	**rs05_20888190**	5	2.1E+07	a	g	-0.601	0.113	9.16E-08	0	Solyc05g018430	Genome polyprotein
225	GUAIA	rs09_69299940	9	6.9E+07	a	g	0.728	0.084	5.90E-18	0	Solyc09g089560	NSGT1
226	HEXAN	**rs01_1083181**	1	1083181	c	g	0.407	0.063	1.45E-10	0	Solyc01g006540	Lipoxygenase
227	HEXAN	**rs04_58252797**	4	5.8E+07	t	c	0.215	0.041	1.10E-07	0	Solyc04g071320	Protein FAM188A
228	HEXAN	**rs08_40214293**	8	4E+07	a	c	-0.611	0.115	1.06E-07	0	Solyc08g028780	AT5G28150-like protein
229	HEXAN	rs09_50436898	9	5E+07	t	c	0.295	0.057	1.75E-07	0	Solyc09g057750	Unknown Protein
230	HEXAN	**rs09_62965099**	9	6.3E+07	a	c	0.324	0.057	1.73E-08	0	Solyc09g065110	Unknown Protein
231	METHY	**rs02_45301168**	2	4.5E+07	t	g	0.447	0.083	7.79E-08	0	Solyc02g081320	SET domain-containing protein
232	METHY	**rs04_51282247**	4	5.1E+07	a	g	-1.048	0.202	2.27E-07	0	Solyc04g052940	Unknown Protein
233	METHY	rs05_9831567	5	9831567	t	c	-0.818	0.143	1.02E-08	0	Solyc05g015100	Unknown Protein
234	METHY	**rs09_38016657**	9	3.8E+07	a	c	0.664	0.131	3.73E-07	0	Solyc09g055240	Monooxygenase family protein
235	METHY	rs09_69293875	9	6.9E+07	a	g	-1.142	0.127	2.34E-19	0	Solyc09g089580	1-aminocyclopropane-1-carboxylate oxidase-like protein
236	PHEAC	**rs02_53977499**	2	5.4E+07	a	c	-0.081	0.015	1.72E-07	56	Solyc02g093080	1-aminocyclopropane-1-carboxylate oxidase
237	PHEAC	rs04_55635636	4	5.6E+07	c	g	0.107	0.016	5.59E-22	98	Solyc04g064490	Glycosyltransferase
238	PHEAC	**rs06_24606676**	6	2.5E+07	t	c	0.278	0.053	1.68E-07	0	Solyc06g035580	Choline dehydrogenase
239	PHEAC	**rs07_18999907**	7	1.9E+07	t	c	-0.286	0.055	1.79E-07	0	Solyc07g021510	Acetyl xylan esterase
240	PHEAC	**rs07_37870453**	7	3.8E+07	a	t	0.280	0.052	6.43E-08	0	Solyc07g032330	Unknown Protein
241	PHEAC	**rs07_45966785**	7	4.6E+07	a	g	-0.295	0.055	7.00E-08	0	Solyc07g039200	Guanine nucleotide-binding protein subunit beta
242	PHEAC	**rs07_53292212**	7	5.3E+07	a	g	-0.289	0.054	9.67E-08	0	Solyc07g041480	Unknown Protein
243	PHEAC	**rs08_52258545**	8	5.2E+07	t	g	-0.157	0.026	2.74E-09	0	Solyc08g062870	Unknown Protein
244	PHEAC	rs09_69805632	9	7E+07	t	c	-0.067	0.011	2.09E-09	71	Solyc10g008450	F-box family protein
245	PHEAC	**rs12_64989295**	12	6.5E+07	t	c	0.109	0.021	2.67E-07	0	Solyc12g095880	Phosphofructokinase family protein
246	X1P3ON	rs05_3036212	5	3036212	a	g	-0.059	0.010	7.07E-09	0	Solyc05g008800	Lipid phosphate phosphatase 3
247	X2M1BU	rs01_3842886	1	3842886	t	g	-0.129	0.025	3.96E-07	0	Solyc01g009610	Unknown Protein

248	X2M1BU	rs02_48923294	2	4.9E+07	c	g	0.456	0.073	5.21E-10	0	Solyc02g086170	Unknown Protein
249	X2M1BU	rs03_62062078	3	6.2E+07	a	g	-0.425	0.074	1.06E-08	0	Solyc03g111460	Nuclear transcription factor Y subunit C-2
250	X2M1BU	**rs05_17569205**	5	1.8E+07	t	c	0.917	0.168	4.82E-08	0	Solyc05g017760	Acetyl-CoA C-acetyltransferase
251	X2M1BU	**rs05_55508538**	5	5.6E+07	a	g	1.139	0.216	1.35E-07	0	Solyc05g043310	GDSL esterase/lipase
252	X2M1BU	**rs05_64274113**	5	6.4E+07	t	c	-0.634	0.122	2.24E-07	0	Solyc05g054350	Epoxide hydrolase
253	X2M1BU	rs06_37782796	6	3.8E+07	a	g	0.179	0.031	5.50E-09	0	Solyc06g059850	3-methyl-2-oxobutanoate dehydrogenase
254	X2M1BU	**rs07_1815826**	7	1815826	t	c	0.098	0.018	1.04E-07	72	Solyc07g007010	Pentatricopeptide
255	X2M1BU	**rs09_65686608**	9	6.6E+07	a	c	-0.796	0.154	2.18E-07	0	Solyc09g073020	Unknown Protein
256	X2M1BU	**rs12_6984**	12	6984	t	c	0.582	0.113	2.65E-07	0	Solyc12g005010	Os04g0625000 protein
257	X3M1BU	**rs01_45493960**	1	4.5E+07	a	g	0.851	0.166	3.08E-07	0	Solyc01g049640	Polynucleotidyl transferase Ribonuclease H
258	X3M1BU	**rs01_72555420**	1	7.3E+07	c	g	-1.333	0.245	5.49E-08	0	Solyc01g065910	Periaxin-like protein
259	X3M1BU	rs02_48923294	2	4.9E+07	c	g	0.513	0.088	5.10E-09	0	Solyc02g086170	Unknown Protein
260	X3M1BU	**rs03_55675376**	3	5.6E+07	a	g	0.212	0.103	4.04E-07	0	Solyc03g094010	Glutamate dehydrogenase
261	X3M1BU	**rs03_59783522**	3	6E+07	t	c	-0.462	0.088	1.66E-07	0	Solyc03g097470	NADH dehydrogenase
262	X3M1BU	rs04_54285533	4	5.4E+07	t	g	-0.572	0.093	7.25E-10	0	Solyc04g056490	Transcription factor bZIP38
263	X3M1BU	rs05_17569205	5	1.8E+07	t	c	1.202	0.202	2.86E-09	0	Solyc05g017760	Acetyl-CoA C-acetyltransferase
264	X3M1BU	rs05_55508538	5	5.6E+07	a	g	1.566	0.265	3.24E-09	0	Solyc05g043310	GDSL esterase/lipase
265	X3M1BU	**rs05_64102511**	5	6.4E+07	t	c	0.804	0.148	5.91E-08	0	Solyc05g054140	Nucleic acid-binding OB-fold
266	X3M1BU	**rs06_34695214**	6	3.5E+07	t	c	-0.523	0.102	2.89E-07	0	Solyc06g051370	Ycf2
267	X3M1BU	**rs07_31901278**	7	3.2E+07	a	g	-1.148	0.218	1.34E-07	0	Solyc07g026850	Unknown Protein
268	X3M1BU	**rs10_11719734**	10	1.2E+07	t	g	-1.092	0.215	3.93E-07	0	Solyc10g019100	Exocyst complex component 5
269	X6MHON	**rs01_9655961**	1	9655961	t	c	-0.491	0.076	9.55E-11	0	Solyc01g011510	Aldehyde dehydrogenase
270	X6MHON	**rs01_20299004**	1	2E+07	t	g	0.445	0.075	2.80E-09	0	Solyc01g016610	Branched-chain alpha keto-acid dehydrogenase E1 alpha subunit
271	X6MHON	**rs01_31491684**	1	3.1E+07	t	c	-0.443	0.077	7.38E-09	0	Solyc01g020530	Unknown Protein
272	X6MHON	**rs01_51561195**	1	5.2E+07	a	c	-0.419	0.076	3.04E-08	0	Solyc01g055200	S8 self-incompatibility ribonuclease
273	X6MHON	**rs01_55678900**	1	5.6E+07	a	g	-0.461	0.078	3.28E-09	0	Solyc01g056670	NADH-quinone oxidoreductase subunit K
274	X6MHON	**rs01_76239450**	1	7.6E+07	t	c	-0.384	0.066	5.38E-09	0	Solyc01g067510	Receptor-like kinase
275	X6MHON	**rs01_95596615**	1	9.6E+07	a	g	0.342	0.057	2.45E-09	0	Solyc01g108230	LAG1 longevity assurance homolog 2
276	X6MHON	rs02_40378276	2	4E+07	a	g	-0.261	0.034	1.33E-14	0	Solyc02g070770	NAD-dependent epimerase/dehydratase
277	X6MHON	rs03_3212583	3	3212583	t	c	-0.600	0.057	6.76E-26	0	Solyc03g025720	Long-chain-fatty-acid--CoA ligase
278	X6MHON	**rs03_55675376**	3	5.6E+07	a	g	-0.354	0.060	3.62E-09	0	Solyc03g094010	Glutamate dehydrogenase
279	X6MHON	**rs04_1185013**	4	1185013	a	t	0.250	0.047	1.34E-07	0	Solyc04g007510	ATP-dependent RNA helicase A-like protein

280	X6MHON	rs04_60345897	4	6E+07	a	t	0.237	0.036	3.00E-11	0	Solyc04g074360	UDP-glucuronosyltransferase
281	X6MHON	**rs05_4185523**	5	4185523	a	g	-0.292	0.051	1.38E-08	0	Solyc05g009930	Chloroplast unusual positioning
282	X6MHON	**rs06_47572597**	6	4.8E+07	t	c	0.312	0.058	9.03E-08	0	Solyc06g076600	Receptor-like protein kinase
283	X6MHON	**rs08_64306757**	8	6.4E+07	a	g	0.231	0.042	2.81E-08	0	Solyc08g081240	Hydroxyproline-rich glycoprotein family protein
284	X6MHON	**rs09_3790788**	9	3790788	a	c	0.284	0.047	2.01E-09	0	Solyc09g010420	Arginine biosynthesis bifunctional protein
285	X6MHON	rs09_60851387	9	6.1E+07	c	g	0.309	0.046	1.85E-11	0	Solyc09g062960	Unknown Protein
286	X6MHON	rs09_66583920	9	6.7E+07	a	t	-0.476	0.062	1.86E-14	0	Solyc09g074770	Unknown Protein
287	X6MHON	**rs10_41126678**	10	4.1E+07	a	g	-0.621	0.102	1.21E-09	0	Solyc10g047570	Polygalacturonase
288	X6MHON	**rs10_61007386**	10	6.1E+07	a	g	-0.205	0.036	9.28E-09	0	Solyc10g079470	L-galactono-1%2C4-lactone dehydrogenase
289	X6MHON	rs11_7652084	11	7652084	t	c	-0.207	0.032	4.40E-11	0	Solyc11g016980	Genomic DNA chromosome 5 TAC clone K21L13
290	X6MHON	rs11_53251102	11	5.3E+07	c	g	0.371	0.057	6.37E-11	0	Solyc11g068580	Germin-like protein 1
291	X6MHON	**rs12_2159033**	12	2159033	a	g	-0.184	0.030	1.17E-09	0	Solyc12g008830	GATA transcription factor 20
292	X6MHON	**rs12_39603482**	12	4E+07	t	c	-0.166	0.031	6.98E-08	0	Solyc12g042850	Unknown Protein
293	X6MHON	**rs12_57980130**	12	5.8E+07	a	c	-0.269	0.053	3.31E-07	0	Solyc12g035470	Glycosyl transferase family 17 protein
294	X6MHON	**rs12_63747014**	12	6.4E+07	a	g	-0.344	0.054	2.04E-10	0	Solyc12g088300	Unknown Protein
295	Z3H1O	**rs01_1134235**	1	1134235	a	g	0.600	0.118	3.94E-07	0	Solyc01g006540	Lipoxygenase
296	Z3H1O	rs03_9251887	3	9251887	a	c	-0.894	0.160	2.36E-08	0	Solyc03g044440	Ribosomal protein L22 family protein
297	Z3H1O	**rs06_40009759**	6	4E+07	t	c	-0.685	0.131	1.78E-07	0	Solyc06g063300	Kelch-domain-containing protein
298	Z3H1O	rs07_20023294	7	2E+07	t	c	1.018	0.168	1.40E-09	0	Solyc07g021610	Unknown Protein
299	Z3H1O	**rs08_2660808**	8	2660808	a	t	0.594	0.117	3.86E-07	0	Solyc08g008210	V-type proton ATPase subunit E
300	Z3H1O	**rs08_54105697**	8	5.4E+07	t	c	0.163	0.032	2.64E-07	0	Solyc08g065810	Lysine-specific demethylase 3B
301	Z3H1O	rs09_69639016	9	7E+07	a	g	-0.766	0.141	4.88E-08	0	Solyc09g090080	Inorganic phosphate transporter
302	Z3H1O	rs10_55323486	10	5.5E+07	a	c	-0.884	0.167	1.11E-07	0	Solyc10g054410	Unknown Protein NA
303	Z3H1O	**rs12_2120769**	12	2120769	a	c	-0.062	0.012	3.69E-07	0	Solyc12g008790	Lipid Acyl-transferase
304	Z3HEX	**rs06_45350584**	6	4.5E+07	a	c	-0.249	0.049	3.14E-07	0	Solyc06g073580	1-aminocyclopropane-1-carboxylate oxidase 1
305	Z3HEX	**rs06_48719569**	6	4.9E+07	a	g	0.206	0.039	1.18E-07	0	Solyc06g083190	Peptidyl-prolyl cis-trans isomerase
306	Z3HEX	**rs08_40214293**	8	4E+07	a	c	-0.528	0.103	3.26E-07	0	Solyc08g028780	AT5G28150-like protein
307	Z3HEX	**rs09_62965099**	9	6.3E+07	a	c	0.270	0.053	3.58E-07	0	Solyc09g065110	Unknown Protein

Note: The newly identified loci are highlighted by bold blue. All SNP positions are aligned on the tomato genome version 2.50.

Supplementary Data 8. Detail information for singular enrichment analysis. Significant enriched processes and groups were indicated in bold.

GO term	Ontology	Description	No. in input list	No. in BG/Ref	P	FDR
GO:0005975	P	**carbohydrate metabolic process**	11	311	0.00017	0.011
GO:0006629	P	lipid metabolic process	6	160	0.004	0.13
GO:0008152	P	metabolic process	47	3665	0.0078	0.16
GO:0055114	P	oxidation reduction	9	504	0.043	0.68
GO:0044267	P	cellular protein metabolic process	6	676	0.58	1
GO:0044260	P	cellular macromolecule metabolic process	12	1688	0.85	1
GO:0044238	P	primary metabolic process	28	2595	0.2	1
GO:0019538	P	protein metabolic process	7	1050	0.85	1
GO:0044237	P	cellular metabolic process	17	2024	0.67	1
GO:0043687	P	post-translational protein modification	5	490	0.46	1
GO:0009987	P	cellular process	18	2440	0.86	1
GO:0009058	P	biosynthetic process	7	965	0.78	1
GO:0006464	P	protein modification process	5	499	0.48	1
GO:0043170	P	macromolecule metabolic process	13	2090	0.95	1
GO:0044249	P	cellular biosynthetic process	5	899	0.91	1
GO:0043412	P	macromolecule modification	5	525	0.52	1
GO:0006807	P	nitrogen compound metabolic process	5	1153	0.98	1
GO:0044281	P	small molecule metabolic process	5	253	0.085	1
GO:0003824	F	**catalytic activity**	87	6061	5.80E-07	9.20E-05
GO:0008194	F	**UDP-glycosyltransferase activity**	10	168	5.10E-06	0.00027
GO:0016757	F	**transferase activity, transferring glycosyl groups**	15	394	4.80E-06	0.00027
GO:0016758	F	**transferase activity, transferring hexosyl groups**	12	262	7.60E-06	0.0003
GO:0016740	F	**transferase activity**	36	2180	0.00033	0.01
GO:0016491	F	**oxidoreductase activity**	20	960	0.00056	0.015
GO:0035251	F	**UDP-glucosyltransferase activity**	6	113	0.00074	0.017
GO:0046527	F	**glucosyltransferase activity**	6	143	0.0023	0.042
GO:0016746	F	**transferase activity, transferring acyl groups**	8	247	0.0023	0.042
GO:0008415	F	acyltransferase activity	6	177	0.0064	0.1
GO:0016747	F	transferase activity, transferring acyl groups other than	6	222	0.018	0.26
GO:0022891	F	substrate-specific transmembrane transporter activity	11	646	0.036	0.49
GO:0015291	F	secondary active transmembrane transporter activity	5	214	0.049	0.6
GO:0022892	F	substrate-specific transporter activity	12	786	0.058	0.62
GO:0022857	F	transmembrane transporter activity	12	786	0.058	0.62
GO:0016829	F	lyase activity	5	250	0.082	0.82
GO:0022804	F	active transmembrane transporter activity	7	419	0.092	0.87
GO:0015075	F	ion transmembrane transporter activity	7	431	0.1	0.91
GO:0008324	F	cation transmembrane transporter activity	5	281	0.12	0.97
GO:0005215	F	transporter activity	13	989	0.12	0.97
GO:0005488	F	binding	50	8209	1	1
GO:0003676	F	nucleic acid binding	9	2178	1	1
GO:0003677	F	DNA binding	7	1425	0.98	1
GO:0043167	F	ion binding	6	1148	0.95	1
GO:0008270	F	zinc ion binding	5	835	0.88	1
GO:0000166	F	nucleotide binding	6	773	0.71	1
GO:0004674	F	protein serine/threonine kinase activity	6	625	0.51	1
GO:0004672	F	protein kinase activity	6	772	0.71	1
GO:0046872	F	metal ion binding	6	1148	0.95	1
GO:0016787	F	hydrolase activity	20	2055	0.41	1
GO:0043169	F	cation binding	6	1148	0.95	1
GO:0016788	F	hydrolase activity, acting on ester bonds	10	838	0.24	1
GO:0046914	F	transition metal ion binding	5	952	0.94	1
GO:0016772	F	transferase activity, transferring phosphorus-containing	8	1098	0.79	1
GO:0016301	F	kinase activity	7	991	0.8	1
GO:0004091	F	carboxylesterase activity	5	299	0.14	1
GO:0016773	F	phosphotransferase activity, alcohol group as acceptor	6	848	0.79	1
GO:0005515	F	protein binding	20	3878	1	1
GO:0032991	C	macromolecular complex	7	631	0.35	1
GO:0044424	C	intracellular part	8	1381	0.94	1
GO:0005737	C	cytoplasm	8	701	0.31	1
GO:0005623	C	cell	15	2766	0.99	1
GO:0016020	C	membrane	5	1114	0.98	1
GO:0044444	C	cytoplasmic part	7	575	0.27	1
GO:0044464	C	cell part	15	2766	0.99	1
GO:0043234	C	protein complex	5	336	0.2	1
GO:0005622	C	intracellular	10	1736	0.96	1

Appendix 3

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-09462-w OPEN

Meta-analysis of genome-wide association studies provides insights into genetic control of tomato flavor

Jiantao Zhao[1], Christopher Sauvage[1,6], Jinghua Zhao[2,7], Frédérique Bitton[1], Guillaume Bauchet[1,8], Dan Liu[3], Sanwen Huang[3,4], Denise M. Tieman[5], Harry J. Klee[5] & Mathilde Causse[1]

Tomato flavor has changed over the course of long-term domestication and intensive breeding. To understand the genetic control of flavor, we report the meta-analysis of genome-wide association studies (GWAS) using 775 tomato accessions and 2,316,117 SNPs from three GWAS panels. We discover 305 significant associations for the contents of sugars, acids, amino acids, and flavor-related volatiles. We demonstrate that fruit citrate and malate contents have been impacted by selection during domestication and improvement, while sugar content has undergone less stringent selection. We suggest that it may be possible to significantly increase volatiles that positively contribute to consumer preferences while reducing unpleasant volatiles, by selection of the relevant allele combinations. Our results provide genetic insights into the influence of human selection on tomato flavor and demonstrate the benefits obtained from meta-analysis.

[1] INRA, UR1052, Génétique et Amélioration des Fruits et Légumes, Domaine Saint Maurice, 67 Allée des Chênes CS 60094, 84143 Montfavet Cedex, France. [2] MRC Epidemiology Unit & Institute of Metabolic Science, University of Cambridge, Addrenbrooke's Hospital, Box 285Hills Road, Cambridge CB2 0QQ, UK. [3] Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, 518124 Shenzhen, Guangdong, China. [4] Key Laboratory of Biology and Genetic Improvement of Horticultural Crops of the Ministry of Agriculture, Sino-Dutch Joint Laboratory of Horticultural Genomics, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, 100081 Beijing, China. [5] Horticultural Sciences, Plant Innovation Center, University of Florida, Post Office Box 110690Gainesville, FL 32611, USA. [6] Present address: Syngenta, 12 Chemin de l'Hobit, Saint Sauveur 31790, France. [7] Present address: Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, Strangeways Research Laboratory, University of Cambridge, Wort's Causeway, Cambridge CB1 8RN, UK. [8] Present address: Boyce Thompson Institute, Cornell University, 533 Tower Rd, Ithaca, NY 14853, USA. Correspondence and requests for materials should be addressed to M.C. (email: mathilde.causse@inra.fr)

The deterioration of tomato flavor has been a source of complaint from consumers for decades[1]. During long-term domestication and breeding history, flavor has not been a priority, in contrast to yield, disease resistance, and postharvest shelf life[1,2]. However, flavor is one of the most important traits for improving tomato sensory quality and consumer acceptability[3]. Flavor is centrally influenced by sugars, acids, amino acids and a diverse set of volatiles[4–6]. Most of these compounds are quantitatively inherited as shown by many QTL studies but only a few QTLs have been positionally cloned[7]. Genome-wide association studies (GWAS) have detected many significant associated loci for tomato flavor related traits[6,8–12] However, reducing a QTL to a causative gene is difficult and only a few candidate genes have been functionally validated[7]. The underlying genetic control of tomato flavor is still incomplete and remains an important breeding target.

Meta-analysis of genome-wide associations is powerful in dissecting complex human diseases[13,14]. A recent meta-analysis in cattle stature also demonstrated its power in non-human species[15]. However, to the best of our knowledge, no GWAS meta-analysis has been reported in major crops, despite the increasing number of GWAS studies in major crops, such as rice. To date, the genomes of over 500 tomato accessions have been fully sequenced[6,12,16–19], making it possible to perform genotype imputation[20,21] and subsequent meta-analysis of GWAS using summary data[14] to decipher the polygenic architecture of agronomic traits. In this study, we perform a meta-GWAS on 775 tomato accessions and 2,316,117 SNPs and discover 305 significant associations for diverse flavor-related traits. Our results provide genetic insights into tomato flavor.

Results

Meta-analysis. Here we report the first meta-analysis of GWAS in tomato using results of three publicly available GWAS panels: 163 tomato accessions from panel S[8], 291 accessions from panel B[11], and 402 accessions from panel T[6] (Fig. 1). We analyzed a large set of tomato flavor-related quality chemicals, including sugars, organic acids, amino acids, and volatiles measured in each of these panels.

First, we used IMPUTE2 software[22] to increase the genome-wide SNP densities of panel S[8] and panel B[11], which were genotyped using SNP arrays (Online methods). After quality control (Supplementary Figs. 1–3, Supplementary Tables 1 and 2, Supplementary Data 1–3), a total of 209,152 and 252,414 SNPs was retained for panel S and B, respectively. Imputation greatly increased the density of genomic coverage (Supplementary Fig. 4) and revealed a similar genetic population structure compared

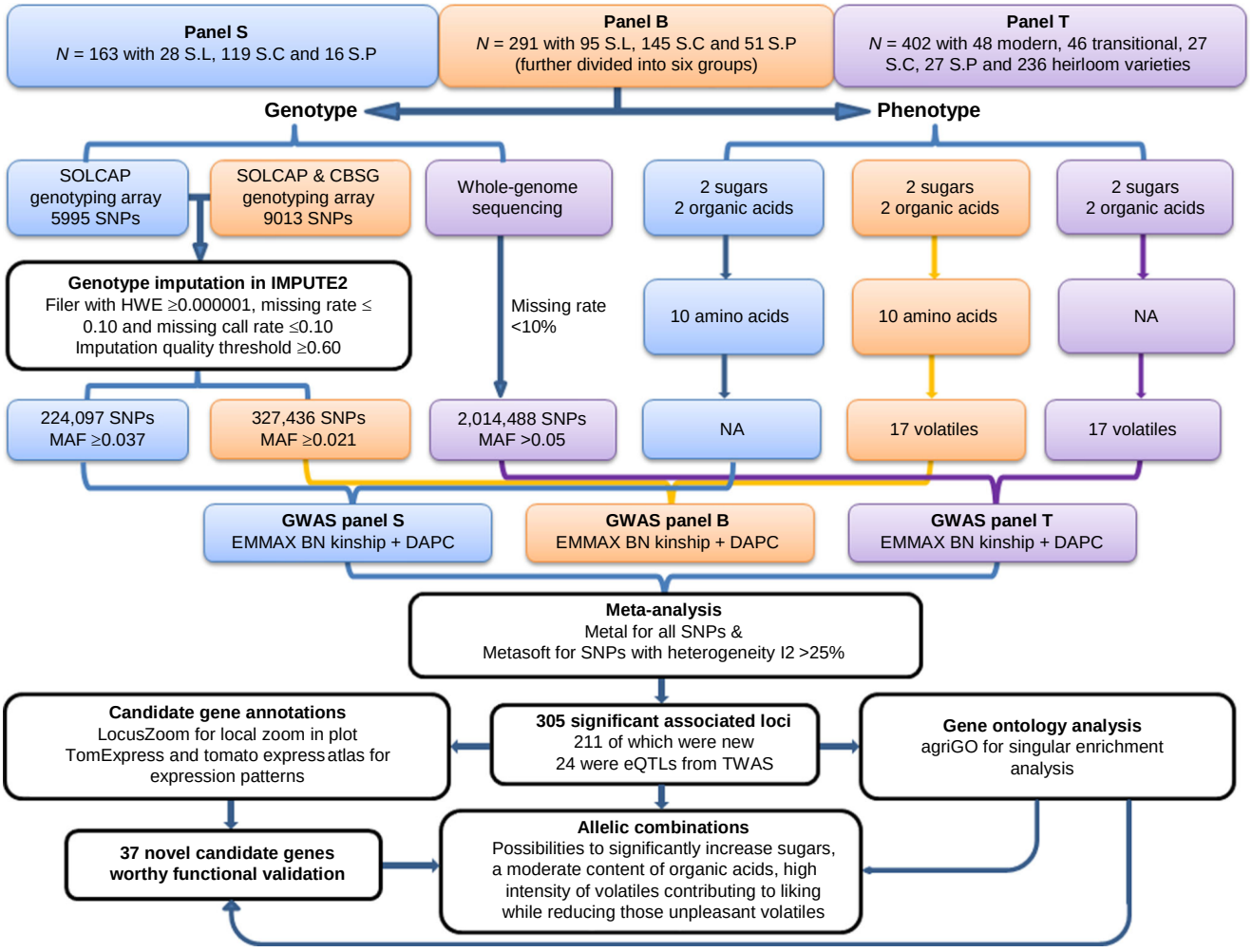


Fig. 1 Overview of study design. N, the number of individuals; S.L, *S. lycopersicum*; S.C, *S. lycopersicum* var *cerasiforme*; S.P, *S. pimpinellifolium*; Genotyping arrays: SOLCAP, Solanaceae Coordinated Agricultural Project; CBSG, Centre of Biosystems Genomics consortium; HWE, Hardy–Weinberg equilibrium; MAF, minor allele frequency; GWAS, genome-wide association study; EMMAX, Efficient Mixed-Model Association eXpedited; DAPC, Discriminant Analysis of Principal Components; eQTL, expression quantitative trait locus; TWAS, transcriptome-wide association study

Table 1 Summary of 37 candidate genes associated with main flavor-related traits in tomato fruit[a]

Trait	Chr	BP	Ref	Alt	*P*	I^2	Locus name	Candidate gene
Citrate	1	1749084	c	g	3.62×10^{-13}	0	Solyc01g007090	Aluminum-activated malate transporter
Citrate	2	47904426	a	g	4.30×10^{-13}	97.9	Solyc02g084820	Glycosyl transferase group 1
Citrate	3	52998165	a	c	1.84×10^{-15}	0	Solyc03g083090	Glycogen synthase
Citrate	6	44955568	a	c	7.46×10^{-27}	98.4	Solyc06g072920	Aluminum-activated malate transporter
Citrate	7	63601724	t	g	4.70×10^{-12}	0	Solyc07g055840	Citrate synthase
Fructose	1	3327330	a	g	6.37×10^{-11}	0	Solyc01g009150	Glycosyl hydrolase
Fructose	5	63485334	c	g	4.68×10^{-10}	0	Solyc05g053400[a]	Glucosyltransferase
Fructose	7	63757414	a	c	4.28×10^{-09}	0	Solyc07g055840	Citrate synthase
Fructose	8	64470216	a	g	2.33×10^{-10}	96.2	Solyc08g081420	Glycosyltransferase-like protein
Fructose	10	422707	a	t	6.27×10^{-10}	0	Solyc10g005510[a]	Glyceraldehyde-3-phosphate dehydrogenase
Fructose	10	65465775	t	c	6.84×10^{-09}	0	Solyc10g086720	Fructose-1 6-bisphosphatase class 1
Glucose	1	1998383	a	g	2.36×10^{-10}	0	Solyc01g007910	Succinyl-CoA ligase
Glucose	2	43844073	t	c	2.87×10^{-09}	96.7	Solyc02g079220	Solute carrier family facilitated glucose transporter member 8
Glucose	4	911809	a	g	6.62×10^{-09}	0	Solyc04g007160	Alpha-glucosidase
Glucose	8	58158082	a	g	4.99×10^{-08}	0	Solyc08g069060	Beta-1 3-galactosyltransferase 6
Glucose	10	332069	t	g	1.20×10^{-09}	0	Solyc10g005510[a]	Glyceraldehyde-3-phosphate dehydrogenase
Malate	1	2650772	t	c	2.08×10^{-15}	0	Solyc01g008550	Cinnamoyl CoA reductase-like protein
Malate	9	72364359	a	t	1.34×10^{-15}	0	Solyc09g098590	Sucrose synthase
Malate	11	55879120	a	c	7.14×10^{-16}	0	Solyc11g072700	Glycosyltransferase-like protein
Malate	12	1824226	t	g	1.75×10^{-19}	0	Solyc12g008430	Malic enzyme
Asparagine	2	54365596	a	g	3.72×10^{-10}	94	Solyc02g093550[a]	Methyltransferase type 11
Asparagine	5	62468569	a	g	8.92×10^{-09}	0	Solyc05g052170	Acetyltransferase GNAT family protein
Asparagine	12	64463407	t	c	1.13×10^{-09}	0	Solyc12g089350	GDSL esterase/lipase
Aspartate	8	60307917	t	c	6.35×10^{-09}	0	Solyc08g076350	Abhydrolase domain-containing protein
Aspartate	11	4008385	t	g	7.24×10^{-11}	0	Solyc11g010960	Alcohol dehydrogenase
Aspartate	12	37536492	a	t	9.16×10^{-08}	0	Solyc12g044940[a]	Short-chain dehydrogenase/reductase
Phenylalanine	11	4002767	t	c	9.57×10^{-09}	0	Solyc11g010960	Alcohol dehydrogenase
Proline	3	66798980	t	g	2.39×10^{-09}	0	Solyc03g117770[a]	Serine incorporator 1
Serine	3	69913055	a	g	3.06×10^{-14}	0	Solyc03g121910	Threonine synthase
Geranyl acetone	2	40883244	a	g	6.00×10^{-15}	0	Solyc02g081330	Phytoene synthase 2
Hexenal	1	1083181	c	g	1.45×10^{-10}	0	Solyc01g006540	Lipoxygenase
Methyl salicylate	9	69293875	a	g	2.34×10^{-19}	0	Solyc09g089580	1-aminocyclopropane-1-carboxylate oxidase-like protein
1-penten-3-one	5	3036212	a	g	7.07×10^{-09}	0	Solyc05g008800[b]	Lipid phosphate phosphatase 3
2-methyl-1-butanol	6	37782796	a	g	5.50×10^{-09}	0	Solyc06g059850	3-methyl-2-oxobutanoate dehydrogenase
6-methyl-5-hepten-2-one	3	3212583	t	c	6.76×10^{-26}	0	Solyc03g025720	Long-chain-fatty-acid--CoA ligase
6-methyl-5-hepten-2-one	4	60345897	a	t	3.00×10^{-11}	0	Solyc04g074360	UDP-glucuronosyltransferase
6-methyl-5-hepten-2-one	10	61007386	a	g	9.28×10^{-09}	0	Solyc10g079470	L-galactono--lactone dehydrogenase

[a]A total of 305 loci for main tomato flavor-related quality traits were identified by meta-analysis of 775 tomato accessions and 2,316,117 SNPs. For each association, associated traits, chromosome (Chr), reference allele (Ref), alternative allele (Alt), the marker-trait association *P* value (*P*), heterogeneity I square (I^2), locus name (International Tomato Annotation Group 2.4) and candidate genes are shown. All SNP positions were aligned on the tomato reference genome version 2.50. The *P-value* is reported from the random-effect model performed using the inverse variance-weighted fixed-effect model in METAL[25]. For those SNPs where heterogeneity occurs ($I^2 > 25$, indicating moderate heterogeneity), we used the Han and Eskin random-effects model (RE2) implemented in METASOFT[26]. We also treated those candidate genes as new if previous GWAS did not report them though the association might be significant
[b]Significant cis expression quantitative trait loci (cis-eQTLs) from a previous transcriptome-wide association study (TWAS)[12] mainly based on panel T

with genotyped SNPs for both panels (Supplementary Figs. 5–12 and Supplementary Data 4–5). We used the Efficient Mixed-Model Association eXpedited (EMMAX) software for association tests for panel S and B[23], as reported for panel T[6] (Online methods, Supplementary Fig. 13). After imputation, we observed a similar or slight statistical increase in terms of the significance and the number of associated loci compared with MLMM[24] (Supplementary Figs. 14–44) and no genomic inflation ($\lambda < 1$) was detected for most (83.3%) of the traits (Supplementary Data 6). For panel T, which was characterized by 2,040,403 SNPs, the association tests had also been performed using EMMAX[6].

By combining the three separate studies, a total of 775 unique tomato accessions were used for the final meta-analysis of 31 flavor-related traits (2 sugars, 2 organic acids, 10 amino acids, and 17 flavor-related volatiles). We performed the meta-analysis with two software: METAL[25] using a fixed effect model and METASOFT[26] for those SNPs where heterogeneity occurred ($I^2 > 25$) using a random effect model. Manhattan plots and quantile–quantile (Q-Q) plots for all traits are shown in Supplementary Figs. 45–75. Meta-analysis identified a total of 305 significant loci ($P < 4 \times 10^{-7}$ for sugars, acids, and volatiles; $P < 2.99 \times 10^{-6}$ for amino acids), among which 211 were new (Supplementary Data 7). A total of 87 strong effect meta-QTLs were identified with high probability ($P < 10^{-9}$). Most of these loci passed the suggestive thresholds in at least one panel (Supplementary Figs. 14–75). Among the identified loci, 35 had a moderate to strong heterogeneity ($I^2 > 25$). We generated a local SQLite dataset for tomato (Online methods) and provided the LocusZoom plots for all the genome-wide significant associated loci (Supplementary Figs. 76–123). Among the 305 loci, 24 loci exhibited cis-eQTLs in a previous transcriptome-wide association study[12] in fruit tissue (Supplementary Data 7). Among the 211 associated loci, we identified 37 promising candidate genes (7 with significant cis-eQTLs[12]) with functional annotations related to the pathways of flavor chemicals (Table 1).

We performed a singular enrichment analysis for all associations using agriGO[27] (http://bioinfo.cau.edu.cn/agriGO/index.php). Up to 10 biological processes were significantly enriched ($P < 0.005$) (Supplementary Data 8). All these enriched processes or groups were closely involved in flavor-related metabolites (in terms of sugars, organic acids, amino acids, and volatiles), such as UDP-glycosyltransferase activity, transferase activity, oxidoreductase activity, and carbohydrate metabolic processes.

Previously reported flavor-related loci in the three panels were all strongly associated in the meta-analysis at a higher significance level, such as *Lin5* (Solyc09g010080, fructose, $P = 6.16 \times 10^{-10}$; glucose, $P = 4.30 \times 10^{-10}$), *TFM6* (Solyc06g072920, malate, $P = 2.26 \times 10^{-37}$), and *Phytoene synthase 1* (Solyc03g031860, geranyl acetone, $P = 6.73 \times 10^{-26}$)[6,28]. In meta-analysis of GWAS, heterogeneity represents the genetic variations observed across combined studies[13]. In this study, strong heterogeneity occurred even for those loci with major effects, such as *Lin5* (fructose, $I^2 = 95.6$, $P = 1.05 \times 10^{-10}$; glucose, $I^2 = 95.3$, $P = 5.85 \times 10^{-10}$). This could be due to population structure, linkage disequilibrium, phenotyping platforms, G × E interactions, etc.[13]. We then focused on loci in regions showing low LD, where one or a few

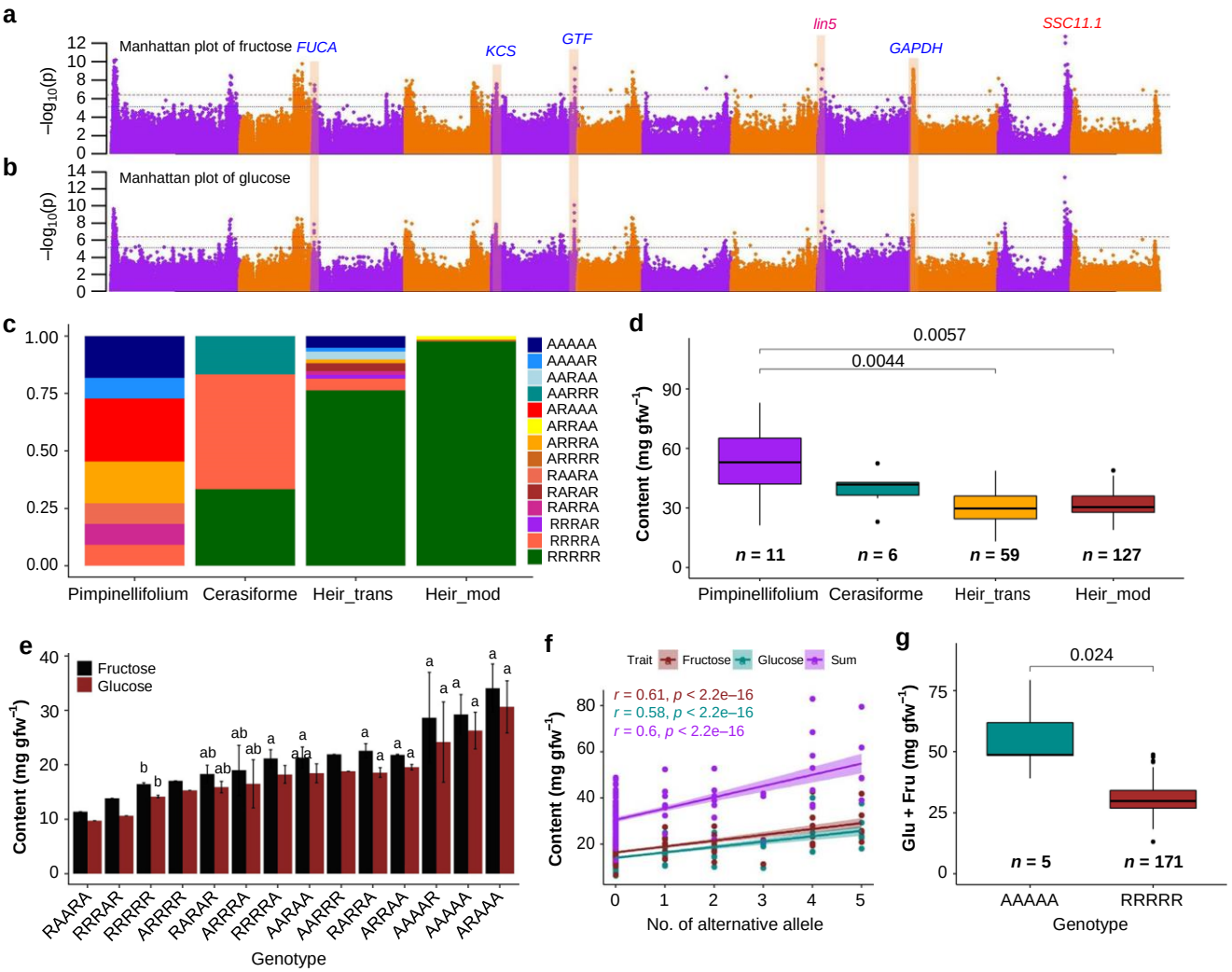


Fig. 2 Combinations of fructose and glucose alleles for the improvement of tomato sugar content. Only alleles that were significantly associated both with fructose and glucose were analyzed. a, b Manhattan plot for meta-analysis of genome-wide association analysis of fructose (a) and glucose (b) content. Candidates and previously identified genes were labeled in blue and red, respectively. FUCA, *alpha-L-fucosidase 1*; KCS, *fatty acid elongase 3-ketoacyl-CoA synthase*; GTF, *glucosyltransferase*; GADPH, *glyceraldehyde-3-phosphate dehydrogenase*. c Allele distribution of fructose/glucose content at positions: chr3:1,506,106, chr5:3,403,706, chr5:63,485,334, chr9:3,477,979, and chr10:422,707 that were both significantly associated with fructose and glucose in *S. lycopersicum* var *cerasiforme* (cerasiforme), heirloom + transitional (heir_trans), heir + modern (heir_mod), and the closest wild species *S. pimpinellifolium* (pimpinellifolium) tomato accessions (see detailed information about groups in online methods). d Comparison of sugar content (fructose + glucose) between different tomato types in cerasiforme, heir_trans, heir_mod, and pimpinellifolium tomato accessions. e Mean (±SE) content of fructose (black) and glucose (brown) at different allele combinations in cerasiforme, heir_trans, heir_mod, and pimpinellifolium tomato accessions. Significant *t*-test *P* values are also provided. f Correlation between the number of alternative alleles and sugar content. Fructose, glucose, and the sum of fructose + glucose were colored in brown4, cyan4, and purple. g Comparison of sugar content (fructose + glucose) between all alternative and reference allele combinations at position chr3: 1,506,106, chr5: 3,403,706, chr5: 63,485,334, chr9: 3,477,979, and chr10: 422,707. Center line and limits of box were the mean and interquartile ranges. Error bars represent the maximum and minimum values. Whiskers indicate variability outside the upper and lower quartiles. Significant *t*-test *P* values are also provided. Source data of Fig. 2c–g are provided in a Source Data file

candidate genes could be identified and regions with medium LD but with candidate genes near the peak SNPs.

Meta-analysis for sugar content. We looked into six candidate genes that were significantly associated both with fructose and glucose. In addition to *Lin5* and *SSC11.1*, we found four loci from the meta-analysis that were significantly associated both with fructose (Fig. 2a) and glucose content (Fig. 2b). These associations are in strong linkage disequilibrium with four candidate genes: *alpha-L-fucosidase 1* (*FUCA*; chr3: 1,506,106; fructose, $P = 3.39 \times 10^{-8}$; glucose, $P = 1.46 \times 10^{-8}$), *fatty acid elongase 3-ketoacyl-CoA synthase* (*KCS*; chr5: 3,403,706, fructose, $P = 2.57 \times 10^{-8}$; chr5: 3,406,424, glucose, $P = 1.49 \times 10^{-8}$), *glucosyltransferase* (*GTF*; chr5: 63,485,334; fructose, $P = 4.68 \times 10^{-10}$; glucose, $P = 8.36 \times 10^{-10}$), and *glyceraldehyde-3-phosphate dehydrogenase* (*GAPDH*; chr10:422,707, fructose, $P = 6.27 \times 10^{-10}$; chr10:332,069, glucose, $P = 1.20 \times 10^{-9}$). Notably, near the region of *FUCA* (up to ten genes), there are two candidate genes (Solyc03g006870, *phosphoglucomutase* and Solyc03g006860, *fructokinase*), which are also promising candidate genes for association with fructose and glucose content. Notably, *GTF* ($P = 7.55 \times 10^{-34}$) and GAPDH

($P = 7.84 \times 10^{-17}$) also showed significant cis-eQTL in a related transcriptome-wide association study[12].

Interestingly, all these loci, except *Lin5* (which falls in the domestication sweep DW149[19]), were not associated with any domestication or improvement sweep[19]. We compared the frequencies of different combinations of alleles of these candidate genes in relation to sugar content in wild, transitional, heirloom and modern accessions (more detailed explanations about group definition in Online Methods). All modern, heirloom, and transitional accessions lost most of the diversity of allele combinations that is present in the wild species group (Fig. 2c). The sugar content of heirloom + transitional (heir_trans) and heirloom + modern (heir_mod) groups were both significantly lower than that of the wild species (Fig. 2d). Fruit sugar content increased gradually as the number of alternative alleles increased (Fig. 2e). We observed significant positive correlations between the number of alternative alleles within allele combinations and sugar content (Fig. 2f). In addition, total sugar content (glucose + fructose) of all alternative allele combinations was significantly higher ($P = 0.024$) than that of all reference allele combinations (Fig. 2g). Together, these results provide insights into possibilities for tomato sugar improvement.

Meta-analysis for organic acids. The meta-analysis also provided several candidate genes for tomato fruit acid content. A strong association ($P = 2.26 \times 10^{-37}$ was detected for malate at an aluminum-activated malate transporter-like gene on chromosome 6, which has been reported to have a major effect on malate content[6,8,11], and was further validated as *Al-Activated Malate Transporter 9* (*Sl-ALMT9*)[28]. We found a strong significant association for citrate (chr6: 44,955,568, $P = 7.46 \times 10^{-27}$), which was 1.54 kb away from *Sl-ALTM9* (Supplementary Fig. 45 and Table 1). We also identified a significant association with another aluminum-activated malate transporter on chromosome 1 (chr1:1,749,084, $P = 3.62 \times 10^{-13}$; Supplementary Fig. 45 and Table 1). The strong linkage with both citrate and malate indicated that *Al-Activated Malate Transporter* also plays an important role in regulating citrate content in tomato fruit.

Candidate genes directly involved in the biosynthesis of citrate and malate were also identified. For example, we identified an association with citrate on chromosome 7, 150 kb away from a gene coding a citrate synthase (Solyc07g055840, $P = 4.70 \times 10^{-12}$). This candidate gene was also significantly associated with fructose ($P = 4.28 \times 10^{-09}$). For malate content, we found one association on chromosome 12 (chr12: 1,824,226, $P = 1.75 \times 10^{-19}$) close (36 kb) to a gene coding a malic enzyme (Solyc12g008430, four genes away from the peak SNP). We then took six candidate genes to analyze the relationships between different allele combinations and citrate and malate content, respectively (Fig. 3). The six candidate genes for citrate were *AIMT* (*Aluminum-activated malate transporter*, chr1: 1,749,084, $P = 3.62 \times 10^{-13}$), *GTF* (*Glycosyl transferase group 1*, chr2: 47,904,426, $P = 4.30 \times 10^{-13}$), *GS* (*Glycogen synthase*, chr3: 52,998,165, $P = 1.84 \times 10^{-15}$), *AIMT* (*Aluminum-activated malate transporter*, chr6: 44,955,568, $P = 7.46 \times 10^{-27}$), *CS* (*Citrate synthase*, chr7: 63,601,724, $P = 4.70 \times 10^{-12}$), and *Rubisco* (*Ribulose-1 5-bisphosphate carboxylase/oxygenase activase 1*, chr10: 65,378,714, $P = 5.35 \times 10^{-09}$). The six candidate genes for malate were *GTF* (*UDP-glucosyltransferase*, chr2: 48,509,791, $P = 3.47 \times 10^{-28}$), *PDHB* (*Pyruvate dehydrogenase E1 component subunit beta*, chr4: 2,156,747, $P = 4.45 \times 10^{-17}$), *AIMT* (*Aluminum-activated malate transporter*, chr6: 44,999,916, $P = 2.26 \times 10^{-37}$), *SS* (*Sucrose synthase*, chr9: 72,364,359, $P = 1.34 \times 10^{-15}$), *ME* (*Malic enzyme*, chr12: 1,824,226, $P = 1.75 \times 10^{-19}$), and *GAPB* (*Glyceraldehyde-3-phosphate dehydrogenase B*, chr12: 64,816,056, $P = 5.99 \times 10^{-16}$). Among the selected candidates, *GTF* on chromosome 2 and *AIMT* on chromosome 6 were associated with both citrate and malate (Fig. 3a, b). Both *GTF* and *GS* are located within improvement sweeps (IS031 and IS044, respectively)[19] and domestication sweeps (DS050 and DS175)[19] were observed for malate on *PDHB* and *ME*. For citrate and malate, the modern tomato accessions presented very different allele combinations than those in wild species and cherry tomatoes (Fig. 3c, d). In comparison, the total number of allele combinations for malate was approximately three times that of citrate. The citrate content was significantly different between some allele combinations (Fig. 3e). With the increase in the total number of alternative alleles in different allele combinations, the citrate content first increased gradually, with a peak at $n = 2$, and then steadily decreased (Fig. 3f). The malate content also showed a wide range of variation among alleles (Fig. 3g and Supplementary Data 9). We observed a weak but significant ($P = 0.02$) positive linear correlation ($r = 0.16$) between the number of alternative alleles and malate content (Fig. 3h).

These results demonstrated that citrate content was more influenced by improvement sweeps while malate was more influenced by domestication sweeps in the long-term breeding history. In addition, citrate has much less allele diversity than malate and a distinct pattern of relationships between the number of alternative alleles and its content.

Meta-analysis for amino acids and volatiles. Many candidate genes associated with amino acid and volatile contents were identified. For example, we found a significant association for serine on chromosome 3 ($P = 3.06 \times 10^{-14}$) (Supplementary Fig. 57 and Table 1), which was only significant in panel B ($P = 2.13 \times 10^{-9}$) (Supplementary Fig. 26). The candidate gene is annotated as a threonine synthase, an enzyme involved in the serine biosynthesis pathway. For proline, we found one associated locus (Solyc03g117770, $P = 2.39 \times 10^{-9}$), which was also reported as a significant eQTL ($P = 1.04 \times 10^{-35}$)[12]. This gene is a serine incorporator, and directly regulates serine content. One locus corresponding to GDSL esterase/lipase (Solyc12g089350) was also significantly associated with four amino acids (asparagine, GABA, glutamine and threonine). For hexanal, we found the strongest association corresponding to the lipoxygenase gene *LoxC* (Solyc01g006540, $P = 1.45 \times 10^{-10}$), which encodes an enzyme that is essential for synthesis of C6 and C5 fatty acid-derived volatiles[29,30]. This candidate gene was also significantly associated with (Z)-3-hexen-1-ol ($P = 3.94 \times 10^{-07}$). For 2-methyl-1-butanol, the strongest association corresponded to a 3-methyl-2-oxobutanoate dehydrogenase gene (Solyc06g059850, $P = 5.50 \times 10^{-09}$), an enzyme associated with branched chain amino acid metabolism.

