Analyse du polymorphisme moléculaire de gènes de composantes de la qualité des fruits dans les ressources génétiques sauvages et cultivées de tomate ; recherche d'associations gènes/QTL

Nicolas N. Ranc

ECOLE NATIONALE SUPERIEURE AGRONOMIQUE DE MONTPELLIER – SUPAGRO

Ecole Doctorale SIBAGHE - Systèmes Intégrés en Biologie, Agronomie, Géosciences, Hydrosciences, Environnement
Formation doctorale EERPG - Evolution, Ecologie, Ressources Génétiques, Paléontologie

THESE

présentée publiquement le 28 janvier 2010

pour obtenir le grade de Docteur en Sciences
de l'Ecole Nationale Supérieure Agronomique de Montpellier

Analyse du polymorphisme moléculaire de gènes de composantes de la qualité des fruits dans les ressources génétiques sauvages et cultivées de tomate ; recherche d'associations gènes/QTL

Nicolas RANC

JURY

Anne-Françoise ADAM-BLONDON	INRA Versailles-Grignon	Rapporteur
Alain CHARCOSSET	INRA Versailles-Grignon	Rapporteur
Dominique BRUNEL	INRA Versailles-Grignon	Examinateur
Rafael FERNÁNDEZ MUÑOZ	CSIC	Examinateur
Jacques DAVID	SupAgro	Examinateur
Mathilde CAUSSE	INRA Avignon	Directeur de thèse

INRA-UR1052 Unité de Génétique et Amélioration des Fruits et Légumes
Centre INRA Avignon – Montfavet - France

ECOLE NATIONALE SUPERIEURE AGRONOMIQUE DE MONTPELLIER – SUPAGRO

Ecole Doctorale SIBAGHE - Systèmes Intégrés en Biologie, Agronomie, Géosciences, Hydrosciences, Environnement
Formation doctorale EERPG - Evolution, Ecologie, Ressources Génétiques, Paléontologie

THESE

présentée publiquement le 28 janvier 2010

pour obtenir le grade de Docteur en Sciences
de l'Ecole Nationale Supérieure Agronomique de Montpellier

Analyse du polymorphisme moléculaire de gènes de composantes de la qualité des fruits dans les ressources génétiques sauvages et cultivées de tomate ; recherche d'associations gènes/QTL

Nicolas RANC

JURY

Anne-Françoise ADAM-BLONDON	INRA Versailles-Grignon	Rapporteur
Alain CHARCOSSET	INRA Versailles-Grignon	Rapporteur
Dominique BRUNEL	INRA Versailles-Grignon	Examinateur
Rafael FERNÁNDEZ MUÑOZ	CSIC	Examinateur
Jacques DAVID	SupAgro	Examinateur
Mathilde CAUSSE	INRA Avignon	Directeur de thèse

INRA-UR1052 Unité de Génétique et Amélioration des Fruits et Légumes
Centre INRA Avignon – Montfavet - France

Résumé
Chez la tomate, l'amélioration pour la qualité du fruit est rendue difficile par la multiplicité et la complexité des caractères. La cartographie de QTL a permis la caractérisation génétique de ces caractères. L'objectif est maintenant d'identifier les gènes sous-jacents aux QTL. Nous avons utilisé la cartographie par déséquilibre de liaison (DL) dans ce but. Pour éviter les fausses associations entre les caractères et les polymorphismes moléculaires, la structure génétique a été prise en compte dans l'analyse. La tomate cultivée montre un faible niveau de diversité génétique, ce qui réduit la résolution de cartographie. Le génome de la tomate de type cerise (*S. lycopersicum* var. *cerasiforme*) est décrit comme une mosaïque entre celui de la tomate cultivée et de l'ancêtre sauvage. Ce mélange devrait augmenter la résolution des études d'association. Nous avons utilisé une « core collection » focalisée sur des accessions de type cerise pour valider la région génomique contenant un QTL pour le nombre de loges. Deux mutations sont associées avec le caractère. Ces deux SNP ont évolué différemment du reste du chromosome 2, en subissant une sélection balancée qui témoigne de l'augmentation de la diversité morphologique lors de la domestication. L'étude d'association, focalisée sur le chromosome 2, a permis d'analyser l'étendue du DL en fonction de la distance génétique et physique. Des associations entre des polymorphismes et les phénotypes étudiés ont été détectés avec des méthodes prenant en compte la structure génétique. Nous avons montré l'intérêt d'utiliser la structure en mosaïque du génome des accessions de type cerise pour surmonter les limitations de résolution dans les analyses d'associations chez une espèce cultivée autogame. Nous avons validé des QTL identifiés précédemment et nous avons trouvé des associations avec de nouveaux QTL et de nouveaux gènes candidats. Un modèle d'évolution incluant un goulet d'étranglement et des flux de gènes entre compartiment sauvage et cultivé de tomate est aussi présenté.

Mots clés : tomate, qualité du fruit, ressources génétiques, déséquilibre de liaison, génétique d'association, diversité moléculaire

Abstract
In Tomato (*Solanum lycopersicum*), breeding for fruit quality is difficult due to the multiplicity and complexity of the traits. QTL mapping has allowed the genetic characterization of these traits. One of the challenges is now to identify the genes underlying these QTLs. Following this aim, we used linkage-disequilibrium (LD) mapping. To avoid hazardous associations between traits and polymorphisms, the genetic structure has to be taken into account for LD mapping. Cultivated tomato showed low genetic diversity reducing mapping resolution. Cherry type tomato (*S. lycopersicum* var. *cerasiforme*) genome is described to be admixture between cultivated tomato and its wild ancestor. Such admixture may increase resolution of association mapping. We used a core collection focused on cherry type accessions to validate a candidate gene for a fruit locule-number QTL. We found that two single nucleotide polymorphisms (SNP) were highly associated with the trait. These two SNP evolved differently from the rest of the chromosome 2. They underwent a balanced selection which testifies a selection for fruit morphology diversity by human. Association mapping, focused on whole chromosome 2, allowed us to assess the extent of linkage disequilibrium over genetic and physical distances. Associations of polymorphisms with phenotypes were detected with structured association methods. We thus showed efficiency of genome admixture to overcome the low-resolution limitation of association mapping for an inbred crop. We validated previously identified QTLs and found associations with new QTLs and new candidate genes. An evolutionary model including bottleneck and gene flow between wild and domesticated forms of tomato is also presented.

Key words: tomato, fruit-quality traits, genetic resources, linkage disequilibrium, association mapping, molecular diversity

Remerciements

En écrivant ces lignes, je conclue trois années de pur plaisir au sein de l'Unité GAFL. Je me souviens encore de l'accueil chaleureux que je recevais lorsque je suis arrivé, ou plutôt lorsque je suis revenu, à Avignon. Ce travail de thèse n'aurait pas été possible sans le soutien de nombreuses personnes qui doivent être remerciées comme il se doit.

Je commencerai donc par remercier les différents membres du jury de ma thèse qui ont accepté de juger ce travail. Je tiens notamment à remercier Jacques David qui m'a insufflé sa passion de la recherche et de l'histoire des plantes. Lorsque je suis arrivé en Master 2, je dois avouer que j'ai été noyé quelques temps dans un océan de concepts qui ne m'étaient pas familiers. Mais les personnes responsables de la formation, dont Jacques, ont su me faire confiance et, aujourd'hui, je leur suis reconnaissant pour l'aboutissement que représente cette thèse.

Je tiens à remercier Mathilde, qui malgré son emploi du temps plus que chargé, a toujours su me consacrer du temps lorsque j'en avais besoin. Elle a su me faire confiance en me proposant un sujet ambitieux et j'espère que j'ai répondu à toutes ses attentes quant à l'apport que j'ai pu offrir à l'équipe dans le domaine de l'étude de la diversité génétique. Je tiens aussi à la remercier pour le temps qu'elle a passé à corriger les 's' qui se baladaient ça et là dans le manuscrit mais qui, malheureusement, n'étaient jamais à leur place. Jusqu'au dernier moment elle a su être vaillante dans les échanges des parties de la thèse à corriger et à relire.

Une autre personne, qui a participé de près à cette thèse, mérite tout le témoignage de ma reconnaissance. Stéphane a su, pendant ces trois années, me montrer qu'il n'est pas de science sans amusement et sans plaisir. Je crois que j'ai compris la leçon et j'espère qu'un jour l'élève dépassera le maître. Les discussions qu'on a pu avoir dans ton bureau, aussi fine qu'un engin de démolition, ont toujours su me faire rire et rendre le travail agréable. J'espère que ta passion pour la trompette ne t'écartera jamais de la recherche (même si ta version de Fort Boyard est d'anthologie).

Je tiens aussi à remercier Sophie qui a toujours été là lorsque le moral n'était pas au plus haut. Plus qu'une collègue de travail, elle est devenue une amie, avec qui, les pires manips (broyage au mortier de 360 plantes) ont toujours été agréables. Avec elle, c'était un petit bout de Bretagne qui a accompagné cette thèse. J'ai essayé d'en faire une vraie provençale mais je pense qu'il y a encore beaucoup de travail pour ça. Alors, Sophie, si jamais la définition des mots 'pègue' et 'escagasser' t'échappe, n'hésite pas à me demander.

Une grande partie du travail dans la génétique d'association est consacrée au phénotypage des plantes. Lorsqu'il a fallu s'occuper d'un tunnel complet d'accessions de tomate, on m'a confié un co-équipier de choc. Esther m'a accompagné dans cette épopée formidable que j'espère avoir rendu aussi agréable que possible pour elle. Les moments passés à peser, mesurer, couper et broyer des fruits, au son de Claude François sur Nostalgie, ont été très productifs car j'ai enfin appris les paroles de 'Magnolia for ever' qui manquaient tant à mon répertoire musical.

Je voudrais aussi remercier Hélène pour sa connaissance des ressources génétiques de tomate et je vous assure que retrouver, de mémoire, que LA1950 est en fait un *S. pimpinellifolium* orange, en vous donnant la date de la dernière multiplication, n'est pas chose facile. Je remercie aussi Yolande pour son engouement et sa précision dans les travaux de plantations et

d'entretien des plantes. L'équipe expérimentale réalise aussi un travail quotidien qu'il faut saluer. Rien de ce qui est présenté dans ce manuscrit ne serait possible sans eux.

La bioinformatique est encore quelque chose qui reste obscur pour moi et je dois avouer que Jean-Paul a bien essayé de me convertir à Unix. Je me suis laissé tenté, puis je suis revenu en courant vers ce bon vieux Windows qui crache plus souvent, certes, mais qui est tellement plus convivial. J'espère qu'un jour on pourra utiliser ce sacré GnpSNP comme on lance un Blast en ligne de commande (c'est dire avec quelle facilité).

Le trio infernal composé de Marianne, Véranne et Manu a permis de soigner les baisses de régime avec le meilleur remède : de grands éclats de rire. Manu a su compléter mes lacunes dans tous les domaines qui ne nécessitent que peu de neurones (domaines qui me tiennent à cœur). Le 'sirop typhon' aura au moins permis de réconcilier non pas deux mais trois générations. Et à la question que penses-tu des Tranformers, il te faudra maintenant répondre : Optimus Prime est le plus grand et il vient d'une autres planète pour nous sauver des Décepticons.

Je tiens maintenant à remercier les personnes qui ont partagé le même bureau que moi. Même si on a souvent manifesté notre mécontentement sur le fait d'être trois dans seulement 12 m², je tenais à remercier Besma et Noé pour les échanges qu'on a pu avoir, au niveau scientifique comme personnel.

Je souhaite un bon courage aux thésards qui suivent : Noé (et oui encore une fois, quel privilégié), Laurence, Sophie et la petite dernière Mélissa. Je tiens aussi à encourager Benoît qui est sur la dernière ligne droite.

Je voudrais aussi remercier les anciens thésards qui ont quitté l'unité. Je veux bien entendu parler de Carine, Julien(s) et Marion. J'ai réalisé une bonne partie de la thèse dans le même bureau que Marion, qui me précédait d'un an. Je tenais à la remercier pour les discussions sur nos avenirs professionnels qui finalement n'ont pas tardé à montrer le bout de leurs nez.

Je voudrais aussi remercier Mireille, Bénédicte, Véronique, Karine, Jean-Marc et toutes les autres personnes du CPER avec qui les discussions ont toujours été fructueuses et les moments de pause toujours agréables (même si les discussions étaient un peu trop féministes à mon goût…).

Le GAFL ne serait rien sans les personnes qui travaillent dans l'ombre de la publication et je pense qu'une thèse et le meilleur moment pour remercier ces gens pour leur engagement. Je voudrais donc remercier l'équipe de l'installation expérimentale (Alain, Christophe, Mara, Robert, etc.) ainsi que la Dream Team qui règne dans le bâtiment administratif : Claudie, Annick, Astrid et Angélique.

Je voudrais aussi remercier l'équipe de course à pied INRAtatouille (Didier, Alexandre, Christophe, Benoît, Patou) qui m'a fait prendre conscience qu'il fallait que j'arrête de fumer et que je me mette au sport. Je tiens notamment à remercier le coach pour tous ses bons conseils et, qu'il ne s'inquiète pas : ce ne sont pas toutes ces blessures qui vont m'arrêter, elles vont juste me ralentir un peu.

Je voudrais aussi souligner, l'accueil chaleureux que j'ai reçu quand je suis arrivé à Montpellier, au début de ma thèse, pour réaliser le génotypage SSR. Je tiens à saluer

l'ambiance dans laquelle, Sylvain Santoni, Isabelle Hochu, Audrey Weber ainsi que toute les autres personnes présentes à l'époque dans le laboratoire Biomol de l'unité DIAPC, savent réaliser un travail performant. Sylvain, je n'oublierai pas les discussions qu'on a pu avoir sur l'histoire et l'évolution des technologies en biologie et ta passion pour cette science que je qualifierais de visionnaire. Je n'oublierai pas non plus que « choisir c'est renoncer ».

Il m'a été donné, durant cette thèse, de naviguer régulièrement vers une contrée que je ne connaissais point : Evry Courcouronnes. C'est dans cette région isolée que j'ai pu rencontrer les personnes de l'équipe EPGV sans qui cette thèse n'aurait pas été la moitié de ce qu'elle est. Je tiens à remercier Marie-Christine Le Paslier, Aurélie Bérard, Aurélie Chauveau et Rémi Bounon qui forment cette équipe, orchestrée majestueusement par Dominique Brunel. Je n'oublierai jamais nos repas népalais avec vue sur le mont Everest.

Je voudrais aussi remercier deux personnes qui ont suivis régulièrement ce travail. Il s'agit de Domenica Manicacci et Joëlle Ronfort qui ont fait parti de mes comités de thèse successifs.

Je voudrais aussi saluer la performance de Laure David que j'ai encadrée pour son stage de DUT. J'ai apprécié qu'elle ne se soit pas enfuie lorsque j'ai commencé à lui parler de déséquilibre de liaison et d'associations statistique entre phénotype et génotype. Le travail qu'elle a rendu était exemplaire et j'ai réellement apprécié de travailler avec elle.

Je tiens maintenant à remercier des personnes qui n'ont pas contribué directement à ce travail mais qui sont là depuis de nombreuses années. Je voudrais remercier Salva, Ben Bedo, Claude Alain dit (Pristiluc ou Prosac, j'ai jamais réellement su), Souris, Angélique, Geiss, JD, le Cèpe, Jean-Charles, Yoch, Christelle, Ben Grand, Momo, David et j'en passe. Je pense qu'on se construit avec les gens qu'on côtoie souvent.

Je voudrais maintenant remercier les personnes sans qui, cette thèse, ni rien d'autres, n'aurait été possible. Je veux bien entendu parler de mes parents et de ma sœur. Mes parents se sont investis corps et âme pour me donner une éducation digne de ce nom et j'aimerai aujourd'hui les remercier pour tout ce qu'ils ont toujours fait pour moi. Ma sœur, quant à elle, a toujours su montrer qu'elle était présente (des fois un peu trop) mais je pense que c'est le bon moment pour lui affirmer tout mon amour.

Je remercie aussi ma Tatie Francette pour sa relecture acharnée du manuscrit mais aussi pour tout ce qu'elle a pu m'apporter depuis ma plus tendre enfance. Tu représentes pour moi plus qu'une tante. J'en profite pour te souhaiter une bonne retraite.

Je voudrais aussi remercier Conchi, Anaïs, Pascal, Véro, Marine et Océane, Daniel, Christiane, Marie-Jo et tous les autres Arlésiens que j'ai pu connaître ces dernières années, pour leur sympathie et leur joie de vivre.

Je tiens, en dernier lieu, à remercier Céline qui me supporte depuis plus de six ans maintenant. Je pense que ces derniers jours ont été encore plus difficiles pour elle. Je la remercie donc pour tout ce qu'elle m'apporte au quotidien et que dire de plus que : « Je t'aime ».

Je dédie cette thèse à mes quatre grands parents : Lucien et Marie-Louise Ranc et André et Andrée Massebœuf, qui ont consacré leur vie à la culture de la terre pour m'offrir ce que j'ai aujourd'hui.

Sommaire

Introduction Générale

La saveur et la texture des tomates (*Solanum lycopersicum*) de frais sont particulièrement critiquées par les consommateurs. La qualité gustative est un problème d'actualité chez de nombreux fruits, notamment en vue de répondre aux préconisations du Programme National Nutrition Santé qui propose de consommer au moins cinq fruits ou légumes par jour. La qualité gustative a été mise de côté par l'amélioration variétale qui s'est focalisée sur le rendement, l'adaptation à différents environnements de culture et les résistances aux stress biotiques et abiotiques. La tomate répond relativement bien aux attentes nutritionnelles de l'organisme car le fruit est relativement pauvre en calories mais riche en eau et en éléments minéraux. Elle contient aussi une grande quantité d'éléments anti-oxydants, comme le lycopène et la vitamine C qui jouent un rôle important dans la qualité nutritionnelle.

Avec l'importance que prend la tomate dans la consommation de fruits et légumes quotidienne (22 Kg/an/personne en France), il est nécessaire que le fruit consommé présente des critères appréciés par le consommateur. Il est donc important de prendre en compte ces caractères dans les schémas d'amélioration variétale. Pour cela, il est nécessaire de connaître les bases moléculaires qui gouvernent leurs variations. L'identification des QTL des composantes de la qualité est donc un enjeu majeur qui a conduit à leur cartographie et à la recherche de gènes candidats liés. Le développement de nombreuses ressources moléculaires et génétiques, associé à l'émergence de la première version de la séquence du génome, ont, par ailleurs, fait de la tomate un modèle pour étudier le développement et la maturation des fruits charnus.

La cartographie génétique utilise une population issue d'un croisement entre deux parents. Ces deux individus ne représentent qu'une part infime de la variation qu'on peut trouver dans une collection de ressources génétiques. La génétique d'association est une méthode émergente chez les plantes pour localiser des polymorphismes impliqués dans la variation de caractères d'intérêt. Cette méthode montre des résultats encourageants chez des espèces allogames ou chez des espèces sauvages. Très peu d'études convaincantes ont été réalisées chez des espèces cultivées autogames. Ces études montrent l'intérêt de prendre en

compte la structuration génétique de l'échantillon dans les modèles utilisés pour détecter les associations. Une collection de ressources génétiques de tomates sauvages et cultivées, est maintenue et caractérisée à l'unité GAFL du Centre de Recherche INRA d'Avignon. Cette collection représente un outil d'une grande richesse, en vue de tenter d'associer certaines variations moléculaires à des composantes de la qualité du fruit.

Des études de diversité génétique chez la tomate cultivée montrent que la diversité est très faible chez celle-ci. Ce manque de diversité diminue énormément la puissance de détection des polymorphismes moléculaires. Un niveau de polymorphisme intermédiaire entre l'espèce sauvage la plus proche et l'espèce cultivée a été identifié chez les accessions de type cerise (*S. lycopersium* var *cerasiforme*). L'objectif de cette thèse a été de valider l'utilisation des ressources génétiques chez la tomate, notamment des accessions de type cerise, dans le but de disséquer les bases génétiques des composantes de la qualité du fruit.

Les objectifs scientifiques de ce travail étaient les suivants :

- Identifier la structure génétique d'un échantillon de 360 accessions de tomates sauvages et cultivées et construire plusieurs « core collections » représentatives de la diversité morphologique et génétique présentes dans l'échantillon initial.
- Identifier, en utilisant les ressources génétiques, le polymorphisme causal d'un QTL d'architecture du fruit en cours de clonage positionnel.
- Estimer les conditions d'utilisation de la génétique d'association chez la tomate en vue de son amélioration.
- Identifier de nouvelles régions génomiques impliquées dans la variation de composante de la qualité.

Nous nous sommes concentrés sur l'ensemble des accessions de types cerise maintenues à l'INRA d'Avignon. Nous avons vérifié que ces accessions représentent un matériel de choix pour réaliser des études d'association chez la tomate. Le taux de polymorphisme présent dans cet échantillon ainsi que la manière dont il se structure ont été étudiés. Nous avons ensuite étudié l'étendue du déséquilibre de liaison et nous avons identifié le modèle le plus pertinent, utilisé pour détecter des associations entre phénotypes et génotypes.

Cette thèse a été soutenue financièrement par le projet européen EUSOL (FOOD-CT-2006-016214). Elle a fait l'objet de plusieurs collaborations avec l'Unité Diversité et Adaptation des Plantes Cultivées (INRA, Montpellier), l'Unité Etude du Polymorphisme des Génomes Végétaux (INRA, Evry) et le laboratoire d'Alisdair Fernie au Max Planck Institute de Golm (Allemagne).

Chapitre 1 : Synthèse bibliographique

1.1. Caractérisation de locus liés à un phénotype d'intérêt.

1.1.1. Une histoire pas si récente.

C'est après dix années de travaux minutieux, que le 8 février 1865, Johann Gregor Mendel (1822-1884) partage ses travaux intitulés *Versuche uber Pflanzen-Hybriden* ou *Recherche sur des hybrides végétaux* qu'il publiera un an plus tard (1866). Mendel fut un des premiers à utiliser un plan d'expérience afin d'expliquer les lois de l'origine et de la formation des hybrides. Même si ses travaux, à l'époque, n'ont pas connu le succès escompté, et bien qu'ils continuent à être critiqués, Gregor Mendel a eu le génie d'établir les bases statistiques de la génétique et de l'hérédité moderne.

L'évocation de la théorie de la sélection naturelle par Charles Robert Darwin (1809-1882) fut une seconde révolution, contemporaine aux travaux de Mendel. Les travaux de Darwin ont été pris très au sérieux dès la publication de *On the Origin of Species by Means of Natural Selection* en 1859 (Darwin 1859). Darwin travaille notamment sur la domestication des animaux mais aussi des plantes pour illustrer ses arguments portant sur la variation d'un caractère, dans un groupe d'espèces soumis à une pression de sélection (Darwin 1868). Il remarque notamment le passage aux générations suivantes de caractères avantageux apparus dans certains groupes d'espèces. Darwin, qui n'est pas au courant des travaux de Mendel, est partisan de la théorie de l'hérédité par mélange. La transmission de caractère d'une génération à une autre était donc acceptée mais le support physique de cette information n'était pas connu. Darwin parlait alors de gemmules ou de pangènes (Darwin 1868).

C'est au début du XXème siècle que Griffith identifie un « facteur » capable de transformer une souche de pneumocoque en une autre, leur conférant de façon héréditaire de nouvelles propriétés génétiques (Griffith 1928). Ce facteur transformant ne sera isolé que 16 ans plus tard. Avery, McLeod et al. (1944) identifient l'ADN : une molécule d'acide désoxyribonucléique. Certains scientifiques continuent alors à penser que l'ADN est un enchaînement monotone de quatre bases (Adénine, Thymine, Cytosine et Guanine) et que la diversité des protéines déjà identifiées suffirait à expliquer l'information génétique.

Des travaux plus poussés sur l'étude de l'infection de bactéries par des phages ont montré que ADN des phages est responsable de leur réplication au sein des bactéries infectées et donc que cette molécule porte l'information génétique (Hershey and Chase 1952). L'élaboration du modèle en double hélice de la molécule d'ADN est sans aucun doute la révolution qui a permis le développement de la biologie moléculaire (Watson and Crick 1953).

Parallèlement à ces études mécanistiques sur le fonctionnement de l'hérédité, d'autres scientifiques s'intéressent à la redécouverte des travaux de Mendel. Les écarts à la ségrégation indépendante des caractères trouvent une explication dans la liaison génétique, établie par Bateson et Punnett en 1910. Cette théorie a ensuite été étendue à plusieurs caractères avec le développement des cartes génétiques (Morgan, Sturtevant et al. 1915). Ces cartes ont d'abord été développées avec des caractères morphologiques (Figure 1-1) puis ont été complétées avec les premiers marqueurs moléculaires comme les isozymes.

En même temps, l'essor des statistiques permet aux généticiens de s'intéresser aux caractères quantitatifs. Ce sont des statisticiens et évolutionnistes, tels que R. A. Fisher, S. Wright et J.B.S. Haldane, qui ont permis le développement des concepts biométriques nécessaires à l'étude de ces caractères. Les deux écoles « biométrique » et « mendelienne » entrent en conflit à propos de la théorie de l'hérédité puis se réconcilient quand, en 1910, il est démontré comment une variation phénotypique continue peut résulter de l'effet combiné de l'environnement et de la ségrégation de plusieurs locus mendéliens (East 1910). La génétique quantitative moderne naît de la fusion entre Mendélisme et Biométrie (Mauricio 2001).

L'amélioration génétique des plantes mais aussi la sélection animale se faisaient par sélection généalogique combinée à la sélection massale. Etant donné qu'il n'était pas possible de prédire la ségrégation des caractères, un nombre important d'individus étaient testé et on ne retenait que les plus intéressants qui étaient recroisés entre eux. L'arrivée des marqueurs génétiques, avec d'abord les isozymes puis les marqueurs moléculaires, a transformé la sélection moderne. La recherche de marqueurs moléculaires aide le sélectionneur à mieux connaître la génétique des caractères importants afin d'optimiser l'efficacité des programmes de sélection. Ces marqueurs sont en effet très précieux car ils permettent de tester rapidement les variétés et de ne retenir que celles qui possèdent les caractéristiques recherchées.

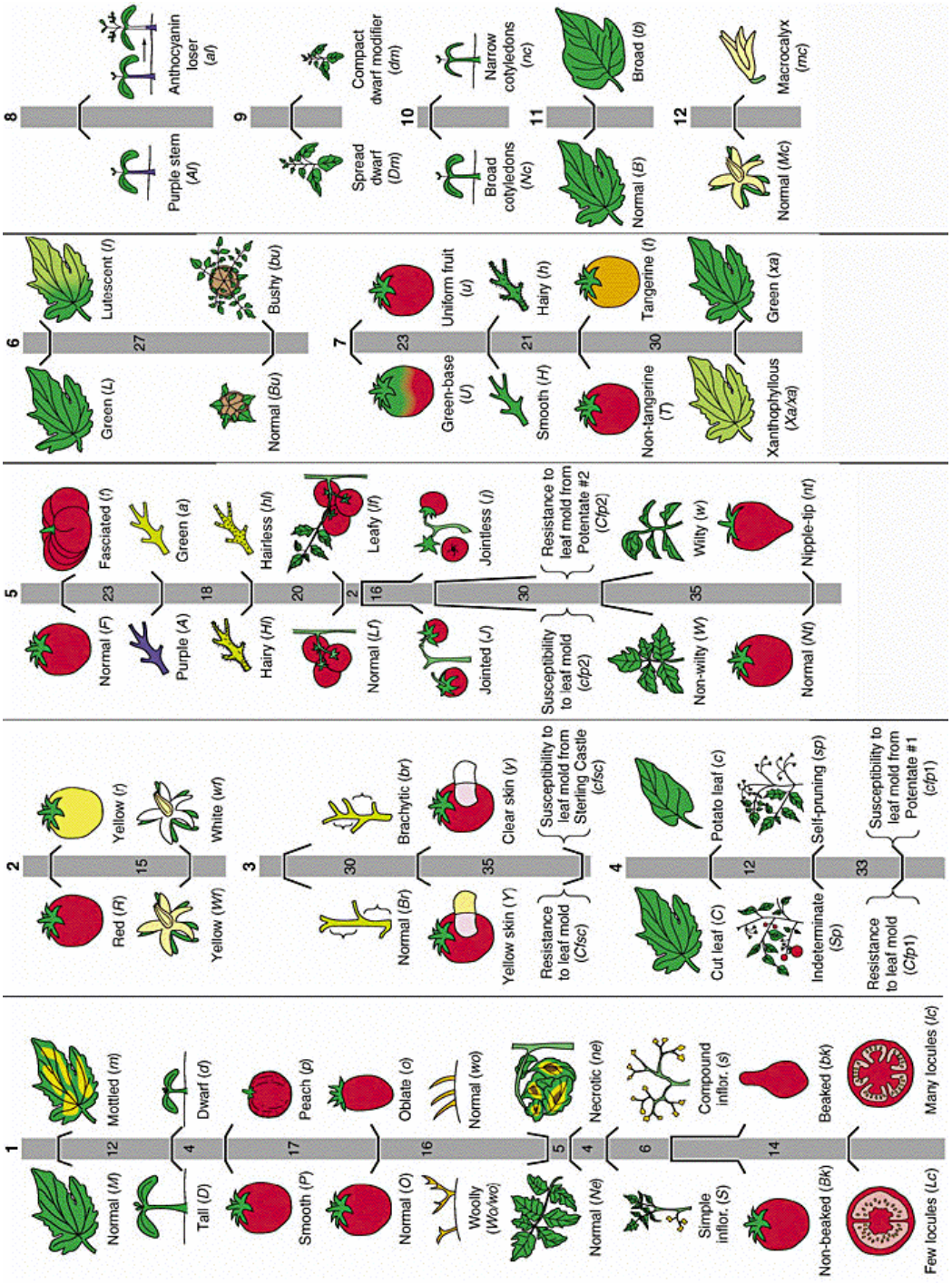


Figure 1-1. Carte Génétique de la tomate basée sur des caractères morphologiques. D'après Butler (1952)

De nombreux caractères importants en sélection, comme le rendement, la précocité de floraison, des traits liés à l'adaptation au milieu de culture, à la qualité de production ainsi que certaines résistances aux pathogènes, sont des phénotypes à distribution continue. Les principes de cartographie par liaison génétique ainsi que les principes de détection de QTL (Quantitative Trait Locus) étaient connus depuis le début du siècle mais faisaient intervenir des calculs très lourds. On peut dire que c'est l'avènement de l'informatique avec la naissance des microprocesseurs en 1971 qui va permettre le développement des biostatistiques. Cette

amélioration du temps de calcul est contemporaine à l'émergence de la biologie moléculaire moderne et ce n'est qu'en combinant ces deux avancées que les scientifiques ont pu développer des cartes génétiques haute densité.

Un enjeu majeur a alors été de cartographier des marqueurs moléculaires et des locus ayant un effet quantitatif sur ces cartes. En même temps, l'identification des gènes, responsables des variations des caractères d'intérêt, devenait un intérêt scientifique. Les végétaux sont considérés comme un matériel biologique de choix : des croisements contrôlés avec des phases d'autofécondation peuvent être effectués et un nombre élevé d'individus issus de ces croisements peuvent être étudiés sur une zone limitée. De nombreux travaux concernant la cartographie de QTL ont donc été initiés sur plusieurs espèces cultivées, comme le maïs (*Zea mays*) (Edwards, Helentjaris et al. 1992) et la tomate (*Solanum lycopersicum*) (Paterson, Lander et al. 1988). Par la suite, ces travaux ont permis d'isoler des gènes contrôlant la variation de caractères domestiqués (Doebley, Stec et al. 1997; Frary, Nesbitt et al. 2000). D'autres approches, dites de génétique inverse, permettent la caractérisation fonctionnelle de gènes, en induisant de façon ciblée ou aléatoire des mutations dans ces gènes et en étudiant l'effet de celles-ci sur le phénotype (Adams and Sekelsky 2002). Ces méthodes ne seront pas présentées ici.

1.1.2. La cartographie génétique de marqueurs, de gènes et de QTL

Comme indiqué ci-dessus, la première stratégie utilisée afin de préciser la position d'un QTL est la cartographie génétique (Figure 1-2). Pour cela, une carte génétique doit être construite en utilisant la descendance d'un croisement entre deux parents phénotypiquement différents. Différents types de descendance peuvent être utilisés comme la génération F2, des populations de lignées recombinantes ou des lignées d'haploïdes doublés. Les individus de la population étudiée doivent être phénotypés et génotypés à l'aide d'un grand nombre de marqueurs moléculaires. Il est important que les parents de la population étudiée soient polymorphes au niveau du caractère d'intérêt mais aussi au niveau génétique afin de pouvoir suivre la ségrégation des marqueurs moléculaires. Les évènements de recombinaison, qui ont eu lieu durant les différentes méioses subies par les individus, sont détectés et on peut alors construire une carte génétique. Le nombre de ces évènements entre deux marqueurs est transformé en distance génétique additive. Le caractère d'intérêt est mesuré sur ces mêmes individus, dans plusieurs lieux et durant plusieurs années, afin de prendre en compte la variation environnementale dans la recherche de QTL. Ensuite différentes approches

statistiques (régression simple, « interval mapping » ou « composite interval mapping ») peuvent être employées afin de détecter les localisations les plus vraisemblables des QTL (Thoday 1961; Haley and Knott 1992; Zeng 1993).

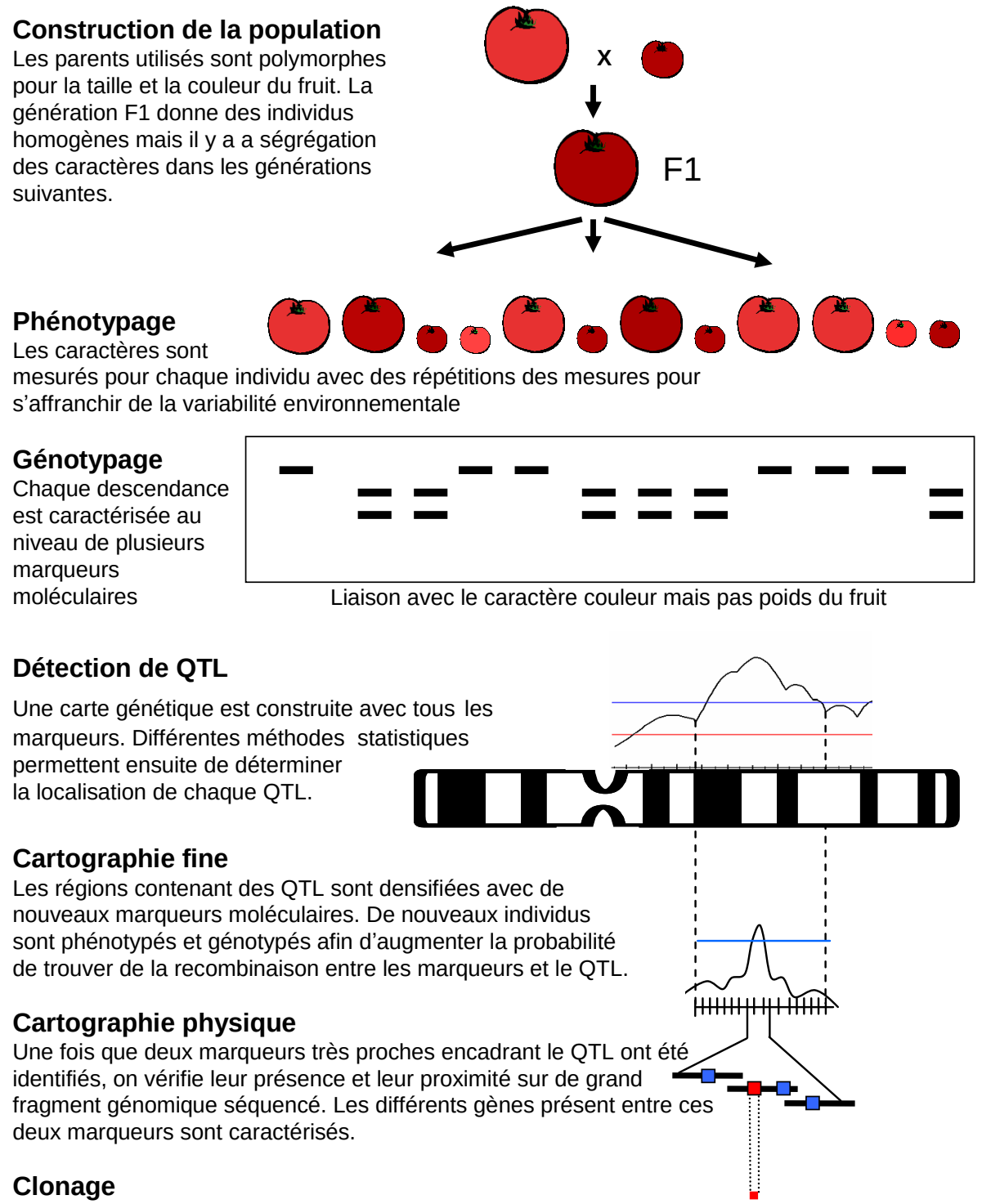


Figure 1-2. Différentes étapes dans la détermination des bases moléculaires d'un caractère par cartographie génétique.

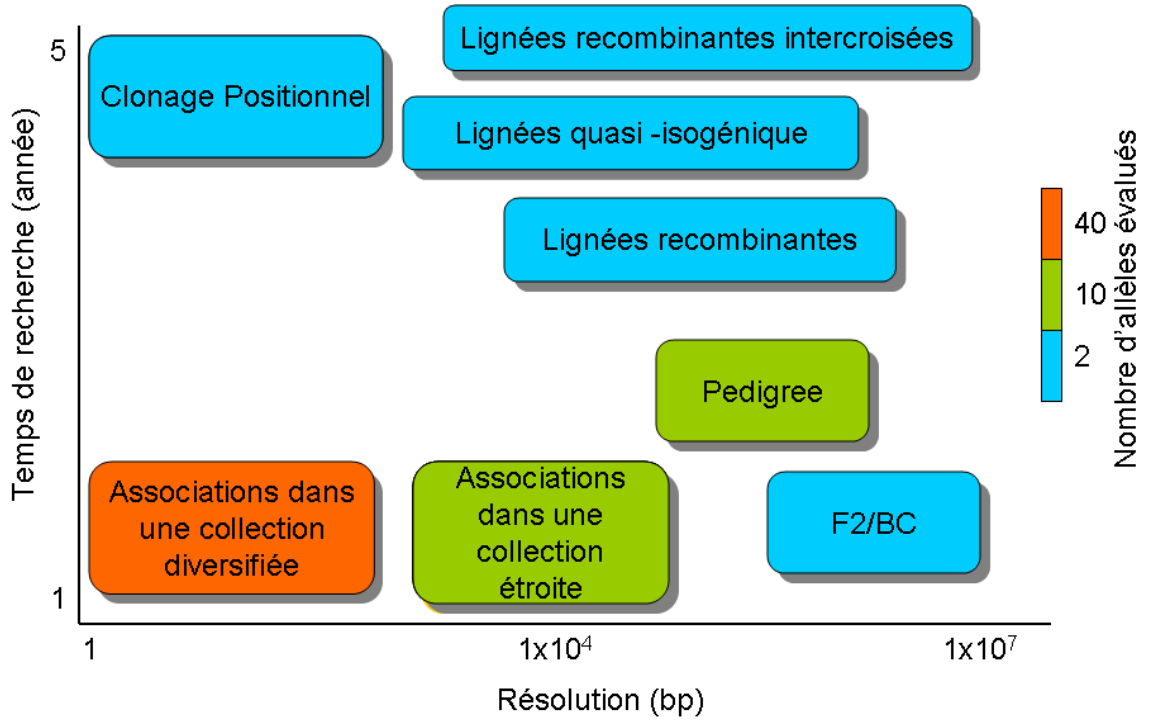
La résolution de cette approche, en termes de taille de fragment chromosomique identifié contenant le QTL, dépend bien entendu du nombre de marqueurs disponibles mais surtout du nombre d'évènements de recombinaisons. Celui-ci sera fonction du type de descendance et du nombre d'individus génotypés. La sélection assistée par marqueur se focalisera sur des marqueurs qui coségrègent avec le caractère d'intérêt ou qui expliquent une part significative de la variation d'un caractère quantitatif.

Afin d'identifier le gène sous-jacent au caractère, la région génétique contenant le QTL doit être affinée jusqu'à pouvoir isoler un gène ou une région génomique contenant le polymorphisme responsable. Le seul moyen d'affiner la région est d'augmenter encore le nombre de recombinaisons dans la région. Pour cela, un nombre très important d'individus sont typés. Ainsi un des QTL les plus importants lors de la domestication du maïs : *tb1* (*teosinte branch 1*) a pu être cloné en combinant une approche de cartographie de QTL, la production de lignées quasi-isogénique et l'utilisation d'une banque de mutants d'insertion (Doebley, Stec et al. 1997). Chez la tomate, plusieurs QTL (qui seront détaillés plus loin) ont été clonés en utilisant des stratégies de cartographie et de clonage positionnel (Frary, Nesbitt et al. 2000; Fridman, Carrari et al. 2004; Manning, Tor et al. 2006). Chez *Arabidopsis thaliana*, les gènes *Frigida* et *CRY2*, responsable respectivement de la réponse à la vernalisation et de la variation de la précocité de floraison, ont été isolés par clonage positionnel (Johanson, West et al. 2000; El-Din El-Assal, Alonso-Blanco et al. 2001). Malgré le fait que la cartographie de QTL continue à être une stratégie de choix pour identifier les gènes responsables des variations quantitatives chez les plantes, elle souffre de certains inconvénients. Tout d'abord ces études sont réalisées sur des populations biparentales, où les deux parents sont, la plupart du temps, homozygotes. Cela implique qu'on détecte ici seulement l'effet de deux allèles à chaque locus. De plus, les populations subissent souvent une seule méiose efficace générant les évènements de recombinaison, les autres générations servent juste à fixer ces évènements. La génération de populations de cartographie peut prendre un temps assez long (notamment pour les arbres qui ne forment pas d'organes reproducteurs jusqu'à quatre ou cinq années après semis).

Même lorsque la séquence du génome de l'espèce travaillée est disponible, le passage de la cartographie génétique au clonage du gène peut demander un travail important pour cribler des milliers d'individus et réaliser le phénotypage des recombinants (Mauricio 2001; Abdurakhmonov and Abdukarimov 2008). De plus, la découverte de nouveaux gènes est

limitée à ceux ayant un effet fort sur ces caractères à variation continue (Buckler and Thornsberry 2002). Une comparaison des différentes stratégies de cartographie est présentée dans la Figure 1-3.

A



B

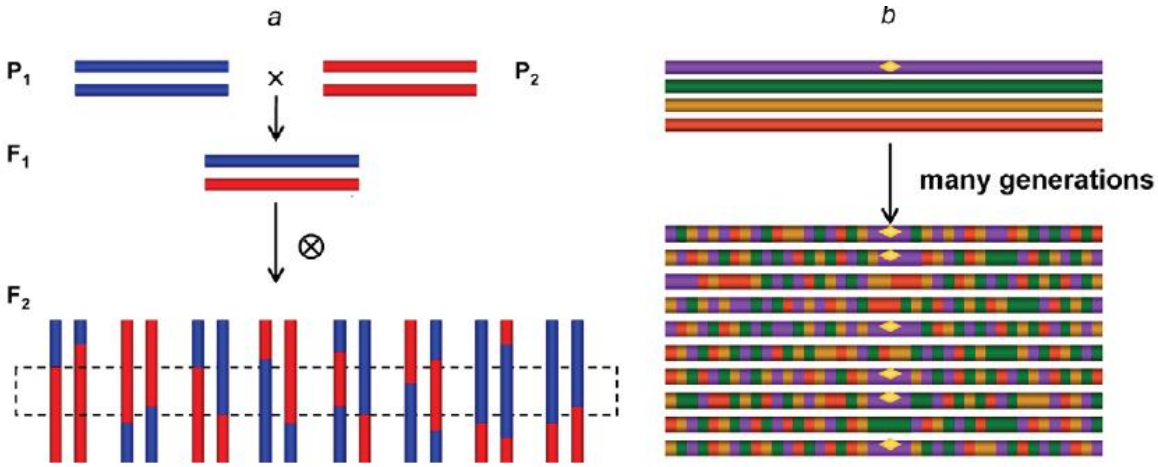


Figure 1-3. Comparaison schématique de différentes stratégies de cartographie. Modifié d'après Yu et Buckler (2006) et d'après Zhu et al. (2008).
(A) La comparaison au niveau du temps de recherche nécessaire, de la résolution et du nombre d'allèles évalués est représentée. La génétique d'association apparaît comme la plus intéressante pour sa rapidité, sa résolution ainsi que le nombre d'allèles étudiés. BC : Backcross. (B) Comparaison de la résolution entre les approches de cartographie classique (à gauche) et les études d'associations (à droite).

1.1.3. La génétique d'association

La cartographie par déséquilibre de liaison (DL), ou génétique d'association, semble être une méthode puissante pour identifier des gènes qui contribuent à la variation de caractères complexes (Long and Langley 1999). Cette stratégie a été développée pour pallier l'impossibilité de mettre en place des dispositifs d'étude de QTL (création de populations en ségrégation) chez l'humain (Spielman, McGinnis et al. 1993). Cette approche utilise comme échantillon un groupe d'individus non apparentés formant deux cohortes aux phénotypes contrastés (par exemple une cohorte saine et une cohorte malade pour une pathologie chez l'humain). Ces individus sont génotypés pour différents marqueurs puis des corrélations sont recherchées entre le phénotype et les allèles aux marqueurs. La suggestion, que les études de génétique d'association sont plus puissantes que les études de cartographie génétique, est basée sur l'hypothèse qu'un des marqueurs typés sera le polymorphisme causal. Cependant, tant que le génotypage des individus ne sera pas exhaustif (reséquençage complet des génomes individuels) il est vraisemblable que la densité des marqueurs utilisés ne sera pas suffisante pour justifier cette hypothèse. La cartographie par génétique d'association s'appuie sur une propriété génétique des populations naturelles : le déséquilibre de liaison (DL). Le DL représente l'association non aléatoire entre les allèles de différents locus, dans une population donnée. L'étendue de celui-ci, sur une région génétique ou physique, rend compte de la résolution de la cartographie par génétique d'association. Le DL est un écart à une population idéale (population panmictique avec absence de mutation, migration et sélection) qui respecte l'équilibre de Hardy-Weinberg pour des locus indépendants. Dans une telle population, le DL observé ne serait dû qu'à la distance génétique entre les deux locus. Différentes statistiques permettent d'estimer le déséquilibre de liaison entre deux locus, r^2 et D' étant les plus couramment utilisées (Figure 1-4) (Flint-Garcia, Thornsberry et al. 2003; Gupta, Rustgi et al. 2005). Les statistiques r^2 et D' reflètent différents aspects du DL. Bien que ni r^2 ni D' ne donnent des résultats convenables lorsqu'ils sont calculés sur de petits échantillons et, ou de faibles fréquences alléliques, chacun possède différents avantages. Alors que r^2 résume à la fois l'histoire des mutations et des recombinaisons, D' mesure uniquement l'histoire des recombinaisons. Cependant, D' est fortement affecté par la taille de l'échantillon car il présente un biais lorsqu'il est utilisé pour comparer des locus avec de faibles fréquences alléliques (diminution de la probabilité de retrouver les quatre combinaisons alléliques même si les locus sont indépendants). Dans le but d'analyser la résolution des études d'associations, on favorise généralement l'utilisation de r^2 (Flint-Garcia, Thornsberry et al. 2003).

Principales statistiques utilisées pour mesurer le DL :
On considère 2 locus avec les allèles *A* et *a* pour l'un et *B* et *b* pour l'autre.
Soit $\pi_A, \pi_B, \pi_a, \pi_b$ les fréquences alléliques et $\pi_{AB}, \pi_{aB}, \pi_{Ab}, \pi_{ab}$ les fréquences haplotypiques.

D représente l'écart entre les fréquences haplotypiques observées et attendues

$$D = (\pi_{AB} - \pi_A \pi_B)$$

Les satistiques r^2 et D' se calculent de la façon suivante

$$r^2 = \frac{(D_{AB})^2}{(\pi_A \pi_B \pi_a \pi_b)}$$

$$D' = \frac{|D_{AB}|}{\min(\pi_A \pi_B, \pi_a \pi_b)} si\ D_{AB} < 0$$

ou

$$D' = \frac{|D_{AB}|}{\min(\pi_A \pi_{b,} \pi_a \pi_B)} si\ D_{AB} > 0$$

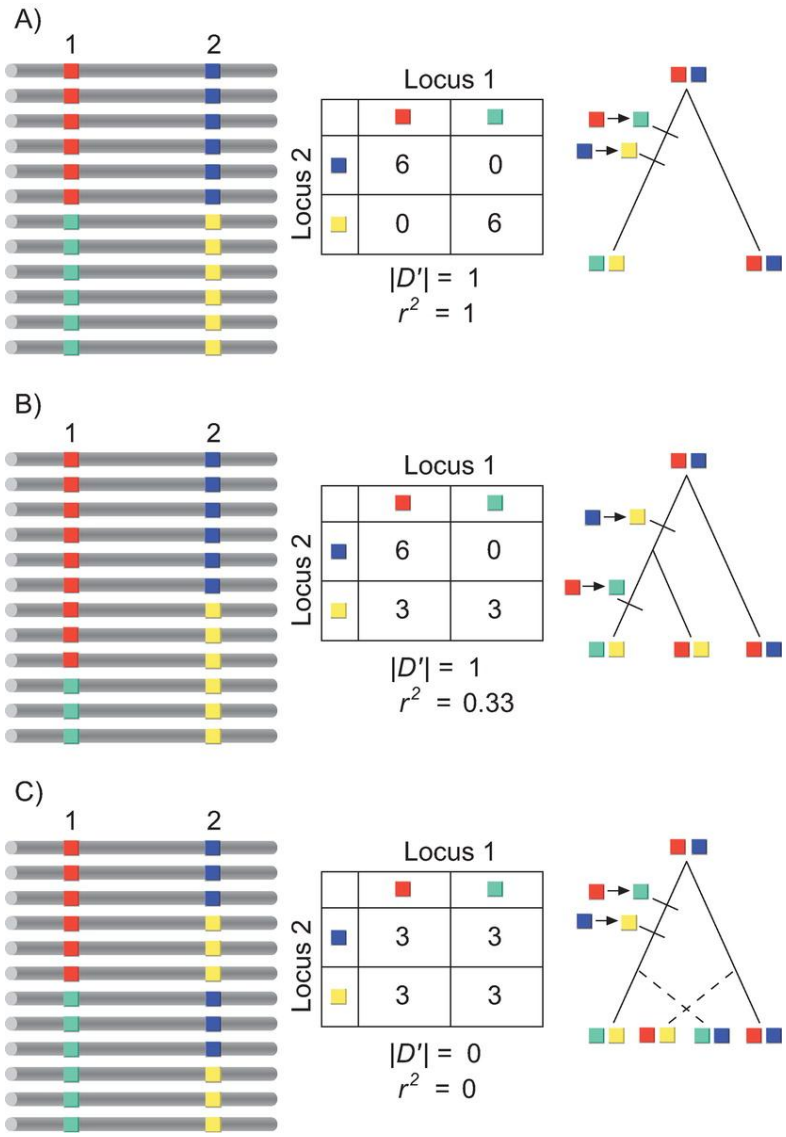


Figure 1-4. Méthode d'estimation du déséquilibre de liaison (DL) et comparaison de différents scenarii expliquant le DL entre deux locus polymorphes liés. Modifié à partir de Flint-Garcia, Thornsberry et al. (2003)
La partie de gauche présente le calcul de deux estimateurs du DL entre deux locus bialléliques. La partie de droite montre le comportement de r^2 et de D' en fonction de l'histoire évolutive des locus, affectés par la mutation et la recombinaison. (A) Les deux locus présentent une histoire mutationnelle similaire sans recombinaison. r^2 et D' sont maximum. (B) Le DL est dû à deux évènements de mutation successifs ayant eu lieu sur deux branches différentes sans recombinaison entre les locus. Le r^2 et D' sont ici très différents. (C) Ici la recombinaison entre les locus, ayant subi chacun un évènement de mutation, tend à annuler le DL. r^2 et D' sont égaux à 0.