We then looked at the possibility that significantly increasing the overall intensity of volatiles contributed to consumer liking as well as significantly reducing the overall content of unpleasant volatiles by combining the strongest loci associated with the contents of six volatiles (Fig. 4). The four volatiles positively contributing to liking included geranyl acetone (chr3: 4,328,514, $P = 6.73 \times 10^{-26}$), hexanal (chr1: 1,083,181, $P = 1.45 \times 10^{-10}$), phenylacetaldehyde (chr4: 55,635,636, $P = 5.59 \times 10^{-22}$), and 6-methyl-5-hepten-2-one (chr3: 3,212,583, $P = 6.76 \times 10^{-26}$). The two unpleasant (or negative) volatiles were guaiacol (chr9: 69,299,940, $P = 5.90 \times 10^{-18}$) and methyl salicylate (chr9: 69,293,875, $P = 2.34 \times 10^{-19}$) (Fig. 4a–f). Modern and heirloom + transitional accessions had the lowest allele diversity,

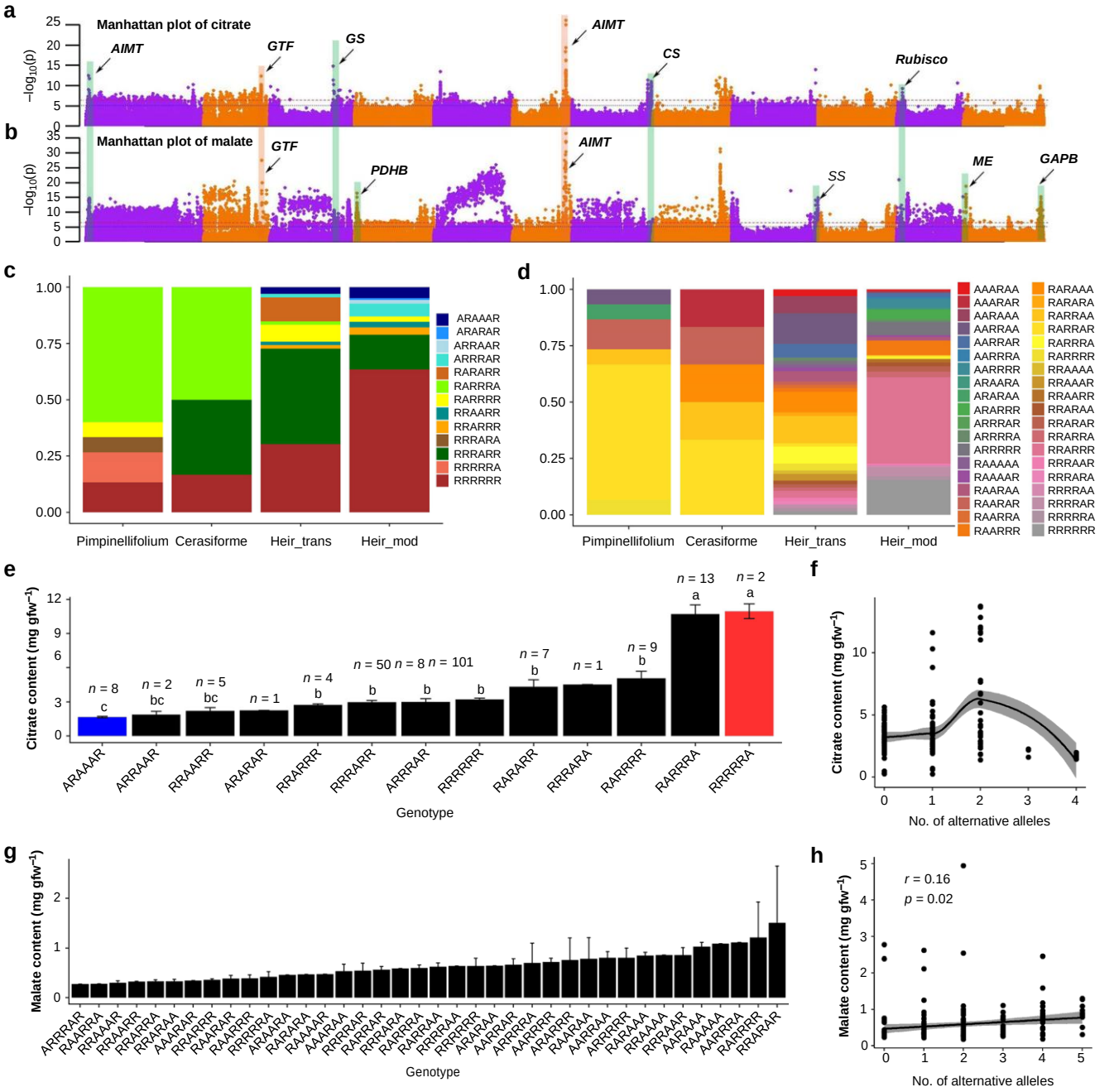
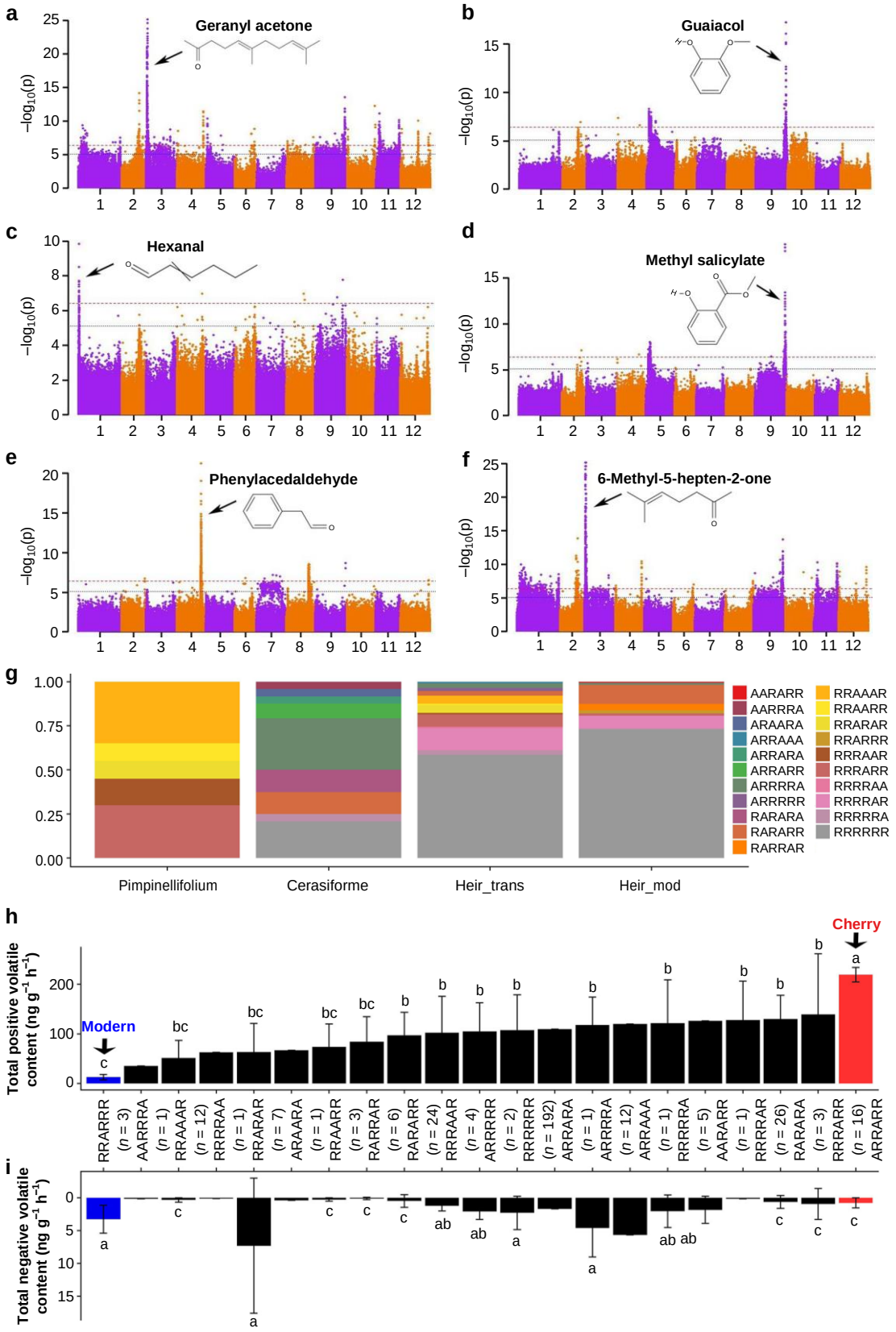


Fig. 3 Combinations of citrate and malate alleles for the improvement of tomato organic acid content. a, b Manhattan plot for meta-analysis of genome-wide association analysis of citrate (a) and malate (b) content. AIMT, *Aluminum-activated malate transporter*; GTF, *Glycosyl transferase group 1*; GS, *Glycogen synthase*; AIMT, *Aluminum-activated malate transporter*; CS, *Citrate synthase*; Rubisco, *Ribulose-1 5-bisphosphate carboxylase/oxygenase activase 1*; PDHB, *Pyruvate dehydrogenase E1 component subunit beta*; SS, *Sucrose synthase*; ME, *Malic enzyme*; GAPB, *Glyceraldehyde-3-phosphate dehydrogenase B*. c Allele distribution of citrate content at positions: chr1:1749084, chr2: 47,904,426, chr3: 52,998,165, chr6: 44,955,568, chr7: 63,601,724, and chr10: 65,378,714 in cerasiforme, heir_trans, heir_mod, and pimpinellifolium tomato accessions. d Allele distribution of malate content at positions: chr2: 48,509,791, chr4: 2,156,747, chr6: 44,999,916, chr9: 72,364,359, chr12: 1,824,226, and chr12: 64,816,056 in cerasiforme, heir_trans, heir_mod, and pimpinellifolium tomato accessions. e Mean (±SE, standard error) content of citrate content at different allele combinations in cerasiforme, heir_trans, heir_mod, and pimpinellifolium tomato accessions. f Correlation between the number of alternative alleles and citrate content. g Mean (±SE) content of malate content at different allele combinations in cerasiforme, heir_trans, heir_mod, and pimpinellifolium tomato accessions. h Correlations between the number of alternative alleles and malate content. Source data of Fig. 3c–h are provided in a Source Data file

especially compared with *S. pimpinellifolium* and cherry tomato accessions (*S. l. cerasiforme*). Interestingly, we also found that cherry tomatoes had the greatest diversity of allele combinations and some of them only appeared in this group (Fig. 4g).

The highest total content of the four positive volatiles was observed in allele combinations of cherry tomato accessions, which were significantly higher than the allele combinations of all modern tomato accessions (Fig. 4h). In contrast, modern



accessions have, on average, a significantly higher content of unpleasant volatiles, compared with the cherry accessions (Fig. 4i). These results revealed the combinations of alleles that have the potential to significantly enhance the total contents of volatiles associated with consumer liking.

Discussion

With the development of next-generation sequencing technology, GWAS has become a classical genetic approach to identify QTLs and causal genes in crops[31]. We herein demonstrate the potential of meta-analysis of GWAS following the detailed protocols first

Fig. 4 Combinations of six volatile alleles for the improvement of tomato volatile content. a–f Manhattan plot for meta-analysis of genome-wide association analysis of geranyl acetone (a), guaiacol (b), hexanal (c), methyl salicylate (d), phenylacetaldehyde (e), and 6-methyl-5-hepten-2-one (f) content. g Allele distribution of six volatiles content at positions: chr3: 4,328,514 (geranyl acetone), chr9: 69,299,940 (guaiacol), chr1: 1,083,181 (hexanal), chr9: 69,293,875 (methyl salicylate), chr4: 55,635,636 (phenylacetaldehyde), and chr3: 3,212,583 (6-methyl-5-hepten-2-one) in cerasiforme, heir_trans, heir_mod, and pimpinellifolium tomato accessions. h, i Mean (±SE, standard error) content of total content of the four positive volatiles (geranyl acetone, hexanal, phenylacetaldehyde and 6-methyl-5-hepten-2-one) (h) and two unpleasant volatiles (lower panel, guaiacol and methyl salicylate) (i) at different allele combinations in cerasiforme, heir_trans, heir_mod and pimpinellifolium tomato accessions. Source data of Fig. 4g–i are provided in a Source Data file

proposed in human genetics[32,33], which can be easily applied in other crops. Meta-analysis of GWAS is used when pooling raw data of separate panels (mega-analysis) is not possible. It has been shown both theoretically and numerically that meta-analysis is statistically as efficient as mega-analysis[34,35]. Even when possible, it is thus not necessary to re-analyze the raw data to perform meta-analysis. Only summary data (beta, standard error and p-values of associations at each SNP) from each panel is needed and should be provided with each GWAS result. For mega-analysis, genotypes and phenotypes from all panels should be first combined and then analyzed, which requires proper management of phenotypic structure (data coming from different studies with different plant growth conditions, different harvesting and sampling procedures, different metabolic analysis protocols etc.) and genotypic structure (such as population structure and kinship). Compared to mega-analysis, meta-analysis can assess the heterogeneity (consistency) of studies, which can be caused by many factors, such as phenotypic structure, genetic structure, linkage disequilibrium, imputation accuracies or G × E interactions[13,34]. Flavor remains a major breeding challenge in tomato[1,6]. Here, we used imputation-driven meta-analysis of genome-wide association studies to greatly increase the number of SNPs linked to chemicals associated with flavor. Among the 305 significantly associated loci, 41% of the SNPs had a low frequency (MAF < 0.1). Very low-frequency (0.01 < MAF < 0.05) SNPs were also detected (3 significant associated loci) (Supplementary Fig. 124). These results demonstrated that a sufficiently large sample size is needed to uncover these low-frequency and less common variants and to account for missing heritability[36–38]. Although hundreds of tomato genome sequences have been published[6,12,16–19], a high sequence depth reference panel is needed, such as the 1000 Genomes Project[39] in humans or the 1135 Arabidopsis genomes[40] in Arabidopsis, to perform genotype imputation[20,21], heritability estimation[36,41–43] and meta-analysis[13,14] with higher accuracy. Also, an imputation server could greatly enhance the integration of genetic resources[44].

In this study, we identified 37 promising candidate genes with functional annotations consistent with their involvement in biosynthesis of flavor chemicals. With the advancement of genome editing technologies, their functional analysis could greatly promote our knowledge of the genetic architecture of tomato flavor, provide fully linked markers for breeding and ensure consumer satisfaction[45–48]. It is also possible now to introduce desirable traits into wild stress-tolerant tomato accessions by genome editing[49,50]. However, tomato flavor can only be significantly improved when multiple genes are modified.

Many consumers are more attracted by small and medium size tomatoes with superior taste[51], as higher sugar content is usually associated with smaller fruit size[6]. In the meta-analysis, we found that modern cultivars have lost the majority of high-sugar alleles that were present in transitional, cherry tomato varieties and wild species. All these loci did not seem to have been influenced by any domestication or improvement sweeps, with the exception of *Lin5*, but some were loosely linked to fruit weight QTLs due to large LD in tomato. These results reflect the fact that sugar content has not been a breeding priority, in contrast to fruit size, yield, biotic, and abiotic resistances[1,6]. Strong positive correlations between the number of alternative alleles and sugar content provide clues on how to select higher sugar content tomato cultivars. However, sugar content can only be significantly improved when almost all the alternative alleles are selected, and will probably be accompanied by reduced fruit size[6] except if precise recombination or genetic modifications limits the linkage drag effect.

Malate and citrate are the main organic acids in most ripe fruits[52]. In tomato, citrate has a stronger impact on consumer preferences. In this study, candidate genes potentially impacting both citrate and malate contents were identified. We also demonstrated that citrate has been more influenced by improvement sweeps and malate by domestication sweeps. These results show that citrate was probably selected for improving tomato flavor.

Flavor-related volatiles are strongly influenced by the environment[53,54]. Nevertheless this meta-analysis illustrates that it should be possible to significantly enhance the content of favorable aromas via replacement of undesirable alleles. However, unlike sugars, the undesirable alleles should be carefully chosen[6]. Cherry tomato varieties have been introduced to the market since the 1990s. Their genomes are an admixture of those of big-fruited tomatoes and *S. pimpinellifolium* species[19,55] and may still contain a large number of favorable alleles. Thus they may serve as the most promising allele reservoir for breeding of high-flavor tomatoes.

In conclusion, we performed the first meta-analysis of genome-wide association analyses in a major vegetable and identified numerous loci involved in tomato flavor that were not identified in the three independent studies. A strong positive correlation between allele combinations and sugar content provides clues for breeding for higher sugar content. Modern cultivars have lost most of the allelic diversity for sugars, acids, and volatiles that is present within the species. Significant improvements should be achieved by replacing undesirable alleles. Taken together, our meta-analysis provides genetic insights into the genetic control of tomato flavor and gives a roadmap for flavor improvement.

Methods

Three GWAS panels. The meta-GWASs approach is based on three different GWAS panels already published and genotyped using different technologies. Our approach consisted in imputing SNP data for panels S[8] and B[11] from a reference panel, then conducting separate GWAS using the same mixed linear model (MLM) as described in[6] and collecting the summary statistics to run a meta-GWAS.

Panel S consists of 163 accessions[8], including 28*S. lycopersicum* (large tomato), 119*S. lycopersicum* var *cerasiforme* (cherry tomato), and 16*S. pimpinellifolium* (closest wild species). This panel was genotyped using the Solanaceae Coordinated Agricultural Project (SOLCAP) genotyping array[56,57], generating 5995 high quality SNPs. The minimal success genotyping rate per accession was fixed at 90%. The minor allele frequency of SNPs ranged from 0.037 to 0.45. Tomato accessions in Panel S were grown in Avignon, France, following a randomized complete block design, in a greenhouse during the summers of 2007 and 2008[8,58].

Panel B consists of 300 accessions with 62*S. pimpinellifolium*, 48*S. lycopersicum*, and 190*S.* l. *cerasiforme* accessions[11]. This panel was genotyped both with the SOLCAP[56,57] and CBSG arrays[59]. After quality control, 9013 SNPs (minor allele frequency, MAF > 0.1) and 291 accessions were kept. Accessions in Panel B were grown in Agadir, Morocco, France, under passive greenhouse irrigated conditions in 2011 and 2012[11]. Each trial followed a randomized complete block design, with three and two blocks, in 2011 and 2012, respectively.

Panel T consists of 402 tomato accessions from two separate panels[6]. Panel T was genotyped by whole genome resequencing technology, generating a number of 2,014,488 SNPs passing quality control (MAF > 0.05, missing rate < 10%). This panel includes five tomato types, including modern (51), transitional (50), cherry (27), heirloom (243), and wild species (27)[6].

Phenotypes. A total of 31 flavor-related quality traits in tomato were analyzed for meta-analysis, including two sugars (fructose and glucose), two organic acids (citrate and malate), 10 amino acids, and 17 flavor-related volatiles. The 10 amino acids were asparagine, aspartate, GABA, glutamine, lysine, methionine, phenylalanine, proline, serine, and threonine. The 17 volatiles were (E)-2-heptenal (E2HEP), (E)-2-hexenal (E2HEX), (E)-2-pentenal (E2PEN), (E,E)-2,4-decadienal (EE24D), (Z)-3-hexen-1-ol (Z3H1X), (Z)-3-hexenal (Z3HEX), 1-octen-3-one (X1O3ON), 1-penten-3-one (X1P3ON), 2-methyl-1-butanol (X2M1BU), 3-methyl-1-butanol (X3M1BU), 6-methyl-5-hepten-2-one (X6MHON), beta-ionone (BIONO), geranylacetone (GRACE), guaiacol (GUAIA), hexanal (XEXAN), phenylacetaldehyde (PHEAC), and methylsalicylate (METHY).

Sugars and organic acids were measured in all three panels. Amino acids were measured both in panel S and B, while flavor-related volatiles were measured both in panel B and T. Briefly, fructose and glucose in panel S were measured using the micro-method. Citrate and malate were measured by gas chromatography-mass spectrometry (GC-MS)[8]. Data distribution was tested using the Shapiro–Wilk test and data with a non-normal distribution were Log_{10} transformed. In panel B, these metabolites were measured within the Product Metabolism and Analytical Sciences Endogenous Metabolite Profiling Platform at Syngenta Jealott's Hill International Research Center, Bracknell, UK. Fructose and glucose were analyzed by high pH ion-exchange chromatography. Citrate and malate were analyzed using electrospray ionization-liquid chromatography (ESI-LC-MS/MS). Fructose and malate were transformed using the Boxcox method. Citrate was transformed using the Log_{10} method. In panel T, citrate and malate were measured using the citrate and malate analysis kits (R-Biopharm, Marshall, MI), according to the manufacturer's instructions[60]. Measurements of amino acids and volatiles in panel S was measured using GC-MS by comparing with a database of authentic standards. Small organic acids and amino acids in panel B were analyzed using electrospray ionization-liquid chromatography (ESI-LC-MS/MS). Volatiles in panel T were first captured by headspace solid phase micro extraction (HS-SPME) coupled GC-MS.

Reference panel for SNP imputation. A reference panel was selected from the 360 re-sequenced tomato accessions[19] to perform SNP imputation in panels S and B. Among this panel, only accessions with genome coverage ≥90% and mean sequencing depth ≥4.0 were kept. Wild tomato species were also removed, generating a total reference set of 221 accessions genotyped with 3,809,156 SNPs (Supplementary Table 1).

Recombination map. A high-density recombination map is required for imputation and computing genomic partitions. However, the available tomato genetic maps EXPIM 2012 and EXPEN 2012[57] have a limited genomic coverage (~3500 mapped SNPs). In order to use a much denser genetic map, we developed a Python script to infer the corresponding genetic positions of the 3,809,156 SNPs in the reference panel. Before calculating the recombination rate, we first compared the physical vs genetic distribution patterns for each chromosome (Supplementary Fig. 1). Comparing with EXPIM 2012, this newly built genetic map had the same distribution pattern (Supplementary Fig. 1). This comparison indicated the inferred genetic positions were accurate and were then used for estimating the recombination rate, as required for imputation. Minor adjustments were also done for some SNPs in order to follow an overall increasing positional order. Extreme recombination rate values were also removed (>2000 cM/Mb).

Genotype imputation. One unphased reference panel from IMPUTE2 (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#home)[22] was adopted for imputation of panel S and B independently. The 221 filtered sequenced accessions passing quality control were used as the reference panel. The newly built recombination map was used instead of EXPIM 2012. The whole genome was then divided into genomic intervals of 5 Mb for imputation and the effective size of population (*Ne*) was set at 2000.

Quality control. After imputation, the minimum MAF for panel S and B was set at 0.037 and 0.021, respectively, according to the formula: [Number of chromosomes/(2 × Number of individuals)][61]. After combining all the imputed data, basic statistic summaries were obtained in QCTOOL v2 (http://www.well.ox.ac.uk/~gav/qctool_v2/) with the following command:./qctool -g GWAS.gen -snp-stats. We then filtered all imputed SNPs with Hardy-Weinberg equilibrium (HWE) ≥ 0.000001, MAF ≥ 0.037 (0.021 for panel B), missing rate ≤ 0.10 and missing call rate ≤ 0.10. After these primary control steps, a total of 224,097 and 327,436 SNPs were retained for panel S and B, respectively.

In order to determine the optimal threshold of imputation quality (Info criteria), we compared the imputed and sequenced genotype data of the nine overlapping accessions in panel S that have been genotyped by SNP arrays and whole-genome sequencing. If the maximum of the three probabilities at a locus was higher than 0.9, we treated it as a certainty. This was done by converting the imputed data to ped/map format via GTOOL (http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html). We then compared the imputed and genotyped values of the nine accessions (Supplementary Fig. 2). Total numbers of corrected SNPs at different MAF and Info thresholds were obtained to validate the optimal threshold of MAF and Info. The average value of Info was 0.882 (with no filtering of MAF). With the increase of Info, the number of correctly genotyped SNPs increased from less than 200 to about 50,000 for panel S (Supplementary Fig. 2a, Supplementary Table 2). On average, 51.45% of the SNPs have been correctly imputed for all Info values. There was no significant difference between the numbers of corrected imputed SNPs for different Info values of the three tomato groups (Supplementary Fig. 2b). The majority of imputed SNPs had a MAF value ranging from 0.037 to 0.25, with a mean value of 0.172 ± 0.103 (with no filtering of Info). The percentage of successfully genotyped SNPs averaged at 57.3% and a higher percentage of corrected imputed SNPs decreased gradually with the increase of MAF (Supplementary Fig. 2c). Similarly, no significant difference was found between the numbers of corrected imputed SNPs for different MAF values of three tomato genetic groups (Supplementary Fig. 2d). Details of the number and percentage of corrected imputed SNPs at different MAF bins among the nine accessions are listed in Supplementary Data 1. We than compared the relationship between MAF and Info. The average value of Info was 0.912 for all values of MAF (Supplementary Fig. 2e). We found that the lowest mean value of Info (0.622) was observed on less common SNPs (0.037 < MAF < 0.05) (Supplementary Fig. 2e, Supplementary Data 2). However, this value is still higher than the proper imputation quality threshold (0.4) in common quality control of meta-analysis of genome-wide association studies[33]. So, we decided to set the Info threshold at 0.60 as the threshold of high imputation quality.

After filtering with imputation quality threshold (Info) ≥ 0.60, total of 209,152 and 252,414 SNPs were retained for panel S and B, respectively. The mean Info value at different MAF values for panel S and B were 0.929 and 0.922, respectively (Supplementary Data 3). The lowest mean value of Info at different MAF value was 0.810 and 0.783, respectively (Supplementary Fig. 2f, Supplementary Fig. 3). These SNPs offered a much denser genomic coverage for both panel S and B (35-fold and 28-fold, respectively) (Supplementary Fig. 4). Only some large genomic gaps still remained where there were few genotyped SNPs over a long genomic region (Supplementary Fig. 4). These results indicated that all the retained SNPs had a high imputation quality and were used for further analyses.

Linkage disequilibrium analysis. For population structure and kinship analyses, only independent SNPs ($r^2 < 0.2$) were used. This was done in PLINK (https://www.cog-genomics.org/plink2) with: --indep-pairwise 50 5 0.2 (windows, step, r2) –maf 0.05, generating a total of 3,602 and 4,294 independent SNPs for panel S and B, respectively.

Principal component analysis. In order to compare the genetic structure revealed before and after imputation, we performed a principal component analysis (PCA) for panels S and B, using all genotyped SNPs and independent imputed SNPs ($r^2 < 0.2$) in PLINK: --pca. Principal component analysis showed that genotype imputation did not lead to significant differences in genetic group composition and pairwise individual distances, for all three accession classes of panel S (S.C., S.L., S.P.) (Supplementary Fig. 5a–c). For the first principal component (PC1), there were strong positive correlations (0.93, 0.82, and 0.93 for S.C, S.L., and S.P. respectively) between genotyped and imputed SNPs (only imputed SNPs) (Supplementary Fig. 5d). By combining genotyped and imputed SNPs together (hereafter called 'All' dataset), a similar strong positive correlation (0.94, 0.82, and 0.94 for S.C, S.L., and S.P. respectively) was also found (Supplementary Fig. 5e). Correlation between imputed and all SNPs was also strong for all tomato classes (Supplementary Fig. 5f). For the panel B, a previous study revealed a population structure composed of six groups[62]. After imputation, we found they had a similar distribution pattern (Supplementary Fig. 6). PC1 between genotyped SNPs and all (genotyped and imputed) SNPs had a strong positive correlation (higher than 0.7 for all six groups) (Supplementary Fig. 6c). In contrast, the second principal component (PC2) had strong negative correlations for all six groups (lower than −0.6 for all six groups) (Supplementary Fig. 6d).

Population structure. In a previous study, the population structure of panel S was evaluated by Structure v2.3.4[63] (https://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/html/structure.html). So we first compared the structure following the same parameters, with 1×10^6 burn-in period and 5×10^6 MCMC steps. Based on the Evanno method[63], the optimal number of ancestral populations was two. Only minor population assignment differences were found for both subpopulations, compared with structure from genotyped SNPs (Supplementary Fig. 7).

We further used discriminant analysis of principal components (DAPC)[64] (http://adegenet.r-forge.r-project.org/files/tutorial-dapc.pdf) using the independent 3,602 and 4,294 SNPs ($r^2 < 0.2$) to infer the optimal population structure for panels S and B. This method partitioned the variance within and among groups without assumptions on LD or Hardy–Weinberg equilibrium[65], which has shown a better performance in clustering individuals[11]. The optimal number of clusters was

determined by Bayesian Information Criteria (BIC) with a minor increase or decrease. All PCs and all discriminant functions were retained to find the optimal number of clusters. In the following DAPC analyses, all discriminant functions and the first 50 PCs were retained in order to achieve 80% of cumulative variance for both panel S and B.

For panel S, the optimal number of clusters was six (Supplementary Fig. 8) and DAPC revealed a clear structure of all the accessions (Supplementary Fig. 9). For panel B, the optimal number of cluster was six, which was the same as that revealed by using genotyped SNPs (Supplementary Fig. 10). Membership of each cluster was also quite similar (Supplementary Fig. 11), compared with that of genotyped SNPs (Supplementary Fig. 12). Detailed information of the membership of each cluster revealed by all independent SNPs for panels S and B is listed in Supplementary Data 4 and Supplementary Data 5, respectively. These results indicated that imputation did not cause significant differences in the genetic structure for both panels S and B. For panel T, the optimal number of clusters was five from DAPC with the first 20 PCs retained and a cross validation run of 100 times[6].

Genome-wide association analysis. Though SNPTEST v2.5.4 (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html#introduction) can use the imputed data from IMPUTE2 to detect associations directly, it cannot however handle too many cofactors in the model. For accessions from each panel used in this study, there is strong genetic structure. We first took one trait (malate) in panel S as an example to choose the optimal association software to perform the association tests.

In order to add kinship as a cofactor in SNPTEST, we performed a principal component analysis of the kinship calculated in SPAGeDi (http://ebe.ulb.ac.be/ebe/SPAGeDi.html) and structure in Structure v2.3.4. We then added the first 20 PCs as cofactors in the frequentist association test model in SNPTEST. In the next step, we used EMMAX (http://genetics.cs.ucla.edu/emmax/index.html) with the BN kinship matrix and DAPC results to conduct association analyses. For BN kinship calculation, the default command was used: emmax-kin -v -h -d 10. A uniform threshold ($P = 1/n$, n is the effective number of independent SNPs) was used as the genome-wide significance threshold for all three panels. The effective number of independent SNPs was calculated in Genetic type 1 Error Calculator (GEC)[66] (http://grass.cgs.hku.hk/gec/download.php). The suggestive p-value for the 224,097 SNPs of panel S was 9.63×10^{-5} and the significant p-value was 4.82×10^{-6}. For the 327,436 SNPs of panel B, the suggestive and significant p-value was 5.99×10^{-5} and 2.99×10^{-6}, respectively.

After comparing the association results for malate of panel S, we found the strongest p-value in SNPTEST was still quite low, compared with other approaches (Supplementary Fig. 13). Results from MLMM (https://github.com/Gregor-Mendel-Institute/MultLocMixMod) and EMMAX were quite similar. So, in the following analyses, we only used SNPTEST to compute summary statistics, not for finding associations. For MLMM, this model adds the marker as co-factor using a window of 10. If too many markers are in full LD, the genetic variance calculation may be biased[24]. So, we used EMMAX for association analyses for all traits with the BN kinship matrix and DAPC results as covariance.

Meta-analysis. A total of 788 tomato accessions and 2,316,117 SNPs from three GWAS panels were used for the final meta-analysis. Since each panel was stratified and a small number of individuals overlapped between panels (38 between panel B and S, 18 between panel S and T, 17 between panel B and T), genomic inflation factor (λ) was corrected before meta-analysis using GenABEL[61] (http://www.genabel.org/packages/GenABEL) in R. Genomic inflation can be caused by population structure, cryptic relatedness, genotyping errors, sample size, LD, trait heritability, number of causal variants and other technical artefacts[67]. Though no adjustment is necessary when λ is lower or equal to one, we still corrected the standard errors of beta coefficients by applying the formula $SE \times \sqrt{\lambda}$ in general for each individual studies to get the chi-squares to its optimal values[68].

METAL[25] (fixed-effect model) (https://genome.sph.umich.edu/wiki/METAL_Documentation) and METASOFT[26] (random-effect model) (http://genetics.cs.ucla.edu/meta/) are two most commonly used meta-analysis software[13]. Meta-analysis was first performed using the inverse variance-weighted fixed-effect model in METAL[25]. The genome-wide significant p-value for meta-analysis was set as 4.0×10^{-7}, except for SNPs that only appeared between panel S and B (the significant p-value was set at 2.99×10^{-6}). For those SNPs where heterogeneity occurs ($I^2 > 25$, indicating moderate heterogeneity), we used the Han and Eskin random-effects model (RE2) in METASOFT[26]. This model assumes no heterogeneity under the null hypothesis and offers greater power under heterogeneity, compared with conventional random-effect models[26].

Local SQLite database for LocusZoom. In order to obtain a regional zoom plot of the candidate SNPs in LocusZoom[69] (https://genome.sph.umich.edu/wiki/LocusZoom_Standalone), a local SQLite database of tomato was required. We thus created a custom SQLite database in LocusZoom with the following steps. SNP positions in the 221 accessions of the reference panel were inserted by: dbmeister.py --db my_database.db --snp_pos my_snp_pos_file. For the gene information, we first downloaded the gene annotation file from Solgenomics (ftp://ftp.solgenomics.net/genomes/Solanum_lycopersicum/annotation/ITAG2.4_release/). We then converted it to genePred file format by gff3ToGenePred (http://hgdownload.cse.ucsc.edu/admin/exe/). Gene names were replaced with short codes instead of providing full names to avoid long names and overlapping. We then inserted the gene information by the following command line: dbmeister.py --db my_database.db --refflat my_refflat_file. For the recombination file, we used the recombination map previously inferred and inserted the data into our database by: dbmeister.py --db my_database.db --snp_set my_snpset_file. We used the 221 reference panel to calculate the linkage disequilibrium (LD) in PLINK by the following parameter: --ld-snp my.snp --ld-window-kb 100000 --ld-window 1000 --r2 --ld-window-r2 0 (windows, step, r2).

LD in candidate gene regions. In order to define the window size of the candidate genes, we first calculated the LD around the significant associated SNP with the window size of 5 Mb in PLINK with the following command line: --ld-window-kb 500000 --ld-window 1000 --r2 --ld-window-r2 0 (windows, step, r2). We then chose LD higher than 0.5 as the threshold of LD decay for the candidate gene region sizes. Within the regions, we chose the candidate genes based on both the distance of the peak SNP as well as the closest genes with known functions related to the trait. If no gene fell in the candidate regions, we provided the closest gene. We further crosschecked the candidate gene expression patterns using the Tomato Expression Atlas[70] (http://tea.solgenomics.net/expression_viewer/input).

Group re-definition of panel T. The relationship between allele combinations and flavor-related metabolites (sugars, organic acids and volatiles) was only based on panel T. For the accessions in panel T, they were previously defined as five clusters, namely *S. lycopersicum* var *cerasiforme*, heirloom, transitional, modern and the closest wild species *S. pimpinellifolium* tomato accessions[6]. However, there were up to 11 accessions with duplicated individual IDs (Supplementary Data 10) and we cross-checked these duplicated lines and only kept one. In addition, some accessions in the group of heirloom, modern and transitional were labeled inappropriately based on the DAPC analysis. In order to correct for this, we generated the principal component analysis (PCA) based on independent SNPs (LD = 0.1) (Supplementary Fig. 125). Based on PCA, some heirloom accessions are mixed with modern accessions and were labeled as heir_mod (heirloom and modern). For the remaining heirloom accessions, they were combined with transitional accessions and labeled as heir_trans (heirloom and transitional) (Supplementary Fig. 126). The accessions of panel T were thus re-defined as four clusters, namely *S. lycopersicum* var *cerasiforme*, (cerasiforme, 26 members), heirloom and modern (heir_mod, 196 members), heirloom and transitional (heir_trans, 138 members), and *S. pimpinellifolium* (27 members) (Supplementary Data 10–11). These redefined groups were then used for allelic combination analyses. Statistical tests were only performed for those allele combinations with at least two observations (either labeled with letters or with p-values).

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data supporting the findings of this work are available within the paper and its Supplementary Information files. All new meta-analysis data associated with the paper are available in a repository [https://doi.org/10.15454/TWFDYW]. The source data underlying Figs. 2c–g, 3c–h, and 4g–i and Supplementary Figs. 5a–f, 6a–d, and 124–126 are provided as a Source Data file. Additional datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Received: 5 November 2018 Accepted: 4 March 2019
Published online: 04 April 2019

References

1. Klee, H. J. & Tieman, D. M. The genetics of fruit flavour preferences. *Nat. Rev. Genet.* 19, 347–356 (2018).
2. Tieman, D. et al. The chemical interactions underlying tomato flavor preferences. *Curr. Biol.* 22, 1035–1039 (2012).
3. Causse, M. et al. Consumer preferences for fresh tomato at the European scale: a common segmentation on taste and firmness. *J. Food Sci.* 75, S531–S541 (2010).
4. Baldwin, E. A., Scott, J. W., Shewmaker, C. K. & Schuch, W. Flavor trivia and tomato aroma: biochemistry and possible mechanisms for control of important aroma components. *HortScience* 35, 1013–1022 (2000).
5. Goff, S. A. & Klee, H. J. Plant volatile compounds: sensory cues for health and nutritional value? *Science* 311, 815–819 (2006).
6. Tieman, D. et al. A chemical genetic roadmap to improved tomato flavor. *Science* 355, 391–394 (2017).
7. Rothan, C., Diouf, I. & Causse, M. Trait discovery and editing in tomato. *Plant J.* 97, 73–90 (2019).

8. Sauvage, C. et al. Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant Physiol.* 165, 1120–1132 (2014).
9. Zhang, J. et al. Genome-wide association mapping for tomato volatiles positively contributing to tomato flavor. *Front. Plant Sci.* 6, 1042 (2015).
10. Zhao, J. et al. Association mapping of main tomato fruit sugars and organic acids. *Front. Plant Sci.* 7, 1–11 (2016).
11. Bauchet, G. et al. Identification of major loci and genomic regions controlling acid and volatile content in tomato fruit: implications for flavor improvement. *New Phytol.* 215, 624–641 (2017).
12. Zhu, G. et al. Rewiring of the fruit metabolome in tomato breeding. *Cell* 172, 249–261.e12 (2018).
13. Evangelou, E. & Ioannidis, J. P. A. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* 14, 379–389 (2013).
14. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* 18, 117–127 (2017).
15. Bouwman, A. C. et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nat. Genet.* 50, 362–367 (2018).
16. Sato, S. et al. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485, 635–641 (2012).
17. Aflitos, S. et al. Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. *Plant J.* 80, 136–148 (2014).
18. Bolger, A. et al. The genome of the stress-tolerant wild tomato species *Solanum pennellii*. *Nat. Genet.* 46, 1034–1038 (2014).
19. Lin, T. et al. Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.* 46, 1220–1226 (2014).
20. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11, 499–511 (2010).
21. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* 48, 1284–1287 (2016).
22. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS. Genet.* 5, e1000529 (2009).
23. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354 (2010).
24. Segura, V. et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44, 825–830 (2012).
25. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190–2191 (2010).
26. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* 88, 586–598 (2011).
27. Tian, T. et al. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* 45, W122–W129 (2017).
28. Wang, B. et al. An InDel in the promoter of Al-ACTIVATED MALATE TRANSPORTER9 selected during tomato domestication determines fruit malate contents and aluminum tolerance. *Plant Cell* 29, 2249–2268 (2017).
29. Chen, G. et al. Identification of a specific isoform of tomato lipoxygenase (TomloxC) involved in the generation of fatty acid-derived flavor compounds. *Plant Physiol.* 136, 2641–2651 (2004).
30. Shen, J. et al. A 13-lipoxygenase, TomloxC, is essential for synthesis of C5 flavour volatiles in tomato. *J. Exp. Bot.* 65, 419–428 (2014).
31. Liu, H. J. & Yan, J. Crop genome-wide association study: a harvest of biological relevance. *Plant J.* 97, 8–18 (2019).
32. Turner, S. et al. *Current Protocols in Human Genetics* Chapter 1, Unit 1. 19 (NIH Public Access, Hoboken, New Jersey, USA, 2011).
33. Winkler, T. W. et al. Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* 9, 1192–1212 (2014).
34. Panagiotou, O. A., Willer, C. J., Hirschhorn, J. N. & Ioannidis, J. P. A. The power of meta-analysis in genome-wide association studies. *Annu. Rev. Genom. Hum. Genet.* 14, 441–465 (2013).
35. Lin, D. & Zeng, D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genet. Epidemiol.* 34, 60–66 (2010).
36. Yang, J., Zeng, J., Goddard, M. E., Wray, N. R. & Visscher, P. M. Concepts, estimation and interpretation of SNP-based heritability. *Nat. Genet.* 49, 1304–1310 (2017).
37. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* 13, 135–145 (2012).
38. Marouli, E. et al. Rare and low-frequency coding variants alter human adult height. *Nature* 542, 186–190 (2017).
39. Gibbs, R. A. et al. A global reference for human genetic variation. *Nature* 526, 68–74 (2015).
40. Alonso-Blanco, C. et al. 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166, 481–491 (2016).
41. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* 461, 747–753 (2009).
42. Eichler, E. E. et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* 11, 446–450 (2010).
43. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S. & Balding, D. J. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* 49, 986–992 (2017).
44. Wang, D. R. et al. An imputation platform to enhance integration of rice genetic resources. *Nat. Commun.* 9, 3519 (2018).
45. Gao, C. The future of CRISPR technologies in agriculture. *Nat. Rev. Mol. Cell Biol.* 19, 275–276 (2018).
46. Rodríguez-Leal, D., Lemmon, Z. H., Man, J., Bartlett, M. E. & Lippman, Z. B. Engineering quantitative trait variation for crop improvement by genome editing. *Cell* 171, 470–480.e8 (2017).
47. Huang, S., Weigel, D., Beachy, R. N. & Li, J. A proposed regulatory framework for genome-edited crops. *Nat. Genet.* 48, 109–111 (2016).
48. Yin, K., Gao, C. & Qiu, J.-L. Progress and prospects in plant genome editing. *Nat. Plants* 3, 17107 (2017).
49. Zsögön, A. et al. De novo domestication of wild tomato using genome editing. *Nat. Biotechnol.* 36, 1211–1216 (2018).
50. Gao, C. et al. Domestication of wild tomato is accelerated by genome editing. *Nat. Biotechnol.* 36, 1160–1163 (2018).
51. Oltman, A. E., Jervis, S. M. & Drake, M. A. Consumer attitudes and preferences for fresh market tomatoes. *J. Food Sci.* 79, S2091–S2097 (2014).
52. Etienne, A., Génard, M., Lobit, P., Mbeguié-A-Mbéguié, D. & Bugaud, C. What controls fleshy fruit acidity? A review of malate and citrate accumulation in fruit cells. *J. Exp. Bot.* 64, 1451–1469 (2013).
53. Cebolla-Cornejo, J. et al. Evaluation of genotype and environment effects on taste and aroma flavor components of Spanish fresh tomato varieties. *J. Agric. Food Chem.* 59, 2440–2450 (2011).
54. Karppinen, K., Zoratti, L., Nguyenquynh, N., Häggman, H. & Jaakola, L. On the developmental and environmental regulation of secondary metabolism in *Vaccinium* spp. berries. *Front. Plant Sci.* 7, 655 (2016).
55. Blanca, J. et al. Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. *BMC Genom.* 16, 257 (2015).
56. Hamilton, J. P. et al. Single nucleotide polymorphism discovery in cultivated tomato via sequencing by synthesis. *Plant Genome J.* 5, 17 (2012).
57. Sim, S. C. et al. Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. *PLoS One* 7, e40563 (2012).
58. Xu, J. et al. Phenotypic diversity and association mapping for fruit quality traits in cultivated tomato and related species. *Theor. Appl. Genet.* 126, 567–581 (2013).
59. Víquez-Zamora, M. et al. Tomato breeding in the genomics era: insights from a SNP array. *BMC Genom.* 14, 354 (2013).
60. Tieman, D. M. et al. Identification of loci affecting flavour volatile emissions in tomato fruits. *J. Exp. Bot.* 57, 887–896 (2006).
61. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 23, 1294–1296 (2007).
62. Bauchet, G. et al. Use of modern tomato breeding germplasm for deciphering the genetic control of agronomical traits by Genome Wide Association Study. *Theor. Appl. Genet.* 130, 875–889 (2017).
63. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620 (2005).
64. Jombart, T. et al. Package ' adegenet'. *Bioinforma. Appl. Note* 24, 1403–1405 (2008).
65. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* 11, 94 (2010).
66. Li, M. X., Yeung, J. M. Y., Cherny, S. S. & Sham, P. C. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* 131, 747–756 (2012).
67. Yang, J. et al. Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* 19, 807–812 (2011).
68. de Bakker, P. I. W. et al. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* 17, 122–128 (2008).
69. Pruim, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 27, 2336–2337 (2011).
70. Fernandez-Pozo, N. et al. The tomato expression atlas. *Bioinformatics* 33, 2397–2398 (2017).

Acknowledgements

J-T.Z. was funded by a Chinese Scholarship Council (CSC) scholarship. We thank Guangtao Zhu from Huang's group in helping by providing the original GWAS results of

panel T and discussions about the results. We thank Qi Wu from the University of Cambridge for detailed theoretical explanations about linkage disequilibrium and population genetics. We thank David Francis from Ohio State University for the positive discussions and cross-checking the misclassification of the accessions in panel T. We thank Rebecca Stevens for the English language editing.

Author contributions

Study design/conception: M.C., J-T.Z., C.S.; supervision: C.S, M.C.; data collection and analysis: J-T.Z., F.B., J-H.Z., D.L., G.B., S.H., D.M.T., H.J.K.; data interpretation: J-T.Z., F.B., C.S., M.C., D.M.T., H.J.K.; first draft of the manuscript: J-T.Z.; critical revisions of the manuscript: all co-authors.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-09462-w.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Journal Peer Review Information: *Nature Communications* thanks Yun Li, and other anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019

Appendix 4

Genomic Designing for Climate Smart Tomato

RUNNING TITLE: CLIMATE-SMART TOMATO

Authors: Mathilde Causse[1], Jiantao Zhao[1], Isidore Diouf[1], Jiaojiao Wang[2], Veronique Lefebvre[1], Bernard Caromel[1], Michel Génard[3], Nadia Bertin[3]

[1]INRA, Centre de Recherche PACA, Génétique et Amélioration des Fruits et Légumes, Domaine Saint Maurice, CS60094, Montfavet 84143, France
[2]INRA and University of Bordeaux, UMR 1332 Biologie du Fruit et Pathologie, 71, av. Edouard Bourlaux - CS 20032 - 33882 Villenave d'Ornon Cedex - France
[3]INRA, Plantes et Systèmes de Culture Horticoles, Institut National de la Recherche Agronomique - Centre de Recherche PACA, Avignon, France

*Corresponding author: Mathilde CAUSSE (Mathilde.causse@inra.fr)

Outline

Abstract
1 Introduction
2 Challenges, Priorities and Breeding objectives
- 2.1 Productivity
- 2.2 Fruit quality
  - 2.2.1 Nutritional quality
  - 2.2.2 Sensory quality
  - 2.2.3 Mild stress as a tool to manage fruit quality
- 2.3 Biotic and abiotic stresses
  - 2.3.1 Biotic stresses
    - 2.3.1.1 Pests and pathogens of tomatoes
    - 2.3.1.2 Impact of climate change on pest and pathogen resistance
    - 2.3.1.3 New emerging tomato diseases
  - 2.3.2 Abiotic stresses
    - 2.3.2.1 Water deficit
    - 2.3.2.2 Salinity stress
    - 2.3.2.3 Temperature stress
    - 2.3.2.4 Mineral nutrition deficiency
  - 2.3.3 Stress combination

3 Genetic and genomic resources for trait breeding
- 3.1 Genetic resources
  - 3.1.1 Origin of tomato and its wild relatives
  - 3.1.2 Genetic resources as sources for adaptation
  - 3.1.3 Natural and induced mutants
- 3.2 Molecular markers and trait dissection
  - 3.2.1 Evolution of molecular markers
  - 3.2.2 Trait mapping
  - 3.2.3 Specific populations to dissect phenotypes
  - 3.2.4 Genes and QTL controlling tomato disease resistance
    - 3.2.4.1 Resistance gene and QTL discovery
    - 3.2.4.2 Resistance gene and QTL architecture
    - 3.2.4.3 Molecular basis of resistance genes and QTLs
- 3.3 Genomic resources
  - 3.3.1 The reference genome sequence
  - 3.3.2 Resequencing tomato accessions
- 3.4 SNP markers
  - 3.4.1 SNP discovery
  - 3.4.2 SNP arrays
  - 3.4.3 Genotype imputation
- 3.5 Diversity analyses
- 3.6 Cloned genes/QTL
- 3.7 New resources for genes and QTL identification
- 3.8 Genome-wide association studies
  - 3.8.1 The conditions for applying Genome Wide Association Studies
  - 3.8.2 Meta-analysis
- 3.9 Genetic dissection of abiotic stress tolerance
- 3.9.1 Genetic control of G x E interaction
- 3.9.2 Grafting as a defense against stresses
- 3.10 Omic studies
  - 3.10.1 Metabolome analyses
  - 3.10.2 Transcriptome analyses for eQTL mapping
  - 3.10.3 Multi-omic approach
  - 3.10.4 miRNA and epigenetic modifications
- 3.11 Databases

4 Breeding for smart tomato
- 4.1 Traditional breeding
- 4.2 Marker-Assisted Selection
  - 4.2.1. Marker-Assisted Backcross for monogenic traits
  - 4.2.2. Marker-assisted selection for QTLs
  - 4.2.3 Advanced backcross for the simultaneous discovery and transfer of new alleles

Abstract
Tomato is the first vegetable consumed in the world. It is grown in very different conditions and areas, mainly in field for processing tomatoes while fresh market tomatoes are often produced in greenhouses. Tomato faces many environmental stresses, both biotic and abiotic. Today many new genomic resources are available allowing an acceleration of the genetic progress. In this chapter, we will first present the main challenges to breed climate smart tomatoes. The breeding objectives relative to productivity, fruit quality and adaptation to environmental stresses will be presented with a special focus on how climate change is impacting these objectives. In a second part the genetic and genomic resources available will be presented. Then traditional and molecular marker breeding techniques will be discussed. A special focus will then be presented on ecophysiological modeling, which could constitute an important strategy to define new ideotypes adapted to breeding objectives. Finally we will illustrate how new biotechnological tools are implemented and could be used to breed climate smart tomatoes.

Key words: Tomato, breeding, productivity, biotic stress, abiotic stress, ideotypes, modeling

1 Introduction

Tomato is the first vegetable consumed worldwide after potato. It has become an important food in many countries. Two main types of tomato varieties are produced, tomatoes for processing industry, with determinate growth produced only in open field and indeterminate growth varieties for fresh market, which may be grown in very diverse conditions, from open field to greenhouses with controlled conditions.
Tomato, *Solanum lycopersicum* L., is a member of the large Solanaceae family, together with potato, eggplant and pepper. It is a self-pollinated crop, with a diploid (2n=2x=24) genome of medium size (950 Mb). A high quality reference genome sequence was published in 2012 (The Tomato Genome Consortium, 2012). Tomato originates from South America as well as 12 wild relative species, which can be crossed with the cultivated tomato species. Several large collections of genetic resources exist and more than 70,000 varieties are conserved in these gene banks. The collections also include scientific resources such as collections of mutants or segregating populations.
Tomato is also a model species for genetic analysis since a long time. Many mutations inducing important phenotype variations were discovered and positionally cloned and many disease resistance genes functionally characterized. Tomato is also a model species for fruit development and physiology. It is easy to transform and it has been the first transgenic food produced and sold (Kramer and Redenbaugh, 1994).
In this chapter, we will first present the main challenges to breed climate smart tomatoes. The breeding objectives relative to productivity, fruit quality and adaptation to environmental stresses will be presented with a special focus on how climate change is impacting these objectives. In a second part the genetic and genomic resources available will be presented. Then traditional and molecular marker breeding techniques will be discussed. A special focus will then be presented on ecophysiological modeling, which could constitute an important strategy to define new ideotypes adapted to breeding objectives. Finally we will illustrate how new biotechnological tools are implemented and could be used to breed climate smart tomatoes.

2 Challenges, priorities and breeding objectives

Tomato crop faces several challenges, which impacts its breeding objectives. Breeders will orientate their main breeding objectives according to the wide diversity of growth conditions and use as fresh or processed. These objectives can be classified in (1) productivity, (2) adaptation to growth conditions in terms of response to biotic and abiotic stresses and (3) fruit quality at both nutritional and sensory levels.

2.1 Productivity

From 1988 to 2017, the tomato world production regularly grew from 64 MT to 182 MT. Since 1995, China increased its production and became the first producer, and since then, its production increased up to 60 MT (**Figure 1**) covering almost 4,800,000 ha. This growth is due to an increase in production area, but also due to improvement in productivity and variety breeding.
With an average yield of 37 T/ha, compared to 16 t/ha in 1961, yield has increased over years but large differences remain according to countries and growth conditions. In south Europe greenhouses, the average yield is 50-80T/ha, while it may be more than 400T/ha in the Netherland and Belgium, with a crop lasting up to 11 months. Expressed per square meter, the average yield is 3.7 kg/m^2, reaching 50 kg/m^2in the Netherland, while it is 5.6 in China where most of the production is in open field although modern Chinese solar greenhouses are developed (Cao et al., 2019).

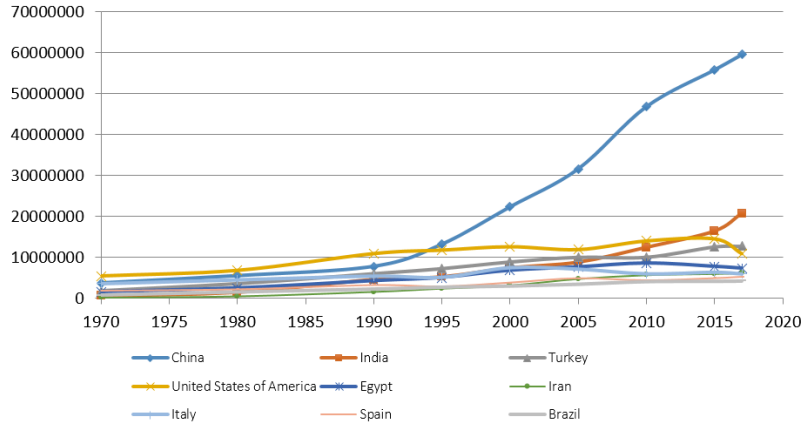


Figure 1 Evolution of tomato production over years in the 9 main producing countries

Yield is strongly dependent on cultivars and growth conditions. Yield results from fruit number and fruit weight. Cultivars for fresh market are classified based on their fruit size and shape from the cherry tomato (less than 20g) to beef tomato (fruit weight higher than 200g). The potential size depends on cell number established in pre-anthesis stage, but final fruit size mainly depends on the rate and duration of cell enlargement (Ho, 1996). Seed number and competition among fruits also affect the final fruit size (Bertin et al., 2002; 2003). Seed and fruit are highly sensitive to biotic and abiotic stresses, which often lead to seed and fruit abortion (Ruan *et al.* 2012). Fruit number is controlled by the truss architecture but the increase in flower number often leads to abortion (Soyk *et al.* 2017). Fruit shape varies from flat to long or ovate fruit and is also determined at the carpel development stage. Mutations in four genes explain most of the tomato fruit shape (Rodriguez et al., 2011).

2.2 Fruit quality

2.2.1 Nutritional quality

Tomato consumption has been shown to reduce the risks of certain cancers and cardiovascular diseases (Giovannucci, 1999). Its nutritional value is related to fruit composition in primary and secondary metabolites (**Table 1**) but is mostly due to its content in lycopene and carotene (Bramley, 2002). Lycopene is responsible of the red fruit color but also acts as a dietary antioxidant. Tomato also constitutes an important source of vitamin C. In spite of considerable efforts in developing cultivars with higher content in carotenoids, or in vitamin C, none has reached a commercial importance, in part because of a negative relation between yield and these traits (Klee, 2010).

Table 1. Average tomato fruit nutritional value and composition (adapted from USDA)

Proximates	Content (per 100g fresh weight)
Water	94.5 g
Energy	18 kcal
Protein	0.88 g
Lipids	0.2 g
Fibers	1.2 g
Sugars	2.63 g
Acids	0.65g
Minerals	
Calcium	10 mg
Magnesium	11 mg
Phosphorus	24 mg
Potassium	237 mg
Sodium	5 mg
Fluoride	□□□□□g
Vitamins	
Vitamin C	14 mg
Choline	6.7 mg
Vitamin A & carotene	0.59 mg
Lycopene	2.57 mg
Lutein & zeaxanthin	123 g
Vitamin K	8 g

(adapted from USDA: https://www.usda.gov/)

In addition to these well-known vitamins and antioxidants, other compounds in tomato fruit with antioxidant properties include chlorogenic acid, rutin, plastoquinones, tocopherol, and xanthophylls. Tomatoes also contribute but to a lesser extent in carbohydrates, fiber, flavor compounds, minerals, protein, fats and glycoalkaloids to the diet (Davies and Hobson 1981). Exhaustive metabolome studies have completed the composition of tomato in both primary and secondary metabolites and shown the wide diversity present among tomato accessions and their wild relatives (Tikunov et al., 2005; Schauer et al., 2006; wells et al., 2013; Rambla et al., 2013, Tieman et al., 2017; Zhu et al., 2018).

Considerable genetic variation exists in tomato for micronutrients with antioxidant activity or other health conferring properties (Hanson *et al.* 2004; Schauer *et al.* 2005). A number of these micronutrients, particularly carotenoids, have long been major objectives of breeding programs because of their contribution to the quality of fresh and processed tomato products. Increased recognition of their health promoting properties has stimulated new research to identify loci that influence their concentration in tomato.

Vitamin A and vitamin C are the principal vitamins in tomato fruit. Tomatoes also provide moderate levels of folate and potassium in the diet and lesser amounts of vitamin E and several water-soluble vitamins. β-carotene is a pro-vitamin A carotenoid. Carotene biosynthesis in tomato has been deciphered and many genes and mutations identified (Ronen et al., 1999). More than 20 genes that influence the type, amount, or distribution of fruit carotenoids have been characterized in tomato (Labate et al., 2007).

Vitamin C pathway in plants has been deciphered by Smirnoff and Wheeler (2000). The variation in ascorbic acid content may depend on varieties and growth conditions (Gest et al., 2013) and a few QTL controlling its variation have been identified (Stevens et al., 2007). The biosynthetic pathway of folate is also well characterized and the genes involved identified (Almeida *et al.* 2011). One of the major QTL controlling its variation has been shown to be due to epigenetic variation (Quadrana et al., 2014).

Glycoalkaloids and their toxic effects are commonly associated with Solanaceous species. Tomato accumulates the glycoalkaloids α-tomatine and dehydrotomatine which are less toxic than glycoalcaloids in potato ((Madhavi and Salunkhe, 1998; Milner et al., 2011). Several genes controlling their variations have been identified (Cardenas et al., 2016; Zhu et al., 2018).

Tomato mineral composition is greatly influenced by plant nutrition (see below), and as a result, has been well characterized in the context of mineral deficiency and the effect of these conditions on plant health. There is significant genotypic variation for mineral content in tomato fruit. Potassium, together with nitrate and phosphorous, constitutes approximately 93% of the total inorganic fruit constituents (Davies and Hobson 1981).

Flavonoids comprise a large group of secondary plant metabolites and include anthocyanins, flavonols, flavones, catechins, and flavonones (Harborne 1994). Numerous efforts have focused on manipulation of transgene expression to enhance fruit flavonoids (Bovy *et al.* 2002; Colliver *et al.* 2002; Muir *et al.* 2001). Willits *et al.* (2005) identified a wild accession that expressed structural genes of the anthocyanin biosynthetic pathway in the fruit peel and fruit flesh. Introgression of the *S. pennellii* accession into tomato produced progeny that accumulated high levels of quercetin in fruit flesh and peel. The mutation responsible for the lack of accumulation of yellow color flavonoid in the pink tomato has been identified (Adato et al., 2009; Ballester et al., 2010). Phenolic acids form a diverse group. Hydroxycinnamic acid esters of caffeic acid predominate in Solanaceous species and chlorogenic acid is the most abundant (Molgaard and Ravn 1988). Rousseaux *et al.* (2005) noted large environmental interactions for fruit antioxidants and identified several QTL for total phenolic concentration in fruit of *S. pennellii* introgression lines.

2.2.2 Sensory quality

Fresh-market tomato breeders improved yield, disease resistances, adaptation to greenhouse conditions, fruit aspect, but have lacked clear targets for improving organoleptic fruit quality. Consumers have complained about tomato taste for years (Bruhn et al., 1991). Nevertheless improving sensory fruit quality is complex as it is determined by a set of attributes, describing external (size, color, firmness) and internal (flavor, aroma, texture) properties.

Flavor is mostly due to sugars and organic acids (Stevens et al., 1977), to their ratio (Stevens et al., 1979; Bucheli et al., 1999), and to the composition in volatile aromas (Klee and Tieman 2013). Sweetness and acidity are related to sugars and acids content (Malundo et al., 1995; Janse and Schols, 1995). Sweetness seems to be more influenced by the content in fructose than in glucose, while acidity is mostly due to the citric acid, present in higher content than malic acid in mature fruits (Stevens et al., 1977). Depending on the studies, acidity is more related to the fruit pH or to the titratable acidity (Baldwin et al., 1998; Auerswald et al. 1999). Both sugars and acids contribute to the sweetness and to the overall aroma intensity (Baldwin et al., 1998). More than 400 volatiles have been identified (Petro-Turza, 1987), a few of them contributing to the particular aroma of tomato fruit (Baldwin et al., 2000; Tieman *et al.* 2017). Texture traits are more difficult to relate to physical measures or to fruit composition, although firmness in mouth is partly related to instrumental measure of fruit firmness (Causse et al., 2002), and mealiness was found related to the texture parameters of the pericarp (Verkeke et al., 1998). Several studies intended to identify the most important characteristics for consumer preferences (Causse et al., 2010).

Although production of high quality fruits is dependent on environmental factors (light and climate) and cultural practices, a large range of genetic variation has been shown, which could be used for breeding tomato quality as reviewed by Davies and Hobson (1981), Stevens (1986) and Dorais *et al.* (2001). Preferences of consumers faced to genetic variability have rarely been studied. Causse *et al.* (2003) showed the importance of flavor and secondarily of texture traits in consumer appreciation. Cherry tomatoes have been identified as a source of flavor (Hobson and Bedford 1989), with fruits rich in acids and sugars. Long shelf life cultivars have been described as generally less tasty than traditional ones (Jones 1986), with lower volatile content (Baldwin et al., 1991). Furthermore quality has a subjective component and there is not a unique expectation (Causse et al., 2010).

Wild relatives of *S. lycopersicum* may be interesting for improving fruit composition. Mutations of enzymes involved in the carbon metabolism were found in *S. chmielewskii* and in *S. habrochaites*, leading to particular sugar compositions: The *sucr* mutation in an invertase gene, in S. *chmielewskii*, provides fruits with sucrose instead of glucose and fructose (Chetelat et al., 1995). In *S. habrochaites*, an allele of the ADP glucose pyrophosphorylase enzyme was identified as much more efficient than the allele of the cultivated species, leading to an increase in the final sugar content of the fruit (Schaffer et al., 2000). Another locus *Fgr* modulates the fructose-glucose ratio in mature fruit, *S. habrochaites* allele yielding higher ratio (Levin et al., 2000). The gene responsible is a sugar transporter of the SWEET family (Shammai et al., 2018). A gene *Lin5* encoding an apoplastic invertase has been shown to be a QTL modulating sugar partitioning, the allele of *S. pennellii* leading to higher sugar concentrations than the *S. lycopersicum* one (Fridman et al., 2000). Wild tomato species may also provide original aromas, either favorable to tomato quality (Kamal *et al.* 2001) or unfavorable (Tadmor et al., 2002). Several genes responsible for the variation of aroma production in tomato have been cloned (Klee 2010; Bauchet et al., 2017; Zhu et al., 2019).

Many efforts for improving fruit quality have failed because of the complex correlations between the various components or between yield or fruit weight and fruit components. The correlation between fruit weight and sugar content is frequently negative (Causse et al., 2001), but may be positive in other samples (Grandillo and Tanksley, 1996a). In several studies involving sensory evaluation and fruit composition analyses, sweetness was positively correlated with reducing sugar content and sourness with titratable acidity (Baldwin et al., 1998; Causse et al., 2002). Firm texture is positively correlated with the instrumental firmness (Lee et al., 1999; Causse et al., 2002). Correlations were also detected between fruit size and antioxidant composition (Hanson et al., 2004). High throughput metabolic profiling allowed getting insight on the whole metabolic changes in tomato fruits during fruit development or in various genotypes (Schauer et al., 2005; Overy et al., 2005; Baxter et al., 2005).

Answering to the demand of producers and retailers of fresh-market tomatoes, breeders have considerably improved external aspect and shelf life of tomato fruit. This improvement was obtained either by the use of the ripening mutations or by the cumulative effect of several genes improving fruit firmness. Several mutations affecting fruit ripening are known, *rin* (ripening inhibitor) the most widely used, *nor* (non ripening), and *alc* (alcobaca). Long shelf life cultivars have invaded the tomato market in the 90's, but consumers have criticized their flavor (Jones, 1986; McGlasson et al., 1987). The corresponding genes have been identified and extensively studied (Vrebalov et al., 2009; Ito et al., 2017; Wang et al., 2019). The impact of the enzymes involved in cell wall modifications during ripening on fruit firmness and shelf life has been extensively studied and modifications of polygalacturonase or pectin methyl esterase activity were proposed to increase fruit shelf life and texture properties (Hobson and Grierson, 1993).

Processing tomato has specific quality attributes. The self pruning mutation (sp), characteristic of all the processing varieties, controls the determinate growth habit of tomato plants. Processing cultivars associate the sp mutation with concentrated flowering, fruit firmness and resistance of mature fruits to over-ripening, allowing a unique mechanical harvest. The sp gene was cloned (Pnueli et al., 1998). This mutation does not only affect plant architecture, but also modulates the expression of genes controlling fruit weight and composition (Stevens, 1986; Fridman et al., 2002; Quinet et al., 2011). This gene belongs to a gene family which is composed of at least six genes (Carmel-Goren et al., 2003). Recently, this gene was also shown to be responsible for the loss of day-length-sensitive flowering (Soyk et al., 2017). The jointless mutations, provided by the j and j2 genes, are also useful to processing tomato production. The j2 mutation has been discovered in a S. cheesmaniae accession, and has no abscission zone in fruit pedicel allowing harvest without calyx and pedicel during vine pick-up (Mao et al., 2000; Budiman et al., 2004).

2.2.3 Mild stress as a tool to manage quality

Tomatoes are produced all year-round under contrasting environmental conditions, triggering seasonal variations in their sensory quality. Over the tomato growth cycle, different factors such as light intensity, air and soil temperatures, plant fruit load, plant mineral nutrition or water availability influence the final fruit quality (reviewed in Davies and Hobson, 1981 and Poiroux-Gonord *et al*. 2010). Variations in temperature and irradiance during ripening affect carotene, ascorbic acid and phenolic compound content in the fruit, although acid and sugar content are not modified considerably by these two factors (Venter *et al.* 1977; Rosales *et al.* 2006 and Gautier *et al.* 2008). Changes in plant fruit load through trust pruning modify fruit dry matter content and final fruit fresh weight by disrupting the carbon flux entering to the fruit (Bertin *et al.* 2000; Guichard *et al.*

2005). Water limitation and irrigation with saline water may impact positively tomato fruit quality, mainly through an increase in sugar content in fruit (either by concentration or accumulation effect) and contrasted effects on the secondary metabolite contents (Mitchell *et al*. 1991; De Pascale *et al.* 2001; Nuruddin *et al*. 2003; Johnstone *et al.* 2005; Gautier *et al.* 2009; Ripoll et al. 2016). The effects reported on fruit composition are associated or not to large yield loss depending upon the intensity and duration of the treatment and the development stage of the plant (see Ripoll *et al*. 2014 for review) and result from modifications of the water and carbon fluxes imported by the fruit during its growth (Guichard *et al.* 2001; Albacete *et al.* 2013; Osorio *et al.* 2014).

Thus, the optimization of the growth practice, in particular water management, is considered in horticultural production as a tool to manage fruit quality while limiting yield losses, offering the opportunity to address simultaneously environmental issues and consumer expectations of tastier fruits (Stikic *et al.* 2003; Fereres *et al.* 2006; Costa *et al.* 2007). The genetic variability of tomato response to water limitations and others abiotic constraints and their combination still need to be deciphered to develop genotypes adapted to these practices (Poiroux-Gonord *et al.* 2010; Ripoll *et al.* 2014). Large phenotypic variation in response to a wide range of climate and nutrition conditions exists in the genus *Solannum* at both inter and intra species levels (reviewed in Labate, 2007).

Several authors attempted to measure genotype by environment (G x E) interactions on tomato fruit quality by repeating a same experiment in different locations or/and under several growing facilities (Auerswald *et al.* 1999; Johansson *et al.* 1999; Causse *et al.* 2003) or by building experimental design to isolate the effect of particular environmental factors on large number of genotypes (see Semel *et al.* 2007; Albert *et al.* 2016a; Gur *et al.* 2011 for water availability and Monforte *et al.* 1996; Monforte *et al.* 1997a, Monforte *et al.* 1997b for salt stress). In the different experiments, the G x E interaction was significant for the fruit quality traits measured (including fruit fresh weight, secondary and primary metabolism contents and fruit firmness), but generally accounted for a low part of the total variation in comparison to the genotype main effect. Albert *et al*. (2016a) dissected further the genotype by watering regime interaction in an intraspecific *S. lycopersicum* recombinant inbred line population grown under two contrasting watering regimes in two locations. In their studies, the interaction resulted from genotype re-ranking across the watering regime rather than scale changes. Besides, they identified large genetic variation and genetic heritabilities under both watering regimes, encouraging the possibility to develop tomato genotypes with an improved fruit quality under deficit irrigation.

2.3 Biotic and abiotic stresses

2.3.1 Biotic stresses

2.3.1.1 Pests and pathogens of tomatoes

Pests and pathogens cause great damage to tomato crops in field and in greenhouse. Tomato is afflicted by at least 200 pests and pathogens, from most major classes such as bacteria, fungi, oomycetes, viruses, nematodes, insects and spider mites (Foolad and Panthee 2012). Insects are as diverse as aphids, thrips, whiteflies, leafminers, fruit borers, caterpillars, leafhoppers; they disturb the foliage development perturbing photosynthesis carbon assimilation, deform fruit appearance, and ultimately reduce the yield. Moreover, several of them may transmit viruses. A few viruses may also be transmitted by contact such as Tobamoviruses. Foolad and Panthee (2012) made a compendium of the most important diseases on tomato caused by 21 fungi, 1 oomycete, 7 bacteria, 7 viruses, and 4 nematodes.

Diseases contribute to almost 40% of tomato yield loss in the field worldwide, whilst the global food production has to be increased by 60% to feed the further 10-billion world population in 2050. The occurrence of those diseases varies according to the geographical regions where tomatoes are grown, environmental conditions and cultural practices. For instance, high relative humidity favors the stem canker and the early blight caused by different species of *Alternaria*, and warm air temperature and damp conditions favor the gray leaf spot caused by different species of *Stemphylium* whilst low soil temperature favors the corky root rot caused by *Pyrenochaeta lycopersici* and cool air temperature the *Fusarium* crown and root rot. Otherwise, high air humidity alternating with cool night temperature is favorable for the development of late blight caused by the Oomycete *Phytophthora infestans* that can easily destroy up to 100% of field or greenhouse tomato crops.

2.3.1.2 Impact of climate change on pest and pathogen resistance

Climatic prediction models indicate severe weather pattern changes, which will result in frequent droughts and floods, rising global temperatures, and decreased availability of fresh water for agriculture. A great challenge is thus to improve the robustness of plant resistance and tolerance to pests and pathogens, to a wide array of combined biotic and abiotic stress combinations. Tomato crops are exposed to multiple abiotic stresses in fields and greenhouses that could attenuate or enhance the response to biotic stress. Recent studies have revealed that the response of plants to combinations of two or more stress conditions is unique and cannot be directly extrapolated from the response of plants to each stress applied individually. Few studies report the tomato responses to biotic x abiotic stress combinations.

It is well known for long time that high temperatures (above 30°C) inhibit plant defense mechanisms making major resistance genes frequently dysfunctional. For instance, the tomato *Mi-1.2* resistance gene to root knot nematode and *Cf-4* / *Cf-9* genes to *Cladosporium fulvum* are inactivated at high temperature (de Jong et al. 2002; Marques de Carvalho et al. 2015). Other abiotic stresses could also modify tomato immunity. For instance, drought stress reduces disease severity to *Botrytis cinerea* and stops the development of *Oidium neolycopersici*. Irrigation with saline water increases disease severity to *Fusarium oxysporum f. sp. radicis-lycopersici* and to *Phytophthora capsici*, does not affect *B. cinerea* infection, and reduces infection by *O. neolycopersici* (Achuo et al. 2006; Dileo et al. 2010). Bai et al. (2018) suggest that salt stress modifies the hormone balance involved in signaling pathway that could decrease the resistance level conferred by the *Ol-1* gene but has no effect on resistance conferred by *ol-2* and *Ol-4* genes, those three genes controlling *O. neolycopersici* responsible for tomato powdery mildew. Limited nitrogen or water supplies increase tomato stem susceptibility to *Botrytis cinerea* (Lecompte et al. 2017). Very high environmental pressure caused by elevated ozone concentration eliminates the effect of potato spindle tuber viroid (PSTVd) on biomass reduction in tomato (Abraitiene and Girgzdiene 2013). The few examples cited here mainly focused on the effect of environmental changes on tomato immunity controlled by major resistance genes. Much less publications concern resistance QTLs yet, even if research on the effect of G x E interactions on resistance to biotic stress is increasing. Actually, there is a knowledge gap in the identification of QTLs involved in responses to combined biotic x abiotic stress.

2.3.1.3 New emerging tomato diseases

Global climate change is supposed to result in the emergence of new pests and pathogens into production areas. Tomato health management is thus challenged by the emergence of new races that overcome resistance genes deployed in cultivars and by novel introductions due to the world agricultural market and the climate change. Several diseases are reemerging or emerging on tomato crops such as the late blight caused by *P. infestans* (Fry and Goodwin 1997), the leafminer *Tuta absoluta*, and new viruses that increasingly affect tomato crops. The Potexvirus *Pepino mosaic virus* (PepMV), mainly mechanically transmitted, emerged around 2000 and causes now significant problems on glasshouse tomato crops worldwide (Hanssen and Thomma 2010). Recently, the *tomato brown rugose fruit virus* (ToBRFV), a new tobamovirus present in Jordania and Israel, was able to break *Tm-2*-mediated resistance in tomato that had lasted 55 years (Maayan et al. 2018). Emergence of new viruses is often coupled to the proliferation of adapting insect vectors. Tomato production in tropical countries is severely constrained by insects and mites, particularly whiteflies (*Bemisia tabaci*) that could transmit begomoviruses (including TYLCV known for long time but also many other emergent begomoviruses) and fruit borers that cause serious problems during the reproductive phase of the crop. Deploying host resistance against viruses, when available, is actually the most effective method for controlling viruses and preventing their spread, even if in recent years resistance-breaking strains of viruses have been characterized, against which these resistance genes are no longer effective. For example, the resistance gene *Sw-5* confers resistance to TSWV transmitted by the thrips *Frankliniella occidentalis*, as well as to related orthotospovirus species such as *Groundnut ringspot virus* (GRSV) and *Tomato chlorotic spot virus* (TCSV) recently emerged in the United States and the Caribbean. But it has been overcome by new virulent TSWV strains (Oliver and Whitfield 2016; Turina et al. 2016).

In addition, the bacteria *Clavibacter michiganense* subsp. *michiganensis* (Cmm), causing the bacterial canker disease devastating tomato production worldwide, is considered as a real plague. This bacteria is one of the few pathogens transmitted by seeds. To fight the spread of this disease, Good Seed and Plant Practices (GSPP; https://www.gspp.eu/), adopted by sites or companies working on tomato breeding and plantlet production, prevent tomato seed and plant lots from being infected by Cmm. GSPP-accredited sites or companies are granted the right to market their tomato seeds and young plants with the GSPP logo. The first GSPP seed and plants have been available since July 2011 in France and the Netherlands.

So far, there is no sufficiently sustainable or effective genetic leverage available for tomato breeding programs to combat these new diseases. Their sustainable control is a goal of global importance, which will probably require combining several genetic strategies associated to cultural practices to effectively managing those novel pathosystems.

2.3.2 Abiotic stresses

Tomato domestication and improvement have focused for a long time on agronomic traits associated to productivity, quality and disease resistances. Crop resilience facing the global climate change nowadays represents one of the most challenging aspects in plant breeding, raising awareness in developing climate-smart crops. It has led to the characterization of new breeding traits related to abiotic stress tolerance. Understanding the complex genetic architecture of plant response to environmental changes appears to be central for the development of new cultivars. Indeed, variations in environmental factors usually induce some disorders at molecular, physiological and morphological levels that may alter agronomic performance of crops. Stress adaptation in plants at the molecular level requires generally the activation of multiple stress-response genes that are involved in different metabolic pathways for growth maintenance and which expression is regulated by various transcription factors (TFs). The genomic era facilitated the characterization of such stress-response genes across plant species that were assigned to diverse family of TFs. The major families of TFs playing significant

roles in stress tolerance that were described in the literature include the basic leucine zipper (bZIP), dehydration-responsive element-binding protein (DREB), APETALA 2 and ethylene-responsive element binding factor (AP2/ERF), zinc fingers (ZFs), basic helix-loop-helix (bHLH), Heat-Shock proteins (Hsp) and the NAC, WRKY, MYB among others (Lindemose *et al.* 2013). The functions covered by these TFs are very common in the plant kingdom, however each species present specificities.
In tomato, Bai *et al.* (2018) characterized the 83 WRKY genes identified in previous studies and displayed their different roles in response to pathogen infection, drought, salt, heat and cold stresses. Some genes were highlighted as being altered in their expression by different stress such as drought and salinity stress (*SlWRKY3; SlWRKY3 and SlWRKY33*) pointing pertinent candidates for further investigation. The expression profiles of other tomato stress-response genes were also investigated for a class of genes belonging to the ERFs family (Klay *et al.* 2018) and Hsp20 gene family (Yu *et al.* 2016). Examples of single genes involved in tomato tolerance to abiotic stress were also described including the *SlJUB1* promoting drought tolerance; *DREB1A* and *VP1.1* playing a role in salinity tolerance and *ShDHN*, *MYB49* and *SlWRKY39* for tolerance to multi-stress factors (Liu *et al.* 2015; Sun *et al.* 2015; Cui *et al.* 2018).
Tomato is a suitable plant model to study the genetics of plant response to the environment and deciphering the genotype-by-interaction (GxE) mechanisms, due to the wide range of environmental conditions – from fields to greenhouse cultivation – for its production highlighting its large adaptability.

2.3.2.1 Water deficit

Tomato is a high water-demanding crop (Heuvelink 2005) making water resource management one of the key factors essential for the crop. The amount of irrigation water in tomato production is usually managed according to the reference evapotranspiration (ET_0) and the developmental stage. When water deficit (WD) occurs during the cropping period, morphological and molecular changes are usually observed that hamper the final yield production. Several studies addressed the impact of WD stress on tomato, most of which establishing WD as a percentage of water restriction, according to the optimal water requirement (Albert *et al.* 2016a,b; Diouf *et al.* 2018; Ripoll *et al.* 2016).
From an agronomic point of view, the main consequence of WD on tomato is yield reduction, that can be severe when stress occurs during fruit development (Chen *et al.* 2013). However, all developmental stages are susceptible to WD to a level depending on the cultivar and stress intensity. Seed germination is the first step exposed to environmental stress. In tomato, a delay or even an inhibition of seed germination was observed with the application of osmotic stress (Bhatt and Rao 1987). Water deficit during vegetative and reproductive development negatively affects the overall economic performance of the crop but positive effects on fruit quality are documented. Indeed, Costa *et al.* (2007) described some trade-off between yield decrease and increase in quality component on fruit trees and vegetables including tomato where enhancement in fruit quality compounds such as vitamin C, antioxidants and soluble sugars was observed under WD stress (Albert *et al.* 2016a; Ripoll *et al.* 2014; Patanè and Cosentino 2010; Zegbe-Domínguez *et al.* 2003). The two groups of accessions constituted of cherry tomato and large fruit accessions usually show different sensitivity to environmental stresses. For instance, a study using a panel of unrelated lines tested under control and WD conditions revealed that large fruit tomato accessions were more susceptible and had higher responsiveness to WD (Albert *et al.* 2016b). This study also showed that the increase in the sugar content in fruit under WD is due to a reduction in fruit water content and not to increased synthesis of sugars. However, Ripoll *et al.* (2016) found higher fructose and glucose synthesis in tomato fruits submitted to WD stress for different stages of fruit development, indicating that both dilution effect and higher sugar synthesis are responsible of fruit quality enhancement in tomato under WD. The Omics approaches allow targeting specific genes and studying their variation in expression level according to different environmental conditions. Some examples of water deficit response genes involved in tomato tolerance to drought are published. This is the case for *SlSHN1* gene that induces tolerance to drought by activating downstream genes involved in higher cuticular wax accumulation on leaves (Al-Abdallat *et al.* 2014). Tolerance to drought induces an early activation of signaling pathways to elicit drought related genes. Wang *et al.* (2018) identified a drought-induced gene (*SlMAPK1*) playing an active role in the antioxidant enzymes activities and ROS scavenging leading to higher drought tolerance.

2.3.2.2 Salinity stress

Soil salinity has become problematic in agriculture especially in the Mediterranean region where soil aridification and non-sustainable irrigation practices tend to increase the surface area of salty soils (Munns and Tester 2008). Munns and Gilliham (2015) defined salinity stress (SS) as the level of salinity up to which the energy for plant growth is redirected into defense response. Considering yield as a measure of tolerance to SS, tomato is a crop that can tolerate up to $2.5 dS.m^{-1}$ of salinity and cherry tomatoes are less salt sensitive than large fruit accessions (Scholberg and Locascio 1999; Caro *et al.* 1991). Over the above-mentioned threshold, a significant yield decrease is observed. Yield reduction under SS in tomato was found to be associated to a reduction in both fruit size and fruit number (Scholberg and Locascio 1999). As for WD, SS also leads to an increase in sugar content in tomato fruits (Mitchell *et al.* 1991). Besides, SS leads to changes in the cation/anion

ratio and the increase in sugar content in fruits of salinized plants likely results from the interaction between reduced fruit water content, increased ion content, and maintained hexose accumulation (Navarro *et al.* 2005). These changes are the consequences of tomato response to osmotic adjustment. The threshold for salinity tolerance defined above was set upon the characterization of few selected tomato cultivars. However, Alian *et al.* (2000) noticed a high genotypic variability in response to salinity in fresh market tomato cultivars. This highlights the possibility and the potentiality for the crop to breed salt-tolerant cultivars.

Facing SS, plants deploy a variety of response to rebalance and reestablish the cellular homeostasis. Physiological responses to SS involve the ionic channels transporters as they are highly needed to regulate the ionic imbalance (Apse *et al.* 1999). In their study, Rajasekaran *et al.* (2000) screened salinity tolerance in a number of tomato wild relatives and associated salinity tolerance mainly to a higher K^+/NA^+ ratio in roots. High genetic variability was observed in *S.pimpinellifolium* accessions for yield and survival traits in response to SS (Rao *et al.* 2013). Among yield component traits, fruit number was the most affected trait in both wild and cultivated populations (Rao *et al.* 2013; Diouf *et al.* 2018). Breeding salt-tolerant variety thus seems possible by using either physiological traits or agronomic performance under salinity, as sufficient genetic variability is available in several tomato genetic resources.

2.3.2.3 Temperature stress

All crop species have an optimal temperature range for growth. Tomato is known as a crop that can grow in a wide range of environments, from elevated areas with low temperatures to tropical and arid zones where high temperatures usually occur. Based on crop simulation model, Boote et al., (2012) indicated that the optimal growth for tomato and its fruit development is about 25°C. Temperatures below 6°C and above 30°C severely limit growth, pollination and fruit development and could negatively impact final fruit yield. Studies on different accessions and wild relative species of tomato helped understanding how the crop its responds to low and high temperature stresses.

High temperature stress

The most visible effect of climate change is the rise in temperature in different areas of the world. The end of the 21st century is expected to come with the increase in global warming causing significant yield decrease in major worldwide cultivated crops (Zhao *et al.* 2017). When plants are exposed to fluctuating high temperatures (HT), ensuing stress are considered as short-term heat stress when the period of exposure to HT is short or long-term heat stress if plants experienced the HT for several consecutive days. The latter has more dramatic effects on agronomic performances of crops, especially when it occurs during the entire cropping season. In open field trials, seed germination is more generally impaired by high temperature of the soil and can differ to effects of elevated air temperatures. However, flowering period is described as the most critical stage under HT stress (Wahid *et al.* 2007). Severe yield decrease caused by HT stress arises from the hampered reproduction performance with a high impact of HT on reproductive organs (Nadeem *et al.* 2018). In tomato, HT stress around flowering was shown to inhibit reproduction by altering male fertility at high degree and female fertility at a lower rate (Xu *et al.* 2017). In areas where the temperature range could be reliably predicted, managing the sowing date to avoid HT stress around anthesis is an important factor to consider. Tomato male fertility could be considered as the main factor limiting reproduction success under HT stress. This has led some studies to use of pollen traits as a measure of heat tolerance instead of only final yield (Driedonks *et al.* 2018). Male reproductive traits were highly variable among wild species and some accessions showed high pollen viability compared to cultivated cultivars. This opens possibilities for transferring heat-tolerance alleles from wild donors to cultivated tomato. A reduction of fruit setting was also observed in cultivated tomato with higher rate of parthenocarpic fruits noticed under HT stress at 26°C in growth chambers (Adams *et al.* 2001). These authors noticed that fruit maturation is accelerated under higher temperature mostly when fruits are exposed themselves to heating periods, that could alter final fruit quality composition.

Considering the important effect of HT on agriculture, numerous studies successfully tackled and identified several heat-response genes (Waters *et al.* 2017; Keller and Simm 2018; Fragkostefanakis *et al.* 2016). Heat-response genes are commonly regulated by the activity of several heat stress transcription factors (HSFs) as described in the literature for different organisms. This has led to the investigation of the roles played by HSFs in thermo-tolerance and majors HSFs depicted across plant species could lead to the development of heat-tolerant tomato via genome editing (Fragkostefanakis *et al.* 2015).

Chilling and cold stress

Chilling stress (CS) is usually considered when plants are growing in temperature below the optimal growth range and above 0°C, just before freezing stress. The geographical distribution of wild tomato species include elevated zones where annual temperatures can be below the optimal growth for cultivated tomatoes (Nakazato *et al.* 2010). This denotes that adaptation to sub-optimal temperature is possible in tomato.

Adams *et al.* (2001) observed that at 14°C, tomato growth was reduced. Lower temperatures equally induce some chilling stress symptoms as reviewed by Ploeg and Heuvelink (2005) who noticed that below 12°C almost no growth is observed for tomato. As for HT stress, fruit set is inhibited in tomato mainly due to poorer pollen viability. Reduction in the number of flowers, number of fruits and final yield was observed with low temperature that also affects the partitioning of photosynthetic products (Meena *et al.* 2018). Indeed,

photosynthesis is highly impacted during CS and several related physiological parameters are described. For example, the relative water content, chlorophyll fluorescence and accumulation of phenolic compounds are associated to mechanisms inducing cold tolerance (Giroux and Filion 1992; Dong *et al.* 2019; Khan *et al.* 2015). By the way, Meena *et al.* (2018) showed that external application of phenolic compounds – notably salicylic acids – significantly increased tomato tolerance to CS. Low temperature stress during plant growth and development adversely affects fruit quality of tomato and reduces non-enzyme antioxidants such as lycopene, β-carotene and α-tocopherol.

Transcriptome analysis depicted some genes responding to CS in tomato. For example, Zhuang *et al.* (2019) identified a cold response tomato gene (*SlWHY1*) whose expression is enhanced under 4°C, playing a role in photosysteme II protection and starch accumulation in chloroplast. For several plant species, signal transmission of CS involves the C-repeat binding factor (CBF) (Jha *et al.* 2017) leading to downstream activation of cold responsive genes for cold tolerance. Major types of CBF are known to regulate cold acclimation in tomato (Mboup *et al.* 2012). In a recent review, Kenchanmane Raju *et al.* (2018) showed that genes related to photosynthesis and chloroplast development were consistently repressed in response to low-temperature and the most conserved set of genes up-regulated in response to low-temperature stress belonged to the CBFs, WRKYs, and AP2/EREBP transcription factors. These results highlighted some genes and family of transcription factors that could be targeted for breeding tomato adapted to low temperature conditions.

2.3.2.4 Mineral nutrition deficiency

The positive effect of mineral nutrition on plant growth has long been recognized and mineral elements are usually classified as essential or non-essential; the latter being however beneficial for plant development (Marschner 1983). The macronutrients are mostly necessary to stimulate growth and nitrogen (N), potassium (K^+), and phosphorus (P) are among the most important in higher plants. Their use has a significant environmental cost and thus selection for reduced need of fertilizer could be useful for the production of smart crops.

Nitrogen

Nitrogen *(N)* is among the most important limiting nutrient for tomato development. Insufficient N nutrition can cause severe consequences to economically important traits. It was shown that N deficiency negatively affect the number of fruits, fruit size, storage quality, color, and taste of tomato (Sainju *et al.* 2003). As evidenced by Groot *et al.* (2004) and Larbat *et al.* (2012), tomato growth rate is linearly correlated to N supply. Low N supply limits growth in leaves but promotes root development and this activity was mainly linked to variation in cytokinin concentration. An increase in accumulation of phenolic compounds is also a notable consequence of N deficiency in tomato. Indeed, Larbat *et al.* (2012) found that sequential limitation of N nutrition resulted in an up-regulation of genes associated to phenolic biosynthetic pathway.

Oversupply of N above the required optimal level is usual in tomato cultivation due to its beneficial effects and the willing to avoid the negative effects of limited N; however, excess of N can overproduce vegetative growth at the expense of fruit development and rapid fruit maturation and inhibits root system development beside its negative effect on groundwater pollution (Du *et al.* 2018). This highlights the necessity to manage N nutrition in tomato cropping that can be achieved through a good characterization of genes involved in nitrogen-use efficiency. Apart from genetic solutions to improve tolerance to N-deficiency, real time greenhouse management technics are now available with the use of computational intelligence systems and definition of new stress tolerance traits like leaf reflectance as proposed by Elvanidi *et al.* (2018).

Phosphorus

Phosphorus *(P)* is usually present in the soil in a form that is not accessible for plants. Fertilization is thus required for major crops including tomato. Plant capacity to acquire P present in the soil is associated to root morphological changes and involves variation in plant-hormone levels. Early plant development is very sensitive to P nutrition and sub-optimal P supply in tomato can lead to impaired growth and plant development (Sainju *et al.* 2003; de Groot *et al.* 2004). Phosphate deficiency induces modification in root architecture morphology via increased auxin sensitivity leading to the activation of P transporter genes to remobilize P from lipids and nucleic acids (Schachtman and Shin, 2007). Long-term adaptation to P starvation appears to be linked to reduced primary root growth at the expanse of lateral root growth that is promoted (Xu et al., 2012). Besides, the net-photosynthesis decreased in the leaves with reduced sucrose content after long exposure to P starvation, while the starch content increased. Theses authors also identified different genes responding to P starvation that belong to the 14-3-3 gene family encoding phosphoserine-binding proteins involved in protein-protein interactions.

In open field conditions, a larger root system development may be required for greater exploration and acquisition of P present in the soil. For greenhouse production where the P input can be managed, the need is more in the characterization of P-deficiency response genes and their correlation to morphological and physiological response for the development of cultivars with higher P-use efficiency.

Potassium

The importance of *Potassium (K^+)* in plant nutrition has been attested with its involvement in important physiological processes such as photosynthesis, osmoregulation and ion homeostasis (Marschner 1983;

Pettigrew 2008). Yield and quality are known to be impacted by the photosynthesis capacity of the plant and thus could be directly linked to the K^+ concentration in plant organs. In tomato, positive effects of K^+ supply have been described for vigorous growth, early flowering, fruit number production and higher rate of titratable acidity (Sainju et al. 2003). Increase in soluble solids, anti-oxidative capacity and ascorbic acid were also observed in tomato fruits (Tavallali *et al.* 2018) with K^+ supply. Alternatively, deficiency in K^+ nutrition induced morphological injuries resulting in brown marginal scorching with interveinal chlorosis and yellowing of tomato leaves. Indeed, plants usually sense external changes in K^+ concentration leading to the activation of signal transduction to reestablish the ion homeostasis. Adaptation to low K^+ supply is achieved through different K^+ movement monitored by different K^+ transporters. The function and role of different transporter channels involved in K^+ movement in plants were described by Wang and Wu (2015) including the HAK/KUP/KT family of transporters seemingly crucial for K^+ transport. The transport of K^+ in plants is initiated in the roots and the major impact of K^+ deficiency is on root architecture (Zhao *et al.* 2018). Improving root system development could then directly alleviate the deleterious effect of K^+ deficiency.

Calcium

Calcium is an important ion involved in diverse metabolic processes central to plant growth and development (Bush, 1995). Several reviews regarding the role of this macronutrient on plants pinpoint its involvement in the cell wall rigidity, cell membrane stability, the control of ion transport and the signaling of abiotic stress (Hepler, 2005; Hirschi, 2004; Wilkins *et al.* 2016). Calcium deficiency is associated to changes in the cell ion homeostasis and had been related to nutritional imbalance incidence, among other problems in plants. The diminution of $Ca2^+$ nutrition as well as environmental stimuli have been considered as leading changes in cytosolic concentration of Ca^{2+} mediating some modifications in Ca^{2+} flux through transporter proteins in order to reestablish the ion homeostasis (Bush, 1995). Besides, plant response to abiotic stresses are tightly linked to modification in Ca^{2+} homeostasis essential to signaling and subsequent plant tolerance deployment (Rengel, 1992; Wilkins *et al.* 2016). In tomato, Ca^{2+} nutrition under salinity stress for example has been shown to alleviate the negative impact induced by salt toxicity on plant and fruit growth (Tuna *et al.* 2007). This was linked to Ca^{2+} use efficiency upon the availability of sufficient Ca^{2+} concentration in the plant. Calcium-use efficiency is an important characteristic for plant adaptation to environmental stress and this trait is genetically variable indicating the possibility for breeding cultivars with high potentiality of adaptation to low Ca^{2+} input (Li and Gabelman, 1990). However, most tomato accessions are susceptible to Ca^{2+} deficiency and among the undesirable effects associated to this stress, a physiological disorder at the fruit named blossom-end rot (BER) has been noticed (Adams and Ho, 1993). Other studies correlate BER incidence to differences in genotype capacity to limit oxidative stress by increasing the synthesis of antioxidant metabolite such as ascorbate (Rached *et al.* 2018) or genotype sensitivity to gibberellin (Gaion *et al.* 2019) suggesting a non-direct effect of Ca^{2+} depletion in the cells to induce BER symptoms. Moreover, through transcriptomic analyses, de Freitas *et al.* (2018) identified candidate genes inhibiting BER in tomato that were mostly associated to resistance against oxidative stress. Tomato BER is thus a complex physiological disorder occurring from the impact of abiotic stresses, genetic, physiological or agronomic factors with possible interaction between them (Hagassou *et al.* 2019). However, regarding the tight link between BER and the level of Ca^{2+} in tomato, the characterization of the channel gene families involved in regulation of Ca^{2+} homeostasis under different environmental stimuli could help to disentangle the underlying molecular mechanisms of the interaction between BER incidence and Ca^{2+} concentration.

2.3.3 Stress combination

Plant responses to individual stress at specific growth stage are well documented and avenues for crop breeding to enhance tolerance to a particular stress were provided. However, observations in the nature and in open field conditions clearly brought to light that stress combination is a common phenomenon, especially with the climate change that has an incidence on co-occurring of environmental stresses such as WD and HT stress. Climate change trend has also an impact on pathogen spreading and new disease appearance and distribution (Harvell *et al.* 2002). Different scenarios of biotic and abiotic stress combination are then expected to arise, according to the geographical regions and areas of crop cultivation. With different crop species exposed to different stress treatments, Suzuki *et al.* (2014) presented a stress matrix with the potential positive and negative effects of various patterns of stress combination. The global effect of combined stresses on yield, morphological and physiological traits on plants can be highly different from those of a single stress. Thus the stress matrix proposed by Suzuki *et al.* (2014) would be highly useful if specified for tomato, to achieve a global view of how stress combinations could be managed in breeding programs.

Examples of studies conducted in tomato to assess the impact of combined stress on different traits are available in the literature. Zhou *et al.* (2017) showed that physiological and growth responses to the combined WD and HT stresses had a similar pattern across different cultivars but the response was different from the single heat response. Combination of HT stress and SS on tomato showed however less damage on growth than the application of SS alone (Rivero *et al.* 2014). Beside morphological changes, some studies conducted on the model species *Arabidopsis thaliana* demonstrated that variations in gene expression under stress combination are highly independent of variation induced by single stress application (Rasmussen *et al.* 2013).

In addition to combination of different environmental stresses, simultaneous biotic and abiotic stresses, which are usually studied separately, are expected, especially in field conditions. Recently, studies were performed to fill the lack of knowledge about the genetic response to biotic and abiotic stress combination compared to single stress effect. In tomato, Kissoudis *et al.* (2015) studied the combined effect of salinity and powdery mildew (*Oidium neolycopersici*) infection and found that salt stress increases the powdery mildew susceptibility in an introgression line population. Anfoka *et al.* (2016) showed that long-term HT stress was accompanied with TYLCV accumulation in tomato reducing by the way the HT response efficiency. Some stress responses such as endogenous phytohormone secretion and ROS production are important physiological processes involved in both abiotic and biotic plant responses (Fujita *et al.* 2006) that could require a-the action of a group of genes regulating both type of stresses. Some genes were shown to be involved in the simultaneous response to biotic and abiotic stress on tomato such as the SlGGP-LIKE gene that Yang *et al.* (2017) found to be correlated to higher ascorbic acid synthesis, less ROS damage and higher tolerance to chilling stress, however its suppression led to higher ROS accumulation and resistance to *P. syringae*. Using genomic data from multiple stress response genes, Ashrafi-Dehkordi *et al.* (2018) performed a comparative transcriptome analysis on tomato and found a set of genes the expression of which is altered under simultaneous biotic and abiotic stresses. Single tomato genes involved in responses to both abiotic stresses and *Pseudomonas syringae* (Sun *et al.* 2015) or *Phytophthora infestans* (Cui *et al.* 2018) were identified making them suitable targets for breeding. However, up to now, stress combination is mostly addressed in a genomic or metabolomics point of view and few examples of genetic response to combined stress are documented except in *A. thaliana* (Thoen *et al.* 2017).

The impact of mineral nutrition on plant pathogen is also important: the enhanced phenolic and volatile compounds accumulated with N fertilization has been shown to interact with tomato disease induced by insect attack such as whitefly, *Bemisia tabaci* (Islam *et al.* 2017) and leafminer *Tuta absoluta* Han *et al.* (2015). Interaction between N supply and tomato resistance to *Botrytis cinerea* has also been described (Lecompte *et al.* 2010). Nitrogen supply not only interacts with biotic tolerance in tomato but has also a different impact according to some abiotic factors.

Among abiotic stresses, salinity is the most important stress in tomato affecting tomato responses. The simultaneous effect of salinity stress and N input was measured by Papadopoulos and Rendig (1983) who showed that the positive effects of N supply on growth and fruit weight was suppressed by salinity stress reaching up to 5 $dS.m^{-1}$.

In an interspecific introgression line (IL) population, (Frary et al., 2011) showed that salinity decreased the leaf $Ca2^{+}$ content by 47% and K^{+} content by 8%. *S. pennellii* alleles were found contributing mostly to higher $Ca2^{+}$ content under both control and salinity stress suggesting this species as a natural resource for salinity and low $Ca2^{+}$ input stress tolerance.

3 Genetic and genomic resources for trait breeding

3.1 Genetic resources

3.1.1 Origin of tomato and its wild relatives

Genetic resources for food and agriculture are keys to global food security and nutrition (FAO, 2015). In crop production, maintaining genetic diversity is an essential strategy not only to breed new varieties, to identify candidate genes of target traits, to dissect the evolutionary history, but also to reduce the effects of biotic and abiotic stresses, etc.

Tomato belongs to the large and diverse *Solanaceae* family also called Nightshades, which includes more than three thousand species. Among them, major crops arose from Old world (eggplant from Asia) and New world (pepper, potato, tobacco, tomato from South America). The *Lycopersicon* clade (**Table 2**) contains the domesticated tomato (*Solanum lycopersicum*) and its 12 closest wild relatives (Peralta et al., 2005). Charles Rick and colleagues started the first prospections and studies on the tomato wild relatives in the 40's.

Tomato clade species are originated from the Andean region, including Peru, Bolivia, Ecuador, Colombia and Chile. Their growing environments range from sea level to 3,300 m altitude, from arid to rainy climate and from Andean Highlands to the coast of Galapagos Islands**.** Their habitats are often narrow and isolated valleys and they were adapted to many climates and different soil types. The large range of ecological conditions contributed to the diversity of the wild species. This broad variation is also expressed at the morphological, physiological, sexual and molecular levels (Peralta et al., 2005).

The domestication of tomato is due to a divergence from S. *pimpinellifolium* that occurred several thousand years ago. It probably happened in two steps, first in Peru, leading to S. *lycopersicum cerasiforme* accessions then in Mexico, leading to large fruit accessions (reviewed in Bauchet and Causse, 2012) and confirmed by molecular analyses (Blanca et al., 2012; Lin et al., 2014; Blanca et al., 2015). Only a few tomato seeds were brought back from Mexico to Europe, leading, after domestication, to a new genetic bottleneck. The tomato cultivation first slowly spread in southern Europe and it is only after the Second World War that its intentional selection started and that it was spread over the world.

Table 2. Tomatoes and their wild relative species of the Lycopersicon section according to Peralta et al. 2008 'Lycopersicon group' corresponds to the red- and orange-fruited species). For further details of crossability and other biological parameters of wild tomatoes see Grandillo et al. (2011).

Species	**Distribution**	**Habitat;(elevational range**	**Section according to Peralta et al. (2008)**
Solanum lycopersicum L.	Globally cultivated domesticate	Cultivated; sea level-4000 m	*Lycopersicon* 'Lycopersicon group'
Solanum pimpinellifolium L.	Southwestern Ecuador to northern Chile (many northern populations in Ecuador are admixture with *S. lycopersicum*; Peralta et al. 2008; Blanca et al. 2013)	Dry slopes, plains and around cultivated fields; sea level-3000 m	*Lycopersicon* 'Lycopersicon group'
Solanum peruvianum L.	Central Peru to northern Chile	Dry coastal deserts and lomas; sea level-3000 m	*Lycopersicon* 'Eriopersicon group'
Solanum cheesmaniae (L.Riley) Fosberg	Galápagos Islands	Dry, open, rocky slopes; sea level-1300 m	*Lycopersicon* 'Lycopersicon group'
Solanum galapagense S.C.Darwin & Peralta	Galápagos Islands	Dry, open, rocky slopes; seashores; sea level-1600 m	*Lycopersicon* 'Lycopersicon group'
Solanum arcanum Peralta	Northern Peru	Dry inter-Andean valleys and in coastal lomas (seasonal fog-drenched habitats); 100-4000 m	*Lycopersicon 'Arcanum group'*
Solanum chmielewskii (C.M.Rick, Kesicki, Fobles & M.Holle) D.M.Spooner, G.J.Anderson & R.K.Jansen	Southern Peru and northern Bolivia	Dry inter-Andean valleys, usually on open, rocky slopes; often on roadcuts; 1200-3000 m	*Lycopersicon 'Arcanum group'*
Solanum neorickii D.M.Spooner, G.J.Anderson & R.K.Jansen	Southern Ecuador to southern Peru	Dry inter-Andean valleys; 500-3500 m	*Lycopersicon 'Arcanum group'*
Solanum chilense (Dunal)Reiche	Coastal Chile and southern Peru	Dry, open, rocky slopes; sea level-4000 m (B. Igic, pers. comm. has suggested the higher elevation plants represent a new species)	*Lycopersicon 'Eriopersicon group'*
Solanum corneliomulleri J.F.Macbr.	Southern Peru (Lima southwards)	Dry, rocky slopes; 20-4500 m (low elevation populations associated with landslides in southern Peru)	*Lycopersicon 'Eriopersicon group'*
Solanum habrochaites S.Knapp & D.M.Spooner	Andean Ecuador and Peru	Montane forests, dry slopes and occasionally coastal lomas; 10-4100 m	*Lycopersicon 'Eriopersicon group'*
Solanum huaylasense Peralta	Río Santa river drainage, north-central Peru	Dry, open, rocky slopes; 950-3300 m	*Lycopersicon 'Eriopersicon group'*
Solanum pennellii Correll	Northern Peru to northern Chile	Dry slopes and washes, usually in flat areas; sea level-4100 m	*Lycopersicon 'Neolycopersicon group'*

3.1.2 Genetic resources as sources for adaptation

There are more than 83,000 tomato accessions stored in different seed banks worldwide (FAO, 2015). These seed banks include the Tomato Genetic Resources Center (TGRC) in Davis, USA (https://tgrc.ucdavis.edu/), the United States Department of Agriculture (USDA) in Geneva, USA (https://www.ars.usda.gov/), the World Vegetable Center in Taiwan, (https://avrdc.org/), the Centre for Genetic Resources, in the Netherlands (https://www.wur.nl/en/Research-Results/Statutory-research-tasks/Centre-for-Genetic-Resources-the-Netherlands-1.htm) and others. These seed banks maintain most of the genetic diversity of tomatoes.

Thanks to the pioneer work of Charles Rick, the Tomato Genetics Resource Center of the University of California, in Davis, maintains the largest collection of wild relative accessions that he prospected during his life. This collection has been an important source of diversity for breeding tomato and for gene discovery. For instance, there is a collection of 46 *Solanum pennellii* that is only found in Peru, and is particularly adapted to dry conditions (**Figure 2**).

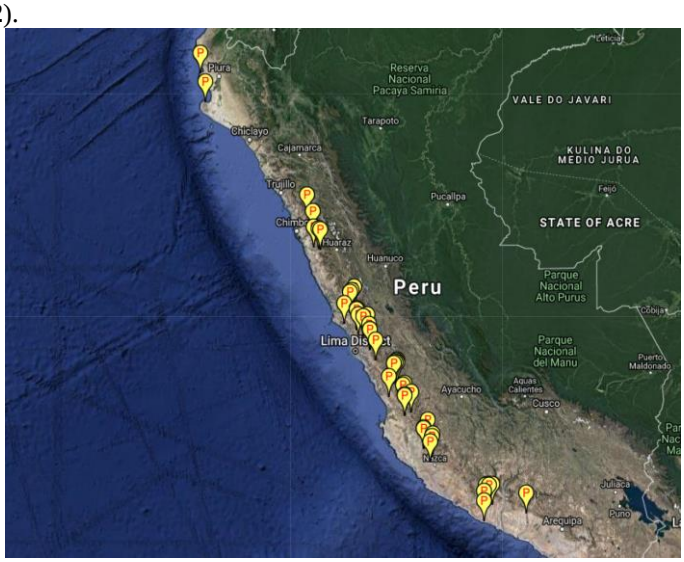


Figure 2. Geographical locations of wild tomato species *Solanum pennellii*. Data were collected from Tomato Genetics Resource Center, University of California, Davis (https://tgrc.ucdavis.edu/Data/Acc/Wildspecies.aspx).

3.1.3 Natural and induced mutants

Natural genetic diversity is the main source for adaptation and crop breeding. Natural mutations appeared in cultivated accessions or were introduced from wild relative species, which provide a great source of genetic diversity for many traits, including disease resistance genes and quality trait-related genes (Bauchet and Causse, 2012; Bauchet et al., 2017a; Rothan et al., 2019). However, the number of cloned genes with detailed functional validations is still limited (Rothan et al., 2019). Some biotechnology tools such as TILLING (Targeting Induced Local Lesions in Genomes; Comai and Henikoff, 2006) provide collections of mutants in a specific accession, accelerating functional genomic research and the discovery of interesting alleles at a given locus (Menda et al., 2004; Baldet et al., 2007; Okabe et al., 2011; Mazzucato et al., 2015; Gauffier et al., 2016). This technology typically uses chemical mutagens such as ethyl methanesulfonate (EMS) to generate several base mutations in the genome. There are several TILLING collections worldwide for tomato, such as the UCD Genome Center TILLING laboratory, University of California, USA (http://tilling.ucdavis.edu/index.php/TomatoTilling); The Microtom collection (Okabe et al., 2011); TOMATOMA database, Japan (http://tomatoma.nbrp.jp/); Repository of Tomato Genomics Resources (RTGR), University of Hyderabad, India (https://www.uohyd.ac.in/images/index.html); The Genes That Make Tomatoes (http://zamir.sgn.cornell.edu/mutants/index.html); the Tilling Platform of Tomato, INRA, France (http://www-urgv.versailles.inra.fr/tilling/tomato.htm) (Minoïa et al., 2010); LycoTILL database, Metapontum Agrobios, Italy (http://www.agrobios.it/tilling/) (Minoia et al., 2010) and others.

3.2 Molecular markers and gene/QTL mapping

3.2.1 Evolution of molecular markers

Tomato has been used for genetic studies and mutation mapping of interesting traits even before the discovery of molecular markers (Butler, 1952). Genes of interest were first mapped thanks to pairs of near isogenic lines differing only in the region of the interesting gene (Philouze, 1991; Laterrot, 1996). Nevertheless, until the 1980s, the location of mutations of interest on genetic maps was not precise. The first isozyme markers were limited in number and rapidly replaced by restriction fragment length polymorphism (RFLP) markers. The first high-density genetic map based on RFLP markers was constructed (Tanksley et al., 1992). With more than 1000 loci, spread on the 12 chromosomes, it allowed the localization of several mutations and genes of interest. Then,

PCR based markers, including RAPD, AFLP and microsatellites, were used, but remained limited in polymorphism level and distribution across the genome. Following the identification of PCR markers linked to the gene of interest, specific PCR markers were set up, simplifying the genotyping step for breeders. Nevertheless, PCR markers such as RAPD or AFLP map in majority close to the centromeres, reducing their potential efficiency for gene mapping in tomato (Grandillo and Tanksley, 1996a; Haanstra et al., 1999; Saliba-Colombani et al., 2001).

3.2.2 Trait mapping

The construction of genetic maps of molecular markers permitted the dissection of quantitative traits into QTL (Quantitative Trait Loci) (Paterson et al., 1988; Tanksley et al., 1992). This strategy also opened the way to investigate physical mapping and molecular cloning of genetic factors underlying quantitative traits (Paterson et al., 1991). The first gene cloned by positional cloning was the *Pto* gene, confering resistance to *Pseudomonas syringae* (Martin *et al.* 1993). Since then, several interspecific progenies with each wild relative species were studied. Due to the low genetic diversity within the cultivated compartment (Miller and Tanksley 1990), most of the mapping populations were based on interspecific crosses between a cultivar and a related wild species from the lycopersicon group (as reviewed by (Foolad, 2007; Labate et al., 2007; Grandillo et al., 2011) or from lycopersicoides (Pertuzé et al., 2003) and juglandifolia group (Albrecht et al., 2010). However, maps based on intraspecific crosses have proved their interest notably for fruit quality aspects (Saliba-Colombani et al., 2001). All those populations allowed the discovery and characterization of a myriad of major genes (Rothan et al., 2019) and QTLs involved in various traits (Grandillo and Tanksley, 1996b; Tanksley et al., 1996; Fulton et al., 1997; Bernacchi et al., 1998; Chen et al., 1999; Grandillo et al., 1999; Frary et al., 2000; Monforte and Tanksley, 2000; Causse et al., 2001; Saliba-Colombani et al., 2001; Causse et al., 2002; Doganlar et al., 2003; Frary et al., 2004; Schauer et al., 2006; Baldet et al., 2007; Jiménez-Gómez et al., 2007; Cagas et al., 2008; Kazmi et al., 2012; Haggard et al., 2013; Alseekh et al., 2015; Pascual et al., 2015; Ballester et al., 2016; Rambla et al., 2016; Kimbara et al., 2018).

The main results of QTL studies can be summarised:

- QTLs are detected in every case, sometimes with strong effects. A few QTLs explaining a large part of the phenotypic variation, acting together with minor QTLs, are frequently detected. Most of the QTLs act in an additive manner, but a few dominant and even over-dominant QTLs were detected (Paterson et al., 1988; DeVicente and Tanksley, 1993).
- QTLs can be separated in two types: QTLs stable over the environments, years or types of progeny, and QTLs more specific of one condition (Paterson et al., 1991).
- Some regions involved in the variation of a trait are found in progenies derived from different accessions of a species, or from different species (Fulton et al., 1997; Bernacchi et al., 1998; Chen et al., 1999; Grandillo et al., 1999; Fulton, 2002).
- The dissection of complex traits in relevant components and the QTL mapping of these components allowed the genetic bases of the variability of complex traits to be understood. For example, a map of QTLs controlling several attributes of organoleptic quality in fresh-market tomato revealed relations between QTLs for sensory attributes and chemical components of the fruit (Causse et al., 2002). The analysis of biochemical composition of a trait is also important.
- Fine mapping experiments allowed to precisely map the QTLs in a chromosome region and to verify the existence of several QTLs linked in the same region (Paterson et al., 1990; Frary et al., 2003; Lecomte et al., 2004a). For example, by reducing the size of an introgressed fragments from *S. pennellii*, (Eshed and Zamir, 1995) identified three linked QTLs controlling fruit weight on a single chromosome arm. Fine mapping is also an important step for cloning QTLs, as first shown by the successes in cloning QTLs controlling fruit weight (Alpert and Tanksley, 1996; Frary et al., 2000), fruit shape (Tanksley, 2004) and soluble solid content (Fridman et al., 2000; Fridman et al., 2004).
- Wild species, in spite of their low characteristics in comparison to cultivars, can carry alleles, which may contribute to the improvement of most of the agronomic traits (DeVicente and Tanksley, 1993).