Le DL présent dans une population naturelle va être créé par la mutation qui va induire l'apparition de nouveaux allèles. Le DL pourra ensuite augmenter s'il y a migration et/ou sélection (facteurs qui vont influencer les fréquences alléliques des nouveaux allèles). Le seul facteur capable de diminuer le DL entre deux locus est la recombinaison. La sélection et la liaison physique (diminution du nombre de recombinaisons) vont tendre à augmenter le DL local (à moins de sélectionner en même temps deux locus indépendants) alors que la dérive (due à la faible taille des populations) et le mélange de populations génétiquement différenciées vont tendre à augmenter le DL au niveau du génome entier (Tableau 1-1).

Facteur	**Effet**
Taux de recombinaison	Diminue le DL
Système de reproduction : espèce autogame	Fort DL
Système de reproduction : espèce allogame	Faible DL
Isolation génétique entre famille	Augmente le DL globalement
Subdivision de la population	Augmente le DL globalement
« admixture »	Augmente le DL globalement
Sélection naturelle et artificielle	Augmente le DL localement
Taille de la population	De petites populations présentent un DL plus fort (dérive)
Sélection balancée	Augmente le DL localement
Taux de mutation	De forts taux de mutation diminuent le DL. Le DL reste fort autour des nouveaux allèles mutés, jusqu'à ce qu'il diminue grâce à la recombinaison
Réarrangements génomiques	Les réarrangements suppriment la recombinaison locale ce qui entraîne une augmentation du DL dans le voisinage
Effets stochastiques (hasard)	Augmente ou diminue le DL

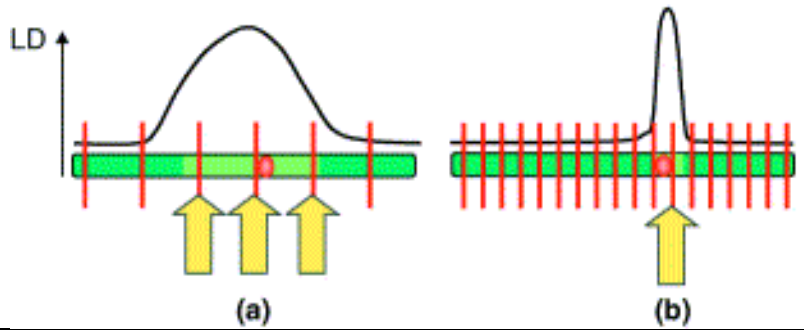
Tableau 1-1. Facteurs affectant le déséquilibre de liaison (DL) dans une population. D'après Rafalski et Morgante (2004)

La conversion génique (transfert non réciproque d'une information de séquence) peut agir comme une double recombinaison. Elle va donc diminuer le DL (Pritchard and Przeworski 2001; Wall and Pritchard 2003). Néanmoins, lors de l'association de deux brins d'ADN issus de chromosomes homologues, des mésappariements peuvent apparaître. Un mécanisme de réparation va venir rétablir la complémentarité des bases. Ce mécanisme va créer un biais dans la réparation en convertissant le plus souvent le mésappariement en une paire G-C. L'enrichissement en nucléotides G-C dans le locus va être corrélé avec une diminution du taux de recombinaison et, naturellement avec une augmentation du DL local (Galtier and Duret 2007).

On peut imaginer que pour des populations à l'équilibre dérive-migration-sélection, le DL avec un locus d'intérêt sera d'autant plus fort qu'on sera physiquement proche de celui-ci. La génétique d'association ou cartographie par DL identifie des marqueurs moléculaires qui sont liés au polymorphisme causal même si celui-ci n'est pas typé. La puissance de détection des associations marqueur/phénotype dépendra de la distribution du DL. En fonction de l'étendue de celui-ci, différentes approches peuvent être imaginées (Rafalski 2002). Si on estime que, pour une population donnée, le déséquilibre de liaison s'étend sur plusieurs centimorgans (peu importe le locus), alors deux marqueurs séparés d'une centaine de paires de bases (bp) apporteront une information redondante. Il sera inutile de génotyper la population pour ces deux marqueurs. Par conséquent, on pourra identifier la plupart des QTL en génotypant la population d'intérêt avec un nombre réduit de marqueurs répartis sur tout le génome. On pourra alors scanner celui-ci de façon large et identifier des zones en DL avec les polymorphismes causaux. Cette approche est nommée analyse du génome entier (Whole Genome Analysis ou WGA). Au contraire, si, pour cette population, le DL ne s'étend que sur une centaine de bp, la résolution sera plus importante. Le nombre de marqueurs moléculaires requis pour une analyse globale sera lui aussi plus important. Il sera néanmoins possible de s'intéresser à des régions particulières qui seront typées avec une densité plus grande de marqueurs selon une approche de type gène candidat (Candidate Gene Analysis ou CGA) (Figure 1-5).

Lorsqu'on utilise des populations naturelles, on est rarement confronté à l'équilibre de Hardy-Weinberg, où seule la distance génétique (nombre de recombinaisons) influe sur le déséquilibre de liaison. En effet, les populations naturelles peuvent avoir subi, dans le passé, différents évènements sélectifs et démographiques qui ont modelé le déséquilibre de liaison

(Nordborg and Tavare 2002). Comme déjà formulé ci-dessus, les effets sélectifs influenceront le DL au niveau local, autour du locus causal. Ils vont donc limiter la résolution d'une approche CGA mais améliorer la puissance d'une analyse de type WGA. L'histoire démographique de la population influencera le DL au niveau global ce qui améliorera les approches de type WGA.



Etendue du DL :	**Fort (a)**	**Faible (b)**
Résolution	Faible	Forte
Nombre de marqueurs requis	Faible	Elevé
Approche de cartographie par association	Analyse « Génome Entier » WGA	Analyse « Gène Candidat » CGA

Figure 1-5. Relation entre l'étendue du DL et la résolution des études d'associations. D'après Rafalski (2002)
Dans (a), le DL décroît lentement avec la distance lorsqu'on s'écarte du gène responsable du phénotype (rond rouge). Dans ce cas, une densité faible de marqueurs est suffisante pour identifier des marqueurs associés (flèches jaunes). Dans (b), le DL décroît très rapidement autour du gène responsable du phénotype et une densité plus grande de marqueurs est nécessaire pour identifier un marqueur associé.

Selon la population étudiée, il est possible d'identifier du DL entre locus localisés sur des chromosomes différents (locus non liés). Ce patron du DL va être causé par le mélange d'individus provenant de populations génétiquement différenciées. La structuration génétique d'une population va augmenter le DL global et va, par conséquent, augmenter le nombre de fausses associations identifiées (augmentation du nombre de faux positifs ou erreur de type I). L'importance de cette structure va diminuer avec les générations successives d'inter-croisements qui suivent les évènements de migration.

Lander et Schork (2006) montrent comment une étude d'association sur une population d'Amérindiens a conduit à une fausse association entre la sensibilité au diabète de type II et le locus G*m* (codant pour des gamma-globulines). Ils montrent que cette erreur est causée par la structuration génétique de la tribu. Des travaux supplémentaires ont montré que

l'association était due à un degré différent d'ascendance caucasienne chez les membres de la tribu. La présence de l'allèle caucasien, à n'importe quel gène, est corrélée avec un risque plus faible de développer un diabète de type II (risque plus faible chez les populations caucasiennes). Cet exemple révèle donc l'importance de prendre en compte la structure génétique de l'échantillon étudié lorsqu'on recherche des associations avec un caractère, et surtout, lorsque ce caractère influence la structure. Chez les végétaux, la première étude utilisant la génétique d'association en prenant en compte la structure génétique a été réalisée par Thornsberry et al. (2001). Cette étude a permis d'identifier une association significative entre *Dwarf8* et la précocité de floraison chez le maïs. Les nouvelles méthodes statistiques permettant d'identifier la structure génétique d'un échantillon sont responsables du succès de cette analyse. Pritchard, Stephens et al. (2000) proposent par exemple d'utiliser des données de génotypage de marqueurs répartis sur tout le génome pour inférer la structure génétique de l'échantillon. Ils ont développé le programme STRUCTURE dans ce but. Thornsberry, Goodman et al. (2001) utilisent cet algorithme pour identifier la structure génétique de leur échantillon et réinjectent l'information dans une régression logistique pour tester l'effet de chaque polymorphisme. Suite à cela, des travaux ont démontré l'efficacité de différents modèles d'analyse chez *A. thaliana* (Aranzana, Kim et al. 2005). Cette étude compare l'efficacité de différentes méthodes d'estimation de la structure génétique (programme STRUCTURE *vs.* coordonnées des individus calculées par analyse multifactorielle), Yu, Pressoir, et al. (2006) ont décrit une nouvelle méthode permettant de compléter les modèles linéaires, en utilisant les données d'apparentement entre individus. Ce modèle a été développé pour diminuer le taux de faux positifs dans des études d'association chez le maïs (allogame) et l'humain mais il semble aussi efficace chez l' espèce autogame *A. thaliana*, (Zhao, Aranzana et al. 2007).

Une seconde étude, qui s'est focalisée sur la précocité de floraison chez le maïs, fait suite aux travaux sur *Dwarf8* (Camus-Kulandaivelu, Chevin et al. 2008). Cette étude utilise un panel d'accessions différent avec des cultivars hybrides entre deux formes génétiquement différenciées. Le rôle adaptatif de *Dwarf8* dans la précocité de floraison est remis en cause car un autre locus, situé en région 5' de *Dwarf8*, est plus significativement lié au phénotype. Un DL important est détecté entre ce marqueur et certains polymorphismes identifiés dans *Dawrf8* lorsque les auteurs ne s'intéressent qu'aux populations initiales. Cela expliquerait les associations significatives identifiées par Thornsberry, Goodman et al. (2001). La rupture du DL, causée par de nombreux évènements de recombinaisons accumulés dans l'échantillon

hybride, expliquerait la disparition de l'association. Il semble donc important de noter que les résultats d'associations peuvent différer en fonction de l'échantillon étudié. Il est nécessaire de valider (ou invalider) des associations en utilisant d'autres échantillons indépendants. Le tableau 1-2 présente quelques résultats d'association chez les plantes.

Il semble cependant important de noter que la difficulté de conduire le phénotypage sur certains caractères présentant une variation phénotypique trop large représente une limite importante de la génétique d'association. La puissance de détection peut aussi être diminuée si les polymorphismes causaux présentent une faible diversité dans le panel étudié.

1.1.4. La génétique d'association chez les espèces cultivées autogames.

Chez les espèces autogames, on s'attend à avoir une forte structuration génétique ainsi qu'à divers niveaux d'apparentement entre individus (Yu, Pressoir et al. 2006). Les espèces autogames, largement homozygotes, peuvent être qualifiées de candidates idéales pour la génétique d'association (Aranzana, Kim et al. 2005). En effet, elles présentent l'avantage de pouvoir être maintenues facilement dans des collections de ressources génétiques, de pouvoir être génotypées une seule fois et phénotypées dans divers environnements. De plus, les autofécondations successives induisent une structuration des polymorphismes en de longs haplotypes. L'autogamie augmente le niveau d'homozygotie des individus et diminue le nombre de recombinaison efficace. Le DL s'en trouve naturellement augmenté. Cette augmentation du DL présente un avantage pour réaliser des analyses de type WGA par rapport à des espèces comme le maïs où le DL décroît de façon drastique au-delà de 1500 bp (Remington, Thornsberry et al. 2001). Par exemple, chez *Arabidopsis thaliana*, le DL diminue rapidement (10 Kb) lorsqu'on analyse des marqueurs répartis sur tout le génome (Kim, Plagnol et al. 2007). Ces résultats concordent avec des simulations réalisées par Nordborg (2000) qui montrent que le DL est réduit après 10 Kb pour des espèces autogames.

Cependant, des différences importantes existent entre espèces autogames sauvages et espèces autogames cultivées. Chez les espèces autogames cultivées, l'augmentation du DL est exacerbée par les pressions sélectives et les changements démographiques qui se sont exercés sur les populations, au cours de la domestication et de la sélection moderne. On se retrouve donc avec des populations d'accessions cultivées ayant des effectifs efficaces très fortement réduits par les goulets d'étranglement successifs liés à leurs histoires évolutives. Chez l'orge (*Hordeum vulgarae*), hautement autogame, le déséquilibre de liaison (seuil : $r^2 < 0.2$) s'étend jusqu'à 2,6 cM (~11 Mb) lorsqu'on étudie 170 cultivars utilisés au Canada durant les 20

dernières années (Zhang, Marchand et al. 2009). Par contre, le DL ne s'étend pas au delà de 250 Kb chez l'espèce sauvage apparentée (Caldwell, Russell et al. 2006).

Espèce	Régime de reproduction	Etendue du DL	Trait(s) étudié(s)	Référence
Maïs	Allogame	200–1500 bp		Remington *et al.* (2001)
			précocité de floraison	Thornsberry *et al.* (2001)
			précocité de floraison	Camus-Kulandaivelu *et al.* (2006)
			couleur du grain	Palaisa *et al.* (2004)
Arabidopsis	Autogame	10-500 Kb		Kim *et al.* (2007)
			précocité de floraison	Aranzana *et al.* (2005)
			résistance	Zhao *et al.* (2007)
			précocité de floraison	Ehrenreich *et al.* (2006)
Sorgho	Autogame	4 cM		Deu *et al.* (2004)
			caractères morphologiques et précocité de floraison	Casa *et al.* (2008)
Orge	Autogame	2,6 cM		Zhang *et al.* (2009)
			stress biotiques et abiotiques	Ivandic *et al.* (2003)
			adaptation saisonnière	Rostoks *et al.* (2006)
			rendement	Kraakman *et al.* (2004)
Lolium perenne	Allogame		500–3,000 bp	Skøt *et al.* (2005)
			précocité de floraison	Skøt *et al.* (2005)
			contenu en sucre soluble	Skøt *et al.* (2007)
Blé	Autogame	2 - 3 cM		Somers *et al.* (2007)
			morphologie du grain	Breseghello *et al.* (2006)
			résistance	Tommasini *et al.* (2007)
			rendement en grain et résistances	Crossa *et al.* (2007)
Pomme de terre	Allogame/clonal	300 bp		Simko *et al.* (2006)
			résistance à *Verticillium*	Simko *et al.* (2004)
			résistance à *Phytophtora*	Pajerowska-Mukhtar *et al.* (2009)
			qualité du tubercule	D'hoop *et al.* (2008)
Riz	Autogame	75-500 Kb		Mather *et al*. (2007)
			morphologie du grain	Iwata *et al.* (2007)
			qualité de l'amidon	Bao *et al.* (2006)
Tomate	Autogame	20 cM		van Berloo *et al.* (2008)
			poids du fruit	Nesbitt *et al*. (2002)
Soja	Autogame	100-600 Kb		Hyten *et al.* (2007)
Vigne	Allogame/clonal	5–10 cM		Barnaud *et al.* (2006)
			couleur de la baie	This *et al.* (2007)
			couleur de la baie	Fournier-Level *et al.* (2009)

Tableau 1-2. Quelques exemples d'études d'associations chez les plantes

Des valeurs intermédiaires concernant l'étendue du DL sont retrouvées chez les cultivars anciens (Caldwell, Russell et al. 2006). Chez le soja, de grandes différences dans l'étendue du DL sont également constatées entre le compartiment sauvage (90 Kb), domestiqué (200 Kb) et cultivé (500 Kb) (Hyten, Choi et al. 2007). Chez le riz (*Oryza sp.*), des étendues du DL entre 75 Kb et 500 Kb ont été décrites en fonction du groupe d'espèce examiné (Mather, Caicedo et al. 2007). Le déséquilibre de liaison chez le blé (*Triticum aestivum*) s'étend aussi sur de grandes régions (1 cM sur le chromosome 2D et jusqu'à 5 cM sur le chromosome 5A) (Breseghello and Sorrells 2006). Chez la tomate une étude visant à caractériser le DL à l'aide de marqueurs AFLP sur des variétés modernes indique qu'il s'étend jusqu'à 20 cM (Figure 1-6) (van Berloo, Zhu et al. 2008).

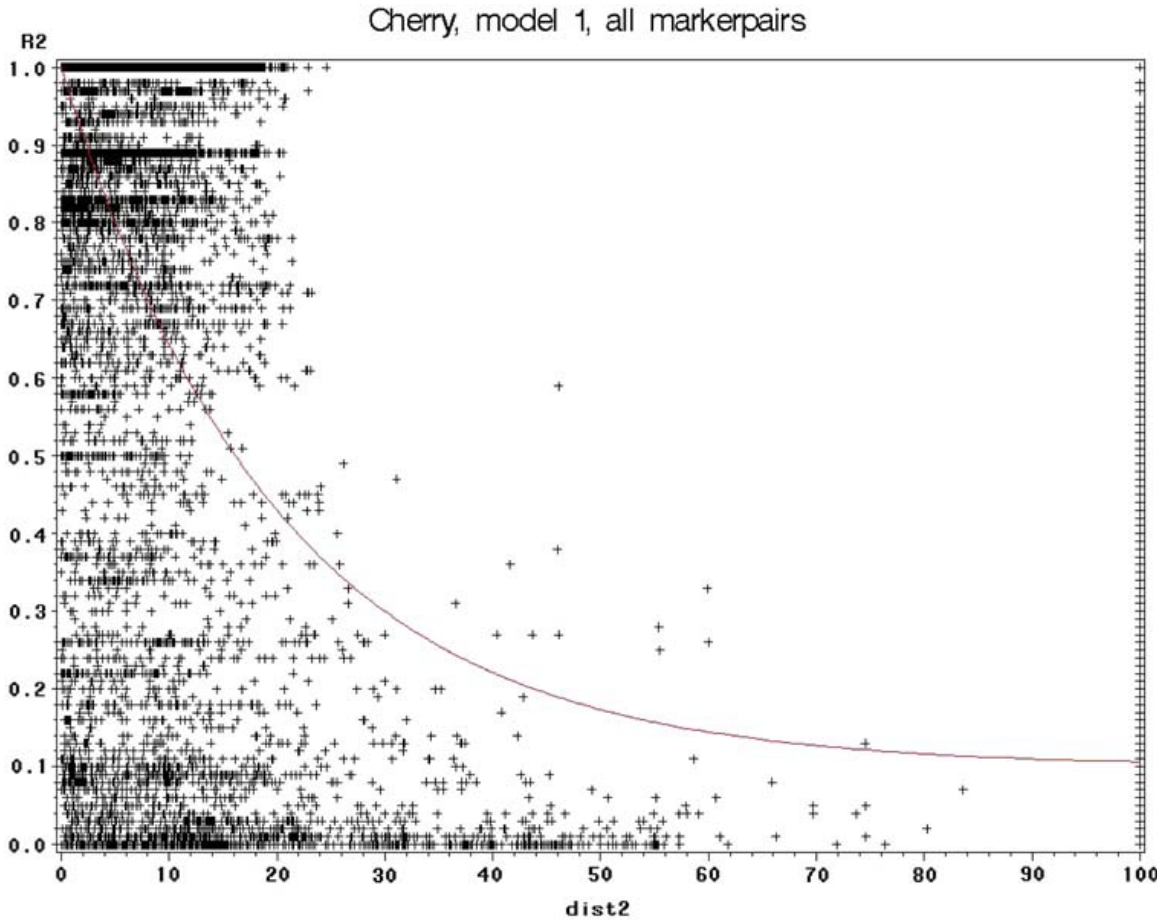


Figure 1-6. Etendue du déséquilibre de liaison chez la tomate mesuré chez 18 accessions de type cerise à l'aide de marqueurs AFLP. D'après van Berloo *et al.* (2008). La courbe représentée est une courbe de lissage des données. Les distances représentées sont des distances génétiques en cM.

La diminution de la diversité moléculaire est un autre problème majeur chez les espèces autogames cultivées. Celle-ci est corrélée à l'effectif efficace (dans le Modèle de

Wright-Fisher) par la relation $\theta = 4Ne\mu$ (où θ est l'estimateur de la diversité nucléotidique de Watterson, *Ne* est la taille efficace et μ le taux de mutation). L'effectif efficace des espèces cultivées a été réduit à cause des goulets d'étranglement inhérents à la domestication.

Tout ceci indique que les espèces autogames semblent pouvoir se prêter facilement à des études de type WGA avec un nombre limité de marqueurs mais ces analyses n'auront qu'une résolution relativement faible. Caldwell, Russell, et al. (2006) décrivent néanmoins la possibilité de réaliser des études préliminaires avec une résolution moyenne en utilisant un groupe d'accessions cultivées, puis de cartographier de façon plus fine les régions d'intérêt en utilisant des accessions sauvages ou des variétés anciennes. L'avantage de travailler avec des accessions domestiquées plutôt qu'avec les apparentées sauvages devient évident lorsqu'on s'intéresse à des caractères sélectionnés lors de la domestication. En effet, si le caractère n'est pas variable chez le groupe sauvage, aucune association ne peut être identifiée. Cette observation montre bien tout l'intérêt de conserver et caractériser, tant au niveau moléculaire que phénotypique, les ressources génétiques des espèces cultivées. La génétique d'association et l'analyse de la diversité génétique ouvrent la possibilité de valoriser à la fois le compartiment cultivé, sauvage mais aussi le compartiment intermédiaire constitué par les premières accessions domestiquées et des hybrides naturels (ou synthétique) entre les deux compartiments.

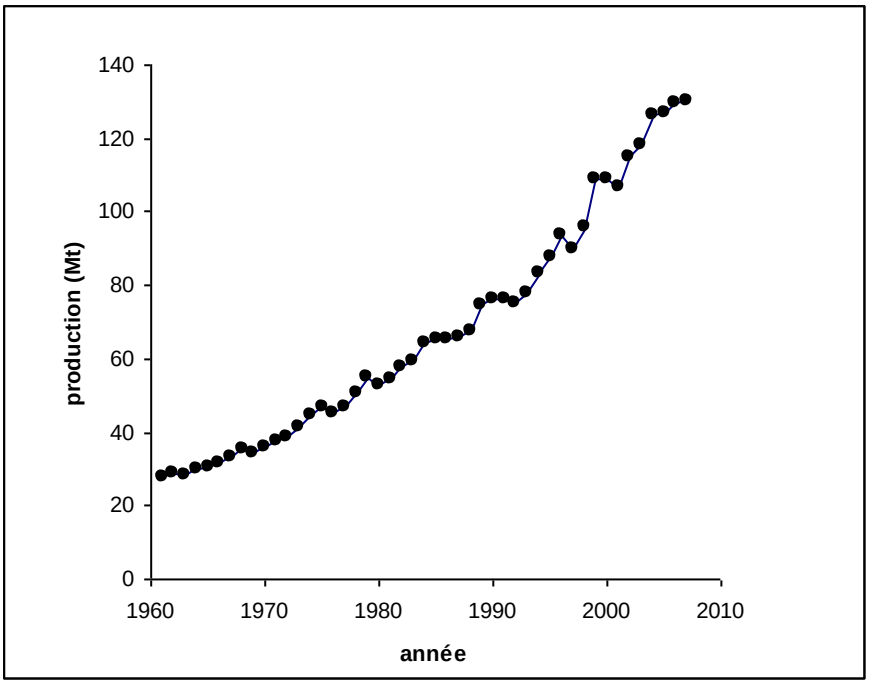
1.2. La Tomate (*Solanum lycopersicum* L. anciennement *Lycopersicon esculentum*)

1.2.1. Description

La tomate est une plante herbacée de la famille des solanacées, cultivée pour son fruit. Le terme désigne à la fois la plante et le fruit charnu qui, bien qu'il soit biologiquement un fruit, est considéré comme un des légumes les plus importants dans l'alimentation humaine. En termes de quantité produite en 2007, il s'agit de la douzième culture au niveau mondial et de la quatorzième au niveau européen (FAO, 2009). Avec près de 130 millions de tonnes produites en 2007 (FAO, 2009), la culture de la tomate est en plein essor au niveau international (Figure 1-7). Ce légume se consomme, soit cru, en mélange avec d'autres ingrédients ou en jus, soit cuit sous la forme de préparations variées à partir de produits frais ou transformés industriellement. De cela se dégagent deux grands types de cultures de la

tomate : la culture de frais, réalisée en général sous abris et récoltée manuellement et la culture d'industrie, réalisée en plein champs et récoltée mécaniquement.

A



B

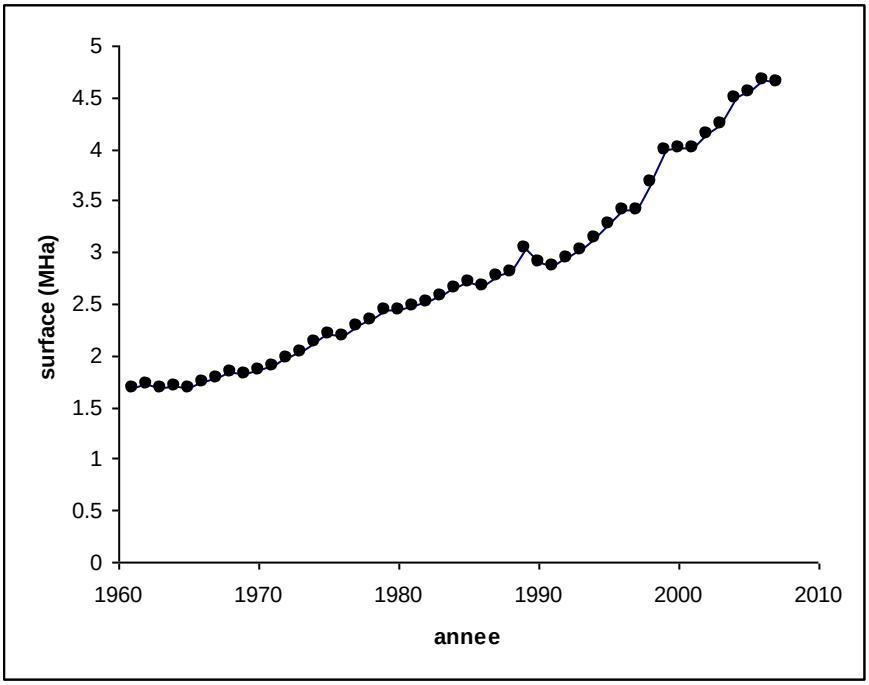


Figure 1-7. Croissance internationale de la production de tomate depuis 1960 en million de tonnes (A) et de la surface cultivée en million d'hectares (B). D'après les données de la FAO (2009)

La tomate est un aliment diététique riche en eau et pauvre en calories. Le fruit renferme aussi beaucoup d'éléments minéraux et de vitamines, dont la plus importante en quantité est la vitamine C (ou acide ascorbique). Lorsque le fruit est mûr, il contient aussi des pigments de la famille des caroténoïdes. Le β-carotène possède une activité de provitamine A. Le lycopène, aussi présent en grande quantité dans le fruit mûr (entre 3 et 8 mg/100 g de matière fraîche) mais surtout dans les concentrés de tomate (30 mg pour 100 g de concentré), joue un rôle d'anti-oxydant dans l'alimentation humaine et la prévention de certains cancers (Nguyen and Schwartz 1998; Giovannucci 1999).

Son importance économique ainsi que la disponibilité d'importantes ressources génomiques et génétiques pour cette plante, font d'elle, un modèle pour l'étude des solanacées et pour les études sur le développement du fruit charnu (Stevens 2007).

1.2.2. Biologie

La tomate est une plante vivace, généralement cultivée comme une annuelle. C'est une plante à croissance indéterminée (tige monopodiale), mais il existe certaines variétés à croissance déterminée (tige monopodiale puis sympodiale après 4 ou 5 feuilles). Le type de croissance déterminée a permis le développement de la récolte mécanisée, impossible sur les autres variétés qui doivent être tuteurées. Les feuilles sont alternes, composées, imparipennées (nombre impair de foliole) et comprennent 5 à 7 folioles aux lobes découpés. L'appareil reproducteur est formé par des inflorescences de type déterminé. La tomate est généralement autogame mais des allofécondations sont possibles. Les fleurs (Figure 1-8) sont hermaphrodites et actinomorphes. Le calice compte cinq sépales ou plus, de couleur verte. La corolle compte autant de pétales que de sépales, soudés à la base. L'androcée compte cinq étamines ou plus, à déhiscence latérale, introrses. Les anthères allongées forment un cône resserré autour du pistil. Ce dernier est constitué de plusieurs carpelles soudés, formant un ovaire supère biloculaire ou multiloculaire et à placentation centrale. Selon le cultivar et les conditions environnementales, le style peut être en position interne dans le cône d'étamine (fleur brévistyle), affleurant, ou dépasser légèrement (fleur longistyle). Cette caractéristique va jouer sur la possibilité du cultivar à subir des inter-croisements naturels. En culture sous abris, la pollinisation est assurée par des bourdons d'élevage (*Bombus terrestris*) ou par vibrage manuel des fleurs. En plein champ, le vent assure le vibrage des fleurs et permet la fécondation. En milieu naturel, une abeille de la famille des *Halictidae* (*Augochloropsis ignita*) a été décrite comme pollinisateur naturel potentiel (Reeves 1973).

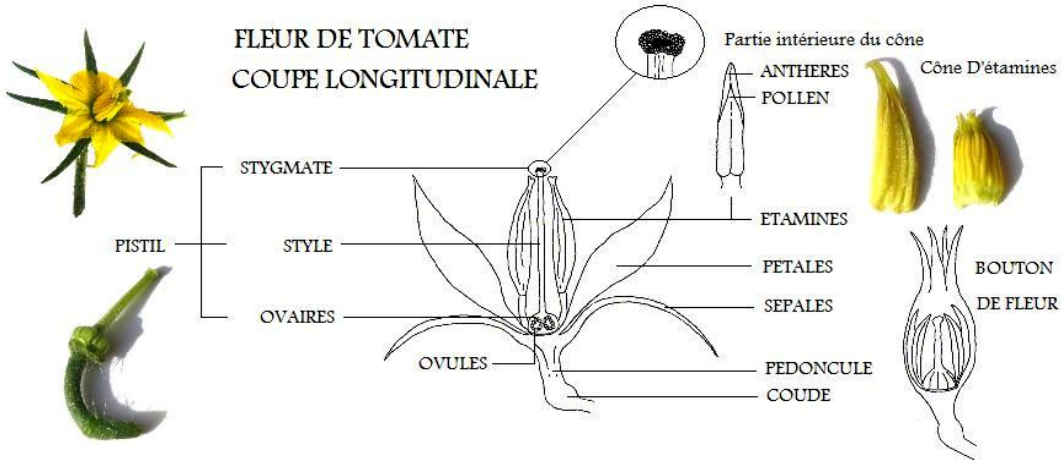


Figure 1-8. Coupe longitudinale d'une fleur de tomate. *Tomodori.com*. 24 nov. 2009. <**http://tomodori.com/phpBB2/viewtopic.php?t=4567**>.

Les fruits charnus sont des baies présentant deux ou plusieurs loges. Ils peuvent peser de quelques grammes à près de deux kilogrammes. Leur forme est généralement sphérique mais peut être plus ou moins aplatie, plus ou moins côtelée, en forme de cœur ou de poire (Figure 1-9). Les fruits sont verts puis virent généralement au rouge à maturité. Ils peuvent cependant être de couleur jaune, rose, orange, blanche, noire voire bicolore à maturité.

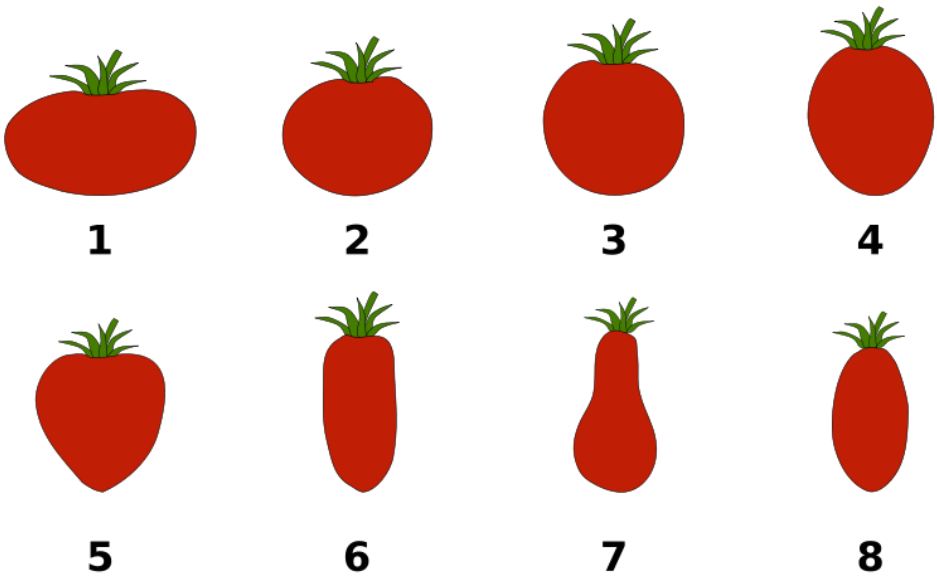


Figure 1-9. Différentes formes de tomates utilisées pour décrire une variété (descripteur IPGRI). *Wikipedia*. 24 nov. 2009. Wikimedia Foundation, Inc.
< http://fr.wikipedia.org/wiki/Fichier:Formes_de_tomates.svg>

1 : aplati
2 : légèrement aplati
3 : arrondi
4 : allongé arrondi (ovoïde)
5 : cordiforme
6 : cylindrique
7 : pyriforme
8 : obovoïde (forme de prune)

1.2.3. Caractéristiques génomiques

La tomate est une espèce diploïde possédant 12 paires de chromosomes (2n =2x =24). La taille de son génome est estimé à 950 Mb (2C = 1,90 pg d'ADN) et il contiendrait environ 35 000 gènes (van der Hoeven, Ronning et al. 2002). Le génome de la tomate est de taille intermédiaire entre celui d'*A. thaliana* (2n=20, 2C = 130 Mb) et celui du maïs (2n = 20, 2C = 2500 Mb). La carte génétique de référence (http://solgenomics.net/) a été construite à partir d'un croisement interspécifique entre un cultivar de l'espèce *S. lycopersicum* et une accession de l'espèce *S. pennellii* (Tanksley, Ganal et al. 1992). Elle a été construite à partir de 80 individus F2 et mesure près de 1300 cM. De nombreux marqueurs (2506 au total) sont cartographiés sur cette carte : marqueurs RFLP (Restriction Fragment Length Polymorphism), marqueurs microsatellites (Simple Sequence Repeat - SSR) ou marqueurs COS (Conserved Orthologous Sequence) issus de séquences conservées entre solanacées et *A. thaliana* (Fulton, Van der Hoeven et al. 2002).

Une première collection de 120,892 EST (Expressed Sequence Tag) a été développée à partir de 23 banques d'ADNc (ADN complémentaire) et réduite à 27,274 séquences consensus uniques : unigènes (van der Hoeven, Ronning et al. 2002). La plupart des gènes semblent être localisés au niveau de l'euchromatine qui représente seulement un quart de l'ADN total. Cette collection d'EST a été complétée par de nouvelles librairies d'ADNc disponibles sur la base de données 'Tomato Gene Index' (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=tomato). Cette ressource est complétée par :

- le SOL Genomics Network http://solgenomics.net/
- la Micro-Tom database (MiBase) http://www.kazusa.or.jp/jsol/microtom/
- la Tomato Stress EST Database (TSED) http://abrc.sinica.edu.tw/tsed/app_all/index.php
- la TomatEST DB http://biosrv.cab.unina.it/tomatestdb/ qui intègre l'information de toutes les bases de données précédentes (D'Agostino, Aversano et al. 2006).

Aujourd'hui une collection de plus de 300 000 EST, répartis dans ces différentes bases de données, est disponible.

Des banques BAC (Bacterial Artificial Chromosome) ont été développées pour initier la cartographie physique. La première banque a été construite à partir du cultivar Heinz 1706. Elle comprend environs 129 000 clones ayant une taille moyenne de 117,5 Kb (Budiman, Mao et al. 2000). Cette banque, ainsi que deux autres non publiées, ont permis d'initier le

séquençage du génome de la tomate à travers une collaboration réunissant 10 pays (dont la France). Le séquençage se concentre sur les 220 Mb d'euchromatine, riche en gènes (Mueller, Tanksley et al. 2005). Aujourd'hui, 50% du séquençage est disponible avec 1203 BAC séquencés sur 2500 initialement prévus (Figure 1-10). Grâce aux nouvelles technologies de séquençages (« Next-Generation Sequencing technologies »), 7409 « scaffolds » représentant 794 Mb de séquence génomique sont désormais disponibles. L'annotation sera disponible courant 2010.

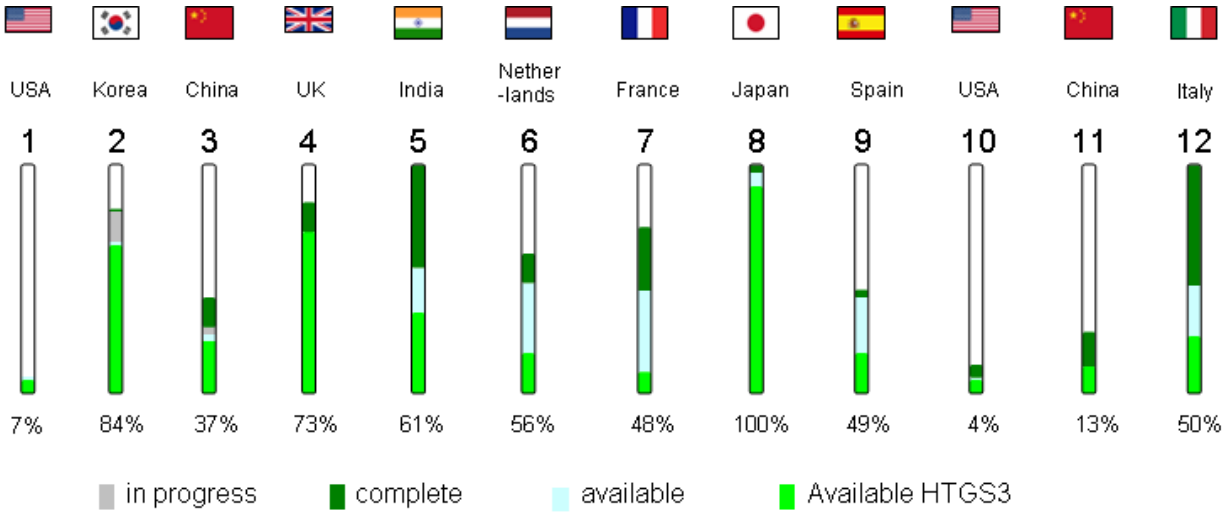


Figure 1-10. Avancée du projet international de séquençage du génome de la tomate. *Sol Genomics Network*. 24 nov. 2009.
<http://solgenomics.net/about/tomato_sequencing.pl>
HTGS3 (High Throughput Genomic Sequence 3) signifie que les séquences génomiques sont finies, sans gaps, avec ou sans annotations.

1.2.4. Ressources génétiques

La collection, la description, la propagation et la distribution de matériel génétique sont d'une grande importance pour l'amélioration de la tomate. Nikolaï Ivanovitch Vavilov (1887-1943) est le premier à démontrer l'importance de créer des collections de ressources génétiques et il lance au début du XX^{ème} siècle des expéditions à travers le monde afin de collecter des semences de différentes espèces (Kurlovich, Rep'ev et al. 2000). De gros efforts ont été initiés à partir de la deuxième moitié du XX^{ème} siècle, pour prospecter dans le centre de diversification du genre *Solanum* et récolter des centaines d'accessions sauvages. Charles M. Rick (1915-2002) a organisé, au sein du Tomato Genetics Resource Center (http://tgrc.ucdavis.edu/) à Davis en Californie, les accessions prospectées dans les Andes (espèces sauvages), au Mexique et dans le monde entier (espèce cultivée). Les accessions cultivées de tomate et les accessions apparentées proches sont majoritairement autogames et donc facilement propagées par autofécondation. Les accessions sauvages apparentées plus

éloignées sont majoritairement allogames. Ces accessions sont maintenues en population : elles sont intercroisées au sein de chaque population. D'autres collections ont été créées pour la tomate réunissant des accessions uniques ainsi que des accessions partagées entre instituts :

- la collection de tomate du North central Regional Plant Introduction Station (USA) : http://www.ars-grin.gov/npgs/searchgrin.html
- la collection du N.I. Vavilov Research Institute of Plant Industry (St Petersbourg, Russie) : http://www.vir.nw.ru/data/dbf.htm
- la collection du Leibniz Institute of Plant Genetics and Crop Plant Research (Gatersleben, Allemagne) : http://pgrc.ipk-gatersleben.de/
- la collection de la Estación Experimental La Mayora (Algarrobo costa, Málaga, Espagne)
- la collection de Ressources génétiques tomate du Centre INRA d'Avignon : http://w3.avignon.inra.fr/rg_tomate/

L'European Cooperative Programme for Plant Genetic Resources Tomato Database (Wageningen, Pays Bas, http://documents.plant.wur.nl/cgn/pgr/tomato/) est une base de données développée pour regrouper l'information contenue dans les collections européennes.
Le projet européen EU-SOL, a permis de réunir près de 7000 lignées de tomate cultivée ainsi que des accessions apparentées sauvages. Ces accessions sont maintenues et caractérisées par l'équipe de Dani Zamir (The Hebrew University of Jerusalem, Rehovot, Israël).

Parallèlement aux ressources naturelles, de nouvelles ressources ont été développées artificiellement dans le but d'étudier la ségrégation de caractères d'intérêts et de permettre leur identification. Ces populations sont issues principalement de croisements entre une accession cultivée et une accession sauvage. Ces populations interspécifiques présentent l'avantage d'être hautement polymorphes au niveau moléculaire. Une population F2 issue d'un croisement entre le cultivar VF36-Tm2a et l'accession sauvage LA716 (*S. pennelli*) a permis de construire la carte génétique haute densité, référence pour la tomate (Tanksley, Ganal et al. 1992). D'autres populations (Tableau 1-3), utilisant différents parents sauvages, ont été créées et ont montré leur efficacité dans la détection de QTL ainsi que dans l'identification des bases moléculaires de caractères d'intérêt (Grandillo, Ku et al. 1999; Paran and van der Knaap 2007).

Parent Sauvage	Trait(s) étudié(s)	Références
S. pennellii	Rendement - poids - SSC	(Eshed and Zamir 1995)
	Poids du fruit - SSC - AT - concentration sucres réducteurs & acides	(Causse, Duffe et al. 2004)
S. habrochaites	Rendement - poids du fruit - SSC -couleur - fermeté	(Bernacchi, Beck-Bunn et al. 1998)
	Rendement - poids du fruit - SSC - couleur - forme du fruit	(Monforte and Tanksley 2000)
S. peruvianum	Rendement - poids du fruit - SSC - AT - dosage sucres réducteurs & acides - fermetée - couleur - précocité - exertion du style	(Fulton, Beck-Bunn et al. 1997)
S. chmielewskii	Poids du fruit – SSC - pH	(Paterson, Lander et al. 1988; Paterson, DeVerna et al. 1990)
	Rendement - poids du fruit - SSC	(Yousef and Juvik 2001)
	poids et composition du fruit - physiologie du fruit - développement de la plante	(Prudent, Causse et al. 2009)
S. neorickii	Rendement - poids du fruit - viscosité - SSC - couleur - concentration Lycopène et β-carotène - forme du fruit - fermeté - précocité - AT - pH	(Fulton, Grandillo et al. 2000)
S. cheesmaniae	Poids du fruit - SSC - pH - croissance - couleur	(Paterson, Damon et al. 1991)
	Poids du fruit - SSC - poids des graines	(Goldman, Paran et al. 1995)
S. pimpinellifolium	Morphologie de la plante - rendement - poids du fruit - viscosité - SSC - couleur- fermeté - forme du fruit	(Grandillo and Tanksley 1996)
	Morphologie de la plante - poids du fruit - SSC - couleur - forme du fruit - nombre de loge - nombre de graine - précocité	(Tanksley, Grandillo et al. 1996)
	Poids du fruit - forme du fruit - SSC - pH - contenu en lycopène	(Chen, Foolad et al. 1999)
	Poids du fruit - nombre de loges - forme du fruit - poids des graines	(Lippman and Tanksley 2001)
	Précocité de floraison - précocité de maturité - nombre de fleur par inflorescence - craquellement - poids du fruit - couleur du fruit - fermeté - SSC - forme du fruit - poids des graines	(Doganlar, Frary et al. 2002)
	Forme du fruit - poids du fruit - nombre de loges - nombre de fleur par inflorescence	(van der Knaap and Tanksley 2003)
S. lycopersicum var. *cerasiforme*	Poids du fruit - fermeté - élasticité - couleur du fruit - SSC - contenu en sucre - pH - AT – contenu en lycopène et β-carotène contenu en arômes volatils	(Saliba-Colombani, Causse et al. 2001)
	Idem + 5 descripteurs de texture et 7 descripteur d'arôme évalués par un panel de dégustateur.	(Causse, Saliba-Colombani et al. 2002)
	Cartographie fine sur le chr. 2 de QTLs liés à poids du fruit - fermeté - SSC - contenu en sucre - AT	(Lecomte, Saliba-Colombani et al. 2004)

SSC : Soluble Solide Content (teneur en solide soluble), AT : acidité titrable.

Tableau 1-3. Résumé des différentes populations de cartographie développées et quelques références associées à la détection de QTL lié à la qualité du fruit.

De tous les systèmes modèles, les croisements interspécifiques chez la tomate ont été les premiers utilisés pour développer des populations de lignées d'introgression. Ces lignées permettent de partitionner les variations quantitatives en composantes à héritabilité mendélienne (Paterson, DeVerna et al. 1990). Une des plus utilisées est la population de lignées quasi isogéniques, développée par D. Zamir (Eshed and Zamir 1995) à partir d'un croisement entre M82 (*S. lycopersicum*) et LA716 (*S. pennellii*) (Paran, Goldman et al. 1995). Depuis sa création, près de 2800 QTL ont été identifiés sur cette population et, maintenant, des sous-lignées sont produites afin de cartographier plus finement une partie de ces QTL (Lippman, Semel et al. 2007). Une seule population intraspécifique a été développée à partir de deux accessions cultivées chez la tomate. Cette population de lignées recombinantes a été construite à partir d'un croisement entre un cultivar moderne aux qualités organoleptiques moyennes et une tomate de type cerise aux qualités organoleptiques supérieures (Saliba-Colombani, Causse et al. 2000).

Des populations de mutants ont été développées à partir de la lignée à croissance déterminée M82 (Menda, Semel et al. 2004), à partir de la lignée à croissance indéterminée Red Setter (Carreiro, Petrozza et al. 2004) et à partir de la lignée naine Micro-Tom (gènes *dwarf*, *miniature* et *SELF-PRUNING*). La lignée de mutant issue de l'accession Micro-Tom permet de phénotyper un grand nombre de mutants sur une surface limitée (Meissner, Chague et al. 2000; David-Schwartz, Badani et al. 2001; Eyal and Levy 2002; Dan, Fei et al. 2007). Ces mutants sont utilisés dans des études de TILLING (Targeting Induced Local Lesions in Genomes) (McCallum, Comai et al. 2000).

1.2.5. Taxonomie

La tomate fait partie de la famille des *Solanaceae* et du genre *Solanum*. La famille des *Solanaceae* comprend 94 genres et environ 2950 espèces cosmopolites. Les *Solanaceae* comprennent notamment le tabac (*Nicotiana tabacum*) et le piment (genre *Capsicum*). D'autres espèces sont connues pour leurs vertus psychotropes comme la mandragore (*Mandragora officinarum*) ou la belladone (*Atropa belladonna*). Le genre *Solanum* est très important dans le monde (environ 1700 espèces recensées) et comporte des plantes alimentaires comme la pomme de terre (*Solanum tuberosum*), et bien sûr la tomate. Ce genre comprend aussi des plantes ornementales comme la morelle faux jasmin (*Solanum jasminoides*). Certaines espèces de la famille des *Solanaceae* sont connues pour leur toxicité comme la morelle noire (*Solanum nigrum*) ou la douce amère (*Solanum dulcamara*).

Espèce	Couleur du Fruit	Mode de Reproduction [a]	Distribution et habitat	Importance pour l'amélioration variétale
S. cheesmaniae [L. cheesmaniae]	Jaune, orange	AC, exclusivement At	Endémique des Iles Galápagos. Du bord de mer à zone volcanique 0-1300 m, zone volcanique.	Tolérance salinité, résistances virus et lépidoptères, épaisseur du péricarpe et jointless
S. chilense [L. chilense]	vert, rayures violettes	SI, Al	Natif du Sud du Pérou au Nord du Chili, 0-3000 m, lits de rivières asséchées.	Résistance à la sécheresse
S. chmielewskii [L. chmielewskii]	vert	SC, Al facultatif	Natif du Sud du Pérou au Nord de la Bolivie, 1500-3000 m, zones drainées.	Augmentation du contenu en sucre
S. galapagense [L. cheesmaniae f. minor]	jaune, orange	AC, exclusivement At	Endémique des Iles Galapagos, bord de mer.	Idem que *S. cheesmaniae*
S. habrochaites [L. hirsutum]	vert	Typiquement Al	Natif du Sud-Ouest de l'Equateur au Centre Sud du Pérou, 500-3300m, régions boisées.	Tolérance au froid et au gel, résistances aux insectes et autres résistances
S. lycopersicum [L. esculentum]	rouge	AC, Al facultatif	Probablement natif de l'Equateur et Pérou, maintenant répandu, large gamme d'habitats.	Tolérance à l'humidité, résistances aux champignons pathogènes
S. neorickii [L. parviflorum]	vert pale	AC, fortement At	Natif du Sud de l'Equateur au Centre Sud du Pérou, 1500-3000m, environnement rocheux, humide et bien drainé.	
S. pennellii [L. pennellii]	vert	Normalement Al	Natif des côtes péruviennes, 500-1500m (50m), habitats chauds et secs, sujet à brouillard.	Résistance à la sécheresse, résistances aux insectes (pilosité dense)
S. peruvianum [L. peruvianum]	vert	SI typique, Al, rare pop. AC, At	Natif du Nord du Pérou au Nord du Chili, 0-3000 m, large gamme d'habitats.	
S. pimpinellifolium [L. pimpinellifolium]	*rouge*	*AC, At, Al facultatif*	*Natif du Sud de l'Equateur et du Nord du Pérou, <1000 m, vallée Sud des côtes pacifiques.*	*Amélioration de la couleur et de la qualité du fruit, résistances à insectes, nématodes.*

[a] AC = auto-compatible, AI = auto-incompatible, At = autogamme, Al = allogamme

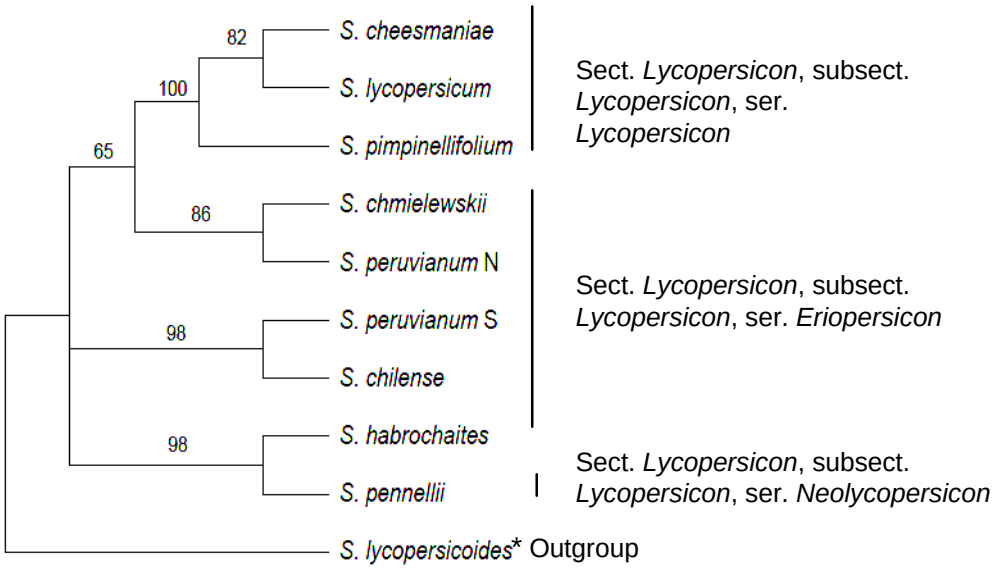
Tableau 1-4. Caractéristiques des espèces sauvages de tomate (*Solanum* L. section *Lycopersicon* subsection *Lycopersicon*). D'après Spooner, Peralta et al. 2005.
Les noms synonymes du genre *Lycopersicon* suivent les noms du genre *Solanum*.