3.2.3 Specific populations to dissect phenotypes

Rapidly, molecular breeding strategies were set up and implemented to try to "pyramid" genes and QTL of interest for agronomical traits, notably using Advanced Backcross QTL method (AB-QTL) (Grandillo and Tanksley, 1996b). Using this approach with a *S. lycopersicum x S. pimpinellifolium* progeny, in which agronomical favorable QTL alleles were detected, Grandillo and colleagues showed how a wild species could contribute to improve cultivated tomato (Grandillo et al., 1996). Introgression Lines (IL) derived from interspecific crosses allowed to dissect the effect of chromosome fragments from a donor (usually from a wild relative) introgressed into a recurrent elite line. IL offers the possibility to evaluate the agronomic performance of a specific set of QTL (Paran et al., 1995). IL was used as a base for fine mapping and positional cloning of several genes and QTL of interest. The first IL library was developed between *S. pennellii* and *S. lycopersicum* (Eshed and Zamir, 1995; Zamir, 2001). QTL mapping power was increased compared to biallelic QTL mapping population, and was again improved by the constitution of sub-IL set with smaller introgressed fragments. This

progeny was successful in identifying QTLs for fruit traits (Causse et al., 2004); anti-oxidants (Rousseaux et al., 2005), vitamin C (Stevens et al., 2007) and volatile aromas (Tadmor et al., 2002). The introgression of a QTL identified in these IL has allowed plant breeders to boost the level of soluble solids (brix) in commercial varieties and largely increased tomato yield in California (Fridman et al., 2004). Complementary genetic resources are now available, including a new backcrossed inbred line (BIL) population generated by repeated backcrosses, followed by selfing (Ofner et al., 2016). This BIL population could be used in combination with ILs for fine-mapping QTLs previously identified and to pinpoint strong candidate genes (Fulop et al., 2016) . Moreover, the *S. pennellii* ILs have been broken into additional sub-lines carrying molecular marker-defined introgressions that are smaller than those carried by the original ILs, further facilitating the identification of candidate genes (Alseekh et al., 2013). These sub-isogenic lines are available to the scientific community and have been used to map loci affecting fruit chemical composition (Alseekh et al., 2015; Liu et al., 2016). Such exotic libraries were also designed with other species, involving *S. pimpinellifolium* (Doganlar et al., 2003), *S. habrochaites* (Monforte and Tanksley, 2000; Finkers et al., 2007) and *S. lycopersicoides* (Canady et al., 2005).
Introgression lines were also used to dissect the genetic basis of heterosis (Eshed and Zamir, 1995). Heterosis refers to phenomenon where hybrids between distant varieties or crosses between related species exhibit greater biomass, speed of development, and fertility than both parents (Birchler et al., 2010). Heterosis involves genome–wide dominance complementation and inheritance model such as locus–specific overdominance (Lippman and Zamir, 2007). Heterotic QTL for several traits were identified in tomato IL (Semel et al., 2006). A unique QTL was shown to display at the heterozygous level improved harvest index, earliness and metabolite content (sugars and amino acids) in processing tomatoes (Gur et al., 2010; 2011). Furthermore, a natural mutation in the SFT gene, involved in flowering (Shalit et al., 2009), was shown to correspond to a single overdominant gene increasing yield in hybrids of processing tomato (Krieger et al., 2010).

3.2.4 Genes and QTLs controlling tomato disease resistance

The excessive use of chemical fungicides and pesticides was for long time most common in tomato crops. Because of environmental, consumer and grower constraints, their elevated costs, and their limited effectiveness, other levers, such as genetic resistance and various cultural practices, have to be integrated for achieving sustainable agriculture (Lefebvre et al. 2018). However, the development of new cultivars with enhanced resistance or tolerance was often hindered by the lack of genetic diversity within the cultivated *S. lycopersicum* germplasm, because of its narrow genetic diversity due its domestication history. Screening the tomato-related wild species germplasm collections enabled to discover many sources of disease resistance traits during the last 80 years (Rick and Chetelat 1995). About 40 major resistance traits were discovered in wild tomato species. Those genes confer resistance to diseases of different pest and pathogen classes. Of the 40 major resistance traits, about 20 have been introgressed into cultivated tomato (Ercolano et al. 2012). *S. peruvianum*, *S. habrochaites*, *S. pimpinellifolium* and *S. chilense* have proved to be the richest sources of resistance genes (Laterrot 2000). The systematic screening of tomato germplasm for disease resistance will probably permit to discover further novel resistance sources and consequently novel resistance loci (major resistance genes and resistance QTLs).

3.2.4.1 Resistance gene and QTL discovery

More than 100 loci underlying the 30 major tomato resistance diseases have been genetically mapped (Foolad and Panthe, 2012 for review). Molecular markers associated with many resistance genes or QTLs have been reported. Up to now, 26 major resistance genes were isolated (*Asc-1*, *Bs-4*, *Cf-2*, *Cf-4*, *Cf-5*, *Cf-9*, *Hero*, *I* (*=I-1*), *I-2*, *I-3*, *I-7*, *Mi-1.2* (*=Mi=Meu*), *ol-2*, *Ph-3*, *pot-1*, *Prf*, *Pto*, *Tm-1*, *Tm-2*, *Tm-2*^{2} (*=Tm-2.2=Tm-2*^{a}), *Ty-1*, *Ty-2*, *Ty-3*, *ty-5*, *Ve-1 (=Ve)*, *Sw-5*) (**Table 3**). Resistance tomato locus has a well-defined nomenclature; written in italic, they are abbreviated by 1 to 3 letters (the first letter in uppercase for dominant resistance alleles and in lowercase for recessive dominant alleles) and separated of a number by a dash, the number indicating the order of discovery of the gene for the target disease. In a few cases, the last figure is followed by a dot and another number indicating different alleles; alleles could also be indicated by a number or a letter in superscript.
Most of reported major effect resistance genes are dominant, except *pot-1*, *ty-5* and *ol-2* conferring resistance to potyviruses (PVY and TEV), *Tomato yellow leaf curl virus* (TYLCV) and to *Oidium neolycoersici*, respectively, that were both cloned (Bai et al. 2008; Lapidot et al. 2015; Ruffel et al. 2005). Another recessive resistance allele *py-1* (also named *pyl*) controlling *Pyrenochaeta lycopersici* responsible for corky root rot was reported but is not cloned yet (Doganlar et al. 1998).
For a few tomato diseases, both major effect resistance genes and resistance QTLs have been identified according to the resistance genitor and the pathogen variant used in the analysis and to environmental conditions. Otherwise, a single major resistance gene was discovered for most tomato diseases. For a few diseases, several major resistance genes have been reported, such as for TSWV, where 6 dominant resistance genes and 3 recessive resistance genes were described (Foolad and Panthee 2012) and for *Meloidogyne* nematodes where several resistance genes have been identified. However generally a single of those genes, such as *Sw-5* and *Mi-1.2*, is currently used in MAS because it confers a broader spectrum resistance than others.

Table 3: Pest and pathogen resistance genes of tomato molecularly characterized. Genes are classified by pest and pathogen latin name inside each pest and pathogen class. For each gene, the ITAG gene model(s) and the Genebank accession number are given when available.

Locus name	Function of cloned gene	Species from which the trait was discovered	Genetic ressources carrying this gene	Tomato chromosome	ITAG gene model	Genebank accession number	Literature
Asc (Asc-1)	LAG1 Longevity Assurance Gene Family	*S. pennellii*	VFNT Cherry, LA716	T3	Solyc03g114600	AJ312131	Brandwagt et al. (2000)
Cf-2	Leucine-rich repeat receptor-like protein kinase LRR- RLP	*S. pimpinellifolium*	LA2244, LA3043	T6	Solyc06g008300	U42444	Dixon et al. (1996)
Cf-4	Leucine-rich repeat receptor-like protein kinase LRR- RLP	*S. habrochaites*	LA2446, LA3045, LA3051, LA3267	T1	Solyc01g006550	AJ002235	Takken et al. 1998, 1999
Cf-5	Leucine-rich repeat receptor-like protein kinase LRR- RLP	*S. lycopersicum*	-	T6	-	AF053993	Dixon et al. (1998)
Cf-9	Leucine-rich repeat receptor-like protein kinase LRR- RLP	*S. pimpinellifolium*	LA3047	T1	Solyc01g005160	AJ002236	Jones et al. (1994)
I (I-1)	Leucine-rich repeat receptor-like protein kinase LRR- RLP	*S. pimpinellifolium*	PI79532	T11	Solyc11g011180		Catanzariti et al. (2017)
I-2	CC-NB-LRR	*S. pimpinellifolium*	PI126915	T11	Solyc11g071430		Ori et al. (1997); Simons et al (1998)
I-3	S-receptor-like kinase 5 (SRLK-5)	*S. pennellii*	LA716	T7	Solyc07g055640	KP082943	Catanzariti et al. (2015)
I-7	Leucine-rich repeat receptor-like protein kinase LRR- RLP	*S. pennellii*	PI414773, Tristar cultivar	T8	Solyc08g77740	KT185194	Gonzalez-Cendales et al. (2016)
ol-2 (SlMlo1)	Loss-of-function mlo	*S. lycopersicum*	LA1230, KNU-12 cultivar	T4	Solyc04g049090	AY967408	Bai et al. (2008)
Ve-1 (Ve)	RLP-type resistance protein	*S. lycopersicum*	VFN8, Craigella GCR 151, PI 303801	T9	Solyc09g005090	AF272367	Kawchuk et al. (2001); Fradin et al (2009)
Ph-3	CC-NB-LRR	*S. pimpinellifolium*	LA4285, LA4286, LA1269(=PI365957), L3708	T9	near Solyc09g092280-Solyc09g092310	KJ563933	Zhang et al. (2013); Zhang et al. (2014)
pot-1	eukaryotic translation initiation factor 4E (eIF4E)	*S. habrochaites*	PI247087	T3	Solyc03g005870	AY723736	Ruffel et al (2005); Piron et al. (2010)
Tm-1	Inhibitor of tobamovirus RNA replication	*S. habrochaites*	PI126445	T2	Solyc02g062560	AB713135, AB713134	Ishibashi et al. (2007)
Tm-2	CC-NB-LRR	*S. peruvianum*	Craigella GCR236	T9	Solyc09g018220	AF536200	Lanfermeijer et al. (2003)
Tm-2^2 (Tm-2^a)	CC-NB-LRR	*S. peruvianum*	Craigella GCR267	T9	Solyc09g018220	AF536201	Lanfermeijer et al. (2005)
Sw-5	CC-NB-LRR	*S. peruvianum*	PI128654 / Stevens cultivar	T9	Solyc09g098130	AY007367	Brommonschenkel et al. (2000)
Ty-1	DFDGD-Class RNA–Dependent RNA Polymerases	*S. chilense*	LA1969	T6	Solyc06g051170, Solyc06g051180, and Solyc06g051190		Verlaan et al. (2013)
Ty-2 (TYNBS1)	CC-NB-LRR	*S. habrochaites*	H9205, TY-Chie, Shurei cultivars	T11	near Solyc11g069660.1 and Solyc11g069670.1	LC126696	Yamaguchi et al., 2018
Ty-3	DFDGD-Class RNA–Dependent RNA	*S. chilense*	LA2279	T6	Solyc06g051170,		Verlaan et al. (2013)

	Polymerases				Solyc06g051180, and Solyc06g051190		
ty-5	messenger RNA surveillance factor Pelota (Pelo)	*S. peruvianum*	Tyking cultivar TY172	T4	Solyc04g009810	KC447287	Lapidot et al., 2015
Pto	Serine/threonine protein kinase	*S. pimpinellifolium*	LA2396, LA2458, LA3472	T5	Solyc05g013300	U02271	Martin et al. (1993)
Prf	CC-NB-LRR	*S. pimpinellifolium*	LA2396, LA2458, LA3472	T5	Solyc05g013280	U65391	Salmeron et al. (1996)
Bs-4	TIR-NB-LRR	*S. lycopersicum*	Money Maker cultivar	T5	Solyc05g007850	AY438027	Schornack et al. (2004)
Hero	CC-NB-LRR	*S. pimpinellifolium*	LA121	T4	Solyc04g008120	AJ457051	Ernst et al. (2002)
Mi-1.2 (Mi, Meu)	CC-NB-LRR	*S. peruvianum*	Motelle cultivar and most of tomato rootstocks	T6	Several homologs on Chr6	AF039682	Vos et al. (1998); Milligan et al., (1998); Nombela et al., 2001; Rossi et al., 1998; Casteel et al., 2007

A few cloned genes correspond to allelic series such as *Ty-1* and *Ty-3* on chromosome T6 (Verlaan et al. 2013), or *Tm-2* and *Tm-2²* on chromosome T9 (Lanfermeijer et al. 2005), to very tightly linked genes such as *Pto* and *Prf* on chromosome T5 both involved in recognition of *Pseudomonas syringae* pv. *tomato* (Salmeron et al. 1996), or else they belong to clusters of major resistance genes such as *Cf-4* and *Cf-9* on chromosome T1 (Takken et al. 1999) or *Cf-2* and *Cf-5* on chromosome T6 (Dixon et al. 1998). Additionally, while resistance genes are often specific to a pest, a pathogen or a variant of a species, in rare cases a same gene can confer resistance to different distantly related pests, such as *Mi-1.2* called also *Meu* that triggers the resistance to root knot nematodes caused by three *Meloigogyne* species (*M. incognita, M. arenaria, M. javanica*), to the aphid *Macrosiphum euphorbiae*, to the whitefly *Bemisia tabaci*, and to the psyllid *Bactericerca cockerelli* (Casteel et al. 2007; Milligan et al. 1998; Nombela et al. 2003; Rossi et al. 1998; Vos et al. 1998).
For many diseases, no major gene has been found yet, or major genes previously discovered were breakdown by virulent pathogen variants. For this reason, several research groups are now willing to focus on quantitative resistance that have the particularity to reduce the development of pests and pathogens rather than to block them totally. Quantitative resistance, also called partial resistance and generally controlled by QTLs, provides in most of the cases a more durable and broad-spectrum resistance (Cowger and Brown 2019); in addition, resistance QTLs are more frequent than major resistance genes in natural genetic resources. Many resistance QTLs have been mapped in the tomato genome, particularly for resistance traits to *P. infestans* (Arafa et al. 2017; Brouwer et al. 2004; Brouwer and St Clair 2004; Foolad et al. 2008; Ohlson et al. 2018; Ohlson and Foolad 2016; Panthee et al. 2017; Smart et al. 2007), *O. lycopersici* (Bai et al. 2003), *Alternaria solani* (Foolad et al. 2002), *Alternaria alternata* (Robert et al. 2001), *Xanthomonas* sp. (Hutton et al. 2010; Sim et al. 2015), *C. michiganensis* (Coaker and Francis 2004; Kabelka et al. 2002), *Ralstonia solanacearum* (Carmeille et al. 2006; Mangin et al. 1999; Wang et al. 2013), *Botrytis cinerea* (Davis et al. 2009; Finkers et al. 2008; Finkers et al. 2007) and *Cucumber mosaic virus* (CMV) (Stamova and Chetelat 2000).
Mainly, 3 genes were described for controlling resistance to late blight, but *Ph-1* is not effective anymore, due to the emergence of evolved races of *P. infestans*, and *Ph-2* and *Ph-3* have both an incomplete penetrance and evolved races of *P. infestans* have been described on plant material carrying those genes. Due to the breakdown of those 3 major resistance genes controlling late blight, many efforts are now underway to identify new resistance sources in tomato relatives and within the cultivated tomato germplasm (Caromel et al. 2015 and work in progress at INRA GAFL; Foolad et al. 2014).
An approach to breed for resistance when there is no natural variants, without transformation with foreign DNA, consists to inactivate by TILLING plant dominant susceptibility genes that permit the pathogen to multiply. A proof of concept of such an approach has allowed the de novo creation of resistance to two potyvirus species in tomato (Piron et al. 2010). Similarly, EcoTILLING allows the detection of natural variability of the allelic variants of a specific gene, an approach that has resulted in the detection in tomato diversity of a new Sw-5 variant controlling TSWV (Belfanti et al. 2015).

3.2.4.2 Resistance gene and QTL architecture

Mapping of resistance loci in the tomato genome highlights several hotspots of resistance genes even if the 12 tomato chromosomes harbor resistance loci **(Figure 3)**. Equally, mapping of the repertoire of major resistance genes evidenced that they are organized in tandem or in clusters (Foolad 2007). It appears that a lot of resistance loci were identified on chromosomes 6 and 9, from a same genitor or from the tomato wild relatives. The chromosome 6 carries major resistance genes to root knot *Meloidogyne* (*Mi-1.2*), *O. neolycopersici* (*Ol-1*, *Ol-3*, *Ol-4*, *Ol-5* and *Ol-6*), *Cladosporium fulvum* (*Cf-2* and *Cf-5*), TYLCV (*Ty-1* and *Ty-3*), *Alfalfa mosaic virus* (*Am*), and resistance QTLs to *Ralstonia solanacearum* and ToMoV (*Tomato mottle virus*) (Agrama and Scott 2006). Identically the chromosome 9 is rich in resistance gene clusters with *Tm-2* and *Tm-2²* controlling the *Tomato mosaic virus* (ToMV) (Pillen et al. 1996) and *Frl* controlling FORL (Vakalounakis et al. 1997) near the centromer, *Sw-5* controlling TSWV (Stevens et al. 1995) and *Ph-3* controlling *P. infestans* (Chunwongse et al. 2002) near a telomere and *Ve* controlling *Verticillium dahliae* near the other telomere (Kawchuk et al. 2001).

3.2.4.3 Molecular basis of resistance genes and QTLs

Many resistance traits in tomato are conferred by single dominant genes, encoding proteins that recognize directly or indirectly avirulent proteins of pests and pathogens and trigger the plant defense response. A few correspond to single recessive genes (*e.g. pot-1*, *ol-2*, generally written with lowercase letters). Recessive resistance alleles are due to loss-of-function or absence of susceptibility that hamper the pathogen's development in the plant; conversely the corresponding susceptible alleles facilitate the development of the pathogen that benefits of the host's machinery. Many of major resistance genes have been cloned by forward genetics and map-based cloning approaches (see section 3.6 below) and most of the dominant cloned genes encode conserved NB-LRR proteins. The conserved molecular structure of resistance genes (NB-LRR R-genes, RLP, RLK...) was used to search for genes homologous to genes already isolated in the same species or in related species, and to discover and isolate new resistance alleles or genes (*e.g. Sw-5* and *Mi* that are homolog, the *Cf* serie genes). More recently, the RenSeq technology, using baits designed from 260 NBS-LRR genes previously identified in Solanaceae, helped to pick-up 105 novel NBS-LRR sequences within the reference genome of tomato (*S.*

lycopersicum) Heinz1706 and 355 novel NBS-LRR novel within the draft of *S. pimpinellifolium* LA1589 genome, to complete the repertoire of genes that encode NB-LRR R-genes in these species (Andolfo et al. 2014). Beside those major effect resistance genes, many genes activated during the tomato disease defense response were also characterized. Several are specific of a plant-pathogen interaction. A few are involved in several plant-pathogen interactions, such as the lipase-like protein EDS1 that is involved in defense mechanisms triggered by Cf-4 and Ve proteins. Equally Prf, I-2 and Bs-3 proteins interact with the RAR1, SGT1 and HSP90 proteins. Beside, transcriptional analysis highlighted several genes involved in Jasmonate Acid or Salicilic Acid signaling pathway regulation. A few of these genes could correspond to resistance QTLs.
Until now, no QTL determining disease resistance has been cloned in tomato. Quantitative plant resistance loci may correspond to a large array of molecular mechanisms that play a role in partial resistance, they may be genes involved in PAMP recognition responsible for basal defense, genes involved in defense signal transduction, genes regulating the phytoalexin synthesis, weak effect alleles of R genes, genes regulating developmental phenotypes, or other genes not yet identified (Poland et al. 2009).

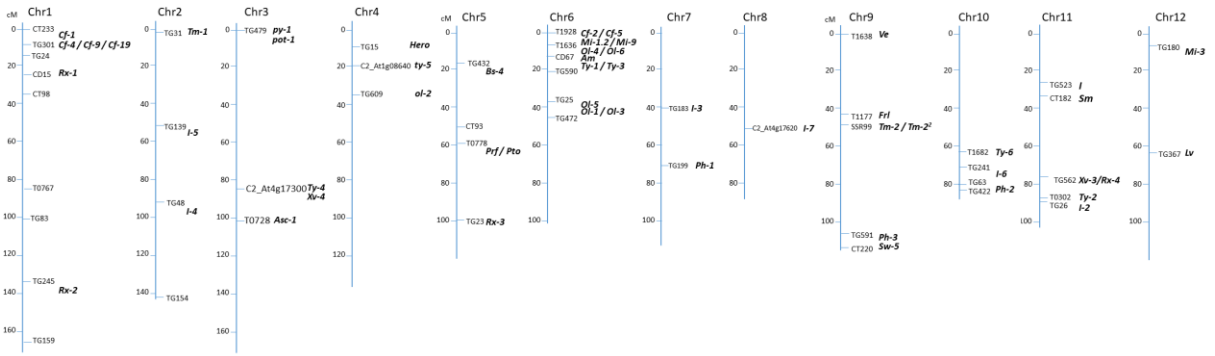


Figure 3: Genetic map of tomato with mapped major resistance genes. Marker names and genetic distances are according to the SGN tomato- EXPEN 2000 map (https://solgenomics.net/). Position of genes are adapted from Foolad 2007, Foolad et al. 2012, Lee et al. 2015, Bai et al. 2018, Gill et al. 2019, and Sharma et al. 2019. When there is no common marker between the publication and the EXPEN 2000 map, the relative position was determined using a blastn search with the linked marker sequences as a query, against tomato chromosomes SL2.50 to identify the nearest marker. Genetic distances (in cM) are indicated on the left of the chromosomes.

3.3 Genomic resources

3.3.1 The reference genome sequence

Genomic information greatly promoted our understanding of the genetic architecture and evolutionary history of modern tomato. The tomato genome sequencing project was initiated as part of the International Solanaceae Project (SOL), which was launched on November 3, 2003 at Washington, USA and gathered a consortium of scientists of 10 countries including China, France, Spain, Italy, USA, UK, the Netherlands, Japan, Korea and India (Mueller *et al.*, 2005). The main reason why tomato was first chosen as the reference genome for the Solanaceae was due to its high level of macro and micro-synteny among over 3000 species. This project was first started with conventional sequencing technologies, such as Sanger sequencing. In order to reduce the cost of producing a high-quality reference, BAC-by-BAC sequencing strategy based on saturated genetic markers was used to select seed BACs within the gene-rich part of the tomato genome for sequencing. However, this process was quite slow and became a serious obstacle, which was greatly accelerated by next-generation sequencing (Pietrella and Giuliano, 2016).
The first tomato genome sequence was published in 2012 for the inbred tomato cultivar ‘Heinz 1706’ (*S. lycopersicum*) together with a draft of its closest wild species *S. pimpinellifolium* (accession LA1589) (The Tomato Genome Consortium, 2012). In the tomato genome, recombination, genes and transcripts are substantially located in the euchromatin regions compared to the heterochromatin regions, whereas chloroplast insertions and conserved microRNA genes were more evenly distributed throughout the genome (The Tomato Genome Consortium, 2012). The tomato genome was highly syntenic with other Solanaceae species, such as pepper, eggplant, potato and *Nicotiana*. Tomato had fewer high-copy, full-length long terminal repeat retrotransposons with older insertion ages compared to *Arabidopsis* and Sorghum. Genome annotation showed that there were a total 34,727 protein-coding genes and 30,855 of them were supported by RNA sequencing data. Chromosomal organization of genes, transcripts, repeats and sRNAs were very similar between tomato and potato. Among all the protein-coding genes, 8615 genes were common to tomato, potato, *Arabidopsis*, rice and grape. A total of 96 conserved sRNAs were predicted in tomato, which could be further divided into 34 families, 10 of which being highly conserved in plants. The potato genome showed more than 8% divergence from tomato, with nine large and several smaller inversions (The Tomato Genome Consortium, 2012). The *Solanum* lineage has experienced one ancient and one more recent consecutive genome triplications. The genome information provides a basic understanding of the genetic bottlenecks that narrowed tomato genetic diversity (The Tomato Genome Consortium, 2012).

Since the first published version, the sequence has been completed, corrected and re-annotated using new sequence data and new RNAseq data and the genome version today is SL3.0 while the annotation is ITAG3.2.

3.3.2 Resequencing tomato accessions

Next generation sequencing technologies made it possible to sequence genomes at large scales (Goodwin et al., 2016). Soon after the availability of the reference tomato genome, the genome of the stress-tolerant wild tomato species *S. pennellii* was published (Bolger et al., 2014). This species is characterized by extreme drought tolerance and unusual morphology. Many stress-related candidate genes were mapped in this wild species. Large gene expression differences were observed between *S. lycopersicum* cv. M82 and *S. pennellii* (LA716) due to polymorphisms at the promoter and/or coding sequence levels. This wild species and others were further re-sequenced and assembled using long read sequencing platforms complemented with Illumina sequencing (Usadel et al., 2017;). After the genome of *S. pennellii*, a panel of diversified tomato accessions and related wild species were sequenced (The 100 Tomato Genome Sequencing Consortium, 2014)(The 100 Tomato Genome Sequencing Consortium, 2014) . The allogamous self-incompatible wild species have the highest level of heterozygosity, which was low for the autogamous self-compatible species (The 100 Tomato Genome Sequencing Consortium, 2014). Almost at the same time, a comprehensive genomic analysis based on resequencing 360 tomato accessions elucidated the history of tomato breeding (Lin et al., 2014). This study showed that domestication and improvement of tomato mainly involved two independent sets of QTLs leading to fruit size increase. Five major QTLs (*fw1.1*, *fw5.2*, *fw7.2*, *fw12.1* and *lcn12.1*) contributed to the enlargement of tomato fruit during domestication process. Then, up to 13 major QTLs (*fw1.1*, *fw2.1*, *fw2.2*, *fw2.3*, *lcn2.1*, *lcn2.2*, *fw3.2*, *fw3.2*, *fw5.2*, *fw7.2*, *fw9.1*, *fw10.1*, *fw11.1*, *fw12.1*, *fw11.3*, *fw12.1* and *lcn12.1*) contributed to the second improvement of tomato fruit. This study also detected several independent mutations in a major gene *SlMYB12* that changed modern red tomato to pink tomato appreciated in Asia. This study also illustrated the linkage drag associated with wild introgressions (Lin et al., 2014).

Since then, low-depth resequencing or genotyping-by-sequencing has become a common practice and is widely applied in many tomato collections. Up to now, around 900 tomato accessions have been re-sequenced, with the sequence depth ranging from low to high (The Tomato Genome Consortium, 2012; Causse et al., 2013; Bolger et al., 2014; Lin et al., 2014; The 100 Tomato Genome Sequencing Consortium, 2014; Tieman et al., 2017; Ye et al., 2017; Tranchida-Lombardo et al., 2018). These genomic resources are freely available (https://solgenomics.net) and will greatly facilitate modern breeding of new climate smart tomato cultivars.

In a recent pan-genome study of 725 phylogenetically and geographically representative tomato accessions, a total of 4,873 genes were newly discovered compared to the reference genome (Gao et al., 2019). Among these, 272 were potential contaminations and were removed from the 'Heinz 1706' reference genome. Substantial gene loss and intensive negative selection of genes and promoters were detected during tomato domestication and improvement. During tomato domestication, a total of 120 favorable and 1213 unfavorable genes were identified, whereas 12 favorable and 665 unfavorable genes were identified during improvement process.

Disease resistance genes were especially lost or negatively selected. Gene enrichment indicated that defense response was the most enriched group of unfavorable genes during both domestication and improvement. No significantly enriched gene families were found in favorable genes during improvement. A rare allele in the *TomLoxC* promoter was found under selected during domestication. In orange-stage fruit, accessions with both the rare and common *TomLoxC* alleles have high expression compared to those homozygous in modern tomatoes. Taken together with other findings, this pan-genome study provides useful knowledge for further biological discovery and breeding (Gao et al., 2019).

3.4 SNP markers

3.4.1 SNP discovery

Single nucleotide polymorphisms (SNPs) are the most abundant molecular markers for major crops. SNPs can be detected in any region of the genome, including coding sequences or non-coding sequences of genes, as well as the intergenic regions. Only the non synonymous SNPs in the coding regions of genes change the amino acid sequences of proteins. However, SNPs in the non-coding region are also likely to affect gene expression through different mechanisms (Farashi et al., 2019). Millions of SNPs can be directly generated via genotyping-by-sequencing (GBS) or resequencing of a few lines (Catchen et al., 2011). Next-generation sequencing-based technologies have also accelerated the identification and isolation of genes associated with agronomic traits in major crops (Nguyen et al., 2018). There are many GBS methods available, including at least 13 reduced-representation sequencing (RRS) approaches and at least four whole-genome resequencing (WGR) approaches (Scheben et al., 2017). Among them, RNA sequencing and exome sequencing based on transcriptome sequences is an important alternative RRS approach (Haseneyer et al., 2011; Scheben et al., 2017). The sequenced data can be used for expression analysis and also does not require prior genomic sequence information (Wang et al., 2010).

Since the availability of the reference tomato genome, whole-genome resequencing of different tomato accessions could directly generate millions of SNPs, covering the whole tomato genome (Bolger et al., 2014; Lin et al., 2014; Menda et al., 2014; The 100 Tomato Genome Sequencing Consortium, 2014; Tieman et al., 2017;

Ye et al., 2017; Zhu et al., 2018). The number of SNPs in the wild tomato species exceeds 10 million, which are 20-folds higher than that in most of the domesticated accessions (The 100 Tomato Genome Sequencing Consortium, 2014). Once the reference genome was available, it became possible to only sequence chromosome regions of interest to screen for SNP. For example, Ranc et al., (2012) sequenced 81 DNA fragments covering the chromosome 2 at different mapping densities in a core collection of 90 tomato accessions and discovered 352 SNPs.

3.4.2 SNP arrays

SNP arrays is another popular and cost-effective genotyping approach, such as the Solanaceae Coordinated Agricultural Project (SolCAP) (Hamilton et al., 2012; Sim et al., 2012b), the Centre of Biosystems Genomics (CBSG) consortium (Víquez-Zamora et al., 2013) or, the Diversity Arrays Technology (DArTseq) (Pailles et al., 2017). However, RNA-seq based SNP arrays, such as SolCAP and ddRAD-Seq (Arafa et al., 2017), have some major limitations: Gene expression is dependent on tissue and time, multiple biases are introduced by library preparation during RNA fragmentation (Wang et al., 2009) and SNP coverage is low in coding regions (Scheben et al., 2017). In tomato, these SNP arrays have been widely used to genotype different tomato collections (Sim et al., 2012a; Víquez-Zamora et al., 2013; Ruggieri et al., 2014; Sauvage et al., 2014; Blanca et al., 2015; Bauchet et al., 2017a; Bauchet et al., 2017b; Pailles et al., 2017; Albert et al., 2016b).

3.4.3 Genotype imputation

When a large diverse reference panel is available, SNP density can be significantly increased by genotype imputation (Guan and Stephens, 2008; Halperin and Stephan, 2009; Iwata and Jannink, 2010; Marchini and Howie, 2010; Pasaniuc et al., 2012; Browning and Browning, 2016; Das et al., 2016; Wang et al., 2018). In human and model plant species, there are some very good reference panels suitable for genotype imputation, such as the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) and the UK10K Project in humans (Danecek et al., 2015; The UK10K Consortium, 2015), the 3000 Rice Genome Project (The 3000 rice genomes project, 2014; McCouch et al., 2016) and the 1001 Genomes Consortium in *Arabidopsis thaliana* (The 1001 Genomes Consortium, 2016). The marker density of SNP arrays in tomato is quite low and many genomic gaps remain, compared with the whole-genome sequencing (Sauvage et al., 2014; Bauchet et al., 2017b; Zhao et al., 2019). After imputation, the SNP number can be increased up to 30-folds and greatly bridged the genomic gaps and genomic coverage (**Figure 4**) (Zhao et al., 2019).

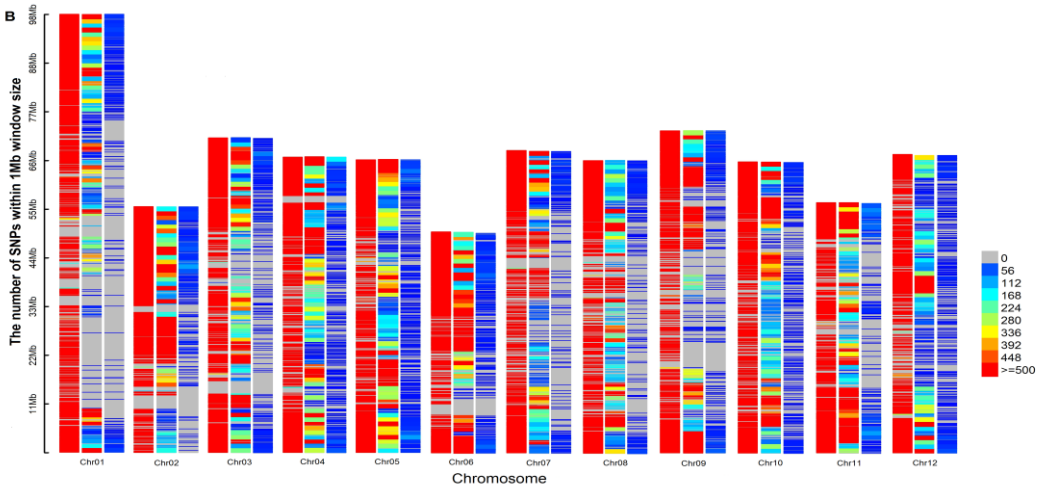


Figure 4. SNP density for the tomato collection reported in Sauvage et al., (2014). Left, middle and right panels represent the SNP density of the reference panel, after and before genotype imputation, adapted from Zhao et al., (2019).

3.5 Diversity analyses

Molecular genetic markers play an important role in the modern breeding (Ramstein et al., 2018). They also provide a new vision of tomato genetic diversity (Bauchet and Causse, 2012). Overall, modern cultivated tomato accessions present a lower polymorphism level compared to wild species, as shown by different types of markers, such as RFLP (Miller and Tanksley, 1990), AFLP (Suliman-Pollatschek et al., 2002; Park et al., 2004; Van Berloo et al., 2008; Zuriaga et al., 2009), RAPD (Grandillo and Tanksley, 1996a; Archak et al., 2002; Tam et al., 2005; Carelli et al., 2006; El-hady et al., 2010; Meng et al., 2010; Length, 2011), SSR (Suliman-Pollatschek et al., 2002; Jatoi et al., 2008; Mazzucato et al., 2008; Albrecht et al., 2010; Meng et al., 2010; Sim et al., 2010; Zhou et al., 2015), ISSR (Vargas-Ponce et al., 2011; Shahlaei et al., 2014) and SNPs (Blanca et al., 2012; Sim et al., 2012a; Lin et al., 2014; The 100 Tomato Genome Sequencing Consortium, 2014).

Whole genome sequencing technology made it possible to detect millions of SNPs and it has revealed that the number of SNPs in wild species is over 10 million and is 20-fold higher than that for most domesticated tomato accessions (The 100 Tomato Genome Sequencing Consortium, 2014), which provides clues on the genetic diversity loss during tomato domestication and improvement. A study based on whole-genome sequencing of wild and cultivated tomato species demonstrated that approximately 1% of the tomato genome has experienced a very strong purifying selection during domestication (Sahu and Chattopadhyay, 2017). At the expression level, domestication has affected up to 1729 differentially expressed genes between modern tomato varieties and the *S. pimpinellifolium* wild species and also affected about 17 gene clusters. Some gene regulation pathways were significantly enriched, such as carbohydrate metabolism and epigenetic regulations (Sauvage et al., 2017).
Cherry tomato accessions (*S. lycopersicum* var. *cerasiforme*) are intermediate between cultivated and wild species with a moderate genetic diversity (Ranc et al., 2012; Xu et al., 2013; Zhang et al., 2016). The linkage disequilibrium of cherry tomatoes is also intermediate between that of cultivated and wild species (Sauvage et al., 2014; Bauchet et al., 2017a). They could thus be helpful to bridge the gaps between low genetic diversity and high morphological diversity of modern cultivated tomato accessions and wild species which may provide interesting genes but also a strong genetic load. Molecular markers could also link the genetic and morphological diversities together and provide insight into the origin of tomato. By phenotyping 272 genetically and morphologically diverse tomato accessions with the SOLCAP genotyping SNP array, Blanca et al., (2012) revealed that cherry tomato accessions were morphologically and genetically intermediate between modern cultivated tomato accessions (*S. lycopersicum*) and wild accessions (*S. pimpinellifolium*). In addition, cherry and wild tomato accessions inhabited strikingly different ecological and climatic regions and a clear relationship was found between the population structure and a geographic map based on the climatic classification (**Figure 5**).

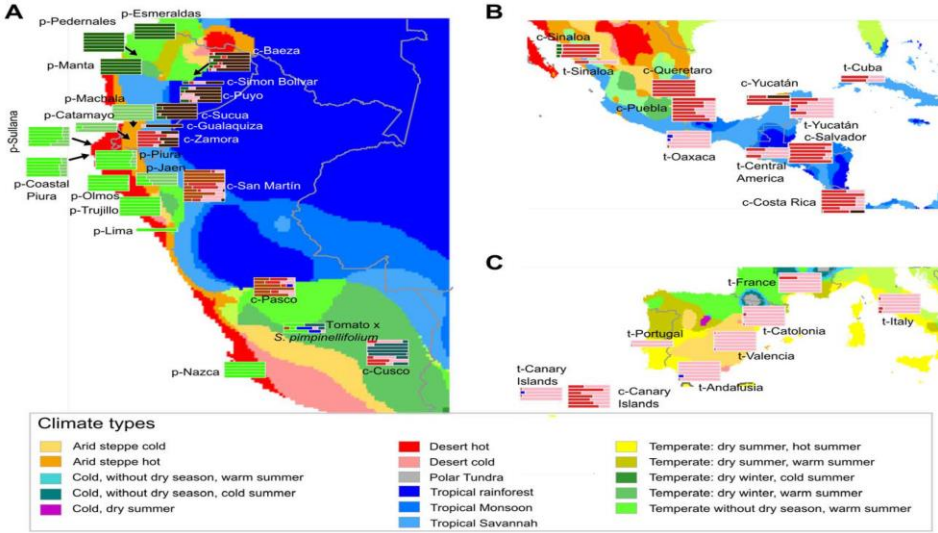


Figure 5. Geographical distributions of the population structure revealed by SOLCAP SNPs, as adapted from Blanca et al., (2012). Different colored bars represent the proportion of the population structure.

3.6 Cloned genes/QTL

Tomato is probably one of the crops with the largest number of single mutations used for its breeding (as reviewed by Grandillo and Cammareri, 2018, and Rothan et al., 2019). Before the SNP discovery, due to the limited genetic diversity of domesticated tomato accessions, the populations used for linkage mapping have been generated by crosses between a cultivated and a close wild tomato species (Foolad, 2007; Foolad and Panthee, 2012). Since the development of molecular markers, these segregating populations have become an effective and efficient tool to construct high density genetic linkage maps (Tanksley et al., 1992), allowing the detection of Quantitative Trait Loci (QTLs). By using different linkage populations and multiple molecular markers, including RFLP, SSR and SNPs, hundreds of QTLs have been reported, for different agronomical, morphological, and quality related traits (Grandillo and Tanksley, 1996b; Tanksley et al., 1996; Fulton et al., 1997; Bernacchi et al., 1998; Chen et al., 1999; Grandillo et al., 1999; Fulton et al., 2000; Monforte and Tanksley, 2000; Saliba-Colombani et al., 2001; Causse et al., 2002; Doganlar et al., 2003; van der Knaap and Tanksley, 2003; Fridman et al., 2004; Baldet et al., 2007; Foolad, 2007; Jiménez-Gómez et al., 2007; Cagas et al., 2008; Dal Cin et al., 2009; Sim et al., 2010; Ashrafi et al., 2012; Haggard et al., 2013; Kinkade and Foolad, 2013).
However, among the detected QTLs, only a few have been cloned and functionally validated (Bauchet and Causse, 2012; Rothan et al., 2019). The first gene cloned by positional cloning in tomato was the *Pto* gene, conferring resistance to *Pseudomonas syringae* races, with the assistance of RFLP markers (Martin et al., 1993). Based on the same RFLP map, *Fen*, another member of this gene family, was also soon reported (Martin et al., 1994). From then on, different resistance genes were identified and cloned based on RFLP markers, such as *Cf-2*,

a leucine-rich repeat protein conferring resistance to *Cladosopum fulvum* strains (Dixon et al., 1996); *Prf*, another resistance gene to *Pseudomonas syringae* pv. tomato (Pst) strains (Salmeron et al., 1996); *Ve* conferring Verticilium wilt resistance, encoding surface-like receptors (Kawchuk et al., 2001) and others. Some other markers were also developed and applied for resistance gene identification, such as *Ph-3* gene from *S. pimpinellifolium* conferring resistance to *Phytophthora infestans*, which was cloned based on cleaved amplified polymorphic sequences (CAPS) or insert/deletion (InDel) markers (Zhang et al., 2014). Sequence-characterized amplified region (SCAR) markers and cleaved amplified polymorphic sequence (CAPS) markers are also applying to map tomato yellow leaf curl virus resistance gene *Ty-2* (Yang et al., 2014).

Some important genes/QTL involved in developmental processes were also identified and cloned with the assistance of molecular markers. Among them, *fw2.2*, a major QTL controlling tomato fruit weight, was one of the first examples. With the benefits of CAPs markers, a single candidate gene ORFX on chromosome 2 was identified and cloned (Frary et al., 2000), which alters tomato fruit size likely by expression regulation rather than sequence and structure variation of the encoded protein (Nesbitt and Tanksley, 2002). Recently, some other major QTLs were functionally validated, such *fw3.2* (corresponding to a cytochrome P450 gene) (Chakrabarti et al., 2013) and *fw11.2* (corresponding to a cell size regulator) (Mu et al., 2017). Some major QTLs closely related to fruit weight were also reported, such as *OVATE*, a negative regulatory gene causing pear-shaped tomato fruits (Liu et al., 2002); SUN, a retrotransposon-mediated gene (Xiao et al., 2008); locule number *fas* (Huang and van der Knaap, 2011) and *lc* (Munos et al., 2011). Other cloned genes related to tomato development are summarized in a recent review paper (Rothan et al., 2019).

Tomato fruits are rich in diverse nutrients and health-promoting compounds, such as sugars, organic acids, amino acids and volatiles (Goff and Klee, 2006; Klee, 2013). However, breed tomatoes with high nutrition and strong flavor still remain a major breeding challenge (Tieman et al., 2012; Klee and Tieman, 2013; Klee and Tieman, 2018; Zhao et al., 2019). *Lin5*, a major QTL modifying sugar content in tomato fruit, was cloned about 20 year ago (Fridman et al., 2000). In various genetic backgrounds and environments, the wild-species allele increased glucose and fructose contents compared to cultivated allele (Fridman et al., 2000). In addition, this gene shared a similar expression pattern in tomato, potato and Arabidopsis (Fridman and Zamir, 2003). Recently a *SWEET* protein, a plasma membrane-localized glucose efflux transporter, was shown to play a role in the ratio of glucose and fructose accumulation (Shammai et al., 2018). A balanced content of sugars and organic acids is crucial for consumer preference (Tieman et al., 2017). Recently, a major QTL regulating malate content was cloned, corresponding to an *Aluminium Activated Malate Transporter 9* (*Sl-ALMT9*) (Ye et al., 2017). In a new recent study, it was further found that this QTL was also likely regulating the content of citrate in tomato fruits (Zhao et al., 2019). Though only a few QTLs regulating sugars and organic acids have been functionally validated, this knowledge is important for understanding the regulation mechanisms. Several genes involved in the variation of volatile production were also characterized (Tieman et al., 2006; Tikunov et al., 2013; Klee 2010; Klee and Tieman, 2018).

3.7 New resources for genes/QTL identification

Lin et al., (2014) demonstrated the benefits of whole-genome resequencing of the two extreme bulk populations from an F_2 population of tomato, where many fruit weight QTLs were identified, including *fw2.1*, *fw2.2*, *fw2.3*, *lcn2.1*, *lcn2.2*, *fw9.1*, *fw9.3*, *fw11.1*, *fw11.2* and *fw11.3*. Whole-genome-sequencing of bulked F_2 plants with contrasted phenotypes offers the opportunity to identify the SNPs that are putatively related to the target phenotypes via aligning the sequenced data to the reference genome (Garcia et al., 2016). This approach has been efficient in identifying mutations, especially generated by EMS (Garcia et al., 2016).

However, the genetic diversity of linkage populations is limited to the two parental accessions used for crossing. In order to overcome this limitation, multi-parent advanced generation intercross (MAGIC) populations offer an alternative, which has been generated for different species, such as Arabidopsis (Kover et al., 2009), rice (Bandillo et al., 2013), wheat (Huang et al., 2012; Mackay et al., 2014), faba bean (Sallam and Martsch, 2015), sorghum (Ongom and Ejeta, 2017) and tomato (Pascual et al., 2015). The first tomato MAGIC population was developed by crossing eight resequenced tomato lines and there was no obvious population structure in this population. The linkage map was 87% larger than those derived from biparental populations and some major fruit quality QTLs were identified by using this approach (Pascual et al., 2015). Recently, this MAGIC population was also used for identifying QTLs under water deficit and salinity stresses and many stress-specific QTLs were identified (Diouf et al., 2018).

3.8 Genome-wide association studies

3.8.1 The conditions for applying Genome-wide association studies

Association mapping is used to detect associations between a given phenotype and genetic markers in a population of unrelated accessions. If the genetic markers cover the whole genome, it is referred to genome-wide association studies (GWAS). This technology was first developed in humans. After the demonstration of GWAS power to analyze human diseases (Klein et al., 2005), it was quickly adopted in major crops (Brachi et al., 2011; Luo, 2015; Liu and Yan, 2019). In tomato, the first reported association study was performed to identify the

SNPs associated with the fruit weight QTL *fw2.2*. However, the authors did not find any positive associated SNP in a small collection of 39 cherry tomato accessions (Nesbitt and Tanksley, 2002).

In order to efficiently apply GWAS in tomato, linkage disequilibrium (LD) in different tomato types was assessed using different molecular markers. In general, the LD in cultivated tomato accessions was larger than that of wild species, which could be up to about 20 Mbs, while cherry tomatoes ranged in between (Van Berloo et al., 2008; Mazzucato et al., 2008; Sim et al., 2010; Ranc et al., 2012; Xu et al., 2013; Sauvage et al., 2014; Zhang et al., 2016; Bauchet et al., 2017a). These results also indicated that modern tomatoes lost genetic diversity during tomato domestication and breeding. Admixture of cherry tomatoes with modern cultivars and wild species could help reduce the large LD and overcome the low resolution of association mapping of modern tomato cultivars (Ranc et al., 2012). The average high degree of LD is beneficial in terms of the minimum number of molecular markers needed to cover the whole genome. For example, (Xu et al., 2013) performed an association mapping on 188 tomato accessions with 121 polymorphic SNPs and 22 SSRs. They successfully identified 132 significant associations for six quality traits. Before the availability of large SNP number, molecular markers such as SSRs were popular for GWAS. In particular, (Zhang et al., 2016) genotyped 174 tomato accessions including 123 cherry tomato and 51 heirlooms with 182 SSRs and performed GWAS for fruit quality traits. A total of 111 significant associations were identified for 10 traits and many previously identified major QTLs were located in/near regions of the significant associated markers. The authors further extended the phenotypes to volatiles (Zhang et al., 2015), as well as sugars and organic acids (Zhao et al., 2016). Many significant associations were also identified and some of them were consistent with other GWAS focusing on the same traits that were based on genome-wide SNPs (Sauvage et al., 2014; Bauchet et al., 2017b; Tieman et al., 2017; Zhao et al., 2019).

With the availability of the reference tomato genome (The Tomato Genome Consortium, 2012), millions of SNPs became available and allowed the identification of causative polymorphisms. For instance, the causative gene *SlMYB12* conferring pink tomato fruit color was identified in a GWAS using 231 sequenced tomato accessions (Lin et al., 2014). Several mutations were further identified in the protein structure of SlMYB12 and the authors identified three recessive alleles of this gene useful for pink tomato breeding (Lin et al., 2014).

However, whole-genome-sequencing is still quite expensive, especially at a large population scale, which greatly limits the wide applications. SNP arrays were thus developed to overcome this limit (Hamilton et al., 2012; Sim et al., 2012b). Sauvage et al., (2014) genotyped 163 tomato accessions composed of large-fruit, cherry and wild tomato accessions with the SolCAP array, generating a total of 5995 high quality SNPs. Then they performed GWAS using a multi-locus mixed model (MLMM; (Segura et al., 2012) for 36 metabolites that were highly correlated during two growth periods and identified 44 candidate loci associated for different fruit metabolites (Sauvage et al., 2014). Among the candidate loci, they identified a gene with unknown function on chromosome 6 that was strongly associated with malate content. This association was further identified in different GWAS and meta-analysis of GWAS based on different populations (Bauchet et al., 2017b; Tieman et al., 2017; Ye et al., 2017; Zhao et al., 2019) and was further validated as an *Al-Activated Malate Transporter 9* (*Sl-ALMT9*) (Ye et al., 2017). In a meta-analysis of GWAS based on three populations, it was further found that this gene was also significantly associated with citrate content in tomato fruits, demonstrating its important role in the regulation of organic acids in tomato (Zhao et al., 2019). In fact, the Al-activated malate transporters are a family of plant-specific proteins, which are important for plant root tissue and function (Delhaize et al., 2007).

Bauchet et al., (2017b) genotyped 300 tomato accessions with both the SolCAP and CBSG arrays, generating a total of 11,012 high quality SNPs, which were used for GWAS using both MLMM and multi-trait mixed model (MTMM) (Korte et al., 2012). A total of 79 significant associations were identified for 13 primary and 19 secondary metabolites in tomato fruits. Among these, two associations involving fruit acidity and phenylpropanoid content were particularly investigated (Bauchet et al., 2017b). The same population was also characterized for agronomic traits and many QTLs were identified, such as *fw2.2* and *fw3.2*. Within this panel, the authors also demonstrated that intermediate accessions shared different haplotype patterns compared to domesticated and wild tomatoes (Bauchet et al., 2017a). GWAS for similar quality traits were also performed in other collections (Ruggieri et al., 2014; Zhang et al., 2016).

With the fast development of whole-genome-sequencing technology and the reduction of cost per genome, it is possible to sequence hundreds of diverse tomato collections. For instance, (Tieman et al., 2017) sequenced 231 new accessions and combined these data with 245 previously sequenced genomes, generating a total of 476 genome sequences. These data were then used for GWAS for diverse flavor-related metabolites, including 27 volatiles, total soluble solids, glucose, fructose, citric acid, and malic acid. A total of 251 significant associations were detected for 20 traits. Two loci were significantly associated with both glucose and fructose, corresponding to two major QTL *Lin5* and *SSC11.1*. By combining with selection analysis, it was further shown that the negative correlation between sugar content and fruit weight was likely caused by the loss of high-sugar alleles during domestication and improvement of ever-larger tomato fruits (Tieman et al., 2017). In addition, some good candidate genes involved in tomato volatile contents were also identified, such as Solyc09g089580 for guaiacol and methylsalicylate. By combining the three significant associated loci for geranylacetone and 6-methyl-5-hepten-2-one, it was shown that the allelic combinations conferring favorable aromas were progressively lost during domestication and breeding (Tieman et al., 2017).

3.8.2 Meta-analysis

However, with the results of several GWAS in tomato for the same trait, only some significant associations could be identified in different studies, indicating strong cross-study heterogeneity, which refers to the non-random variance in the genetic effects between different GWASs. The main sources of heterogeneity include population structure, linkage disequilibrium, phenotyping measurement methods, environmental factors, genotyping methods, G × E interactions … (Evangelou and Ioannidis, 2013). Meta-analysis of GWAS is a new approach to combine different GWAS properly handling the heterogeneity.

(Zhao et al., 2019) reported the meta-analysis of GWAS from three tomato populations (Sauvage et al., 2014; Bauchet et al., 2017b; Tieman et al., 2017). Following genotype imputation, a total of 775 tomato accessions and 2,316,117 SNPs were used in the meta-analysis and a total of 305 significant associations were identified for the contents of sugars, organic acids, amino acids and flavor-related volatiles. By looking at the five loci associated with both fructose and glucose, they showed that sugar contents significantly increased with the number of wild alleles. The authors also demonstrated that domestication and improvement have had an impact on citrate and malate content. In particular, the major QTL *Al-Activated Malate Transporter 9* of malate was also significantly associated with citrate and another malate transporter was identified for citrate content on chromosome 1. This study also identified many new significant associations for flavor-related volatiles. By targeting six significant associations, it was further demonstrated that modern tomato accessions had a limited flavor due to a lower content of pleasant volatiles but also a higher content of unpleasant volatiles compared to cherry tomatoes (Zhao et al., 2019).

3.9. Genetic dissection of abiotic stress tolerance

3.9.1 Genetic control of G x E interaction

In section 2.3.2 above, the impact of different abiotic stresses on tomato was described. Nevertheless a large diversity of response has been shown notably between the wild species and the cultivated one, but also across cultivated accessions. Several studies were conducted to understand the genetic mechanisms leading to such variation in tomato response to environmental stresses. Elucidating the genetic determinants of tomato response to abiotic stress was possible thanks to the high genetic diversity present in the *S. lycopersicum* clade.

A large panel of genetic resources is available for the tomato community, including both cultivated and wild species (section 3.1). Screening the genetic diversity in both compartments brought to light high loss of diversity within the cultivated group (Lin *et al.* 2014) due to extensive directional selection towards agronomic performance traits. However, substantial diversity for environmental response genes remains in the cultivated group that could be attributed to local adaptations during the diversification for both climatic conditions and growth conditions. This is identified by the presence of substantial genotype-by-environment (GxE) interactions, as observed in different intraspecific experimental tomato populations (Villalta *et al.* 2007; Mazzucato *et al.* 2008; Albert *et al.* 2016a; Diouf *et al.* 2018).

Besides, wild species constitute a reservoir of specific genes related to abiotic stress tolerance, derived from adaptation to their growing and typically harmful local habitats. For example the two wild relative species *S. habrochaites* and *S. pennellii* are more tolerant to chilling stress (Bloom *et al.* 2004) and to drought and salinity stress conditions (Bolger *et al.* 2014), compared to cultivated species. The presence of tolerance genes in the wild species and the genetic diversity of stress response genes in cultivated clade give clues to achieve considerable progress in tomato breeding for climate-smart cultivars.

Several studies investigated the genetic nature of tomato response to abiotic stresses since a high density genetic map was made available. Grandillo *et al.* (2013) and Grandillo and Cammareri (2016) reported a summary of the QTLs that were identified under different abiotic stress conditions. The **table 4** summarizes abiotic stress QTL identified during the last decade only. These QTLs were mapped in different population types and with different mapping methods covering the wide range of mapping strategies available in plant genetics. These studies highlighted several phenotypic traits that were defined to assess tomato response to abiotic factors due to the complexity of stress response mechanisms. For example, Kazmi *et al.* (2012) used seed quality traits to identify QTLs associated with tomato germination capacity under WD, CS, SS and HT stress. They identified no less than 90 seed-quality QTLs under stress conditions. Physiological parameters under WD and nitrogen-deficiency conditions were mapped in sub-NILs (Arms *et al.* 2016) and 130 F10 RILs (Asins *et al.* 2017) populations, respectively. Metabolite variation in tomato seeds under SS was studied by Rosental *et al.* (2016) and several QTLs were identified in 72 ILs derived from the introgression of chromosome fragments of *S. pennellii* LA716 into the domesticated tomato cultivar M82. A recent study used gene expression data under WD and control conditions and identified some WD interactive eQTLs (Albert *et al.* 2018). This approach permitted the distinction between *cis* and *trans* regulatory eQTL clarifying the patterns of expression regulation in tomato under WD leading to genotype-by-environment interaction. Combining expression data with QTL analysis thus helped to identify candidate stress-response genes and could be useful for the optimal choice of genetic markers to conduct MAS for stress adaptation.

However, the majority of the studies used agronomic traits instead of physiological parameters or metabolic traits to evaluate the impact of abiotic stress. This has led to the definition of different stress index according to

breeding objectives (**Table 4**); thus QTL identified for such stress index could be directly used in breeding programs.

Table 4. QTL studies on tomato abiotic stress published during the last decade. For each study, the number of genotypes analyzed, the population cross-design and the number and type of markers used are displayed. The columns "Stress treatment" and "Stress period" present the level of stress applied and the period on which stress was applied. The column "Phenotypes" highlights the phenotypic traits that were evaluated to conduct the QTL/association analysis. The phenotypic traits usually correspond to different traits: Seed quality (germination ability); Fruit quality (SSC, Vitamin C, pH, firmness, organic acids); Plant architecture and vegetative growth (diameter, leaf length, height, dry matter content, specific leaf area, biomass); Phenology (flowering, ripening time); Productivity (yield, fruit weight, number of fruits); Physiological traits (WUE); Model parameters (Maximum cell wall extensibility, membrane conductivity, sugar active uptake, membrane reflection, Pedicel conductivity, soluble sugar concentration, fruit dry weight, fruit water content, xylem conductivity).

Treatment	Number of individuals	Marker types	Stress treatment	Stress period	Cross-design	Phenotypes	Number of QTLs	Reference
Cold stress (CS)								
CS	83 RILs	865 SNP	Cold stress (12 °C)	Seed germination	Bi-parental (Interspecific)	Seed quality	12 QTLs	Kazmi et al. 2012
CS	146 RILs	120 SSR	Cold stress (11°C)	Seed germination	Bi-parental (Interspecific)	Germinatin ratio	5 QTLs	Liu et al. 2016
CS	146 RILs	120 SSR	2°C for 48 hours	4 - 5 true leaves	Bi-parental (Interspecific)	Chilling injuries	9 QTLs	Liu et al. 2016
Hight temperature stress (HT)								
HT	192 F2	106 AFLP markers	Minimal /Maximal T° > 25°C /40°C	Transplanting - end of the experiment	Bi-parental (Intraspecific)	Fruit set	6 QTLs	Grilli et al. 2007
HT	160 F2	62 RAPD, ISSR and AFLP markers	Day/Night T° = 37.2°C /24.7°C	All growing season	Bi-parental (Interspecific)	Yield; Fruit quality; Reproductive traits	21 QTLs	Lin et al. 2010
HT	83 RILs	865 SNP	Heat stress (35-36°C)	Seed germination	Bi-parental (Interspecific)	Seed quality	16 QTLs	Kazmi et al. 2012
HT	180 F2	96 SNP	Day/Night T° = 31°C /25°C	From 1st inflorescences appearance	Bi-parental (Intraspecific)	Reproductive traits	13 QTLs	Xu et al. 2017
HT	98 F8 RILs	727 SNP	37°C	Seed germination	Bi-parental (Interspecific)	Thermo-tolerance, Thermo-inhibition, Thermo-dormancy	9 QTLs	Geshnizjani et al. 2018
Salinity stress (SS)								
SS	123 RILs	156 SSR, SCAR markers	125 mM NaCl	15 days after transplanting to the end of the experiment	Bi-parental (Interspecific)	Root-stock induced physiological parameters; Vegetative growth	57 QTLs	Asins et al. 2010
SS	52 ILs	!!	150 mM NaCl	21 days from the seven true leaf stage	Bi-parental (Interspecific)	Plant architecture; antioxidant content	71 QTLs	Frary et al. 2010
SS	52 ILs	!!	150 mM NaCl	15 days of treatment	Bi-parental (Interspecific)	Plant architecture; Vegetative growth	225 QTLs	Frary et al. 2011
SS	78 ILs	!!	700 mM NaCl + 70 mM CaCl2	4 days after transplanting	Bi-parental (Interspecific)	Survival performance	4 QTLs	Li et al. 2011
SS	90 ILs	!!	700 mM NaCl + 70 mM CaCl2	4 days after transplanting	Bi-parental (Interspecific)	Survival performance	6 QTLs	Li et al. 2011
SS	100 RILs	134 SSR, SCAR markers	75 mM NaCl	15 days after transplanting to the end of the experiment	Bi-parental (Interspecific)	Root-stock induced physiological parameters; Vegetative growth	2 QTLs	Asins et al. 2010

SS	83 RILs	865 SNP	Two levels of SS (-0.3 & -0.5 MPa NaCl)	Seed germination	Bi-parental (Interspecific)	Seed quality	32 (26) QTLs	Kazmi et al. 2012
SS	124 RILs	2059 SNPs	8.94 dS/m.	10 days after the transplanting	Bi-parental (Interspecific)	Yield; Fruit quality; Biomass	54 QTLs	Asins et al. 2015
SS	72 ILs	!!	EC = 6 dS/m	Planting - end of the experiment	Bi-parental (Interspecific)	Seed weight; Seed Germination; Metabolites	131 QTLs	Rosental et al. 2016
SS	253 MAGIC RILs	1345 SNP	Two lelvels of SS (Ec=3.7 dS/m-1 & Ec=6.5 dS/m-1)	Transplanting - end of the experiment	MAGIC (Intraspecific)	Fruit quality; Plant architecture and vegetative growth; Phenology; Productivity	35 QTLs	Diouf et al. 2018
Water deficit stress (WD)								
WD	75 ILs	!!	WD (30m3 of water irrigation for 1000m2)	Transplanting - end of the experiment	Introgression Line (Interspecific)	Fruit quality; Plant architecture and vegetative growth; Productivity	114 QTL	Gur et al. 2011
WD	83 RILs	865 SNP	Two levels of Osmotic stress (-0.3 & -0.5 MPa PEG)	Seed germination	Bi-parental (Interspecific)	Seed quality	23 (19) QTLs	Kazmi et al. 2012
WD	119 RILs	679 SNP	WD (40% ETP)	Transplanting - end of the experiment	Bi-parental (Intraspecific)	Fruit quality; Plant architecture and vegetative growth; Phenology; Productivity	36 QTL	Albert et al. 2016a
WD	141 small-fruit accessions	6100 SNPs	WD (40% ETP)	Transplanting - end of the experiment	GWAS-panel	Fruit quality; Plant architecture and vegetative growth; Phenology; Productivity	100 QTLs	Albert et al. 2016b
WD	18 sub-NILs	10 markers (SNP; SCAR; CAP)	WD (33%ETP)	Transplanting - end of the experiment	Near-Introgression Line (Interspecific)	Physiological traits; Plant architecture	2 QTLs regions	Arms et al. 2016
WD	117 F7 RILs	501 SNP	WD (49% ETP)	Transplanting - end of the experiment	Bi-parental (Intraspecific)	Model parameters	8 QTLs	Constantinescu et al. 2016
WD	241 MAGIC RILs	1345 SNP	WD (50% ETP)	Transplanting - end of the experiment	MAGIC (Intraspecific)	Fruit quality; Plant architecture and vegetative growth; Phenology; Productivity	22 QTLs	Diouf et al. 2018
WD	124 RILs	501 SNP	WD (60% ETP)	Transplanting - end of the experiment	Bi-parental (Intraspecific)	Fruit quality; Plant architecture and vegetative growth; Phenology; Productivity	23 QTLs	Albert et al. 2018
WD	124 RILs	501 SNP	WD (60% ETP)	Transplanting - end of the experiment	Bi-parental (Intraspecific)	Gene expression level for 274 genes	103 e-QTL	Albert et al. 2018
Other abiotic stress								
Oxidative stress	83 RILs	865 SNP	Oxidative stress (300 mm H2O2)	Seed germination	Bi-parental (Interspecific)	Seed quality	17 QTLs	Kazmi et al. 2012
N- deficiency	130 F10 lines	1899 SNP	N deficiency (NH4+: 0.1mM & NO3-: 1mM)	Transplanting - 1st truss fruit set	Bi-parental (Interspecific)	Vegetative growth, Leaf nitrogen content; Xylème sap hormone content	40 QTLs	Asins et al. 2017

Until now, most QTL studies on tomato were conducted on single stress evaluation, achieving a better characterization of genetic loci involved in tomato response to a given abiotic stress. Further studies should target genomic regions that interfere in response to stress combinations. Few examples of such studies are available in plants (Davila Olivas *et al.* 2017).
Genotype-by-environment (GxE) interaction usually occurs in cultivated crops exposed to abiotic stresses. Two strategies are commonly adopted by breeders to deal with GxE: (i) developing some elite cultivars for specific targeted environment or (ii) breeding stable cultivars for a wide range of environmental conditions. The first strategy will allow to reach high yield in predictable environments (likely controlled environments) while the second strategy will be more efficient for reducing at an optimized level, the yield decrease in unpredictable environments. This has led plant geneticists into the question of genetic control of phenotypic plasticity related to GxE phenomenon. Some studies addressed this question in major crop species and identified different plasticity QTLs. Kusmec *et al.* (2017) for example suggested that in maize, genes controlling plasticity for different environments are in majority distinct from genes controlling mean trait variation, assuming a possible co-selection for stability and yield performance concurrently. In tomato, plasticity QTLs were also identified in intraspecific populations under WD and SS conditions (Albert *et al.* 2016a; Diouf *et al.* 2018). Extending the environmental range to different stress conditions could be a way to reliably identify multi-stress response genes that would be useful in the task of breeding climate-smart tomato.

3.9.2 Grafting as a defense against stresses

For many plant species specially vegetables and fruit trees, grafting has been considered as a solution to manage soil-borne disease and to improve crop response to a variety of abiotic stresses (King et al. 2010). For stress induced by extreme soil conditions, grafting elite cultivars onto genetic resistant rootstocks is an attractive alternative to introgression from wild resources due to the side effects of linkage drag and the polygenic nature of abiotic stress tolerance. However, grafting requires paying specific attention to the scion x rootstock combination in order to achieve better performance. In tomato interactions between the scion and the rootstock were detected in different grating operations with alteration in fruit quality components, plant vigor, plant hormonal status and final yield (Kyriacou et al. 2017). This highlights the necessity to test different combinations of scion-rootstocks in one hand, and in the other hand to have a better understanding of how grafting impact the targeted breeding traits for efficient utilization of rootstocks under stressful environments.
Different tomato rootstock populations were developed and characterized accordingly. This involves populations generated from interspecific crosses between a cherry tomato accession and two wild relatives from *S. pimpinellifolium* and *S. cheesmaniae* (Estañ et al. 2009). These populations were studied under salinity (Albacete et al. 2009; Asins et al. 2010, 2015; Asins et al. 2013) and N-deficiency stress conditions (Asins et al. 2017). They revealed that grafting could induce variation in leaf hormonal content and ion concentrations correlated to vegetative growth and yield under salinity. The effect mediated by rootstock under salinity has a polygenic nature and is controlled by different QTLs among which one, located on chromosome 7, was related to two HTK candidate genes, involved in ion transport and cell homeostasis regulation. However, while grafting under salinity present a promising approach to maintain or increase tomato yield, some drawbacks were recorded concerning higher incidence of BER and delayed fruit ripening.
The hormonal status changs induced by rootstock was also shown as being potentially exploitable to increase tomato WUE (Cantero-Navarro et al. 2016). More generally, Nawaz et al. (2016) reviewed the effect of grafting on ion accumulation within horticultural crops highlighting the need for deeper characterization of rootstock x scion x environment interaction both at phenotype and genetic levels for effective utilization of grafting as a technique to manage extreme soil conditions for crops.
Beside of the direct use of genetic control of pests and pathogens, grafting susceptible cultivars onto selected vigorous rootstocks may counteract soilborne biotic stresses as well as abiotic stresses. Grafting was also proposed for improving virus resistance by enhancing RNA-silencing (Spano et al. 2015). A great challenge is consequently to breed for rootstocks that can withstand combined biotic and abiotic stresses.

3.10 Omic approaches

3.10.1 Metabolome analyses

Metabolomics has an important role to play in characterization of natural diversity in tomato (Schauer et al., 2005; Fernie et al., 2011). Metabolome analysis can be done in a targeted way to better characterize known metabolites (Tieman et al., 2006) or untargeted manner to identify new metabolites (Tikunov, 2005). As well, it can boost the biochemical understanding of fruit content and be an enhancer for quality breeding (Fernie and Schauer, 2009; Allwood et al., 2011). Metabolome analyses were used to analyse fruit composition at a high-throughput level. Metabolite QTL (mQTL) have been identified for non-volatiles metabolites like sugars, pigments or volatiles compounds (Bovy et al., 2007; Klee 2010; 2013; Klee and Tieman, 2018). This was done on several interspecific populations, notably on *S. pennelli* (Alseek et al., 2015, 2017) and *S. chmielewskii* (Do et al., 2010; Ballester et al., 2016) introgression lines and intraspecific crosses (Saliba-Colombani et al., 2001;

Causse et al., 2002; Zanor et al., 2009). The interaction between tomato plant and thrips was also studied by metabolome profiling (Mirnezhad *et al.*, 2010).

3.10.2 Transcriptome analyses for eQTL mapping

Several studies analysed the transcriptome changes along fruit development (Patison et al. 2015; Giovanonni et al., 2017; Shinozaki et al., 2018) revealing key changes in gene expression during the different stages. Analysis of the genetic control of such variations in segregating populations was also performed (Ranjan et al., 2016; Coneva et al., 2017). Characterizing the natural diversity of gene expression across environments is also n important step in understanding genotype-by-environment interactions. Albert et al. (2018) identified some eQTL in response to water stress and showed the large differences between the transcriptome of leaf and fruit under well irrigated and water stress conditions. The authors also studied allele-specific expression (ASE) in the F1 hybrid

To reveal genes deviating from the 1/1 allele ratio expected and showed a large range of genes whose variation exhibited significant ASE-by-watering regime interaction, among which ~80% presented a response to water deficit mediated through a majority of trans-acting.

3.10.3 Multi-omic approach

Combining metabolome and transcriptome may give clues about the genetic control of fruit composition as underlined by Prudent et al. (2011). Zhu et al., (2018) performed a multi-omic study by integrating data of the genomes, transcriptomes and metabolomes. Up to 3,526 significant associations were identified for 514 metabolites and 351 of them were associated with unknown metabolites. Correlation analysis between genomes and transcriptomes identified a total of 2,566 cis-eQTL and 93,587 trans-eQTL. Rigorous multiple correction tests between transcriptomes and metabolomes identified 232,934 expression-metabolite correlations involving 820 chemicals and 9,150 genes. By integrating these three groups, a total of 13,361 triple relationships (metabolite-SNP-gene) were further identified, including 371 metabolites, 970 SNPs, and 535 genes. Selection analysis discovered 168 domestication sweeps and 151 improvement sweeps, representing 7.85% and 8.19% of the tomato genome, respectively. A total of 4,095 and 4,547 genes were located within the identified domestication and improvement sweeps. In addition, a total of 46 steroidal glycoalkaloids were identified and five significant associations were located within domestication or improvement sweeps. They also showed that the introgression of resistance genes also introduced significant differences in some metabolites.

3.10.4 miRNA and epigenetic modifications

Epigenome is the complete set of epigenetic marks at every genomic position in a given cell at a given time (Taudt et al., 2016). These marks fall into six categories, including DNA modifications, histone modifications, chromatin variants, nucleosome occupancy, RNA modifications, non-coding RNAs, chromatin domains and interactions (Stricker et al., 2017). Technological advances nowadays make it possible to achieve high-resolution measurements of epigenome variation at a genome-wide scale and great achievements have been made in human, rat, yeast, maize, tomato, Arabidopsis and soybeans (Taudt et al., 2016; Giovannoni et al., 2017).

Most of epigenome studies in tomato focused on the molecular regulations of fruit ripening and development (Gallusci et al., 2016; Giovannoni et al., 2017). Among these, histone post-translational modifications play an important role, which include phosphorylation, methylation, acetylation and mono-ubiquitination of lysine residues (Berr et al., 2011). In Arabidopsis, histone post translational modifications are involved in many aspects of plant development and stress adaptation (Ahmad et al., 2010; Mirouze and Paszkowski, 2011). In tomato, at least nine DNA methyltransferases and four DNA demethylases have been identified (Gallusci et al., 2016). Expression patterns of different histone modifiers in some fresh fruits have also been identified, such as histone deacetylases, histone acetyltransferase, and histone methyltransferases (Gallusci et al., 2016). Repression of tomato Polycomb repressive complex 2 (PRC2) components *SlEZ1* altered flower and fruit morphology (How Kit et al., 2010) and *SlEZ2* altered fruit morphology, such as texture, color and storability (Boureau et al., 2016). These results demonstrated that epigenetic regulations are important for many biological processes.

Very few phenotypes have been associated to epi-mutations. Manning et al., (2006) identified a naturally occurring methylation epigenetic mutation in the SBP-box promoter residing at the colorless non-ripening (*Cnr*) locus, a major component in the regulatory network controlling tomato fruit ripening (Eriksson et al., 2004). Quadrana *et al.* (2014) identified an epi-mutation responsible of the variation in vitamin E in the fruit. In order to determine whether the process of tomato fruit ripening involves epigenetic remodeling, Zhong et al., (2013) found that tomato ripen prematurely under methyltransferase inhibitor 5-azacytidine. Up to 52,095 differentially methylated regions were identified, representing 1% of the tomato genome. In particular, demethylation regions were identified in the promoter regions of numerous ripening genes. In addition, the epigenome status was not static during tomato fruit ripening (Zhong et al., 2013). Shinozaki et al., (2018) performed a high-resolution spatio-temporal transcriptome mapping during tomato fruit development and ripening. Some tissue-specific ripening-associated genes were identified, such as SlDML2. Together with other analyses, these results indicate that spatio-temporal methylations play an important role during tomato fruit development and ripening (Shinozaki et al., 2018).

Lü et al., (2018) investigated the functional elements of seven climacteric fruit species (apple, banana, melon, papaya, peach, pear and tomato) and four non-climacteric fleshy fruit species (cucumber, grape, strawberry and watermelon). By analyzing 361 transcriptome, 71 accessible chromatin, 147 histone and 45 DNA methylation profiles from the fruit ENCODE data, three types of transcriptional feedback circuits were identified controlling ethylene-dependent fruit ripening (Lü et al., 2018). In particular, H3K27me3, associated with silencing of the flowering regulator FLOWERING LOCUS C and floral homeotic gene AGAMOUS (He, 2012), played a conserved role in dry and ethylene-independent fruits by restricting ripening genes and their orthologs.

MicroRNA (miRNAs) is another type of epigenetic regulation. miRNAs are a class of 20- to 24-nucleotide noncoding endogenous small RNAs that are important in transcriptional or post-transcriptional regulation by transcript cleavage and translation repression (Chen, 2005; Chen, 2009; Rogers and Chen, 2013; Sanei and Chen, 2015). miRNAs are encoded by miRNA genes, which contain the TATA-box motif and transcription factor binding motifs, and are regulated by general and specific transcription factors (Xie et al., 2005; Megraw et al., 2006; Rogers and Chen, 2013; Yu et al., 2017). miRNAs play an important role in many biological processes, including physiological, developmental, defense and environmental changes both in humans (Calin and Croce, 2006; Mendell and Olson, 2012; Cui et al., 2017b; Hill and Tran, 2018), animals (Ambros, 2004; Rajewsky, 2006; Grimson et al., 2008) and plants (Rogers and Chen, 2013; Won et al., 2014; Sanei and Chen, 2015; Cui et al., 2017a; You et al., 2017; Yu et al., 2017). Some regulatory mechanisms of the core components of the dicing complex, such as DICER-LIKE1 (DCL1) and HYPONASTIC LEAVES1 (HYL1) have been uncovered (Manavella et al., 2012; Cho et al., 2014; Zhang et al., 2017). Proteins promoting pre-miRNA processing and reducing miRNA levels have also been identified, such as CAP-BINDING PROTEIN 80 (CBP80), CAP-BINDING PROTEIN 20 (CBP20), STABILIZED1 (STA1) and others (Gonatopoulos-Pournatzis and Cowling, 2015; Yu et al., 2017). Some proteins could reduce the accumulation of both mature pre-miRNA and mature miRNA, such as CDC5, NOT2, Elongator, and DDL (Yu et al., 2008; Wang et al., 2013; Zhang et al., 2013; Fang et al., 2015). Though many processes involved in miRNA biogenesis, degradation and activity have been discovered, our knowledge regarding the subcellular locations of these processes is still largely unknown (Yu et al., 2017).

During the tomato genome sequencing, a total of 96 conserved miRNA genes were predicted. Among them, 34 miRNA have been identified and 10 are highly conserved in both tomato and potato (The Tomato Genome Consortium, 2012). Several studies focused on the characterizations of miRNAs in tomato during fruit development (Moxon et al., 2008; Zuo et al., 2012; Gao et al., 2015). The dominant sRNAs were 21- to 24-nt sRNAs (Mohorianu et al., 2011; Zuo et al., 2012; Gao et al., 2015). Many ripening-associated gene transcription factors were regulated by certain miRNA families, such as miR156/157, miR159, miR160/167, miR164, miR171 and miR172 families (Moxon et al., 2008; Karlova et al., 2013; Zuo et al., 2013). miRNA precursor genes are also regulated by many transacting factors (Rogers and Chen, 2013). Ethylene might be involved in the regulation of miRNA and also their corresponding precursor genes, such as TAS3-mRNA, miR156, miR159, miR160, miR164, miR171, miR172, miR390, miR396, miR4376 and miR5301 (Gao et al., 2015). RIN (ripening inhibitor) regulates tomato fruit ripening-related genes through of the post-transcriptional regulations of related genes via miRNA and ethylene. In addition, the ethylene can also regulate miRNA by modulating the abundance of mRNA (Gao et al., 2015). miRNAs specifically induced in response to biotic or abiotic stresses have also been identified and could be interesting targets for tomato adaptation (Jin et al., 2012; Cao et al., 2014; Liu et al., 2018; Sarkar et al., 2017; Shi et al., 2019). Though epigenome regulation is important during fresh fruit development and ripening, additional investigations about epigenome dynamics during fruit maturation and ripening or under environmental stresses are still needed (Giovannoni et al., 2017).

3.11 Databases

Databases are essential to access the wide range of data produced and shared on tomato. Tomato community has benefited for years of the will to gather genetic and later genomic data into one single free access database,

known as Solanaceae Genome Network, as the resource concern several Solanaceae species. Since the first RFLP genetic map, the database hosts information about markers, genes and QTL and now a genome browser where several genomes and SNP can be found. Several other databases can be useful to tomato geneticists. They describe genetic resources and mutant collections or information about gene expression (**Table 5**).

Table 5 Main databases useful for tomato genetics and genomics

Name	Address	Characteristics
Solanaceae Genome Network (SGN)	https://solgenomics.net	Central hub for sol genomics (genome sequences, loci, phenotypes …
Tomato Genetic Resource Center (TGRC)	https://tgrc.ucdavis.edu/	Charles Rick Tomato Genetic Resource Collection in UC Davis
Tomatoma	http://tomatoma.nbrp.jp/	Microtom mutants and genome archive
Mibase Tomato DB	http://www.kazusa.or.jp/jsol/microtom	Microtom genomic resources
SolCAP	http://solcap.msu.edu/	SNP, genotype and phenotypes
Tomato Expression Database	http://ted.bti.cornell.edu/	Gene expression analysis results
Tomato Expression Atlas	http://tea.solgenomics.net/	High resolution map of gene expression
Tomexpress	http://tomexpress.toulouse.inra.fr/	RNAseq data
Tomato EFP browser	http://bar.utoronto.ca/efp_tomato	Tomato gene expression viewer
Solcyc	http://solcyc.solgenomics.net/	Pathway/genome DB

4 Breeding for smart tomato

4.1 Traditional breeding

Tomato is a self-pollinated crop. The first varieties were landraces and the intensive breeding started in the 1930s in the USA. As a self-pollinated crop, for years tomato has been bred through a combination of pedigree and backcross selection. Very early, introgressions from wild species were proposed to introduce disease resistances but also to improve fruit firmness and other fruit quality traits (Bai and Lindhout 2007). Recurrent selection (successive rounds of selection and intercrossing of the best individuals) also proved efficient to simultaneously increase fruit sugar content and fruit size and break the negative relationship between both traits (Causse et al., 2007).

Although tomato exhibits a low heterosis for yield, F1 hybrid varieties progressively replaced the pure lines since the 1970s. This was first shown to be interesting for fruit shape and size homogeneity and then for combining several dominant resistance genes. Today F1 hybrids combine 6 to 8 disease resistance genes. For the production of F1 seeds, a set of nuclear recessive male sterility genes have been described, but are not used for a commercial purpose. The use of a functional male sterility gene, controlled by the positional sterile mutation (*ps2*) whose anthers do not naturally open, has been proposed (Atanassova, 1999). Nevertheless, due to the difficulty of carrying sterility genes along the selection schemes and to the rapid turnover of tomato cultivars, F1 hybrids are more frequently produced by hand pollination, in countries with low labor cost.

4.2 Marker-Assisted Selection

Many important loci have been mapped and tagged with molecular markers. Marker-Assisted Selection (MAS) allows breeders to follow genomic regions involved in the expression of traits of interest. The efficiency and complexity of MAS depend on the genetic nature of the trait (monogenic or polygenic). For monogenic traits, marker-assisted backcross (MABC) is the most straightforward strategy, whereas for polygenic traits various strategies are available.

4.2.1. Marker-Assisted Backcross for monogenic traits

The principle of MABC for a single gene is simple. First, molecular markers tightly linked to the target gene are identified, allowing the efficient detection of the presence of the introgressed gene ("foreground selection"). Other markers may be also used in order to accelerate the return to the recipient parent genotype at other loci ("background selection"). Background selection is based not only on markers located on the chromosomes carrying the gene to introgress (carrier chromosome), but also on other chromosomes. Markers devoted to background selection on a carrier chromosome allow the identification of individuals for which recombination events took place on one or both sides of the gene, in order to reduce the length of the donor type segment of genome dragged along with the gene (Young and Tanksley, 1989). In three generations of MABC, isogenicity is higher than that obtained by classical methods. By comparison, traditional approach would require approximately two more generations to obtain such an isogenicity (Hospital et al., 1992). Many important genes have been mapped or even cloned and specific markers for favorable alleles developed (Rothan et al., 2019 for a

recent review). Today, tomato breeders use molecular markers for the introgression of several monogenic traits such as disease resistances or fruit specific traits. The reduction of the cost of genotyping allows today the screening of a large number of plants to accelerate the selection process.

4.2.2. Marker-assisted selection for QTLs

Traits showing a quantitative variation are usually controlled by several QTLs, each with different individual effect. Due to the genetic complexity of such traits, several QTLs with limited effects must be simultaneously manipulated. Depending on their number, the nature and range of their effect, the origin of favorable alleles, different MAS strategies were proposed.

As for monogenic traits, MABC is the most effective strategy when a small number of QTLs, coming all from the same parent, must be transferred into an elite line. Hospital and Charcosset (1997) determined the optimal number and positions of the markers needed to control the QTLs during the foreground selection step and the maximum possible number of QTLs that could be simultaneously monitored with realistic population sizes (a few hundred individuals). In average, using at least three markers per QTL allows a good control over several generations, providing a low risk to have the donor type alleles at the markers without having the desired genotype at the QTL. However, as the minimum number of individuals that should be genotyped at each generation depends on (i) the confidence interval length, (ii) the number of markers and (iii) the number of QTLs, it seems illusive to transfer more than four or five QTLs with this simultaneous design unless a very large population can be considered, or the precision of the QTL location is very high.

After the identification of QTL for fruit quality traits (Saliba-Colombani et al., 2001; Causse et al., 2001), several clusters of QTLs were identified. As most of the favorable alleles for quality improvement came from the cherry tomato parental line, a MABC scheme has then been set up in order to transfer the five regions of the cherry tomato genome with the largest effects on fruit quality into three recurrent lines (Lecomte et al., 2004b). The population size allowed a successful transfer of the five segments into each recurrent line, and the MAS scheme allowed reducing the proportion of donor genome on the non-carrier chromosomes under the level expected without selection. Plants carrying from one to five QTLs were selected in order to study their individual or combined effects. Most of the QTLs were recovered in lines carrying one introgression region and new QTLs were detected (Causse et al., 2007). Introgressed lines had improved fruit quality, in comparison to parental lines, promising a potential improvement. Nevertheless, fruit weight in these genotypes was always lower than expected due to the effect of unexpected QTLs, whose effect was masked in the RIL population, suggesting that negative alleles at fruit weight QTLs were not initially detected.

4.2.3 Advanced backcross for the simultaneous discovery and transfer of new alleles

The advanced backcross QTL analysis is another strategy tailored for the simultaneous discovery and transfer of valuable QTL alleles from unadapted donor lines into established elite inbred lines (Tanksley and Nelson, 1996). The QTL analysis is delayed until an advanced generation (BC_3 or BC_4), while negative selection is performed to reduce the frequency of deleterious donor alleles during the preliminary steps. The use of BC_3 / BC_4 populations reduces linkage drag by reducing the size of introgressed fragments, limits epistatic effects and decreases the amount of time later needed to develop near isogenic lines carrying the QTL (Fulton et al., 1997). Tanksley and colleagues have applied this strategy for screening positive alleles in 5 wild species, *S. pimpinellifolium* (Tanksley et al., 1996), *S. habrochaites* (Bernacchi et al., 1998a), *S. peruvianum* (Fulton et al., 1997), *S. pennellii* (Eshed et al., 1996) et *S. parviflorum* (Fulton et al., 2000). They identified a number of important transgressions potentially useful for processing tomato and demonstrated that beneficial alleles could be identified in unadapted germplasm and simultaneously transferred into elite cultivars, thus exploiting the hidden value of exotic germplasm (Bernacchi et al., 1998b, Tanksley and Nelson, 1996).

4.2.4 Pyramidal design

When the number of QTLs to introgress becomes important, Hospital and Charcosset (1997) proposed to use a pyramidal design. QTLs are first monitored one by one by MABC, to benefit from higher background selection intensity, and then the selected individuals are intercrossed, to cumulate favorable alleles at the QTLs in the same genotype. When favorable alleles come from different sources, van Berloo and Stam (1998) proposed an index method to select among recombinant inbred lines those to be crossed, to obtain a single genotype containing as many favorable quantitative trait alleles as possible. Plants showing the optimal index are crossed together. This strategy was shown efficient to obtain transgression in offspring populations of *Arabidopsis* (van Berloo and Stam, 1999).

The benefit of MAS for QTL pyramiding was shown but limited by the number of QTL easily managed (Lecomte et al., 2004b; Gur and Zamir 2015; Sacco *et al.* 2013). This can be overcome by fine mapping experiment and/or validating the QTL effect in other backgrounds (Lecomte et al., 2004a). Today SNP availability and genomic selection open new ways to marker-assisted selection for quantitative traits.

4.2.5 Breeding for resistance to pests and pathogens

Despite decades of conventional breeding and phenotypic selection, there are still a large number of pests and pathogens that make tomato production challenging in various parts of the world. It is why the most prominent issue of tomato breeding remains pest and pathogen resistance. Current advances in tomato genetics and genomics can be combined with conventional plant breeding methods to introgress resistance loci or genes and expedite the breeding process.

Phenotypic (*e.g.* sensitivity to the Fenthion insecticide linked to resistance to *Pseudomomas syringae* pv. *Tomato* (Laterrot and Moretti 1989)), enzymatic (*e.g.* Aps-1^1 linked to root knot nematode resistance (Aarts et al. 1991; Messeguer et al. 1991)) and DNA markers tightly linked to resistance loci have long been used for MAS to incorporate resistance loci in new tomato cultivars. MAS is valuable for increasing the efficiency of selection, particularly when it is difficult to perform disease resistance assay, for instance with quarantine pathogens requiring controlled experimental infrastructures, and when disease resistance is controlled by recessive genes, or when genes display a weak penetrance or are strongly influenced by environment. Markers help to carry on a more efficient and precise introgression of the targeted loci, reducing the negative effects of linkage drag. MAS has also permitted to pyramid several resistance loci with other desirable traits. Because most of resistance genes are clustered on the tomato genome, introgression of resistance traits by phenotyping selection or by using MAS with markers at both sides of the major resistance gene permitted to introgress a kind of cassettes of resistance alleles when they are in coupling linkage and to create multi-resistant cultivars. For instance, most of *Tm-2²* tomato cultivars hitchhiked the *Frl* gene responsible for the Fusarium crown and root rot resistance caused by FORL (Foolad and Panthee 2012). Inversely, when resistance alleles are linked in repulsion phase, breeding selection may be hindered by the difficulty to select for homozygous coupling-phase recombinant lines, as illustrated for the association of *Sw-5* and *Ph-3* (Robbins et al. 2010). Thanks to MAS, the rate of improvement has been significantly enhanced in tomato even if many challenges remain.

Nowadays, DNA markers have been made available for about 30 genes controlling single gene inherited resistance traits important for tomato breeding (https://solgenomics.net/; Foolad and Panthee 2012). DNA markers for complex inherited resistance traits are much less abundant and they have rarely been used. MAS is thus routinely employed for selecting major effect resistance genes (*I*, *I-2*, and more recently *I-3*, *Ve*, *Mi-1.1*/*Mi1.2*, *Asc*, *Sm*, *Pto*, *Tm-*2^2, *Sw-5*) and many commercial cultivars now are resistant to *Fusarium oxysporum f. sp. lycopersici*, *Verticillium dahlia*, *Meloigogyne incognita*, *Alternaria alternata* f.sp. *lycopersici*, *Stemphyllium*, *Pseudomonas syringae* pv. *tomato*, ToMV and TSWV. Also markers for *Rx-3* and *Rx-4*, and for *Ty-1*, *Ty-2*, *Ty-3*, *Ty-4* are more and more used to deliver resistant cultivars to *Xanthomonas* spp. and TYLCV.

Although markers have been identified for many disease resistances in tomato, not all of them are useful because of absence of polymorphism within breeding populations that are often based on intraspecific crosses or because markers are too far from genes or QTLs of interest permitting unwanted crossing-overs. However, advances in next generation sequencing make possible to identify linked SNPs from which new PCR-based markers can be developed for trait association within breeding populations. The whole plant genome technologies greatly help to identify useful markers linked to resistance traits within the wild germplasm by eco-tilling, allele mining, or GWAS. Tomato breeders are thus now able to select the best combinations of genotypes to inter-cross in order to associate favorable traits and design elite ideotypes.

4.3 Genomic selection

Many traits are controlled by a large number of QTL with low effect. Both linkage mapping and GWAS have limitations in identifying and quantifying small effect and also rare QTLs or associations that are highly susceptible to environmental conditions (Crossa et al., 2017). In contrast, genomic selection (GS), which has been proposed for about two decades (Meuwissen et al., 2001; Crossa et al., 2017) uses all the genetic information from markers spread over the whole genome, such as SNPs and phenotypic data, in a training population, to predict the genetic estimated breeding values (GEBVs) of unphenotyped individuals in a test population. The main advantages of GS include cost reduction and time saving compared to phenotype-based selection (Crossa et al., 2017).

Several factors influence the accuracy of genomic prediction (GP), including the size, structure and genetic diversity of the training population, trait heritability, the number and distribution of molecular markers, linkage disequilibrium, prediction method and number of QTLs (Isidro et al., 2015; Spindel et al., 2015; Duangjit et al., 2016; Kooke et al., 2016; Yamamoto et al., 2016; Boison et al., 2017; Crossa et al., 2017; Minamikawa et al., 2017; Müller et al., 2017; Yamamoto et al., 2017; Crain et al., 2018; Edwards et al., 2019; Mangin et al., 2019; Sun et al., 2019). In order to improve the prediction accuracy, complex GS models were developed in order to handle different factors, such as the multi-trait and multi-environment G × E interactions (Montesinos-López et al., 2016; Fernandes et al., 2018). To date, many models for GS are available and the prediction accuracy vary according to traits and conditions (Heslot et al., 2012; Jonas and de Koning, 2013; Yamamoto et al., 2016; Yamamoto et al., 2017).

The first GS test in tomato was focused on a simulation-based breeding design and phenotypic prediction, where a theoretical method was proposed to apply GS to actual breeding schemes of simultaneous improvement of yield and flavor (Yamamoto et al., 2016). Briefly, 96 big-fruited tomato varieties were selected and 20 agronomic traits were measured, which can be divided into four categories, including yield, quality,

physiological disorder of fruit and others, with the broad-sense heritability ranging from 0.10 to 1.00. Seven GP models were compared, including five linear methods, Ridge regression (RR) (Endelman, 2011), Bayesian Lasso (BL) (Park and Casella, 2008), extended Bayesian Lasso (EBL) (Mutshinda and Sillanpää, 2010), weighted Bayesian shrinkage regression (wBSR) (Hayashi and Iwata, 2010), and Bayes C (Habier et al., 2011), and two nonlinear methods, reproducing kernel Hilbert space regression (RKHS) (Gianola and Kaam, 2008) and random forest (RF) (Breiman, 2001). The highest prediction accuracy for different traits varied and the accuracy of Bayes C was highest for up to eight traits, ranking the best among all models. Some individuals with high GEBV of total fruit weight and soluble solid contents were selected as parents to simulate later generations. Simulations demonstrated that after five generations, the simulated GEBVs were comparable with parental varieties. Breeding selections of target traits could also have impacts on some non-target traits. In particular, simultaneous selection for yield and flavor resulted in morphological changes, such as the increase in plant height. These results demonstrated the benefits of simulations for real breeding design.

Yamamoto et al., (2017) then used big-fruited F1 population to construct the GS models to assess its potential for the improvement of total fruit weight and soluble solid content in a practical experiment. By testing six GS models and 10-fold cross-validation, the prediction accuracy for soluble solid content was higher than for total fruit weight. GBLUP and BL had significantly higher predictability compared to other models for soluble solid content. In contrast, RKHS and RF had significantly higher predictability compared to other linear models for total fruit weight. The authors further developed four progeny populations to predict trait segregations and demonstrated that all individuals in the four progeny populations were genetically distinct from each other but intermediate between their parental varieties. However, the genetic diversity within each population was much lower compared to the training population.

Duangjit et al., (2016) investigated the impacts of some key factors on the efficiency of GP, including the size of training population, the number and density of SNPs and individual relatedness. Based on the analysis of 163 tomato accessions, the optimal size of the training population was 122. The prediction accuracy also increased with the increase of marker density and number, but weakly. Individual relatedness also influenced the prediction accuracy, and predictions were better in closer individual relatedness. However, there are some limitations in this study: 1) it only tested the ridge regression best linear unbiased prediction (rrBLUP) statistical model (Endelman, 2011); 2) the number of SNPs was relatively small and the genomic coverage in certain genomic regions was quite limited (Zhao et al., 2019); 3) Population structure existed and the number of wild accessions was quite small compared to cherry and large-fruited tomato accessions.

Most of the GS models rely on marker-based information and are unable to exploit local epistatic interactions among markers. Molecular markers can also be combined into haplotypes by combining linkage disequilibrium and linkage analysis to improve prediction accuracy (Clark, 2004; Calus et al., 2008; Jiang et al., 2018), which has been recently shown especially in animals (Calus et al., 2008; Cuyabano et al., 2014; Cuyabano et al., 2015a; Cuyabano et al., 2015b; Hess et al., 2017; Karimi et al., 2018). Haplotype-based genome-wide prediction models make it possible to exploit local epistatic effects inside haplotype blocks (Wang et al., 2012; de Los Campos et al., 2013; He et al., 2016; Jiang et al., 2018). The benefits of haplotype-based GS remain to be investigated in major crops (Jiang et al., 2018).

Genomic selection should permit to breed for a combination of traits related to qualitative resistance to biotic stresses as well as quantitative resistance and tolerance to biotic and abiotic stress combinations in considering also the genetic architecture of yield and fruit quality related traits. Both foreground and background selection should promote a sustained performance under diverse changing environments. Until now, disease quantitative resistance does not seem to be actively pursued by breeders because the complex polygenic control has generally hampered a wide deployment of QTL introgression. The development of post-genomics should help to foster tomato breeding for multiple polygenic traits including multi-resistance to pests and pathogens.

5 Designing ideotypes by ecophysiological modelling

Until the 1970s, genetic advances have favored the creation of high-yielding varieties adapted to mechanized and high-input production systems. Since the 90s, the context of global change instigates to renew the breeding goals by taking into account multiple environmental, economic and social issues. These multidisciplinary and integrative approaches have combined genetics and ecophysiology or agronomy skills, taking into account the mechanisms linking phenotypes to genotypes, and their modulation by the environment (essentially defined by soil, climate and pests) and cultural practices. Such approaches have allowed for a meaningful assessment of genotype-environment interactions and plant performances in terms of yield, quality and environmental impact in current production contexts. They have also made it possible to combine genetic information (available through the emergence of genetic and genomic tools) with phenotypic traits that determine variables of agronomic interest. In this context, the notion of ideotype has progressively developed to design plants able to perform in a given production context and finally to define breeding targets. To this end, process-based predictive models have proven their efficiency to unravel the mechanisms behind genetic variability of complex traits (Reymond et al., 2003; Tardieu, 2003, Yin et al., 2010; Quilot et al., 2005; Struik *et al.* 2005), to analyze

Genotype x Environment x Management (GxExM) interactions (Génard *et al.* 2007; Bertin *et al.* 2010; Martre *et al.* 2011), or to design new ideotypes adapted to specific environments (Kropff *et al.* 1995; Quilot *et al.* 2016; Martre *et al.* 2015; Génard *et al.* 2016).

5.1 What is an ideotype?

The ideotype concept, first proposed for wheat and then extended to several domesticated crops, is 'a theoretical biological model which is expected to perform or behave in a predictable manner within a defined environment' (Donald, 1968). Martre *et al.* (2015) extended the ideotype definition, to 'the combination of morphological and physiological traits (or their genetic bases) conferring to a crop a satisfying adaptation to a particular biophysical environment, crop management, and end use'.

Application for breeding may be straightforward for monogenic traits such as some biotic stress resistance. For instance, Zsögöna *et al.* (2017) proposed to take advantage of genome editing techniques in order to tailor such monogenic traits in cultivated cultivars or, on the opposite, to manipulate yield-related traits in wild relatives harboring polygenic stress resistance. Things are more complicated in case of traits with polygenic basis, for which geneticist has to face major issues. One of them is the complexity of some selection targets, such as yield, quality, nitrogen use-efficiency or adaptation to water deficit, etc. Indeed, these traits result from numerous nested processes with feedback effects and therefore, they are controlled by many genes. Another issue lies in the fact that the expression of these characters also depends on the environment and farming practices. This often results in strong GxExM interactions that make genetic work and their breeding application difficult. In a first empirical approach, optimal combinations of traits adapted to one specific environment and production system could be easily designed. For extrapolation to many different contexts, process-based predictive models may play a major role as discussed below (Quilot *et al.* 2012; Génard *et al.* 2016).

5.2 Current process-based models of tomato for the prediction of GxExM interactions

The plant and its organs can be seen as complex systems in which many processes interact at different scales under the control of GxExM interactions. Process-based predictive models are formal mathematical descriptions of this system and they have the potential to mimic its complexity in interaction with the environment, by integrating processes at several organizational levels (from cell to plant). The so-called component traits, which are underlying the predicted complex traits, are characterized in terms of model parameters, which instead of the complex trait itself, may subsequently be linked to underlying genetic variations (Struik *et al.* 2005; Bertin *et al.* 2010). This usually consists in forward genetics approaches such as QTL-mapping, in which one searches for co-localisations between QTL for traits and QTL for model parameters (e.g., Yin *et al.* 1999; Reymond *et al.* 2003; Quilot *et al.* 2005; Prudent *et al.* 2011; Constantinescu *et al.* 2016). Thus, a preliminary step is the identification of specific genotype-dependent parameters of the model in opposition to other generic parameters that do not vary among genotypes. Then each combination of genes or alleles is represented by a set of parameters and the phenotype can then be simulated *in silico* under various environmental and management conditions. In order to extend the range of prediction beyond known genotypes, it is necessary to estimate the values of the genotypic parameters depending on combinations of QTLs (QTL-based models), alleles or genes (gene-based models) involved in the modelled process (Martre *et al.* 2015). By formalizing each individual trait as a combination of genotypic and environmental effects, the model-based approach allows to detect more QTL that tend to be more stable than traditional QTL mapping. However, up to day, only few genotypic parameters (i.e. allelic variants) have been advantageously introduced into simulation models of tomato (Prudent *et al.* 2011; Constantinescu *et al.* 2016).

Several process-based simulation models that predict the processes underlying fruit growth and quality are now available and allow exploring the myriad of GxExM combinations (Génard and Lescourret, 2004; Bertin *et al.* 2010; Martre *et al.* 2011; Kromdjik *et al.* 2014). For tomato, several plant models are driven by processes of carbon assimilation and allocation among sinks according to different rules of priority (Heuvelink and Bertin, 1994; Jones *et al.* 1991; Boote 2016; Fanwoua *et al.* 2013), while only a few models simulate the water transfer and accumulation. For instance, Lee (1990) considers a unidirectional and constant flux of water uptake and transpiration per unit of fruit area. Bussières (1994) developed a model of water import in tomato fruit, based on water potential gradients and resistances. Yet, only rare models of fruit growth integrate both dry matter and water accumulation within the fruit. A virtual fruit model developed for peach (Fishman and Génard, 1998) has been adapted to predict processes involved in tomato fruit growth and composition (Liu *et al.* 2007). This model relies on a biophysical representation of one big cell, in which sugars are transported from the fruit's phloem by mass flow, diffusion and active transport. Incoming water flows are regulated, in particular, by differences in water potential and growth is effective only when the flow balance induces a sufficient turgor pressure on the cell walls. These models have been further modified and coupled to a stem model to estimate the contribution of xylem and phloem (Hanssens *et al.* 2015) and evaluate the effect of crop load on fruit growth (De Swaef *et al.* 2014).

The Virtual Fruit model has been also combined with a structural plant model to predict water and carbon allocation within the plant architecture, as well as the induced gradients of water potential and phloem sap concentration in carbon (Baldazzi *et al.* 2013). Because the cell level is the elementary level for mechanistic

modeling of fruit (Génard *et al.* 2007), a crucial issue is to model the way cell division and expansion developmentally progress (Baldazzi *et al.* 2012; Okello *et al.* 2015). The rare models of tomato fruit, which integrate cell division, cell expansion and DNA endoreduplication, have been used to better understand the emergence of fruit size and cell distribution (Fanwoua *et al.* 2013; Baldazzi *et al.* 2017; 2019). A virtual fruit model that predicts interactions among cell growth processes would be able to integrate sub-cellular models (Beauvoit *et al.* 2018), such as the ones proposed for tomato fruit to describe metabolic shifts during fruit development (Colombié *et al.* 2015, 2017) and pericarp soluble sugar content based on enzyme activity and compartmentation (Beauvoit *et al.* 2014). Indeed, except for sugar metabolism (Prudent *et al.* 2011), there is still a lack of predictive models of fruit composition, which is a major issue for fruit quality. For instance, no mechanistic model predicts the main compounds involved in tomato health value, like carotenoids, polyphenols or vitamins, which deserves further development. Such models exist for peach acidity (Lobit *et al.* 2003; 2006) and could be tailored to tomato.

Such integrated models centered on the fruit, integrating cellular processes and connected to a plant model open major perspectives to integrate information on the molecular control of fruit growth and composition regulations and to analyze the effects of GxExM interactions on yield and quality (Martre *et al.* 2011). Indeed, integrated models are important tools to phenotype plant *in silico*. They do not only allow to predict plant and organ traits such as yield or fruit composition, but also to asses physiological variables that are not easily measured on large panels such as xylem and phloem fluxes, active sugar transport… (Génard *et al.* 2007). So, process-based models enable to better understand genetic variability and identify candidate genes. They can also assist breeders to identify the most relevant traits and appropriate developmental stages to phenotype plants, and provide necessary links between genotype and phenotype in a given environmental context (Struik *et al.* 2005).

5.3 Process-based models design of tomato ideotypes

An important issue of simulating GxExM interactions is the *in silico* design of ideotypes, i.e. combinations of QTL/genes/alleles relevant to optimize fruit growth and quality under specific conditions, by multi-criteria optimization methods (Quilot-Turion *et al.* 2016). Therein lies the interest of process-based predictive models for developing breeding strategies.

A process-based model breeding program could break down into 3 successive steps (**Figure 6**): the first step consists in determining the values of the genetic coefficients of the model that makes it possible to obtain the desired characters for the ideotypes (virtual phenotype), in a given context of production (for instance low water supply, plant pruning…). The second step is to assess the values of the genetic coefficients from the genetic point of view (virtual genotypes), which requires identifying the combinations of alleles associated with each genetic coefficient. The last step is either to search among the existing genotypes for those that are the closest to the ideotype defined for a given environment, or to propose breeding strategies to obtain new genotypes on the basis of these ideotypes. For this last step, process-based models can be coupled with genetic models accounting for the genetic architecture of the genetic coefficients to simulate the genotypic changes that are expected to occur during the breeding program. Quilot-Turion *et al.* (2016) further proposed to add genetic constraints to improve ideotype realism and to optimize directly the alleles controlling the parameters, taking into consideration pleiotropic and linkage effects. This approach enabled reproducing relationships between parameters as observed in a real progeny and could be very useful to find out the best combinations of alleles in order to improve fruit phenotype in a given environment.

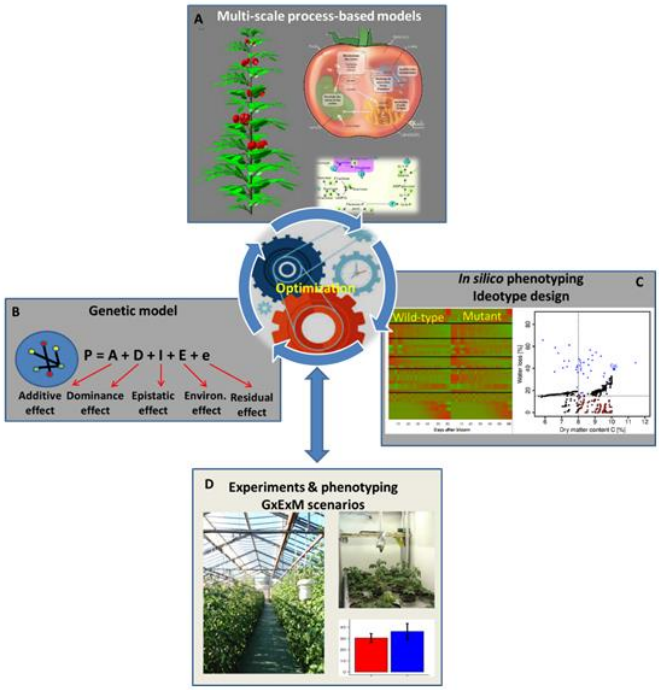


Figure 6. Overall scheme of the process-based design of tomato ideotypes. Plant and organ phenotypes measured in controlled environment or phenotyping platforms under different GxExM combinations (D) can be predicted by coupling process-based models that describe water and carbon fluxes in the plant, growth processes and primary and secondary fruit metabolism (A). On the right, figure (C) illustrates the use of the coupled model for phenotyping plant and fruits and for designing ideotypes. The heatmap shows the effect on all the simulated processes of a virtual mutation controlling one genetic parameter of the model, while the plot shows the position of ideotypes generated by the model according to fruit dry matter content and fruit water loss due to water deficit. On the left (B), the genetic model is dependent on several effects, which control the genotypic parameters of the process-based models in (A). The genetic model enables to predict the genotype of ideotypes selected in (C). The optimization procedure applies both to estimate the genotypic parameters of the models and to design the ideotypes.

Despite clear benefits and perspectives, only a few tomato ideotypes have been designed through modeling. Using a static functional structural plant model, Sarlikioti *et al.* (2011) looked for optimal plant architecture of greenhouse-grown tomato with respect to light absorption and photosynthesis. They concluded that an ideotype with long internodes and long and narrow leaves would improve crop photosynthesis. A second example based on the virtual fruit model of tomato described above (Constantinescu *et al.* 2016), suggested that a successful strategy to maintain yield and quality of large fruit genotypes under water deficit conditions could be to combine high pedicel conductance and high active uptake of sugars. Through the model calibration, the authors could identify some genotypes of the studied population, which were close to the ideotypes and thus, which may bring interesting traits and alleles for breeding plant adapted to low water supply.

As seen above, predictive models used for the design of ideotypes are expected to be highly mechanistic and detailed, therefore very complex, often combining different scales of description. Model parameters are ideally measured through adequate phenotyping, or more currently estimated through model calibration. Yet, a major difficulty is their parameterization based on extensive and heavy experiments on large genetic panels, which is rather prohibitive (Cournède *et al.* 2013). Similarly, the prediction of model parameters from QTL, alleles or genes relies on a calibration step that also suffers from the relatively limited number of parameterized genotypes (Letort *et al.* 2008; Migault *et al.* 2017). Instead of measuring extensive sets of physiological traits on all genotypes of the studied population, one can select a set of genotypes that well represents the genetic diversity and then predict the parameters for the whole selection of genotypes by QTL or genomic prediction models (van Eeuwijk et al., 2019). Alternatively, a representative training set of genotypes can be selected based on relevant morpho-physiological traits for estimating model parameters, as done in Constantinescu *et al.* (2016). From the mathematical point of view, the design of ideotypes is complex and relies on multi-objective optimization methods, which are complex due to dimensional problem (increasing number of genotypes and variables) and to the fact that ideotypes usually combine antagonistic nonlinear traits, such as yield and quality for tomato fruit. To solve the optimization problems large panels of meta-heuristics exist, based on different algorithms that can provide satisfactory solutions in a reasonable amount of time (Ould-Sidi and Lescourret. 2011). These methods can also apply to the model calibration step.