Dès que la tomate a été introduite en Europe, les botanistes ont inclus cette espèce dans le même genre que la pomme de terre : le genre *Solanum*. Anguillara, nomme en 1561 cette plante nouvellement introduite, *Lycopersicon* qui signifie « pêche de loup » mais il y a confusion pour ce terme qui pourrait désigner le datura, l'aubergine ou la tomate. En 1694, Tournefort est le premier à distinguer la tomate cultivée et à créer un nouveau genre pour classer cette espèce : le genre *Lycopersicum*, (Peralta and Spooner 2007). Linné, en 1753, revoie la taxonomie de la tomate et l'intègre à nouveau dans le genre *Solanum* sous le nom spécifique de *Solanum lycopersicum*. Un an après, Miller reconsidère la classification évoquée par Tournefort et réactualise le genre *Lycopersicon* dans la quatrième édition de *The Gardener's Dictionary*. En 1768, Miller inclut dans ce genre trois espèces de tomate : *L. esculentum*, *L. peruvianum* et *L. pimpinellifolium* ainsi que la pomme de terre qu'il renomme *L. tuberosum*. Dans une version posthume de *The Gardener's and Botanist's Dictionary*, l'éditeur, Thomas Martyn, reconnaît la classification de Linné établie 50 ans plus tôt. Il inclut *Lycopersicon* dans le genre *Solanum* (Labate, Grandillo et al. 2007). Même si la tomate (nommée *Lycopersicon esculentum*) a été reconnue jusqu'à récemment, comme faisant partie du genre *Lycopersicon* par la plupart des taxonomistes, la classification de la tomate au sein du genre *Solanum* est aujourd'hui largement acceptée. Elle est appuyée par les résultats convergents de plusieurs études phylogénétiques portant sur des critères morphologiques et moléculaires (Bohs and Olmstead 1997; Spooner, Peralta et al. 2005). L'espèce cultivée *Solanum lycopersicum* se décline en deux variétés botaniques : la variété *esculentum* et la variété *cerasiforme*. Ces deux variétés se distinguent essentiellement par la taille du fruit (comprise entre 1,5 et 2,5 cm pour la variété *cerasiforme*) et le nombre de loges (deux loges pour *cerasiforme* et deux loges ou plus pour *esculentum*) (Rick, Latterot et al. 1990). La tomate de type cerise (*S. lycopersicum* var. *cerasiforme*) est décrite comme l'ancêtre domestiquée de la tomate cultivée (Miller and Tanksley 1990; Rick and Holle 1990; Bai and Lindhout 2007).

Toutes les espèces sauvages de tomates sont natives du Sud-Ouest de l'Amérique. Elles se distribuent le long des côtes andines, de l'Equateur jusqu'au nord du Chili avec deux espèces endémiques des Iles Galápagos. Le système de reproduction est variable entre ces espèces et va d'allogame auto-incompatible à autogame auto-compatible en passant par allogame facultative auto-compatible (Tableau 1-4). Toutes les tomates sauvages peuvent être croisées avec la tomate cultivée (parfois avec certaines difficultés ; en prenant l'espèce cultivée comme femelle). Elles jouent un rôle capital dans l'amélioration variétale car elles

représentent des sources de résistance génétique ou d'allèles favorables pour des caractères agronomiques (Spooner, Peralta et al. 2005). Des études phylogénétiques incluent ces espèces sauvages dans la section *Lycopersicon* au sein du genre *Solanum* (Figure 1-11).

A



B

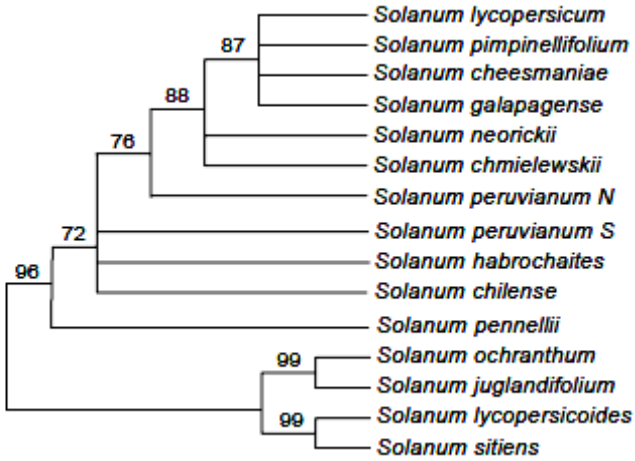


Figure 1-11. Arbre phylogénétique des espèces apparentées à la tomate (*S. lycopersicum*). D'après Spooner, Peralta et al.(2005)
(A) Les informations de 65 accessions sont résumées dans une analyse cladistique combinant l'information de marqueurs AFLP, de la séquence du gène GBSSI, de profil de restriction de l'ADN chloroplastique et de fragment ITS (Iternal Transcribed Spacer) de l'ADN ribosomique. (B) Analyse cladistique prenant en compte seulement des caractères morphologiques. Les valeurs sur les branches représentent les valeurs de bootstrap.

Au sein de cette section, les taxonomistes ont créé une division entre (i) la sous-section *Eulycopersicon* avec des espèces avec des fruits contenant des caroténoïdes à maturité

et (ii) la sous-section *Eriopersicon* regroupant les autres espèces apparentées à la tomate dont les fruits restent verts à maturité (Peralta and Spooner 2005). Cette subdivision de la section *Lycopersicon* a été revue récemment et on classe les dix espèces sauvages et cultivées en séries *Lycopersicon*, *Eriopersicon* et *Neolycopersicon* (Peralta and Spooner 2005). La série *Neolycopersicon* ne contient que *S. pennellii*. La dernière classification taxonomique établie reconnaît 13 espèces différentes. *S. peruvianum* se subdivise en *S. arcanum, S. huaylasense, S. peruvianum* et *S. corneliomulleri*. *S. cheesmaniae* se subdivise en *S. galapagense* et *S. cheesmaniae* (Peralta and Spooner 2005; Peralta, Spooner et al. 2007). Le Tableau 1-4 donne quelques caractéristiques biologiques de chaque espèce (Spooner, Peralta et al. 2005).

1.2.6. Domestication

Le processus de domestication implique qu'une espèce acquière des caractéristiques nouvelles par rapport à un ancêtre sauvage, qui apportent un avantage dans l'utilisation de l'espèce par les communautés humaines (Diamond 2002). Le syndrome de domestication, chez les plantes, réunit ces caractères sélectionnés par les communautés humaines et différenciant les espèces sauvages des espèces cultivées. Le syndrome de domestication inclut généralement un port de la plante plus compact, une précocité accrue, une réduction ou une perte de la dormance et de la dispersion des graines, un gigantisme et une augmentation de la diversité des parties consommées (Frary and Doganlar 2003). Chez les plantes où le fruit est la partie consommée, les caractères influant sur la morphologie du fruit sont les principaux critères qui ont évolué avec la domestication. Chez la tomate, la domestication a entraîné une augmentation de la diversité de la forme et de la couleur de fruits, une amélioration de la saveur avec une augmentation du contenu en sucres et acides, une augmentation de la taille des fruits voire un gigantisme (Bai and Lindhout 2007). L'homme a aussi sélectionné la capacité de la tomate à donner des plantes identiques lors du passage aux générations suivantes, en fixant l'autogamie. Chez la tomate, bien que les espèces apparentées sauvages soient maintenant bien connues, le processus de domestication reste encore flou. Jenkins (1948) décrit deux hypothèses quant à l'origine de la tomate cultivée. La première, soutenue notamment par de Candolle (1778-1841), suppose que la tomate a été transportée vers l'Europe à partir du Pérou quelques temps après la découverte de cette région. La seconde hypothèse soutien une origine mexicaine de la tomate.

Dans *Les Origines des Plantes Cultivées* (1882), de Candolle écrit que la reconstruction de l'histoire des plantes cultivées relève de la probabilité et il applique, pour la

première fois, une méthode d'évaluation des preuves disponibles se rapportant à chaque plante. Ces preuves peuvent être classées dans les thèmes suivants : la botanique (distribution géographique des espèces du même genre), la diversité génétique, l'archéologie, l'histoire et la linguistique.

1.2.6.1. Botanique

Nikolaï Ivanovitch Vavilov (1887-1943) étudie l'origine et la distribution géographique des plantes cultivées. Il est le premier à décrire les « centres de diversité » des plantes cultivées : zones géographiques où les espèces végétales cultivées développent pour la première fois leurs caractères distinctifs. Vavilov estimait que les premières étapes, dans la domestication d'une plante, prenaient place au sein du centre primaire. Le transport hors de ces centres est dû aux migrations humaines. Le fait que la plupart des espèces sauvages de tomates soient endémiques de l'Amérique du Sud, et restreintes à une zone limitée, favorise l'hypothèse d'une domestication dans cette région. Vavilov admet cependant que la domestication peut avoir lieu loin du centre de diversité primaire. Ce schéma semble caractéristique des plantes domestiquées tardivement, pour lesquelles on observe une expansion graduelle des espèces sauvages, dans des zones perturbées par l'activité humaine (Jenkins 1948). La participation de l'homme dans l'expansion de l'espèce serait donc inconsciente. Jenkins (1948) souligne qu'il n'est pas question de douter du centre d'origine du genre *Solanum* sect. *Lycopersicon* mais qu'il n'est pas évident que le centre de diversité de la tomate cultivée (zone vraisemblable de domestication) soit identique au centre de diversité des espèces apparentées sauvages. Pour cet auteur, il semble évident que la région de Mexico est le centre d'origine de l'espèce cultivée. Il s'appuie sur la diversité morphologique des accessions rencontrées là-bas, la présence d'accessions de type cerise (adventices et cultivées) et la présence de formes transitoires entre les formes sauvages et cultivées. Sachant que *S. pimpinellifolium*, trouvé entre le Pérou et l'Equateur, est l'espèce sauvage la plus proche de l'espèce cultivée, les accessions de tomates introduites au Mexique devaient avoir subi les premiers stades de la domestication. En effet, il serait étonnant que de petites baies rouges parcourent 5000 Km (des régions andines au Mexique) sans recevoir la moindre attention de la part des humains qui les transportent (Brücher 1989).

1.2.6.2. Diversité Génétique

Rick (1976) émet l'hypothèse que la domestication s'est réalisée à partir d'accessions sauvages *Solanum lycopersicum var. cerasiforme* (*S. l. cerasiforme*), qui aurait évolué directement à partir des espèces sauvages auto-incompatibles (Figure 1-12).

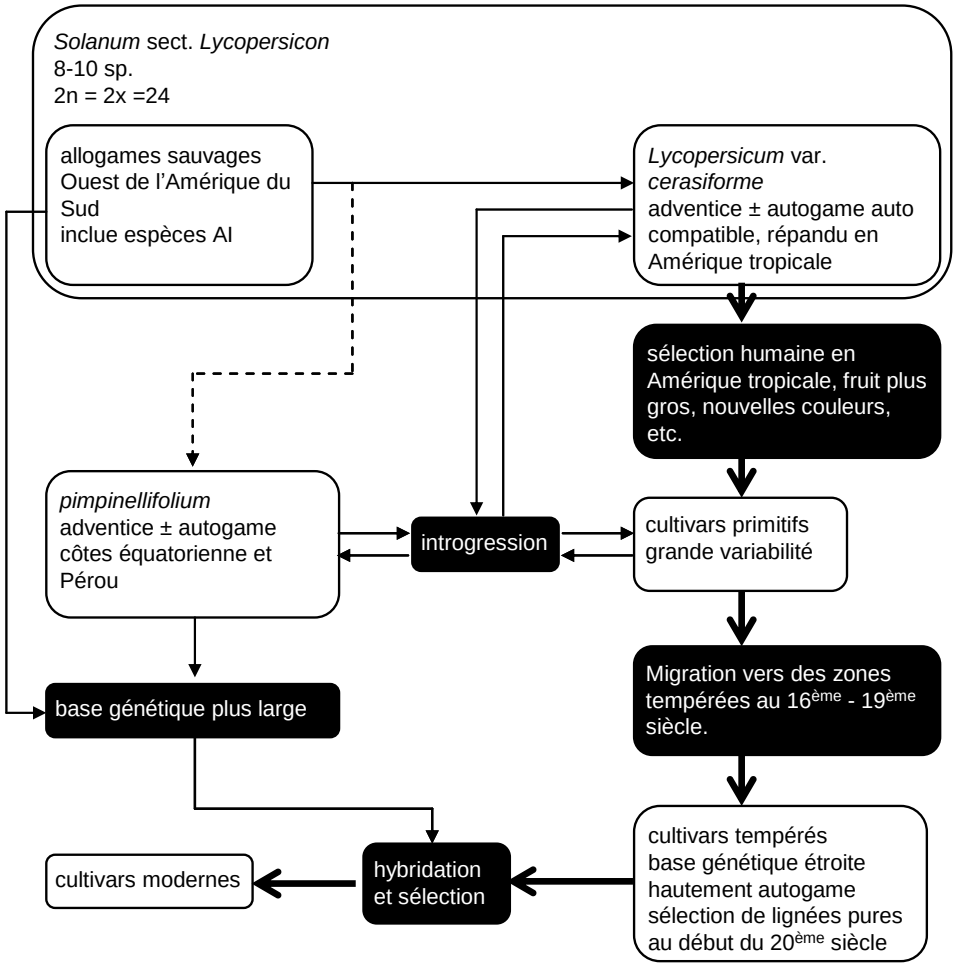


Figure 1-12 Evolution de la tomate cultivée *Solanum. Lycopersicum*. Traduit de Rick (1976)

L'auteur suggère même que l'espèce *S. pimpinellifolium* serait issue d'une branche parallèle lors de l'évolution de la tomate. Cependant, les travaux de phylogénie sur des caractères morphologiques et sur les marqueurs moléculaires montrent une affiliation directe entre *S. pimpinellifolium*, *S. l. cerasiforme* et *S. l. esculentum*. Aujourd'hui, l'espèce *S. l. cerasiforme* est considérée comme un type primitif de tomate cultivée ou comme une forme transitoire entre *S. l. cerasiforme* sauvage et *S. l. esculentum* cultivé. Il semblerait aussi que de nombreuses accessions soient d'origine férale (Rick and Holle 1990; Peralta and Spooner 2007).

L'aire de répartition de *S. l. cerasiforme* s'étend dans toutes les régions tropicales et sub-tropicales du monde. Cette expansion post-colombienne gomme les traces de la diffusion historique de cette espèce pendant la domestication. La nature adventice de *S. l. cerasiforme* peut expliquer pourquoi cette espèce est si souvent associée avec l'homme. Rick et Fobes (1975) utilisent des isozymes pour analyser la diversité génétique chez l'espèce cultivée de tomate ainsi que chez les espèces proches. Ils mettent en évidence l'uniformité génétique des cultivars de tomate par rapport à la tomate de type cerise. Ils soulignent aussi que, pour *S. l. cerasiforme* et *S. l. esculentum*, une variabilité génétique supérieure est trouvée pour des accessions prospectées au Pérou ou en Equateur par rapport aux accessions provenant du Mexique. Ils montrent également, que dans ces régions où la sympatrie avec l'espèce *S. pimpinellifolium* est courante, les accessions de *S. l. cerasiforme* et de *S. l. esculentum* présentent des allèles issus de l'espèce *S. pimpinellifolium*. Ces allèles communs témoignent d'évènements d'intercroisements récents plutôt que de parenté potentielle (Miller and Tanksley 1990). Rick et Holle (1990) montrent, sur de nouvelles accessions de *S. l. cerasiforme* prospectées en zone andine, qu'elles sont plus polymorphes que les accessions extra-andines. Ces accessions, prospectées en dehors du centre de diversité de la tomate sauvage, sont monomorphes et possèdent le même zymotype commun à la majorité des accessions de *S. l. esculentum*. Ces données soutiennent, mais ne prouvent aucunement, que *S. l. esculentum* aurait été domestiqué à partir de populations andines de *S. l. cerasiforme*. La tomate de type cerise étant morphologiquement intermédiaire entre *S. pimpinellifolium* sauvage et *S. l. esculentum* cultivé, on peut alors imaginer que la domestication s'est faite dans le sens suivant :

S. pimpinellifolium → *S. l. cerasiforme* → *S. l. esculentum*. Néanmoins, des inter-croisements récurrents entre l'espèce sauvage et les variétés botaniques ont eu lieu. D'autres auteurs pensent que le chaînon manquant, entre les espèces sauvages et la tomate cultivée, serait représenté par une espèce poussant naturellement au Nord de l'Equateur, dans un environnement humide. Ils décrivent une nouvelle espèce : *Lycopersicon humboldtii*, identifiée seulement par deux fois et ressemblant fortement à *S. l. cerasiforme* (Brücher 1989).

1.2.6.3. Archéologie – Histoire

Aucune donnée archéologique n'a pu être rattachée de près ou de loin à la culture de la tomate. Aucun reste de plante de tomate n'a été retrouvé dans les fouilles archéologiques du Pérou, de l'Equateur, ni même du Mexique. L'unique représentation potentielle retrouvée en

Colombie (McMeekin 1992), serait une bobine pour filer la laine décorée d'une fleur de tomate, qui daterait de 900 à 500 ans av. J.C. (Figure 1-13).

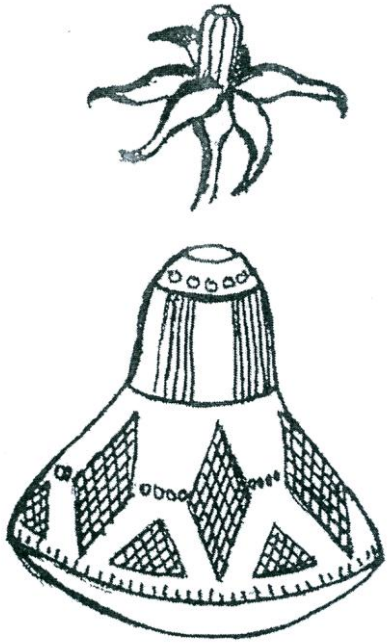


Figure 1-13. Bobine à filer colombienne (500-900 BC) reproduisant une fleur de tomate. D'après McMeekin (1992).

L'interprétation de cette représentation peut être contestée (Daunay, Laterrot et al. 2007). Alors qu'aucune trace relative à la culture de la tomate n'est disponible en Amérique du Sud, on retrouve des informations dans des écrits contemporains à la conquête du Mexique par les Espagnols. Bernardino de Sahagún recueille les témoignages d'Aztèques ayant connu la société indigène avant la conquête espagnole. Il retranscrit ses récits dans *Histoire générale des choses de la Nouvelle-Espagne*. La tomate y est décrite comme un aliment commun chez les Aztèques. Elle est notamment utilisée comme plante médicinale et elle semble répandue, à la fois dans les offrandes religieuses et sur les marchés (Daunay, Laterrot et al. 2007). Les premières descriptions de tomate apparaissent dans les écrits des naturalistes du XVI^{ème} siècle mais celles-ci sont brèves car elles témoignent de l'introduction récente de la tomate en Europe. Matthiolus (1544) décrit à la fois l'aubergine et la tomate dans la première édition des *Dioscorides*. La définition qu'il donne en 1554 de la tomate a été traduite en français en 1572 par Guillaume Rouillé (texte traduit en français moderne) : « De plus, il n'y a pas longtemps qu'on a commencé à voir une autre sorte d'aubergine plate, rondes comme des pommes, divisée en côtes comme des pompons, d'abord vertes, puis étant mûres, jaunes comme l'or sur

quelques plantes et rouges sur d'autres. On les appelle vulgairement *Pomi d'oro*, pommes d'or [actuel nom commun de la tomate en Italie]. On les mange comme les sudistes [frites dans l'huile] : mais elles donnent envie de vomir et font souvent vomir. ». Il est intéressant de voir que la première tomate décrite était fasciée et de couleur jaune ce qui lui a valu le nom de « *Pomi d'oro* ». Il semble donc que les premières tomates, introduites en Europe depuis Mexico, avaient déjà subi un stade avancé dans la domestication avec l'apparition de fruit de taille moyenne, du caractère fascié et de la couleur jaune. La Figure 1-14 montre la première image publiée de tomate réalisée par Dodoens en 1553. D'autres images, plus précises seront publiées dans divers guides d'herboristes mais toutes montrent des fruits fasciés.

A B

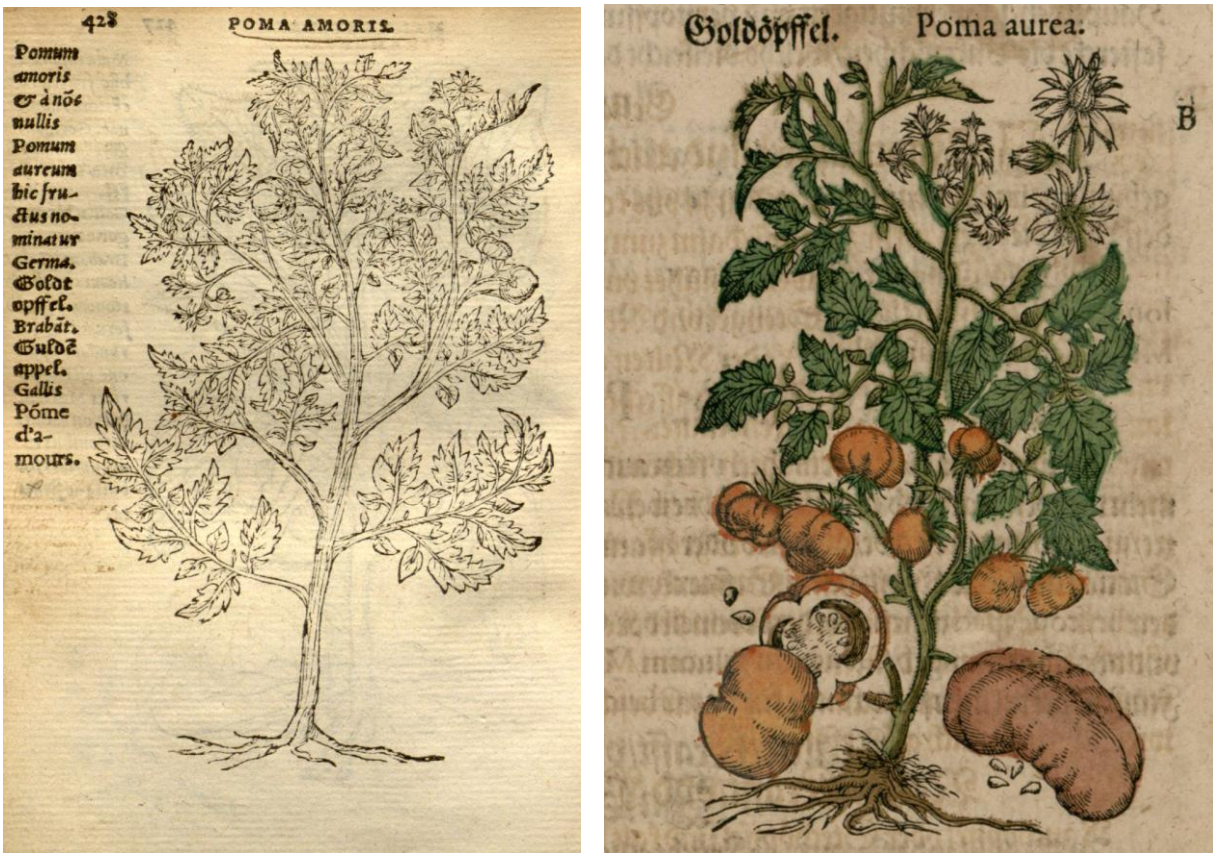


Figure 1.14. Premières images de tomate publiées.
(A) Image publiée par Dodoens en 1553. Tiré de Daunay et al. (2007).
(B) Planche de tomate dessinée par Mattioli en 1590, dans une édition en allemand et en couleur des *Commentaires sur Dioscorides*. *Service de la documentation de l'Université de Strasbourg*. 24 nov. 2009.
<http://imgbase-scd-ulp.u-strasbg.fr/displayimage.php?album=28&pos=772>

1.2.6.4. Linguistique

Le mot tomate proviendrait du mot espagnol « tomate » (prononcer tomaté), lui même issu du suffixe *tomatl*, utilisé en langage nahualt (langue parlée en Amérique Centrale et toujours utilisée au Mexique). Le mot *tomatl* est souvent accompagné d'un préfixe qui précise l'identification de la plante décrite. On retrouve par exemple les mots *xalttomatl* ou *xitomatl*. D'après le *Dictionnaire de la langue nahuatl ou mexicaine* de Rémi Siméon (1885) (http://sites.estvideo.net/malinal/nahuatl.page.html), le mot *tomatl* traduit la tomate actuelle ou la tomatille qui est l'espèce *Physalis sp*., une autre *Solanaceae*. L'espèce décrite forme des fruits plus petits que ceux connus communément en Europe. En nahuatl, le préfixe *toma* indiquent « faire sortir de prison » ou « ouvrir une chose ». Il semblerait donc que le mot *tomatl* désigne principalement le *Physalis* dont le fruit est enfermé dans une capsule et, était consommé par les Aztèques. Le mot *xitomatl* semble désigner la tomate telle qu'on la connaît en Europe et ce mot est utilisé par Tezozomoc (1531) dans *Cronica mexicayotl* pour décrire la tomate cultivée. Le mot *xalttomatl* désigne la « tomate des sables » (*Saracha jaltomata*). Le mot *miltomatl* est relatif à une espèce de petite tomate verte et douce employée à des fins médicinales. De nos jours, a Mexico, les tomates cultivées sont toujours nommées « xitomatl », « gitomate » ou « jitomate », les deux derniers provenant sans doute du premier. Les peuples Maya (présents plus au Sud) utilisent une étymologie différente (avec le préfixe « p'aak ») et n'ont jamais adopté les noms nahualt. Le fait que des noms si différents soient utilisés dans différentes régions ethniques semble appuyer une introduction ancienne de la tomate en Amérique Centrale. Il est aussi assez hasardeux de conclure sur une origine de la tomate en s'appuyant sur les noms donnés à cette plante par les européens. Jenkins (1948) montre qu'il est peu évident de se fier au premiers noms latins donnés à la tomate : *Pomi del Peru* et *Mala peruviana*. En effet, ces noms étaient utilisés auparavant en France pour nommer le datura (*Datura stramonium*). Il se peut aussi qu'il y ait confusion sur le terme *Lycopersicon*, utilisé par Galien au IIe siècle alors que la tomate n'était pas présente sur le nouveau monde. Galien utilisait ce terme pour désigner le datura ou l'aubergine.

1.2.6.5. Usage

La tomate était donc utilisée dans la vie courante par les Aztèques. Elle était utilisée comme plante médicinale et elle servait sans doute de condiment avec le piment dans la réalisation de sauces. L'absence de représentation de la tomate, laissée par les différentes civilisations précolombiennes, traduit le fait que cette plante ne devait pas être d'un intérêt majeur dans leur alimentation. C'est après la conquête de Mexico qui débute en 1529 que la

tomate est ramenée par les conquistadors vers le Vieux Monde. La tomate a d'abord été conservée dans les jardins des herboristes avant d'être disséminée dans tout le pourtour méditerranéen où elle reste peu consommée. Sa proximité botanique avec la mandragore lui vaut d'être utilisée comme une plante médicinale. On lui soupçonne même des propriétés aphrodisiaques. Certains auteurs commencent à s'interroger sur les qualités nutritionnelles qu'apporterait la tomate au corps alors que d'autres continuent à penser qu'il s'agit d'un poison (Daunay, Laterrot et al. 2007). L'expansion de la tomate progresse vers le Nord de l'Europe mais le fruit n'est consommé que dans les pays du pourtour méditerranéen. De plus en plus de types différents apparaissent : en 1623, on connaît quatre couleurs de fruit : rouge, jaune, orange et doré (Cox 2000). Des livres de recettes mentionnent l'utilisation de la tomate dans la réalisation de sauces, de soupes ou bien dans des recettes où la tomate est frite dans de l'huile ou consommée crue avec de l'huile et du vinaigre (Daunay, Laterrot et al. 2007). Cette espèce continue à être considérée comme vénéneuse notamment en Grande Bretagne. Ce n'est qu'à la fin du XVIIIe siècle que les premières variétés potagères apparaissent (Doré and Varoquaux 2006). Parallèlement, la tomate est amenée sur le Nouveau Continent par les colons britanniques mais celle-ci était encore utilisée comme plante ornementale. La consommation du fruit de tomate s'étend lentement aux Etats-Unis. Certains commerçants de semences incluent alors quelques variétés (au sens agronomique du terme) dans leurs catalogues. La consommation et la production de ce fruit connaissent ensuite une popularité sans précédent et les sociétés semencières sélectionnent certains cultivars pour des caractères d'intérêt agronomique dès la fin du XVIIIe siècle. La nature autogame de cette espèce permet aux générations successives de ressembler aux générations précédentes. Les cultivars, possédés par certaines familles de cultivateurs, sont transmis aux générations suivantes ce qui leur vaut le nom de « Heirloom » (traduction anglaise de « héritage ») (Bai and Lindhout 2007). Ces mêmes familles sélectionnent alors des mutations ponctuelles apportant au fruit ou à la plante, des caractéristiques nouvelles, comme par exemple de nouvelles couleurs du fruit (noir, blanc, violet foncé, orange avec des rayures vertes, etc.), de nouvelles formes (forme de piment), mais surtout une augmentation de la taille de celui-ci. Les vagues d'immigration récurrentes vers les Etats-Unis complètent la diversité des cultivars déjà présents là-bas. Comme avec n'importe quelle plante autogame, le croisement entre lignées pures pour donner un hybride F1 entraîne un gain dans la production grâce à l'effet d'hétérosis ou vigueur hybride. Les avantages apportés par les hybrides ont conquis les cultivateurs qui étaient prêts à payer le prix fort pour un lot de semences malgré le fait que la propagation de ces plantes était impossible pour eux (ségrégation des caractères à la génération F2). A partir des années

70, les sociétés semencières se lancent dans des programmes de sélection ambitieux et travaillent particulièrement sur la résistance aux stress biotiques et abiotiques (locus introgressés à partir d'espèces sauvages, Tableau 1-4), le rendement, l'adaptation aux conditions de culture et plus récemment sur la qualité gustative et nutritionnelle du fruit.

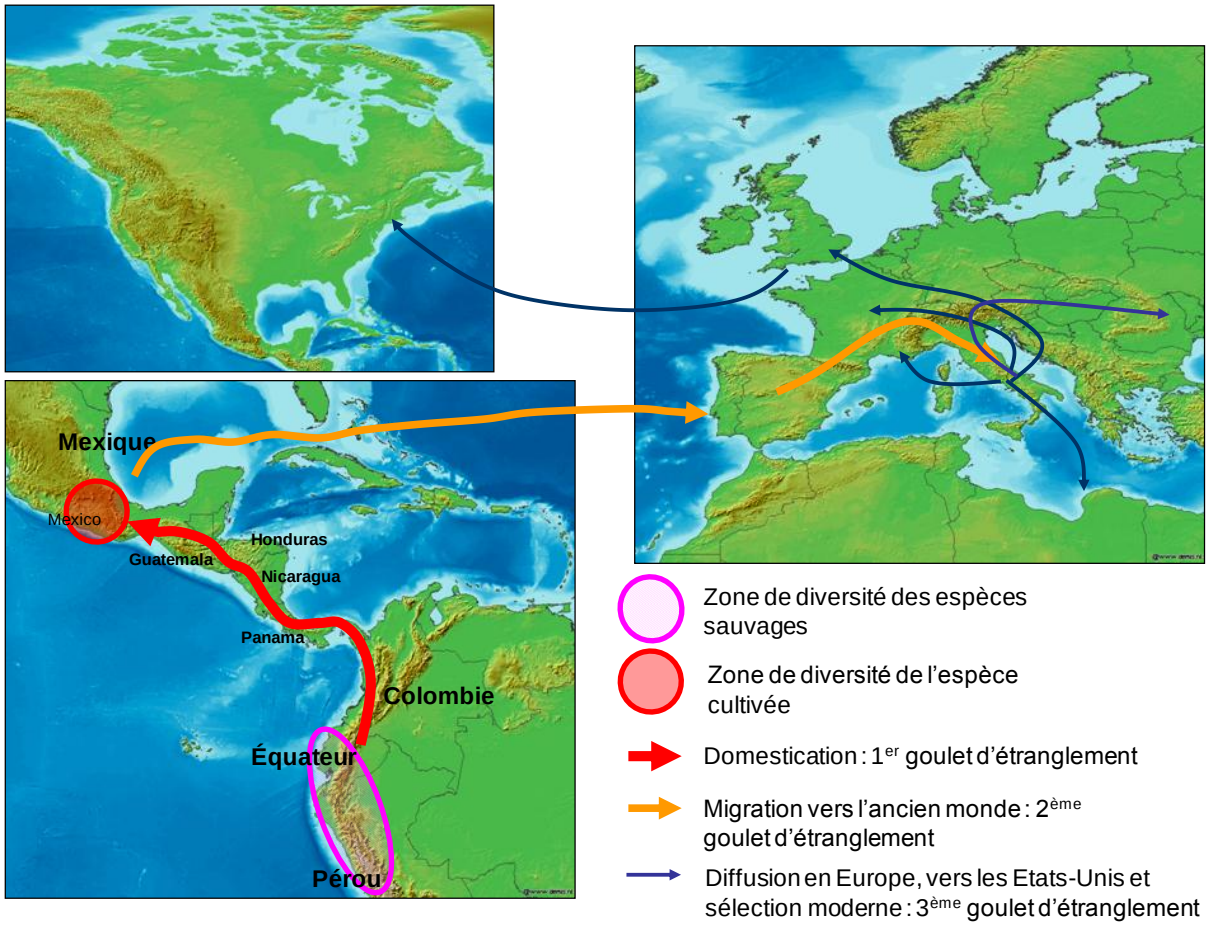


Figure 1-15. Histoire hypothétique de la domestication de la tomate.
La répartition géographique des espèces sauvages proches est limitée aux côtes andines du Pérou et de l'Equateur. C'est à Mexico que l'on retrouve la plus grande diversité phénotypique pour la tomate cultivée. Seul un faible nombre d'accessions a dû être rapporté en Europe au 16ème siècle et ce faible nombre d'accessions est à l'origine de du « pool » de tomate cultivée moderne.
Les fonds de carte proviennent du DEMIS World Map Server.
<http://www2.demis.nl/mapserver/mapper.asp>

En conclusion, on s'aperçoit que malgré l'importance qu'a pris la tomate dans notre alimentation depuis son arrivée en Europe (et ensuite aux Etats-Unis), peu de choses sont connues sur son histoire ancienne. Il semblerait que lorsque la tomate est arrivée à la Cour Espagnole, elle avait été domestiquée et qu'elle était déjà utilisée par les Aztèques. Son histoire à travers les différents continents a entraîné une réduction considérable de la diversité génétique malgré l'explosion de la variation morphologique qu'on lui connaît. Cette réduction

de la diversité est due à trois goulets d'étranglement que la tomate a subit : (i) la domestication qui a sans doute débuté inconsciemment dans les vallées andines jusqu'au Mexique, (ii) le passage de quelques cultivars en Europe et (iii) l'amélioration moderne en Europe et aux Etats-Unis qui a encore diminué le « pool » de géniteurs utilisés dans les programmes de sélection (Figure 1-15).

1.3. La qualité du fruit chez la tomate

Plusieurs définitions peuvent être données au terme « qualité » :

- Ensemble des caractéristiques d'une entité qui lui confère l'aptitude à satisfaire des besoins exprimés et implicites. (ISO 9000 1994)
- Aptitude d'un ensemble de caractéristiques intrinsèques à satisfaire des exigences. (ISO 9000 2000)

La qualité d'une production végétale sera perçue différemment selon qu'on s'adresse au producteur, au transformateur (production industrielle), au grossiste (production en frais) ou au consommateur. Pour le producteur, la qualité d'une production sera liée au rendement, à l'homogénéité de la récolte ainsi qu'à l'adaptation au système cultural (résistance aux pathogènes et faible besoin en intrants). Le grossiste s'intéressera plutôt à des critères comme la résistance de la production au conditionnement et au transport, l'homogénéité de la production et les caractéristiques visuelles du produit. Pour l'industriel, ce sont les capacités à la transformation mais aussi les composantes physico-chimiques du produit qui détermineront la qualité finale du produit transformé. Le consommateur qui est le dernier maillon de la chaîne de distribution sera plus sensible à l'aspect visuel du produit ainsi qu'à sa valeur santé lors de l'achat. Lors de la consommation du produit, c'est la qualité organoleptique qui est retenue.

La qualité chez la tomate est donc un critère multi-composite qui doit être pris en compte à tous les niveaux d'utilisation lors de la sélection variétale. La qualité organoleptique du fruit, plus particulièrement, est plus difficile à caractériser ce qui complique la sélection pour ce caractère. Elle rassemble l'ensemble des critères liés à l'apparence, la flaveur et la texture du produit. La flaveur est une combinaison des saveurs et des arômes et elle se caractérise par les composantes chimiques associées à la teneur en sucres, en acides et à la composition en arômes. La composition d'un fruit de tomate à maturité est la suivante : 95 % d'eau, 5% de matière sèche comprenant entre autre 50% de sucres, 25% d'acides organiques,

8% de minéraux, 2% d'acides aminés, de caroténoïdes et autres métabolites secondaires (Davies and Hobson 1981). La perception sucrée provient majoritairement du fructose qui possède un pouvoir sucrant supérieur au glucose (le saccharose étant présent en faible quantité à maturité). La perception acide provient majoritairement de l'acide citrique.

Certains composés volatils ont été corrélés avec des arômes détectés par un panel de dégustation (Baldwin, Goodner et al. 2004). Il semble aussi qu'il existe une grande variabilité au sein des consommateurs quant à leurs préférences (Causse, Buret et al. 2003). Ainsi on peut réaliser une cartographie des consommateurs européens en fonction de leurs préférences par rapport à la taille du fruit, la fermeté, la douceur et l'acidité (Causse, Friguet et al. *in prep.*).

La sélection moderne a consacré beaucoup d'efforts à l'amélioration des résistances aux stress biotiques et abiotiques ainsi qu'à l'augmentation du rendement, notamment en exploitant l'effet d'hétérosis (expliqué par les effets additifs des gènes et par la superdominance) (Bai and Lindhout 2007). Beaucoup de gènes ont été introgressés dans les lignées parentales à partir d'accessions sauvages (Rick and Chetelat 1995). Ces inrogression ont introduit par la même occasion des gènes potentiellement défavorables pour la qualité organoleptique. L'utilisation des mutants naturels *nor* (non ripening) et *rin* (ripening inhibitor) a permis d'augmenter la capacité au stockage et donc de rallonger les circuits de distribution. Ceci, lié à des récoltes plus précoces lors de la maturation du fruit pour augmenter encore la fermeté, participe à une diminution de sa qualité organoleptique. Les conditions de culture, le stade de récolte et les conditions de conservation sont autant de variables environnementales qui ont un impact fort sur la qualité organoleptique de la tomate (Baldwin, Nisperos-Carriedo et al. 2002; Aguayo, Escalona et al. 2004).

Un autre phénomène responsable de la diminution de la qualité est l'antagonisme qui existe entre les différents caractères. Par exemple, plusieurs études de QTL ont révélé une corrélation négative entre le poids du fruit et le contenu en sucres et en acides (Lecomte, Saliba-Colombani et al. 2004; Prudent 2008). Ce phénomène peut être expliqué par des variations de l'équilibre entre les entrées d'eau et d'assimilats dans le fruit (Guichard, Bertin et al. 2001). Il a aussi été démontré une instabilité de l'expression de certains QTL (liés à la fermeté du fruit) ce qui peut rendre difficile leur caractérisation (Chaïb, Lecomte et al. 2006) . Il existe aussi de nombreux phénomènes d'épistasie entre les QTL (Eshed and Zamir 1996)

ainsi que des interactions entre certains QTL et le fond génétique (Lecomte, Duffé et al. 2004).

De par la complexité inhérente à l'étude de la qualité du fruit, l'identification des bases moléculaires des composantes de la qualité, est d'un intérêt majeur pour mieux comprendre les relations entre la génétique du caractère et l'effet de l'environnement.

1.3.1. Poids du fruit

Un des caractères de qualité les plus étudiés chez la tomate est sans aucun doute le poids du fruit. Des QTL de la variation du poids du fruit ont été détectés sur toutes les populations de cartographie existantes. Toutes les espèces sauvages possèdent des fruits de 10 à 15 mm de diamètre alors que des fruits de plus gros calibre (diamètre supérieur à 3 cm) sont trouvés uniquement chez l'espèce cultivée. La variation du caractère est telle qu'elle peut être analysée dans toutes les populations en ségrégation et tous les chromosomes semblent porter des QTL de poids frais (Paterson, DeVerna et al. 1990; Paterson, Damon et al. 1991; Eshed and Zamir 1995; Fulton, Beck-Bunn et al. 1997; Bernacchi, Beck-Bunn et al. 1998; Chen, Foolad et al. 1999; Lippman and Tanksley 2001; Saliba-Colombani, Causse et al. 2001; Prudent, Causse et al. 2009).

Quatre QTL majeurs sont responsables d'une part importante de la variation du poids du fruit : FW1.1, FW2.2, FW3.1 et FW4.1, mais seul FW2.2 a pour l'instant été cloné (Tanksley 2004). Celui-ci se situe en position basale du chromosome 2 et il a été cartographié dans pratiquement toutes les populations. Ce QTL explique 10 à 50% de la variation phénotypique (Lippman and Tanksley 2001). Frary et al. (2000) ont identifié le gène sous-jacent au QTL. Il s'agit d'*ORFX* qui code pour une protéine similaire à un oncogène humain. *ORFX* est surexprimé dans les carpelles au stade pré-anthèse chez les lignées recombinantes portant des petits fruits. Cette surexpression de l'*ORFX* entraîne une réduction de la durée de la division cellulaire dans les carpelles. Une étude plus approfondie sur l'expression de l'*ORFX* montre que la différence est due à un décalage du pic d'expression d'une semaine. Ce décalage est associé à des changements concomitants dans l'activité mitotique pendant les stades précoces du développement du fruit (Cong, Liu et al. 2002). FW2.2 n'a pas d'effet démontré sur le rendement ni sur la forme du fruit. Un autre QTL a été identifié comme responsable d'une part importante de la variation du poids du fruit mais il est aussi impliqué dans l'organisation structurale du fruit et donc de sa forme. Le QTL FW11.3 co-localise avec

le QTL *fasciated*, qui est responsable du phénotype fascié. Ce phénotype est observé chez un grand nombre de variétés anciennes (Figure 1-16).

Figure 1-16. Différence extrême du poids du fruit entre deux espèces. D'après Tanksley (2004).
Un fruit issu de l'espèce sauvage *S. pimpinellifolium* est présenté à gauche. Un fruit du cultivar Giant Red (*S. lycopersicum*) est représenté à droite. Ce cultivar possède à la fois les allèles « gros-fruit » pour *fw2.2* et pour *fas*.

Il s'agit en fait de fruits possédant un grand nombre de loges (plus de 6 loges). C'est une modification de l'expression d'un facteur de transcription de type YABBY qui entraîne la différence phénotypique (Cong, Barrero et al. 2008). Ces facteurs de transcription avaient déjà été décrits comme contrôlant le nombre d'organes chez d'autres espèces. Ces deux QTL étaient majeurs et il est maintenant nécessaire d'identifier les autres QTL ayant un effet plus faible car il semble que ce soit la combinaison allélique à plusieurs locus qui soit responsable du passage de petites baies sauvages à des fruits de calibre supérieur.

1.3.2. Forme du fruit

Les cultivars de tomates produisent majoritairement des fruits ronds mais de nombreuses variantes sont connues. Comme nous l'avons déjà vu les fruits peuvent être aplatis ou allongés jusqu'à des formes pyriformes ou ressemblant très fortement à des poivrons. Les QTL liés à la morphologie se répartissent eux aussi sur plusieurs chromosomes (van der Knaap and Tanksley 2003; Gonzalo and van der Knaap 2008). Le QTL *fas* est aussi impliqué dans la variation morphologique car l'augmentation du nombre de loges tend à aplatir les fruits. Un autre QTL modifiant le nombre de loges, LCN2.1, a été cartographié sur le chromosome 2 mais son effet est beaucoup moins fort que celui de *fas* (3 à 4 loges pour LCN2.1 et jusqu'à plus de 15 pour *fas*). Quand les allèles avantageux de ces deux locus sont présents en même temps, ils interagissent de façon épistatique pour produire des fruits avec un

nombre encore plus élevé de loges (Barrero and Tanksley 2004). Les QTL responsables de la variation du nombre de loges ont un effet sur le poids du fruit et sur la forme générale.

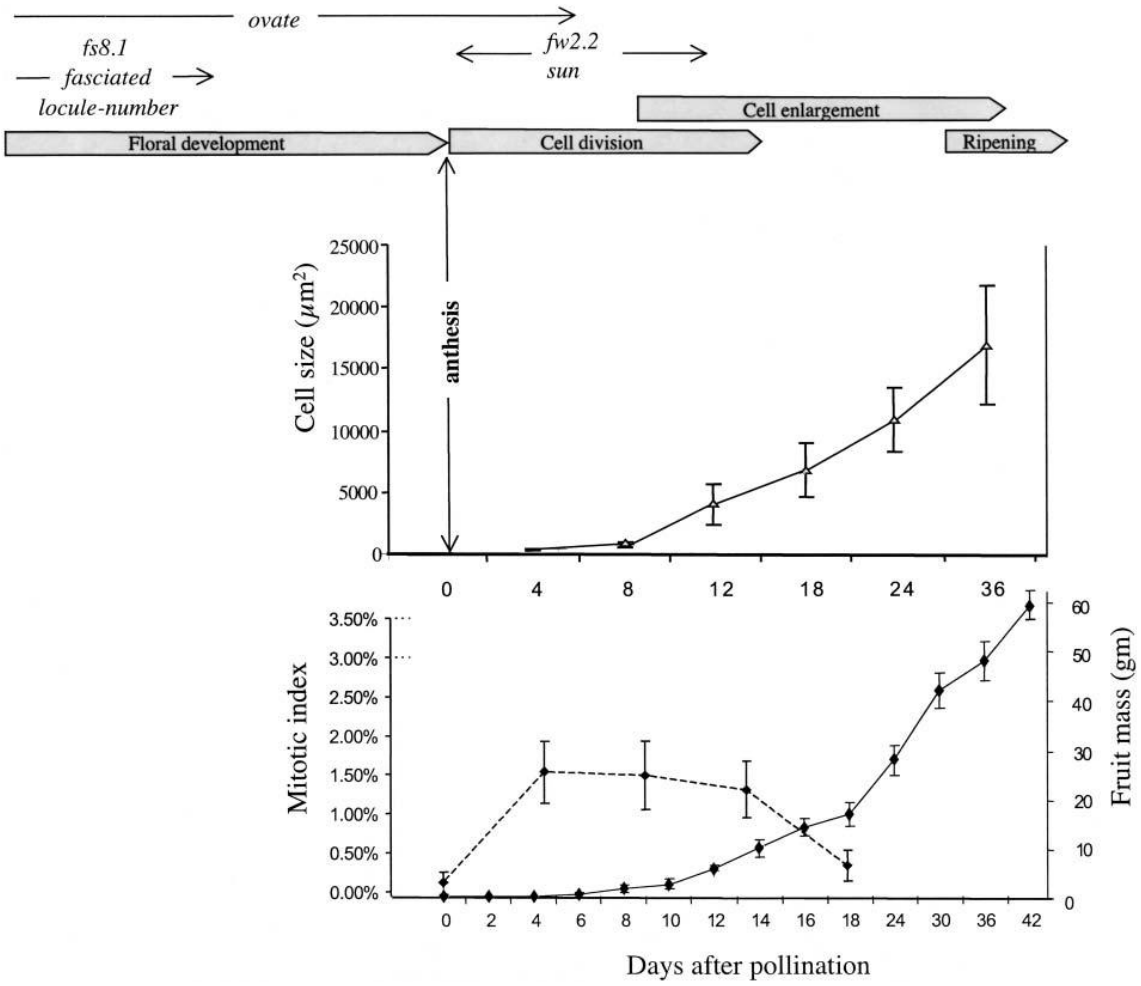


Figure 1-17. Etapes du développement du fruit de tomate. D'après Tanksley (2004).
Le déroulement de l'expression des gènes qui affecte le poids et la forme du fruit est indiqué en haut de la figure. D'autres gènes affectent le poids et la forme du fruit mais le déroulement de leur activité sur le contrôle du développement n'est pas connu.

Si on ne s'intéresse qu'à la forme du fruit, trois QTL nommés *ovate*, *sun* et FS8.1 sont responsables d'une part importante de la variation du ratio longueur du fruit *versus* largeur du fruit (60% d'effet pour Sun et 40% d'effet pour Ovate dans deux populations différentes). Le QTL *ovate* a été cloné et le gène responsable de la variation de l'indice longueur/largeur correspond à une nouvelle classe de protéine régulatrice présente dans le noyau (Liu, Van Eck et al. 2002). Ce gène est exprimé précocement dans le développement du fruit (moins de 10 jours après pollinisation) et jusqu'à deux semaines après anthèse. La mutation responsable du caractère allongé du fruit induit un codon stop dans le deuxième exon du gène. Bien que les locus *ovate* et *sun* causent une forme allongée du fruit, il existe des différences

morphologique et au niveau du développement. Le QTL *ovate* crée une élongation asymétrique telle que, l'élongation du haut du fruit est plus importante que celle du bas (forme pyriforme), alors que l'effet de *sun* est uniforme (forme ovoïde). L'effet de *sun* est aussi plus tardif que celui d'*ovate* et n'a lieu que durant la phase de division cellulaire. Le QTL FS8.1 induit une forme « carré » du fruit. Ce caractère a été utilisé en amélioration variétale des cultivars destinés à la production industrielle et a notamment permis d'augmenter la résistance mécanique du fruit pendant la récolte mécanique. Des études sur le développement utilisant des lignées quasi-isogéniques ont montré que les changements dans la forme du fruit induits par FS8.1 commençaient très tôt au cours du développement floral et carpellaire (Ku, Grandillo et al. 2000). La Figure 1-17 montre les différentes étapes du développement du fruit avec l'expression des différents gènes décrits ci-dessus.

1.3.3. Couleur du fruit

La majeure partie des variations dans la couleur du fruit est due à des mutations dans les enzymes de la voie de biosynthèse des caroténoïdes. La couleur du fruit est la résultante des couleurs de la chair et de l'épiderme. Ces mutations sont responsables de phénotypes relativement faciles à lire et qui accélèrent l'identification des gènes sous-jacents (Paran and van der Knaap 2007). Les protéines ayant un rôle dans la biosynthèse des caroténoïdes peuvent être prédites par des études biochimiques. Ces gènes sont donc identifiés en construisant des mutants où l'expression de certains gènes est éteinte. Les fruits verts de tomate contiennent de la chlorophylle et des caroténoïdes comme la lutéine. Chez les accessions de la série *Eulycopersicon*, lors de la maturation, les chloroplastes se transforment en chromoplastes donnant la couleur rouge du fruit mûr. La voie de biosynthèse des caroténoïdes ainsi que les mutations impliquées dans la couleur du fruit sont indiquées dans la Figure 1-18.