Our ability to phenotype large panels has increased in the last decades, with the emergence of high throughput genotyping and phenotyping platforms that generate large datasets on plant morphology and physiology at high temporal and spatial resolution. The way phenotyping information can be advantageously incorporated in different classes of genotype-to-phenotype models has been recently illustrated for field crops (van Eeuwijk *et al.* 2019). However, in case of tomato and other horticultural plants, the range of phenotyped traits should go

well beyond the traits that are routinely measured on such platforms, for instance by including fruit growth and composition alongside with plant and fruit development.

5.4 Prospects on the use of model-based plant design

Model-based design of plants offers promising opportunities for both crop management and breeding of plants able to cope with different environments and to answer multiple objectives. Tomato is particularly relevant for such approach. Its sequenced genome, the large number of genetic resources, available process-based models integrating process-networks at different organization levels, a strong societal demand for high quality fruits are all key-assets for the successful design of tomato ideotypes. Yet, some progress is still necessary. The integration of cellular and molecular levels can help refine plant models, and shed light onto the complex interplay between different spatial and temporal scales that control the traits of interest. For this, small networks of genes involved in the modelled processes might be helpful, as they could boost our capacity to link process-based model parameters to their genetic basis.

While the proof of concept is validated, it is clear that up-to-date, rare or no plant improvement has grounded in *in silico* design of ideotypes. To this end, closer collaborations among modelers, agronomists, geneticists and breeders are necessary to combine approaches and in particular to couple process-based models and genetic models of tomato. Furthermore, the development of new process-based sub-modules predicting important tomato quality traits such as texture, carotenoid, polyphenol and vitamin contents will be essential.

Finally, we could question the dominant paradigm according which genetic improvement relies on gene pyramiding. Indeed, stacking multiple genes in one variety might efficiently increase multiple resistances to biotic stresses, but may fail for other traits depending on the number of genes and their genetic architecture, the nature of germplasm... etc (Kumar *et al.* 2016). Instead, a new issue could be to bet on multi-genotype crops to stabilize their performances and reduce the inputs. This will require better understanding interactions among genomes within a population.

6 Biotechnology and Genetic engineering

6.1 A brief history of genetic engineering in tomato

According to the annual report of ISAAA (International Service for the Acquisition of Agri-biotech Applications) of 2017, 17 million farmers in 24 countries planted 189.8 million hectares biotech/GM crops. In 22 years, the planted area increased over 100 times. Nowadays there is no genetic engineered tomato available in market, whereas the first genetically engineered and commercialized food has been tomato, with a cultivar named FLAVR SAVRTM, which was approved by FDA (USA) on May 18, 1994, and just 3 days later, was available in two stores. It was created by scientists in Calgene company via antisense RNA of polygalacturonase (PG), one of the most abundant protein that had long been thought to be responsible for softening in ripe tomatoes (Kramer et al., 1994). FLAVR SAVRTM showed 99% decrease of PG protein and significant decrease in softening during storage, and increased resistance to fungi, which normally infect ripe fruits, thus providing a longer shelf life. Scientists expected that this tomato could be vine-ripened for enhanced flavor, and still suitable for the traditional distribution system (Kramer et al., 1992). At the same year, Zeneca commercialized a tomato puree made from tomatoes silenced PG with sense gene, with improved viscosity and flavor, and reduced waste (Grierson, 2016). The success was not as expected. FLAVR SAVR was removed of the market in 1999. Later a dozen of genetic engineering events were registered up to 1999, but none of them was commercialized (**Table 6**). Since 2000, not any new transgenic tomato was registered (http://www.isaaa.org/gmapprovaldatabase/default.asp).

6.2 Toolkit for genetic engineering tomato

Tomato genetic transformation was initially established in the 1980s (McCormick et al. 1986). The primary mode of transformation is Agrobacterium-mediated procedures by incubating with tomato explants such as leaf, hypocotyl or cotyledon, followed by the regeneration of plants via shoot organogenesis from callus. Based on reported protocols and the review by Bhatia et al. (2004), a general genetic engineering program for tomato requires (**Figure 7**):

1) Vectors to deliver engineering modules into agrobacteria and plants;
2) Integration of the introduced engineering modules into the genome for stable transformation;
3) *In vitro* regeneration and selection of transformed plants.

The effective transformation and regeneration are prerequisite steps for utilizing genetic engineering. Transformation efficiency is strongly dependent on the genotype, explant and plant growth regulators in the medium (reviewed by Gerszberg et al., 2015).

Successful transformation can also be performed either by dipping developing floral buds in the Agrobacterium suspension or by injecting Agrobacterium into the floral buds. Yasmeen et al. (2009) observed a high transformation frequency, 12% to 23% for different constructs, while for Sharada et al. (2017), a much lower transformation efficiency (0.25–0.50%) was obtained on floral dips/floral injections. Unlike in Arabidopsis, for

which flower-dipping method became a widely used transformation way (Clough et al., 1998), in tomato, this methodology has not been efficient.

Gene silencing or expression of heterologous genes in tomato have been used for decades in research. Different from those two conventional genetic engineering methods, genome editing based on CRISPR/Cas9 (clustered regularly interspaced short palindromic repeats) was first proposed on tomato a few years ago (Brooks et al., 2014), but rapidly showed a large potential and wide application for functional gene characterizing, breeding and domestication.

Table 6. Transgenic tomato varieties approved for commercialization, reproduced from Gerszberg et al (2015)

Event	Developer	Traits	Year	Approved for	Country
FLAVR SAVR	Calgene	Delayed softening (developed by additional PG gene expressed)	1994	All uses in USA; Japan and Mexico for feed and for environment	USA
1345-4	DNA Plant Technology Corporation	Delayed ripening (developed by a truncated aminocyclopropane cyclase synthase gene)	1994	All uses in USA; food in Canada and Mexico	USA
Da,V,F tomato	Zeneca Seeds	Delayed ripening (developed by additional PG gene expressed)	1994	All uses in USA; food in Canada and Mexico	USA
8338	Monsanto Company	Delayed ripening (developed by introduction of 1-aminocyclopropane-1-carboxylic acid deaminase (accd) gene)	1995	All uses in USA	USA
351N	Agritope	Delayed ripening (developed by introduction the S-adenosylmethionine hydrolase (SAMK) gene)	1995	All uses in USA	China
Huafan No 1	Huazhong Agricultural University	Delayed ripening (developed by introduction anti-sense EFE gene)	1996	Data not available	China
5345	Monsanto Company	Insect resistant (developed by introduction of one cry1Ac gene)	1997	All uses in USA; food in Canada	USA
PK-TM8805R (8805R)	Beijing University	Delayed ripening	1999	Food, feed, cultivation in China	China

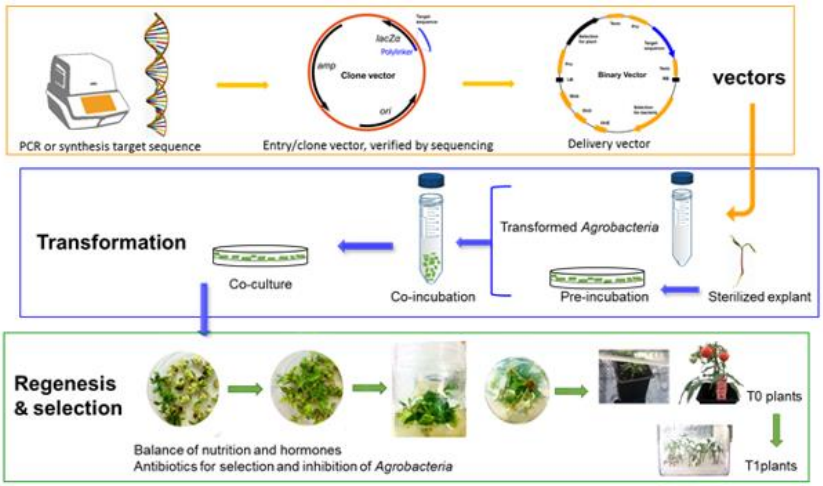


Figure 7. A general workflow for transformation based on widely used protocols. The target sequence could be obtained by PCR or commercial synthesis, and then different cloning methods used to transfer it into the clone vector. After verifying the clone vector, target sequence could be transferred to delivery vector, which is adapted for agrobacteria transformation. Tomato seeds are germinated in sterilized medium. When cotyledons appear, they are cut for pre-culture. After pre-culture, cotyledons (or other explants) are co-incubated with Agrobacteria that carry delivery vector and Ti plasmid, following a short period (such as 2 days) for co-culture. Then explants are transferred to medium suitable for regenesis and selection. For different step of regenesis, different nutrition and hormones are needed. When roots appear, transgenic plants are introduced to greenhouse. For T0 plants, the insertion of exogenous modules should be checked. The seeds of T0 plants are planted on medium with selection antibiotic for selecting the transgenic plants.

6.2.1 Gene silencing and homologous/heterologous expression

Gene silencing is usually obtained via antisense (as for FLAVR SAVR), sense or RNA interfering (RNAi). Scientists have used it to inhibit the unfavorable ripening/softening after tomato harvesting and during long distance transportation, to remove compounds stimulating allergies (Le et al., 2006), or block seed production resulting in parthenocarpic fruit (Schijlen et al., 2007). Inhibition or better control of fruit ripening and softening is still one of the major challenges for breeders and scientists for commercial perspectives. This purpose was achieved to different degrees by silencing different genes, including those coding pectin methylesterase (Tieman et al., 1994), expansin protein (Brummell et al., 1999), beta-galactosidase (Smith et al., 2002), ACC synthase (Gupta et al., 2013), transcription factor SlNAC1 (Meng et al., 2016), pectate lyase (Uluisik et al., 2016).

Different from gene silencing strategies which aim to down regulate endogenous genes of tomato, over expression of endogenous or exogenous genes can also be manipulated to study promoters and gene expression, enhance tolerance to biotic/abiotic stresses, and increase the accumulation of secondary metabolites… Promoters

(endogenous or exogenous) can be fused with GUS or florescent protein to follow the gene expression pattern. Fernandez et al. (2009) generated novel Gateway destination vectors based on the detailed characterization of series promoters' expression pattern during fruit development and ripening, facilitating tomato genetic engineering. Redox sensitive GFP (roGFP) was also developed to better study the *in vivo* redox state in tomato (Huang et al., 2014).

Researchers who work on perennial trees such as apple, peach, banana, et al., often used tomato to do heterologous expression of target genes to *in vivo* study the gene function, since the transformation and regeneration techniques are difficult to apply on those species and even when possible, it is time-consuming to pass juvenile phase to obtain fruit phenotypes. In return, the genes from other species, which showed a phenotype on tomato, can be interesting resources for genetic engineering. For instance, apple vacuolar H+-translocating inorganic pyrophosphatase (MdVHP1) overexpressed in tomato, improved tolerance to salt and drought stress (Dong, 2011). Overexpression of banana MYB TF MaMYB3 inhibited starch degradation and delayed fruit ripening (Fan et al., 2018).

Fusing abiotic-driven promoter with functional TF responding to abiotic stress was a promising strategy for improving stress tolerance. Transgenic plants with the transcription factor CBF driven by ABA-responsive complex (ABTC1) showed enhanced tolerance to chilling, water deficit and salt stresses without affecting the growth and yield under normal growing conditions (Lee et al., 2003).

The metabolism flux can also be altered to improve fruit qualities, such as volatiles and nutrition compounds. Domínguez et al. (2010) overexpressed genes coding ω-3 fatty acid desaturases, FAD3 and FAD7, resulting in an increase in the 18:3/18:2 ratio in leaves and fruit, and a significant alteration of (Z)-hex-3-enal/hexanal ratio. AtMYB12 under the fruit-specific E8 promoter was inserted into tomato genome, activating the genes related to flavonol and hychoxycinnamic ester biosynthesis, leading to an accumulation as much as 10% of fruit dry weight (Zhang et al., 2015).

In addition to those remarkable progresses of genetic engineering since 1980s, the most notable progress has been made since the emerging and development of genome-editing tools, such as CRISPR/Cas9.

6.2.2 Genome editing

Unlike genome editing tools, Zinc-finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs), which are based on protein-DNA recognition, CRISPR/Cas9 relies on simple RNA-DNA base pairing and the PAM (protospacer adjacent motif) sequence recognition (Gaj et al., 2013). All these tools result in DNA double-strand breaks (DSBs), but CRISPR/Cas9 showed higher efficiency than ZFN and TALEN (Adli, 2018). DSB can be repaired either by error-prone non-homology end joining (NHEJ) or homology-directed repair (HDR). Organisms recruit NHEJ or HDR repairing system to induce indel mutations or precise substitution, resulting in knockout or precise-genome editing, respectively. Besides studying the mechanism of CRISPR/Cas9 genome-editing system, scientists also showed enthusiasm of re-engineering CRISPR/Cas9 tools to make them more flexible and increase their fidelity, via making Cas9 nucleases smaller, expanding the targeting scope, and decreasing the off-target rate.

In 2014, the first CRISPR/Cas9 case was reported in tomato (Brooks et al., 2014) and later scientists have explored CRISPR-based engineering on several topics. As CRISPR/Cas9 system can efficiently introduce knockout mutation, it is a useful method to characterize candidate genes from forward genetics or natural mutation. An elegant case of using CRISPR/Cas9 was the production of RIN knockout mutant, shedding light on an old topic. Tomato *rin* mutants remain firm after harvest and fail to produce red pigmentation and ethylene, thus RIN has long been believed to be indispensable for the induction of ripening. Ito et al. (2017) used CRISPR/Cas9 gene editing to obtain RIN-knockout mutant, which showed moderate red coloring, different from *rin*'s completely fail-to-ripening phenotype. Moreover, using CRISPR/Cas9 to edit *rin* mutant allele partially restored the induction of ripening. Therefore, they showed that RIN is not essential for the initiation of ripening and is a gain-of-function mutation producing a protein actively repressing ripening, rather than a null mutation. This technology has also been used on methylation/demethylation study. A DNA demethylase gene of tomato SlDML2 was mutated by CRISPR/Cas9 to generate loss-of-function mutants, showing a critical role of SlDML2 in tomato fruit ripening possibly via active demethylation of ripening induced genes and the inhibition of ripening-repressed genes (Lang et al., 2017).

Second generation of CRISPR gene-editing tools include base-editing, CRISPR-mediated gene expression regulation, CRISPR-mediated live cell chromatin imaging (Adli, 2018). The probability of gene insertion was increased by the production of landing pad (Danilo et al., 2018) as well as gene knock-in by precise base mutations (Danilo et al. 2019; Veillet et al. 2019). All these strategies are based on manipulation of Cas9, by turning nuclease Cas9 to nickase Cas9 (nCas9) or dead Cas9 (dCas9, catalytically inactive Cas9), but still keeping the capability to recognize specific sequences. The engineered Cas9 can be fused with other enzymes or proteins to enable base editing, gene regulation or chromatin imaging.

Shimatani et al. (2017) generated marker-free plants with homozygous heritable DNA substitutions by using D10A mutant nCas9At fused with either a human codon-optimized PmCDA1 (nCas9At-PmCDA1Hs) or a version codon-optimized for Arabidopsis (nCas9At-PmCDA1At). It should be mentioned that the offspring of

T0 generation also revealed indels, moreover the rate of substitution was much lower than the rate of indel mutation. It demonstrated the feasibility of base editing for crop improvement even though with a lower rate. Dreissig et al. (2017) showed visualization of telomere repeats in live leaf cells of *Nicotiana benthamiana* by fusing eGFP/mRuby2 to dCas9, and also DNA-protein interactions *in vivo* via combining CRISPR-dCas9 with fluorescence-labelled proteins. Researchers developed CRISPR interference (CRISPRi) approach with dCas9 binding activity blocking the transcriptional process and thus down regulating gene expressions (Qi et al., 2013).
CRISPR/Cas9 and related second-generation genome-editing tools increase the feasibility and enlarge the applicable scope of biotechnology. With those progresses and the conventional transgenic tools (RNAi, overexpression and so on), it allows comprehensive breeding to face multiple challenges towards increasing population and climate changes.

6.2.3 Comprehensive genomic engineering on tomato

Rodriguez-Leal et al. (2017) focused on three major productivity traits in tomato: fruit size, inflorescence branching, and plant architecture, and used CRISPR/Cas9 to do genome editing of promoters to generate several cis-regulatory alleles. They evaluated the phenotypic impact of those variants and provided an efficient approach to select and fix novel alleles controlling the quantitative traits.
Genome editing can also accelerate domestication, as shown by two groups. Li et al. (2018) selected four stress-tolerant wild-tomato accessions to introduce desirable traits by using multiplex CRISPR/Cas9 editing. They targeted coding sequences, cis regulatory regions or upstream open reading frames of genes associated with morphology, flower and fruit production, and ascorbic acid synthesis. The progeny of edited plants showed domesticated phenotypes yet retained parental disease resistances and salt tolerance. In the same time, Zsögön et al. (2018) chose wild *S. pimpinellifolium* as starting material to combine agronomically desirable traits with useful wild line traits via editing of six loci that are important for yield and productivity. Engineered tomatoes showed remarkable increase of fruit size, number, and lycopene content. As the researchers said, those impressive *de novo* domestication cases pave the way to exploit the genetic diversity present in wild plants.
Genome editing tools also show big potential for achieving tomato ideotype, for which the concept and design strategies have been explained in chapter 5. Recently Naves et al. (2019) proposed to engineer tomato to be the biofactory of secondary metabolites, such as capsaicinoids (the metabolites responsible of the burning sensation of hot pepper). Considering that tomato genome presented all the necessary genes for capsaicinoid production, two strategies, transcriptional activator-like effectors (TALEs), or genome engineering for targeted replacement of promoters were suggested to be used in tandem to activate capsaicinoid biosynthesis in the tomato (Naves et al., 2019).

6.3 Genetic engineering for improving pest and pathogen resistance

A few tomato diseases remains orphan, that is to say that no natural resistance genes or QTLs have been discovered yet. Moreover, although available from crop wild relatives, breeders may be unable to fully utilize the resistance genes from genetic diversity because of interspecific barriers or because of linkage drag associated to an introgression from a distant species. In that case, resistance might be engineered through biotechnology.
To circumvent the absence of natural resistance, transgenic technologies relying on RNA interference or expression of pathogen- derived sequence have been used to engineer resistance to a number of pathogens. Besides, the ectopic expression of resistance gene could enhance resistance as shown with the introgression of *pvr1*, a recessive gene from *Capsicum chinense*, in tomato that results in dominant broad- spectrum potyvirus resistance (Kang et al. 2007). Nekrasov et al. (2017) also created a transgene-free powdery mildew resistant tomato by genome deletion.
The CRISPR/Cas technology is also expected to accelerate the breeding of cultivars resistant to diseases. Recently, CRISPR/Cas9 system has been used to engineer tomato plants that target the TYLCV genome with Cas9-single guide RNA at the sequences encoding the coat protein (CP) or replicase (Rep) resulting in immunity against TYLCV (Tashkandi et al. 2018). In addition, although still in its infancy, gene-editing by CRISPR-nCas9-cytidine deaminase technology might be used to design *de novo* synthetic functional resistance alleles in tomato, using knowledge about the natural evolution of resistance genes in related species, as demonstrated by Bastet et al. (2019) in *Arabidopsis thaliana*.

6.4 Regulatory status of gene edited plants

Since 2013, CRISPR/Cas9 systems allowed considerable progress in plant genome editing, giving access to cost-effective and efficient transformation compared with previous technologies and making it rapidly accessible to many researchers. However, this emerging method is still developing and scientific efforts continue to be made in order to realize the full potential of the technology. It offers great opportunities, but also creates regulatory challenges. Concerns have been raised over the status of the plants produced by gene editing and classical GMOs as the technology generates transgene-free plants. Many plant breeders and scientists consider that gene-editing techniques such as CRISPR/Cas9 should be considered as mutagenesis, and thus be exempt from the GMO directive, because they can induce only changes of DNA sequences and not the insertion of foreign genes. But people opposed to GM organisms contend that the deliberate nature of alterations made through gene editing

means that they should fall under the GMO directive. In the U.S.A., Canada and several other countries, CRISPR/Cas induced mutations are exempt from GMO laws and regarded as equivalent to traditional breeding. In Europe, on 25 July 2018 the European Court of Justice (ECJ) ruled that gene-edited crops should be subject to the same regulations as conventional GMOs (Callaway, 2018). This may have strong consequences on the breeding developments in the different countries.

7 Conclusion and prospects

Tomato is a crop widely adapted to very different conditions. Subsequently it has to respond to many stresses. Molecular markers have permitted the dissection of the genetic bases of complex traits into individual components, the location of many genes/QTLs on chromosomes, which became accessible to selection. Molecular markers have also allowed breeders to access to wild species in a more efficient way than in the past. Exotic libraries, which consist of marker-defined genomic regions taken from wild species and introgressed onto the background of elite crop lines, provide plant breeders with an important opportunity to improve the agricultural performance of modern varieties. Several research consortiums (for genome sequencing, but also for the valorization of genetic resources and traditional varieties) were gathered to study tomato diversity and adaptation.

Since the availability of the reference genome many new resources (genome sequences, millions of SNPs), tools (databases, methodological tools) and methods (genome editing, crop modeling and genomic selection) became available and thus breeding should be more efficient.

Better knowledge of physiological processes, metabolic pathways, genes involved as well as the genetic variability of candidate genes, mutant identification and translational genetics may be used to go further. New growth conditions such as urban horticulture must be taken into account.

It will be important to combine the empirical approach of breeders based on an intimate knowledge of the tomato crop with the power of biotechnologies. Integration of related disciplines will be more and more important to (1) develop more efficient methods to evaluate the impact of environment on the crop, (2) enhance knowledge of the biochemical and molecular bases of the traits, and (3) better understand G x E and to increase the adaptation of new varieties to new conditions.

Some complex questions remain for research: how several stresses interact, how to deal with new pathogens and pests, root x rootstock interaction, reduction of fertilizers. Finally modeling can help taking into account these aspects and designing new ideotypes optimized to the adverse variable or optimal conditions.

References

Aarts J, Hontelez JGJ, Fischer P, Verkerk R, Vankammen A, Zabel P (1991) Acid phosphatase-11, a tightly linked molecular marker for root-knot nematode resistance in tomato - from protein to gene, using pcr and degenerate primers containing deoxyinosine. Plant Molecular Biology 16: 647-661

Abraitiene A, Girgzdiene R (2013) Impact of the short-term mild and severe ozone treatments on the potato spindle tuber viroid-infected tomato (*Lycopersicon esculentum* Mill.). Zemdirbyste-Agriculture 100: 277-282

Achuo EA, Prinsen E, Hofte M (2006) Influence of drought, salt stress and abscisic acid on the resistance of tomato to *Botrytis cinerea* and *Oidium neolycopersici*. Plant Pathology 55: 178-186

Adams SR, Cockshull KE, Cave CRJ (2001) Effect of Temperature on the Growth and Development of Tomato Fruits. Ann Bot 88: 869–877

Adato A, Mandel T, Mintz-Oron S, Venger I, Levy D, Yativ M, Domínguez E, Wang Z, De Vos RC, Jetter R, Schreiber L, Heredia A, Rogachev I, Aharoni A (2009) Fruit-surface flavonoid accumulation in tomato is controlled by a SlMYB12-regulated transcriptional network. PLoS Genet. e1000777. doi: 10.1371/journal.pgen.1000777

Adli M (2018) The CRISPR tool kit for genome editing and beyond. Nature communications 9(1):1911.

Agrama HA, Scott JW (2006) Quantitative trait loci for *tomato yellow leaf curl virus* and *tomato mottle virus* resistance in tomato. Journal of the American Society for Horticultural Science 131: 267-272

Ahmad A, Zhang Y, Cao X-F (2010) Decoding the epigenetic language of plant development. Mol Plant 3: 719–28

Albacete A, Cantero-Navarro E, Großkinsky DK, Arias CL, Balibrea ME, Bru R, Fragner L, Ghanem ME, González M de la C, Hernández JA, et al. (2015) Ectopic overexpression of the cell wall invertase gene CIN1 leads to dehydration avoidance in tomato. J Exp Bot 66: 863–878

Albert E, Duboscq R, Latreille M, Santoni S, Beukers M, Bouchet JP, Bitton F, Gricourt J, Poncet C, Gautier V, et al. (2018) Allele-specific expression and genetic determinants of transcriptomic variations in response to mild water deficit in tomato. Plant J. 96(3):635-650

Albert E, Gricourt J, Bertin N, Bonnefoi J, Pateyron S, Tamby J-P, Bitton F, Causse M (2016a) Genotype by watering regime interaction in cultivated tomato: lessons from linkage mapping and gene expression. Theor Appl Genet 129: 395–418

Albert E, Segura V, Gricourt J, Bonnefoi J, Derivot L, Causse M (2016b) Association mapping reveals the genetic architecture of tomato response to water deficit: focus on major fruit quality traits. J Exp Bot 67: 6413–6430

Albrecht E, Escobar M, Chetelat RT (2010) Genetic diversity and population structure in the tomato-like nightshades *Solanum lycopersicoides* and *S. sitiens*. Ann Bot 105: 535–554

Alian A, Altman A, Heuer B (2000) Genotypic difference in salinity and water stress tolerance of fresh market tomato cultivars. Plant Sci 152: 59–65

Allwood JW, De Vos RCH, Moing A, Deborde C, Erban A, Kopka J, Goodacre R, Hall RD (2011) Plant metabolomics and its potential for systems biology research: Background concepts, technology, and methodology, 1st ed. Methods Enzymol. doi: 10.1016/B978-0-12-385118-5.00016-5

Almeida J, Quadrana L, Asís R, et al. (2011) Genetic dissection of vitamin E biosynthesis in tomato. *J Exp Bot*. 62(11): 3781–3798.

Alpert KB, Tanksley SD (1996) High-resolution mapping and isolation of a yeast artificial chromosome contig containing fw2.2: a major fruit weight quantitative trait locus in tomato. Proc Natl Acad Sci U S A 93: 15503–7

Alseekh S, Fernie AR (2018) Metabolomics 20 years on: what have we learned and what hurdles remain? Plant J 94: 933–942

Alseekh S, Ofner I, Pleban T, Tripodi P, Di Dato F, Cammareri M, Mohammad A, Grandillo S, Fernie AR, Zamir D (2013) Resolution by recombination: Breaking up Solanum pennellii introgressions. Trends Plant Sci 18: 536–538

Alseekh S, Tohge T, Wendenberg R, Scossa F, Omranian N, Li J, Kleessen S, Giavalisco P, Pleban T, Mueller-Roeber B, et al. (2015) Identification and Mode of Inheritance of Quantitative Trait Loci for Secondary Metabolite Abundance in Tomato. Plant Cell 27: 485–512

Alseekh S, Tong H, Scossa F, Brotman Y, Vigroux F, Tohge T, et al. (2017). Canalization of tomato fruit metabolism. The Plant Cell 29(11), 2753–2765.

Ambros V (2004) The functions of animal microRNAs. Nature 431: 350–355

Andolfo G, Jupe F, Witek K, Etherington GJ, Ercolano MR, Jones JDG (2014) Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. Bmc Plant Biology 14

Anfoka G, Moshe A, Fridman L, Amrani L, Rotem O, Kolot M, Zeidan M, Czosnek H, Gorovits R (2016) Tomato yellow leaf curl virus infection mitigates the heat stress response of plants grown at high temperatures. Sci Rep 6: 19715

Apse MP, Aharon GS, Snedden WA, Blumwald E (1999) Salt tolerance conferred by overexpression of a vacuolar Na+/H+ antiport in Arabidopsis. Science 285: 1256–8

Arafa RA, Rakha MT, Soliman NEK, Moussa OM, Kamel SM, Shirasawa K (2017) Rapid identification of candidate genes for resistance to tomato late blight disease using next-generation sequencing technologies. PLoS One 12: e0189951

Archak S, Karihaloo JL, Jain A (2002) RAPD markers reveal narrowing genetic base of Indian tomato cultivars. Curr Sci 82: 1139–1143

Arms EM, Lounsbery JK, Bloom AJ, St. Clair DA (2016) Complex Relationships among Water Use Efficiency-Related Traits, Yield, and Maturity in Tomato Lines Subjected to Deficit Irrigation in the Field. Crop Sci 56: 1698

Ashrafi H, Kinkade MP, Merk HL, Foolad MR (2012) Identification of novel quantitative trait loci for increased lycopene content and other fruit quality traits in a tomato recombinant inbred line population. Mol Breed 30: 549–567

Ashrafi-Dehkordi E, Alemzadeh A, Tanaka N, Razi H (2018) Meta-analysis of transcriptomic responses to biotic and abiotic stress in tomato. PeerJ 6: e4631

Asins MJ, Albacete A, Martinez-Andujar C, Pérez-Alfocea F, Dodd IC, Carbonell EA, Dieleman JA (2017) Genetic analysis of rootstock-mediated nitrogen (N) uptake and root-to-shoot signalling at contrasting N availabilities in tomato. Plant Sci 263: 94–106

Asins MJ, Bolarín MC, Pérez-Alfocea F, Estañ MT, Martínez-Andújar C, Albacete A, et al. (2010) Genetic analysis of physiological components of salt tolerance conferred by Solanum rootstocks. What is the rootstock doing for the scion? Theor Appl Genet. 121: 105–115.

Asins MJ, Raga V, Roca D, Belver A, Carbonell EA (2015) Genetic dissection of tomato rootstock effects on scion traits under moderate salinity. Theor Appl Genet. 128: 667–679.

Atanassova B (1999) Functional male sterility (ps2) in tomato (*Lycopersicon esculentum* Mill.) and its application in breeding and seed production. Euphytica 107: 1, 13-21

Auerswald H, Schwarz D, Kornelson C, Krumbein A, Brückner B (1999) Sensory analysis, sugar and acid content of tomato at different EC values of the nutrient solution. Sci Hortic (Amsterdam) 82: 227–242

Bai Y, Lindhout P (2007) Domestication and breeding of tomatoes: what have we gained and what can we gain in the future? Ann Bot. 100(5):1085-94.

Bai YL, Huang CC, van der Hulst R, Meijer-Dekens F, Bonnema G, Lindhout P (2003) QTLs for tomato powdery mildew resistance (*Oidium lycopersici*) in *Lycopersicon parviflorum* G1.1601 co-localize with two qualitative powdery mildew resistance genes. Molecular Plant-Microbe Interactions 16: 169-176

Bai YL, Kissoudis C, Yan Z, Visser RGF, van der Linden G (2018) Plant behaviour under combined stress: tomato responses to combined salinity and pathogen stress. Plant Journal 93: 781-793

Bai YL, Pavan S, Zheng Z, Zappel NF, Reinstadler A, Lotti C, De Giovanni C, Ricciardi L, Lindhout P, Visser R, Theres K, Panstruga R (2008) Naturally occurring broad-spectrum powdery mildew resistance in a central American tomato accession is caused by loss of *Mlo* function. Molecular Plant-Microbe Interactions 21: 30-39

Baldazzi V, Bertin N, Jong H, Genard M (2012) Towards multiscale plant models: integrating cellular networks. Trends in Plant Science 17:728-736.

Baldazzi V, Génard M, Bertin N (2017) Cell division, endoreduplication and expansion processes: setting the cell and organ control into an integrated model of tomato fruit development. Acta Horticulturae, 1182

Baldazzi V, Pinet A, Vercambre G, Benard C, Biais B, Génard M (2013) In-silico analysis of water and carbon relations under stress conditions. A multi-scale perspective centered on fruit. Frontiers in Plant Science 4. doi: 10.3389/fpls.2013.00495.

Baldazzi V, Valsesia P, Génard M, Bertin N (2019) Organ-wide and ploidy-dependent regulations both contribute to cell size determination: evidence from a computational model of tomato fruit. Journal of Experimental Botany in press

Baldet P, Stevens R, Causse M, Duffe P, Buret M, Rothan C, Garchery C, Duffé P, Carchery C, Baldet P, et al. (2007) Candidate Genes and Quantitative Trait Loci Affecting Fruit Ascorbic Acid Content in Three Tomato Populations. Plant Physiol 143: 1943–1953

Baldwin E, Scott J, Shewmaker C, Schuch W (2000) Flavor trivia and tomato aroma: biochemistry and possible mechanisms for control of important aroma components. Hortscience 35: 1013–1022

Baldwin EA, Nisperos-Carriedo MO, Baker R, Scott JW (1991) Quantitative analysis of flavor parameters in six Florida tomato cultivars (*Lycopersicon esculentum* Mill). J Agric Food Chem 39: 1135–1140

Baldwin EA, Scott JW, Einstein MA, Malundo TMM, Carr BT, Shewfelt RL, Tandon KS (1998) Relationship between sensory and instrumental analysis for tomato flavor. J Am Soc Hortic Sci 123: 906–915

Ballester A-R, Bovy AG, Viquez-Zamora M, Tikunov Y, Grandillo S, de Vos R, de Maagd RA, van Heusden S, Molthoff J (2016) Identification of Loci Affecting Accumulation of Secondary Metabolites in Tomato Fruit of a *Solanum lycopersicum* × *Solanum chmielewskii* Introgression Line Population. Front Plant Sci 7: 1428

Bandillo N, Raghavan C, Muyco P, Sevilla MAL, Lobina IT, Dilla-Ermita C, Tung C-W, McCouch S, Thomson M, Mauleon R, et al. (2013) Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding.

Rice 6: 11
Bastet A, Zafirov D, Giovinazzo N, Guyon-Debast A, Nogué F, Robaglia C, Gallois J-L (2019) Mimicking natural polymorphism in eIF4E by CRISPR-Cas9 base editing is associated with resistance to potyviruses. Plant Biotechnology Journal doi: 10.1111/pbi.13096
Bauchet G, Causse M (2012) Genetic Diversity in Tomato (*Solanum lycopersicum*) and Its Wild Relatives. Genetic Divers Plants. Intech, doi: 10.5772/33073
Bauchet G, Grenier S, Samson N, Bonnet J, Grivet L, Causse M (2017a) Use of modern tomato breeding germplasm for deciphering the genetic control of agronomical traits by Genome Wide Association study. Theor Appl Genet 130: 875–889
Bauchet G, Grenier S, Samson N, Segura V, Kende A, Beekwilder J, Cankar K, Gallois J-L, Gricourt J, Bonnet J, et al. (2017b) Identification of major loci and genomic regions controlling acid and volatile content in tomato fruit: implications for flavor improvement. New Phytol 215: 624–641
Baxter CJ, Liu JL, Fernie AR, Sweetlove LJ (2007) Determination of metabolic fluxes in a non-steady-state system. Phytochemistry 68: 2313–2319
Beauvoit B, I Belouah, N Bertin, C Belmys Cakpo, S Colombié, Z Dai, H Gautier, M Génard, A Moing, L Roch, G Vercambre, Y Gibon (2018) Putting primary metabolism into perspective to obtain better fruits. Annals of Botany 122 (1), 1–21.
Beauvoit BP, Colombié S, Monier A, Andrieu MH, Biais B, Bénard C, Chéniclet C, Dieuaide-Noubhani M, Nazaret C, Mazat JP et al. (2014) Model-assisted analysis of sugar metabolism throughout tomato fruit development reveals enzyme and carrier properties in relation to vacuole expansion. The Plant cell 26(8): 3224-3242.
Beecher GR (1998) Nutrient Content of Tomatoes and Tomato Products. Exp Biol Med 218: 98–100
Belfanti E, Malatrasi M, Orsi I, Boni AG (2015) Isolated nucleotide sequence from *solanum lycopersicum* for improved resistance to *tomato spotted wilt virus*, TSWV. Patent WO/2015/090468; International Application No: PCT/EP2013/077799.
Bellec-Gauche A, Chiffoleau et Y (2015) Construction des stratégies et des performances dans les circuits courts alimentaires : entre encastrement relationnel et gestionnaire. Rev d'Etudes en Agric Environ 96: 653–676
Bernacchi D, Beck-Bunn T, Emmatty D, Eshed Y, Inai S, Lopez J, Petiard V, Sayama H, Uhlig J, Zamir D, Tanksley S (1998). Advanced backcross QTL analysis in tomato. II. Evaluation of near-isogenic lines carrying single-donor introgressions for desirable wild QTL-alleles derived from *Lycopersicon hirsutum* and *L. pimpinellifolium*. Theor Appl Genet 97 (1/2): 170-180; erratum 97(7): 1191-1196
Bernacchi D, Beck-Bunn T, Eshed Y, Lopez J, Petiard V, Uhlig J, Zamir D, Tanksley S (1998) Advanced backcross QTL analysis in tomato. I. Identification of QTLs for traits of agronomic importance from *Lycopersicon hirsutum*. Theor Appl Genet 97: 381–397
Berr A, Shafiq S, Shen WH (2011) Histone modifications in transcriptional activation during plant development. Biochim Biophys Acta - Gene Regul Mech 1809: 567–576
Bertin N, C Borel, B Brunel, C Cheniclet, M Causse (2003) Do genetic make- up and growth manipulation affect tomato fruit size by cell number, or cell size and DNA endoreduplication? Annals of Botany 92 (3), 415-424
Bertin N, Gary C, Tchamitchian M, Vaissiere BE (1998) Influence of cultivar, fruit position and seed content on tomato fruit weight during a crop cycle under low and high competition for assimilates. J Hortic Sci Biotechnol 73: 541–548
Bertin N, Guichard S, Leonardi C, Longuenesse JJ, Langlois D, Navez B (2000) Seasonal Evolution of the Quality of Fresh Glasshouse Tomatoes under Mediterranean Conditions, as Affected by Air Vapour Pressure Deficit and Plant Fruit Load. Ann Bot 85: 741–750
Bertin N, H Gautier, C Roche (2002) Number of cells in tomato fruit depending on fruit position and source-sink balance during plant development. Plant Growth Regulation 36 (2), 105-112
Bertin N, Martre P, Génard M, Quilot B, Salon C (2010) Why and how can process-based simulation models link genotype to phenotype for complex traits? Case-study of fruit and grain quality traits. Journal of Experimental Botany 61: 955-967
Bhatia P, Ashwath N, Senaratna T, Midmore D (2004) Tissue culture studies of tomato (*Lycopersicon esculentum*). Plant Cell, Tissue and Organ Culture 78(1):1-21.
Bhatt RM, Srinivasa Rao NK (1987) Seed germination and seedling growth responses of tomato cultivars to imposed water stress. J Hortic Sci 62: 221–225
Birchler JA, Yao H, Chudalayandi S, Vaiman D, Veitia RA (2010) Heterosis. Plant Cell 22: 2105–2112
Blanca J, Cañizares J, Cordero L, Pascual L, Diez MJ, Nuez F (2012) Variation Revealed by SNP Genotyping and Morphology Provides Insight into the Origin of the Tomato. PLoS One 7: e48198
Blanca J, Montero-Pau J, Sauvage C, Bauchet G, Illa E, Díez MJ, Francis D, Causse M, van der Knaap E, Cañizares J (2015) Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. BMC Genomics 16: 257
Bloom AJ, Zwieniecki MA, Passioura JB, Randall LB, Holbrook NM, St. Clair DA (2004) Water relations under root chilling in a sensitive and tolerant tomato species. Plant, Cell Environ 27: 971–979
Boison SA, Utsunomiya ATH, Santos DJA, Neves HHR, Carvalheiro R, Mészáros G, Utsunomiya YT, do Carmo AS, Verneque RS, Machado MA, et al. (2017) Accuracy of genomic predictions in Gyr (Bos indicus) dairy cattle. J Dairy Sci 100: 5479–5490
Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S, Sørensen I, Lichtenstein G, et al. (2014b) The genome of the stress-tolerant wild tomato species Solanum pennellii. Nat Genet 46: 1034–1038
Boote K (2016) Modelling crop growth and yield in tomato cultivation. ID: 9781786760401-010
Boureau L, How-Kit A, Teyssier E, Drevensek S, Rainieri M, Joubès J, Stammitti L, Pribat A, Bowler C, Hong Y, et al. (2016) A CURLY LEAF homologue controls both vegetative and reproductive development of tomato plants. Plant Mol Biol 90: 485–501
Bovy A, de Vos R, Kemper M, Schijlen E, Pertejo MA, Muir S, Collins G, Robinson S, Verhoeyen M, Hughes S, Santos-Buelga C (2002) High-flavonol tomatoes resulting from the heterologous expression of the maize transcription factor genes *LC* and *C1*. Plant Cell 14:2509-2526
Bovy A, Schijlen E, Hall RD (2007) Metabolic engineering of flavonoids in tomato (*Solanum lycopersicum*): The potential for metabolomics. Metabolomics 3: 399–412
Brachi B, Morris GP, Borevitz JO (2011) Genome-wide association studies in plants: The missing heritability is in the field. Genome Biol 12: 232
Bramley PM (2000) Is lycopene beneficial to human health? Phytochemistry 54: 233–236
Breiman L (2001) Random Forests. Mach Learn 45: 5–32
Brooks C, Nekrasov V, Lippman ZB, Van Eck J (2014) Efficient gene editing in tomato in the first generation using the clustered regularly interspaced short palindromic repeats/CRISPR-associated9 system. Plant physiology 166(3):1292-7.
Brouwer DJ, Jones ES, St Clair DA (2004) QTL analysis of quantitative resistance to *Phytophthora infestans* (late blight) in tomato and comparisons with potato. Genome 47: 475-492
Brouwer DJ, St Clair DA (2004) Fine mapping of three quantitative trait loci for late blight resistance in tomato using near isogenic lines (NILs) and sub-NILs. Theoretical and Applied Genetics 108: 628-638
Browning BL, Browning SR (2016) Genotype Imputation with Millions of Reference Samples. Am J Hum Genet 98: 116–126
Bruhn CM, Feldman N, Garlitz C, Harwood J, Ivans E, Marshall M, Riley A, Thurber D, Williamson E (1991) Consumer Perceptions of Quality: Apricots, Cantaloupes, Peaches, Pears, Strawberries, and Tomatoes. J Food Qual 14: 187–195
Brummell DA, Harpster MH, Civello PM, Palys JM, Bennett AB, Dunsmuir P (1999) Modification of expansin protein abundance in tomato fruit alters softening and cell wall polymer metabolism during ripening. The Plant Cell 11(11):2203-16.
Bucheli P, Voirol E, De La Torre R, López J, Rytz A, Tanksley SD, Pétiard V (1999) Definition of nonvolatile markers for flavor of tomato (*Lycopersicon esculentum* Mill.) as tools in selection and breeding. J Agric Food Chem 47: 659–664
Budiman MA, Chang S-B, Lee S, Yang TJ, Zhang H-B, de Jong H, Wing RA (2004) Localization of jointless-2 gene in the centromeric region of tomato chromosome 12 based on high resolution genetic and physical mapping. Theor Appl Genet 108: 190–196
Bussières P (1994) Water import rate in tomato fruit:A resistance model. Ann. Bot. 73:75-82.
Butler L (1952) The linkage map of the tomato. J Hered 43: 25–36
Cagas CC, Lee ON, Nemoto K, Sugiyama N (2008) Quantitative trait loci controlling flowering time and related traits in a *Solanum lycopersicum* × *S. pimpinellifolium* cross. Sci Hortic (Amsterdam) 116: 144–151
Calin GA, Croce CM (2006) MicroRNA signatures in human cancers. Nat Rev Cancer 6: 857–866
Callaway (2018) CRISPR plants now subject to t ough GM l aws in European Union. Nature 560, 16 doi: 10.1038/d41586-018-05814-6
Calus MPL, Meuwissen THE, Roos APW de, Veerkamp RF (2008) Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. Genetics 178: 553–561
Canady MA, Meglic V, Chetelat RT (2005) A library of *Solanum lycopersicoides* introgression lines in cultivated tomato . Genome 48: 685–697
Cao K, Xu H, Zhang R, Xu D, Yan L, Sun Y, Xia L, Zhao J, Zou Z, Bao E (2019) Renewable and Sustainable Strategies for Improving the Thermal Environment of Chinese Solar Greenhouses. Energy Build. In Press
Cárdenas PD, Sonawane PD, Pollier J, Vanden Bossche R, Dewangan V, Weithorn E, Tal L, Meir S, Rogachev I, Malitsky S, Giri AP, Goossens A, Burdman S, Aharoni A (2016) GAME9 regulates the biosynthesis of steroidal alkaloids and upstream isoprenoids in the plant mevalonate pathway. Nat Commun. 7:10654.
Carelli BP, Gerald LTS, Grazziotin FG, Echeverrigaray S (2006) Genetic diversity among Brazilian cultivars and landraces of tomato *Lycopersicon esculentum* Mill. revealed by RAPD markers. Genet Resour Crop Evol 53: 395–400
Carmeille A, Caranta C, Dintinger J, Prior P, Luisetti J, Besse P (2006) Identification of QTLs for *Ralstonia solanacearum* race 3-phylotype II resistance in tomato. Theoretical and Applied Genetics 113: 110-121

Carmel-Goren L, Liu YS, Lifschitz E, Zamir D (2003) The *SELF-PRUNING* gene family in tomato. Plant Mol Biol 52: 1215–1222
Caro M, Cruz V, Cuartero J, Estañ MT, Bolarin MC (1991) Salinity tolerance of normal-fruited and cherry tomato cultivars. Plant Soil 136: 249–255
Caromel B, Hamers C, Touhami N, Renaudineau A, Bachellez A, Massire A, Damidaux R, Lefebvre V (2015) Screening tomato germplasm for resistance to late blight. INNOHORT, Innovation in Integrated & Organic Horticulture. ISHS International Symposium, Avignon (France), 8-12 June 2015, pp 15-16
Carrari F, Baxter C, Usadel B, Urbanczyk-Wochniak E, Zanor M-I, Nunes-Nesi A, Nikiforova V, Centero D, Ratzka A, Pauly M, et al. (2006) Integrated Analysis of Metabolite and Transcript Levels Reveals the Metabolic Shifts That Underlie Tomato Fruit Development and Highlight Regulatory Aspects of Metabolic Network Behavior. Plant Physiol 142: 1380–1396
Casteel CL, Walling LL, Paine TD (2007) Effect of *Mi-1.2* gene in natal host plants on behavior and biology of the tomato psyllid *Bactericerca cockerelli* (Sulc) (*Hemiptera : Psyllidae*). Journal of Entomological Science 42: 155-162
Catchen JM, Boone JQ, Davey JW, Hohenlohe PA, Etter PD, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12: 499–510
Causse M, Buret M, Robini K, Verschave P (2003) Inheritance of nutritional and sensory quality traits in fresh market tomato and relation to consumer preferences. J Food Sci 68: 2342–2350
Causse M, C Friguet, C Coiret, M Lépicier, B Navez, M Lee, N Holthuysen, F Sinesio, E Moneta and S Grandillo (2010) Consumer Preferences for Fresh Tomato at the European Scale: A Common Segmentation on Taste and Firmness. Journal of Food Science 75, 9, 531–541
Causse M, Chaïb J, Lecomte L, Buret M, Hospital F (2007) Both additivity and epistasis control the genetic variation for fruit quality traits in tomato. Theor Appl Genet 115: 429–442
Causse M, Duffe P, Gomez MC, Buret M, Damidaux R, Zamir D, Gur A, Chevalier C, Lemaire-Chamley M, Rothan C (2004) A genetic map of candidate genes and QTLs involved in tomato fruit size and composition. J Exp Bot 55: 1671–1685
Causse M, R Damidaux, P Rousselle (2007) Traditional and enhanced breeding for fruit quality traits in tomato. In *Genetic Improvement of Solanaceous Crops, Vol.2: Tomato*. Eds: M.K.Razdan and A. K. Mattoo, Science Publishers, Enfield, USA, 153-192.
Causse M, Saliba-Colombani V, Lesschaeve I, Buret M (2001) Genetic analysis of organoleptic quality in fresh market tomato. 2. Mapping QTLs for sensory attributes. Theor Appl Genet 102: 273–283
Causse M, Saliba- Colombani V, Lecomte L, Duffé P, Rousselle P, Buret M (2002) QTL analysis of fruit quality in fresh market tomato: a few chromosome regions control the variation of sensory and instrumental traits. J Exp Bot 53: 2089–2098
Chakrabarti M, Zhang N, Sauvage C, Muños S, Blanca J, Cañizares J, Diez MJ, Schneider R, Mazourek M, McClead J, et al. (2013) A cytochrome P450 regulates a domestication trait in cultivated tomato. Proc Natl Acad Sci U S A 110: 17125–30
Chen FQ, Foolad MR, Hyman J, St. Clair DA, Beelaman RB (1999) Mapping of QTLs for lycopene and other fruit traits in a *Lycopersicon esculentum* × *L. pimpinellifolium* cross and comparison of QTLs across tomato species. Mol Breed 5: 283–299
Chen J, Kang S, Du T, Qiu R, Guo P, Chen R (2013) Quantitative response of greenhouse tomato yield and quality to water deficit at different growth stages. Agric Water Manag 129: 152–162
Chen X (2005) microRNA biogenesis and function in plants. FEBS Lett 579: 5923
Chen X (2009) Small RNAs and Their Roles in Plant Development. Annu Rev Cell Dev Biol 25: 21–44
Chetelat RT, DeVerna JW, Bennett AB (1995) Introgression into tomato (*Lycopersicon esculentum*) of the *L. chmielewski*i sucrose accumulator gene (sucr) controlling fruit sugar composition. Theor Appl Genet 91: 327–333
Cho SK, Chaabane S Ben, Shah P, Poulsen CP, Yang SW (2014) COP1 E3 ligase protects HYL1 to retain microRNA biogenesis. Nat Commun 5: 5867
Chunwongse J, Chunwongse C, Black L, Hanson P (2002) Molecular mapping of the *Ph-3* gene for late blight resistance in tomato. Journal of Horticultural Science & Biotechnology 77: 281-286
Clark AG (2004) The role of haplotypes in candidate gene studies. Genet Epidemiol 27: 321–333
Clough SJ, Bent AF (1998) Floral dip: a simplified method for Agrobacterium- mediated transformation of Arabidopsis thaliana. The plant journal 16(6):735-743.
Coaker GL, Francis DM (2004) Mapping, genetic effects, and epistatic interaction of two bacterial canker resistance QTLs from *Lycopersicon hirsutum*. Theoretical and Applied Genetics 108: 1047-1055
Colliver S, Bovy A, Collins G, Muir S, Robinson S, de Vos CHR, Verhoeyen ME (2002) Improving the nutritional content of tomatoes through reprogramming their flavonoid biosynthetic pathway. Phytochem Rev 1: 113–123
Colombié S, Beauvoit B, Nazaret C, Bénard C, Vercambre G, Le Gall S, Biais B, Cabasson C, Maucourt M, Bernillon S, Moing A, Dieuaide-Noubhani M, Mazat J-P. and Gibon Y (2017) Respiration climacteric in tomato fruits elucidated by constraint-based modelling. New Phytol, 213: 1726–1739.
Colombié S, Nazaret C, Bénard C, Biais B, Mengin V, Solé M, Fouillen L, Dieuaide-Noubhani M, Mazat J-P, Beauvoit B and Gibon Y (2015) Modelling central metabolic fluxes by constraint-based optimization reveals metabolic reprogramming of developing *Solanum lycopersicum* (tomato) fruit. Plant J 81: 24–39.
Comai L, Henikoff S (2006) TILLING: Practical single-nucleotide mutation discovery. Plant J 45: 684–694
Coneva V, Frank MH, Balaguer MAL, Li M, Sozzani R, Chitwood DH (2017) Genetic Architecture and Molecular Networks Underlying Leaf Thickness in Desert-Adapted Tomato Solanum pennellii. Plant Physiol. 175(1):376-391.
Constantinescu D, Memmah M-M, Vercambre G, Génard M, Baldazzi V, Causse M, et al. (2016) Model-Assisted Estimation of the Genetic Variability in Physiological Parameters Related to Tomato Fruit Growth under Contrasted Water Conditions. Frontiers in Plant Science, 7, 1841. http://doi.org/10.3389/fpls.2016.01841
Costa JM, Ortuño MF, Chaves MM (2007) Deficit Irrigation as a Strategy to Save Water: Physiology and Potential Application to Horticulture. J Integr Plant Biol 49: 1421–1434
Cournède P-H et al. (2013) Development and evaluation of plant growth models: Methodology and implementation in the pygmalion platform, Math. Mod. of Nat. Phen. 8(4), 112–130.
Cowger C, Brown JKM (2019) Durability of quantitative resistance in crops: greater than we know? Annual Review of Phytopathology: in press
Crain J, Mondal S, Rutkoski J, Singh RP, Poland J (2018) Combining High-Throughput Phenotyping and Genomic Information to Increase Prediction and Selection Accuracy in Wheat Breeding. Plant Genome 11: 0
Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, Burgueño J, Camacho-González JM, Pérez-Elizalde S, Beyene Y, et al. (2017) Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends Plant Sci. doi: 10.1016/j.tplants.2017.08.011
Cui J, Jiang N, Zhou X, Hou X, Yang G, Meng J, Luan Y (2018) Tomato MYB49 enhances resistance to Phytophthora infestans and tolerance to water deficit and salt stress. Planta 248: 1487–1503
Cui J, You C, Chen X (2017a) The evolution of microRNAs in plants. Curr Opin Plant Biol 35: 61–67
Cui J, Zhou B, Ross SA, Zempleni J (2017b) Nutrition, microRNAs, and Human Health. Adv Nutr 8: 105–112
Cuyabano BC, Su G, Lund MS (2014) Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC Genomics. doi: 10.1186/1471-2164-15-1171
Cuyabano BCD, Su G, Lund MS (2015a) Selection of haplotype variables from a high-density marker map for genomic prediction. Genet Sel Evol 47: 61
Cuyabano BCD, Su G, Rosa GJM, Lund MS, Gianola D (2015b) Bootstrap study of genome-enabled prediction reliabilities using haplotype blocks across Nordic Red cattle breeds. J Dairy Sci 98: 7351–7363
Dal Cin V, Kevany B, Fei Z, Klee HJ (2009) Identification of Solanum habrochaites loci that quantitatively influence tomato fruit ripening-associated ethylene emissions. Theor Appl Genet 119: 1183–1192
Danecek P, Huang J, Min JL, Timpson NJ, Trabetti E, Richards JB, Durbin R, Howie B, Gambaro G, Zheng H-F, et al. (2015) Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat Commun 6: 8111
Danilo B, Perrot L, Botton E, Nogué F, Mazier M (2018) The DFR locus: A smart landing pad for targeted transgene insertion in tomato. PLoS ONE 13(12): e0208395.
Danilo B, Perrot L, Mara K, Botton E, Nogué F, Mazier M. (2019) Efficient and transgene-free gene targeting using Agrobacterium-mediated delivery of the CRISPR/Cas9 system in tomato. Plant Cell Rep. 38(4):459-462.
Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, et al. (2016) Next-generation genotype imputation service and methods. Nat Genet 48: 1284–1287
Davies JN, Hobson GE (1981) The constituents of tomato fruit — the influence of environment, nutrition, and genotype. Crit Rev Food Sci Nutr 15: 205–280
Davies JN, Hobson GE and McGlasson WB (1981) The constituents of tomato fruit — the influence of environment, nutrition, and genotype. C R C Crit. Rev. Food Sci. Nutr., 15, 205–280.
Davila Olivas NH, Kruijer W, Gort G, Wijnen CL, van Loon JJA, Dicke M (2017) Genome-wide association analysis reveals distinct genetic architectures for single and combined stress responses in *Arabidopsis thaliana*. New Phytol 213: 838–851
Davis J, Yu DZ, Evans W, Gokirmak T, Chetelat RT, Stotz HU (2009) Mapping of loci from *Solanum lycopersicoides* conferring resistance or susceptibility to *Botrytis cinerea* in tomato. Theoretical and Applied Genetics 119: 305-314
de Groot CC, Marcelis LFM, van den Boogaard R, Lambers H (2004) Response of growth of tomato to phosphorus and nitrogen nutrition.