La couleur jaune du fruit est contrôlée par le locus *R* (chair jaune). L'allèle récessif *r* possède une mutation dans le gène *PSY1* qui code pour une phytoène synthase 1. Cette mutation induit une protéine tronquée incapable de convertir le geranylgeranyl diphosphate en phytoène (Ray, Moureau et al. 1992; Fray and Grierson 1993). Plusieurs locus sont responsables de la couleur orangée du fruit. Le locus *delta* correspond à un gène codant pour la lycopène δ-cyclase (Ronen, Carmel-Goren et al. 2000). L'allèle donnant la couleur orange provient de *S. pennellii* et confère une surexpression de la lycopène δ-cyclase et donc une accumulation de δ-carotène plutôt que de lycopène.

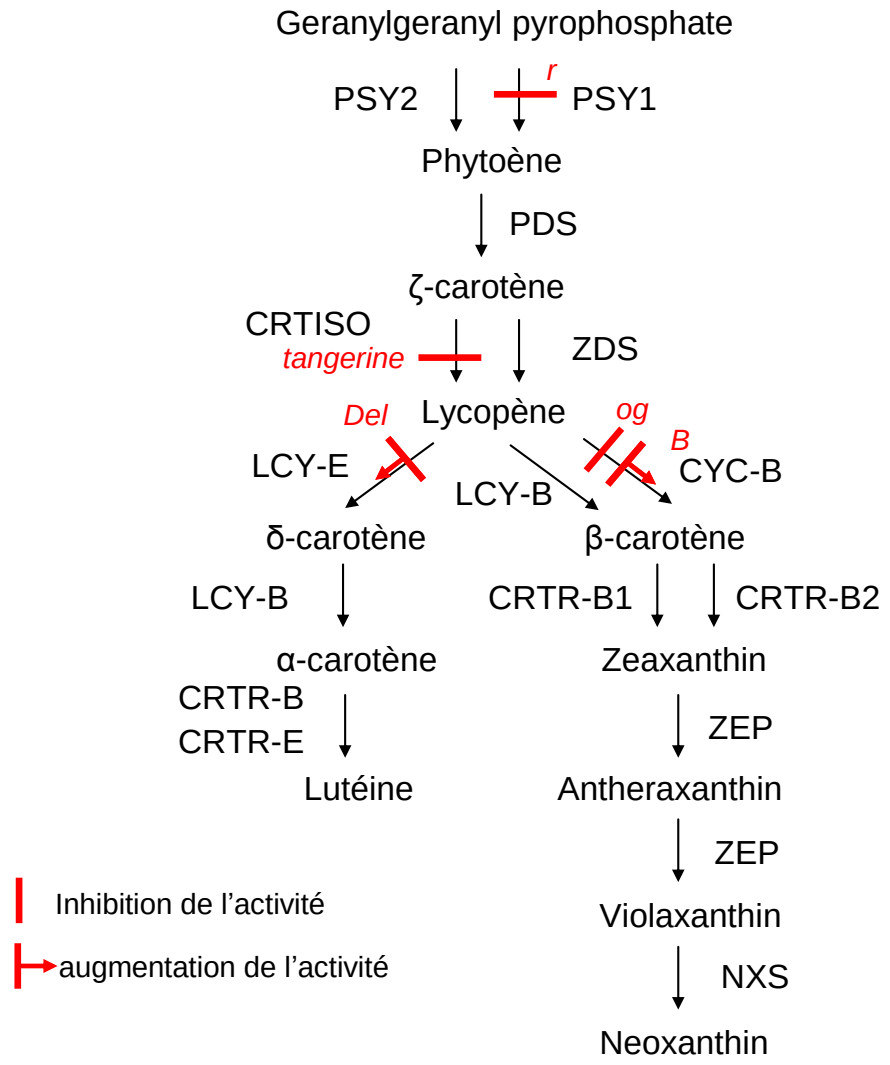


Figure 1-18. Voie de biosynthèse des caroténoïdes chez la tomate et localisation des mutations ayant un effet sur la couleur du fruit. D'après Paran et van der Knaap (2007). Les mutations ayant un effet sur le phénotype sont indiquées par les barres rouges et les noms des phénotypes associés sont aussi colorés en rouge. Les enzymes de la voie de biosynthèse sont les suivantes : PSY1, chromoplast-specific phytoene synthase ; PSY2, chloroplast-specific phytoene synthase ; PDS, phytoene desaturase ; ZDS, ζ-carotene desaturase ; CRTISO, carotenoid isomerase ; CYCB, chromoplast-specific lycopene β-cyclase; LCY-B, chloroplast-specific lycopene β-cyclase; CRTR-B1, chloroplast-specific b-ring hydroxylase;CRTR-B2, chromoplast- specific β-ring hydroxylase ; ZEP, zeaxanthinepoxidase; NXS, neoxanthin synthase ; LCY-E, lycopene d-cyclase; CRTR-E, δ-ring hydroxylase.

Un autre mutant entraîne la coloration orange du fruit : *tangerine* (Isaacson, Ronen et al. 2002). *TANGERINE* code pour une caroténoïde isomérase dont l'expression est éteinte chez le mutant. Un autre locus est responsable de la couleur orangée naturelle des fruits de *S. cheesmaniae*. Les fruits de cette espèce contiennent 5 à 10 fois plus de β-carotène que les fruits des autres espèces sauvages (hormis *S. pimpinellifolium*). Le gène *BETA*, responsable,

code pour une lycopène β-cyclase spécifique du chromoplaste. Le phénotype *old-gold* (*og*) est dû à un allèle récessif de ce gène. Les mutants *og* montrent une augmentation significative de la concentration en lycopène dans le fruit ce qui donne des fruits avec une couleur rouge profond (Ronen, Carmel-Goren et al. 2000).

En plus de gènes affectant directement la biosynthèse des caroténoïdes, s'ajoutent d'autres mutations qui influencent l'intensité de la couleur du fruit (Paran and van der Knaap 2007). La couleur marron de certains fruits est due à la concentration dans le fruit à la fois de caroténoïdes rouges et de pigments chlorophylliens verts (normalement dégradés à maturité). Les deux locus *green flesh* (*gf*) et *chlorophyll retainer* (*cl*) sont responsables de ce phénotype. Les fruits violacés accumulent plus d'anthocyane que les fruits rouges. Ce phénotype est dû au locus *Anthocyanin fruit* (*Aft*), une mutation dominante introgressée à partir de *S. chilense*. *Aft* est un gène orthologue au gène *A*, responsable du même phénotype chez le piment. Ce gène est lui-même homologue à *ANTHOCYANIN2*, un facteur de transcription contrôlant l'expression des gènes dans la voie de biosynthèse des anthocyanes chez *A. thaliana*.

La couleur rose des fruits de tomate est due à l'absence de pigmentation de l'épiderme. Ce phénotype est conditionné par la présence du locus *y*. La combinaison des mutations aux locus *y* et *r* confère au fruit une couleur qui peut aller du jaune pâle au blanc. *hp-1* et *hp-2* (*high pigment* 1 et 2) sont deux locus responsables de l'augmentation de la production de caroténoïdes, flavonoïdes et vitamines. *HP-1* code pour un gène orthologue de *DEETIOLATED1* (*DET1*) et *HP-2* code pour un orthologue de *UV DAMAGED DNA BINDING* protein (*DDB1*), tous deux identifiés chez *A. thaliana*.

La figure 1-19 résume les locus majeurs identifiés, responsables de la diversité morphologique du fruit chez la tomate.

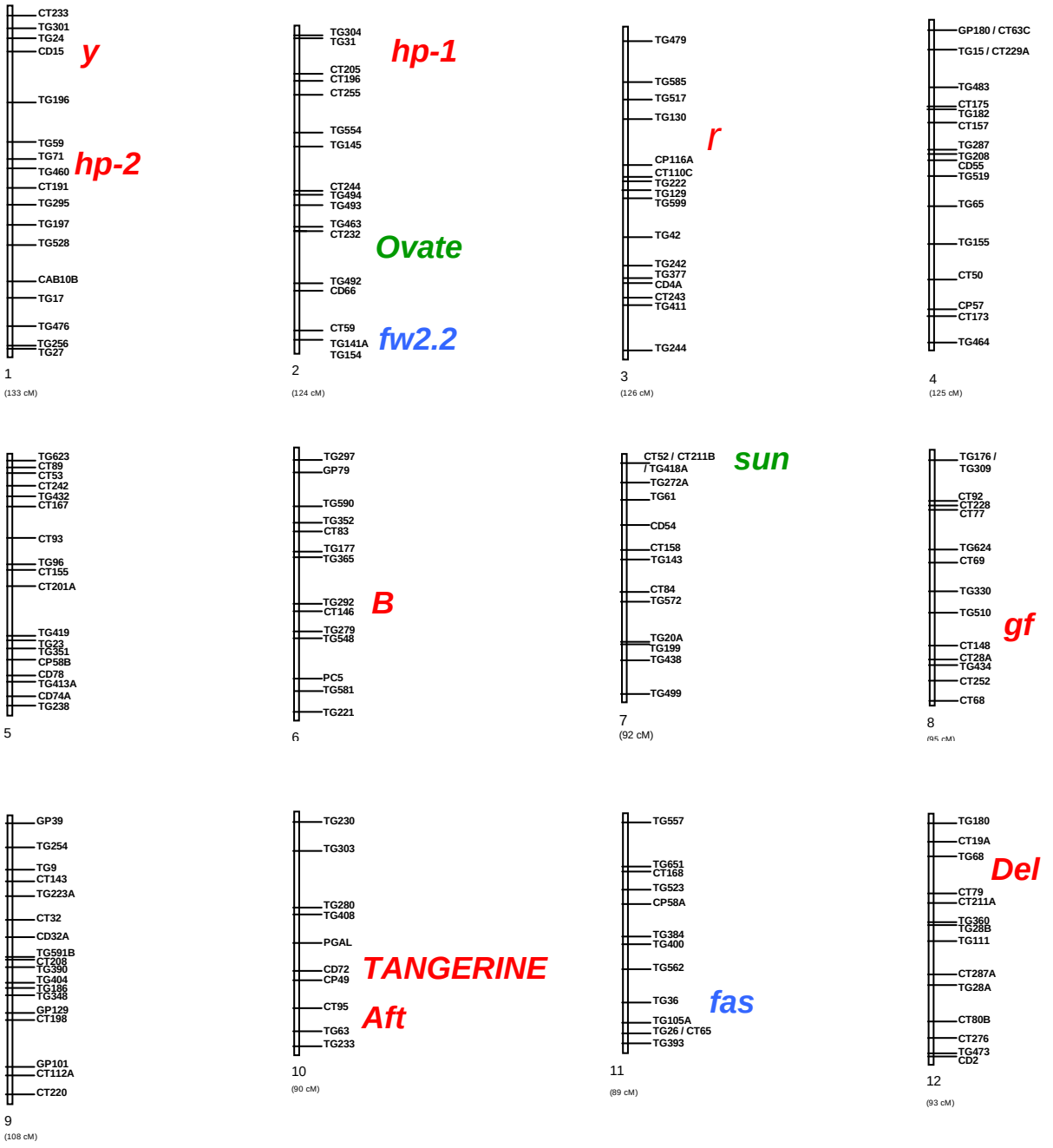


Figure 1-19. Carte de synthèse des différents gènes identifiés, responsables de la diversité morphologique du fruit chez la tomate.
Les gènes impliqués dans la variation de la couleur du fruit sont indiqués en rouge, ceux impliqués dans la variation du poids du fruit sont indiqués en vert et enfin ceux impliqués dans la variation de la forme du fruit sont indiqués en bleu. Pour chaque chromosome, les marqueurs sont indiqués en noir. Le gène *fas* en augmentant le nombre de loges modifie le poids et la forme du fruit.

Les variations de couleur ne sont pas contrôlées uniquement par des gènes majeurs. Il existe une variation continue qui est mesurée en utilisant une échelle colorimétrique. On cherche à mettre en évidence la déviation des teintes par réflexion et on convertie cette déviation en coordonnées, dans un espace colorimétrique. On utilise notamment l'espace CIE Lab, développé par la Commission Internationale de l'Eclairage. Cet espace permet de caractériser une couleur à l'aide d'un paramètre d'intensité et de deux paramètres décrivant la teinte. Des QTL liés à ces trois composantes ont été cartographiés sur la population de lignées recombinantes *S. lycopersicum* x *S. pennellii* (Yong-Sheng, Amit et al. 2003). Seize QTL ont été identifiés mais seulement trois correspondent à des mutations déjà connues (*r*, *B* et *Del*). Les séquences de 23 gènes codant pour des enzymes de la voie de biosynthèse des caroténoïdes ont été cartographiées et seulement cinq co-localisent avec des QTL. Ceci implique que les gènes de la voie de biosynthèse n'expliquent pas toute la variation observée. Des études génétiques de la variation de ces trois composantes de la couleur ont été réalisées chez une population de lignées d'introgression issues du croisement *S. l. esculentum* x *S. l. cerasiforme* (Saliba-Colombani, Causse et al. 2001). Sur huit QTL identifiés, seuls deux co-localisent entre les études précédentes. Le QTL a2.1 est commun entre ces deux études, et co-localise avec une phytoène synthase et une Plastid-lipid-associated protein qui est impliquée dans le transport des caroténoïdes vers le chromoplaste (Yong-Sheng, Amit et al. 2003).

1.3.4. Contenu en sucres et acides

Le contenu en sucres peut être mesuré de plusieurs façons. Les sucres comme le glucose et le fructose peuvent être dosés par réactions enzymatiques et par HPLC (High pressure Liquid Chromatography). Ces mesures sont lourdes à mettre en place alors la plupart des études génétiques se sont concentrées sur l'évaluation du contenu en solides solubles (Soluble Solid Content ou SSC), mesuré par réfractométrie sur la pulpe du fruit de tomate. Le SSC mesuré en degré brix est fortement corrélé à la concentration totale en sucres mais aussi en acides organiques (Fulton, Bucheli et al. 2002). Eshed et Zamir (1995) ont détecté 23 QTL pour le contenu en solides solubles et 14 QTL pour le rendement en sucres solubles (en combinant rendement en fruit par plante et SSC) dans la population de lignées d'introgression *S. pennellii* x *S. lycopersicum*. La plupart de ces QTL co-localisent avec des QTL de poids de fruit avec des effets alléliques opposés. Ceci est dû à la corrélation négative qui existe entre ces deux caractères (Eshed and Zamir 1996). La Figure 1-20 représente une synthèse des différents QTL identifiés pour le poids frais et la teneur en solides solubles. Il y a pour la

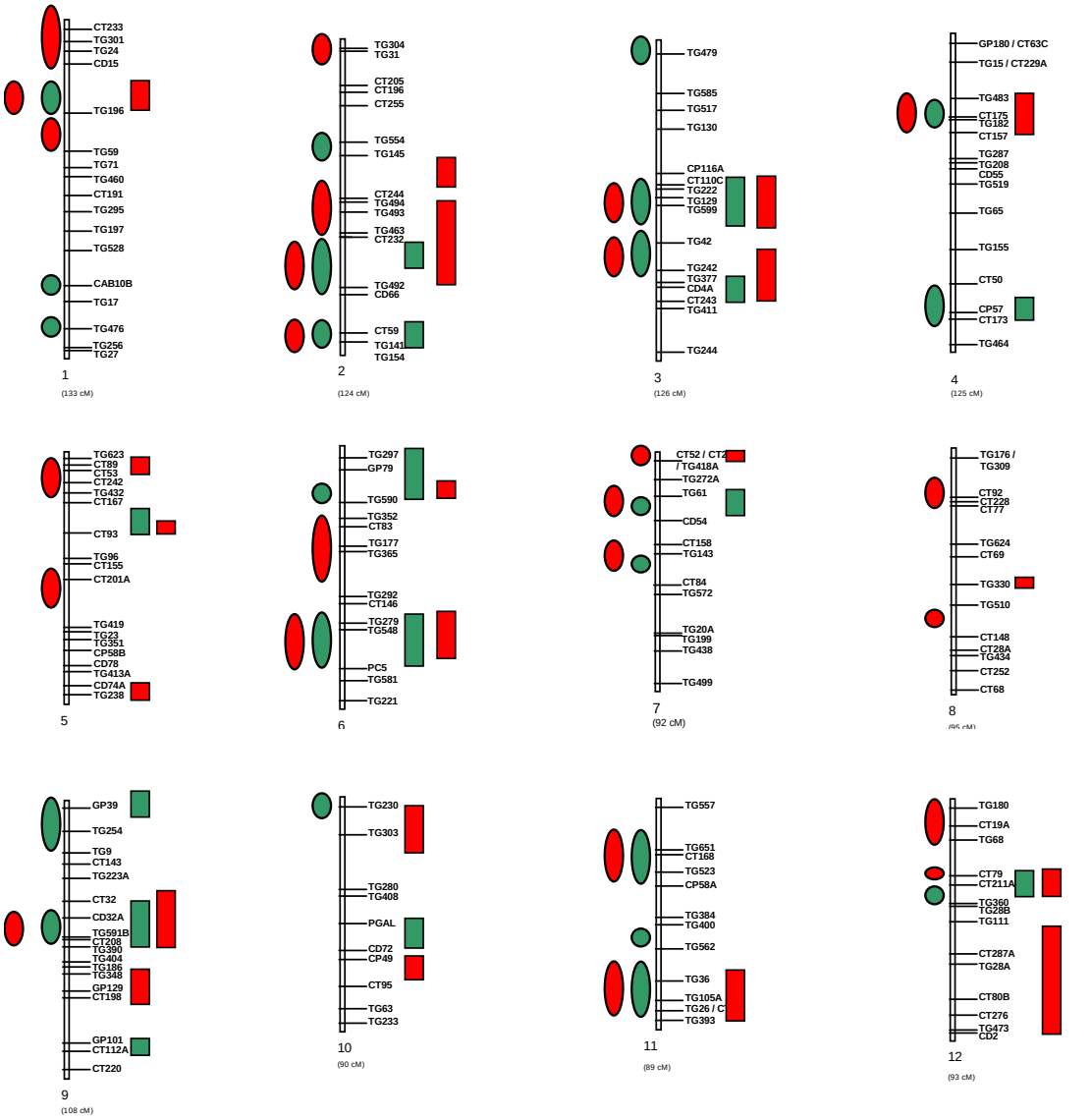


Figure 1-20. Carte de synthèse des différents QTL responsables dans la variation du poids du fruit et de la teneur en solides solubles chez la tomate. Modifiés d'après Prudent (2008).

Pour chaque chromosome, les marqueurs sont indiqués en noir. Les QTL de poids frais sont positionnés à gauche de chaque chromosome et les QTL de teneur en solides solubles sont positionnés à droite. En rouge, sont indiqués les QTL détectés dans au moins une population utilisant comme deuxième parent une accession de la série *Eulycopersicon* (fruit rouge à maturité) .En vert sont indiqués les QTL détectés dans au moins deux populations utilisant comme parent sauvage une accession de la série *Eriopersicon* ou *Neolycopersicon* (fruit vert à maturité). Les différentes populations sont indiquées dans le tableau 1-3.

plupart des régions chromosomiques impliquées, co-localisation entre QTL de poids et QTL de teneur en solides solubles (sucres et acides).

Un de ces QTL, brix9-2-5 ou *lin5* (haut du chromosome 9), a été cloné et la différence phénotypique entre *S. pennellii* et le cultivar M-82 réside dans une mutation au sein d'un gène codant pour une invertase pariétale (Fridman, Pleban et al. 2000). Cette mutation entraîne le changement d'un acide aminé proche du site catalytique ce qui accélère la cinétique de l'enzyme issue de l'accession *S. pennellii*. A ce jour, *lin5* est le seul QTL responsable d'une différence du contenu en solides solubles cloné chez la tomate. Des QTL liés à la teneur en sucres (concentrations obtenues par dosage enzymatique) ont été cartographiés sur la population *S. l. cerasiforme* x *S. l. esculentum* développée à Avignon. Quatorze QTL ont été identifiés et la plupart d'entre eux co-localisent avec des QTL de teneur en solides solubles (Saliba-Colombani, Causse et al. 2001; Causse, Saliba-Colombani et al. 2002).

Tous les gènes codant pour des enzymes qui entrent dans le cycle de Krebs, le cycle de Calvin ainsi que dans les flux d'eau, le transport des sucres ou le transport des acides sont potentiellement des gènes candidats pour la teneur en sucres et en acides organiques et donc pour la teneur en solides solubles. Causse, Duffe et al. (2004) ont cartographié 63 gènes codant pour des enzymes impliqués dans le cycle de Calvin, la glycolyse, le cycle de Krebs, le métabolisme des sucres et de l'amidon, le transport ainsi que d'autres fonctions reliées au métabolisme primaire, sept gènes spécifiques du cycle cellulaire, 14 gènes exprimés durant l'étape de division cellulaire dans le fruit en formation et 23 gènes exprimés durant l'étape d'expansion cellulaire. Des QTL de poids du fruit et contenu en sucres et en acides organiques ont été cartographiés sur la même population. Certains QTL de poids de fruit co-localisent avec des cyclines impliquées dans l'activité mitotique. Des QTL de teneur en sucres et de contenu en solide soluble co-localisaient avec *lin5* identifié auparavant, avec des gènes codant pour la grande et la petite sous unité de l'ADP-glucose pyrophosphorylase, une invertase vacuolaire et une fructokinase. Des QTL de teneur en acides organiques co-localisaient avec une phosphoenolpyruvate carboxylase impliquée dans les voies contrôlant la biosynthèse de l'acide malique et citrique, la sucrose synthase ainsi que la G6P isomérase qui ont un rôle central dans la glycolyse, l'enzyme NADP-malique, la PEP carboxykinase, ainsi que des enzymes liées au transport des sucres ou acides (phosphatase vacuolaire, ATPase vacuolaire et aquaporines spécifiques du fruit).

Une combinaison d'analyse de séquences de marqueurs, d'analyse de variation allélique de gènes candidats et d'évaluation de corrélation entre expression de gène et composition en métabolites a été utilisée dans une étude plus récente pour identifier des gènes candidats responsables de certains QTL de teneur en métabolite et de rendement (Bermudez, Urias et al. 2008). Ces approches de cartographies combinées de QTL et de gènes candidats ne valident en aucun cas leurs rôles mais les gènes co-localisant avec des QTL peuvent par la suite être validés génétiquement ou fonctionnellement.

1.3.5. Maturation du fruit – fermeté

La maturation des fruits implique de nombreux processus intervenant dans la production des caroténoïdes et des composés aromatiques, dans des modifications du contenu en sucres et acides ainsi que dans des modifications de la texture du fruit (Giovannoni 2004; Barry and Giovannoni 2007). La tomate est un fruit climactérique, caractérisé par une augmentation brusque de la respiration au début du processus de maturation qui est accompagné par la production d'éthylène (Figure 1-21).

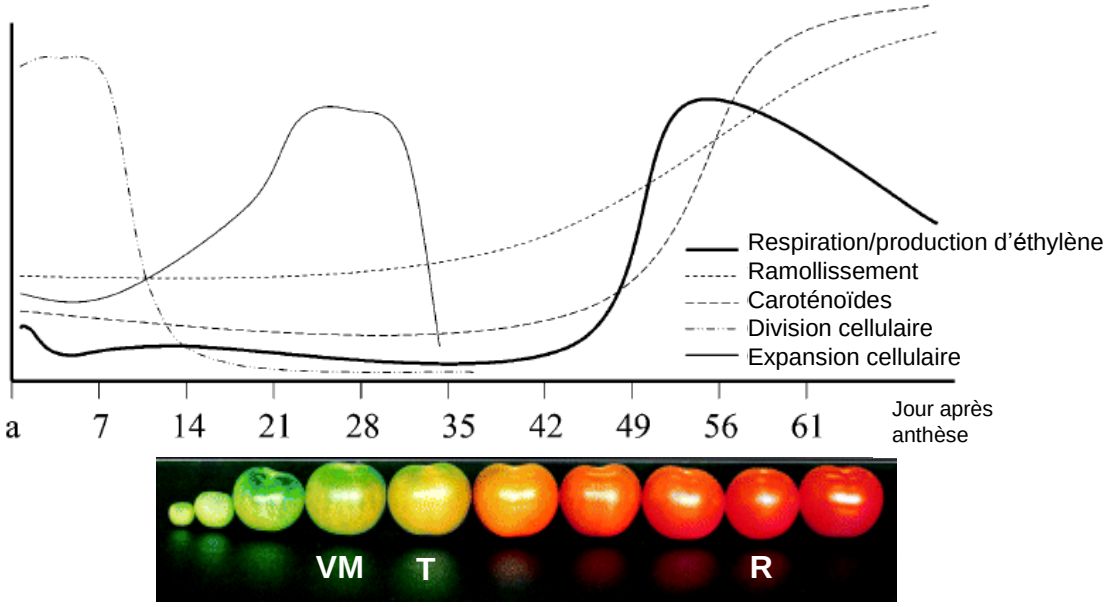


Figure 1-21. Changements majeurs dans le développement du fruit pendant la croissance et la maturation. D'après Ronen et al. (1999) et Giovannoni (2004).
Le délai entre l'anthèse (a) jusqu'au stade vert mature (VM ; fruit ayant atteint sa taille finale avec graines mûres), tournant (T ; première accumulation visible des caroténoïdes) et rouge mûr (R) peut varier de façon importante entre cultivars. Les lignes indiquées sur ce schéma correspondraient à un cultivar ayant des fruits moyens à larges (5-7 cm de diamètre).

Les mutants de maturation *ripening inhibitor* (*rin*) et *non-ripening* (*nor*) ne produisent pas d'éthylène et possèdent une faible quantité de caroténoïdes dans le fruit mûr. Les fruits produits par ces mutants sont beaucoup plus fermes et peuvent être conservés pendant un temps supérieur par rapport aux fruits produits par des plantes ne possédant pas le phénotype. Les gènes *RIN* et *NOR* codent respectivement pour une protéine de la famille des MADS-box protein et un facteur de transcription. Ces deux gènes ne semblent pas être impliqués dans la voie hormonale de la maturation (indépendants de la voie de régulation de l'éthylène) (Moore, Vrebalov et al. 2002; Vrebalov, Ruezinsky et al. 2002). Les plantes homozygotes pour *rin* et *nor* montrent un phénotype extrême et ces mutations ne sont donc utilisées, en amélioration variétale, qu'à l'état hétérozygote dans les hybrides F1. D'autres mutations dominantes ont été identifiées mais ont un effet beaucoup plus fort que les deux précédentes. Le gène *Never-ripe* (*Nr*) code pour un récepteur éthylénique (Wilkinson, Lanahan et al. 1995) et les mutants *Green-ripe* et *Never-ripe2* présentent une réduction de la sensibilité à l'éthylène (Barry and Giovannoni 2006). Pour les mutants *Green-ripe* et *Never-ripe2*, il s'agit d'une seule mutation dans une protéine codant pour une protéine de fonction inconnue mais qui participe à la signalisation hormonale par l'éthylène. *Cnr*, un autre phénotype mutant pour la maturation du fruit est apparu spontanément dans une population commerciale et se cartographie sur le haut du chromosome 2 (Thompson, Tor et al. 1999). Les fruits de ce mutant montrent une réduction de la production d'éthylène, une inhibition du ramollissement du fruit, une peau jaune ainsi qu'un péricarpe non pigmenté. Le phénotype mutant est dû à une modification épigénétique spontanée du promoteur d'un gène codant pour une SBP-box (Manning, Tor et al. 2006).

La fermeté du fruit implique principalement des mécanismes biochimiques et physiologiques relatifs à la structure des tissus : pression osmotique des cellules, adhésion cellulaire, rigidité et élasticité des parois cellulaires (Chaib, Devaux et al. 2007). Toutes les enzymes intervenant dans le métabolisme des parois cellulaires comme les polygalacturonases, les pectine méthylestérases, les β-galactosidases, les xyloglucane endotransglycosylases et les endo-β-1,4-glucanases sont des protéines candidates pouvant affecter la fermeté du fruit à maturation (Brummell and Harpster 2001). La composition des espaces inter-membranaires, notamment dans la lamelle moyenne, pourra modifier la texture et la fermeté du fruit. A maturité, ce sont les polygalacturonases qui hydrolysent les pectines et diminuent la cohésion cellulaire. La pression osmotique des cellules est dépendante des échanges en solutés et en eau entre les cellules et le milieu extérieur. Ces échanges ont lieu au

niveau de la membrane plasmique et sont assurés par des protéines canaux tel que les aquaporines. Toutes ces enzymes sont donc potentiellement candidats dans l'établissement de la fermeté du fruit à maturité.

Plusieurs études ont utilisé la seule population intra-spécifique développée chez la tomate. Il s'agit d'une population faisant intervenir une tomate moderne (*S. l. esculentum*) possédant des fruits de calibre élevé mais des qualités organoleptiques très moyennes et une tomate de type cerise (*S. l. cerasiforme*) aux qualités organoleptiques appréciées par les consommateurs mais avec des fruits relativement petits. Cette population a permis de cartographier des QTL liés à la qualité du fruit. Etant donné que les analyses sensorielles nécessitent un large panel de dégustateurs entraînés pour plusieurs semaines d'analyse, elles sont excessivement lourdes à mettre en place. Les études de Saliba-Colombani, Causse et al (2001) et de Causse, Saliba-Colombani et al.(2001) permettent de relier les QTL sensoriels à des QTL de caractères physico-chimiques (fermeté, concentration en sucre) (Figure 1-22). On remarquera que les QTL liés à la qualité s'organisent en « cluster », en particulier sur le chromosome 2, le chromosome 4 et 9 (Causse, Saliba-Colombani et al. 2002). Les composantes de la texture et de la flaveur ont, ensuite, été évaluées directement par des mesures physiques ou des dosages. Cette population présente l'avantage d'être très proche de la tomate cultivée tout en étant génétiquement polymorphe. De plus, les QTL identifiés sont potentiellement des QTL impliqués dans la domestication de la tomate.

L'étude des bases génétiques de la qualité du fruit s'intéresse à plusieurs caractères différents qui intègrent, eux même, de nombreux processus indépendants ou interconnectés. La complexité de ces caractères explique sans doute pourquoi ils étaient relativement peu travaillés en amélioration variétale par rapport aux résistances génétiques, aux capacités d'adaptation à différents modes de culture ou encore au rendement. Cette complexité reflète aussi le besoin de compléter les schémas de sélection traditionnels par la SAM (Sélection Assistée par Marqueurs - présentée dans la première partie de ce chapitre) afin de suivre plus facilement les différents caractères lors des générations successives. Ceci implique donc une meilleure connaissance à la fois moléculaire et physiologique des caractères qui contribuent à la qualité du fruit.

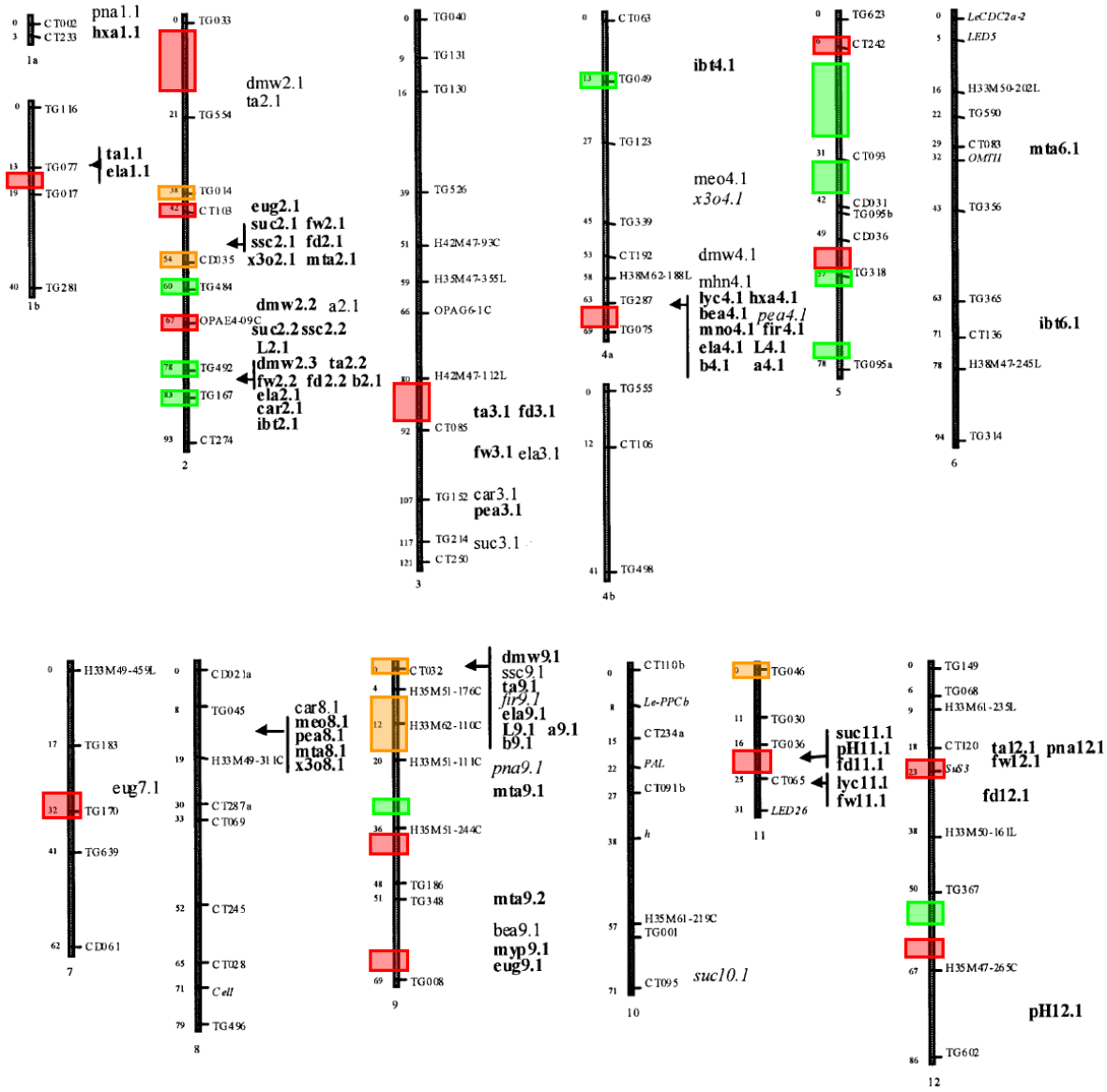


Figure 1-22. Carte de synthèse des QTL de qualité de la tomate basée sur une population intraspécifique de RIL (Recombinant Inbred Lines) obtenue à partir d'un croisement entre une lignée de type cerise et une lignée de type moderne. D'après Saliba-Colombani, Causse et al (2001) et Causse, Saliba-Colombani et al.(2001).

Les noms des marqueurs sont indiqués à droite des chromosomes. Les QTL détectés par CIM (Composite Interval Mapping) et IM (Interval Mapping) sont indiqués en lettres normales et ceux détectés uniquement par IM sont indiqués en italique. Les caractères analysés sont : poids du fruit (fw), diamètre du fruit (fd), fermeté (fir), élasticité du fruit (ela), couleur (L, a, b), poids de matière fraîche (dmw), teneur en solide soluble (ssc), teneur en sucre (suc), acidité titrable (ta), pH (pH), teneur en lycopène (lyc), teneur en carotène (car). Les autres QTL indiqués correspondent à 18 composés volatils aromatiques.

Les carrés représentent les QTL détectés par analyse sensorielle. En rouge sont indiqués les QTL liés à la saveur, en vert les QTL liés à la texture et en orange les régions présentant à la fois des QTL de saveur et de texture.

1.4. Contexte et objectif de l'étude.

Il y a eu ces dernières années un intérêt croissant pour l'utilisation de populations naturelles avec différents objectifs (Weigel and Nordborg 2005). Tout d'abord, il semble évident que l'étude d'un caractère par cartographie génétique ou par génétique inverse permet d'identifier les bases moléculaires de ce caractère mais seule la cartographie génétique reflète les mutations impliquées *in natura* dans l'évolution de celui-ci. Ainsi, si on s'intéresse aux bases moléculaires de l'adaptation à un environnement ou à la réponse phénotypique d'une espèce à la sélection, il faudra vérifier après identification des locus responsables, que les polymorphismes identifiés sont bien présents dans un échantillon plus large que les deux lignées de départ. Il faudra aussi valider la corrélation entre les variations moléculaires et phénotypiques dans cet échantillon. Lippman, Cohen et al. (2008) offrent un exemple de clonage positionnel chez la tomate où le polymorphisme identifié est responsable du phénotype « inflorescence composée » dans la population de cartographie mais aussi dans une collection de ressource génétique. Les populations en ségrégation sont, en principe, construites à partir de deux lignées fixées et offrent la possibilité d'identifier l'effet de variations présentes uniquement entre ces deux lignées. Il y a donc de forte chance de n'identifier qu'une faible part des variations responsables du caractère. De plus, malgré l'efficacité de la cartographie génétique de QTL, celle-ci n'aboutit généralement pas jusqu'au clonage des gènes responsables, ou demande un travail spécifique. Il semble aussi que cette approche ne soit pas efficace pour identifier des gènes avec des effets relativement faibles.

Un avantage conséquent dans l'utilisation des populations naturelles est le fait que les ressources génétiques sont directement exploitables. Il n'est pas nécessaire de créer de populations synthétiques ce qui implique néanmoins que l'échantillonnage n'est contrôlé qu'*a posteriori*. Par ailleurs, durant leur histoire évolutive, les populations naturelles accumulent théoriquement de nombreux évènements de recombinaison, ce qui augmente la résolution qu'offrent ces populations dans une approche de cartographie. Un des freins à l'utilisation des ressources génétiques, pour identifier directement les bases moléculaires d'un QTL, est le nombre de marqueurs moléculaires nécessaires, pour couvrir l'ensemble du génome de l'espèce travaillée. A la fois, l'identification des marqueurs et leur génotypage sur un grand échantillon, représentent un effort humain et financier important. Cependant avec la diminution des coûts de séquençage et de génotypage (notamment grâce à l'arrivée des nouvelles technologies de séquençage ou technologie « Next-Generation ») ce frein est en train d'être levé et il semble nécessaire, dès maintenant, de commencer à concevoir

l'utilisation des ressources génétiques pour identifier les bases moléculaires de caractères d'intérêt, parallèlement à la cartographie génétique.

Cette méthodologie ne semble pas applicable à des espèces fortement autogames, si elle est utilisée dans le but d'identifier le polymorphisme causal du phénotype étudié. En effet les recombinaisons, dont le taux détermine la résolution de l'analyse, ont bien lieu pendant la méiose mais elles sont inefficaces car les individus sont homozygotes pour une grande partie du génome. La tomate ne semble donc pas être, *a priori*, un modèle efficient pour utiliser la génétique d'association afin d'identifier les gènes sous-jacent aux QTL. Cet argument est appuyé par une étude réalisée par van Berloo et al. (2008) qui utilisent des marqueurs AFLP et étudient le déséquilibre de liaison entre ces marqueurs au sein d'une population de 18 accessions de tomate cerise. Le DL s'étend sur 20 cM, ce qui n'offre aucun avantage par rapport à des populations en ségrégation en termes de résolution (Figure 1-7).

Nesbitt et Tanksley (2002) ont utilisé une méthode cladistique afin d'identifier des associations statistiques entre le poids du fruit et des polymorphismes moléculaires identifiés dans le gène *fw2.2* (Frary, Nesbitt et al. 2000). Ils ont, pour cela, séquencé le gène *fw2.2*, son promoteur ainsi que d'autres locus, sur des individus de différentes espèces. L'étude s'est focalisée sur un nombre limité d'accessions de tomate de type cerise. Les auteurs se focalisent sur ces accessions car elles présentent des poids du fruit intermédiaires entre l'espèce sauvage *S. pimpinellifolium* et l'espèce cultivée *S. l. esculentum*. Le poids du fruit fait parti du syndrome de domestication de cette espèce. On s'attend donc à avoir une distribution, chez l'ancêtre de la tomate cultivée, des allèles « gros fruit » et « petit fruit » au niveau du locus *fw2.2*. C'est en se basant sur cette hypothèse que les auteurs espèrent identifier le polymorphisme responsable du phénotype par étude d'association.

Aucun polymorphisme ne ressort associé au poids du fruit. Nesbitt et Tanksley découvrent cependant que la diversité moléculaire est supérieure dans la partie 5' du gène que dans le gène lui-même (Figure 1-23). Les auteurs estiment donc que le polymorphisme causal, dans la variation d'expression de *fw2.2*, se trouve hors de la région promotrice séquencée. Une autre hypothèse, plus vraisemblable, admet que la variation du poids du fruit chez les accessions *S. l. cerasiforme* serait imputable à la combinaison d'autres QTL de poids du fruit. Les auteurs montrent aussi que le locus des accessions de tomate de type cerise est une mosaïque de polymorphismes d'origine sauvage et cultivée.

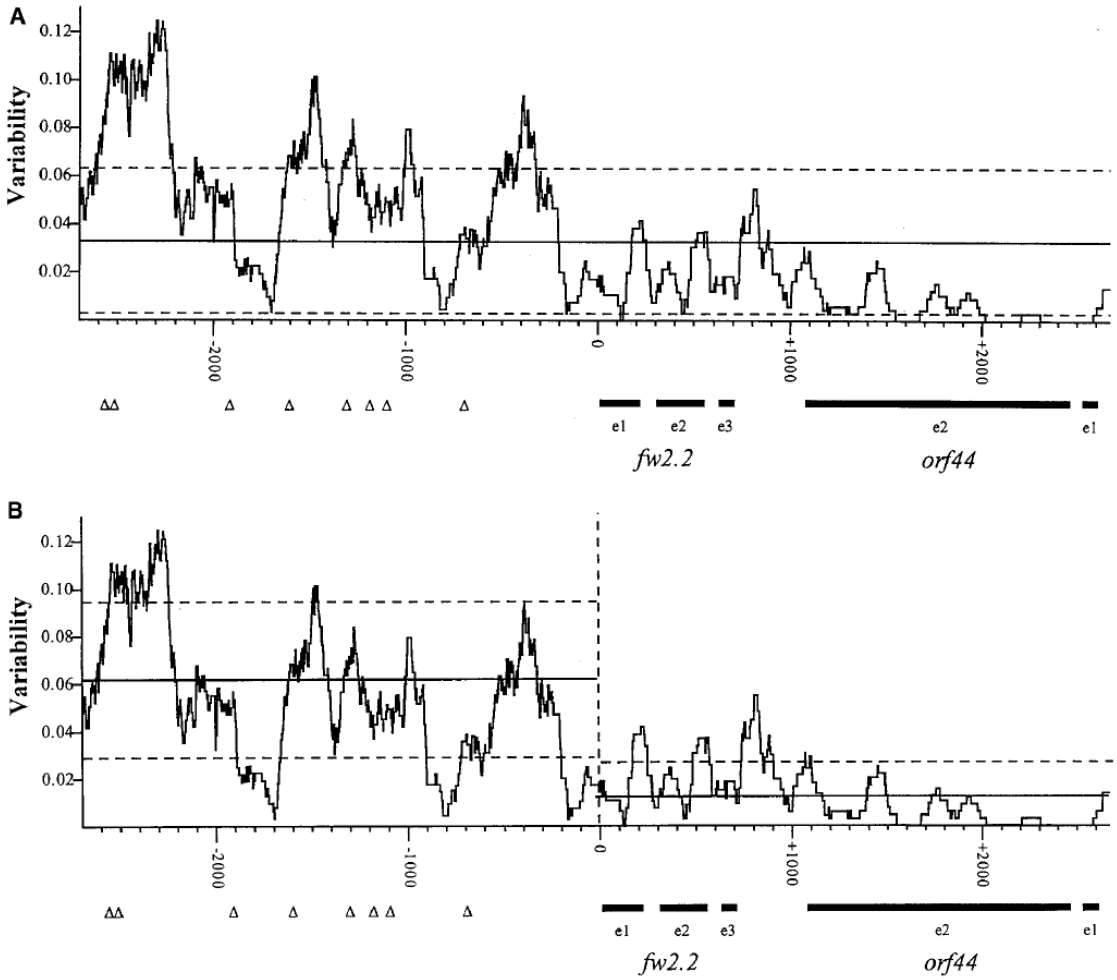


Figure 1-23. Etude de la diversité nucléotidique dans la région de *fw2.2*. D'après (Nesbitt and Tanksley (2002).
L'analyse est réalisée par fenêtre glissante (Sliding Window Analysis of Nucleotid variability) sur la région de *fw2.2* incluant le gène, la région promotrice et l'ORF (Open Reading Frame) adjacente *orf44*. Les codons des gènes sont indiqués par des barres noires sous le graphe. (A) La moyenne et l'écart type sont calculés sur toute la séquence. (B) Les moyennes et les écarts types sont calculés séparément pour les régions en amont et en aval du premier nucléotide de *fw2.2* (abscisse 0). Les barres horizontales pleines représentent les variabilités moyennes, les barres horizontales pointillées représentent les écarts types. Les séquences utilisées pour le calcul ne comprennent que des accessions strictement sauvages (pas d'accessions *S. l. esculentum* ni *S. l. cerasiforme*). Les mutations partagées par toutes les accessions « gros fruit » (synapomorphies) sont indiquées par un Δ en dessous du graphe.

Cette position « admixture » des tomates de type cerise implique l'existence d'hybridations fréquentes entre le compartiment sauvage et le compartiment cultivé. Cette observation avait déjà été faite en analysant la diversité génétique de tomates sauvages et cultivées avec des isozymes (Rick 1958; Rick and Fobes 1975; Rick and Holle 1990) ou avec des marqueurs RFLP (Miller and Tanksley 1990).

L'utilisation de populations en « admixture » chez l'humain a permis d'identifier plusieurs locus liés à la prédisposition à certaines maladies (Seldin 2007). Chez l'humain, le DL s'étend sur une dizaine de Kb (Reich, Cargill et al. 2001) ce qui implique un grand nombre de marqueurs moléculaires pour identifier les polymorphismes responsables de pathologies. Un mélange récent de deux ou plusieurs populations différenciées génétiquement, va avoir pour conséquence une augmentation du DL et va donc réduire le nombre de marqueurs nécessaires. La cartographie par « admixture » (admixture mapping) est semblable à une étude de cartographie utilisant des lignées d'intercroisements avancés (AIL ou Advanced Intercross Lines) (Figure 1-24) (Darvasi and Shifman 2005). Cette méthode permet de réduire le nombre de marqueurs. Elle est décrite comme efficace, relativement robuste (Seldin 2007) et elle présente des caractéristiques intermédiaires entre les analyses d'association et la cartographie génétique (Tableau 1-5). Chez le peuplier (*Populus sp.*) ainsi que chez le tournesol (*Helianthus sp.*), deux espèces préférentiellement allogames, l'utilisation d'individus recombinants naturels entre deux espèces différentes permettent d'augmenter le DL et donc, de développer des approches de cartographie par « admixture » avec un nombre de marqueurs réduits (Rieseberg and Buerkle 2002; Lexer, Buerkle et al. 2006).

Les avantages apportés par cette stratégie chez l'humain ou chez les espèces préférentiellement allogames ne semblent pas convenir à une espèce autogame. La tomate cultivée, comme nous l'avons vu, présente une très faible diversité génétique mais une grande variabilité phénotypique pour des caractères liés à la qualité du fruit. La génétique d'association exploite la diversité génétique des populations pour simplifier des caractères complexes à des gènes ou polymorphismes uniques (Zhu, Gore et al. 2008). Or, chez la tomate cultivée, la faible diversité génétique ne permet pas d'utiliser cette méthodologie pour identifier des locus d'intérêt.

La position « admixée» de *S. lycopersicum* var *cerasiforme* permettrait de profiter à la fois du polymorphisme moléculaire présent chez l'espèce sauvage et de la variabilité phénotypique présente chez l'espèce cultivée. La collection d'accessions de tomate de type cerise maintenue à l'Unité de Génétique et Amélioration des Fruits et Légumes de l'INRA d'Avignon semble donc être un échantillon d'intérêt pour valider l'utilisation de la cartographie par « admixture » chez une espèce cultivée autogame. C'est une hypothèse que nous avons voulu tester.

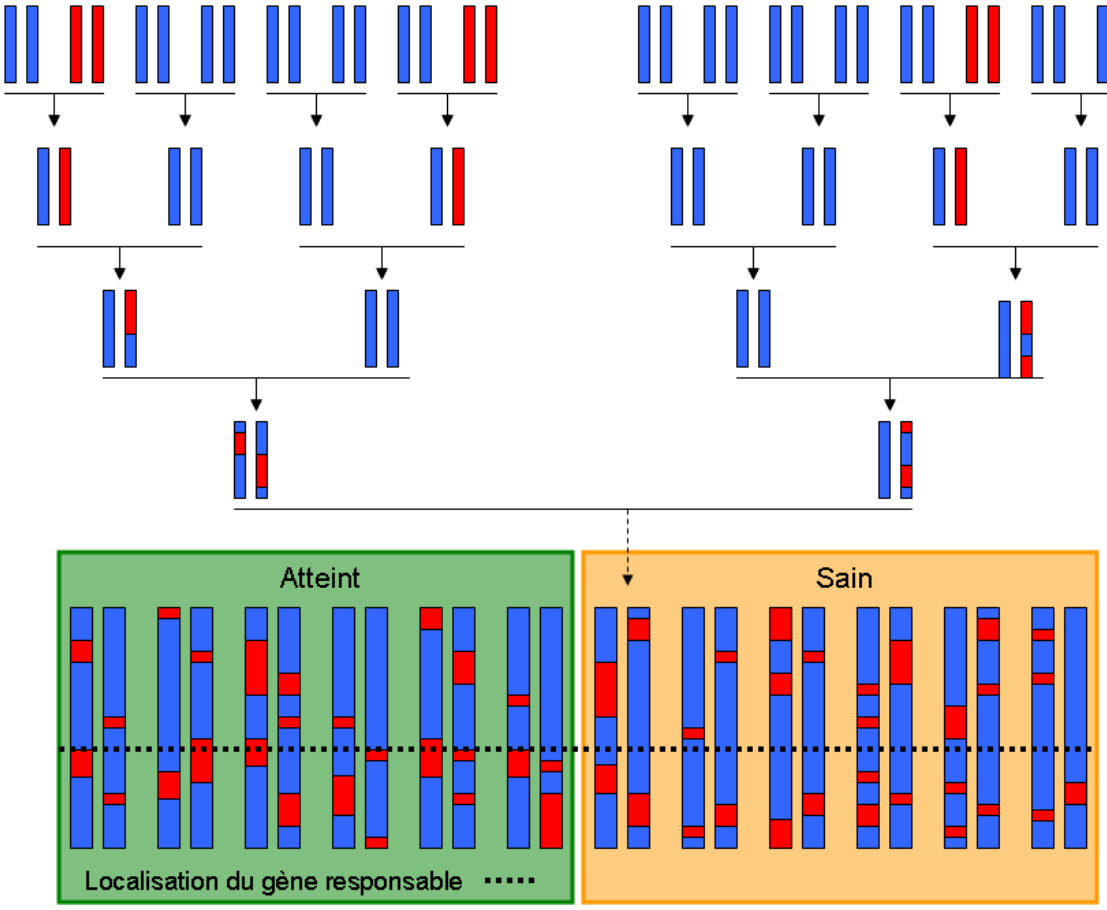


Figure 1-24. Schéma d'une paire de chromosomes pour chacun des individus d'une population en « admixture ». D'après Darvasi and Shifman (2005).
Un groupe atteint (pour une maladie donnée) et un groupe sain sont représentés séparément en bas du schéma. Pour un individu sain (flèche pointillée), tous les ancêtres durant les 4 dernières générations sont représentés. La stratégie de cartographie par « admixture » consiste à scanner le génome et à identifier des régions avec un excès d'apparentement rouge dans le groupe atteint par rapport au groupe sain, en supposant que la population rouge possède l'allèle de prédisposition.

	Cartographie génétique	**Cartographie par « admixture »**	**Analyse d'association**
Puissance statistique	Faible	Forte	Forte
Nombre de SNP requis pour un scan du génome entier	Faible	Faible	Forte
Sensibilité à l'hétérogénéité génétique	Faible	Modérée	Forte
Résolution de cartographie	Faible	Intermédiaire	Bonne

T**ableau 1-5. Principales caractéristiques des stratégies de cartographies.** D'après Darvasi et Shifman (2005)

Mes travaux de thèse ont visé à étudier la structure de la diversité génétique dans un échantillon de la collection de tomates cultivées et sauvages, maintenue et caractérisée à l'Unité de Génétique et Amélioration des Fruits et Légumes. L'échantillon utilisé était principalement composé d'accessions des types *S. l. esculentum* et *S. l. cerasiforme* ainsi que d'accessions de l'espèce *S. pimpinellifolium*. Le but de la thèse était de valider la position mosaïque du génome des accessions de tomate cerise en utilisant des marqueurs microsatellites (SSR) répartis sur tout le génome. Les données de génotypage couplées aux données de phénotypage de toutes les tomates de type cerise et de certaines accessions sauvages et cultivées ont permis de définir différentes « core collections » emboîtées, maximisant à la fois la diversité génétique et phénotypique. Deux approches ont été menées en parallèles. Tout d'abord on a voulu tester l'approche « gène candidat » sur un locus identifié précédemment par clonage positionnel. Une « core collection » a permis d'analyser la diversité moléculaire du QTL contrôlant la variation du nombre de loges. Une deuxième approche, que l'on peut qualifier de « Whole Chromosome Analysis », a été testée. Cette étude s'est focalisée sur le chromosome 2 pour estimer l'étendue du DL sur plusieurs échelles : distance génétique large, distance génétique fine et distance physique. Le chromosome 2 a été choisi car de nombreux QTL organisés en cluster ont été détectés sur ce chromosome (Figure 1-25).

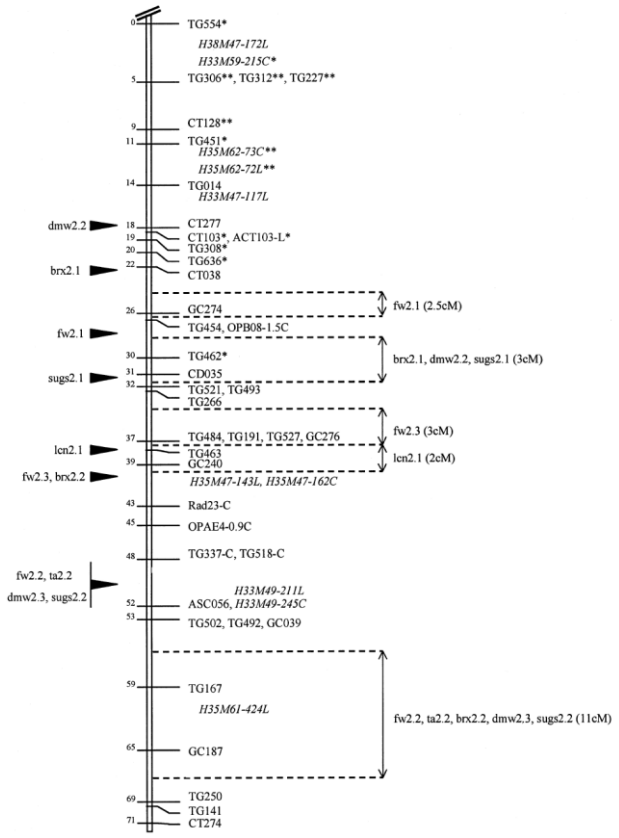


Figure 1-25. Carte synthétique des QTL de la région du chromosome 2 étudié pour l'architecture et la composition du fruit du fruit. D'après Lecomte, Saliba-Colombani et al. (2004)
Les QTL ont été identifiés avec une population de RIL issue du croisement entre une tomate de type cerise et une tomate de type moderne. Ils sont indiqués à gauche du chromosome et la position de chaque QTL a été déterminée par la valeur maximale de LOD. Les intervalles portant les QTL identifiés par cartographie de substitution sont représentés par une flèche, à droite du chromosome. La longueur de la région est indiquée entre parenthèses après le nom des QTL

Les conditions optimales pour mener de façon efficace une étude d'association sur tout le génome chez la tomate sont inférées à partir des résultats obtenus sur ce chromosome. Des associations entre polymorphismes moléculaires et phénotype ont été recherchées. Les informations de séquence obtenues pour l'analyse de la diversité ont aussi permis d'estimer la vraisemblance de différents modèles évolutifs en utilisant une approche basée sur la coalescence. Certains SNP identifiés par séquençage ont permis de construire quatre panels SNPlex® ce qui représente 192 polymorphismes qui seront génotypés sur la collection totale. Malheureusement ces résultats n'ont pas pu être obtenus dans le temps imparti.

Chapitre 2 : Matériel et méthodes

2.1. Matériel Végétal

L'Unité de Recherche de Génétique et Amélioration des Fruits et Légumes (GAFL) de l'INRA d'Avignon est en charge de conserver des ressources génétiques de différentes espèces maraîchères et fruitières. Les ressources génétiques de tomates comprennent plus d'un millier d'accessions provenant d'une collection nationale et d'échanges avec d'autres centres de ressources génétiques comme le TGRC, (Davis, Californie), le CGN (Wageningen, Pays-Bas), le NCRPIS (Geneva, Etats-Unis) et le N.I Vavilov Research Institute of Plant Industry (St Petersburg, Russie). Une partie des accessions de la collection est disponible sur simple demande et une autre partie est limitée car les accessions font partie d'un réseau de multiplication et d'évaluation faisant intervenir des sélectionneurs privés.

Trois cent quatre vingt accessions ont été échantillonnées dans la collection totale sur la base de l'appartenance à différentes espèces. L'échantillon étudié comprend :

- 130 accessions de tomates cultivées à gros fruit appartenant au groupe *S. lycopersicum* var. *esculentum*
- 144 accessions de tomates cultivées ou sauvages à petits fruits classées dans le groupe *S. lycopersicum* var. *cerasiforme*.
- 66 accessions de tomates sauvages apparentées proches appartenant à l'espèce *S. pimpinellifolium*.
- 20 accessions d'espèces sauvages éloignées génétiquement (9 *S. habrochaites*, 2 *S. pennellii*, 2 *S. chmielewskii*, 2 *S. cheesmaniae*, 2 *S. chilense*, 2 *S. peruvianum* et 1 *S. neorickii*)

L'ensemble de la collection est caractérisé pour des descripteurs IPGRI listés dans la Figure 2-1.

Seule une partie de cette collection (201 accessions) a été phénotypée plus finement durant deux années consécutives. Nous avons focalisé l'effort de phénotypage sur les accessions de type cerise (140 accessions), quelques représentants d'accessions à gros fruit (42 accessions) et des accessions de l'espèce *S. pimpinellifolium* (19 accessions).

Le détail des accessions et leur utilisation sont présentées dans l'annexe 1.

- **NOTATIONS PLANTES**
  - Anthocyane de l'hypocotyle (présente / absente)
  - Pilosité (présente / absente)
  - Type de plante (*dwarf* / normale)
  - Forme des folioles (entière / découpées)
  - Port du feuillage (1 / 2 dressé / horizontal / retombant)
  - Type de croissance (déterminée / indéterminée)
  - Nombre d'inflorescences avant détermination (2 à 4 / 4 à 6 / >6)
  - Longueur des entre-nœuds (courts / moyens / longs)

- **NOTATIONS FLEURS, INFLORESCENCES, PEDONCULES**
  - Couleur de la fleur (faune / orangée / blanche)
  - Type d'inflorescence (simple / simple-ramifié / ramifiée)
  - Abscission du pédoncule (présente / absente)

- **NOTATIONS FRUITS**
  - Couleur
    - Collet vert (présente / absente)
    - Intensité verte du collet (claire / moyenne / foncée)
    - Couleur du fruit avant maturité (claire / moyenne / foncée)
    - Couleur du fruit à maturité (vert / blanc / jaune / orange / rose / rouge / autre)
    - Couleur de la chair (verte / blanche / jaune / orange / Crimson / rouge / autre)
    - Couleur de l'épiderme (incolore / coloré)
  - Taille, Forme
    - Taille du fruit (<5g / 5-20g / 20-60g / 60-100g / 100-140g / 140-180g / >180g)
    - Homogénéité de la taille (faible / moyenne / bonne)
    - Forme du fruit (aplati / légt.aplati / rond / cordiforme / rectangulaire / ovoïde / cylindrique / ovale / allongé / pyriforme / parallélép. / obovoïde / variable)
    - Homogénéité de la forme (faible / moyenne / bonne)
  - Caractéristiques externes
    - Côtes pédonculaires (absentes / faibles / moyennes / fortes / très fortes / variables)
    - Taille de l'attache pistillaire (très petite / petite / moyenne / grande / très grande / variable)
    - Forme du sommet (en creux / creux-plat / plat / plat-pointu / pointu)
  - Caractéristiques internes
    - Nombre de loges (2 / 2 à 3 / 3 à 4 / 4 à 6 / <6)
    - Epaisseur du péricarpe (mince / moyen / épais)
    - Fermeté (très faible / faible / moyenne / bonne / très bonne)

Figure 2-1 Ensemble des descripteurs de l'International Plant Genetic Resources Institute (IPGRI) utilisés pour caractériser les accessions de tomate à Avignon.
L'ensemble des descripteurs est disponible sur le site :
<http://www.bioversityinternational.org/fileadmin/bioversity/publications/pdfs/488.pdf>

Toutes les accessions *S. lycopersicum* (var. *cerasiforme* et *esculentum*) et *S. pimpinellifolium* ont déjà subie au moins quatre cycles d'autofécondation ce qui induit une fixation des caractères phénotypiques analysées mais aussi une diminution de la diversité lorsque ces accessions étaient conservées ou prospectées à l'état de population. Les accessions de type sauvages (fruits verts à maturité) sont multipliées par autofécondation ou intercroisement selon leur régime de reproduction.

2.2. Génotypage des marqueurs microsatellites et analyse des données

L'échantillonnage des 360 individus a été réalisé avant le début de ma thèse pour réaliser une étude de la structuration de la diversité en utilisant 21 marqueurs microsatellites (SSR) répartis sur le génome. Ces marqueurs ont été choisis sur le site du Sol Genomics Network (http://solgenomics.net/). La figure 2-2 illustre la position des marqueurs sur la carte génétique de référence de la tomate.

Le génotypage de ces marqueurs sur la collection a été réalisé dans le laboratoire de biologie moléculaire de l'Unité Mixte de Recherche « Diversité et Adaptation des Plantes Cultivées », à Montpellier. Les conditions nécessaires à l'amplification de ces marqueurs par PCR (Poly Chain Reaction) et la lecture sur séquenceur ABI 3710 Xl (Applied Biosystems, Foster City, Etats-Unis) sont indiquées dans l'annexe 2.

Divers outils informatiques ont été utilisés pour étudier la structure de la diversité génétique de cet échantillon. L'outil DARWIN 5.0 (Perrier and Jacquemoud-Collet 2006) (disponible à l'adresse http://darwin.cirad.fr/darwin/Home.php) a permis de représenter la diversité de l'échantillon en représentant les dissimilarités, calculées entre individus, sur des dendrogrammes ou par analyse en coordonnées principales (ACoP). Le logiciel STRUCTURE 2.1 développé par Pritchard et al. (2000) (http://pritch.bsd.uchicago.edu/structure.html) a permis d'analyser la structuration de la diversité génétique au sein de l'échantillon. Cette analyse a été complétée par la méthode d'Evanno, Regnaut et al. (2005). Les données de génotypage combinées aux données de phénotypage issues de la première année de culture ont été utilisées pour échantillonner différentes « core collections » emboîtées, maximisant la diversité au niveau moléculaire et au niveau phénotypique, grâce à l'outil MSTRAT 4.1 (http://www.ensam.inra.fr/gap/MSTRAT/mstratno.htm).

Les AMOVA (Analysis of Molecular Variance) qui permettent d'estimer la part de variation génétique imputable à la structure définie, ont été réalisées avec le logiciel ARLEQUIN 3.0 (http://lgb.unige.ch/arlequin/).

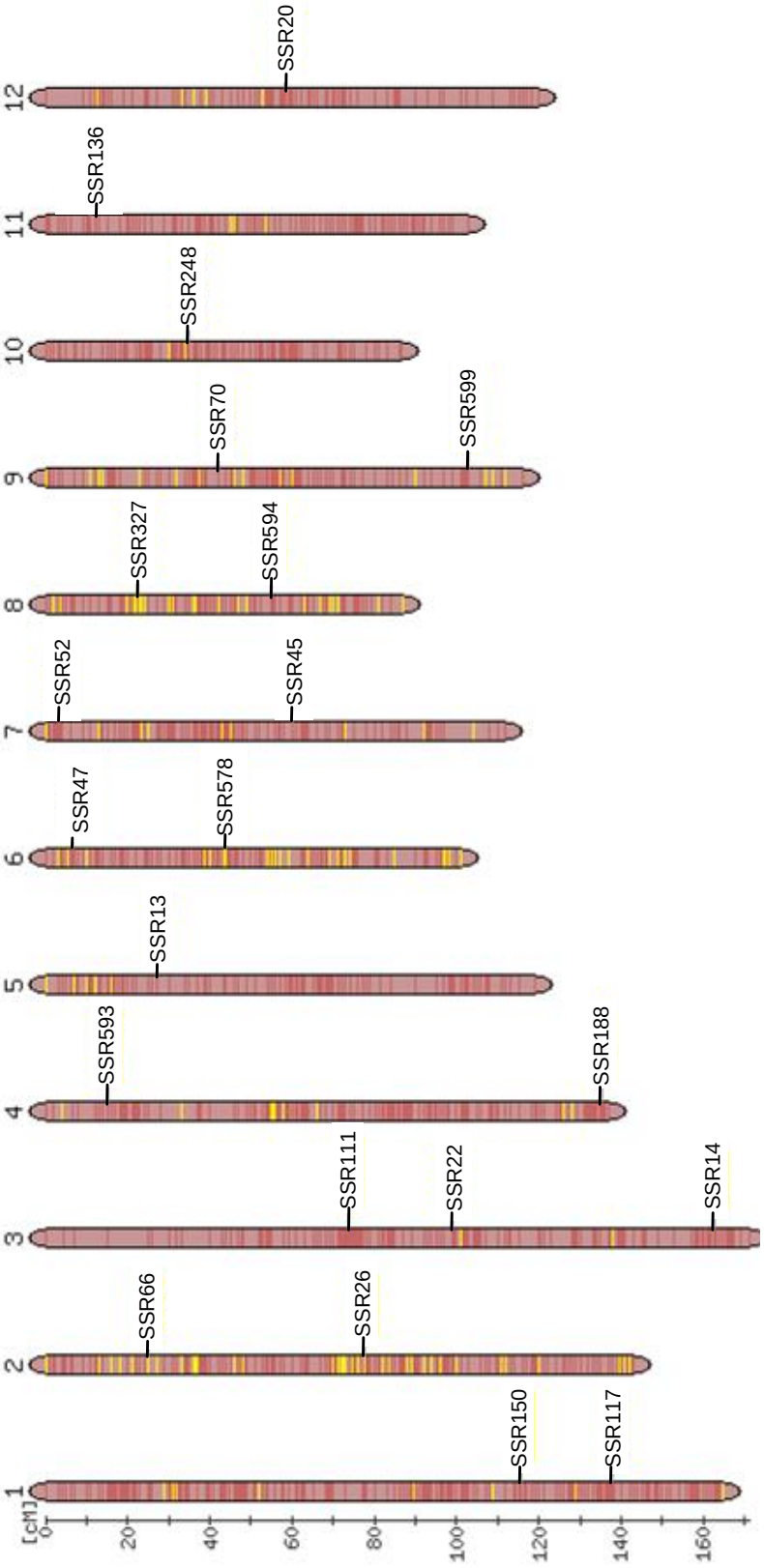


Figure 2-2. Carte génétique des marqueurs microsatellites utilisés dans l'étude de la structure de la diversité d'une collection de tomates cultivées et sauvages.
Les marqueurs sont placés sur la carte de référence de la tomate (carte Exp2000. <http://solgenomics.net/>.

2.3. Séquençage allélique et analyse des données

Afin de détecter du polymorphisme moléculaire, deux « core collections », de 24 et 96 individus, ont été utilisées. La core collection de 24 accessions a été utilisée pour séquencer des fragments de gènes candidats liés à la qualité du fruit qui co-localisent avec des QTL (Causse, Duffé et al. 2004).

Les polymorphismes identifiés par séquençage allélique dans le laboratoire ou dans la bibliographie (van Deynze, Stoffel et al. 2007) ont permis de construire 4 panels de 48 marqueurs chacun qui permettront de génotyper l'ensemble de la collection avec la technologie SNPlex® Genotyping System (Applied Biosystems, Foster City, Etats-Unis). La « core collection » de 96 individus a été utilisée afin d'identifier du polymorphisme sur un échantillon plus large de locus, dans le but d'étudier le déséquilibre de liaison sur le chromosome 2 et de réaliser des tests d'association préliminaires. Cette étude s'est focalisée sur le chromosome 2 car il présente de nombreux QTL liés à la qualité du fruit. Certains QTL du chromosome 2 sont en cours de cartographie fine et de clonage positionnel.

Etant donné la faible diversité moléculaire observée chez la tomate cultivée (Nesbitt and Tanksley 2002; Yang, Bai et al. 2004; Labate and Baldo 2005; van Deynze, Stoffel et al. 2007) nous nous sommes focalisés sur le séquençage de fragments non codant : introns et régions intergéniques. Nous avons pour cela utilisé un programme développé au sein de l'unité, CGIS (Bres, Bouchet et al. 2005). Ce programme utilise les informations de séquence des unigènes disponibles sur le site du Sol Genomics Network (http://solgenomics.net/). Il réalise un alignement de la séquence de l'unigène sur la séquence génomique d'*Arabidopsis thaliana* par TBLASTX. L'homologie entre *Arabidopsis* et la tomate est assez proche pour que la localisation des introns soit conservée (mais pas leurs longueurs). Le programme utilise Primer3 (Rozen and Skaletsky 2000) pour définir des amorces sur les exons (séquences conservées chez les différentes espèces) autour de la position prédite des introns qui seront séquencés préférentiellement. Primer3 a été utilisé pour définir des amorces dans les régions intergéniques.

La figure 2-3 décrit le fonctionnement du logiciel.

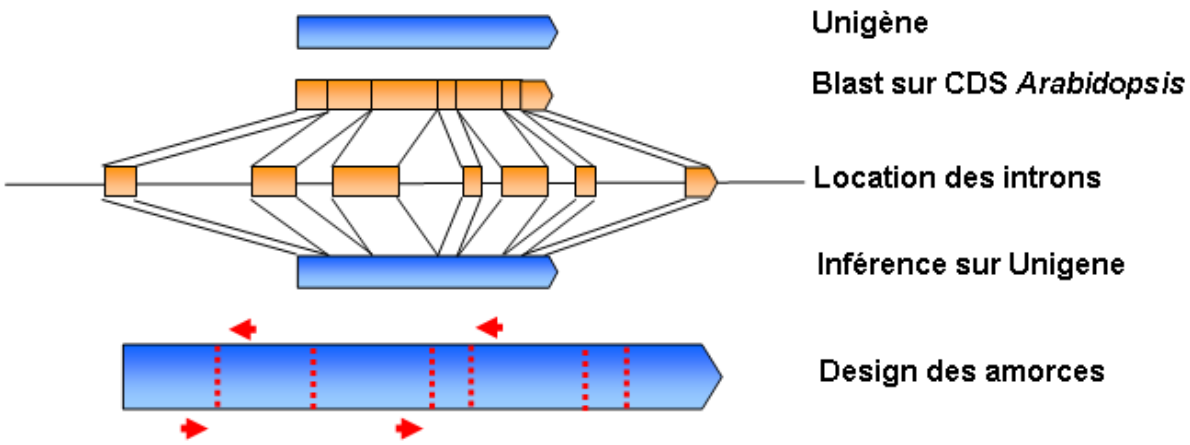


Figure 2-3. Principe de fonctionnement du programme CGIS utilisé pour définir des amorces sur les exons afin d'amplifier préférentiellement les introns des gènes. L'unigène de tomate (en bleue) est aligné par TBlastX sur la séquence génomique d'*A. thaliana*. Lorsque le CDS homologue (en orange) est identifié, la localisation des introns chez *A. thaliana* est inférée sur l'unigène de tomate. Primer3 est utilisé pour définir des amorces sur les exons afin d'amplifier spécifiquement les introns.

Les tests d'amplification par PCR ont été réalisés à Avignon. Seuls les marqueurs donnant des profils simple bande ont été séquencés. Le séquençage a été réalisé par l'unité « Etude du Polymorphisme des Génomes Végétaux » localisée au Centre National de Génotypage (CEA) à Evry. Une seule extrémité de chaque amplicon est séquencée. Pour chaque fragment, l'alignement des séquences, la vérification des chromatogrammes et la recherche des polymorphismes sont réalisés grâce au logiciel GENALYS2.0 (http://software.cng.fr/) (Takahashi, Matsuda et al. 2003).

Les séquences vérifiées ont été formatées sous le format fasta et les alignements ont été vérifiés par l'algorithme d'alignement MUSCLE (http://www.ebi.ac.uk/Tools/muscle/). Les alignements sont visualisés grâce à l'outil BIOEDIT 7.09 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) (Hall 1999).

Les simulations de coalescence permettant de tester la vraisemblance de différents modèles évolutifs ont été réalisées avec le logiciel BayeSSC qui est une modification du logiciel SIMCOAL 1.0 (Excoffier, Novembre et al. 2000). Le logiciel est disponible à l'adresse suivante : http://www.stanford.edu/group/hadlylab/ssc/index.html. La

vraisemblance de chaque modèle a ensuite été vérifiée par vraisemblance selon la méthode de Belle, Ramakrishnan et al. (2006).

Les données de diversité moléculaire qui ont permis l'implémentation du logiciel BayeSSC ont été obtenu avec le logiciel DNAsp 5.10 (http://www.ub.edu/dnasp/) (Librado and Rozas 2009) qui permet d'obtenir pour une séquence donnée les statistiques S le nombre de site polymorphe par fragment, H le nombre d'haplotype par fragment, π la diversité nucléotidique et D de Tajima (Ces statistiques sont présentées dans le chapitre 6).

2.4. Phénotypage

2.4.1. Conditions de culture

Les cultures des 201 accessions, phénotypées finement, ont été menées en 2007 et 2008. Les plantes ont été mises à germer dans des terrines au début du mois d'avril, repiquées en motte de 5 cm dix jours après le semis et plantées en pleine terre sous tunnel non chauffé (orienté Nord-Sud) à la fin du mois d'avril dans l'unité expérimentale de l'Unité GAFL. Quatre plantes par accessions ont été plantées à des densités de 2,6 plantes par m^2, conduites sous ferti-irrigation par goutte-à-goutte. Les plantes ont été tuteurées jusqu'à atteindre une hauteur de 2 m et les bourgeons axillaires ont été éliminés durant toute la période de croissance de la plante (sauf sur les plantes à croissance déterminée). Pour éviter des effets d'ombrages des plantes qui peuvent modifier la maturation des fruits, la ligne Est du tunnel a toujours été composée de lignées à croissance déterminée (*sp*). Un programme combiné de lutte intégré a été mis en place en prévention pour lutter contre des ravageurs de culture et pratiquement aucun moyen de lutte chimique n'a été utilisé.

2.4.2. Récolte des fruits

La récolte des fruits s'est effectuée deux jours après que l'appréciation du stade tournant ait été effectuée (les fruits à maturité étaient éliminés afin de ne pas récolter de fruits trop mûrs). Trois récoltes différentes ont été effectuées sur chaque plante à une semaine d'intervalle. Pour chaque récolte, cinq fruits étaient prélevés pour les accessions à gros fruits (supérieur à 80 g) et dix fruits ont été prélevés sur les accessions produisant des fruits petits à moyens (inférieur à 80 g).

2.4.3. Phénotypage

Des caractères morphologiques des fleurs ainsi que des fruits ont été notées. Pour les fleurs, nous avons réalisé une notation du nombre de pétales, de la longueur des pétales (en mm) et de la position du style par rapport au cône d'étamines sur un ensemble de dix fleurs réparties sur les quatre plantes représentant chaque accession.

A maturité, les mesures suivantes ont été réalisées :

- Poids moyen du lot de fruit (g)

- Fermeté, prise en deux points équatoriaux opposés sur chaque fruit avec un duromètre Durofel (http://www.setop.fr/). Cet appareil portable mesure la pression exercée par un piston sur une surface avec une échelle arbitraire (0 indique une résistance minimale 100 indique une résistance maximale)

- Colorimétrie, prise en deux points différents sur chaque fruit avec un colorimètre Konica Minolta CR300. Les coordonnées suivant les axes L, a et b sont données pour chaque mesure (Figure 2-4).

- La combinaison **L*** est la clarté, qui va de 0 (noir) à 100 (blanc).
- La composante **a*** représente la gamme de couleur sur l'axe rouge (valeur positive) → vert (négative) en passant par le blanc (0) si la clarté vaut 100.
- La composante **b*** représente la gamme de couleur sur l'axe jaune (valeur positive) → bleu (négative) en passant par le blanc (0) si la clarté vaut 100.

Les fruits ont ensuite été coupés transversalement afin de pouvoir compter le nombre de loges puis ils ont été broyés jusqu'à l'obtention d'une purée. Les purées ont été conservées à -20°C et ont été utilisées ensuite pour mesurer :

- le pH

- l'indice réfractométrique IR (ou degré brix) grâce à un réfractomètre digital (Palette PR 101). Cet indice correspond à la teneur en solides solubles qui est corrélée avec la teneur en sucres réducteurs et acides organiques.

- l'acidité titrable mesurée par la quantité de NaOH (en mmol/L) nécessaire pour ramener le pH d'une solution à 8,1 (zone de virage de la phénolphtaléine). Les mesures sont réalisées grâce à un titrateur équipé d'un passeur d'échantillons (Crinson compact titrator). Une masse m_p de pulpe est pesée, diluée dans 50 mL d'eau et on

rajoute la soude jusqu'au pH seuil. L'acidité titrable est ensuite calculée de la façon suivante : $C = [(C_{NaOH} \times V_{NaOH}) \times 100] / m_p$

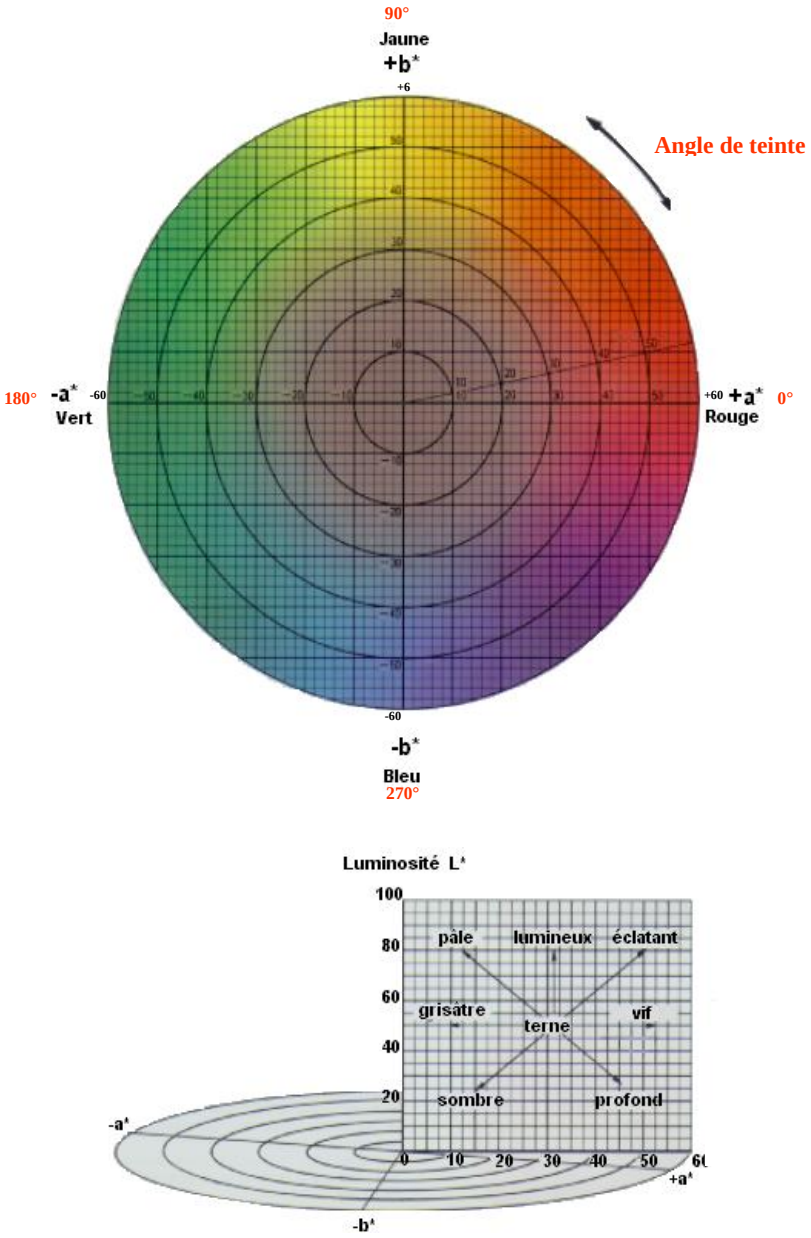


Figure 2-4. Espace couleur L a b.
a varie entre le vert et le rouge, b entre le bleu et le jaune et L représente la clarté qui varie de 0 à 100 (sombre au clair).

Des lots indépendants de fruits récoltés à maturité ont été soigneusement débarrassés des graines et du gel puis ils ont été broyés. Une partie de ces échantillons a été utilisée dans l'équipe d'Alisdair Fernie (Max-Planck-Institut für molekulare Pflanzenphysiologie, Potsdam, Allemagne) afin d'établir le profil métabolomique des 201 accessions. Un dosage relatif d'une cinquantaine de métabolites primaires et secondaires est réalisé par chromatographie gazeuse couplée à un spectromètre de masse (GC-MS). Seule une partie de ces données a été utilisée

au cour de cette thèse afin d'identifier des corrélations entre contenus en métabolites primaires et les caractères de contenu en solides solubles et acidité titrable.

2.5. Analyses statistiques et tests d'association

Les analyses des données phénotypiques ont toutes été réalisées avec le programme R (R Development Core Team 2005). Les coefficients de corrélations de rang de Spearman ont été calculés pour toutes les variables deux à deux. L'effet génotypique et environnemental pour chaque phénotype a été calculé par ANOVA en utilisant la fonction « glm » implémentée dans R, selon le modèle linéaire suivant : $Y_{ij} = \mu + \alpha_i + \beta_j + \alpha_i.\beta_j + \varepsilon$ où Y_{ij} représente la moyenne de l'accession *i* durant l'année *j*, μ la moyenne de la population, α_i l'effet du génotype *i* (effet génétique), β_j l'effet de l'année *j*, $\alpha_i.\beta_j$ l'interaction entre les génotypes et les années et ε l'erreur résiduelle.

Les héritabilités au sens large ont été calculées avec la formule suivante : $h_F^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2 / 2)$ où σ_g^2 et σ_e^2 représentent respectivement la variance génétique et résiduelle. σ_g^2 et σ_e^2 ont été estimées respectivement par (MSc-MSe)/2 et MSe. MSc et MSe sont l'estimation des carrés moyens des cultivars et de la résiduelle.

Pour les tests d'associations j'ai utilisé principalement le logiciel TASSEL 2.1 (http://www.maizegenetics.net/) (Bradbury, Zhang et al. 2007). Ce logiciel implémente de nombreux outils pour réaliser des études de diversité et des tests d'associations. Ce logiciel permet d'étudier le déséquilibre de liaison entre marqueurs moléculaires et dans une séquence génomique.

Plusieurs modèles ont été utilisés pour rechercher des associations entre les marqueurs et les phénotypes d'intérêt :

- Le modèle de Structured Association (SA) proposé par Pritchard et Donnely (2001) et utilisé sur maïs par Thornsberry et al. (2001).
- Le modèle linéaire généralisé prenant en compte la structure génétique (modèle Q).
- Le modèle linéaire mixte prenant en compte la structure génétique ainsi que le lien de parenté entre individus proposés par Yu et al. (2006) (modèle Q+K).

Le modèle Structured Association a été développé initialement pour des études en génétique humaine du type « case-control » et a été adapté à des études s'intéressant plus particulièrement à des variations quantitatives. Etant donné que certains caractères ne suivent pas une distribution de type gaussienne (tel que le nombre de loge), il semble intéressant d'intégrer ce modèle pour nos analyses car il peut être utilisé même si les variables ne suivent pas une loi normale. Il utilise un ratio de vraisemblance Λ qui compare la probabilité d'une hypothèse nulle (H0), dans laquelle les polymorphismes candidats sont indépendants du phénotype, et la probabilité d'une hypothèse alternative (H1), où les polymorphismes candidats sont associés au phénotype :

$$\Lambda = \frac{\Pr_0(C;T;\hat{Q})}{\Pr_1(C;\hat{Q})}$$

Dans cette équation, C représente la matrice génotypique ; T représente la matrice phénotypique et Q représente la matrice de structure de la population estimée par le logiciel STRUCTURE 2.1. Les probabilités correspondant aux deux hypothèses sont calculées par régression logistique avec pour variable cible le polymorphisme candidat et comme variable indépendante T et Q. Des permutations sont effectuées de manière à créer des associations aléatoires auxquelles sont confrontées les données réelles afin de calculer une *p*-value pour chaque association possible.

Le modèle linéaire généralisé à effets fixes (ou generalized linear model, GLM) intègre comme covariable l'information d'appartenance de chaque génotype à une sous-population, estimée par le logiciel STRUCTURE 2.1. Le modèle utilisé s'écrit alors $T = C + Q + \varepsilon$ où T, C, Q et ε représentent respectivement la valeur de caractère, le génotype au polymorphisme candidat, la matrice de structure et l'erreur résiduelle. Le seuil de significativité est estimé par permutations.

Le modèle linéaire mixte (ou mixed linear model, MLM) a été développé car une perte de puissance du modèle linéaire généralisé était due à la non prise en compte de l'apparentement entre individus (Yu, Pressoir et al. 2006). Le modèle MLM utilise une matrice d'estimation des apparentements entre individus (matrice K) comme variable dans le modèle. Le modèle utilisé, s'écrit : $\mathbf{y} = X\boldsymbol{\beta} + S\boldsymbol{\alpha} + Q\mathbf{v} + Z\mathbf{u} + \mathbf{e}$.

Le vecteur $\mathbf{y}$ représente les valeurs phénotypiques. Tous les effets fixes autres que les polymorphismes testés et la structure génétique sont représentés dans le vecteur $\boldsymbol{\beta}$. Le vecteur

$\boldsymbol{\alpha}$ est le vecteur des effets des polymorphismes, $\mathbf{v}$ est un vecteur de l'effet de la structure de la population et **u** est un vecteur de l'effet de divers QTL présents dans le fond génétique des individus. La matrice Q est définie par STRUCTURE 2.1 et elle relie **y** à Q et X. Les matrices S et Z sont des matrices d'incidence composées de 1 et de 0 reliant respectivement **y** à $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ et **u**. Le vecteur **e** représente l'erreur résiduelle. La variance des effets aléatoires est supposée être $\mathrm{Var}(\mathbf{u}) = 2KV_g$ où V_g représente la variance génétique et K est une matrice (de dimension $n \times n$, n étant le nombre d'individus testés) des coefficients de parenté, qui définissent le degré de covariance génétique entre paires d'individus (matrice kinship, K). Cette matrice est calculée selon la méthode de Ritland (1996), implémentée dans le logiciel SPAGeDi (http://ebe.ulb.ac.be/ebe/Software.html) (Hardy and Vekemans 2002). Les valeurs négatives entre les individus sont redéfinies à 0. Les composantes de la variance ont été calculées avec la méthode EMMA (Efficient Mixed Model Association) développée par Kang, Zaitlen et al. (2008) et implémentée dans TASSEL 2.1. Pour contrôler l'inflation du taux de faux positifs détectés (False Discovery Rate, FDR) due aux tests multiples, nous avons corrigé les *p*-value associés à chaque polymorphisme par la procédure de Benjamini et Hochberg (2000).

Chapitre 3 : Analyse de la structure de la diversité d'une collection de tomates sauvages et cultivées

3.1. Introduction

La conservation et la caractérisation des ressources génétiques peuvent être valorisées par l'utilisation de la diversité naturelle en vue de disséquer les bases moléculaires de caractères d'intérêt. Des études d'associations permettent d'utiliser ces ressources qui offrent de nombreux avantages comme une résolution supérieures et un gain de temps (population directement utilisable). Cependant, ces études présentent aussi des difficultés supplémentaires notamment liées au fait que certaines associations peuvent être en réalité des faux positifs. La source d'erreur la plus importante est due au déséquilibre de liaison qui peut exister entre des marqueurs non liés. Ce DL peut être causé par la structure génétique de l'échantillon. Il s'avère donc important d'étudier cette structuration de la diversité afin que l'information décrite puisse être incorporée directement dans les tests d'associations.

La tomate est une espèce majoritairement autogame ce qui implique qu'une structure importante doit exister dans l'échantillon de ressources génétiques comprenant à la fois des accessions sauvages et cultivées. Chez cette espèce, les études de diversité sont limitées à des études descriptives démontrant l'apport des marqueurs moléculaires dans la distinction entre cultivars (Smulders, Bredemeijer et al. 1997; Areshchenkova and Ganal 1999; Bredemeijer, Cooke et al. 2002) et à des études de phylogénies interspécifiques (Alvarez, Wiel et al. 2001; Spooner, Peralta et al. 2005). Les seules études réelles de structuration de la diversité de collection de tomate ont été réalisées à l'aide d'un panel restreint d'isozymes ou à l'aide de marqueurs RFLP (Rick 1958; Rick and Fobes 1975; Rick, Fobes et al. 1977; Miller and Tanksley 1990; Rick and Holle 1990). Une autre étude utilisant des marqueurs AFLP s'est consacré à l'étude de variétés modernes (van Berloo, Zhu et al. 2008). Malgré le régime de reproduction de la tomate, ces études ont mis en évidence des taux non négligeables d'hybridation dans des zones où la tomate cultivée (*S. lycopersicum*) et l'espèce sauvage apparentée (*S. pimpinellifolium*) se retrouvent en sympatrie. Cette observation a été confirmée par l'étude de Nesbitt et Tanksley (2002) sur la diversité moléculaire du locus fw2.2 qui

montrent que ce locus chez la tomate cerise (*S. l. cerasiforme*) est une mosaïque des génomes de l'espèce cultivée à gros fruit (*S. l. esculentum*) et de l'espèce *S. pimpinellifolium*.

Nous avons donc voulu vérifier cette observation au niveau du génome entier en génotypant 21 marqueurs microsatellites sur 360 accessions décrites dans le chapitre Matériel et Méthodes. De plus, il n'était pas envisageable de réaliser la détection du polymorphisme moléculaire (qui sera utilisé pour les tests d'associations dans les chapitres suivants) par séquençage sur l'ensemble de la collection. Nous avons donc utilisé les informations génétiques ainsi que les informations de phénotypage de l'échantillon réduit (201 accessions, année 2007 seulement) pour construire des « core collections » emboîtées. Des accessions des deux types cultivés (*S. l. esculentum* et *S. l. cerasiforme*) et du type sauvage (*S. pimpinellifolium*) constituent ces « core collections ». Ces échantillons réduits, maximisent la diversité génétique et morphologique. Ils peuvent être combinés à souhait et représentent un panel de référence disponible pour la communauté scientifique. Cette étude a été publiée dans la revue *BMC Plant Biology*.

3.2. A clarified position for *Solanum lycopersicum* var. *cerasiforme* in the evolutionary history of tomatoes (*Solanaceae*)

BMC Plant Biology

BioMed Central

Research article

Open Access

A clarified position for *solanum lycopersicum* var. *cerasiforme* in the evolutionary history of tomatoes (solanaceae)

Nicolas Ranc[1], Stéphane Muños[1], Sylvain Santoni[2] and Mathilde Causse*[1]

Address: [1]INRA, UR1052, Unité de Génétique et d'Amélioration des Fruits et Légumes, Montfavet 84 143, France and [2]INRA, UMR 1097 Diversité et Adaptation des Plantes Cultivées, Montpellier 34602, France

Email: Nicolas Ranc - Nicolas.Ranc@avignon.inra.fr; Stéphane Muños - Stephane.Munos@avignon.inra.fr; Sylvain Santoni - Sylvain.Santoni@supagro.inra.fr; Mathilde Causse* - Mathilde.Causse@avignon.inra.fr

* Corresponding author

Published: 20 December 2008

BMC Plant Biology 2008, 8:130 doi:10.1186/1471-2229-8-130

Received: 7 August 2008
Accepted: 20 December 2008

This article is available from: http://www.biomedcentral.com/1471-2229/8/130

Abstract

Background: The natural phenotypic variability present in the germplasm of cultivated plants can be linked to molecular polymorphisms using association genetics. However it is necessary to consider the genetic structure of the germplasm used to avoid false association. The knowledge of genetic structure of plant populations can help in inferring plant evolutionary history. In this context, we genotyped 360 wild, feral and cultivated accessions with 20 simple sequence repeat markers and investigated the extent and structure of the genetic variation. The study focused on the red fruited tomato clade involved in the domestication of tomato and confirmed the admixture status of cherry tomatoes (*Solanum lycopersicum* var. *cerasiforme*). We used a nested sample strategy to set-up core collection maximizing the genetic diversity with a minimum of individuals.

Results: Molecular diversity was considerably lower in *S. lycopersicum* i.e. the domesticated form. Model-based analysis showed that the 144 *S. lycopersicum* var. *cerasiforme* accessions were structured into two groups: one close to the domesticated group and one resulting from the admixture of the *S. lycopersicum* and *S. pimpinellifolium* genomes. SSR genotyping also indicates that domesticated and wild tomatoes have evolved as a species complex with intensive level of hybridization. We compiled genotypic and phenotypic data to identify sub-samples of 8, 24, 32 and 64 cherry tomato accessions that captured most of the genetic and morphological diversity present in the entire *S. lycopersicum* var. *cerasiforme* collection.

Conclusion: The extent and structure of allelic variation is discussed in relation to historical events like domestication and modern selection. The potential use of the admixed group of *S. lycopersicum* var. *cerasiforme* for association genetics studies is also discussed. Nested core collections sampled to represent tomato diversity will be useful in diversity studies. Molecular and phenotypic variability of these core collections is defined. These collections are available for the scientific community and can be used as standardized panels for coordinating efforts on identifying novel interesting genes and on examining the domestication process in more detail.

BMC Plant Biology 2008, **8**:130 http://www.biomedcentral.com/1471-2229/8/130

Background

Advances in molecular marker development and in genome mapping have resulted in high-density molecular-marker linkage maps in crops, and have provided tools for dissecting the genetic variation of complex traits. Map-based strategies were successfully used for the positional cloning of genes that underlie Quantitative Trait Loci (QTL) [1-3]. Despite the success of these strategies, gene discovery is still limited to those loci that have large effects upon quantitative variation [4].

Over the last few years, there has been renewed interest in the study of naturally occurring variation in crop genetic collections. Motivations for such studies are (i) to use natural allelic diversity for the evaluation of gene function, (ii) to find new genes or new alleles involved in specific aspects of plant physiology or development and (iii) to try to understand the molecular basis of adaptation to local environments [5]. Association genetics or linkage disequilibrium studies test for a statistical association between genotypes at a marker locus and the phenotypes in a set of unrelated individuals [6]. Polymorphisms of interest are detected in a large range of genetic backgrounds. The extent of linkage disequilibrium (LD), the non-random association of alleles at two or more loci, is a sample specific property and depends on the biological model studied. In contrast to the situation in multigenerational pedigrees, LD in natural populations is not broken artificially and we need to overcome this restriction.

The primary obstacle to successful association studies or linkage disequilibrium (LD) mapping is the nature of the genetic structure of populations [7]. The presence of subgroups with different allele frequencies, within the population studied, can lead to spurious associations. Domestication of most of modern crops occurred between 10,000 and 5,000 years ago and shaped the allelic frequencies distribution among plant populations. Knowledge about genetic structure can aid in inference of evolutionary history like domestication [8].

The large sample size to be analyzed constitutes another constraint in diversity studies, whereas studying a subset might be more efficient if this sample spans the full range of variation [9]. The first challenge in molecular diversity analysis is thus to sample core collections that better fit the range of morphological and genetic variations found in the global collection. For example, Single Nucleotide Polymorphism (SNP) candidate markers, discovered in a small number of accessions, can be easily genotyped on a larger sample for diversity analysis and association mapping. Several methods have been proposed for constructing core-collections. Some of these take advantage of molecular markers [10] and seem to perform better when used for sampling autogamous plants [11]. The genetic structure of a core collection has to be checked to avoid spurious correlation between molecular polymorphisms and traits in association studies.

Tomato (*Solanum lycopersicum*, formerly *Lycopersicon esculentum*) emerged as a model species for the study of fleshy fruited plants because of the extent of genetic and genomic resources available [12,13]. The large range of phenotypic variation and large collections of genetic resources available for crops are prerequisites for using an association strategy. The cultivated tomato is highly autogamous and shows a large range of morphological diversity but low genetic diversity compared to other *Solanum* relatives [14]. This can be explained by successive bottlenecks: (i) domestication associated with isolation of the crop from the Andes (centre of diversity) to Central America, (ii) transfer of few cultivars to the Mediterranean basin by conquistadors in the 16th century and (iii) modern breeding [15]. Cherry tomato, i.e. *S. lycopersicum* var. *cerasiforme* (*S. l. cerasiforme*), is the expected ancestor of the domesticated form. In its native Andean region, wild and feral forms can be found and *S. l. cerasiforme* is also described as highly invasive [16]. Cherry tomato accessions are also found as landraces from temperate to subartic regions. In Coastal Ecuador and Peru, *S. pimpinellifolium*, genetically close to *S. lycopersicum* and strictly wild, is found growing in sympatry with tomato landraces and cherry tomato (and also with *S. peruvianum* and *S. hirsutum*, two green-fruited species). Wild and feral *S. l. cerasiforme* (i.e. cherry type) exhibit two allozyme-diversity patterns: one similar to the allozyme-diversity pattern exhibited by cultivated tomato and another one similar to the wild *S. pimpinellifolium* allozyme-diversity pattern [17]. Based on isozymes, *S. l. cerasiforme* accessions also show an outcrossing rate comparable to the rate of outbred species [18]. Rick and Holle (1990) suggest that tomato should have undergone natural introgressions from wild and feral accessions. Moreover, Nesbitt and Tanksley [19] demonstrated that, around the *fw2.2* locus, the *S. l. cerasiforme* genome is a mosaic between *S. lycopersicum* and *S. pimpinellifolium* genomes due to frequent hybridizations between the two species. This is evidence of frequent hybridizations in this autogamous complex of species. The admixture hypothesis of *S. l. cerasiforme* has never been tested on the whole genome and would be further evidence of a natural high rate of hybridization. Moreover, *S. l. cerasiforme* and *S. pimpinellifolium* are involved in the domestication of tomato but the process remains to be clarified.

Molecular markers like simple sequence repeat (SSR) markers have often been used to clarify genetic structure in plants [20-25]. In tomato several studies used SSR markers but focused only on wild relatives [26,27] or on elite germplasm [28,29]. No study used a broad sample of

BMC Plant Biology 2008, **8**:130 http://www.biomedcentral.com/1471-2229/8/130

cultivated, landraces, and wild accessions. The goal of the present study is to clarify the domestication process of tomato and to confirm the admixture status of *S. l. cerasiforme*. To achieve this goal we analyzed the genetic structure of a genetic resource collection, that includes predominantly *S. l. cerasiforme* accessions, and we compared this to the genetic structure of *S. lycopersicum* and *S. pimpinellifolium*. We assessed the amount of genetic diversity in the collection and sampled nested core collections of wild and cultivated tomato that will be used in future diversity studies. For this purpose we used a set of 20 SSR markers dispersed over the genome to survey the genetic diversity present in a sample of 360 accessions.

Results

Microsatellite diversity

The Microsatellite markers used (table 1) revealed different diversity patterns in the total collection including green fruited species, *S. cheesmaniae* (N = 20) and red-fruited accessions (N = 340) (table 2). SSR markers revealed 2 to 26 different alleles and an average of 12.45 alleles per locus. This mean dropped to 3 alleles per locus when rare alleles (i.e. with a frequency lower than 0.05) were removed. In the red-fruited tomatoes group, the average allele number per locus was N_A = 7.7 but was equivalent to the total collection when removing rare alleles (N_A = 3.3). The average expected heterozygosity over all loci was 0.496 with large variation among loci (SD = 0.225). Rare heterozygous genotypes were found for all loci in the total collection ($H_O > 0$) but were distributed across individuals.

A much higher genetic diversity was found in wild *S. pimpinellifolium* (H_E = 0.58) than in the cultivated *S. lycopersicum* (H_E = 0.25) (table 3). The observed heterozygosity was also higher for *S. pimpinellifolium* (H_O = 0.0591) than for *S. lycopersicon* (H_O = 0.0098). The reason for these heterozygosity patterns could be the difference in the reproductive regime between *S. pimpinellifolium* accessions and *S. lycopersicum*. The *S. l. cerasiforme* exhibited an intermediate pattern of diversity.

Genetic structure of the sample

The genetic structure in the red-fruited accession sample was analyzed with the model-based clustering algorithm implemented in the Structure2.0 software (see Methods section for details). To avoid redundancy in the collection, we kept only one individual when several accessions were identified with the same SSR fingerprint at all loci. Hence, 23 individuals (18 *S. lycopersicum* and 5 *S. pimpinellifolium*) were removed. Thus, we detected the genetic structure of a sample of 318 accessions. Because *S. l. cerasiforme* genome was described as a mosaic between *S. lycopersicum* and *S. pimpinellifolium* genomes, all the red-fruited accessions were used as a broad sample. *S. cheesmaniae* and *S. galapagense* accessions have not taken part in the domestication process of tomato and were not included in this analysis.

The Evanno et al. (2005) correction of the Structure2.0 outputs was used (Figure 1). The first peak of ΔK, for K = 2, corresponded to the presence of two main clusters and a potential sublevel of clustering was suggested by the secondary peak of ΔK, for K = 4. The classification of accessions into clusters by the model-based method was used to study the sublevel clustering of the red-fruited tomato sample. For all K_{opt}, memberships were consistent between all runs.

For K_{opt} = 2, clustering divided the total sample into two groups. Group 1 consisted of the main part of *S. pimpinellifolium* (Table 4) with 20 accessions from *S. l. cerasiforme* whereas group 2 consisted of the main part of S. *lycopersicum* and of the *S. l. cerasiforme* samples. Group 1 represented the 'wild' part of the sample whereas group 2 represented the 'domesticated' part of the sample. This classification accounted for 35% ($p < 0.000001$) of the total genetic variance; individuals within group accounted for 51% ($p < 0.000001$) of the total variance and the variance within individuals explained five percent ($p < 0.000001$) of the total variance. When individuals were assigned with a minimal membership of 70% into a corresponding cluster, twenty three percent (i.e. 35 individuals) of the *S. l. cerasiforme* accessions was in admixture between 'wild' and 'domesticated' groups.

For K_{opt} = 4, the group 1 divided into subgroups A and B and the group 2 divided into subgroups C and D. When individuals with a membership lower than 70% were not taken into account, the hierarchical AMOVA indicated that 37% ($p < 0.000001$) of the variance was due to variation among groups, 13% ($p < 0.000001$) of the variance was due to variation among subgroup within groups and 45% ($p < 0.000001$) of the variance was due to variation among individuals within subgroup (only five percent ($p < 0.000001$) was due to variation within individual). Pairwise estimates of F_{ST} indicated a high degree of differentiation between the four clusters with values ranging from 0.21 between clusters C and D to 0.64 between clusters A and D (Table 5).