Acta Hortic 357–364
de Jong CF, Takken FLW, Cai XH, de Wit P, Joosten M (2002) Attenuation of *Cf*-mediated defense responses at elevated temperatures correlates with a decrease in elicitor-binding sites. Molecular Plant-Microbe Interactions 15: 1040-1049
de Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL, Kirst M, Huber D, Peter GF (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193: 327–45
De Swaef T, Mellisho CD, Baert A, De Schepper V, Torrecillas A, Conejero W, Steppe K (2014) Model-assisted evaluation of crop load effects on stem diameter variations and fruit growth in peach. Trees, 28, 1607–1622.
Delhaize E, Gruber BD, Ryan PR (2007) The roles of organic anion permeases in aluminium resistance and mineral nutrition. FEBS Lett 581: 2255–2262
DeVicente MC, Tanksley SD (1993) QTL analysis of transgressive segregation in an interspecific tomato cross. Genetics 134: 585–596
Dileo MV, Pye MF, Roubtsova TV, Duniway JM, MacDonald JD, Rizzo DM, Bostock RM (2010) Abscisic Acid in Salt Stress Predisposition to *Phytophthora* Root and Crown Rot in Tomato and Chrysanthemum. Phytopathology 100: 871-879
Diouf IA, Derivot L, Bitton F, Pascual L, Causse M (2018) Water Deficit and Salinity Stress Reveal Many Specific QTL for Plant Growth and Fruit Quality Traits in Tomato. Front Plant Sci 9: 279
Dixon MS, Hatzixanthis K, Jones DA, Harrison K, Jones JDG (1998) The tomato *Cf-5* disease resistance gene and six homologs show pronounced allelic variation in leucine-rich repeat copy number. Plant Cell (The) 10: 1915-1925
Dixon MS, Jones DA, Keddie JS, Thomas CM, Harrison K, Jones JD. (1996) The Tomato Cf-2 Disease Resistance Locus Comprises Two Functional Genes Encoding Leucine-Rich Repeat Proteins. Cell 84: 451–459
Do PT, Prudent M, Sulpice R, Causse M, Fernie AR (2010) The Influence of Fruit Load on the Tomato Pericarp Metabolome in a Solanum chmielewskii Introgression Line Population. Plant Physiol 154: 1128–1142
Doganlar S, Dodson J, Gabor B, Beck-Bunn T, Crossman C, Tanksley SD (1998) Molecular mapping of the *py-1* gene for resistance to corky root rot (*Pyrenochaeta lycopersici*) in tomato. Theoretical and Applied Genetics 97: 784-788
Doganlar S, Frary A, Ku H-M, Tanksley SD (2003) Mapping quantitative trait loci in inbred backcross lines of *Lycopersicon pimpinellifolium* (LA1589). Genome 45: 1189–1202
Domínguez T, Hernández ML, Pennycooke JC, Jiménez P, Martínez-Rivas JM, Sanz C, Stockinger EJ, Sánchez-Serrano JJ, Sanmartín M. (2010) Increasing ω-3 desaturase expression in tomato results in altered aroma profile and enhanced resistance to cold stress. Plant physiology 153(2): 655-65.
Donald 1968 C.M. The breeding of crop idéotypes. Euphytica. 17 (1968). pp. 385-403
Dong QL, Liu DD, An XH, Hu DG, Yao YX, Hao YJ (2011) MdVHP1 encodes an apple vacuolar H+-PPase and enhances stress tolerance in transgenic apple callus and tomato. Journal of plant physiology 168(17):2124-33.
Dong Z, Men Y, Li Z, Zou Q, Ji J (2019) Chlorophyll fluorescence imaging as a tool for analyzing the effects of chilling injury on tomato seedlings. Sci Hortic (Amsterdam) 246: 490–497
Dorais M, Papadopoulos AP, Gosselin A (2001) Greenhouse tomato fruit quality. *Hortic Rev* 26:239-319
Dreissig S, Schiml S, Schindele P, Weiss O, Rutten T, Schubert V, Gladilin E, Mette MF, Puchta H, Houben A. (2017) Live- cell CRISPR imaging in plants reveals dynamic telomere movements. The Plant Journal 91(4):565-73.
Driedonks N, Wolters-Arts M, Huber H, de Boer G-J, Vriezen W, Mariani C, Rieu I (2018) Exploring the natural variation for reproductive thermotolerance in wild tomato species. Euphytica 214: 67
Du Y-D, Niu W-Q, Gu X-B, Zhang Q, Cui B-J (2018) Water- and nitrogen-saving potentials in tomato production: A meta-analysis. Agric Water Manag 210: 296–303
Duangjit J, Causse M, Sauvage C (2016) Efficiency of genomic selection for tomato fruit quality. Mol Breed 36: 36:29
Edwards SM, Buntjer JB, Jackson R, Bentley AR, Lage J, Byrne E, Burt C, Jack P, Berry S, Flatman E, et al. (2019) The effects of training population design on genomic prediction accuracy in wheat. Threotical Appl Genet 443267
El-hady E, Haiba A, El-hamid NRA, Rizkalla A, Phylogenetic AR (2010) Phylogenetic Diversity and Relationships of Some Tomato Varieties by Electrophoretic Protein and RAPD analysis.
Elvanidi A, Katsoulas N, Augoustaki D, Loulou I, Kittas C (2018) Crop reflectance measurements for nitrogen deficiency detection in a soilless tomato crop. Biosyst Eng 176: 1–11
Endelman JB (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Genome J 4: 250
Ercolano MR, Sanseverino W, Carli P, Ferriello F, Frusciante L (2012) Genetic and genomic approaches for R-gene mediated disease resistance in tomato: retrospects and prospects. Plant Cell Reports 31: 973-985
Eriksson EM, Bovy A, Manning K, Harrison L, Andrews J, De Silva J, Tucker GA, Seymour GB, Thompson J, Tor M, et al. (2004) Effect of the Colorless non-ripening Mutation on Cell Wall Biochemistry and Gene Expression during Tomato Fruit Development and Ripening 1[w]. Plant Physiol 136: 4184–4197
Eshed Y, Gera G, Zamir D (1996) A genome-wide search for wild-species alleles that increase horticultural yield of processing tomato. Theor Appl Genet 93: 877-886
Eshed Y, Zamir D (1995) An Introgression Line Population of *Lycopersicon pennellii* in the Cultivated Tomato Enables the Identification and Fine Mapping of Yield-Associated Qtl Genetics 141: 1147–1162
Evangelou E, Ioannidis JPA (2013) Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet 14: 379–389
Fan ZQ, Ba LJ, Shan W, Xiao YY, Lu WJ, Kuang JF, Chen JY (2018) A banana R2R3- MYB transcription factor MaMYB3 is involved in fruit ripening through modulation of starch degradation by repressing starch degradation- related genes and MabHLH6. The Plant Journal 96(6):1191-205.
Fang C, Fernie AR, Luo J (2018) Exploring the Diversity of Plant Metabolism. Trends Plant Sci 24: 83–98
Fang X, Cui Y, Li Y, Qi Y (2015) Transcription and processing of primary microRNAs are coupled by Elongator complex in Arabidopsis. Nat Plants 1: 15075
Fanwoua J, de Visser PHB, Heuvelink E, Yin X, Struik PC, Marcelis LFM (2013) A dynamic model of tomato fruit growth integrating cell division, cell growth and endoreduplication. Functional Plant Biology 40(11) 1098-1114.
FAO (2015) Coping with climate change – the roles of genetic resources for food and agriculture.
Farashi S, Kryza T, Clements J, Batra J (2019) Post-GWAS in prostate cancer: from genetic association to biological contribution. Nat Rev Cancer 19: 46–59
Fereres E, Soriano MA (2006) Deficit irrigation for reducing agricultural water use. J Exp Bot 58: 147–159
Fernandes SB, Dias KOG, Ferreira DF, Brown PJ (2018) Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. Theor Appl Genet 131: 747–755
Fernandez AI, Viron N, Alhagdow M, Karimi M, Jones M, Amsellem Z, Sicard A, Czerednik A, Angenent G, Grierson D, May S (2009) Flexible tools for gene expression and silencing in tomato. Plant Physiology 151(4):1729-40.
Fernie AR, Aharoni A, Willmitzer L, Stitt M, Tohge T, Kopka J, Carroll AJ, Saito K, Fraser PD, DeLuca V (2011) Recommendations for Reporting Metabolite Data. Plant Cell 23: 2477–2482
Fernie AR, Schauer N (2009) Metabolomics-assisted breeding: a viable option for crop improvement? Trends Genet 25: 39–48
Finkers R, Bai YL, van den Berg P, van Berloo R, Meijer-Dekens F, ten Have A, van Kan J, Lindhout P, van Heusden AW (2008) Quantitative resistance to *Botrytis cinerea* from *Solanum neorickii*. Euphytica 159: 83-92
Finkers R, van den Berg P, van Berloo R, ten Have A, van Heusden AW, van Kan JAL, Lindhout P (2007) Three QTLs for *Botrytis cinerea* resistance in tomato. Theoretical and Applied Genetics 114: 585-593
Finkers R, Van Heusden AW, Meijer-Dekens F, Van Kan JAL, Maris P, Lindhout P (2007) The construction of a Solanum habrochaites LYC4 introgression line population and the identification of QTLs for resistance to Botrytis cinerea. Theor Appl Genet 114: 1071–1080
Foolad MR (2007) Genome mapping and molecular breeding of tomato. Int J Plant Genomics 2007: 64358
Foolad MR, Merk HL, Ashrafi H (2008) Genetics, genomics and breeding of late blight and early blight resistance in tomato. Critical Reviews in Plant Sciences 27: 75-107
Foolad MR, Panthee DR (2012) Marker-Assisted Selection in Tomato Breeding. CRC Crit Rev Plant Sci 31: 93–123
Foolad MR, Sullenberger MT, Ohlson EW, Gugino BK (2014) Response of accessions within tomato wild species, Solanum pimpinellifolium to late blight. Plant Breeding 133: 401-411
Foolad MR, Zhang LP, Khan AA, Nino-Liu D, Lin GY (2002) Identification of QTLs for early blight (*Alternaria solani*) resistance in tomato using backcross populations of a *Lycopersicon esculentum* x *L. hirsutum* cross. Theoretical and Applied Genetics 104: 945-958
Fragkostefanakis S, Mesihovic A, Simm S, Paupière MJ, Hu Y, Paul P, Mishra SK, Tschiersch B, Theres K, Bovy A, et al. (2016) HsfA2 Controls the Activity of Developmentally and Stress-Regulated Heat Stress Protection Mechanisms in Tomato Male Reproductive Tissues. Plant Physiol 170: 2461–77
Fragkostefanakis S, Röth S, Schleiff E, Scharf KD (2015) Prospects of engineering thermotolerance in crops through modulation of heat stress transcription factor and heat shock protein networks. Plant Cell Environ 38: 1881–1895
Frary A, Doganlar S, Daunay MC, Tanksley SD (2003) QTL analysis of morphological traits in eggplant and implications for conservation of gene function during evolution of solanaceous species. Theor Appl Genet 107: 359–370
Frary A, Fulton TM, Zamir D, Tanksley SD (2004) Advanced backcross QTL analysis of a *Lycopersicon esculentum* × *L. pennellii* cross and identification of possible orthologs in the Solanaceae. Theor Appl Genet 108: 485–496
Frary A, Keleş D, Pinar H, Göl D, Doğanlar S (2011) NaCl tolerance in *Lycopersicon pennellii* introgression lines: QTL related to physiological responses. *Biol Plant.* 55: 461–468.

Frary A, Nesbitt TC, Frary A, Grandillo S, Van Der Knaap E, Cong B, Liu J, Meller J, Elber R, Alpert KB, et al. (2000) fw2.2: A quantitative trait locus key to the evolution of tomato fruit size. Science (80-) 289: 85–88
Frary A, Göl D, Keleş D, Ökmen B, Pınar H, Şığva HÖ et al. (2010) Salt tolerance in Solanum pennellii: antioxidant response and related QTL. *BMC Plant Biol*. 10: 58.
Fridman E, Carrari F, Liu YS, Fernie AR, Zamir D (2004) Zooming in on a quantitative trait for tomato yield using interspecific introgressions. Science (80-) 305: 1786–1789
Fridman E, Liu YS, Carmel_Goren L, Gur A, Shoresh M, Pleban T, Eshed Y, Zamir D (2002) Two tightly linked QTLs modify tomato sugar content via different physiological pathways. Mol Genet Genomics 266: 821–826
Fridman E, Pleban T, Zamir D (2000) A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. Proc Natl Acad Sci 97: 4718–4723
Fridman E, Zamir D (2003) Functional divergence of a syntenic invertase gene family in tomato, potato, and Arabidopsis. Plant Physiol 131: 603–9
Fry WE, Goodwin SB (1997) Re-emergence of potato and tomato late blight in the United States. Plant Disease 81: 1349-1357
Fujita M, Fujita Y, Noutoshi Y, Takahashi F, Narusaka Y, Yamaguchi-Shinozaki K, Shinozaki K (2006) Crosstalk between abiotic and biotic stress responses: a current view from the points of convergence in the stress signaling networks. Curr Opin Plant Biol 9: 436–442
Fulop D, Ranjan A, Ofner I, Covington MF, Chitwood DH, West D, Ichihashi Y, Headland L, Zamir D, Maloof JN, et al. (2016) A New Advanced Backcross Tomato Population Enables High Resolution Leaf QTL Mapping and Gene Identification. G38; Genes|Genomes|Genetics 6: 3169–3184
Fulton TM (2002) Identification, Analysis, and Utilization of Conserved Ortholog Set Markers for Comparative Genomics in Higher Plants. Plant Cell 14: 1457–1467
Fulton TM, Beck-Bunn T, Emmatty D, Eshed Y, Lopez J, Petiard V, Uhlig J, Zamir D, Tanksley SD (1997) QTL analysis of an advanced backcross of *Lycopersicon peruvianum* to the cultivated tomato and comparisons with QTLs found in other wild species. Theor Appl Genet 95: 881–894
Fulton TM, Grandillo S, Beck-Bunn T, Fridman E, Frampton A, Lopez J, Petiard V, Uhlig J, Zamir D, Tanksley SD (2000) Advanced backcross QTL analysis of a *Lycopersicon esculentum* ×*Lycopersicon parviflorum* cross. Theor Appl Genet 100: 1025–1042
Gaj T, Gersbach CA, Barbas CF (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol. 31(7):397-405.
Gallusci P, Hodgman C, Teyssier E, Seymour GB (2016) DNA Methylation and Chromatin Regulation during Fleshy Fruit Development and Ripening. Front Plant Sci 7: 807
Gao C, Ju Z, Cao D, Zhai B, Qin G, Zhu H, Fu D, Luo Y, Zhu B (2015) MicroRNA profiling analysis throughout tomato fruit development and ripening reveals potential regulatory role of RIN on microRNAs accumulation. Plant Biotechnol J 13: 370–382
Garcia V, Bres C, Just D, Fernandez L, Tai FWJ, Mauxion JP, Le Paslier MC, Bérard A, Brunel D, Aoki K, et al. (2016) Rapid identification of causal mutations in tomato EMS populations via mapping-by-sequencing. Nat Protoc 11: 2401–2418
Gauffier C, Lebaron C, Moretti A, Constant C, Moquet F, Bonnet G, Caranta C, Gallois J-L (2016) A TILLING approach to generate broad-spectrum resistance to potyviruses in tomato is hampered by *eIF4E* gene redundancy. Plant J 85: 717–729
Gautier H, Diakou-Verdin V, Bénard C, Reich M, Buret M, Bourgaud F, Poëssel JL, Caris-Veyrat C, Génard M (2008) How Does Tomato Quality (Sugar, Acid, and Nutritional Quality) Vary with Ripening Stage, Temperature, and Irradiance? J Agric Food Chem 56: 1241–1250
Génard M, Bertin N, Gautier H, Lescourret F, Quilot B (2010) Virtual profiling: a new way to analyse phenotypes. Plant J 62: 344-355
Génard M, Lescourret F (2004) Modelling fruit quality: ecophysiological, agronomical and ecological perspectives. In : R. Dris and S.M. Jain (eds), Production practices and quality assessment of food crops, Vol. 1, "Preharvest practice", Kluwer Academic Publisher, Netherlands, 47-82.
Génard M, Memmah M-M, Quilot-Turion B, Vercambre G, Baldazzi V, Le Bot J, Bertin N, Gautier H, Lescourret F, Pagès L (2016) Process-Based Simulation Models Are Essential Tools for Virtual Profiling and Design of Ideotypes: Example of Fruit and Root. In Crop Systems Biology: Narrowing the gaps between crop modelling and genetics. Yin, Xinyou, Struik, Paul C. (Eds.), pp 83-104
Gerszberg A, Hnatuszko-Konka K, Kowalczyk T, Kononowicz AK (2015) Tomato (*Solanum lycopersicum* L.) in the service of biotechnology. Plant Cell, Tissue and Organ Culture 120(3): 881-902.
Geshnizjani N, Ghaderi-Far F, Willems LAJ, Hilhorst HWM, and Ligterink W (2018) Characterization of and genetic variation for tomato seed thermo-inhibition and thermo-dormancy. *BMC Plant Biol*. 18: 229.
Gest N, Gautier H, Stevens R (2013) Ascorbate as seen through plant evolution: the rise of a successful molecule? Journal of Experimental Botany 64: 33-53
Gianola D, Kaam JBCHM van (2008) Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits. Genetics 178: 2289
Giovannoni J, Nguyen C, Ampofo B, Zhong S, Fei Z (2017) The Epigenome and Transcriptional Dynamics of Fruit Ripening. Annu Rev Plant Biol 68: 61–84
Giovannucci E (1999) Tomatoes, Tomato-Based Products, Lycopene, and Cancer: Review of the Epidemiologic Literature. J Natl Cancer Inst 91: 317–331
Giroux RW, Filion WG (1992) A comparison of the chilling-stress response in two differentially tolerant cultivars of tomato (*Lycopersicon esculentum*). Biochem Cell Biol 70: 191–198
Goff SA, Klee HJ (2006) Plant volatile compounds: Sensory cues for health and nutritional value? Science 311: 815–819
Gonatopoulos-Pournatzis T, Cowling VH (2015) Cap-binding complex (CBC). Biochem J 458: 185–185
Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: Ten years of next-generation sequencing technologies. Nat Rev Genet 17: 333–351
Goyer A, Illarionova V, Roje S, Fischer M, Bacher A, Hanson AD (2004) The Role of Phylogenetics in Comparative Genetics. Plant Physiol 135: 103–111
Grandillo S, Cammareri M (2016) Molecular Mapping of Quantitative Trait Loci in Tomato. The tomato genome. Springer, Berlin, Heidelberg, pp 39–73
Grandillo S, Chetelat R, Knapp S, Spooner D, Peralta I, Cammareri M, Perez O, Termolino P, Tripodi P, Chiusano ML, et al. (2011) Solanum sect. Lycopersicon. Wild Crop Relat. Genomic Breed. Resour. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 129–215
Grandillo S, Ku HM, Tanksley SD (1996) Characterization of fs8.1, a major QTL influencing fruit shape in tomato. Mol Breed 2: 251–260
Grandillo S, Ku HM, Tanksley SD (1999) Identifying the loci responsible for natural variation in fruit size and shape in tomato. Theor Appl Genet 99: 978–987
Grandillo S, Tanksley SD (1996a) Genetic analysis of RFLPs, GATA microsatellites and RAPDs in a cross between L. esculentum and L. pimpinellifolium. Theor Appl Genet 92: 957–965
Grandillo S, Tanksley SD (1996b) QTL analysis of horticultural traits differentiating the cultivated tomato from the closely related species *Lycopersicon pimpinellifolium*. Theor Appl Genet 92: 935–951
Grandillo S, Termolino P, van der Knaap E (2013) Molecular Mapping of Complex Traits in Tomato. Genet. Genomics, Breed. Tomato. Science Publishers, pp 150–227
Grierson D (2016) Identifying and silencing tomato ripening genes with antisense genes. Plant biotechnology journal 14(3): 835-838.
Grilli G, Trevizan Braz L, Gertrudes E, and Lemos M (2007) QTL identification for tolerance to fruit set in tomatoby fAFLP markers. Crop Breed Appl Biotechnol. 7: 234–241.
Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature 455: 1193–1197
Grolier P, Rock E (1998) Composition of tomato in antioxidants: variations and methodology. Proc. Tomato Heal. Semin. Pamplona, Spain. pp 25–28
Grotewold E, Chamberlin M, Snook M, Siame B, Butler L, Swenson J, Maddock S, St Clair G, Bowen B, Hughes S, et al. (1998) Engineering secondary metabolism in maize cells by ectopic expression of transcription factors. Plant Cell 10: 721–40
Guan Y, Stephens M (2008) Practical issues in imputation-based association mapping. PLoS Genet. doi: 10.1371/journal.pgen.1000279
Guichard S, Bertin N, Leonardi C, Gary C (2001) Tomato fruit quality in relation to water and carbon fluxes. Agronomie 21: 385–392
Guichard S, Gary C, Leonardi C, Bertin N (2005) Analysis of Growth and Water Relations of Tomato Fruits in Relation to Air Vapor Pressure Deficit and Plant Fruit Load. J Plant Growth Regul 24: 201–213
Gundersen V, McCall D, Bechmann IE (2001) Comparison of major and trace element concentrations in Danish greenhouse tomatoes (*Lycopersicon esculentum* Cv. Aromata F1) cultivated in different substrates. J Agric Food Chem 49: 3808–15
Gupta A, Pal RK, Rajam MV (2013) Delayed ripening and improved fruit processing quality in tomato by RNAi-mediated silencing of three homologs of 1-aminopropane-1-carboxylate synthase gene. Journal of Plant Physiology 170(11):987-95.
Gur A, Osorio S, Fridman E, Fernie AR, Zamir D (2010) hi2-1, A QTL which improves harvest index, earliness and alters metabolite accumulation of processing tomatoes. Theor Appl Genet 121: 1587–1599
Gur A, Semel Y, Osorio S, Friedmann M, Seekh S, Ghareeb B et al. (2011) Yield quantitative trait loci from wild tomato are predominately expressed by the shoot. Theor Appl Genet. 122: 405–420.
Gur A, Zamir D (2015) Mendelizing all Components of a Pyramid of Three Yield QTL in Tomato. Front Plant Sci. 6:1096. doi: 10.3389/fpls.2015.01096.

Haanstra JPW, Wye C, Verbakel H, Meijer-Dekens F, Van Den Berg P, Odinot P, Van Heusden AW, Tanksley S, Lindhout P, Peleman J (1999) An integrated high-density RFLP-AFLP map of tomato based on two *Lycopersicon esculentum x L. pennellii* F2 populations. Theor Appl Genet 99: 254–271
Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12: 186
Haggard JE, Johnson EB, St. Clair DA (2013) Linkage Relationships Among Multiple QTL for Horticultural Traits and Late Blight (P. infestans) Resistance on Chromosome 5 Introgressed from Wild Tomato Solanum habrochaites. G3 Genes|Genomes|Genetics 3: 2131–2146
Halperin E, Stephan DA (2009) SNP imputation in association studies. Nat Biotechnol 27: 349–351
Hamilton JP, Sim S-C, Stoffel K, Van Deynze A, Buell CR, Francis DM (2012) Single Nucleotide Polymorphism Discovery in Cultivated Tomato via Sequencing by Synthesis. Plant Genome J 5: 17
Han P, Lavoir A-V, Le Bot J, Amiens-Desneux E, Desneux N (2015) Nitrogen and water availability to tomato plants triggers bottom-up effects on the leafminer Tuta absoluta. Sci Rep 4: 4455
Hanson AD, Gregory III JF (2002) Synthesis and turnover of folates in plants. Curr Opin Plant Biol 5: 244–249
Hanson PM, Yang R, Wu J, Chen J, Ledesma D, Tsou SCS, Lee T-C (2004) Variation for Antioxidant Activity and Antioxidants in Tomato. J Am Soc Hortic Sci 129: 704–711
Hanssen IM, Thomma B (2010) *Pepino mosaic viru*s: a successful pathogen that rapidly evolved from emerging to endemic in tomato crops. Molecular Plant Pathology 11: 179-189
Hanssens J, De Swaef T, Steppe K (2015) High light decreases xylem contribution to fruit growth in tomato. Plant, Cell and Environment, 38, 487–498.
Harborne JB (1988) The Flavonoids : advances in research since 1980. Chapman and Hall
Harborne JB (1994) The Flavonoids. Advances in research since 1986, 1st ed. Chapman Hall, London
Harborne JB, Williams CA (2000) Advances in flavonoid research since 1992. Phytochemistry 55: 481–504
Harvell CD, Mitchell CE, Ward JR, Altizer S, Dobson AP, Ostfeld RS, Samuel MD (2002) Climate warming and disease risks for terrestrial and marine biota. Science 296: 2158–62
Haseneyer G, Schmutzer T, Seidel M, Zhou R, Mascher M, Schön CC, Taudien S, Scholz U, Stein N, Mayer KFX, et al. (2011) From RNA-seq to large-scale genotyping - genomics resources for rye (Secale cereale L.). BMC Plant Biol 11: 131
Hayashi T, Iwata H (2010) EM algorithm for Bayesian estimation of genomic breeding values. BMC Genet 11: 3
He S, Schulthess AW, Mirdita V, Zhao Y, Korzun V, Bothe R, Ebmeyer E, Reif JC, Jiang Y (2016) Genomic selection in a commercial winter wheat population. Theor Appl Genet 129: 641–651
He Y (2012) Chromatin regulation of flowering. Trends Plant Sci 17: 556–562
Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: A comparison of models. Crop Sci 52: 146–160
Hess M, Druet T, Hess A, Garrick D (2017) Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. Genet Sel Evol 49: 54
Heuvelink E (2005) Tomatoes. CABI Pub
Heuvelink E. and Bertin N (1994) Dry matter partitioning in a tomato crop: comparison of two simulation models. J. Hort. Sci. 69:885-903.
Hill M, Tran N (2018) MicroRNAs Regulating MicroRNAs in Cancer. Trends in Cancer 4: 465–468
Ho LC (1996) The mechanism of assimilate partitioning and carbohydrate compartmentation in fruit in relation to the quality and yield of tomato. J Exp Bot 47: 1239–1243
Hobson G, Grierson D (1993) Tomato. Biochem. Fruit Ripening. Springer Netherlands, Dordrecht, pp 405–442
Hobson GE, Bedford L (1989) The composition of cherry tomatoes and its relation to consumer acceptability. J Hortic Sci 64: 321–329
Hospital F, Charcosset A (1997) Marker-assisted introgression of Quantitative Trait Loci. Genetics 147: 1469-1485.
Hospital F, Chevalet C, Mulsant P (1992) Using markers in gene introgression breeding programs. Genetics 132: 1199-1210.
How Kit A, Boureau L, Stammitti-Bert L, Rolin D, Teyssier E, Gallusci P (2010) Functional analysis of SlEZ1 a tomato Enhancer of zeste (E(z)) gene demonstrates a role in flower development. Plant Mol Biol 74: 201–213
Huang BE, George AW, Forrest KL, Kilian A, Hayden MJ, Morell MK, Cavanagh CR (2012) A multiparent advanced generation inter-cross population for genetic analysis in wheat. Plant Biotechnol J 10: 826–839
Huang WJ, Liu HK, McCormick S, Tang WH (2014) Tomato pistil factor STIG1 promotes in vivo pollen tube growth by binding to phosphatidylinositol 3-phosphate and the extracellular domain of the pollen receptor kinase LePRK2. The Plant Cell 26(6):2505-23.
Huang Z, van der Knaap E (2011) Tomato fruit weight 11.3 maps close to fasciated on the bottom of chromosome 11. Theor Appl Genet 123: 465–474
Hutton SF, Scott JW, Yang WC, Sim SC, Francis DM, Jones JB (2010) Identification of QTL associated with resistance to bacterial spot race T4 in tomato. Theor and Appl Genet 121: 1275-1287
Isidro J, Jannink J-L, Akdemir D, Poland J, Heslot N, Sorrells ME (2015) Training set optimization under population structure in genomic selection. Theor Appl Genet 128: 145–158
Islam MN, Hasanuzzaman ATM, Zhang Z-F, Zhang Y, Liu T-X (2017) High Level of Nitrogen Makes Tomato Plants Releasing Less Volatiles and Attracting More Bemisia tabaci (Hemiptera: Aleyrodidae). Front Plant Sci 8: 466
Ito Y, Nishizawa-Yokoi A, Endo M, Mikami M, Shima Y, Nakamura N, Kotake-Nara E, Kawasaki S, Toki S (2017) Re-evaluation of the rin mutation and the role of RIN in the induction of tomato ripening. Nat Plants 3(11):866-874.
Iwata H, Jannink JL (2010) Marker genotype imputation in a low-marker-density panel with a high-marker-density reference panel: Accuracy evaluation in barley breeding lines. Crop Sci 50: 1269–1278
Janse J, Schols M (1995) Une préférence pour un goût sucré et non farineux. *Groenten + Fruit* 26:16-17.
Jatoi SA, Fujimura T, Yamanaka S, Watanabe J, Watanabe KN, Watanabe KN (2008) Potential loss of unique genetic diversity in tomato landraces by genetic colonization of modern cultivars at a non-center of origin. Plant Breed 127: 189–196
Jha UC, Bohra A, Jha R (2017) Breeding approaches and genomics technologies to increase crop yield under low-temperature stress. Plant Cell Rep 36: 1–35
Jiang Y, Schmidt RH, Reif JC (2018) Haplotype-Based Genome-Wide Prediction Models Exploit Local Epistatic Interactions Among Markers. G3 Genes|Genomes|Genetics 8: g3.300548.2017
Jiménez-Gómez JM, Alonso-Blanco C, Borja A, Anastasio G, Angosto T, Lozano R, Martínez-Zapater JM (2007) Quantitative genetic analysis of flowering time in tomato. Genome 50: 303–315
Johansson L, Haglund A, Berglund L, Lea P, Risvik E (1999) Preference for tomatoes, affected by sensory attributes and information about growth conditions. Food Qual Prefer 10: 289–298
Johnstone PR, Hartz TK, LeStrange M, Nunez JJ, Miyao EM (2005) Managing fruit soluble solids with late-season deficit irrigation in drip-irrigated processing tomato production. HortScience 40: 1857–1861
Jonas E, de Koning D-J (2013) Does genomic selection have a future in plant breeding? Trends Biotechnol 31: 497–504
Jones JB (1986) Survival of Xanthomonas campestris pv. vesicatoria in Florida on Tomato Crop Residue, Weeds, Seeds, and Volunteer Tomato Plants. Phytopathology 76: 430
Jones JW, Dayan E, Allen LH, Van Keulen H, Challa H (1991) A dynamic tomato growth and yie1d model (Tomgro). American Society Agricultural Engineers34: 663-672
Kabelka E, Franchino B, Francis DM (2002) Two loci from *Lycopersicon hirsutum* LA407 confer resistance to strains of *Clavibacter michiganensis* subsp *michiganensis*. Phytopathology 92: 504-510
Kader AA, Morris LL, Stevens MA, Albright-Holton M (1978) Composition and flavor quality of fresh quality of fresh market tomatoes as influenced by some postharvest handling procedures. J Am Soc Hortic Sci 103: 6–13
Kamal HM, Takashina T, Egashira H, Satoh H, Imanishi S (2001) Introduction of aromatic fragrance into cultivated tomato from the "peruvianum complex." Plant Breed 120: 179–181
Kang BC, Yeam I, Li HX, Perez KW, Jahn MM (2007) Ectopic expression of a recessive resistance gene generates dominant potyvirus resistance in plants. Plant Biotechnology Journal 5: 526-536
Karimi Z, Sargolzaei M, Robinson JAB, Schenkel FS (2018) Assessing haplotype-based models for genomic evaluation in Holstein cattle. Can J Sci 1–10
Karlova R, Van Haarst JC, Maliepaard C, Van De Geest H, Bovy AG, Lammers M, Angenent GC, De Maagd RA (2013) Identification of microRNA targets in tomato fruit development using high-throughput sequencing and degradome analysis. J Exp Bot 64: 1863–1878
Kawchuk LM, Hachey J, Lynch DR, Kulcsar F, Van Rooijen G, Waterer DR, Robertson A, Kokko E, Byers R, Howard RJ, et al. (2001) Tomato Ve disease resistance genes encode cell surface-like receptors. Proc Natl Acad Sci U S A 98: 6511–6515
Kazmi RH, Khan N, Willems LAJ, Van Heusden AW, Ligterink W, Hilhorst HWM (2012) Complex genetics controls natural variation among seed quality phenotypes in a recombinant inbred population of an interspecific cross between *Solanum lycopersicum* × *Solanum pimpinellifolium*. Plant, Cell Environ. 35: 929–951.
Kazmi RH, Khan N, Willems LAJ, Van Heusden AW, Ligterink W, Hilhorst HWM (2012) Complex genetics controls natural variation among seed quality phenotypes in a recombinant inbred population of an interspecific cross between *Solanum lycopersicum*×*Solanum pimpinellifolium*. Plant, Cell Environ 35: 929–951
Keller M, Simm S (2018) The coupling of transcriptome and proteome adaptation during development and heat stress response of tomato pollen. BMC Genomics 19: 447

Kenchanmane Raju SK, Barnes AC, Schnable JC, Roston RL (2018) Low-temperature tolerance in land plants: Are transcript and membrane responses conserved? Plant Sci 276: 73–86
Kenneth J. Boote; Maria R. Rybak; Johan M.S. Scholberg; James W. Jones (2012) Improving the CROPGRO-Tomato Model for Predicting Growth and Yield Response to Temperature. HortScience 47:1038-1049
Khan TA, Fariduddin Q, Yusuf M (2015) *Lycopersicon esculentum* under low temperature stress: an approach toward enhanced antioxidants and yield. Environ Sci Pollut Res 22: 14178–14188
Kimbara J, Ohyama A, Chikano H, Ito H, Hosoi K, Negoro S, Miyatake K, Yamaguchi H, Nunome T, Fukuoka H, et al. (2018) QTL mapping of fruit nutritional and flavor components in tomato (*Solanum lycopersicum*) using genome-wide SSR markers and recombinant inbred lines (RILs) from an intra-specific cross. Euphytica 214: 210
Kinkade MP, Foolad MR (2013) Validation and fine mapping of lyc12.1, a QTL for increased tomato fruit lycopene content. Theor Appl Genet 126: 2163–2175
Kissoudis C, Chowdhury R, van Heusden S, van de Wiel C, Finkers R, Visser RGF, Bai Y, van der Linden G (2015) Combined biotic and abiotic stress resistance in tomato. Euphytica 202: 317–332
Klee HJ (2010) Improving the flavor of fresh fruits: genomics, biochemistry, and biotechnology. New Phytol. 187, 44–56.
Klee HJ (2013) Purple tomatoes: Longer lasting, less disease, and better for you. Curr Biol 23: R520–R521
Klee HJ, Tieman DM (2013) Genetic challenges of flavor improvement in tomato. Trends Genet 29: 257–262
Klee HJ, Tieman DM (2018) The genetics of fruit flavour preferences. Nat Rev Genet 19: 347–356
Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, Henning AK, Paul SanGiovanni J, Mane SM, Mayne ST, et al. (2005) Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science 308: 385–389
Kooke R, Kruijer W, Bours R, Becker F, Kuhn A, van de Geest H, Buntjer J, Doeswijk T, Guerra J, Bouwmeester H, et al. (2016) Genome-Wide Association Mapping and Genomic Prediction Elucidate the Genetic Architecture of Morphological Traits in Arabidopsis. Plant Physiol 170: 2187–2203
Kopeliovitch E, Mizrahi Y, Rabinowitch HD, Kedar N (1980) Physiology of the tomato mutant alcobaca. Physiol Plant 48: 307–311
Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M (2012) A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet 44: 1066–1071
Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, Durrant C, Mott R (2009) A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in *Arabidopsis thaliana*. PLoS Genet 5: e1000551
Kozukue N, Friedman M (2003) Tomatine, chlorophyll, β-carotene and lycopene content in tomatoes during growth and maturation. J Sci Food Agric 83: 195–200
Kramer M, Sanders R, Bolkan H, Waters C, Sheeny RE, Hiatt WR (1992) Postharvest evaluation of transgenic tomatoes with reduced levels of polygalacturonase: processing, firmness and disease resistance. Postharvest Biology and Technology 1(3):241-55.
Kramer MG, Redenbaugh K (1994) Commercialization of a tomato with an antisense polygalacturonase gene: The FLAVR SAVR? tomato story. Euphytica 79: 293–297
Krieger U, Lippman ZB, Zamir D (2010) The flowering gene SINGLE FLOWER TRUSS drives heterosis for yield in tomato. Nat Genet 42: 459–463
Kromdijk J, Bertin N, Heuvelink E, Molenaar J, de Visser PHB, Marcelis LFM, Struik PC (2013) Crop management impacts the efficiency of QTL detection and use - case study of fruit load x QTL interactions. Journal of Experimental Botany doi: 10.1093/jxb/ert365.
Kropff MJ, Haverkort AJ, Aggarwal PK and Kooman PL (1995) Using systems approaches to design and evaluate ideotypes for specific environments. In: J. Bouma, BAM. Bouman, JC. Luyten and HG. Zandstra (eds.). Eco-regional approaches for sustainable land use and food production. Dordrecht, Netherlands: Kluwer Academic Publ. 417-435.
Kumar M, Ashok, I, Chandrawat S (2016) Gene Pyramiding: An Overview. International Journal of Current Research in Biosciences and Plant Biology. DOI 10.20546/ijcrbp.2016.307.004
Kusmec A, Srinivasan S, Nettleton D, Schnable PS (2017) Distinct genetic architectures for phenotype means and plasticities in Zea mays. Nat Plants 3: 715–723
Labate JA, Grandillo S, Fulton T, Muños S, Caicedo AL, Peralta I, Ji Y, Chetelat RT, Scott JW, Gonzalo MJ, et al. (2007) Tomato. *In* C Kole, ed, Vegetables. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 1–125
Lanfermeijer FC, Warmink J, Hille J (2005) The products of the broken *Tm-2* and the durable *Tm-2(2)* resistance genes from tomato differ in four amino acids. Journal of Experimental Botany 56: 2925-2933
Lang Z, Wang Y, Tang K, Tang D, Datsenka T, Cheng J, Zhang Y, Handa AK, Zhu JK (2017) Critical roles of DNA demethylation in the activation of ripening-induced genes and inhibition of ripening-repressed genes in tomato fruit. Proceedings of the National Academy of Sciences 114(22):E4511-9.
Langlois D, Etiévant PX, Pierron P, Jorrot A (1996) Sensory and instrumental characterisation of commercial tomato varieties. Eur Food Res Technol 203: 534–540
Lapidot M, Karniel U, Gelbart D, Fogel D, Evenor D, Kutsher Y, Makhbash Z, Nahon S, Shlomo H, Chen L, Reuveni M, Levin I (2015) A Novel Route Controlling Begomovirus Resistance by the Messenger RNA Surveillance Factor *Pelota*. PLoS Genetics 11
Larbat R, Olsen KM, Slimestad R, Løvdal T, Bénard C, Verheul M, Bourgaud F, Robin C, Lillo C (2012) Influence of repeated short-term nitrogen limitations on leaf phenolics metabolism in tomato. Phytochemistry 77: 119–128
Laterrot H (1996) Twenty-one near isogenic lines in Moneymaker type with different genes for disease resistances. Rep Tomato Genet Coop 46: 34
Laterrot H (2000) Disease resistance in tomato: practical situation. Acta Physiologiae Plantarum 22: 328-331
Laterrot H, Moretti A (1989) Linkage between *Pto* and susceptibility to fenthion. Tomato Genet. Coop. Rep. 39: 21-22
Le LO, Lorenz Y, Scheurer S, Fötisch K, Enrique E, Bartra J, Biemelt S, Vieths S, Sonnewald U (2006) Design of tomato fruits with reduced allergenicity by dsRNAi- mediated inhibition of ns- LTP (Lyc e 3) expression. Plant Biotechnology Journal 4(2):231-42.
Lecompte F, Abro MA, Nicot PC (2010) Contrasted responses of Botrytis cinerea isolates developing on tomato plants grown under different nitrogen nutrition regimes. Plant Pathol 59: 891–899
Lecompte F, Nicot PC, Ripoll J, Abro MA, Raimbault AK, Lopez-Lauri F, Bertin N (2017) Reduced susceptibility of tomato stem to the necrotrophic fungus *Botrytis cinerea* is associated with a specific adjustment of fructose content in the host sugar pool. Annals of Botany 119: 931-943
Lecomte L, V Saliba-Colombani, A Gautier, MC Gomez-Jimenez, P Duffé, M Buret, M Causse (2004a) Fine mapping of QTLs for the fruit architecture and composition in fresh market tomato, on the distal region of the long arm of chromosome 2. Molecular Br 13: 1-14
Lecomte L, P Duffé, M Buret, B Servin, F Hospital, M Causse (2004b) Marker-assisted introgression of 5 QTLs controlling fruit quality traits into three tomato lines revealed interactions between QTLs and genetic backgrounds. Theor. Appl. Genet. 109: 658-668
Lee DR (1990) A unidirectional water flux model of fruit growth. Can. J. Bot. 68:1286-1290
Lee JT, Prasad V, Yang PT, Wu JF, David Ho TH, Charng YY, Chan MT (2003) Expression of Arabidopsis CBF1 regulated by an ABA/stress inducible promoter in transgenic tomato confers stress tolerance without affecting yield. Plant, Cell & Environment 26(7):1181-90.
Lee SY, Luna-Guzman I, Chang S, Barrett DM, Guinard JX (1999) Relating descriptive analysis and instrumental texture data of processed diced tomatoes. *Food Qual Pref* 10:447-455.
Lefebvre V, Boissot N, Gallois J-L (2018) Host plant resistance to pests and pathogens, the genetic leverage in integrated pest and disease management. In: Gullino ML, Albajes R, Nicot P, van Lenteren JC (eds) Pest and Disease Management in Greenhouse Crops. Developments in Plant Pathology. Springer International Publishing
Length F (2011) Genetic diversity in 14 tomato (*Lycopersicon esculentum* Mill .) varieties in Nigerian markets by RAPD-PCR technique. African Journal Of Biotechnology 10(11):4961-4967
Leonardi C, Ambrosino P, Esposito F, Fogliano V (2000) Antioxidative activity and carotenoid and tomatine contents in different typologies of fresh consumption tomatoes. J Agric Food Chem 48: 4723–4727
Letort V, Mahe P, Cournede PH, De Reffye P and Courtois B (2008) Quantitative genetics and functional-structural plant growth models: Simulation of quantitative trait loci detection for model parameters and application to potential yield optimization. Ann Bot-London 101: 1243-1254
Levin I, Gilboa N, Yeselson E, Shen S, Schaffer AA (2000) Fgr, a major locus that modulates the fructose to glucose ratio in mature tomato fruits. Theor Appl Genet 100: 256–262
Li L, Liu L, Bai Y, Zhang P, Finkers R, D. Y, et al. (2011) Seedling salt tolerance in tomato. *Euphytica*. 178: 403–414
Li T, Yang X, Yu Y, Si X, Zhai X, Zhang H, Dong W, Gao C, Xu C (2018) Domestication of wild tomato is accelerated by genome editing. Nature biotechnology 36, 1160–1163
Lin KH, Yeh WL, Chen HM, Lo HF (2010) Quantitative trait loci influencing fruit-related characteristics of tomato grown in high-temperature conditions. Euphytica. 174: 119–135.
Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, Zhang Z, Lun Y, Li S, Wang X, et al. (2014) Genomic analyses provide insights into the history of tomato breeding. Nat Genet 46: 1220–1226
Lippman ZB, Zamir D (2007) Heterosis: revisiting the magic. Trends Genet 23: 60–66
Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie AR (2006) Gas chromatography mass spectrometry-based metabolite profiling in plants. Nat Protoc 1: 387–396

Liu H, Genard M, Guichard S, Bertin N (2007) Model-assisted analysis of tomato fruit growth in relation to carbon and water fluxes. Journal of Experimental Botany 58:3567-3580.
Liu HJ, Yan J (2019) Crop genome-wide association study: a harvest of biological relevance. Plant J 97: 8–18
Liu J, Van Eck J, Cong B, Tanksley SD (2002) A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. Proc Natl Acad Sci 99: 13302–13306
Liu Y, Zhou T, Ge H, Pang W, Gao L, Ren L et al. (2016) SSR Mapping of QTLs Conferring Cold Tolerance in an Interspecific Cross of Tomato. Int J Genomics. 2016: 1–6.
Liu Z, Alseekh S, Brotman Y, Zheng Y, Fei Z, Tieman DM, Giovannoni JJ, Fernie AR, Klee HJ (2016) Identification of a Solanum pennellii Chromosome 4 Fruit Flavor and Nutritional Quality-Associated Metabolite QTL. Front Plant Sci 7: 1–15
Lobit P, Génard M, Soing P, Habib R (2006) Modelling malic acid accumulation in fruits: relationships with organic acids, potassium, and temperature. J. Exp. Bot. 57,1471–1483.
Lobit P, Génard M, Wu BH, Soing P, Habib R (2003) Modelling citrate metabolism in fruits: responses to growth and temperature. J. Exp. Bot. 54, 2489–2501.
Lü P, Yu S, Zhu N, Chen Y-R, Zhou B, Pan Y, Tzeng D, Fabi JP, Argyris J, Garcia-mas J, et al. (2018) Genome encode analyses reveal the basis of convergent evolution of fleshy fruit ripening. Nat Plants 1
Luo J (2015) Metabolite-based genome-wide association studies in plants. Curr Opin Plant Biol 24: 31–38
Maayan Y, Pandaranayaka EPJ, Srivastava DA, Lapidot M, Levin I, Dombrovsky A, Harel A (2018) Using genomic analysis to identify tomato *Tm-2* resistance-breaking mutations and their underlying evolutionary path in a new and emerging tobamovirus. Archives of Virology 163: 1863-1875
Mackay IJ, Bansept-Basler P, Barber T, Bentley AR, Cockram J, Gosman N, Greenland AJ, Horsnell R, Howells R, O'Sullivan DM, et al. (2014) An Eight-Parent Multiparent Advanced Generation Inter-Cross Population for Winter-Sown Wheat: Creation, Properties, and Validation. G3 Genes|Genomes|Genetics 4: 1603–1610
Madhavi DL, Salunkhe DK (1998) Handbook of Vegetable Science and Technology. Handb Veg Sci Technol. doi: 10.1201/9781482269871
Malundo TMM, Shewfelt RL, Scott JW (1995) Flavor quality of fresh tomato (*Lycopersicon esculentum* Mill.) as affected by sugar and acid levels. Postharvest Biol Technol 6: 103–110
Manavella PA, Hagmann J, Ott F, Laubinger S, Franz M, Macek B, Weigel D (2012) Fast-forward genetics identifies plant CPL phosphatases as regulators of miRNA processing factor HYL1. Cell 151: 859–870
Mangin B, Rincent R, Rabier CE, Moreau L, Goudemand-Dugue E (2019) Training set optimization of genomic prediction by means of EthAcc. PLoS One 14: e0205629
Mangin B, Thoquet P, Olivier J, Grimsley NH (1999) Temporal and multiple quantitative trait loci analyses of resistance to bacterial wilt in tomato permit the resolution of linked loci. Genetics 151: 1165-1172
Manning K, Tör M, Poole M, Hong Y, Thompson AJ, King GJ, Giovannoni JJ, Seymour GB (2006) A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. Nat Genet 38: 948–952
Mao L, Begum D, Chuang H, Budiman MA, Szymkowiak EJ, Irish EE, Wing RA (2000) JOINTLESS is a MADS-box gene controlling tomato flower abscissionzone development. Nature 406: 910–913
Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11: 499–511
Marques de Carvalho L, Benda ND, Vaughan MM, Cabrera AR, Hung K, Cox T, Abdo Z, Allen LH, Teal PE (2015) *Mi-1*-Mediated Nematode Resistance in Tomatoes is Broken by Short-Term Heat Stress but Recovers Over Time. Journal of nematology 47: 133–140
Marschner H (1983) General Introduction to the Mineral Nutrition of Plants. Inorg. Plant Nutr. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 5–60
Martin GB, Brommonschenkel SH, Chunwongse J, Frary A, Ganal MW, Spivey R, Wu T, Earle ED, Tanksley SD, Sipvey R, et al. (1993) Map-based cloning of a protein kinase gene conferring disease resistance in tomato. Science (80-) 262: 1432–1436
Martin GB, Frary A, U TW, Brommonschenkel S, Chunwongse J, Earle ED, Tanksley SD (1994) A Member of the Tomato Pto Gene Family Confers Sensitivity to Fenthion Resulting in Rapid Cell Death.
Martre P, Bertin N, Salon C, Génard M (2011) Modelling the size and composition of fruit, grain and seed by process-based simulation models. New Phytologist Tansley Review 191: 601–618
Martre P, Quilot-Turion B, Luquet D, Ould-Sidi M-M, Chenu K, Debaeke P (2015) Chapter 14 - Model-assisted phenotyping and ideotype design. In Crop Physiology (Second Edition) Applications for Genetic Improvement and Agronomy 2015 Ac. Press. pp349-373
Mazzucato A, Cellini F, Bouzayen M, Zouine M, Mila I, Minoia S, Petrozza A, Picarella ME, Ruiu F, Carriero F (2015) A TILLING allele of the tomato Aux/IAA9 gene offers new insights into fruit set mechanisms and perspectives for breeding seedless tomatoes. Mol Breed 35: 22
Mazzucato A, Papa R, Bitocchi E, Mosconi P, Nanni L, Negri V, Picarella ME, Siligato F, Soressi GP, Tiranti B, et al. (2008) Genetic diversity, structure and marker-trait associations in a collection of Italian tomato (*Solanum lycopersicum* L.) landraces. Theor Appl Genet 116: 657–669
Mboup M, Fischer I, Lainer H, Stephan W (2012) Trans-Species Polymorphism and Allele-Specific Expression in the CBF Gene Family of Wild Tomatoes. Mol Biol Evol 29: 3641–3652
McCormick S, Niedermeyer J, Fry J, Barnason A, Horsch R, Fraley R. (1986) Leaf disc transformation of cultivated tomato (*L. esculentum*) using *Agrobacterium tumefaciens*. Plant Cell Reports 5(2):81-4.
McCouch SR, Wright MH, Tung C-W, Maron LG, McNally KL, Fitzgerald M, Singh N, DeClerck G, Agosto-Perez F, Korniliev P, et al. (2016) Open access resources for genome-wide association mapping in rice. Nat Commun 7: 10532
McGlasson WB, Last JH, Shaw KJ, Meldrum SK (1987) Influence of the non-ripening mutant rin and nor on the aroma of tomato fruits. HortScience 22: 632–634
Meena YK, Khurana DS, Singh K (2018) Towards enhanced low temperature stress tolerance in tomato : An approach. J Environ Biol. doi: 10.22438/jeb/39/4/MRN-590
Megraw M, Baev V, Rusinov V, Jensen ST, Kalantidis K, Hatzigeorgiou AG (2006) MicroRNA promoter element discovery in Arabidopsis. RNA 12: 1612–1619
Menda N, Semel Y, Peled D, Eshed Y, Zamir D (2004) *In silico* screening of a saturated mutation library of tomato. Plant J 38: 861–872
Menda N, Strickler SR, Edwards JD, Bombarely A, Dunham DM, Martin GB, Mejia L, Hutton SF, Havey MJ, Maxwell DP, et al. (2014) Analysis of wild-species introgressions in tomato inbreds uncovers ancestral origins. BMC Plant Biol 14: 287
Mendell JT, Olson EN (2012) MicroRNAs in stress signaling and human disease. Cell 148: 1172–87
Meng C, Yang D, Ma X, Zhao W, Liang X, Ma N, Meng Q. (2016) Suppression of tomato SlNAC1 transcription factor delays fruit ripening. Journal of plant physiology. 193:88-96.
Meng FJ, Xu XY, Huang FL, Li JF (2010) Analysis of genetic diversity in cultivated and wild tomato varieties in Chinese market by RAPD and SSR. Agric Sci China 9: 1430–1437
Messeguer R, Ganal M, de Vicente MC, Young ND, Bolkan H, Tanksley SD (1991) High resolution RFLP map around the root knot nematode resistance gene (*Mi*) in tomato. Theoretical and Applied Genetics 82: 529-536.
Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829
Migault V, Pallas B, Costes E (2017) Combining genome-wide information with a functional structural plant model to simulate 1-year-old apple tree architecture. Frontiers in plant science https://doi.org/10.3389/fpls.2016.02065
Miller JC, Tanksley SD (1990) RFLP analysis of phylogenetic relationships and genetic variation in the genus Lycopersicon. Theor Appl Genet 80: 437–448
Milligan SB, Bodeau J, Yaghoobi J, Kaloshian I, Zabel P, Williamson VM (1998) The root knot nematode resistance gene *Mi* from tomato is a member of the leucine zipper, nucleotide binding, leucine-rich repeat family of plant genes. Plant Cell (The) 10: 1307-1319
Milner S. *et al.* (2011) Bioactivities of glycoalkaloids and their aglycones from Solanum species. J. Agric. Food Chem. 59, 3454–3484
Minamikawa MF, Nonaka K, Kaminuma E, Kajiya-Kanegae H, Onogi A, Goto S, Yoshioka T, Imai A, Hamada H, Hayashi T, et al. (2017) Genome-wide association study and genomic prediction in citrus: Potential of genomics-assisted breeding for fruit quality traits. Sci Rep 7: 4721
Minoïa S, Bendahmane A, Piron F, Salgues A, Moretti A, Caranta C, Piednoir E, Nicolaï M, Zamir D (2010) An Induced Mutation in Tomato eIF4E Leads to Immunity to Two Potyviruses. PLoS One 5: e11313
Minoia S, Cellini F, Bendahmane A, D'Onofrio O, Petrozza A, Carriero F, Piron F, Mosca G, Sozio G (2010) A new mutant genetic resource for tomato crop improvement by TILLING technology. BMC Res Notes. doi: 10.1186/1756-0500-3-69
Mirnezhad, M., Romero-Gonzalez, R. R., Leiss, K. A., Choi, Y. H., Verpoorte, R., & Klinkhamer, P. G. (2010). Metabolomic analysis of host plant resistance to thrips in wild and cultivated tomatoes. Phytochemical Analysis, 21(1), 110–117.
Mirouze M, Paszkowski J (2011) Epigenetic contribution to stress adaptation in plants. Curr Opin Plant Biol 14: 267–274
Mitchell J, Shennan C, Grattan S (1991) Developmental-Changes in Tomato Fruit Composition in Response To Water Deficit and Salinity. Physiol Plant 83: 177–185
Mohorianu I, Schwach F, Jing R, Lopez-Gomollon S, Moxon S, Szittya G, Sorefan K, Moulton V, Dalmay T (2011) Profiling of short RNAs during fleshy fruit development reveals stage-specific sRNAome expression patterns. Plant J 67: 232–246
Molgaard P, Ravn H (1988) Evolutionary aspects of caffeoyl ester distribution in dicotyledons. Phytochemistry 27:2411-2421
Monforte AJ, Asíns MJ, Carbonell EA (1996) Salt tolerance in Lycopersicon species. IV. Efficiency of marker-assisted selection for salt