The cluster A consisted of moderate to large fruited individuals with a large part of *S. lycopersicum* accessions, whereas cluster B consisted of small fruited accessions with the cherry type accessions representing the main part of this subgroup. The 'wild' group was divided into the cluster A and B; both consisted of *S. l. cerasiforme* and *S. pimpinellifolium* accessions. When individuals were assigned with a minimal membership of 70% into a corresponding cluster, individuals were found in admixture

BMC Plant Biology 2008, **8**:130 http://www.biomedcentral.com/1471-2229/8/130

Table 1: Characteristics of microsatellite loci

Locus name	Motif	Linkage group[a]	Map position[a] (cM)	Primer sequences (5'- 3')
SSR599	$[TCATTA]_2[TCA]_6$	9	103.00	GGATTTCTCATGGAGAATCAGTC TCCCTTGATCTTGATGATGTTG
SSR111	$[TC]_6[TCTG]_6$	3	73.90	TTCTTCCCTTCCATCAGTTCT TTTGCTGCTATACTGCTGACA
SSR14	$[ATA]_9$	3	162.50	TCTGCATCTGGTGAAGCAAG CTGGATTGCCTGGTTGATTT
SSR248	$[TA]_{21}$	10	35.00	GCATTCGCTGTAGCTCGTTT GGGAGCTTCATCATAGTAACG
SSR52	$[AAC]_9$	7	3.00	TGATGGCAGCATCGTAGAAG GGTGCGAAGGGATTTACAGA
SSR150	$[CTT]_7$	1	115.50	ATGCCTCGCTACCTCCTCTT AATCGTTCGTTCACAAACCC
SSR117	$[TC]_{11}$	1	138.00	AATTCACCTTTCTTCCGTCG GCCCTCGAATCTGGTAGCTT
SSR66	$[ATA]_8$	2	25.00	TGCAACAACTGGATAGGTCG TGGATGAAACGGATGTTGAA
SSR136	$[CAG]_7$	11	11.00	GAAACCGCCTCTTTCACTTG CAGCAATGATTCCAGCGATA
SSR578	$[AAC]_6[ATC]_5$	6	44.00	ATTCCCAGCACAACCAGACT GTTGGTGGATGAAATTTGTG
SSR47	$[AT]_{14}$	6	6.50	TCCTCAAGAAATGAAGCTCTGA CCTTGGAGATAACAACCACAA
SSR594	$[TCT]_8$	8	55.00	TTCGTTGAAGAAGATGATGGTC CAAAGAGAACAAGCATCCAAGA
SSR22	$[AT]_{11}$	3	99.00	GATCGGCAGTAGGTGCTCTC CAAGAAACACCCATATCCGC
SSR327	$[AAT]_7$	8	22.50	TCAGGATCAGGAGCAGGAGT TGGACTTGTTCCATGAACCC
SSR593	$[TAC]_7$	4	15.00	TGGCATGAACAACAACCAAT AGGAAGTTGCATTAGGCCAT
SSR26	$[CGG]_7$	2	77.50	CGCCTATCGATACCACCACT ATTGATCCGTTTGGTTCTGC
SSR45	$[AAT]_{14}$	7	80.00	TGTATCCTGGTGGACCAATG TCCAAGTATCAGGCACACCA
SSR20	$[GAA]_8$	12	37.00	GAGGACGACAACAACAACGA GACATGCCACTTAGATCCACAA

BMC Plant Biology 2008, **8**:130 http://www.biomedcentral.com/1471-2229/8/130

Table 1: Characteristics of microsatellite loci *(Continued)*

SSR70	$[AT]_{20}$	9	42.00	TTTAGGGTGTCTGTGGGTCC GGAGTGCGCAGAGGATAGAG
SSR188	$[AT]_{11}$	4	135.50	TGCAGTGAGTCTCGATTTGC GGTCTCATTGCAGATAGGGC

[a] Linkage group and map position are based on the tomato EXPEN 2000 map http://www.sgn.cornell.edu/.

Table 2: Microsatellite diversity detected in the total collection and in the red fruited subgroup

Locus name	N_A [a]		$N_{A,P}$ [b]		H_E [c]		H_O [d]	
	total	red-fruited	total	red-fruited	total	red-fruited	total	red-fruited
SSR599	6	4	1	1	0.117	0.023	0.011	0.002
SSR111	14	8	4	5	0.654	0.615	0.033	0.029
SSR14	11	9	4	4	0.621	0.603	0.036	0.035
SSR248	25	18	7	8	0.899	0.888	0.067	0.050
SSR52	5	2	1	1	0.070	0.012	0.003	0.000
SSR150	10	6	2	2	0.294	0.220	0.022	0.021
SSR117	13	6	3	3	0.533	0.478	0.050	0.038
SSR66	8	4	3	3	0.421	0.363	0.017	0.024
SSR136	8	6	4	3	0.457	0.396	0.036	0.029
SSR578	6	2	2	2	0.372	0.309	0.008	0.009
SSR47	26	25	3	3	0.725	0.710	0.046	0.048
SSR594	13	5	2	3	0.517	0.463	0.042	0.032
SSR22	17	9	2	3	0.580	0.532	0.061	0.047
SSR327	14	5	2	2	0.275	0.219	0.039	0.021
SSR593	9	6	2	5	0.574	0.537	0.047	0.038
SSR26	2	2	2	2	0.452	0.425	0.017	0.018
SSR45	20	14	5	5	0.795	0.776	0.081	0.062
SSR20	4	3	3	3	0.334	0.281	0.031	0.021
SSR70	21	17	5	5	0.848	0.832	0.069	0.065
SSR188	17	3	3	3	0.386	0.324	0.025	0.015
Mean (SD)	12.45	7.7	3	3.3	0.496 (0.225)	0.4503 (0.2442)	0.037 (0.022)	0.0303 (0.0178)

[a] number of allele per locus ([b] number of allele with frequency higher than 5%), [c] expected heterozygosity, [d] observed heterozygosity

BMC Plant Biology 2008, **8**:130 http://www.biomedcentral.com/1471-2229/8/130

Table 3: Pattern of genetic diversity inferred from simple sequence repeat markers among tomato species.

sample	number of individual	Total number of alleles	specific allele number[a]	H_E [b]	H_O [c]
S. lycopersicum	130	88	6	0.2479	0.0098
S. lycopersicum var.*cerasiforme*	144	99	6	0.3816	0.0370
S. pimpinellifolium	66	130	13	0.5781	0.0591
total red-fruited sample	340	154	-	0.4503	0.0300

[a] specific alleles are identified when they are found only in one sample. [b] expected heterozygosity, [c] observed heterozygosity

between intra-specific groups but most admixture accessions were inter-specific admixes (Table 6).

Groups 1 and 2 were considered as main samples and analyzed separately using the same hypothesis. The optimum number of sublevel populations within the groups 1 and 2 was two, which is consistent with the K_{opt} of 4 for the whole sample. Classification of individuals in each cluster was consistent with results based on Structure2.0 outputs of the total sample. For $K_{opt} = 4$, there were differences between individual's memberships and species classification (Figure 2). Some individuals were misclassified.

We also analyzed the genetic structure of each species separately (see Additional file 1: Determination of Kopt for each species) and the memberships of individuals was consistent with clustering found in the whole red-fruited tomato sample. Individuals previously found in admixture clustered in independent groups.

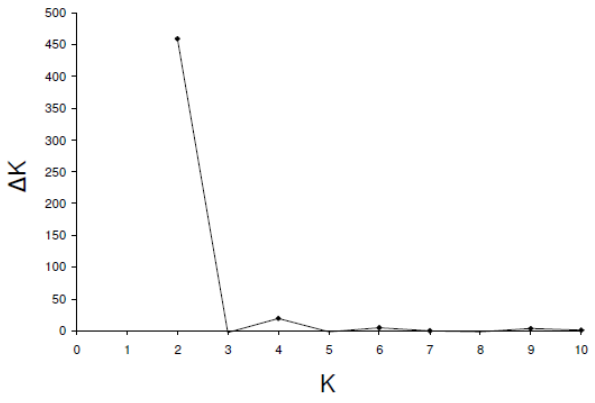


Figure 1
Determination of Kopt following the method of Evanno et al. (2005). The rate of change of the posterior probability of the data given the number of clusters is plotted against K, the number of clusters. ΔK was calculated as |L"(K)|/s[Pr(x|k)] (see Materials and Methods). The first peak (K = 2) corresponds to the optimum number of clusters. The secondary peak (K = 4) indicates a sublevel clustering.

The pattern of genetic diversity within the subdivision was analyzed (Table 6). The two 'wild' clusters presented the highest H_E but subgroup A had a low value of H_E compared to subgroup B. The numbers of statistical pairwise comparisons for non random association of alleles (Table 6) are homogeneous among subgroup A, C and D but much higher for subgroup B and for the 'wild' and 'domesticated' admixed part of *S. l. cerasiforme*. The clustering allowed linkage disequilibrium to decrease in each subgroup compared to the whole sample.

The first axis of Principal Coordinate Analysis of the red-fruited tomatoes separated 'wild' *S. pimpinellifolium* from 'domesticated' *S. lycopersicum* (Figure 3). The second axis separated subgroups A and B on one hand and subgroups C and D on the other hand. The *S. l. cerasiforme* accessions were divided among subgroups B and the admixed cluster. The interspecific admixed group showed a *continuum* between 'wild' and 'domesticated' clusters.

Sampling of the Core collection

Core collections of *S. l. cerasiforme* accessions were built using the Maximization or M strategy algorithm implemented in MStrat software v.4.1. Analyses were first performed on all cherry tomato accessions only (144 accessions). Before sampling the core collections, the whole sample was analyzed to compare two sampling strategies. We also determined the size of the smallest subset that captured all molecular and phenotypic alleles present in the whole sample. Both molecular and phenotypic data were used for these analyses. The phenotypic quantitative variables were split into 5 classes of equal dimension (see Methods). Random and M sampling strategies were compared. SSR allelic richness (number of alleles captured if sampling a core collection of n individuals) was calculated for each core collection size. The 20 SSR alleles were used both as markers, to implement the M and random strategy, and target variables, to compare these two strategies (Figure 4a). The difference between the random and M curves indicated that the M strategy performed better in sampling a core collection for the *S. l. cerasiforme* sample. The optimal size for the core collection, obtained at the plateau of the M curve, was reached for 37 *S. l. cerasiforme* accessions.

BMC Plant Biology 2008, **8**:130 http://www.biomedcentral.com/1471-2229/8/130

Table 4: Species classification among clusters described by Structure2.0 based on maximal individual-membership for each cluster.

species	K_{opt} = 2		K_{opt} = 4			
	Pop1	Pop2	popA	PopB	PopC	popD
S. lycopersicum	1	112	1	0	23	89
S. lycopersicum var. *cerasiforme*	20	124	13	10	78	43
S. pimpinellifolium	58	3	21	35	3	2
Number of pairs of loci in LD (# of comparisons)[a]	158(189)	87(189)	62(152)	143(189)	74(135)	58(189)

[a] Pairs of markers were considered in significant LD using the threshold p-value < 0.001.

The phenotypic diversity captured when sampling only with SSR alleles is shown in figure 4b. The plateau of the M curve was reached for 51 individuals and a weak difference in performance between the two strategies was observed.

When both molecular and phenotypic data were used as marker variables (i.e. to sample the core collection), the M strategy showed higher performance in sampling procedure than a random strategy and gave an optimal size of 51 individuals (figure 4c). Finally, core collections were sampled using both molecular and morphological data. To define the final core collection, accessions were classified by the number of times they were sampled in the fifteen replicates and the most frequently sampled accessions were chosen.

Four nested core collections composed of 8, 24, 32 and 64 *S. l. cerasiforme* accessions were sampled (see Additional file 2: Cerasiforme and mixed core collections). Fourty to 98% of SSR alleles were captured when accession's number increased from 8 to 64 (table 7). The number of phenotypic classes captured, increased from 18 (60% of the classes from the *S. l. cerasiforme* sample) to 27 (90% of the classes from the *S. l. cerasiforme* sample) when accession's number increased from 8 to 64. The 64 accession sample did not show any genetic structure when it was analyzed with the model-based method.

Table 5: Subgroup pairwise F_{ST} for K_{opt} = 4

	popA	popB	popC
popA	0.00000		
popB	0.23358	0.00000	
popC	0.57111	0.39977	0.00000
popD	0.64958	0.48797	0.21444

Individuals were clustered corresponding to their maximal membership. All comparisons were significant (p < 0.001)

For fruit weight (FW), soluble solid content (SSC) and titratable acidity (TA), the core collection of 64 accessions best represented the phenotypic variability of the global sample even though extreme phenotypes were not represented (Figure 5). The sample consisting of 32 accessions seemed to be the best compromise because of its small number of accessions and its representativeness.

To complete representativeness of these core collections, ten accessions from *S. lycopersicum*, two accessions from *S. pimpinellifolium* and four wild related accessions (*S. chesmaniae, S. habrochaites, S. pennellii and S. chmielewskii*) were added to each of the core collections to constitute mixed interspecific core collections. The core collection of 64 accessions was also completed with seven other accessions from *S. lycopersicum* and eight accessions from *S. pimpinellifolium* also sampled separately using 20 SSR alleles and 12 morphological traits with the M strategy.

Discussion

Previous studies on the genetic structure of tomato collections focused on cultivated accessions [29,30] or on the relationship between cultivated and wild relatives [26,27] but did not use a broad sample of wild and cultivated tomatoes with *S. l. cerasiforme* as the main sample. SSR markers have already been shown to be useful for genetic analysis in studies focusing on inferring interspecific relationships or confirming SSR reliability for genetic mapping [26,28,31-34].

Differences were observed among SSR markers. For example, a higher number of alleles was identified in the two-base motif markers compared to other three-base or complex motif markers (P-value = 0.039). A significant difference was observed between the number of alleles with AT-rich motifs and non AT-rich motif markers (P-value = 0.032). Two base AT-rich motif markers also displayed higher expected heterozygosity. This kind of SSR marker might be useful for inferring fine relationships between close accessions. Because of the higher mutation rate in

BMC Plant Biology 2008, **8**:130 http://www.biomedcentral.com/1471-2229/8/130

Table 6: Individual clustering, allelic diversity and proportion of loci in linkage disequilibrium in the four clusters inferred using Structure2.0 (n = 318)

	Group 1: 'wild'			Group 2: 'domesticated'			'wild/domesticated' Admixed
Number of individuals in each cluster	**Subgroup A**	**Subgroup B**	**AB Admixed**	**Subgroup C**	**Subgroup D**	**CD Admixed**	
S. lycopersicum[a]	1	0	0	16	81	12	3
cerasiforme[a]	3	6	1	59	19	22	34
S. pimpinellifolium[a]	13	30	7	1	0	0	10
total	17	36	8	76	100	34	47
N_A [b]	2.9	6.2	2.8	4.15	3.4	3.3	4.7
H_E [c]	0.3275	0.5960	0.3852	0.2816	0.2245	0.2772	0.4595
LD (number of comparison)	23/135	135/190	16/90	36/119	25/152	80/119	83/152

[a] Individuals were classified in a cluster if their membership for this cluster was higher than 70%. [a] number of individual in each cluster. [b] allele number. [c] Expected heterozygosity

the AT-rich motif markers, some misevaluation might occur because of homoplasy (i.e. alleles identical in terms of state but not by descent) for distant individuals [35]. SSR markers with lower mutation rates with three-base or complex motifs are more reliable markers for inferring interspecific relationships.

SSR markers had between two and 26 different alleles in the total collection (including eighteen wild green-fruited accessions, one *S. galapagense* and one *S. cheesmaniae* accessions) and the allele number decreased between one and five alleles when looking in the red-fruited tomato sample and only for allelic frequency higher than 5%. The pattern of genetic diversity inferred from SSR alleles also showed an important decrease in diversity (i.e. expected heterozygosity) when comparing *S. pimpinellifolium* and *S. lycopersicum* accessions. Furthermore, the observed heterozygosity is lower than expected for all species due to the reproductive regime of red-fruited accessions, but also to the way genetic resources were maintained. The red-fruited accessions are mainly autogamous (except a few highly allogamous *S. pimpinellifolium* accessions) and the green fruited accessions are mainly self-incompatible (except *S. chmielewskii* and *S. neorickii* which are self-compatible) [27]. The decrease of allele number and diversity in red fruited accessions is probably due to the restriction of allogamy. The drop in diversity between the wild and domesticated species has been previously described [14,16,17] and was explained by successive bottlenecks starting from domestication and continuing with modern breeding of *S. lycopersicum*. This species presents a high selfing rate which hampers restoration of genetic diversity lost during domestication. *S. pimpinellifolium* showed higher diversity because of its wild status (weak anthropic restriction in the effective population size compared to domesticated species) and because it benefited from intercrossing. In fact, partial allogamous populations of *S. pimpinellifolium* were described in Northwestern Peru. While they migrated away from that territory, selection has favored self pollination [36]. The higher rate of observed heterozygosity shown by *S. pimpinnellifolium* is thus a residue of intercrossing from allogamous accessions.

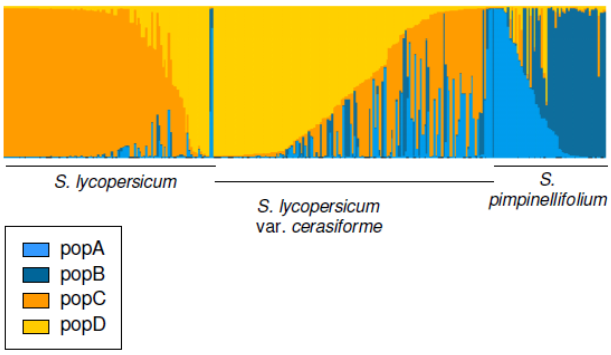


Figure 2
Classification of individuals using *Structure2.0* according to the previous classification into species. The distribution of the individuals to different clusters by the model-based method is indicated by the color code in the legend box.

BMC Plant Biology 2008, **8**:130 http://www.biomedcentral.com/1471-2229/8/130

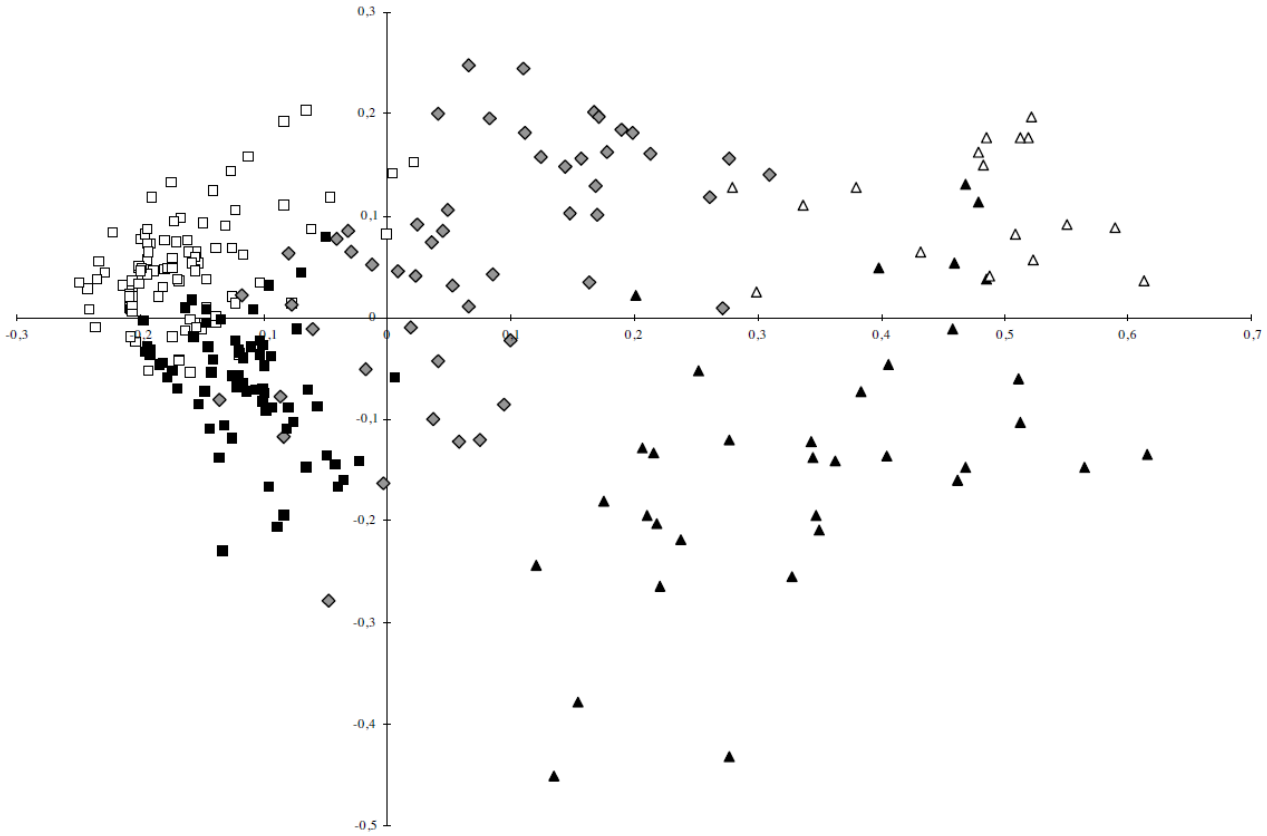


Figure 3
Principal Coordinate Analysis of the *Eulycopersicon* sample with Structure2.0 clustering information. The 'red-fruited' sample did not contain *S. cheesmaniae* accessions. The subdivision of the collection assuming Kopt = 2 separates group 1 (triangle) and group 2, (square) accessions. When assuming Kopt = 4, large fruited accessions: subgroup D (black square) and small-size fruit accessions: subgroup C (white square) are divided. For wild accessions, subgroup A (white triangle) and subgroup D (black triangle) were divided. 'Wild'/'domesticated' admixed accessions are represented by grey diamonds. Intra-specific admixed accessions are not identified. Inertia values are 22.09% and 4.84% for factorial coordinates axes 1 and 2, respectively.

All red-fruited plants used are progenies from self-compatible accessions where seeds are produced through self-pollination. Because of the inbred nature of most accessions, only one plant was used for genotyping. The amount of diversity in this sample is thus underestimated. There is a bias when analyzing observed heterozygosity on artificially self-pollinated accessions but the residue of intercrossing observed testifies to ancestral intercrossing. The estimation of observed heterozygosity should be done on the initial population (from prospecting) to assess the intercrossing rate of these populations. However, the homozygosity of accessions will help in dissecting the genetic bases of agronomical traits using diversity studies.

S. l. cerasiforme showed an intermediate amount of genetic diversity between *S. lycopersicum* and *S. pimpinellifolium*. This particular position has already been described using allozymic variation [17] and both patterns of genetic variation close to *S. pimpinellifolium* and S. *lycopersicum* were encompassed. Cherry type tomatoes, found in coastal Peru or Ecuador and which were described as feral, wild, or used as cultivated landraces, may have played an important role in the evolution of domesticated tomato [37]. This variety characterized by morphological traits like fruit size and seed weight spans a genetic *continuum* between 'wild' and 'domesticated' forms of the crop. Current results suggest that this group of *S. l. cerasiforme* evolved through hybridization between *S. lycopersicum*

BMC Plant Biology 2008, **8**:130 http://www.biomedcentral.com/1471-2229/8/130

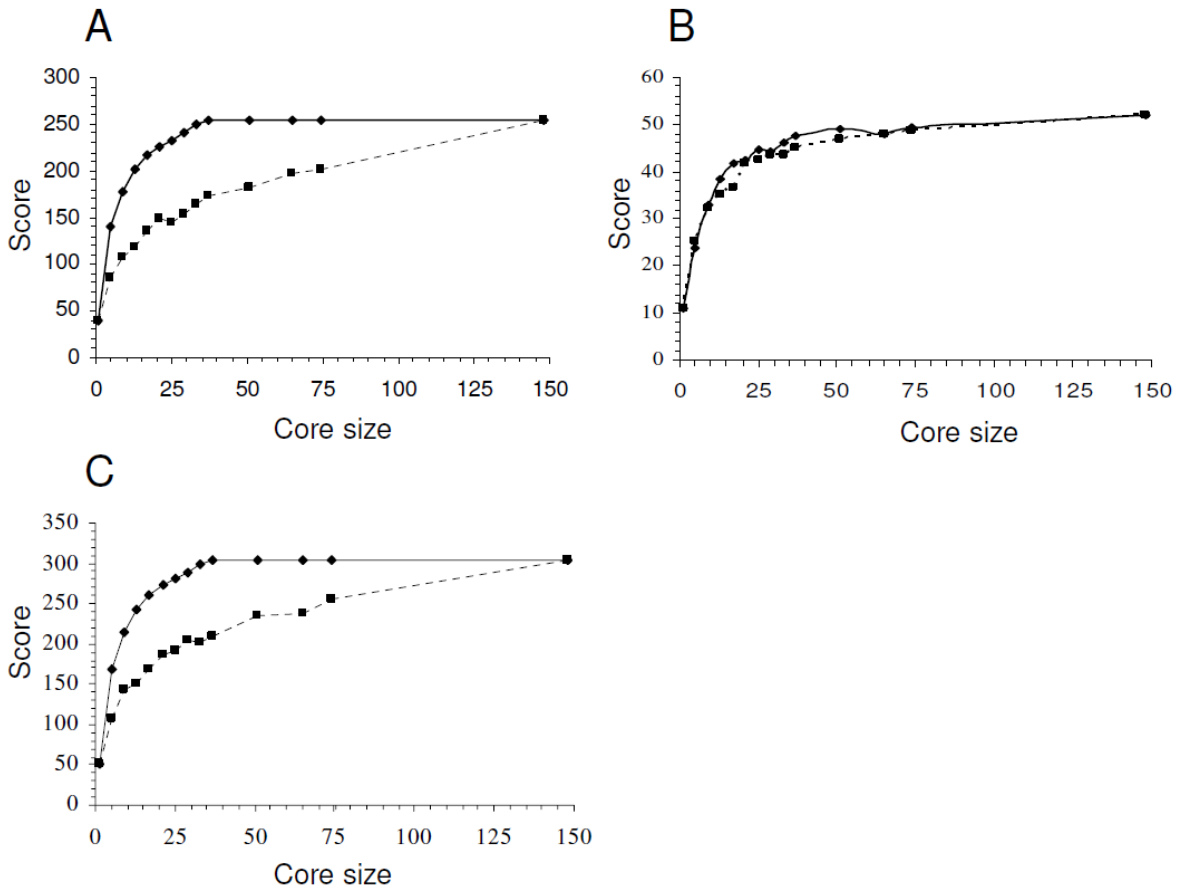


Figure 4
Comparison of efficiency of random and maximization (M) sampling strategy in *S. l. cerasiforme* sample (n = 143 accessions). Score, which represents allelic richness, is plotted against size of core collection. The efficiency of the M strategy is represented by a straight line and the random strategy is represented by a dashed line. A. Core collections were sampled with alleles from 20 SSR loci and were cross validated by the same alleles. B. Core collections were sampled with alleles from 20 SSR loci and were cross validated by alleles from twelve phenotypic data split in 5 classes. C. Core collections were sampled with alleles from 20 SSR markers and twelve phenotypic data and were cross validated by the same alleles.

and *S. pimpinellifolium*. The wild and feral parts of *S. l. cerasiforme* accessions, which have been described as highly invasive, adapted rapidly thanks to the increase in genetic variance, new gene interactions, masking or unloading of deleterious recessive alleles, or the transfer of favourable genes [38].

Genetic structure was highlighted by the model-based method developed by Pritchard et al., (2000) for human genetics. This method performed better than clustering methods based on pairwise genetic distance because only a modest number of loci was used [6]. The higher level of genetic structure allowed most of the *S. lycopersicum* and a part of *S. l. cerasiforme* accessions to be assigned to a 'domesticated' group and most of the *S. pimpinellifolium* to the 'wild' group. The other part of the cherry tomato sample was classified in an admixture position, which is consistent with the distance-based method. The subdivision of the 'domesticated' group in large and small fruit size accessions is consistent with the results of van Berloo et al. (2008) with AFLP markers. These authors found higher differences between cherry versus beef and round tomatoes than between round and beef tomatoes themselves. Homozygosity creates departure from Hardy-Weinberg equilibrium which is one of the hypotheses to apply the model-based method. This limitation was overcome using haploid genotypes. Simulations showed that dominant markers can give results as accurate as codominant

BMC Plant Biology 2008, **8**:130 http://www.biomedcentral.com/1471-2229/8/130

Table 7: Phenotypic and molecular representativeness of the four *cerasiforme* core collections.

	SSR allele		Number of classes for phenotypic quantitative trait[a]						
Size of core collection	number	%	Active variables				Target variables[b]		Total
			Fruit weight	Fruit locule number	Firmness	Color (a*)[c]	SSC[c]	titratable acidity[c]	
8	52	40	3	2	4	2	4	3	18
24	105	76.9	4	3	5	4	4	3	23
32	119	91.5	4	3	5	5	4	5	26
64	128	98.5	4	3	5	5	5	5	27
total	130	100	5	5	5	5	5	5	30

[a] Each quantitative variable was split into 5 classes of equal dimension (see Materials and Methods) when looking for the whole *cerasiforme* sample and number of classes. [b] Soluble Solid Content (SSC) and titratable acidity were only used to analyse core collections representativeness and were not used as active variables for the sampling of these core collections. [c] a* describes how red/green a color is.

markers [39]. We can thus validate our clustering though genotypes were coded in haploid setting. However, caution must be exerted when interpreting biological significance of the clustering because results are sensitive to the type of genetic marker used, the number of loci scored, the number of population sampled and the number of individual typed in each sample.

No relationship between the geographical origin and genetic structure was found within the wild group. Geographic distributions of genetic variability were highlighted for *S. pimpinellifolium* across coastal Peru and Ecuador, using isozyme markers [36]. Regional distribution of isozyme allelic variants and morphological traits for *S. l. cerasiforme* was also described [17,18]. This could be explained by differences in property for markers used. Allozyme markers and morphological traits may be under selective constraint in natural populations in contrast with SSR markers which are usually described following the neutrality hypothesis. Moreover, the results cited above, were obtained for offspring directly collected from natural populations. We employed a different approach using highly inbred plants: diversity patterns were compared among clusters and not among natural populations. The SSR markers presented in this study should be genotyped for natural populations of *S. pimpinellifolium* or *S. l. cerasiforme* to elucidate the correlation of geographical and genetic structures.

The lower amount of diversity and the highest number of alleles in LD in the subgroup A could be explained by reproductive isolation with a high frequency of short-style flowers in the original population (data not shown). This trait is characteristic of strictly autogamous tomato accessions [16,40]. This morphological change, that favors selfing over outcrossing, could also explain the genetic structure [41]. The higher genetic diversity of subgroup C in the 'domesticated' groups could be due to a more ancient and less drastic genetic bottleneck caused by domestication. The drop in genetic diversity in subgroup D is likely due to modern selection which focused on yield and fruit size. The interspecific admixed cluster presented high value of diversity index which is inconsistent with highly autogamous and domesticated forms but confirmed the hypothesis of frequent recombination between cultivated *S. lycopersicum* and wild *S. pimpinellifolium*. These results suggest a two-step selection for fruit size during domestication of tomato from *S. pimpinellifolium* to *S. lycopersicum*. A first step may have allowed selection of cherry type with moderate fruit size probably with fixation of autogamy. The human migration may have resulted in transfer of cultivated tomato from the Andes to Central America with selection for larger fruit size. In Mexico, tomato reached a fairly advanced stage of domestication before being taken to the Old World by conquistador [15,42]. The role of the 'admixed' part of *S. l. cerasiforme*, in tomato domestication can't be established because hybrid pattern could be due to ancient or recent outcrossing events.

The admixed *S. l. cerasiforme* cluster is of particular interest for mapping complex traits. This subsample could be used in an admixture mapping strategy that falls between linkage analysis and association mapping, and is a good approach for initial genome scan [43]. The extent of difference in allele frequency between the ancestral populations is crucial in detecting strong associations between phenotypes and molecular polymorphisms. This difference in allele frequency was obvious in 'wild' and 'domesticated' tomato groups as it represented the main genetic structure level

BMC Plant Biology 2008, **8**:130 http://www.biomedcentral.com/1471-2229/8/130

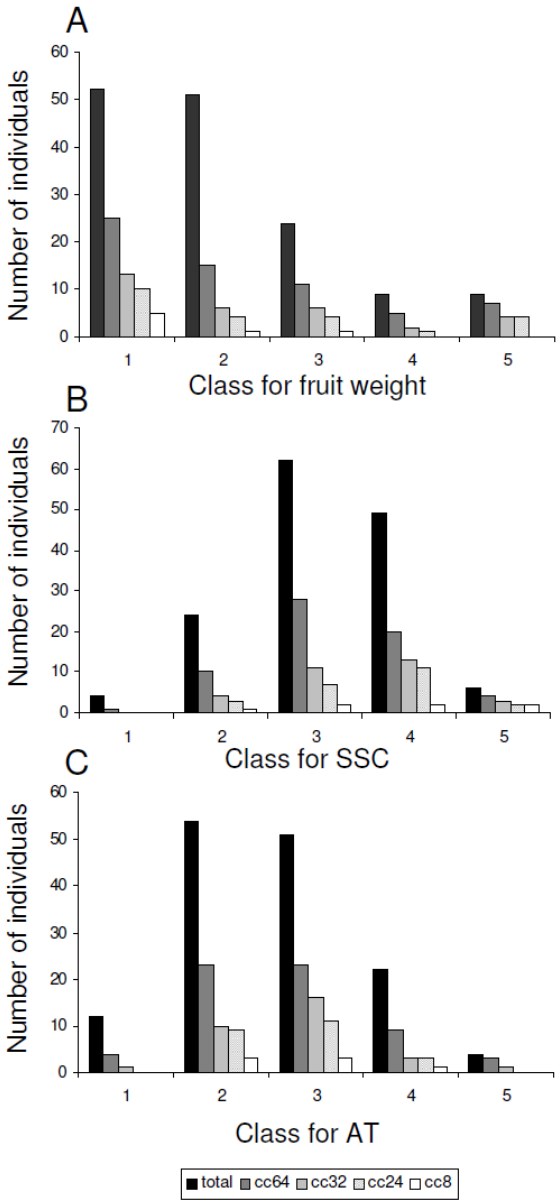


Figure 5
Core collection representativeness for fruit weight, Soluble Solid Content and Titratable Acidity. Classes are those used for core collections (cc) design.

highlighted with the model-based method. In humans, admixture mapping has already been performed to map two loci responsible for hypertension [44]. This method will be assessed in future studies for identifying new QTLs or candidate genes linked to fruit quality traits.

The number of pairwise markers at linkage disequilibrium (LD) decreased in the different groups compared to the total red-fruited accessions. Strong LD between distant or independent markers arises as a consequence of genetic linkage, of the rate of recombination, drift or non-random mating, and as a consequence of population structure. Information on genetic structure of the collection and the membership information for all individuals will be useful in future association mapping to avoid spurious associations due to strong LD over the genome [6].

However, more markers are needed to efficiently tag the genome and better unravel the genetic structure of the cultivated *S. lycopersicum* and *S. l. cerasiforme*. Furthermore, more markers will also be of great interest for estimating individual's kinships. New statistical methods for association studies use both genetic structure information and kinship estimation [45].

This study provided a set of nested core collections for *S. l. cerasiforme* accessions which was completed by selected accessions of *S. lycopersicum*, *S. pimpinellifolium* and wild relatives representing parents of different mapping populations. We focused on *S. l. cerasiforme* because of (i) its close relationship with *S. lycopersicum*, (ii) its genetic diversity which is higher than that of *S. lycopersicum* and (iii) its high range of variation in fruit quality traits. Because of differences in genetic and morphological diversity patterns in 'wild' versus 'domesticated' forms of the tomato *continuum*, core collections were sampled using both phenotypic and molecular diversity. For sampling core collections, the gain when scoring with the Maximization strategy was higher than with the Random strategy. This is not surprising given the high level of selfing in *S. l. cerasiforme* and the pattern of genetic structure uncovered in our sample, both factors that favor the marker assisted sampling strategies over pure random strategies [10,11]. Moreover 20 SSR markers were not sufficient to differentiate all accessions based on their genotype. Markers with higher mutation rates will be more accurate in differentiating individuals based on fingerprinting but will decrease the accuracy of sampling core collections with the M strategy.

Moreover, the M strategy sampled molecular diversity but also morphological diversity even for traits that were not used as markers for sampling the collection.

The four core collections proposed will have different goals. The 24 mixed core collection (including cultivated and wild mapping population parents) will be useful for detecting SNPs by sequencing. SNP markers will then be genotyped on the whole tomato collection for association studies or on mapping populations for QTL localization. Sampling this collection was a preliminary step for future studies on exploring the natural diversity of tomato that will unfold as the tomato genome sequence becomes available [13]. For example, Simon et al. [46] crossed *Arabidopsis thaliana* reference genotypes (i.e. whole genome

BMC Plant Biology 2008, **8**:130 http://www.biomedcentral.com/1471-2229/8/130

sequenced genotypes) with several accessions from a previously defined core collection [47]. The authors built 15 Recombinant Inbred Line family and this new RIL set offered improved accuracy for QTL localization than previous RIL families.

The 64 *S. l. cerasiforme* core collection will be useful for direct association studies. This core collection maximizes the power of associations between phenotypes and allele frequencies. The core collection was test with the model based methods and showed no genetic structure. A broad geographic origin (available for wild accessions) and large phenotypic variation for fruit quality traits were represented. The 96 mixed core collection will help in understanding domestication of tomato from *S. pimpinellifolium*. Identified alleles of interest in admixed *S. l. cerasiforme* could be assigned to *S. pimpinellifolium* or to *S. lycopersicum* to identify their wild or cultivated origins. Core collections will be used to detect genes associated with domestication i.e. under differential selective constraints in domesticated and wild clusters, and to test their potential for breeding [48]. The 8 and the 32 *S. l. cerasiforme* core collection are interesting for rapid sequencing and identifying SNPs and for evolutionary genomics studies, respectively. These core collections will be of interest for new high-throughput analysis of fruit quality integrating 'omic' information such as metabolomic, proteomic or transcriptomic analysis.

Conclusion

This study highlighted the unknown genetic structure of our wild and cultivated germplasm, enhancing the understanding of the history of the tomato complex. It clarified the position of *S. l. cerasiforme* in the evolution of the cultivated tomato. Part of this sub-species is genetically close to the cultivated *S. lycopersicum* group and the other part is in admixture between cultivated and wild related groups. This admixed cluster is of high interest for increasing resolution of association genetics. We created nested core collections implemented with accessions from *S. lycopersicum* and *S. pimpinellifolium* that maximize genetic diversity. These core collections are available for the tomato community and can be used as standardized panels for identifying novel interesting genes or polymorphism. Future studies will focus on the characterization of *S. l. cerasiforme* to understand the domestication process in more detail and to prospect for new interesting alleles.

Methods

Plant Material

The French collection of wild and cultivated tomato maintained in Institut National de Recherche Agronomique in Avignon (South of France) was used for genotyping. In this collection, most tomato accessions are inbred lines maintained by selfing and characterized for vegetative and reproductive traits. The whole collection consists of nearly 2000 accessions containing inbred cultivars, landraces, and representatives of wild related species. It collates accessions from French researchers' prospecting, from breeders' collections, from the Tomato Genetics Resource Center (Davis, California USA), the Centre for Genetic Resources (Wageningen, Netherlands), the North Central Regional Plant Introduction Station (USA) and from the N.I. Vavilov Research Institute of Plant Industry (St Petersburg, Russia). We used a subset of 360 accessions (see Additional file 3: Individuals information and SSR genotypes) with a majority of *S. lycopersicum* (130 accessions), *S. l. cerasiforme* (144 accessions) and *S. pimpinellifolium* (66 accessions). For the red-fruited accessions, classification in different species was based essentially on fruit size [49]. We added one *S. cheesmaniae* and one *S. galapagense* (formerly *L. cheesmanii f. minor*) which are part of red-fruited tomatoes but not included in the studied sample for domestication, because they are assumed to have evolved separately and to be endemic in the Galapagos Islands. Eighteen representatives of wild and green-fruited related were represented by *S. neorickii* (1), *S. chmielewskii* (2), *S. peruvianum* (2), *S. chilense* (2), *S. pennellii* (2) and *S. habrochaites* (2). All red-fruited accessions underwent from one to three cycles of self-pollination. Because of the inbred nature of these accessions, only one plant per accession was used for genotyping. All accessions are available on request from the corresponding author.

S. l. cerasiforme accessions (144 accessions) with 39 accessions of *S. lycopersicum* and 19 accessions of *S. pimpinellifolium* were grown in Avignon (South of France) and were phenotyped for growth habit (determinate: *sp* or indeterminate: sp^+), flower stigma insertion (+) or exertion (-), petal length, petal number, green shoulder (presence/absence), stem hairiness (presence/absence), fruit locule number, fruit weight (FW), color in L*a*b* color space: one measure for lightness (L), one measure for the position between red and green (a) and one measure for the position between yellow and blue (b) with a Konica Minolta CR-300 chromameter, firmness with a Durofel durometer http://www.setop.fr, soluble solid content (SSC) and titratable acidity (TA). Phenotypic data were only used for core collection sampling. Quantitative data were split into 5 classes of equi-spaced breaks with class size calculated as [max(X)-min(X)]/5 with X the quantitative variable.

DNA extraction and Microsatellite genotyping

DNA was isolated from 100 mg frozen leaves using to the DNeasy Plant Mini Kit (Qiagen, Valencia, California, USA). Twenty microsatellite loci were used for genotyping (Table 1). These SSR markers were selected from Sol Genomics Network webpage at http://www.sgn.cornell.edu/.

BMC Plant Biology 2008, **8**:130 http://www.biomedcentral.com/1471-2229/8/130

Amplification reactions were performed according to Ronfort et al. [25]. Samples were prepared by adding 3 μL of diluted PCR product to 6.875 μL formamide and 0.125 μL Gensize 400 HD Rox Size Standard (Applied Biosystems, Foster City, USA). Amplified products were detected on an ABI 3710 × l (Applied Biosystems, Foster City, USA) capillary sequencer. Analyses were performed using the GeneMapper 3.7 sofware (Applied Biosystems, Foster City, USA).

Diversity analysis

For each microsatellite locus, the number of alleles (N_A), allelic frequency, the expected (He) and observed (H_o) heterozygosities were estimated considering both the whole collection and the red-fruited accessions using Genetix 4.05.2 software [50]. Heterozygosity was also compared between subsets at the species level.

Inference of population structure

To infer the population structure of the tomato collection, we used a model-based clustering algorithm implemented in the computer program Structure version 2.0 (Pritchard, Stephens, and Donnelly, 2000). This algorithm uses a multilocus genotype to identify a predetermined number (K) of clusters that have distinct allele frequencies and assigns portions of individual genomes to these clusters. It proceeds by assuming that observations are randomly drawn from a parametric model and inference for the parameters allows estimation of ancestry probability from each putative cluster, for all individuals. Only *S. lycopersicum*, *S. l. cerasiforme* and *S. pimpinellifolium* accessions were included in this analysis. Since tomato accessions used are highly homozygous (autogamy plus self-pollination of accessions), we used a haploid setting [25,51]. Given the hybrid hypothesis for the *S. l. cerasiforme* variety we used the admixture model assuming correlation among allele frequencies. Ten runs were taken into account for each K value (K is the number of clusters to be inferred), for K ranging from 1 to 15. For each run, we used a burn-period of 500,000 Markov Chain Monte Carlo iterations and then 250,000 iterations for estimating the parameters. Pr(X|K) (i.e. the posterior probability of the data (X) given K) and the associated standard deviation was computed for each simulation and K_{opt} was inferred from the formula established by Evanno et al. (2005); K_{opt} being the mode of the first peak of $\Delta K = |L"(K)|/s[Pr(x|k)]$, with $|L"(K)|$ the absolute value of the second order rate of change of Pr(X|K) with respect to K; and s[Pr(x|k)] the variance of the posterior probability of the data given K. To avoid genetic classification at the species level, Structure2.0 runs were also performed with the same parameters on sub-groups defined by the software but for K ranging from one to ten. For each K_{opt}, individuals were assigned into a cluster according to their proportion of membership into this group. Graphical representation of the individual assignation into groups was performed with distruct1.1 software [52]. Analysis of locus by locus MOlecular VAriance (AMOVA) was performed (1000 permutations) and F_{ST}, the correlation of alleles within subpopulatons, was calculated (1000 permutations) with Arlequin3.11 [53]. Pairwise comparisons of linkage disequilibrium (LD) among loci were computed with the dedicated procedure of the TASSEL software, using 1,000 permutations.

Graphical diversity analysis

Genetic uniqueness of each accession was determined with pairwise comparison of multilocus DNA profiles. When two or more accessions had the same profile, only one was taken into account in subsequent analyses. Dissimilarity matrices were built according to the simple matching coefficient [54,55]:

$$d_{ij} = 1 - \frac{1}{L}\sum_{l=1}^{L}\frac{ml}{}$$

where L is the locus number, is the ploidy level and m the number of common alleles between individuals i and j. Bootstraping was performed using 500 replicates for each dissimilarity matrix. Principal coordinate analysis (PCoA) offered graphical representation of genetic distance between accessions and was performed using Darwin 5.0 software [56].

Core collection sampling

For sampling core collections, we used the Maximization (M) algorithm implemented in MSTRAT software version 4.1 [57], and compared the result to a random strategy. The minimum number of accessions in the core collection to capture all alleles present in the whole sample was evaluated by sampling simulations of this collection. The core collections were built using all SSR data and phenotypic data from 12 morphological traits: growth habit, flower stigma insertion or exertion, petal length, petal number, green shoulder, hairiness, fruit locule number, fruit weight, color in L*a*b* color space and firmness. Soluble Solid Content and Titratable Acidity were used only to validate capture of phenotypic diversity. For evaluation of core collection's minimal size and for individual sampling of the collections, 15 replicates of 30 iterations for each replicate were performed.

Authors' contributions

NR participated in the conception of the study, analyzed the data and wrote the manuscript. SS participated in the design of the study and was responsible, with NR, for obtaining the molecular data. SM critically revised the manuscript for intellectual content. MC participated in the conception and the coordination of the study and helped to draft the manuscript. All authors read and approved the final manuscript.

BMC Plant Biology 2008, **8**:130 http://www.biomedcentral.com/1471-2229/8/130

Additional material

Additional file 1

Determination of Kopt for each species. This file provides the graphical determination (Evanno, 2005) of Kopt for each species.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-8-130-S1.ppt]

Additional file 2

Cerasiforme and mixed core collections. This file provides a list of the different core collections for S. lycopersicum *var.* cerasiforme *implemented with* S. lycopersicum *and* S. pimpinellifolium.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-8-130-S2.xls]

Additional file 3

Individuals information and SSR genotypes (n = 360). This file provides a list of all individual genotypic data for 20 SSR markers.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-8-130-S3.xls]

Acknowledgements

We thank anonymous reviewers for their valuable comments on the manuscript. We are grateful to Tomato Genetics Ressource Center (Davis, California USA), to the Centre for Genetic Resources (Wageningen, Netherlands), to the North Central Regional Plant Introduction Station (Geneva, New York USA) and to the N.I. Vavilov Research Institute of Plant Industry (St Petersburg, Russia) for providing tomato accessions. We thank Rebecca Stevens and Cindy Morris for English revision of the manuscript. The authors thank Hélène Burck for characterizing and maintaining the INRA tomato Genetic Resources collection. We thank Karine Leyre for plant DNA extraction and Microsatellite tests. We would also like to thank Isabelle Hochu for her help in high throughput SSR genotyping of the tomato collection. We are grateful to Esther Pelpoir for her help with growing and phenotyping cherry tomato accessions.

This work was supported by French INRA AIP Sequençage and Nicolas RANC was supported by EUSOL European project PL016214-2.

References

1. Doebley J, Stec A, Hubbard L: **The evolution of apical dominance in maize.** *Nature* 1997, **386**:485-488.
2. Frary A, Nesbitt TC, Frary A, Grandillo S, Knaap Evd, Cong B, Liu J, Meller J, Elber R, Alpert KB, Tanksley SD: **fw2.2: A Quantitative Trait Locus Key to the Evolution of Tomato Fruit Size.** *Science* 2000, **289**:85-88.
3. Tanksley SD: **The Genetic, Developmental, and Molecular Bases of Fruit Size and Shape Variation in Tomato.** *Plant Cell* 2004, **16 Suppl**:S181-S189.
4. Buckler I, Edward S, Thornsberry JM: **Plant molecular diversity and applications to genomics.** *Curr Opin Plant Biol* 2002, **5**:107-111.
5. Weigel D, Nordborg M: **Natural Variation in Arabidopsis. How Do We Find the Causal Genes?** *Plant Physiol* 2005, **138**:567-568.
6. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: **Association Mapping in Structured Populations.** *Am J Hum Genet* 2000, **67**:170-181.
7. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, *et al.*: **Assessing the impact of population stratification on genetic association studies.** *Nat Genet* 2004, **36**:388-393.
8. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW: **Genetic Structure of Human Populations.** *Science* 2002, **298**:2381-2385.
9. Whitt SR, Buckler ES: **Using Natural Allelic Diversity to Evaluate Gene Function.** *Methods Mol Biol* 2003, **236**:123-140.
10. Schoen D, Brown A: **Conservation of Allelic Richness in Wild Crop Relatives is Aided by Assessment of Genetic Markers.** *PNAS* 1993, **90**:10623-10627.
11. Bataillon TM, David JL, Schoen DJ: **Neutral Genetic Markers and Conservation Genetics: Simulated Germplasm Collections.** *Genetics* 1996, **144**:409-417.
12. Labate JA, Grandillo S, Fulton TM, Munos S, Caicedo A, Peralta IE, Ji Y, Chetelat R: **Tomato.** In *Genome mapping and molecular breeding in plants Volume 5*. Edited by: Kole C. NY: Springer Publishing; 2007:1-125.
13. Mueller LA, Tanksley SD, Giovannoni JJ, van Eck J, Stack S, Choi D, Kim BD, Chen M, Cheng Z, Li C, *et al.*: **The Tomato Sequencing Project, the first cornerstone of the International Solanaceae Project (SOL).** *Comp Funct Genom* 2005, **6(3)**:153-158.
14. Miller JC, Tanksley SD: **RFLP analysis of phylogenetic relationships and genetic variation in the genus Lycopersicon.** *Theor Appl Genet* 1990, **80**:437-448.
15. Bai Y, Lindhout P: **Domestication and Breeding of Tomatoes: What have We Gained and What Can We Gain in the Future?** *Ann Bot* 2007, **100**:1085-1094.
16. Rick CM: **Tomato *Lycopersicon esculentum* (Solanaceae).** In *Evolution of Crop Plants* Edited by: Simmonds NW. Longman; 1976:268-273.
17. Rick CM, Fobes F: **Allozyme Variation in the Cultivated Tomato and Closely Related Species.** *Bull Torr Bot Club* 1975, **102**:376-384.
18. Rick CM, Holle M: **Andean *Lycopersicum esculentum* var. *cerasiforme*: Genetic Varaition and Its Evolutionary Significance.** *Econ Botany* 1990, **44**:69-78.
19. Nesbitt TC, Tanksley SD: **Comparative Sequencing in the Genus Lycopersicon: Implications for the Evolution of Fruit Size in the Domestication of Cultivated Tomatoes.** *Genetics* 2002, **162**:365-379.
20. Matsuoka Y, Mitchell SE, Kresovich S, Goodman M, Doebley J: **Microsatellites in Zea – variability, patterns of mutations, and use for evolutionary studies.** *Theor Appl Genet* 2002, **104**:436-450.
21. Genlou S, Björn S: **Microsatellite variability and heterozygote deficiency in the arctic-alpine Alaskan wheatgrass (*Elymus alaskanus*) complex.** *Genome* 2003, **46**:729-737.
22. Djè Y, Heuertz M, Lefèbvre C, Vekemans X: **Assessment of genetic diversity within and among germplasm accessions in cultivated sorghum using microsatellite markers.** *Theor Appl Genet* 2000, **100**:918-925.
23. Green JM, Barker JHA, Marshall EJP, Froud-Williams RJ, Peters NCB, Arnold GM, Dawson K, Karp A: **Microsatellite analysis of the inbreeding grass weed Barren Brome (*Anisantha sterilis*) reveals genetic diversity at the within- and between-farm scales.** *Mol Ecol* 2001, **10**:1035-1045.
24. Menz MA, Klein RR, Unruh NC, Rooney WL, Klein PE, Mullet JE: **Genetic Diversity of Public Inbreds of Sorghum Determined by Mapped AFLP and SSR Markers.** *Crop Sci* 2004, **44**:1236-1244 [http://crop.scijournals.org/cgi/content/abstract/44/4/1236].
25. Ronfort J, Bataillon T, Santoni S, Delalande M, David J, Prosperi J-M: **Microsatellite diversity and broad scale geographic structure in a model legume: building a set of nested core collection for studying naturally occurring variation in Medicago truncatula.** *BMC Plant Biol* 2006, **6**:28.
26. Alvarez AE, Wiel CCMvd, Smulders MJM, Vosman B: **Use of microsatellites to evaluate genetic diversity and species relationships in the genus Lycopersicon.** *Theor Appl Genet* 2001, **103**:1283-1292.
27. Spooner DM, Peralta IE, Knapp S: **Comparison of AFLPs with other markers for phylogenetic inference in wild tomatoes**

[Solanum L. section Lycopersicon (Mill.) Wettst.]. *Taxon* 2005, **54**:43-61 [http://www.ingentaconnect.com/content/iapt/tax/2005/00000054/00000001/art00006].
28. Bredemeijer G, Cooke R, Ganal M, Peeters R, Isaac P, Noordijk Y, Rendell S, Jackson J, Röder M, Wendehake K, *et al.*: **Construction and testing of a microsatellite database containing more than 500 tomato varieties.** *Theor Appl Genet* 2002, **105(6&#457)**:1019-1026.
29. van Berloo R, Zhu A, Ursem R, Verbakel H, Gort G, van Eeuwijk F: **Diversity and linkage disequilibrium analysis within a selected set of cultivated tomatoes.** *Theor Appl Genet* 2008, **117**:89-101.
30. Mazzucato A, Papa R, Bitocchi E, Mosconi P, Nanni L, Negri V, Picarella ME, Siligato F, Soressi GP, Tiranti B, Veronesi F: **Genetic diversity, structure and marker-trait associations in a collection of Italian tomato (Solanum lycopersicum L.) landraces.** *Theor Appl Genet* 2008, **116**:657-669.
31. Frary A, Xu Y, Liu J, Mitchell S, Tedeschi E, Tanksley S: **Development of a set of PCR-based anchor markers encompassing the tomato genome and evaluation of their usefulness for genetics and breeding experiments.** *Theor Appl Genet* 2005, **111(2)**:291-312.
32. Smulders MJM, Bredemeijer G, Rus-Kortekaas W, Arens P, Vosman B: **Use of short microsatellites from database sequences to generate polymorphisms among Lycopersicon esculentum cultivars and accessions of other Lycopersicon species.** *Theor Appl Genet* 1997, **94**:264-272.
33. Tam SM, Mhiri C, Vogelaar A, Kerkveld M, Pearce SR, Grandbastien M-Al: **Comparative analyses of genetic diversities within tomato and pepper collections detected by retrotransposon-based SSAP, AFLP and SSR.** *Theor Appl Genet* 2005, **110(5)**:819-831.
34. Takezaki N, Nei M: **Genetic Distances and Reconstruction of Phylogenetic Trees From Microsatellite DNA.** *Genetics* 1996, **144**:389-399.
35. Estoup A, Jarne P, Cornuet J-M: **Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis.** *Mol Ecol* 2002, **11**:1591-1604.
36. Rick CM, Fobes JF, Holle M: **Genetic variation in *Lycopersicon pimpinellifolium*: Evidence of evolutionary change in mating systems.** *Plant Syst Evol* 1977, **127**:139-170.
37. Jarvis DI, Hodgkin T: **Wild relatives and crop cultivars: detecting natural introgression and farmer selection of new genetic combinations in agroecosystems.** *Mol Ecol* 1999, **8**:S159-S173.
38. Lee CE: **Evolutionary genetics of invasive species.** *Trends Ecol Evol* 2002, **17**:386-391.
39. Evanno G, Regnaut S, Goudet J: **Detecting the number of clusters of individuals using the software structure: a simulation study.** *Mol Ecol* 2005, **14**:2611-2620.
40. Chen K-Y, Cong B, Wing R, Vrebalov J, Tanksley SD: **Changes in Regulation of a Transcription Factor Lead to Autogamy in Cultivated Tomatoes.** *science* 2007, **318**:643-645.
41. Gao H, Williamson S, Bustamante CD: **An MCMC Approach for Joint Inference of Population Structure and Inbreeding Rates from Multi-Locus Genotype Data.** *Genetics* 2007.
42. Daunay M-C, Laterrot H, Janick J: **Iconography and History of Solanaceae: Antiquity to the 17th Century.** *Hort Rev* 2007, **34**:1-119.
43. Darvasi A, Shifman S: **The beauty of admixture.** *Nat Genet* 2005, **37(2)**:118-119.
44. Zhu X, Luke A, Cooper RS, Quertermous T, Hanis C, Mosley T, Charles Gu C, Tang H, Rao DC, Risch N, Weder A: **Admixture mapping for hypertension loci with genome-scan markers.** *Nat Genet* 2005, **37**:177-181.
45. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, *et al.*: **A unified mixed-model method for association mapping that accounts for multiple levels of relatedness.** *Nat Genet* 2006, **38**:203-208.
46. Simon M, Loudet O, Durand S, Berard A, Brunel D, Sennesal F-X, Durand-Tardif M, Pelletier G, Camilleri C: **Quantitative Trait Loci Mapping in Five New Large Recombinant Inbred Line Populations of Arabidopsis thaliana Genotyped With Consensus Single-Nucleotide Polymorphism Markers.** *Genetics* 2008, **178**:2253-2264.
47. McKhann HI, Camilleri C, Berard A, Bataillon T, David JL, Reboud X, Le Corre V, Caloustian C, Gut IG, Brunel D: **Nested core collections maximizing genetic diversity in Arabidopsis thaliana.** *Plant J* 2004, **38**:193-202.
48. Le Cunff L, Fournier-Level A, Laucou V, Vezzulli S, Lacombe T, Adam-Blondon A-F, Boursiquot J-M, This P: **Construction of nested genetic core collections to optimize the exploitation of natural diversity in Vitis vinifera L. subsp. sativa.** *BMC Plant Biol* 2008, **8**:31.
49. Rick CM, Latterot H, Philouze J: **A revised key for the *Lycopersicum* and related *Solanum* species.** *Tom Genet Coop Rep* 1990, **40**:31.
50. Belkir K, Borsa P, Chikhi L: **GENETIX 4.05.02, logiciel sous WindowsTM pour la génétique des populations.** *Laboratoire Génome, Populations, Interactions CNRS UMR 5000, Montpellier* 2004.
51. Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, *et al.*: **The Pattern of Polymorphism in Arabidopsis thaliana.** *PLoS Biol* 2005, **3**:e196.
52. Rosenberg NA: ***Distruct*: a program for the graphical display of population structure.** 2007.
53. Excoffier L, Schneider S: **Arlequin ver. 3.0: An integrated software package for population genetics data analysis.** *Evol Bioinfo Online* 2005, **1**:47-50.
54. Sneath PHA, Sokal RR: *Numerical Taxonomy. The principles and practice of numerical classification* San Francisco: W.H. Freeman an Co; 1973.
55. Perrier X, Flori A, Bonnot F: **Data analysis methods.** In *Genetic diversity of cultivated tropical plants* Edited by: Hamon P, Seguin M, Perrier X, Glaszmann JC. Montpellier: Enfield, Science Publishers; 2003:43-76.
56. Perrier X, Jacquemoud-Collet JP: **DARwin software.** 2006.
57. Gouesnard B, Bataillon TM, Decoux G, Rozale C, Schoen DJ, David JL: **MSTRAT: An Algorithm for Building Germ Plasm Core Collections by Maximizing Allelic or Phenotypic Richness.** *J Hered* 2001, **92**:93-94.

Les fichiers supplémentaires S1 et S3 sont placés en annexe 3. Le fichier S2 est combiné dans l'annexe 1.

3.3. Complément d'analyse sur la classification des accessions sauvages et sur l'utilisation du programme Instruct.

L'échantillon utilisé lors de cette analyse comptait, en plus des espèces à fruit coloré, des accessions d'espèces apparentées génétiquement distantes de la tomate cultivée. Les marqueurs SSR ont permis de représenter la distance génétique entre ces accessions en utilisant la méthode Neighbour-Joining (Figure 3-6). Cette représentation est en accord avec les différentes phylogénies proposées dans les études précédentes (Miller and Tanksley 1990; Nesbitt and Tanksley 2002; Spooner, Peralta et al. 2005). Ces phylogénies ont pu être établies avec différents types de marqueurs (AFLP, RFLP) mais aussi grâce à des informations de séquences génomiques nucléaires et chloroplastiques.

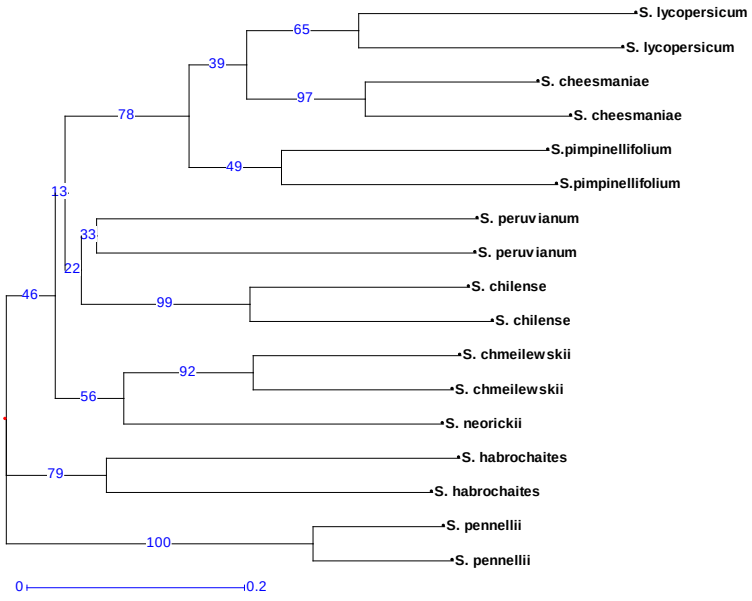


Figure 3-6. Arbre Neighbour-Joining des espèces apparentées à la tomate calculé sur la base de la dissimilarité entre accessions. La matrice de dissimilarité a été calculée sur l'information des 20 marqueurs SSR. Seulement deux accessions représentatives de l'espèce cultivée *S. lycopersicum* (accessions avec gros fruit) et des espèces sauvages *S. pimpinellifolium* et *S. habrochaites* ont été choisies. L'arbre est raciné sur l'espèce *S. pennellii*. Les valeurs de bootstrap sont représentées au niveau des branches correspondantes.

La figure 3-7 représente un arbre Neighbour-Joining des individus avec leur appartenance aux populations identifiées. On remarque que les accessions se regroupent en

fonction de leur appartenance. Les individus cultivés à gros fruits et à petits fruits forment deux sous-populations génétiquement proches.

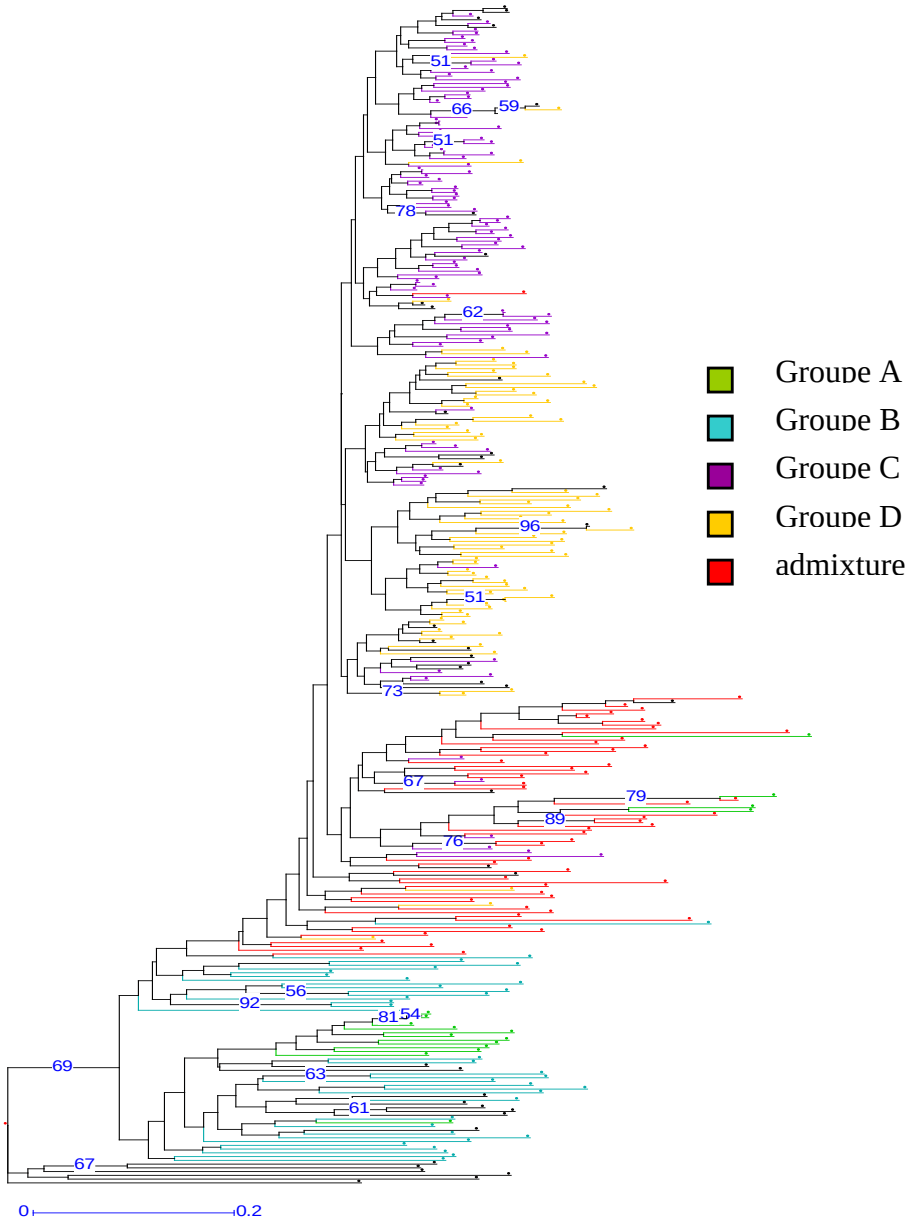


Figure 3-7. Arbre Neighbour-Joining des accessions appartenant à la série *Eulycopericon*. La matrice de dissimilarité a été calculée sur l'information des 20 marqueurs SSR. L'arbre est raciné sur les accessions sauvages de la série *Eriopersicon*. Seules les valeurs de bootstrap supérieures à 50 sont représentées au niveau des branches correspondantes.

De nouveaux programmes d'analyse de la diversité prennent maintenant en compte la recombinaison en incluant la réticulation dans les arbres (Bandelt and Dress 1992; Huson 1998), mais les graphes obtenus par ces méthodes sont peu lisibles à cause d'un nombre trop élevé de nœuds quand l'analyse dépasse dix unités. Cette figure montre aussi la différence de diversité génétique qui existe entre le compartiment sauvage (branches longues entre individus) et le compartiment cultivé (branches courtes entre individus). L'échantillon « admixture » montre bien des branches longues entre individus ce qui caractérise un niveau de diversité équivalent à celui du compartiment sauvage. Un dendrogramme représentant la diversité de différents cultivars de tomate déterminée à partir de marqueurs SSAP, montre un regroupement en fonction du poids du fruit mais aussi de la forme et du type de fruit (frais vs. industriel) (Tam, Causse et al. 2007). Ces regroupements ne sont pas identifiés dans notre étude, notamment du fait d'un plus grand nombre d'accessions qui induisent du bruit de fond empêchant un regroupement aussi efficace. Le nombre de marqueurs est aussi limitant dans ce type d'analyse et il serait intéressant de doubler le nombre de marqueurs SSR utilisés pour reconduire une analyse de la structuration de la diversité. Il est aussi possible d'utiliser un panel de SNP issus de différentes publications afin de recalculer la structure avec des marqueurs différents. Ces marqueurs permettraient d'échantillonner une part plus importante du génome et de donner une estimation de la structure plus précise.

Le logiciel STRUCURE 2.1 et l'analyse de la diversité par construction d'arbre Neighbour-Joining donnent des résultats similaires. L'analyse graphique ne peut pas représenter les recombinaisons entre les individus alors que STRUCURE 2.1 prend en compte l' « admixture » potentielle de l'échantillon. Ce logiciel répond à des hypothèses assez lourdes qui peuvent être surmontées en travaillant avec des haplotypes pour les espèces autogames. Une évolution de la méthode utilisée dans STRUCURE 2.1 a été développée pour prendre en compte la nature autogame de certaines espèces (Gao, Williamson et al. 2007). Le nouveau programme, INSTRUCT, prend en compte le taux d'autofécondation qu'il estime afin d'améliorer la classification des individus dans chaque sous-population. La comparaison des sorties des deux logiciels (Figure 3-8) montre que la dernière méthode n'apporte aucun gain lorsqu'on étudie la collection entière. Un gain substantiel, sur la définition du nombre K de sous-population, est obtenu lorsqu'on s'intéresse à la « core collection » de 92 accessions. En effet, pour les faibles valeurs de K, ce modèle conduit à des vraisemblances plus élevées que STRUCURE 2.1. Ce gain substantiel est à mettre en parallèle avec un nombre de paramètre estimé plus important que dans le modèle (estimation de paramètres liés à l'autogamie). Pour

l'échantillon de 92 accessions, le plateau est plus facilement identifiable avec INSTRUCT qu'avec STRUCTURE 2.1. La méthode d'Evanno donne des résultats consistants dans la correction des sorties de chacune des deux méthodes d'analyse.

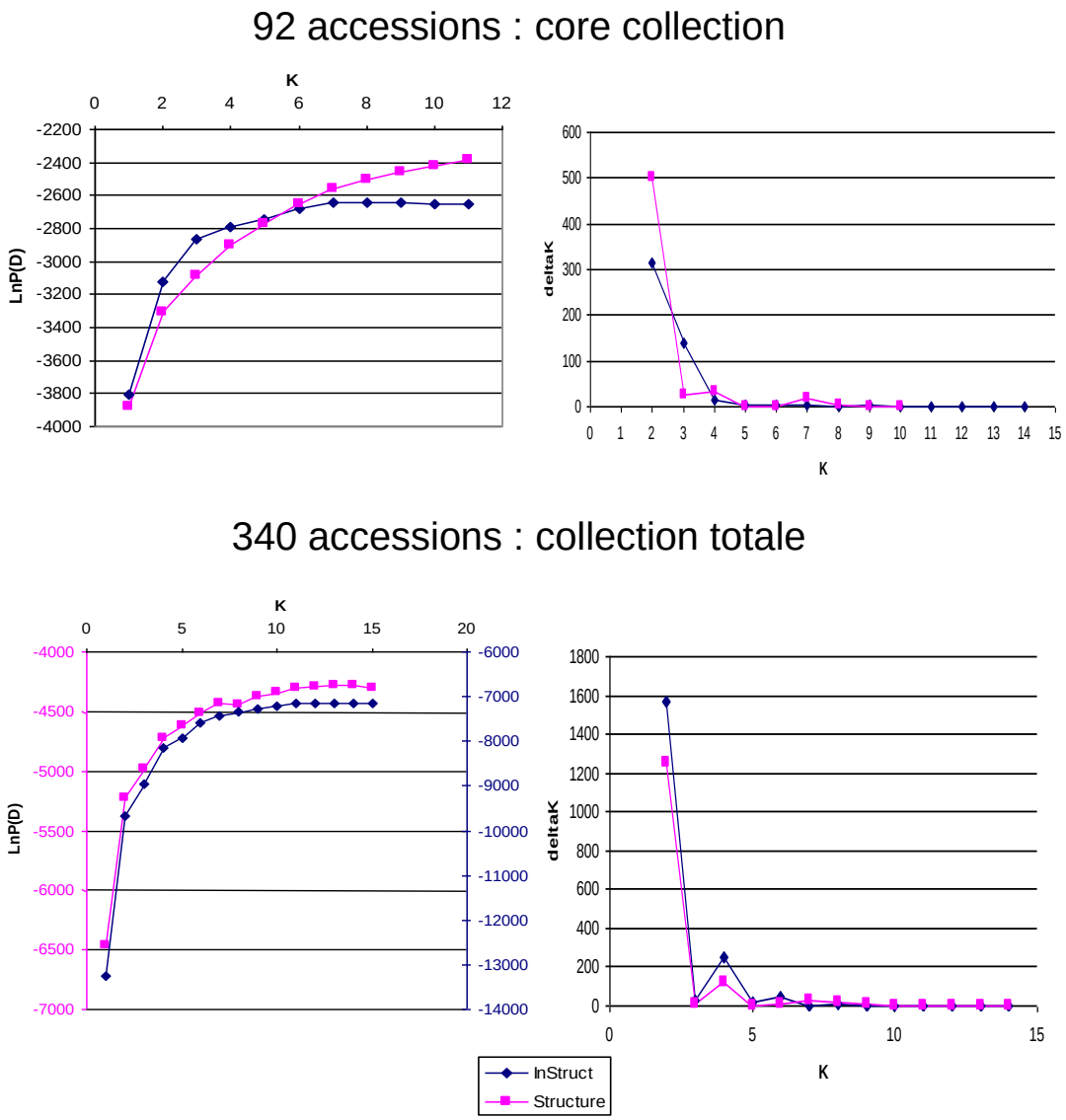


Figure 3-8. **Comparaison de l'efficacité des logiciels INSTRUCT et STRUCTURE dans l'analyse de la structure génétique de deux échantillons de tailles différentes.** La méthode traditionnelle est décrite par les schémas de gauche alors que la méthode corrigée (Evanno, Regnaut et al. 2005) est décrite par les schémas de droite. LnP(D) représente la vraisemblance d'avoir K sous-populations.

Les résultats des deux logiciels en termes de groupement des individus dans chaque sous-population ont ensuite été comparés. Lorsqu'on s'intéresse à la « core collection » de 92 individus, les deux logiciels classent les individus de la même façon si on prend en compte deux sous-populations (2 accessions mal classées dont une accession en « admixture ») mais

ils donnent des résultats différents si on prend en compte quatre sous-populations (29 accessions mal classées dont 14 en « admixture »). Les résultats montrent aussi des différences lorsqu'on s'intéresse à la population totale de 340 accessions. Si on prend en compte deux sous-populations, 27 accessions (dont 10 accessions en « admixture ») ce classent différemment. Si on prend en compte quatre sous-populations, ce sont alors 74 accessions (dont 25 en « admixture ») qui se classent différemment. La comparaison entre ces deux sorties a été analysée au niveau graphique à l'aide de l'ACoP et il semblerait que la classification donnée par STRUCTURE soit concordante avec la structuration graphique alors que celle donnée par INSTRUCT ne l'est pas.

Zhao, Aranzana et al. (2007), dans une étude d'association chez *A. thaliana*, montrent qu'il est possible d'utiliser l'information de coordonnées des individus dans une analyse multifactorielle comme information de structure génétique dans le modèle d'analyse d'association. Il semble donc plus judicieux d'utiliser les sorties de STRUCTURE 2.1 qui coïncident avec les résultats de l'ACoP. C'est pour cela que dans la suite du travail de thèse, seules les structures établies par le logiciel STRUCTURE 2.1 seront prise en compte pour les tests d'association.

Les « core collections » échantillonnées lors de cette étude représentent un panel de référence, disponible pour la communauté scientifique qui s'intéresse à la diversité naturelle rencontrée chez la tomate. L'échantillon de 24 accessions permettra de détecter du polymorphisme par séquençage allélique. Cet échantillon permettra d'inférer la fréquence des allèles rencontrés dans la collection totale et d'éviter de sélectionner des singletons (mutation ponctuelle présente chez une seule accession) pour un génotypage à plus large échelle. La « core collection » de 96 accessions (92 accessions *Eulycopersicon*) représente un échantillon de taille limitée. La puissance d'identification de QTL sera testée dans le chapitre 5. La figure 3-9 montre comment cet échantillon capture la diversité de la collection totale. Dans le reste du manuscrit, on n'utilisera que la « core collection » de 96 accessions en se focalisant uniquement sur les 92 accessions de la série *Eulycopersicon*.

3.4. Conclusion

Un résultat important de ce travail est la confirmation de la position « admixture » d'une partie des accessions de tomate cerise. La distinction entre les deux types de *S. lycopersicum* est basée sur la taille du fruit mais cette structure n'est pas complètement conservée au niveau génotypique. Les tomates de type cultivé, possédant des gros fruits, forment un groupe différencié. Les accessions de type cerise suivent deux patrons de structuration différents. Une première partie forme un groupe différencié, relativement proche du groupe cultivé à gros fruits, alors que l'autre partie se distribue entre les accessions sauvages et les accessions cultivées. Cette structuration de la diversité témoigne de deux histoires évolutives différentes qui conduisent à un même type botanique.

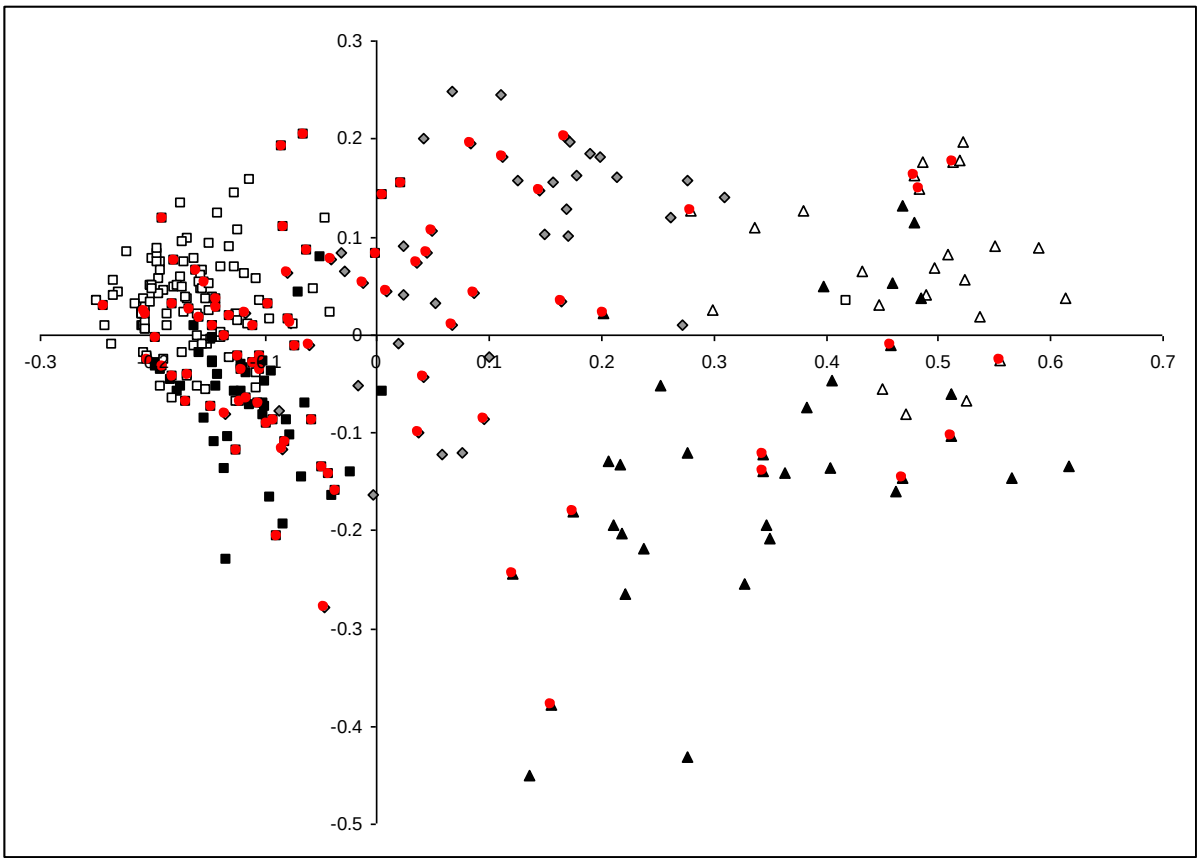


Figure 3-9. Représentativité génétique de la « core collection » de 96 accessions. Les différents groupes sont ceux indiqués dans la figure 3 de l'article. Les carrés blancs représentent la population D, les carrés noirs représentent la population C, les triangles blancs représentent la population A et les triangles noirs représentent la population B. Les points rouges représentent les individus sélectionnés dans la « core collection » de 96 accessions (seulement les 92 accessions appartenant à la série *Eulycopersicon*).

Les accessions en « admixture » entre espèce cultivée et espèce sauvage sont représentées, à la fois, par des accessions collectées dans la région des côtes andines et par des accessions cultivées. N'ayant aucune information de pedigree sur ces dernières, ces

accessions peuvent avoir deux origines différentes issues de l'amélioration moderne. Il peut s'agir d'accessions sauvages, fixées par la sélection variétale, ou d'individus issus de croisement synthétique entre *S. pimpinellifolium* et *S. esculentum* et qui sont sortis du schéma de sélection traditionnel, pour donner une variété à part entière. Qu'ils soient naturels ou artificiels, les évènements de recombinaison qui ont formé ce patron de diversité chez les accessions de type cerise, ont été fixés par l'autogamie prédominante et par les autofécondations successives liées au maintien de ces accessions dans la collection de ressources génétiques. La position intermédiaire entre sauvage et cultivé n'est vérifiée que pour une partie des accessions de type cerise et il est maintenant important de vérifier si la diversité identifiée est suffisante pour réaliser des études d'association. Pour cela, la prochaine partie s'intéresse à un QTL caractérisé par clonage positionnel dans le laboratoire. Le fait d'avoir un *a priori* sur la région va permettre de vérifier les conditions d'utilisation de l'échantillon en vue de détecter le polymorphisme causal.

Chapitre 4 : Utilisation de la diversité naturelle chez la tomate en vue d'identifier le polymorphisme causal d'un QTL cloné

4.1. Introduction

Le clonage positionnel d'un QTL aboutit rarement à l'identification précise du polymorphisme causal. L'exemple de *fw2.2* montre bien que le QTL isolé dans une population biparentale ne peut pas facilement être simplifié au polymorphisme causal (Nesbitt and Tanksley 2002). Ces polymorphismes sont facilement identifiés s'ils influencent directement la fonction d'une protéine en modifiant la séquence en acides aminés (mutation non synonyme) ou en créant une protéine tronquée (mutation créant un codon stop) (Fridman, Carrari et al. 2004; Lippman, Cohen et al. 2008).

Les QTL impliqués dans la variation de caractères sélectionnés lors de la domestication des plantes sont majoritairement des locus où la fonction des protéines n'est pas modifiée mais où des mécanismes de régulation modifient l'expression des gènes (Hirano, Eiguchi et al. 1998; Carroll 2000; Clark, Linton et al. 2004). Chez la tomate, la plupart des QTL décrits, impliquent des modifications de la régulation de différents gènes (Paran and van der Knaap 2007). C'est notamment le cas pour *fw2.2*.

En 2004, l'équipe « Bases Génétiques et Moléculaires de la Qualité des Fruits » de l'unité GAFL, s'est lancée dans le clonage positionnel d'un QTL contrôlant le nombre de loges chez la tomate. Ce QTL, appelé *lcn2.1*, a été cartographié sur le chromosome 2 dans la population issue le croisement intra-spécifique développée à Avignon : Cervil (C) x Levovil, (L). Comme nous l'avons déjà vu dans le chapitre 1, ce locus est responsable du passage de deux loges dans les fruits de l'accession de type cerise (C) à trois loges et plus dans les fruits de l'accession de type moderne (L). Ce QTL a déjà été décrit dans des études précédentes et co-localise avec un QTL de poids de fruit. En effet, en augmentant le nombre de loges dans le fruit, ce locus est aussi impliqué dans l'augmentation de la taille de celui-ci (Lippman and Tanksley 2001).

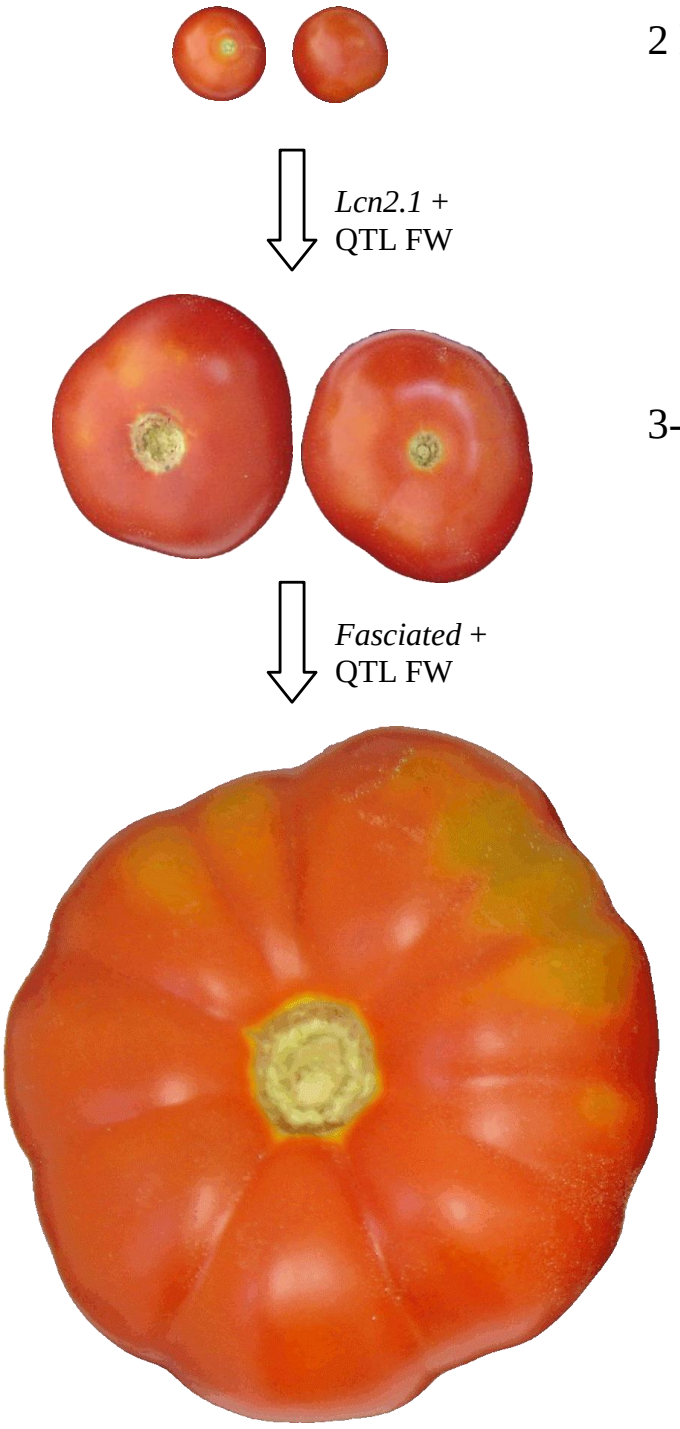


Figure 4-1. Différences phénotypiques de fruit de tomate en fonction du nombre de loge et locus impliqués dans les différences de phénotype. Les deux QTL du nombre de carpelle explique la plus grande part de variation pour ce caractère. Le poids du fruit est aussi contrôlé par des QTL (FW) ayant un rôle uniquement dans l'augmentation de la masse.

Ce QTL n'explique pas la plus grande part de la variation du caractère mais il interagit fortement avec *fas* qui, lui, explique près de 37% de la variation (Barrero and Tanksley 2004). Ce dernier est responsable du phénotype fascié. Ce phénotype extrême est présent chez les accessions de tomate cultivée présentant un nombre de loges supérieur à 6. Le QTL *lcn2.1*, qui agit sur un caractère différenciant la tomate moderne de son ancêtre *S. l. cerasiforme*, est potentiellement impliqué dans la domestication de cette espèce. Le QTL *fas*, qui différencie uniquement certaines accessions de l'espèce cultivée, a dû apparaître après la domestication (Figure 4-1). Le QTL *fasciated* a récemment été cloné. Il s'agit d'un facteur de transcription de la famille des YABBY, déjà identifié dans d'autres espèces comme intervenant dans la modification du nombre d'organes (Cong, Barrero et al. 2008).

Le clonage positionnel a débuté par une étude de cartographie fine des QTL localisés sur le chromosome 2. Celle-ci a permis de générer deux lignées quasi-isogéniques, différant uniquement au niveau de la région d'intérêt (Lecomte, Saliba-Colombani et al. 2004). Ces lignées ont servi de parents à une nouvelle population de cartographie. Lorsque les travaux de cette thèse ont débuté, la cartographie haute résolution avait permis d'identifier un BAC (Bacterial Artificial Chromosome) qui contenait le locus (Tricon 2005). Une annotation de ce BAC, ainsi que l'étude de la micro-synténie avec *A. thaliana* a permis d'identifier *LeWUSCHEL* comme gène candidat potentiel. Ce gène a été caractérisé chez *A. thaliana* comme un régulateur de la transcription, intervenant dans le maintien de l'identité des cellules méristématiques au niveau des bourgeons apicaux (Mayer, Schoof et al. 1998). La co-localisation entre *LeWUSCHEL* (*LeWUS*) et *lcn2.1* avait déjà été décrite auparavant chez la tomate (Barrero, Cong et al. 2006), mais aucune modification de l'expression de ce gène n'a été identifiée entre des cultivars phénotypiquement différents pour le nombre de loges.

4.2. Analyse préliminaire

Le but des travaux sur ce QTL était de vérifier la possibilité d'affiner la position du QTL sur le BAC en utilisant la génétique d'association. Cette analyse préliminaire ne s'est appuyée que sur les données phénotypiques issues de la première année d'expérimentation. Sept fragments répartis sur le BAC 139K19 ont été séquencés sur la « core collection » de 96 accessions. Les fragments ont été répartis sur une région de 50 Kb, ciblée par clonage positionnel (Figure 4-2).

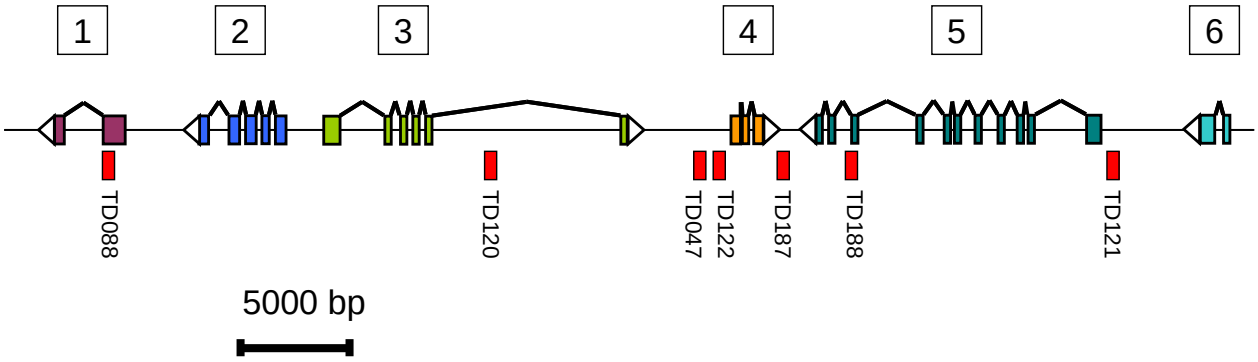


Figure 4-2. Représentation de la région génomique contenant le QTL *lcn2.1* et position des différents fragments séquencés. Chaque gène est représenté par une couleur différente et code pour des fonctions diverses : (1) une protéine kinase calcium-dépendante CPK1, (2) une endo-1,4-beta-glucanase, (3) une oxydoreductase (2OG-Fe(II) oxygenase), (4) un facteur de transcritption homologue à *WUSCHEL*, (5) une transducine avec motif WD40 et (6) une protéine à la fonction inconnue. Les différents fragments séquencés sont représentés en rouge.

Cette région comprenait toujours le candidat *LeWUS* mais contenait aussi un gène codant pour une transducine potentiellement impliquée dans des phénomènes de régulation (de Vetten, Quattrocchio et al. 1997) ainsi qu'un gène de fonction inconnue. Chacun de ces gènes pouvait potentiellement jouer un rôle dans la variation du phénotype. Les fragments TD047 et TD122 ont été définis dans la séquence du promoteur de *LeWUS* car une délétion avait été identifiée dans cette zone et représentait un polymorphisme candidat.

Certaines accessions présentaient un phénotype *fasciated*, or il est connu que ce phénotype est dû à un autre QTL localisé sur le chromosome 11 (*fas*). Les premières analyses ont donc été réalisées sans ces individus. Un panel de 85 accessions a été analysé pour le nombre de loges (lcn), le nombre de pétales (npet) mais aussi la longueur des pétales (lpet). La morphologie florale est prise en compte dans l'analyse afin de vérifier si le déterminisme du nombre de carpelles a lieu précocement, pendant l'organogenèse de la fleur. En effet si le nombre de pétales est corrélé au nombre de loges alors l'hypothèse d'une mise en place très précoce du phénotype (au niveau méristématique) pourra être validée.

Le nombre de loges et le nombre de pétales ne suivent pas une distribution continue, contrairement à la longueur des pétales et au poids du fruit (Figure 4-3). Pour tous les caractères l'effet cultivar est très élevé (p-value < 2.2 e-16). Les héritabilités sont elles aussi très fortes (0,94, 0,99, 0,86 et 0,98 pour lcn, FW, npet et lpet respectivement).

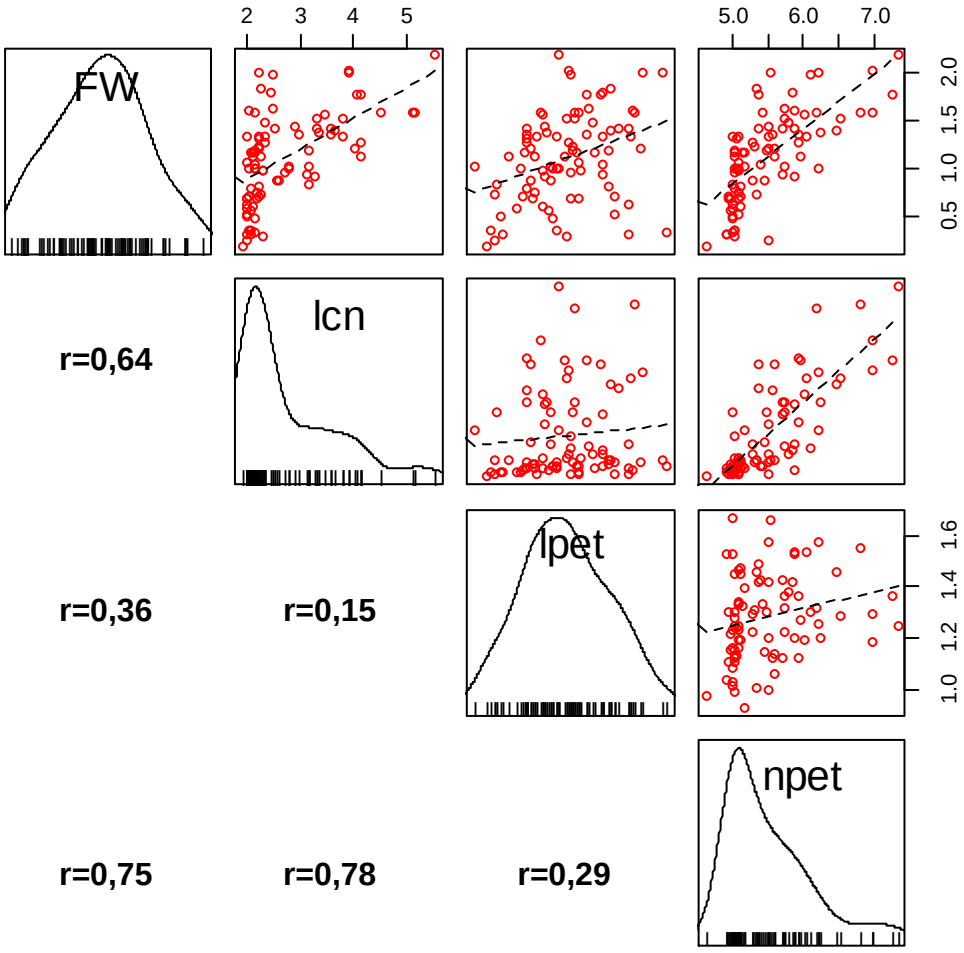


Figure 4-3. Distributions et corrélations des différents phénotypes. Les graphes de densité pour le poids du fruit (FW), le nombre de loges (lcn), la longueur des pétales (lpet) et le nombre de pétales (npet) sont représentés en diagonale. Les corrélations deux à deux (coefficient de Spearman) sont indiquées symétriquement aux graphes en nuage de point.

La détection du polymorphisme ainsi que les comparaisons du polymorphisme moléculaire entre les différentes espèces seront reprises dans le Chapitre 5 et ne seront donc pas discutées dans cette partie du manuscrit. Seuls les sites polymorphes (SNP et indels) présentant une fréquence allélique minimale (minimal allele frequency, MAF) de 0,05 sont retenus dans l'analyse.

La première méthode utilisée pour détecter les associations entre les caractères et les différents polymorphismes identifiés par séquençage, a été la régression logistique utilisée par Thornsberry, Goodman et al. (2001). Cette méthode a permis aux auteurs de détecter une association significative entre la précocité de floraison et *Dwarf8* chez le maïs. Nous avons pris en compte la structure génétique de l'échantillon en utilisant la matrice calculée par le logiciel Structure pour K=2 (K étant le nombre de sous-populations estimées dans le chapitre

3). Un test utilisant 10000 permutations a été utilisé pour calculer la significativité des associations.

Le nombre de loges est un caractère discret qui devient continu lorsqu'on en calcule la moyenne. Pour le clonage positionnel du QTL, celui-ci a dû être mendélisé en transformant les valeurs phénotypiques de chaque accession en allèle provenant du parent C ou L afin de suivre la ségrégation. Nous avons donc voulu voir si la mendélisation du QTL avait un effet sur les résultats d'association. Nous avons remarqué que lorsque le phénotype était mendélisé, un polymorphisme du TD187 était associé avec le caractère (*p*-value = 0,003). La régression logistique associe le caractère au TD120 (*p*-value = 0,01) si les valeurs continues sont prises en compte. Ce phénomène peut être dû à la modification de la gamme de variation du caractère par la mendélisation. En effet, en travaillant sur la base du phénotype mendélisé, on se rapproche de la gamme de variation utilisée pour le clonage positionnel qui est liée au seul locus étudié. Or, on sait que le nombre de loges est contrôlé par plusieurs locus. Même si *lcn2.1* semble être un locus à effet fort, il est évident que l'effet additif des autres locus peut perturber l'étude d'association. Il est donc plus judicieux de travailler avec les mêmes notations que celles utilisées pour le clonage qui a été réalisé avec une population en ségrégation uniquement pour *lcn2.1*.

Lorsqu'on s'intéresse au phénotype « nombre de pétales », la régression logistique identifie là aussi le marqueur TD120 lorsqu'on travaille avec les valeurs discrètes et le marqueur TD187 lorsqu'on travaille avec des classes alléliques. Aucun polymorphisme n'est associé au phénotype « longueur des pétales » La Figure 4-4 présente le déséquilibre de liaison au niveau de la région étudiée. Contrairement à ce qui était attendu chez la tomate, le déséquilibre de liaison semble être relativement faible dans la région, ce qui implique qu'il sera difficile d'identifier le polymorphisme causal si celui-ci n'est pas échantillonné dans le jeu de polymorphismes testés.

Il semble donc bien que la région encadrant l'homologue de *WUSCHEL* soit liée au phénotype « nombre de loges » chez la tomate. Aucun polymorphisme détecté dans le promoteur du candidat n'est responsable de la variation du caractère.

Avant que nous décidions de séquencer ce gène sur l'intégralité des accessions de la « core collection », le criblage de recombinants utilisés pour le clonage positionnel avait

éliminé le gène de la région contenant le QTL *lcn2.1* grâce à deux lignées quasi-isogéniques, différentes pour une région de 1800 bp, distinctes au niveau phénotypique. Cette région se situe en partie 3' de *LeWUS* et de la transducine dans une zone où aucune séquence codante ou régulatrice n'a pu être identifiée.

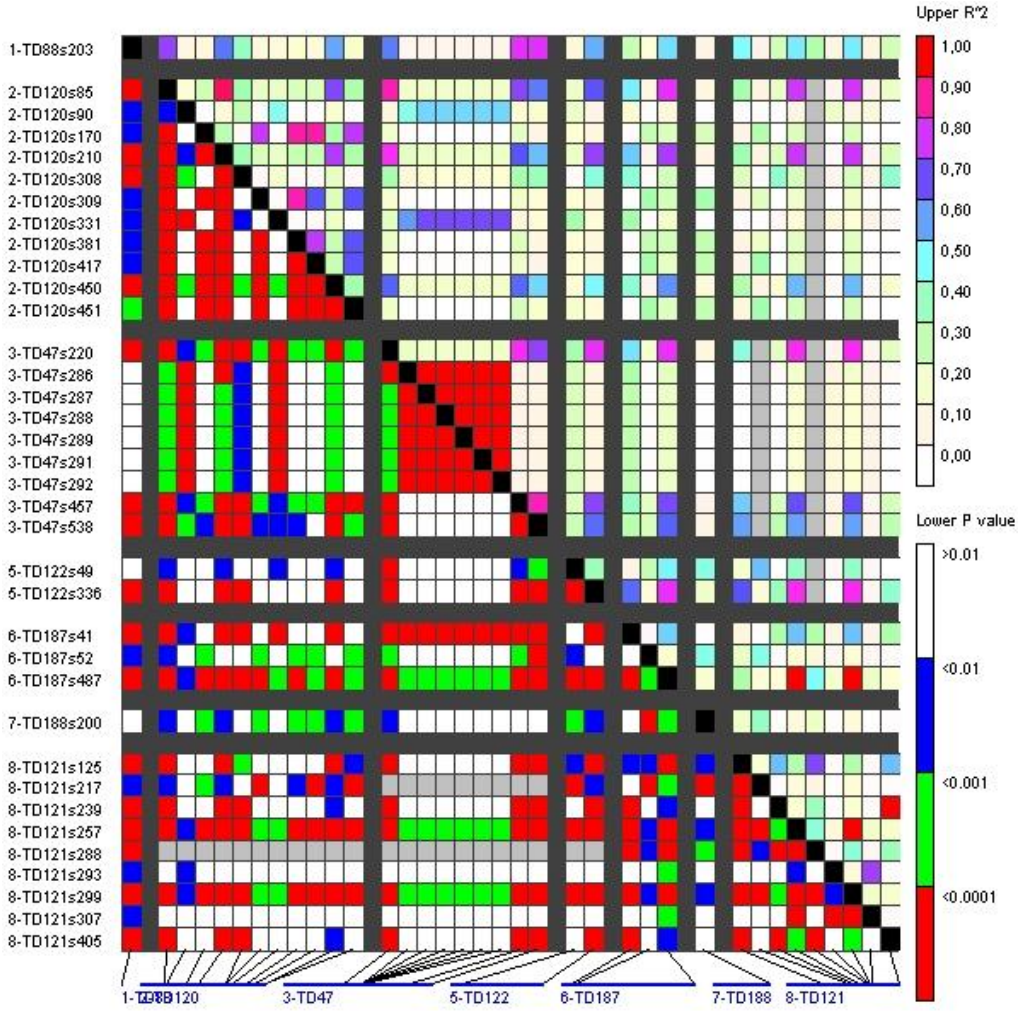


Figure 4-4. Matrice représentant le déséquilibre de liaison au niveau de la région contenant le QTL *lcn2.1* calculé sur l'ensemble de la core collection (92 individus). La partie supérieure de la diagonale correspond au r^2 tandis que la partie inférieure correspond à la *p*-value associée. Seuls les SNP sont représentés.