tolerance improvement. Theor Appl Genet 93–93: 765–772
Monforte AJ, Asíns MJ, Carbonell EA (1997a) Salt tolerance in Lycopersicon species VI. Genotype-by-salinity interaction in quantitative trait loci detection: constitutive and response QTLs. Theor Appl Genet 95: 706–713
Monforte AJ, Asíns MJ, Carbonell EA (1997b) Salt tolerance in Lycopersicon species. V. Does genetic variability at quantitative trait loci affect their analysis? Theor Appl Genet 95: 284–293
Monforte AJ, Asíns MJ, Carbonell EA (1999) Salt tolerance in Lycopersicon spp. VII. Pleiotropic action of genes controlling earliness on fruit yield. Theor Appl Genet 98: 593–601
Monforte AJ, Tanksley SD (2000) Fine mapping of a quantitative trait locus (QTL) from *Lycopersicon hirsutum* chromosome 1 affecting fruit characteristics and agronomic traits: breaking linkage among QTLs affecting different traits and dissection of heterosis for yield. Theor Appl Genet 100: 471–479
Montesinos-López OA, Montesinos-López A, Crossa J, Toledo FH, Pérez-Hernández O, Eskridge KM, Rutkoski J (2016) A Genomic Bayesian Multi-trait and Multi-environment Model. G3 Genes|Genomes|Genetics 6: 2725–2744
Moxon S, Jing R, Szittya G, Schwach F, Rusholme Pilcher RL, Moulton V, Dalmay T (2008) Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening. Genome Res 18: 1602–9
Mu Q, Huang Z, Chakrabarti M, Illa-Berenguer E, Liu X, Wang Y, Ramos A, van der Knaap E (2017) Fruit weight is controlled by Cell Size Regulator encoding a novel protein that is expressed in maturing tomato fruits. PLOS Genet 13: e1006930
Muir SR, Collins GJ, Robinson S, Hughes S, Bovy A, Ric De Vos CH, van Tunen AJ, Verhoeyen ME (2001) Overexpression of petunia chalcone isomerase in tomato results in fruit containing increased levels of flavonols. Nat Biotechnol 19: 470–474
Müller BSF, Neves LG, de Almeida Filho JE, Resende MFR, Muñoz PR, dos Santos PET, Filho EP, Kirst M, Grattapaglia D (2017) Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of Eucalyptus. BMC Genomics 18: 524
Munns R, Gilliham M (2015) Salinity tolerance of crops - what is the cost? New Phytol 208: 668–673
Munns R, Tester M (2008) Mechanisms of Salinity Tolerance. Annu Rev Plant Biol 59: 651–681
Mutshinda CM, Sillanpää MJ (2010) Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. Genetics 186: 1067–75
Nadeem M, Li J, Wang M, Shah L, Lu S, Wang X, Ma C, Nadeem M, Li J, Wang M, et al. (2018) Unraveling Field Crops Sensitivity to Heat Stress : Mechanisms, Approaches, and Future Prospects. Agronomy 8: 128
Nakatani N (2000) Phenolic antioxidants from herbs and spices. BioFactors 13: 141–146
Nakazato T, Warren DL, Moyle LC (2010) Ecological and geographic modes of species divergence in wild tomatoes. Am J Bot 97: 680–693
Nardini M, D'Aquino M, Tomassi G, Gentili V, Di Felice M, Scaccini C (1995) Inhibition of human low-density lipoprotein oxidation by caffeic acid and other hydroxycinnamic acid derivatives. Free Radic Biol Med 19: 541–552
Navarro JM, Flores P, Carvajal M, Martinez V (2005) Changes in quality and yield of tomato fruit with ammonium, bicarbonate and calcium fertilisation under saline conditions. J Hortic Sci Biotechnol 80: 351–357
Naves ER, de Ávila Silva L, Sulpice R, Araújo WL, Nunes-Nesi A, Peres LE, Zsögön A. Capsaicinoids: pungency beyond Capsicum. Trends in plant science. 2019 Jan 7.
Nekrasov V, Wang C, Win J, Lanz C, Weigel D, Kamoun S (2017) Rapid generation of a transgene-free powdery mildew resistant tomato by genome deletion. Sci. Rep. 7, 482
Nesbitt TC, Tanksley SD (2002) Comparative Sequencing in the Genus Lycopersicon: Implications for the Evolution of Fruit Size in the Domestication of Cultivated Tomatoes. Genetics 162: 365–379
Nguyen K Le, Grondin A, Courtois B, Gantet P (2018) Next-Generation Sequencing Accelerates Crop Gene Discovery. Trends Plant Sci 24: 263–274
Nombela G, Williamson VM, Muniz M (2003) The root-knot nematode resistance gene *Mi-1.2* of tomato is responsible for resistance against the whitefly *Bemisia tabaci*. Molecular Plant-Microbe Interactions 16: 645-649
Nuruddin MM, Madramootoo CA, Dodds GT (2003) Effects of Water Stress at Different Growth Stages on Greenhouse Tomato Yield and Quality. HortScience 38: 1389–1393
Ofner I, Lashbrooke J, Pleban T, Aharoni A, Zamir D (2016) Solanum pennellii backcross inbred lines (BILs) link small genomic bins with tomato traits. Plant J 87: 151–160
Ohama N, Sato H, Shinozaki K, Yamaguchi-Shinozaki K (2017) Transcriptional Regulatory Network of Plant Heat Stress Response. Trends Plant Sci 22: 53–65
Ohlson EW, Ashrafi H, Foolad MR (2018) Identification and Mapping of Late Blight Resistance Quantitative Trait Loci in Tomato Accession PI 163245. Plant Genome 11
Ohlson EW, Foolad MR (2016) Genetic analysis of resistance to tomato late blight in Solanum pimpinellifolium accession PI 163245. Plant Breeding 135: 391-398
Okabe Y, Asamizu E, Saito T, Matsukura C, Ariizumi T, Brès C, Rothan C, Mizoguchi T, Ezura H (2011) Tomato TILLING Technology: Development of a Reverse Genetics Tool for the Efficient Isolation of Mutants from Micro-Tom Mutant Libraries. Plant Cell Physiol 52: 1994–2005
Okello RCO,Heuvelink E, de Visser PHB, Struik PC, Marcelis LFM (2015) What drives fruit growth? Functional Plant Biology 42, 817-827
Oliver JE, Whitfield AE (2016) The Genus Tospovirus: Emerging Bunyaviruses that Threaten Food Security. In: Enquist LW (ed) Annual Review of Virology 3, 101-124
Ongom PO, Ejeta G (2017) Mating Design and Genetic Structure of a Multi-Parent Advanced Generation Intercross (MAGIC) Population of Sorghum (*Sorghum bicolor* (L.) Moench). G3 Genes|Genomes|Genetics 8: 331–341
Osorio S, Ruan Y-L, Fernie AR (2014) An update on source-to-sink carbon partitioning in tomato. Front Plant Sci 5: 516
Ould-Sidi M-M, Lescourret F (2011) Model-based design of innovative cropping systems: state of the art and new prospects. Agronomy for sustainable development 31, 3, 571-588.
Overy SA, Walker HJ, Malone S, Howard TP, Baxter CJ, Sweetlove LJ, Hill SA, Quick WP (2004) Application of metabolite profiling to the identification of traits in a population of tomato introgression lines. J Exp Bot 56: 287–296
Pailles Y, Ho S, Pires IS, Tester M, Negrão S, Schmöckel SM (2017) Genetic Diversity and Population Structure of Two Tomato Species from the Galapagos Islands. Front Plant Sci 8: 138
Panthee DR, Piotrowski A, Ibrahem R (2017) Mapping Quantitative Trait Loci (QTL) for Resistance to Late Blight in Tomato. International Journal of Molecular Sciences 18
Papadopoulos I, Rendig V V. (1983) Interactive effects of salinity and nitrogen on growth and yield of tomato plants. Plant Soil 73: 47–57
Paran I, Goldman I, Tanksley SD, Zamir D (1995) Recombinant inbred lines for genetic mapping in tomato. Theor Appl Genet 90: 542–548
Park T, Casella G (2008) The Bayesian Lasso. J Am Stat Assoc 103: 681–686
Park YH, West MA., St. Clair DA (2004) Evaluation of AFLPs for germplasm fingerprinting and assessment of genetic diversity in cultivars of tomato (*Lycopersicon esculentum* L.) . Genome 47: 510–518
Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, Gupta N, Neale BM, Daly MJ, Sklar P, et al. (2012) Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet 44: 631–635
Pascale S De, Maggio A, Fogliano V, Ambrosino P, Ritieni A (2001) Irrigation with saline water improves carotenoids content and antioxidant activity of tomato. J Hortic Sci Biotechnol 76: 447–453
Pascual L, Desplat N, Huang BE, Desgroux A, Bruguier L, Bouchet JP, Le QH, Chauchard B, Verschave P, Causse M (2015) Potential of a tomato MAGIC population to decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. Plant Biotechnol J 13: 565–577
Patanè C, Cosentino SL (2010) Effects of soil water deficit on yield and quality of processing tomato under a Mediterranean climate. Agric Water Manag 97: 131–138
Paterson AH, Damon S, Hewitt JD, Zamir D, Rabinowitch HD, Loncoln SE, Lander ES, Tanksley SD (1991) Mendelian factors underlying quantitative traits in tomato: Comparison across species, generations, and environments. Genetics 127: 181–197
Paterson AH, DeVerna JW, Lanini B, Tanksley SD (1990) Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes, in an interspecies cross of tomato. Genetics 124: 735–742
Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD (1988) Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature 335: 721–726
Pattison RJ, Csukasi F, Zheng Y, Fei Z, van der Knaap E, Catalá C. (2015) Comprehensive Tissue-Specific Transcriptome Analysis Reveals Distinct Regulatory Programs during Early Tomato Fruit Development. Plant Physiol.168(4):1684–1701.
Peralta IE, Knapp S, Spooner DM (2005) New Species of Wild Tomatoes (Solanum Section Lycopersicon: Solanaceae) from Northern Peru. Syst Bot 30: 424–434
Pertuzé RA, Ji Y, Chetelat RT (2003) Comparative linkage map of the *Solanum lycopersicoides* and *S. sitiens* genomes and their differentiation from tomato . Genome 45: 1003–1012
Petró-Turza M (1986) Flavor of tomato and tomato products. Food Rev Int 2: 309–351
Pettigrew WT (2008) Potassium influences on yield and quality production for maize, wheat, soybean and cotton. Physiol Plant 133: 670–681
Philouze J (1991) Description of isogenic lines, except for one, or two, monogenically controlled morphological traits in tomato, *Lycopersicon esculentum* Mill. Euphytica 56: 121–131

Pillen K, Ganal MW, Tanksley SD (1996) Construction of a high-resolution genetic map and YAC-contigs in the tomato *Tm-2a* region. Theoretical and Applied Genetics 93: 228-233
Piron F, Nicolai M, Minoia S, Piednoir E, Moretti A, Salgues A, Zamir D, Caranta C, Bendahmane A (2010) An Induced Mutation in Tomato *eIF4E* Leads to Immunity to Two Potyviruses. Plos One 5
Pnueli L, Carmel-Goren L, Hareven D, Gutfinger T, Alvarez J, Ganal M, Zamir D, Lifschitz E (1998) The SELF-PRUNING gene of tomato regulates vegetative to reproductive switching of sympodial meristems and is the ortholog of CEN and TFL1. Development 125(11):1979-89
Poiroux-Gonord F, Bidel LPR, Fanciullino A-L, Gautier H, Lauri-Lopez F, Urban L (2010) Health Benefits of Vitamins and Secondary Metabolites of Fruits and Vegetables and Prospects To Increase Their Concentrations by Agronomic Approaches. J Agric Food Chem 58: 12065–12082
Powell ALT, Kalamaki MS, Kurien PA, Gurrieri S, Bennett AB (2003) Simultaneous Transgenic Suppression of LePG and LeExp1 Influences Fruit Texture and Juice Viscosity in a Fresh Market Tomato Variety. J Agric Food Chem 51: 7450–7455
Prudent M, Lecomte A, Bouchet JP, Bertin N, Causse M, Génard M (2011) Combining ecophysiological modelling and quantitative trait loci analysis to identify key elementary processes underlying tomato fruit sugar concentration. Journal of Experimental Botany 62: 907–919
Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell 152(5):1173-83.
Quadrana L, Almeida J, Asís R, Duffy T, Dominguez PG, Bermúdez L, Conti G, Corrêa da Silva J V., Peralta IE, Colot V, et al. (2014) Natural occurring epialleles determine vitamin E accumulation in tomato fruits. Nat Commun 5: 4027
Quilot B, Kervella J, Genard M, Lescourret F.(2005) Analysing the genetic control of peach fruit quality through an ecophysiological model combined with a QTL approach.Journal of Experimental Botany 56:3083–3092.
Quilot-Turion B, Génard M, Valsesia P, Memmah M-M. (2016) Optimization of allelic combinations controlling parameters of a Peach quality model. Frontiers in Plant Science, 7, 1873.
Quilot-Turion B, Ould-Sidi M-M, Kadrani A, Hilgert N, Génard M, Lescourret F (2012) Optimization of parameters of the 'Virtual Fruit' model to design peach genotype for sustainable production systems. European Journal of Agronomy, 42: 34-48.
Quinet M, Kinet J-M, Lutts S (2011) Flowering response of the uniflora:blind:self-pruning and jointless:uniflora:self-pruning tomato (*Solanum lycopersicum*) triple mutants. Physiol Plant 141: 166–176
Rajasekaran LR, Aspinall D, Paleg LG (2000) Physiological mechanism of tolerance of *Lycopersicon* spp. exposed to salt stress. Can J Plant Sci 80: 151–159
Rajewsky N (2006) microRNA target predictions in animals. Nat Genet 38: S8–S13
Rambla JL, Medina A, Fernández-del-Carmen A, Barrantes W, Grandillo S, Cammareri M, López-Casado G, Rodrigo G, Alonso A, García-Martínez S, et al. (2016) Identification, introgression, and validation of fruit volatile QTLs from a red-fruited wild tomato species. J Exp Bot 68: erw455
Rambla JL, Tikunov YM, Monforte AJ, Bovy AG and Granell A (2014) The expanded tomato fruit volatile landscape. , 1–11.
Ramstein GP, Jensen SE, Buckler ES (2018) Breaking the curse of dimensionality to identify causal variants in Breeding 4. Theor Appl Genet 1–9
Ranc N, Muños S, Xu J, Le Paslier M-C, Chauveau A, Bounon R, Rolland S, Bouchet J-P, Brunel D, Causse M (2012) Genome-Wide Association Mapping in Tomato (*Solanum lycopersicum*) Is Possible Using Genome Admixture of *Solanum lycopersicum* var. *cerasiforme*. G3 Genes|Genomes|Genetics 2: 853–864
Ranjan A, Budke JM, Rowland SD, Chitwood DH, Kumar R, Carriedo L et al. (2016) eQTL Regulating Transcript Levels Associated with Diverse Biological Processes in Tomato. Plant Physiol. 172(1):328-40.
Rao ES, Kadirvel P, Symonds RC, Ebert AW (2013) Relationship between survival and yield related traits in Solanum pimpinellifolium under salt stress. Euphytica 190: 215–228
Rasmussen S, Barah P, Suarez-Rodriguez MC, Bressendorff S, Friis P, Costantino P, Bones AM, Nielsen HB, Mundy J (2013) Transcriptome Responses to Combinations of Stresses in Arabidopsis. Plant Physiol 161: 1783–1794
Renard CM, Ginies C, Gouble B, Bureau S, Causse M (2013) Home conservation strategies for tomato (*Solanum lycopersicum*): Storage temperature vs. duration - Is there a compromise for better aroma preservation? Food Chem 139(1-4):825–836
Reymond M, Muller B., Leonardi A, Charcosset A, Tardieu F (2003) Combining quantitative trait loci analysis and an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to temperature and water deficit. Plant Physiol. 131:664-675.
Rick CM, Chetelat RT (1995) Utilization of related wild species for tomato improvement. In: FernandezMunoz R, Cuartero J, GomezGuillamon ML (eds) First International Symposium on Solanacea for Fresh Market, pp 21-38
Ripoll J, Urban L, Brunel B, Bertin N (2016) Water deficit effects on tomato quality depend on fruit developmental stage and genotype. J Plant Physiol 190: 26–35
Ripoll J, Urban L, Staudt M, Lopez-Lauri F, Bidel LPR, Bertin N (2014) Water shortage and quality of fleshy fruits—making the most of the unavoidable. J Exp Bot 65: 4097–4117
Rivero RM, Mestre TC, Mittler R, Rubio F, Garcia-Sanchez F, Martinez V (2014) The combined effect of salinity and heat reveals a specific physiological, biochemical and molecular response in tomato plants. Plant Cell Environ 37: 1059–1073
Robbins MD, Masud MAT, Panthee DR, Gardner RG, Francis DM, Stevens MR (2010) Marker-assisted Selection for Coupling Phase Resistance to *Tomato spotted wilt virus* and *Phytophthora infestans* (Late Blight) in Tomato. Hortscience 45: 1424-1428
Robert VJM, West MAL, Inai S, Caines A, Arntzen L, Smith JK, St Clair DA (2001) Marker-assisted introgression of blackmold resistance QTL alleles from wild *Lycopersicon cheesmanii* to cultivated tomato (*L. esculentum*) and evaluation of QTL phenotypic effects. Molecular Breeding 8: 217-233
Rodríguez GR, Muños S, Anderson C, Sim S-C, Michel A, Causse M, Gardener BBM, Francis D, Knaap E van der (2011) Distribution of SUN, OVATE, LC, and FAS in the Tomato Germplasm and the Relationship to Fruit Shape Diversity. Plant Physiol 156: 275–285
Rodríguez-Leal D, Lemmon ZH, Man J, Bartlett ME, Lippman ZB (2017) Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing. Cell 171: 470-480.
Rogers K, Chen X (2013) Biogenesis, Turnover, and Mode of Action of Plant MicroRNAs. Plant Cell 25: 2383–2399
Ronen G, Cohen M, Zamir D and Hirschberg J (1999) Regulation of carotenoid biosynthesis during tomato fruit development: expression of the gene for lycopene epsilon-cyclase is down-regulated during ripening and is elevated in the mutant Delta. Plant J, 17, 341–351.
Rosales MA, Rubio-Wilhelmi MM, Castellano R, Castilla N, Ruiz JM, Romero L (2007) Sucrolytic activities in cherry tomato fruits in relation to temperature and solar radiation. Sci Hortic (Amsterdam) 113: 244–249
Rosental L, Perelman A, Nevo N, Toubiana D, Samani T, Batushansky A, Sikron N, Saranga Y, Fait A (2016) Environmental and genetic effects on tomato seed metabolic balance and its association with germination vigor. BMC Genomics 17: 1047
Rossi M, Goggin FL, Milligan SB, Kaloshian I, Ullman DE, Williamson VM (1998) The nematode resistance gene *Mi* of tomato confers resistance against the potato aphid. Proceedings of the National Academy of Sciences of the United States of America 95: 9750-9754
Rothan C, Diouf I, Causse M (2019) Trait discovery and editing in tomato. Plant J 97: 73–90
Rousseaux MC, Jones CM, Adams D, Chetelat R, Bennett A, Powell A (2005) QTL analysis of fruit antioxidants in tomato using *Lycopersicon pennellii* introgression lines. Theor Appl Genet 111: 1396–1408
Ruan Y-L, Patrick JW, Bouzayen M, Osorio S, Fernie AR (2012) Molecular regulation of seed and fruit set. Trends Plant Sci 17: 656–665
Ruffel S, Gallois JL, Lesage ML, Caranta C (2005) The recessive potyvirus resistance gene *pot-1* is the tomato orthologue of the pepper *pvr2-eIF4E* gene. Molecular Genetics and Genomics 274: 346-353
Ruggieri V, Francese G, Sacco A, Alessandro AD, Rigano MM, Parisi M, Milone M, Cardi T, Mennella G, Barone A (2014) An association mapping approach to identify favourable alleles for tomato fruit quality breeding. BMC Plant Biol 14: 1–15
Sacco A, Di Matteo A, Lombardi N, Trotta N, Punzo B, Mari A, Barone A (2013) Quantitative trait loci pyramiding for fruit quality traits in tomato. Mol Breed. 31(1):217-222
Sahu KK, Chattopadhyay D (2017) Genome-wide sequence variations between wild and cultivated tomato species revisited by whole genome sequence mapping. BMC Genomics 18: 430
Sainju UM, Dris R, Singh B (2003) Mineral nutrition of tomato. Food Agric Environ 1: 176–183
Saliba-Colombani V, Causse M, Langlois D, Philouze J, Buret M (2001) Genetic analysis of organoleptic quality in fresh market tomato: 1. Mapping QTLs for physical and chemical traits. Theor Appl Genet 102: 259-272.
Sallam A, Martsch R (2015) Association mapping for frost tolerance using multi-parent advanced generation inter-cross (MAGIC) population in faba bean (Vicia faba L.). Genetica 143: 501–514
Salmeron JM, Oldroyd GE., Rommens CM., Scofield SR, Kim H-S, Lavelle DT, Dahlbeck D, Staskawicz BJ (1996) Tomato Prf Is a Member of the Leucine-Rich Repeat Class of Plant Disease Resistance Genes and Lies Embedded within the Pto Kinase Gene Cluster. Cell 86: 123–133
Salmeron JM, Oldroyd GED, Rommens CMT, Scofield SR, Kim HS, Lavelle DT, Dahlbeck D, Staskawicz BJ (1996) Tomato *Prf* is a member of the leucine-rich repeat class of plant disease resistance genes and lies embedded within the *Pto* kinase gene cluster. Cell 86: 123-133
Sanei M, Chen X (2015) Mechanisms of microRNA turnover. Curr Opin Plant Biol 27: 199–206
Sarlikioti V, de Visser PHB, Buck-Sorlin GH, Marcelis LFM (2011) How plant architecture affects light absorption and photosynthesis in tomato: towards an ideotype for plant architecture using a functional–structural plant model. Annals of Botany 108, 6, 1065–1073

Sato S, Tabata S, Hirakawa H, et al. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485, 635–641.
Sauvage C, Rau A, Aichholz C, Chadoeuf J, Sarah G, Ruiz M, Santoni S, Causse M, David J, Glémin S (2017) Domestication rewired gene expression and nucleotide diversity patterns in tomato. Plant J 91: 631–645
Sauvage C, Segura V, Bauchet G, Stevens R, Do PT, Nikoloski Z, Fernie AR, Causse M (2014) Genome-Wide Association in Tomato Reveals 44 Candidate Loci for Fruit Metabolic Traits. Plant Physiol 165: 1120–1132
Schachtman DP, Shin R (2007) Nutrient Sensing and Signaling: NPKS. Annu Rev Plant Biol 58: 47–69
Schaffer AA, Levin I, Oguz I, Petreikov M, Cincarevsky F, Yeselson Y, Shen S, Gilboa N, Bar M (2000) ADPglucose pyrophosphorylase activity and starch accumulation in immature tomato fruit: the effect of a *Lycopersicon hirsutum*-derived introgression encoding for the large subunit. Plant Sci 152: 135–144
Schauer N, Semel Y, Roessner U, Gur A, Balbo I, Carrari F, Pleban T, Perez-Melis A, Bruedigam C, Kopka J, et al. (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. Nat Biotechnol 24: 447–454
Schauer N, Zamir D, Fernie AR (2005) Metabolic profiling of leaves and fruit of wild species tomato: A survey of the *Solanum lycopersicum* complex. J Exp Bot 56: 297–307
Scheben A, Batley J, Edwards D (2017) Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. Plant Biotechnol J 15: 149–161
Schijlen EG, de Vos CR, Martens S, Jonker HH, Rosin FM, Molthoff JW, Tikunov YM, Angenent GC, van Tunen AJ, Bovy AG (2007) RNA interference silencing of chalcone synthase, the first step in the flavonoid biosynthesis pathway, leads to parthenocarpic tomato fruits. Plant physiology 144(3): 1520-30.
Scholberg JMS, Locascio SJ (1999) Growth response of snap bean and tomato as affected by salinity and irrigation method. HortScience 34: 259–264
Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, Nordborg M (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44: 825–830
Semel Y, Nissenbaum J, Menda N, Zinder M, Krieger U, Issman N, Pleban T, Lippman Z, Gur A, Zamir D (2006) Overdominant quantitative trait loci for yield and fitness in tomato. Proc Natl Acad Sci 103: 12981–12986
Semel Y, Schauer N, Roessner U, Zamir D, Fernie AR (2007) Metabolite analysis for the comparison of irrigated and non-irrigated field grown tomato of varying genotype. Metabolomics 3: 289–295
Shahlaei A, Torabi S, Khosroshahli M (2014) Efficiacy of SCoT and ISSR marekers in assesment of tomato (*Lycopersicum esculentum* Mill.) genetic diversity. Int J Biosci 5: 14–22
Shalit A, Rozman A, Goldshmidt A, Alvarez JP, Bowman JL, Eshed Y, Lifschitz E (2009) The flowering hormone florigen functions as a general systemic regulator of growth and termination. Proc Natl Acad Sci 106: 8392–8397
Shammai A, Petreikov M, Yeselson Y, Faigenboim A, Moy-Komemi M, Cohen S, Cohen D, Besaulov E, Efrati A, Houminer N, et al. (2018) Natural genetic variation for expression of a SWEET transporter among wild species of *Solanum lycopersicum* (tomato) determines the hexose composition of ripening tomato fruit. Plant J 96: 343–357
Sharada MS, Kumari A, Pandey AK, Sharma S, Sharma P, Sreelakshmi Y, Sharma R (2017) Generation of genetically stable transformants by Agrobacterium using tomato floral buds. Plant Cell, Tissue and Organ Culture 129(2):299-312
Shimatani Z, Kashojiya S, Takayama M, Terada R, Arazoe T, Ishii H, Teramura H, Yamamoto T, Komatsu H, Miura K, Ezura H (2017) Targeted base editing in rice and tomato using a CRISPR-Cas9 cytidine deaminase fusion. Nature biotechnology 35(5):441-443.
Shinozaki Y, Nicolas P, Fernandez-Pozo N, Ma Q, Evanich DJ, Shi Y, Xu Y, Zheng Y, Snyder SI, Martin LBB, et al. (2018) High-resolution spatiotemporal transcriptome mapping of tomato fruit development and ripening. Nat Commun 9: 364
Sim S-C, Durstewitz G, Plieske J, Wieseke R, Ganal MW, van Deynze A, Hamilton JP, Buell CR, Causse M, Wijeratne S, et al. (2012b) Development of a large snp genotyping array and generation of high-density genetic maps in tomato. PLoS One. doi: 10.1371/journal.pone.0040563
Sim S-C, Robbins MD, Van Deynze A, Agee M, Francis DM (2010) Population structure and genetic differentiation associated with breeding history and selection in tomato (*Solanum lycopersicum* L.). Heredity (Edinb) 106: 927–935
Sim S-C, Van Deynze A, Stoffel K, Douches DS, Zarka D, Ganal MW, Chetelat RT, Hutton SF, Scott JW, Gardner RG, et al. (2012) High-Density SNP Genotyping of Tomato (*Solanum lycopersicum* L.) Reveals Patterns of Genetic Variation Due to Breeding. PLoS One 7: e45520
Sim SC, Robbins MD, Wijeratne S, Wang H, Yang WC, Francis DM (2015) Association Analysis for Bacterial Spot Resistance in a Directionally Selected Complex Breeding Population of Tomato. Phytopathology 105: 1437-1445
Smart CD, Tanksley SD, Mayton H, Fry WE (2007) Resistance to *Phytophthora infestans* in *Lycopersicon pennellii*. Plant Disease 91: 1045-1049
Smirnoff N, Wheeler GL (2000) Ascorbic acid in plants: biosynthesis and function. *Critical Reviews in Biochemistry and Molecular Biology* 35, 291–314.
Smith DL, Abbott JA, Gross KC (2002) Down-regulation of tomato β-galactosidase 4 results in decreased fruit softening. Plant physiology 129(4): 1755-62.
Soyk S, Lemmon ZH, Oved M, et al. (2017) Bypassing Negative Epistasis on Yield in Tomato Imposed by a Domestication Gene. *Cell*, 1–14.
Soyk S, Müller NA, Park SJ, Schmalenbach I, Jiang K, Hayama R, Zhang L, Van Eck J, Jiménez-Gómez JM, Lippman ZB (2017) Variation in the flowering gene SELF PRUNING 5G promotes day-neutrality and early yield in tomato. Nat Genet 49: 162–168
Spano R, Mascia T, Kormelink R, Gallitelli D (2015) Grafting on a Non-Transgenic Tolerant Tomato Variety Confers Resistance to the Infection of a *Sw5*-Breaking Strain of *Tomato spotted wilt virus* via RNA Silencing. Plos One 10
Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redoña E, Atlin G, Jannink JL, McCouch SR (2015) Genomic Selection and Association Mapping in Rice (Oryza sativa): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. PLoS Genet 11: 1–25
Stamova BS, Chetelat RT (2000) Inheritance and genetic mapping of *cucumber mosaic virus* resistance introgressed from *Lycopersicon chilense* into tomato. Theoretical and Applied Genetics 101: 527-537
Stevens MA (1972) Citrate and malate concentration in tomato fruits: Genetic control and maturational effects. J Am Soc Hortic Sci 97: 655–658
Stevens MA (1986) Inheritance of Tomato Fruit Quality Components. Plant Breed Rev 4: 273–311
Stevens MA, Kader AA, Albright M (1979) Potential for increasing tomato flavor via increased sugar and acid content. J Am Soc Hortic Sci 104: 40–42
Stevens MA, Kader AA, Albright-Holton M (1977) Intercultivar variation in composition of locular and pericarp portions of fresh market tomatoes. J Am Soc Hortic Sci 102: 689–692
Stevens MR, Lamb EM, Rhoads DD (1995) Mapping the *Sw-5* locus for *tomato spotted wilt virus*-resistance in tomatoes using RAPD and RFLP analyses. Theoretical and Applied Genetics 90: 451-456
Stevens R, Buret M, Duffe P, Garchery C, Baldet P, Rothan C, Causse M (2007) Candidate Genes and Quantitative Trait Loci Affecting Fruit Ascorbic Acid Content in Three Tomato Populations. Plant Physiol 143: 1943–1953
Stikic R, Popovic S, Srdic M, Savic D, Jovanovic Z, Zdravkovic J (2003) Partial root drying (PRD): A new technique for growing plants that saves water and improves the quality of fruit. Bulg J Plant Physiol 164–171
Stommel JR (2001) USDA 97L63, 97L66, and 97L97: Tomato breeding lines with high fruit beta-carotene content. HortScience 36: 387–388
Stommel JR, Abbott JA, Saftner RA (2005) USDA 02L1058 and 02L1059: Cherry tomato breeding lines with high fruit β-carotene content. HortScience 40: 1569–1570
Stricker SH, Köferle A, Beck S (2017) From profiles to function in epigenomics. Nat Rev Genet 18: 51–66
Struik PC, Yin XY and de Visser P (2005) Complex quality traits: now time to model. Trends Plant Sci 10: 513-516
Suliman-Pollatschek S, Kashkush K, Shats H, Hillel J, Lavi U (2002) Generation and mapping of AFLP, SSRs and SNPs in *Lycopersicon esculentum*. Cell Mol Biol Lett 7: 583–97
Sun J, Poland JA, Mondal S, Crossa J, Juliana P, Singh RP, Rutkoski JE, Jannink J-L, Crespo-Herrera L, Velu G, et al. (2019) High-throughput phenotyping platforms enhance genomic selection for wheat grain yield across populations and cycles in early stage. Theor Appl Genet 1–16
Sun X, Gao Y, Li H, Yang S, Liu Y (2015) Over-expression of SlWRKY39 leads to enhanced resistance to multiple stress factors in tomato. J Plant Biol 58: 52–60
Suzuki N, Rivero RM, Shulaev V, Blumwald E, Mittler R (2014) Abiotic and biotic stress combinations. New Phytol 203: 32–43
Tadmor Y, Fridman E, Gur A, Larkov O, Lastochkin E, Ravid U, Zamir D, Lewinsohn E (2002) Identification of malodorous, a wild species allele affecting tomato aroma that was selected against during domestication. J Agric Food Chem 50: 2005–2009
Takken FLW, Thomas CM, Joosten M, Golstein C, Westerink N, Hille J, Nijkamp HJJ, De Wit P, Jones JDG (1999) A second gene at the tomato *Cf-4* locus confers resistance to *Cladosporium fulvum* through recognition of a novel avirulence determinant. Plant Journal 20: 279-288
Tam SM, Mhiri C, Vogelaar A, Kerkveld M, Pearce SR, Grandbastien MA (2005) Comparative analyses of genetic diversities within tomato and pepper collections detected by retrotransposon-based SSAP, AFLP and SSR. Theor Appl Genet 110: 819–831
Tanksley SD (2004) The Genetic , developmental, and molecular bases of fruit size in tomato and shape variation. Plant Cell 16: 181–190
Tanksley SD, Ganal MW, Prince JP, De Vicente MC, Bonierbale MW, Broun P, Fulton TM, Giovannoni JJ, Grandillo S, Martin GB, et al.

(1992) High density molecular linkage maps of the tomato and potato genomes.
Tanksley SD, Grandillo S, Fulton TM, Zamir D, Eshed Y, Petiard V, Lopez J, Beck-Bunn T (1996) Advanced backcross QTL analysis in a cross between an elite processing line of tomato and its wild relative L. pimpinellifolium. Theor Appl Genet 92: 213–224
Tanksley SD, Nelson JC. 1996. Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. Theor. Appl. Genet 92: 191-203
Tardieu F (2003) Virtual plants: modelling as a tool for the genomics of tolerance to water deficit. Trends Plant Sci 8: 9-14
Tashkandi M, Ali Z, Aljedaani F, Shami A, Mahfouz MM (2018) Engineering resistance against *Tomato yellow leaf curl virus* via the CRISPR/Cas9 system in tomato. Plant Signaling & Behavior 13
Taudt A, Colomé-Tatché M, Johannes F (2016) Genetic sources of population epigenomic variation. Nat Rev Genet 17: 319–332
Tavallali V, Esmaili S, Karimi S (2018) Nitrogen and potassium requirements of tomato plants for the optimization of fruit quality and antioxidative capacity during storage. J Food Meas Charact 12: 755–762
The 100 Tomato Genome Sequencing Consortium (2014) Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. Plant J 80: 136–148
The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073
The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65
The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526: 68–74
The 1001 Genomes Consortium (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell 166: 481–491
The 3000 rice genomes project (2014) The 3,000 rice genomes project. Gigascience 3: 7
The Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485: 635–641
The UK10K Consortium (2015) The UK10K project identifies rare variants in health and disease. Nature 526: 82–89
Thoen MPM, Davila Olivas NH, Kloth KJ, Coolen S, Huang PP, Aarts MGM, Bac-Molenaar JA, Bakker J, Bouwmeester HJ, Broekgaarden C, et al. (2017) Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping. New Phytol 213: 1346–1362
Tieman D, Bliss P, McIntyre LMM, Blandon-Ubeda A, Bies D, Odabasi AZZ, Rodríguez GRR, Van Der Knaap E, Taylor MGG, Goulet C, et al. (2012) The chemical interactions underlying tomato flavor preferences. Curr Biol 22: 1035–1039
Tieman D, Taylor M, Schauer N, Fernie AR, Hanson AD, Klee HJ (2006) Tomato aromatic amino acid decarboxylases participate in synthesis of the flavor volatiles 2-phenylethanol and 2-phenylacetaldehyde. Proc Natl Acad Sci U S A 103: 8287–92
Tieman D, Zhu G, Resende MFR, Lin T, Nguyen C, Bies D, Rambla JL, Beltran KSO, Taylor M, Zhang B, et al. (2017) A chemical genetic roadmap to improved tomato flavor. Science (80-) 355: 391–394
Tieman DM, Handa AK. Reduction in pectin methylesterase activity modifies tissue integrity and cation levels in ripening tomato (*Lycopersicon esculentum* Mill.) fruits. Plant Physiology (1994) 106(2):429-36.
Tikunov Y, Lommen A, Vos CHR, de Verhoeven HA, Bino RJ, Hall RD and Bovy AG (2005) A Novel Approach for Nontargeted Data Analysis for Metabolomics. Large-Scale Profiling of Tomato Fruit Volatiles. *Plant physiol.*, 139, 1125–1137.
Tikunov YM, J Molthoff, RCH de Vos, J Beekwilder, A van Houwelingen, et al. (2013) Non-smoky glycosyltransferase1 prevents the release of smoky aroma from tomato fruit. The Plant Cell 25 (8), 3067-3078
Toubiana D, Fernie AR, Nikoloski Z, Fait A (2013) Network analysis: Tackling complex data to study plant metabolism. Trends Biotechnol 31: 29–36
Tranchida-Lombardo V, Aiese Cigliano R, Anzar I, Landi S, Palombieri S, Colantuono C, Bostan H, Termolino P, Aversano R, Batelli G, et al. (2018) Whole-genome re-sequencing of two Italian tomato landraces reveals sequence variations in genes associated with stress tolerance, fruit quality and long shelf-life traits. DNA Res 25: 149–160
Turina M, Kormelink R, Resende RO (2016) Resistance to Tospoviruses in Vegetable Crops: Epidemiological and Molecular Aspects. In: Leach JE, Lindow S (eds) Annual Review of Phytopathology, Vol 54. pp 347-371
Uluisik S, Chapman NH, Smith R, Poole M, Adams G, Gillis RB, Besong TM, Sheldon J, Stiegelmeyer S, Perez L, Samsulrizal N (2016) Genetic improvement of tomato by targeted control of fruit softening. Nature Biotechnology 34(9):950.
Usadel B, Chetelat R, Koren S, Maumus F, Fernie AR, Aury J-M, Maß J, Schmidt MH-W, Denton AK, Wormit A, et al. (2017) De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing . Plant Cell 29: 2336–2348
Vakalounakis DJ, Laterrot H, Moretti A, Ligoxigakis EK, Smardas K (1997) Linkage between *Frl* (*Fusarium oxysporum* f sp *radicis-lycopersici* resistance) and *Tm-2* (*tobacco mosaic virus resistance-2*) loci in tomato (*Lycopersicon esculentum*). Annals of Applied Biology 130: 319-323
van Berloo R and Stam P (1998) Marker-assisted selection in autogamous RIL populations: a simulation study. Theor. Appl. Genet. 96: 147-154.
van Berloo R and Stam P (1999) Comparison between marker-assisted selection and phenotypical selection in a set of *Arabidopsis thaliana* recombinant inbred lines. Theor. Appl. Genet. 98: 113-118.
Van Berloo R, Zhu A, Ursem R, Verbakel H, Gort G, van Eeuwijk FA (2008) Diversity and linkage disequilibrium analysis within a selected set of cultivated tomatoes. Theor Appl Genet 117: 89–101
van der Knaap E, Tanksley SD (2003) The making of a bell pepper-shaped tomato fruit: identification of loci controlling fruit morphology in Yellow Stuffer tomato. Theor Appl Genet 107: 139–147
van Eeuwijk Fred A., Bustos-Korts D, Millet EJ, Boer MP, Kruijer W, Thompson A et al. (2019) Modelling strategies for assessing and increasing the effectiveness of new phenotyping techniques in plant breeding. Plant Science 282 :23-39
Van Ploeg D, Heuvelink E (2005) Influence of sub-optimal temperature on tomato growth and yield: a review. J Hortic Sci Biotechnol 80: 652–659
Vargas-Ponce O, Pérez-Álvarez LF, Zamora-Tavares P, Rodríguez A (2011) Assessing Genetic Diversity in Mexican Husk Tomato Species. Plant Mol Biol Report 29: 733–738
Veillet F, Perrot L, Chauvin L, Kermarrec M-P, Guyon-Debast A, Chauvin J-E, Nogué F, Mazier M (2019) Transgene-Free Genome Editing in Tomato and Potato Plants Using Agrobacterium-Mediated Delivery of a CRISPR/Cas9 Cytidine Base Editor. International Journal of Molecular Sciences 20 (2), 402
Verkerke W, Janse J, Kersten M (1998) Instrumental Measurement and Modelling of Tomato Fruit Taste. Acta Hortic 199–206
Verlaan MG, Hutton SF, Ibrahem RM, Kormelink R, Visser RGF, Scott JW, Edwards JD, Bai YL (2013) The *Tomato Yellow Leaf Curl Virus* Resistance Genes *Ty-1* and *Ty-3* Are Allelic and Code for DFDGD-Class RNA-Dependent RNA Polymerases. PLoS Genetics 9
Villalta I, Bernet GP, Carbonell EA, Asins MJ (2007) Comparative QTL analysis of salinity tolerance in terms of fruit yield using two solanum populations of F7 lines. Theor Appl Genet 114: 1001–1017
Víquez-Zamora M, Vosman B, van de Geest H, Bovy A, Visser RGF, Finkers R, van Heusden AW (2013) Tomato breeding in the genomics era: Insights from a SNP array. BMC Genomics 14: 354
Vos P, Simons G, Jesse T, Wijbrandi J, Heinen L, Hogers R, Frijters A, Groenendijk J, Diergaarde P, Reijans M, Fierens-Onstenk J, de Both M, Peleman J, Liharska T, Hontelez J, Zabeau M (1998) The tomato *Mi-1* gene confers resistance to both root-knot nematodes and potato aphids. Nature Biotechnology 16: 1365-1369
Vrebalov J, Ruezinsky D, Padmanabhan V, White R, Medrano D, Drake R, Schuch W, Giovannoni J (2002) A MADS-box gene necessary for fruit ripening at the tomato ripening-inhibitor (rin) locus. Science (80-) 296: 343–346
Wahid A, Gelani S, Ashraf M, Foolad MR (2007) Heat tolerance in plants: An overview. Environ Exp Bot 61: 199–223
Wang D, Salah El-Basyoni I, Stephen Baenziger P, Crossa J, Eskridge KM, Dweikat I (2012) Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. Heredity (Edinb) 109: 313–9
Wang DR, Agosto-Pérez FJ, Chebotarov D, Shi Y, Marchini J, Fitzgerald M, McNally KL, Alexandrov N, McCouch SR (2018) An imputation platform to enhance integration of rice genetic resources. Nat Commun 9: 3519
Wang JF, Ho FI, Truong HTH, Huang SM, Balatero CH, Dittapongpitch V, Hidayati N (2013) Identification of major QTLs associated with stable resistance of tomato cultivar 'Hawaii 7996' to *Ralstonia solanacearum*. Euphytica 190: 241-252
Wang K, Li M, Hakonarson H (2010) ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38: e164
Wang L, Song X, Gu L, Li X, Cao S, Chu C, Cui X, Chen X, Cao X (2013) NOT2 proteins promote polymerase II-dependent transcription and interact with multiple MicroRNA biogenesis factors in Arabidopsis. Plant Cell 25: 715–27
Wang R, Tavano ECDR, Lammers M, Martinelli AP, Angenent GC, de Maagd RA. (2019) Re-evaluation of transcription factor function in tomato fruit development and ripening with CRISPR/Cas9-mutagenesis. Sci Rep. 8:1696.
Wang Y, Wu W-H (2015) Genetic approaches for improvement of the crop potassium acquisition and utilization efficiency. Curr Opin Plant Biol 25: 46–52
Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: A revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63
Waters AJ, Makarevitch I, Noshay J, Burghardt LT, Hirsch CN, Hirsch CD, Springer NM (2017) Natural variation for gene expression responses to abiotic stress in maize. Plant J 89: 706–717
Wells T, Ward JL, Corol DI, Baker JM, Gerrish C, Michael H, Seymour GB, Fraser PD and Bramley PM (2013) Metabolite profiling of introgression lines of Solanum habrochaites using targeted and non-targeted approaches reveals novel quantitative trait loci.
Whitaker BD (2008) Postharvest flavor deployment and degradation in fruits and vegetables. *In* B Bruckner, S Grant Willie, eds, Fruit Veg.

flavour. CRC Press, Cambridge, UK, pp 103–131
Wilkinson JQ, Lanahan MB, Yen H-C, Giovannoni JJ, Klee HJ (1995) An ethylene-inducible component of signal transduction encoded by never-ripe. Science (80-) 270: 1809
Willits MG, CM Kramer, RT Prata, V De Luca , BG Potter, JC Steffens, G Graser (2005) Utilization of the Genetic Resources of Wild Species To Create a Nontransgenic High Flavonoid Tomato. J Agric Food Chem 53:1231-1236
Won SY, Yumul RE, Chen X (2014) Small RNAs in Plants. Mol. Biol. Springer New York, New York, NY, pp 95–127
Xiao H, Jiang N, Schaffner E, Stockinger EJ, Knaap E van der (2008) A Retrotransposon-Mediated Gene Duplication Underlies Morphological Variation of Tomato Fruit. Science 319: 1527–1530
Xie Z, Allen E, Fahlgren N, Calamar A, Givan SA, Carrington JC (2005) Expression of Arabidopsis MIRNA genes. Plant Physiol 138: 2145–54
Xinyou Y, P Stam, M J. Kropff, Ad HCM. Schapendonk, Yin X, Struik PC (2010) Modelling the crop: from system dynamics to systems biology. Journal of Experimental Botany 61: 2171 - 2183.
Xu J, Driedonks N, Rutten MJM, Vriezen WH, de Boer GJ, Rieu I (2017) Mapping quantitative trait loci for heat tolerance of reproductive traits in tomato (*Solanum lycopersicum*). *Mol Breed*. 37: 58.
Xu J, Ranc N, Muños S, Rolland S, Bouchet J-PP, Desplat N, Le Paslier M-CC, Liang Y, Brunel D, Causse M (2013) Phenotypic diversity and association mapping for fruit quality traits in cultivated tomato and related species. Theor Appl Genet 126: 567–581
Xu J, Wolters-Arts M, Mariani C, Huber H, Rieu I (2017) Heat stress affects vegetative and reproductive performance and trait correlations in tomato (*Solanum lycopersicum*). Euphytica 213: 156
Xu WF, Shi WM, Yan F (2012) Temporal and Tissue-Specific Expression of Tomato 14-3-3 Gene Family in Response to Phosphorus Deficiency. Pedosphere 22: 735–745
Yamamoto E, Matsunaga H, Onogi A, Kajiya-Kanegae H, Minamikawa M, Suzuki A, Shirasawa K, Hirakawa H, Nunome T, Yamaguchi H, et al. (2016) A simulation-based breeding design that uses whole-genome prediction in tomato. Sci Rep 6: 19454
Yamamoto E, Matsunaga H, Onogi A, Ohyama A, Miyatake K, Yamaguchi H, Nunome T, Iwata H, Fukuoka H (2017) Efficiency of genomic selection for breeding population design and phenotype prediction in tomato. Heredity (Edinb) 118: 202–209
Yang D-Y, Li M, Ma N-N, Yang X-H, Meng Q-W (2017) Tomato SlGGP-LIKE gene participates in plant responses to chilling stress and pathogenic infection. Plant Physiol Biochem 112: 218–226
Yang X, Caro M, Hutton SF, Scott JW, Guo Y, Wang X, Rashid MH, Szinay D, de Jong H, Visser RGF, et al. (2014) Fine mapping of the tomato yellow leaf curl virus resistance gene Ty-2 on chromosome 11 of tomato. Mol Breed 34: 749–760
Yasmeen A, Mirza B, Inayatullah S, Safdar N, Jamil M, Ali S, Choudhry MF (2009) In planta transformation of tomato. Plant molecular biology reporter 27(1):20-8.
Ye J, Wang X, Hu T, Zhang F, Wang B, Li C, Yang T, Li H, Lu Y, Giovannoni JJ, et al. (2017) An InDel in the Promoter of Al-ACTIVATED MALATE TRANSPORTER9 Selected during Tomato Domestication Determines Fruit Malate Contents and Aluminum Tolerance. Plant Cell 29: 2249–2268
Yin X, MJ Kropff, P Stam. (1999) The role of ecophysiological models in QTL analysis: The example of specific leaf area in barley. Heredity 82:415–421
You C, Cui J, Wang H, Qi X, Kuo L-Y, Ma H, Gao L, Mo B, Chen X (2017) Conservation and divergence of small RNA pathways and microRNAs in land plants. Genome Biol 18: 158
Young ND, Tanksley SD (1989) RFLP analysis of the size of chromosomal segments retained around the Tm-2 locus of tomato during backcross breeding. Theor Appl Genet 77: 353-359
Yu B, Bi L, Zheng B, Ji L, Chevalier D, Agarwal M, Ramachandran V, Li W, Lagrange T, Walker JC, et al. (2008) The FHA domain proteins DAWDLE in Arabidopsis and SNIP1 in humans act in small RNA biogenesis. Proc Natl Acad Sci 105: 10073–10078
Yu Y, Jia T, Chen X (2017) The 'how' and 'where' of plant microRNAs. New Phytol 216: 1002–1017
Zamir D (2001) Improving plant breeding with exotic genetic libraries. Nat Rev Genet 2: 3–9
Zanor MI, Rambla JL, Chaib J, Steppa A, Medina A, Granell A, Fernie AR, Causse M (2009) Metabolic characterization of loci affecting sensory attributes in tomato allows an assessment of the influence of the levels of primary metabolites and volatile organic contents. J Exp Bot 60: 2139–2154
Zegbe-Dominguez J., Behboudian M., Lang A, Clothier B. (2003) Deficit irrigation and partial rootzone drying maintain fruit dry mass and enhance fruit quality in 'Petopride' processing tomato (*Lycopersicon esculentum*, Mill.). Sci Hortic (Amsterdam) 98: 505–510
Zhang B, Tieman DM, Chen J, Xu Y, Chen K, Fei Z, Giovannoni J, Klee HJ (2016) Loss of tomato flavor quality during chilling is associated with reduced expression of volatile biosynthetic genes and a transient alteration in DNA methylation. *Proc. Natl. Acad. Sci USA* 113: 12580-84.
Zhang C, Liu L, Wang X, Vossen J, Li G, Li T, Zheng Z, Gao J, Guo Y, Visser RGF, et al. (2014) The Ph-3 gene from Solanum pimpinellifolium encodes CC-NBS-LRR protein conferring resistance to Phytophthora infestans. Theor Appl Genet 127: 1353–1364
Zhang J, Zhao J, Liang Y, Zou Z (2016) Genome-wide association-mapping for fruit quality traits in tomato. Euphytica 207: 439–451
Zhang J, Zhao J, Xu Y, Liang J, Chang P, Yan F, Li M, Liang Y, Zou Z (2015) Genome-Wide Association Mapping for Tomato Volatiles Positively Contributing to Tomato Flavor. Front Plant Sci 6: 1042
Zhang S, Xie M, Ren G, Yu B (2013) CDC5, a DNA binding protein, positively regulates posttranscriptional processing and/or transcription of primary microRNA transcripts. Proc Natl Acad Sci U S A 110: 17588–93
Zhang Y, Butelli E, Alseekh S, Tohge T, Rallapalli G, Luo J, Kawar PG, Hill L, Santino A, Fernie AR, Martin C, (2015) Multi-level engineering facilitates the production of phenylpropanoid compounds in tomato. Nature Communications 6:8635.
Zhang Z, Guo X, Ge C, Ma Z, Jiang M, Li T, Koiwa H, Yang SW, Zhang X (2017) KETCH1 imports HYL1 to nucleus for miRNA biogenesis in Arabidopsis. Proc Natl Acad Sci U S A 114: 4011–4016
Zhao C, Liu B, Piao S, Wang X, Lobell DB, Huang Y, Huang M, Yao Y, Bassu S, Ciais P, et al. (2017) Temperature increase reduces global yields of major crops in four independent estimates. Proc Natl Acad Sci 114: 9326–9331
Zhao J, Sauvage C, Zhao J, Bitton F, Bauchet G, Liu D, Huang S, Tieman DM, Klee HJ, Causse M (2019) Meta-analysis of genome-wide association studies provides insights into genetic control of tomato flavor. Nat Commun 10: 1534
Zhao J, Xu Y, Ding Q, Huang X, Zhang Y, Zou Z, Li M, Cui L, Zhang J (2016) Association Mapping of Main Tomato Fruit Sugars and Organic Acids. Front Plant Sci 7: 1–11
Zhao X, Liu Y, Liu X, Jiang J (2018) Comparative Transcriptome Profiling of Two Tomato Genotypes in Response to Potassium-Deficiency Stress. Int J Mol Sci 19: 2402
Zhong S, Fei Z, Chen Y, Zheng Y, Huang M, Vrebalov J, McQuinn R, Gapper N, Liu B, Xiang J, et al. (2013) Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. Nat Biotechnol 31: 154–159
Zhou R, Wu Z, Cao X, Jiang F (2015) Genetic diversity of cultivated and wild tomatoes revealed by morphological traits and SSR markers. Genet Mol Res 14: 13868–13879
Zhou R, Yu X, Ottosen C-O, Rosenqvist E, Zhao L, Wang Y, Yu W, Zhao T, Wu Z (2017) Drought stress had a predominant effect over heat stress on three tomato cultivars subjected to combined stress. BMC Plant Biol 17: 24
Zhu G, J Gou, H Klee, S Huang (2019) Next-Gen Approaches to Flavor-Related Metabolism. Annual Review of Plant Biology 70: 187-212
Zhu G, Wang S, Huang Z, Zhang S, Liao Q, et al. (2018) Rewiring of the Fruit Metabolome in Tomato Breeding. Cell 172, 249–261
Zhuang K, Kong F, Zhang S, Meng C, Yang M, Liu Z, Wang Y, Ma N, Meng Q (2019) Whirly1 enhances tolerance to chilling stress in tomato via protection of photosystem II and regulation of starch degradation. New Phytol 221: 1998–2012
Zsögön A, Čermák T, Naves ER, Notini MM, Edel KH, Weinl S, Freschi L, Voytas DF, Kudla J, Peres LE (2018) De novo domestication of wild tomato using genome editing. Nature Biotechnology 36: 1211-1216.
Zsögön A, Cermak T, Voytas D, Pereira Peres LE (2017) Genome editing as a tool to achieve the crop ideotype and de novo domestication of wild relatives: Case study in tomato. Plant Science 256, 120-130
Zuo J, Fu D, Zhu Y, Qu G, Tian H, Zhai B, Ju Z, Gao C, Wang Y, Luo Y, et al. (2013) SRNAome parsing yields insights into tomato fruit ripening control. Physiol Plant 149: 540–553
Zuo J, Zhu B, Fu D, Zhu Y, Ma Y, Chi L, Ju Z, Wang Y, Zhai B, Luo Y (2012) Sculpting the maturation, softening and ethylene pathway: the influences of microRNAs on tomato fruits. BMC Genomics 13: 7
Zuriaga E, Blanca J, Nuez F (2009) Classification and phylogenetic relationships in *Solanum* section *Lycopersicon* based on AFLP and two nuclear gene sequences. Genet Resour Crop Evol 56: 663–678