L'article suivant présente le clonage positionnel conduit par Stéphane Muños et l'étude de la diversité, que j'ai réalisée sur la région identifiée.

4.3. Increases in tomato fruit size and locule number is controlled by two key SNP located near Wuschel.

Muños Stéphane[1], Ranc Nicolas[1], Botton Emmanuel[1], Bérard Aurélie[2], Rolland Sophie[1], Duffé Philippe[1,3], Carretero Yolande[1], Le Paslier Marie-Christine[2], Delalande Corine[4], Bouzayen Mondher[4], Brunel Dominique[2] and Causse Mathilde[1]

1 : INRA, UR1052 Génétique et amélioration des fruits et légumes, Domaine Saint-Maurice, BP 94, 84143 MONTFAVET CEDEX, France.
2 : INRA, UR1279 Étude du polymorphisme des génomes végétaux, CEA - Institut de Génomique - Centre National de Génotypage, 2 rue Gaston Crémieux, CP 5721, 91057 ÉVRY CEDEX, France
3 : present adress : UMR118 Amélioration des plantes et biotechnologies végétales, INRA-Agrocampus Rennes-Univ. Rennes, Domaine de la Motte au Vicomte, BP 35327, 35653 LE RHEU CEDEX, France.
4 : UMR990 Génomique et biotechnologie des fruits, INRA-INPT/ENSAT, Chemin de Borde-Roug-Auzeville, BP 52627, 31326 CASTANET-TOLOSAN CEDEX, France

Abstract:

In tomato fruits, the number of locules (cavities containing seeds and derived from carpels) has been increased from 2 to more than 10 during domestication. Locule number affects fruit shape and size. Its variation is controlled by several QTLs. Two of them, *fasciated* and *lc*, explain the large majority of the phenotype variation. *Fasciated* has been recently cloned and described as a key mutation in the increase of fruit size in modern varieties. Here, we report the cloning of another QTL, *lc*. The map based cloning, performed using 9456 F2 plants, identified a 1600pb region responsible of *lc* and located 1080 bp far from the 3' end of Wuschel. This locus underwent an extreme reduction in the diversity in cultivated accessions except for 2 SNPs that were necessary to increase locule number during tomato domestication.

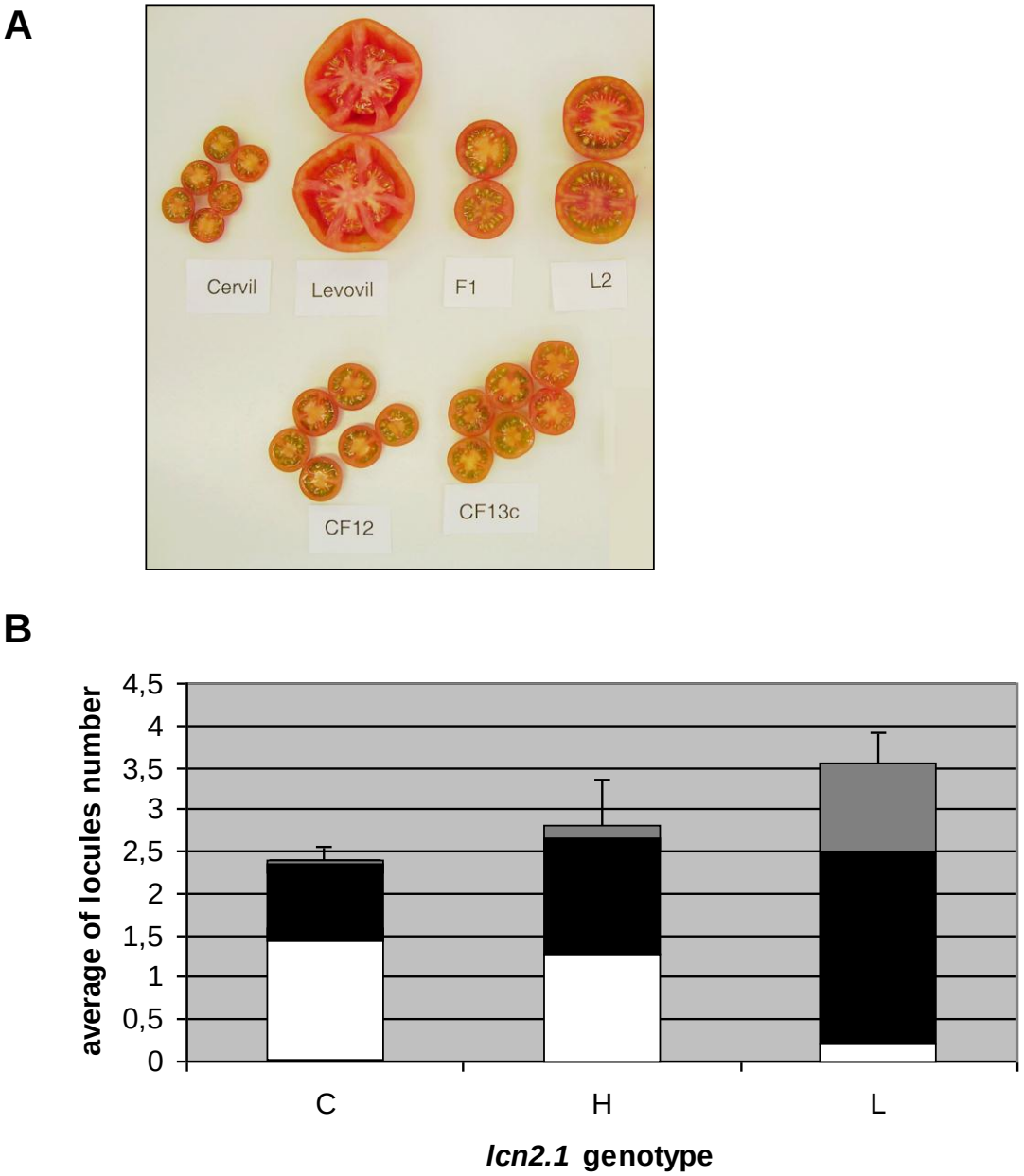


Figure 1: Phenotypic analysis and mendelization of *lc* QTL.
A. Global view of fruit morphology of lines used in the study.
The parental lines Cervil and Levovil were crossed in order to obtain the F1 line and the quasi-isogenic F8 lines CF12, CF13c. L2 is a line derived from Levovil in which the region of the chromosome 2 containing lc QTL from Cervil has been introgressed by marker assisted selection.
B. Phenotypic effect of the QTL
The data are those of F2 heterozygous plants for H or F3 plants for Cervil (C) and Levovil (L). The plants have been selected based on genotyping data. Heterozygous plants were selected by analyzing genotypic and phenotypic segregation in their corresponding F3 progeny. Each histogram is partitioned according to the proportion of fruits with 2 locules (white), 3 locules (black) and 4 or more locules (grey).

One of the major changes in human behaviour during Human history is its transition from hunter-gatherer to farmer. This transition has been accompanied with phenotypic modifications of the plants and animals which composed human food. In plants, a consequence of domestication of wild species was a reduction of molecular diversity together with an increase of morphological diversity. As an example, plant yield has considerably increased, particularly during the last 50 years. But modern plant breeding seems to have reached its limits as the increase is slowing. The erosion of molecular diversity, at least for important loci, that has accompanied the yield increase is one of the hypotheses that could explain this observation. It is therefore essential to return to wild species or old germplasms to find new alleles or new allelic combinations. For this purpose it is necessary to understand crop domestication. Linking molecular and phenotypic evolutions is one of the main challenges of modern genetics to understand domestication. This could contribute to improve crop yield, simultaneously taking the world climactic change into account.

The cultivated tomato (*Solanum lycopersicum* var. *esculentum*) is hypothesized to derive from its closest wild ancestor *Solanum pimpinellifolium* (Nesbitt and Tanksley 2002). Domestication has been accompanied with an explosion of morphological diversity together with a strong erosion of the diversity at the molecular level (van Deynze, Stoffel et al. 2007). The extraordinary phenotypic diversity of fruit shape and size and the few loci controlling their variation (Paran and van der Knaap 2007) is a perfect illustration of this observation. Several genes controlling some of these traits have been cloned: *fw2.2* for fruit weight (Frary, Nesbitt et al. 2000), *ovate* for fruit shape (Liu, Van Eck et al. 2002) and *fasciated* for locule number and fruit size (Cong, Barrero et al. 2008). In tomato fruits, locule number influences several traits such as fruit shape and size. Two major QTLs are responsible of the phenotype: *fasciated* and *lc*. An epistatic interaction between the two QTL drastically influences the phenotype, *lc* being necessary to fully express *fasciated* (Barrero and Tanksley 2004). We herein report the cloning and characterization of *lc* and show that *lc* has a marked diversity pattern and has been necessary to increase locule during domestication.

The *lc* locus, also named *lcn2.1*, was first described by Yeager (1937) and mapped in the ovate region (Butler 1952). The QTL is a major locus known to reduce the number of locules. The QTL location on chromosome 2 was detailed thanks to molecular markers (Lippman and Tanksley 2001; van der Knaap and Tanksley 2003).

We used a recombinant inbred lines population obtained from an intraspecific cross between two cultivated tomatoes: Cervil, a cherry tomato line and Levovil a classic fresh

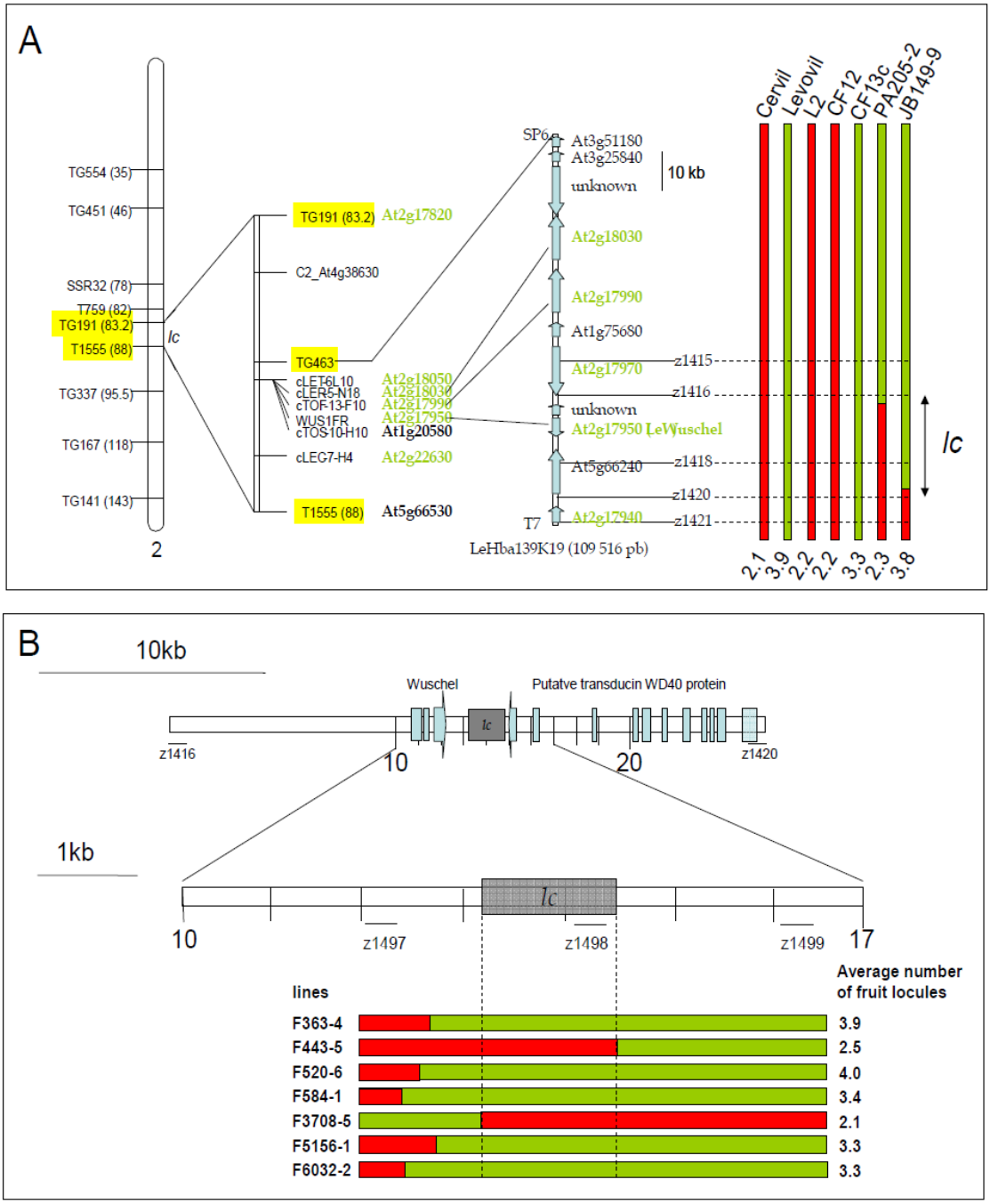


Figure 2: Fine mapping and ultra-high resolution mapping of *lc* QTL

A. Fine mapping of lcn2.1

All sequences in green are those of tomato genes syntenic with the *At2g18000* region in *Arabidopsis thaliana*. Markers highlighted in yellow were used to screen for recombinant plants. PA205-2 and JB149-9 recombinant lines allowed the identification of the BAC clone containing the QTL. LeHBa139K19 contains 11 putative ORF with some of them showing homology to known genes or expressed sequences. *lc* is located in the region between z1416 and z1420.

B. ultra-high resolution mapping of lc

z1416 and z1420 markers have been used to screen for recombinant plants. Design of new markers finally located lcn2.1 between z1497 and z1499. Seven recombinant plants were identified between these two markers. Molecular characterization of these 7 lines restricted lcn2.1 to a 1608 bp uncoding DNA region.

market accession (Causse, Saliba-Colombani et al. 2002). The QTL was first mapped in the region of TG463 (Lecomte, Saliba-Colombani et al. 2004), thanks to two near isogenic lines CF12 and CF13c, genetically identical except for a region of 30 cM containing *lc*. CF12 contains the low locule allele (*lc*) from Cervil; and CF13c the allele coming from Levovil (*LC*) and produces fruits with more locules (figure 1A).

An F2 segregating population (2688 plants) derived from the cross between CF12 and CF13c have been used to identify 215 plants with a recombination event between T1555 and TG191 surrounding *lc* in a 4.8 cM region. The segregation of the phenotype was progeny tested in the self-cross of each recombinant line. This step allowed the mendelization of the QTL (figure 1). The QTL is co-dominant with heterozygous plants having an intermediate phenotype (2.7 locules) between homozygous low loculated plants (2.4 locules) and high loculated plants (3.5 locules). Homozygous plants carrying both alleles from Cervil produce a majority of fruits with two locules and in a very rare cases fruits with more than 3 locules. In contrast, the plants containing the high locule allele from Levovil produce a majority of fruits with 3 or more locules. The QTL being codominant and because of the tight variability of the phenotype, only the homozygous plants were unambiguously discriminated one from each other.

In the fine mapping step, sequence analysis of the region around the closest linked marker TG463 indicates that *lc* region could be syntenic to the region of *A. thaliana* around *At2g1800* (figure 2A). TG463 clone was sequenced; it showed homology with the SP6 sequence of a BAC clone Le_HBa0139K19. The mapping of a deletion in the other BAC-end confirmed the location of the BAC in the *lc* region. PCR amplification also located ctof-13-f10 marker in the BAC. Two recombinant lines PA205-2 and JB149-9 showed a recombination event on the BAC Le_HBa0139K19. PA205-2 produced low loculated fruits (2.3 locules) and JB149-9 high loculated fruits (3.8 locules). We then concluded that Le_HBa0139K19 BAC contained *lc* QTL.

Le_HBa0139K19 (109.5kb) contains 11 putative ORF. Genotyping markers from the BAC showed that PA205-2 and JB149-9 only differ for a 26.6 kb region (figure 2A). This region contains 3 ORF coding for a regulatory protein (transducin WD40 repeat domain protein similar to At5g66240), an unknown protein and *LeWuschel*.

Wuschel (Laux, Mayer et al. 1996; Reinhardt, Frenz et al. 2003) was a good candidate because its function is associated with floral organ number (Mayer, Schoof et al. 1998) and to meristem size by regulating Clavata3 (Brand, Fletcher et al. 2000; Schoof, Lenhard et al. 2000). In a previous work with another progeny (Barrero *et al.*, 2006), Wuschel has been used

as a candidate gene for *lc* but was discarded by the authors due to the lack of polymorphism on either the sequence or its expression. The WD40 repeat protein was also a good candidate because WD proteins are known to be linked to diverse important functions in eukaryotes (Smith, Gaitatzes et al. 1999) such as cell division (Feldman, Correll et al. 1997) or chromatin remodelling (Vitaliano-Prunier, Menant et al. 2008).

To go further in the fine mapping, we decided to genotype 6768 additional F2 plants from the same cross between CF12 and CF13c by using z1416 and z1420 markers surrounding the 26.6kb region containing *lc*. 52 recombinant lines were identified. New markers, designed every 2kb in the 26.6kb region, restricted *lc* to a 3kb region between z1497 and z1499 in which 7 recombinant plants were identified (figure 2B). Sequence analysis of the region restricted *lc* to 1608 bp with 14 polymorphisms (13 SNPs and a 1 bp indel).

This region corresponds to an uncoding region, 1200 bp far from the stop codon of *Wushel*. All *in silico* analysis of the region did not give any information about its putative function. It did not show any homology with known miRNA and had no clear secondary structure which could predict a new one. Expression analysis did not reveal a possible expression of the locus by using either Northern Blot or several primer pairs in RT-PCR experiments (data not shown).

We checked if the *lc* locus could act on the expression of the two adjacent genes by studying their expression pattern in several lines and tissues (figure 3). Wuschel expression was restricted to flower buds compared to the WD repeat protein which was expressed in all tested tissues. We did not reveal any significant differences of expression between lines. These results did not select between the two candidate genes. However, Wuschel expression was restricted to floral buds.

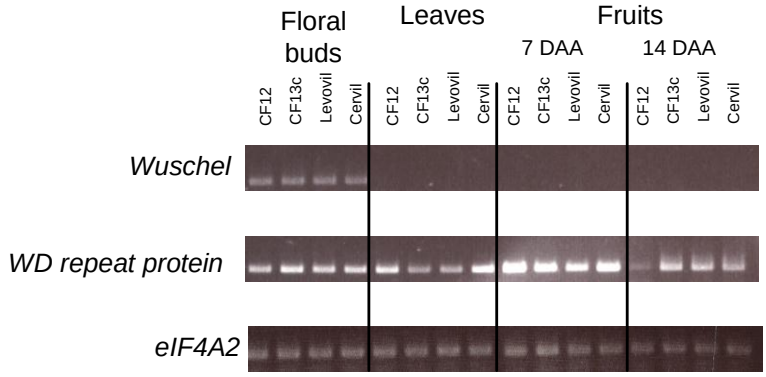


Figure 3: Expression analysis of the two candidate genes
RT PCR was obtained on total RNA extracted from either floral buds, leaves or fruits. Primers were designed on Wuschel (AJ538329 accession number), the WD repeat protein (U332412 tomato unigene). eIF4A2 (U213502 tomato unigene) is used as a reference gene.

We then evaluated the effect of *lc* on floral development. We analyzed floral organ numbers in several representative near-isogenic lines (supplemental data 1). The increase of locule number was systematically associated with an increase of petal number and this increase was not due to an increase of flower size. These results indicate that *lc* functionally acts by increasing the meristem size.

Our results show that the 1608bp region is responsible of *lc* and acts on the flower meristem development. However, we can't identify a clear function, even if the region is located closed to Wuschel.

To go further in the characterization of *lc*, we sequenced the 1608 bp in 89 accessions composed of 16 *S. lycopersicum* var. *esculentum*, 63 *S. lycopersicum* var. *cerasiforme* and 10 *Solanum pimpinellifolium*. The panel of the varieties was shown to cover a large spectrum of tomato diversity (Ranc, Munos et al. 2008) with the genome of *S. lycopersicum* var. *cerasiforme* varieties composed of an admixture of the ones from wild (*Solanum pimpinellifolium*) and cultivated (*S. lycopersicum* var. *esculentum*) accessions. Sequence analysis revealed 25 new polymorphic sites; the majority of them were present in wild species. Two SNPs were found to be associated with locule number (pvalue<10E-6). The correlation was perfect except for 3 lines. Pescio, Muchamiel and Stupicke Polni Rane cultivars produce high loculated fruits (3.8, 5.5 and 4.2, respectively) but contain the low locule allele. These 3 lines contain the other major QTL *fasciated*, explaining their phenotype. To validate the functional effect of these 2 SNPs by association genetics, a 235 bp region containing the 2 SNPs was sequenced in 92 additional lines. The correlation was highly significant (*p*-value<1.30E-12), demonstrating that the two SNPs located in this uncoding sequence are responsible of *lc* which can be considered as a Quantitative Trait Nucleotide (QTN) (Fridman, Carrari et al. 2004). We also check if *lc* should affect fruit weight. The two SNPs explained more than 12 % of fruit weight variation (*p*-value of association < 3.5E-6). In contrast, one SNP in the 5'-UTR of *fw2.2* explained the same variation in the trait but with weaker association between the trait and the polymorphism (*p*-value < 3.0E-4).

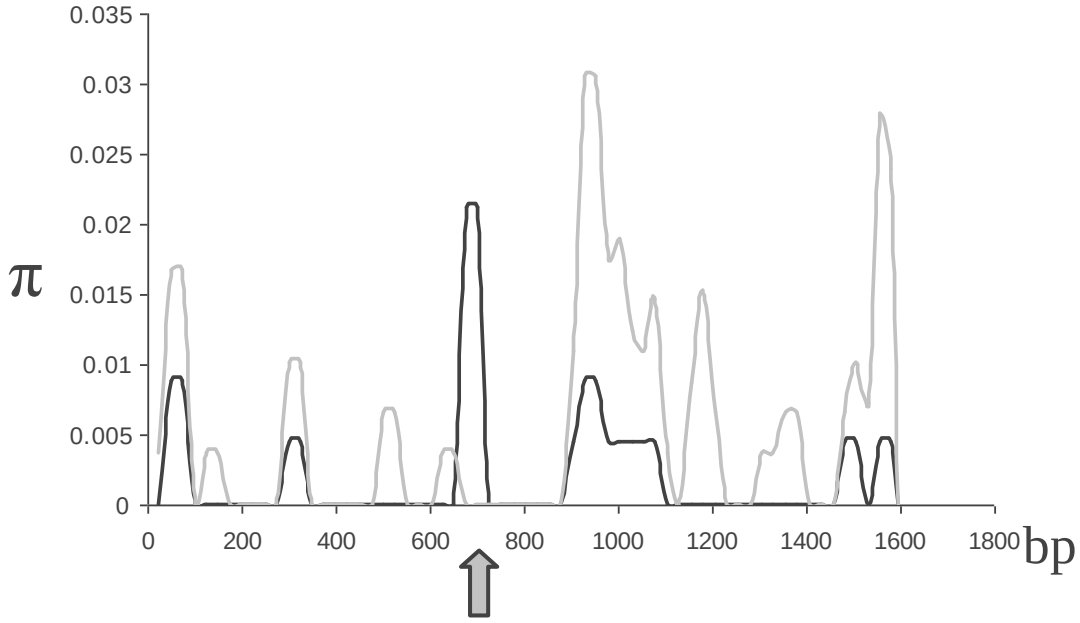


Figure 4: Molecular diversity of lcn2.1 on wild and cultivated lines.
Sliding window analysis of lcn2.1 diversity. Black and gray lines represent cultivated and wild tomato population, respectively. 17 cultivated and 11 wild lines were used. The diversity of the locus is drastically reduced in cultivated lines except for two SNPs (arrow).

Detailed sequence analysis revealed that the diversity of the locus was drastically reduced in the cultivated species except for the two SNPs responsible of *lc* (figure 4, figure 5A). This was particularly clear for a 311 bp window surrounding the 2 SNPs where no diversity was observed in the cultivated accessions. This result indicates that *lc* locus is under a high selection pressure as, in a lesser extend, for others domesticated loci like *tb1* in maize (Wang et al., 1999) or *fw2.2* in tomato (Nesbitt and Tanksley 2002). Analysis of 8 additional loci located on the Le_HBa0139K19 BAC and 16 molecular markers along the entire chromosome 2 indicated that the 2 SNPs responsible for *lc* has evolved differently during domestication. These two SNPs were under balanced selection during tomato domestication (Fig 5B) with Tajima's D = 2.22691, p value<0.05) indicating that human selected both alleles with the same strength. To go further in our investigation, we tried to draw the history of the trait by analysing a new panel of 87 modern cultivars producing fruits with more than 6 locules, considered as fasciated fruits. From these 87 cultivars, only 3 contained the low locule allele of *lc* confirming that *lc* is necessary to fully express the high loculated fruit phenotype.

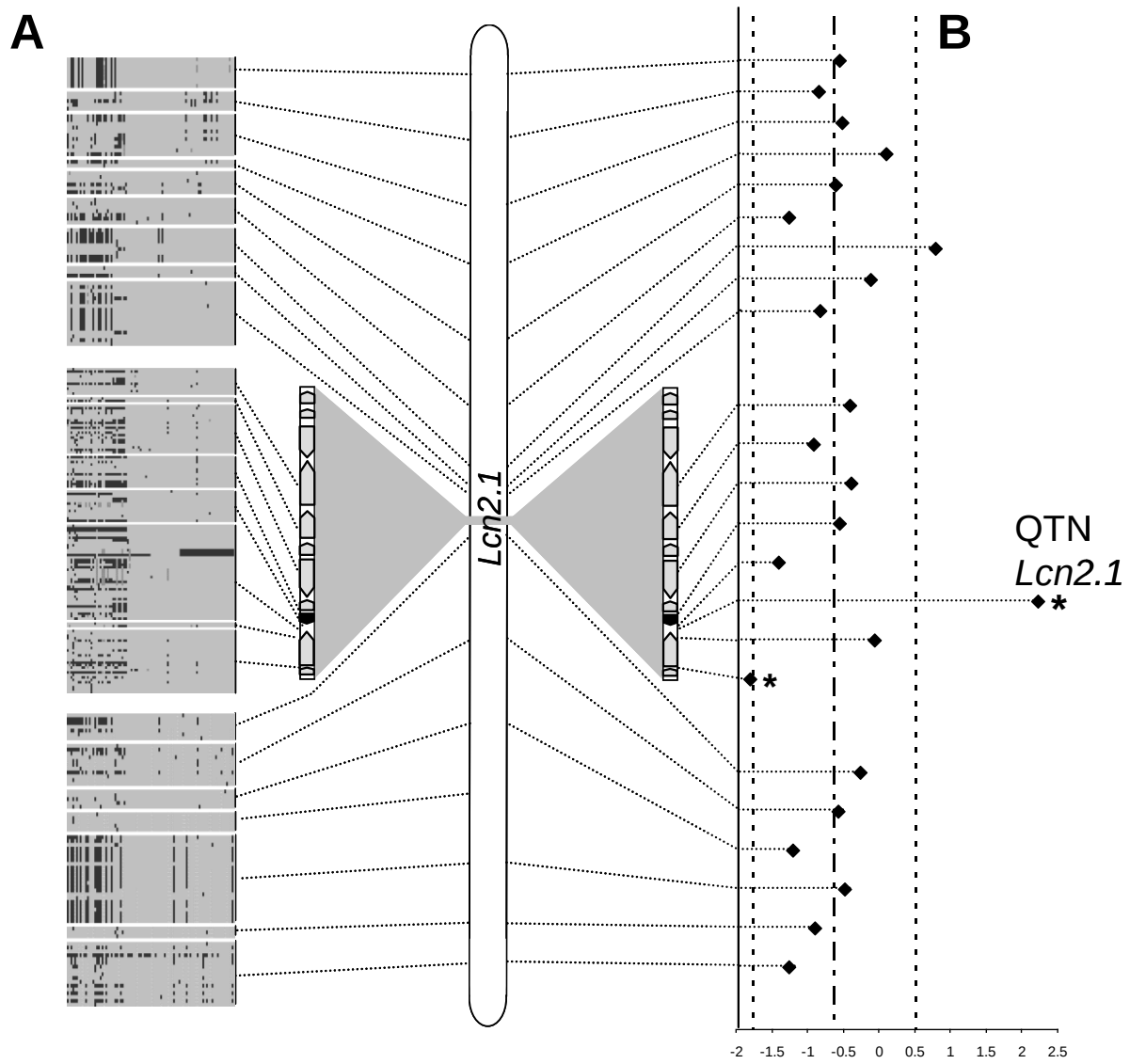


Figure 5: Selection signature detected on lcn2.1.
A. Tomato haplotypic structure of lcn2.1.
Each block corresponds to an amplicon. Columns and rows represent individuals and SNPs respectively. With Heinz1706 used as reference, polymorphisms are indicated either in gray (identical allele) or black (different allele).
B. Selection test over the whole chromosome and on lcn2.1.
Stars indicate significant departure from neutrality hypothesis. A significant positive value of Tajima's D for lcn2.1 indicates either a balanced selection of the locus or a population expansion effect. We looked to the entire chromosome with dashed lines representing means and standard error for the whole chromosome and showed that lcn2.1 has evolved differently from the entire chromosome. lcn2.1 locus undergone a balanced selection whereas chromosome 2 evolved following a population expansion.

Because the first described lines known to be imported from Mexico to Europe produced fasciated fruits (Daunay, Laterrot et al. 2007), we propose a model to explain the evolution of locule number during domestication and breeding (figure 6). From our investigations, around 97% of the varieties producing high loculed fruits have got the allele of lc producing high locule number. Surprisingly, only 42 cultivars (48%) from the 87 cultivar producing fasciated fruits have the *fas* allele. This indicates that *lc* is necessary to fully express fasciated phenotype: not only *fas* but also other loci. Even if it has little effect on the phenotype when it is alone, *lc* is the main locus needed to fully express others. Because almost all the cultivars producing high loculated fruits contain the allele of *lc* producing high locule number, it suggests that *lc* should have appeared before *fas* or other loci mimicking fas phenotype. All tested wild species contain the allele of *lc* producing low locules. This observation makes us think that the two SNPs responsible of the increase of locule number would have appeared during domestication of the *S. lycopersicum* var. *cerasiforme* group. Then, *fas* appeared more recently; the combination of the two loci producing the fasciated phenotype, cultivars introduced in Europe. The modern breeding mixed *lc*, either high or low locule alleles, with other loci.

Although we were unable to identify the function of the 2 polymorphic nucleotides responsible of *lc* at the molecular level, we can assume that they play an important role in meristem development. The region is not expressed but the 2 SNPs have a pleiotropic effect on locule number, floral organ number and fruit weight and should act by regulating meristem size. These QTN are surrounded by two genes acting on development. It is possible that the 2 SNPs could act on the activity of these two genes.

The 2 QTL *fasciated* and *lc* make tomato a good model to study floral meristem and fleshy fruits development. It will be of great interest to look how meristems have naturally evolved. Understanding these mechanisms will allow increasing fruit weight and yield.

METHODS

Plant material and phenotyping

All plants were grown in glasshouses in Montfavet (south of France) between 2004 and 2008.

Fine mapping and ultra-high resolution mapping were performed on a F2 population derived from the cross between the two near isogenic lines CF12 and CF13c (described as F8-V-C and F8-V-L, respectively, in Lecomte *el al.*, 2004). Plants were sown in 96 plates and transferred in 3L container after selection by genotyping. The plants were grown for an

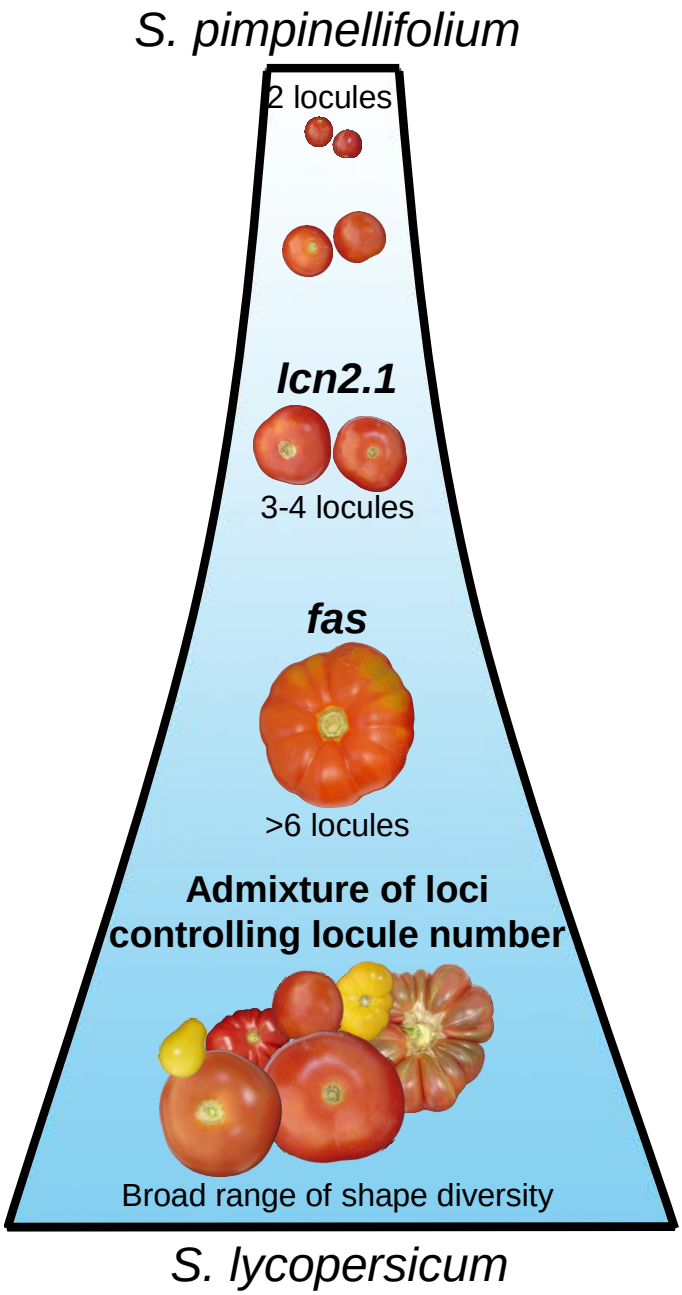


Figure 6: Model of locule number evolution in tomato fruits during domestication.
Solanum pimpinellifolium is the wild ancestor of the cultivated tomato *Solanum lycopersicum*. Based on the analysis of 268 tomato accessions, we propose a model which could explain the history of locule number evolution during tomato domestication. In our study, only 4.6% of the high loculated accessions (i.e. >3 locules) had the low locule allele of *lcn2.1* and 96.9% of the fasciated accessions (i.e. >6 locules) had the high locule allele of *lcn2.1*; but only 49.4% had the *fas* locus. These results indicate that *lcn2.1* has been necessary in the increase of locule number in tomato fruits during domestication. *lcn2.1* could have appeared before *fasciated*. The two QTLs are the major loci controlling locule number. The modern breeding has used other loci to expand phenotypic diversity.

additional week before transfer to glasshouse. For each plant, locule number was determined by phenotyping 20 mature fruits (10 on the second truss and 5 on the third and fourth truss). For each plant, 20 flowers were analyzed to determine petals number and flower diameter.

Genotyping

z100_CAPS and z274 primers were used to amplify T1555 and TG191 markers, the PCR products were digested with *BamHI* and *NdeI* for mapping respectively. The corresponding SNPs were used to develop Taqman markers and screen 2688 F2 plants in the fine mapping. The plants were selected if the genotype at T1555 and TG191 were different. The same markers were used to screen 8 F3 plants from the selfprogeny of each F2 recombinant plant. Two homozygous plants for the segregating marker were then selected for phenotyping.

Polymorphisms in z1416 and z1420 markers were used to develop Taqman marker in order to screen 6768 F2 plants in the ultra-high resolution mapping. For each recombinant F2 plant, the makers were also used to genotype 8 F3 plants from their self-progeny.

Alleles and BAC sequencing

All sequencing reactions were made either manually as described (Le Paslier *et al.*, 2009) or provided from Genome express (11 chemin des pres, 38240 Meylan, France), GATC (http://www.gatc-biotech.com) or MWG (http://www.eurofinsdna.com). TG463 sequence was obtaining by sequencing the plasmid. All other sequencing reactions were performed on PCR products.

LeHBa0139K19 BAC sequencing was performed by Genome Express.

Sequence and polymorphism analysis

Sequence alignment and SNP detection were performed manually using Genalys software (Takahashi *et al.*, 2003), available at http://software.cng.fr.

Accession numbers

Sequences have been submitted to EMBL database (http://www.ebi.ac.uk/embl/). Accession numbers will be available.

Diversity analysis

181 accessions (39 *S. lycopersicum*, 144 *S. lycopersicum var. cerasiforme*, 19 *S. pimpinellifolium*) were phenotyped during summer 2007 and 2008. A part of this sample

represents a core collection of 89 individuals sampled to maximize the diversity of the whole collection (Ranc *et al.* 2008). 24 genomic fragments (8 on the BAC identified by positional cloning and 16 on the rest of the chromosome) were sequenced on this core collection. Logistic regression (Thornsberry et al., 2001) was used for association tests on the 89 individuals' sample. Genetic structure of the sample was taken into account and calculated with Structure2.0 software. We then tested other models to perform association on this sample like in Yu et al. (2006). In order to validate associations we also used a Mixed Linear Model implemented in Tassel using fruit weight as covariable. In order to confirm association we genotyped the two SNPs identified in the whole phenotyped sample.

DNAsp v.4 software allowed molecular diversity (π) to be compared between *S. lycopersicum* (25 accessions) and *S. pimpinellifolium* (8 accessions) using a sliding window analysis methods. Sequences obtained were used to estimate Tajima's D for comparing selection applied to the *lc* locus or to the entire chromosome for cultivated accessions (65 accessions). *lc* fragment was restricted to 500 bp surrounding the SNPs to represent the same length as the others amplicons.

Acknowledgement

Several Master students have helped in experiments: Abir Youssef, David Tricon, Xavier Titeca, and Claire-Emmanuelle Modin. Aurélie Chauveau and Rémi Bounon have helped in allele sequencing. Leydet Florent and Laure David have participated in the fine mapping experiment and in the association genetics, respectively. Hélène Burck, who manages and characterizes all the accessions of the tomato collection at INRA in Avignon, provided the seeds. Several persons, particularly Christophe Giraud, from the greenhouse experimental team of the GAFL help for the growing of all plants. Jean-Paul Bouchet helped on primer design. The Plant Breeding and Genetics Department has financially supported this work. We gratefully thank Esther Van Der Knaap for critical review of the manuscript and for providing one the two markers used to genotype for fas.

References:

Barrero LS, Tanksley SD. 2004. Evaluating the genetic basis of multiple-locule fruit in a broad cross section of tomato cultivars. Theor Appl Genet. 2004 109:669-79.

Barrero LS, Cong B, Wu F, Tanksley SD. 2006. Developmental characterization of the fasciated locus and mapping of Arabidopsis candidate genes involved in the control of floral meristem size and carpel number in tomato. Genome 49:991-1006.

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.

Cong B, Barrero LS, Tanksley SD. 2008. Regulatory change in YABBY-like transcription factor led to evolution of extreme fruit size during tomato domestication. Nat Genet. 40:800-4.

Feldman RM, Correll CC, Kaplan KB, Deshaies RJ (1997) A complex of Cdc4p, Skp1p, and Cdc53p/cullin catalyzes ubiquitination of the phosphorylated CDK inhibitor Sic1p. Cell. 91(2):221-30.

Frary A, Nesbitt TC, Grandillo S, Knaap E, Cong B, Liu J, Meller J, Elber R, Alpert KB, Tanksley SD. 2000. fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science 289:85-8.

Fridman E, Carrari F, Liu YS, Fernie AR, Zamir D (2004) Zooming in on a quantitative trait for tomato yield using interspecific introgressions. Science. 305: 1786-1789.

Laux T, Mayer KF, Berger J, Jürgens G. (1996) The WUSCHEL gene is required for shoot and floral meristem integrity in Arabidopsis. Development. 122(1):87-96.

Lecomte L, Saliba-Colombani V, Gautier A, Gomez-Jimenez MC, Duffé P, Buret M and Causse M 2004. Fine mapping of QTLs of chromosome 2 affecting the fruit architecture and composition of tomato. Mol. Breeding 13: 1-14.

Le Paslier MC, Bounon R, Chauveau A, Thareau V, Canaguier A, Clainche I, Houel C, Moroldo M, Rolland S, Ranc N , Bresson A, Paolucci I (2009) High Throughput SNPs Discovery In Grapevine, Poplar And Tomato. Plant and Animal Genome XVII Conference, January 2009, San Diego, US.

Lippman Z, Tanksley SD. 2001. Dissecting the genetic pathway to extreme fruit size in tomato using a cross between the small-fruited wild species Lycopersicum pimpinellifolium and L. esculentum var. Giant Heirloom. Genetics 158:413-22.

Liu J, Van Eck J, Cong B, Tanksley SD. 2002. A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. PNAS 99:13302-6.

Mayer KFX, Schoof H, Haecker A, Lenhard M, Jürgens G and Laux T (1998) Role of WUSCHEL in regulating stem cell fate in the Arabidopsis shoot meristem, Cell 95, pp. 805–815

Nesbitt TC and Tanksley SD (2002). "Comparative Sequencing in the Genus Lycopersicum: Implications for the Evolution of Fruit Size in the Domestication of Cultivated Tomatoes." Genetics 162(1): 365-379.

Paran I, van der Knaap E. 2007. Genetic and molecular regulation of fruit and plant domestication traits in tomato and pepper. J Exp Bot. 58:3841-52.

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

Reinhardt D, Frenz M, Mandel T, Kuhlemeier C. 2003. Microsurgical and laser ablation analysis of interactions between the zones and layers of the tomato shoot apical meristem. Development 130:4073-83.

Smith TF, Gaitatzes C, Saxena K, Neer EJ (1999) The WD repeat: a common architecture for diverse functions. Trends Biochem Sci. 1999 May;24(5):181-5. Trends Biochem Sci. 24(5):181-5.

Takahashi M, Matsuda F, Margetic N, Lathrop M (2003) Automated identification of single nucleotide polymorphisms from sequencing data. *J Bioinform Comput Biol* 1:253-265.

Thornsberry, J. M., M. M. Goodman, Doebley J, Kresovich S, Nielsen D and Buckler ES (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet 28: 286-289.

Van Deynze A, Stoffel K, Buell CR, Kozik A, Liu J, van der Knaap E, Francis D. 2007. Diversity in conserved genes in tomato. BMC Genomics 8:465.

Vitaliano-Prunier A, Menant A, Hobeika M, Géli V, Gwizdek C and Dargemont C (2008) Ubiquitylation of the COMPASS component Swd2 links H2B ubiquitylation to H3K4 trimethylation.. Nature Cell Biology 10:1365-1371.

Wang RL, Stec A, Hey J, Lukens L, Doebley J. (1999) The limits of selection during maize domestication. Nature. 1999 398:236-239.

La figure supplémentaire est donnée en annexe 5.

4.4. Conclusion

L'utilisation de la diversité naturelle présente dans la « core collection » a permis de simplifier rapidement l'intervalle autour du QTL de 1600 bp (ce qui reste une résolution tout à fait intéressante) à deux SNP distants de 6 bp. Nous sommes donc passés de la notion de Quantitative Trait Locus à la notion de Quantitative Trait Nucleotides, ce qui aurait été relativement lourd sans l'utilisation de ces ressources génétiques.

Comme nous l'avons vu dans l'introduction de ce chapitre, le nombre de loges du fruit de tomate a beaucoup évolué depuis la domestication du fruit. En effet, on est passé d'accessions sauvages dont les fruits possèdent deux loges à des accessions domestiquées où une grande variabilité en termes de nombre de loges est observée. On retrouve des cultivars avec des gros fruits possédant deux loges (en majorité dans les accessions d'industrie), des cultivars avec des fruits possédant 3 à 5 loges, puis des phénotypes extrêmes où les fruits sont complètement déformés à cause d'un trop grand nombre de carpelles (figure 4-7).

Figure 4-7. Photo d'un fruit de tomate contenant plus d'une trentaine de loges. Le fruit est complètement déformé dès le début du développement (lignée LA409).

Pour que le phénotype *fasciated* s'exprime pleinement, il est nécessaire d'avoir l'allèle « fort nombre de loge » au locus *lcn2.1*. Afin que certaines accessions qui présentent le phénotype *fasciated* aient pu être sélectionnées pour donner des fruits de très grande taille, il a

donc fallu que ces lignées soient fixées au niveau du locus *lcn2.1*. Il est donc évident que la mutation impliquant un fort nombre de loges au locus *lcn2.1* est antérieure en terme évolutif à la mutation impliquant un fort nombre de loges au locus *fas*. Les deux SNP, impliqués dans la variation du nombre de loges, montrent un pattern évolutif différent de celui du reste du chromosome. Le D de Tajima moyen du chromosome est de -0,633 ce qui implique une expansion démographique récente. Les deux SNP présentent un D de Tajima qui s'écarte significativement de la moyenne chromosomique. Ce patron n'est donc pas dû à l'histoire démographique de l'échantillon mais plutôt à une histoire sélective particulière. Le D de Tajima est positif ce qui implique une sélection balancée. Ce type de sélection est expliqué par un maintien, à fréquence équilibrée, des deux allèles au locus. Ce maintien par l'homme, des deux états alléliques, peut être expliqué par l'effet du locus qui crée une variation au niveau d'un caractère morphologique. L'homme a donc sélectionné la variabilité morphologique du fruit de tomate. L'augmentation de la variabilité morphologique de l'organe consommé, par sélection, a déjà été observée chez d'autres espèces (Frary and Doganlar 2003).

Malgré le fait que ces deux SNP semblent important pour le développement du fruit, aucune fonction n'a pu leur être associée. Le mécanisme d'interaction entre *lcn2.1* et *fas* n'a pas pu être défini. Cette région doit intervenir dans la régulation d'un ou plusieurs gènes. La structure tridimensionnelle de l'ADN peut avoir une influence sur la régulation de certains gènes. En effet, la succession des quatre bases A, C, T et G forme des ventres et des dos à la surface de l'amas de chromatine. Certaines séquences sont donc plus accessibles que d'autres pour se lier à des protéines régulatrices. Cette étude a montré que des mutations au niveau d'une séquence non codante, mais contrainte au niveau évolutif, pouvaient avoir des effets drastiques sur la topographie locale de l'ADN ce qui pouvait affecter certaines fonctions biologiques (Parker, Hansen et al. 2009). Il serait donc envisageable que *lcn2.1* se situe sur un site de fixation d'une protéine régulatrice et que les deux mutations diminuent ou augmentent l'affinité de la région avec une protéine de régulation. L'interaction entre une protéine régulatrice et *lcn2.1* sera difficile à mettre en évidence étant donné la précocité et la fugacité de la mise en place du phénotype.

Un point important concernant le déséquilibre de liaison semble intéressant à soulever. En effet, ce locus est un cas typique où la génétique d'association permet de valider un candidat déjà identifié. Il semble, cependant, qu'il aurait été difficile voire impossible d'affiner la région contenant *lcn2.1* uniquement en utilisant la génétique d'association car les

deux SNP ne sont pas en déséquilibre de liaison avec les polymorphismes adjacents (Figure 4-8).

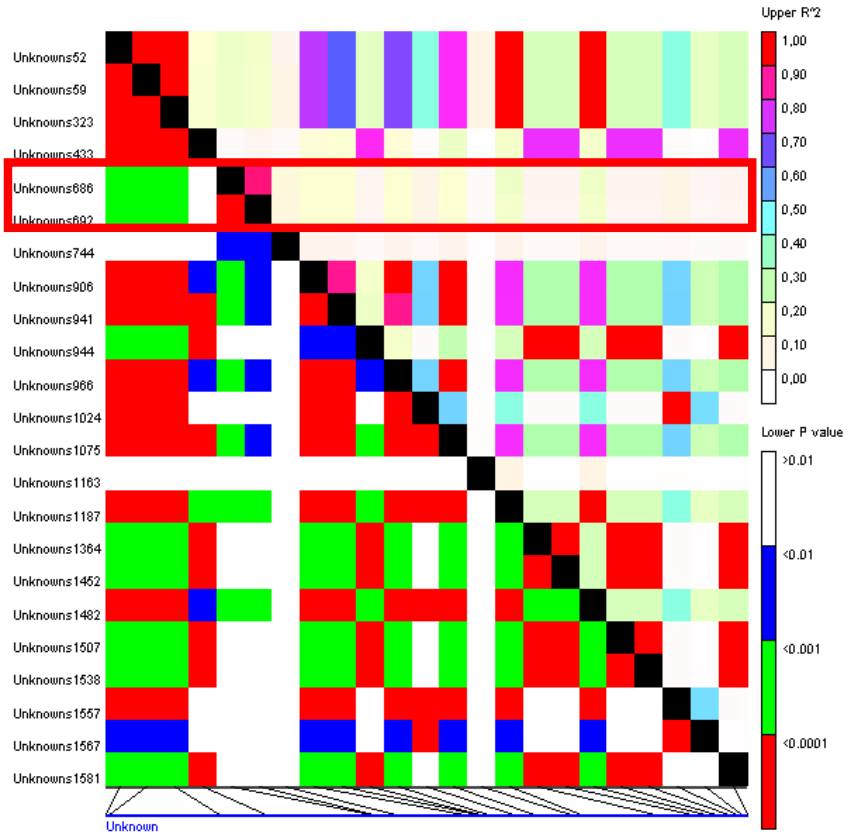


Figure 4-8. Représentation du déséquilibre de liaison au niveau des sites polymorphes de la séquence de *lcn2.1*. Le cadre rouge indique les deux SNP associés avec le nombre de loges.

Par contre la combinaison des approches de cartographie haute résolution et de génétique d'association permet d'identifier les polymorphismes causaux. Cette étude montre tout l'intérêt d'utiliser les ressources génétiques afin d'identifier les polymorphismes causaux. Il est aussi possible d'identifier de nouveaux individus présentant des combinaisons alléliques intéressantes en vue de créer d'autres populations de cartographie. L'utilisation de ces ressources a aussi permis d'étudier le patron de diversité des différents marqueurs et d'identifier le type de sélection exercée par l'homme lors de la domestication. Cependant aucune information n'est disponible sur la capacité à identifier de nouvelles zones du génome, impliquées dans la variation de caractère d'intérêt, sans *a priori* sur la localisation. Le chapitre suivant se focalise sur le chromosome 2 afin d'utiliser une densité en marqueurs

suffisante pour évaluer la résolution potentielle d'une étude d'association de type Whole Genome Association.

Chapitre 5 : Etablissement des conditions optimales permettant de réaliser des études d'associations chez la tomate

5.1. Introduction

Les études d'association ont montré tout leur potentiel chez le maïs ainsi que chez *A. thaliana* (Aranzana, Kim et al. 2005; Yu, Pressoir et al. 2006; Zhao, Aranzana et al. 2007). Chez les espèces hautement autogames, très peu d'études ont déjà abouti à l'identification du polymorphisme causal par association. Chez la tomate, nous avons déjà démontré l'intérêt d'utiliser des ressources génétiques pour étudier une région génétique impliquée dans la variation d'un caractère quantitatif. Cependant, lors de cette étude, la région avait déjà été isolée par clonage positionnel. Aucune étude n'a donc été réalisée sans *a priori* sur la région à cibler.

Avant de définir un grand nombre de marqueurs sur la totalité du génome, il nous semblait plus opportun de se concentrer sur un seul chromosome. Nous nous sommes donc focalisés sur le chromosome 2 qui présente plusieurs QTL liés à la qualité, organisés en cluster. Ce chromosome porte notamment le QTL *lcn2.1* déjà identifié, ainsi que d'autres QTL en cours de clonage, contrôlant la teneur en sucres et le poids du fruit. Des fragments génomiques placés sur tout le chromosome ont été séquencés sur la « core collection » de 96 accessions. Ces fragments ont été répartis tous les cinq cM, puis certaines régions d'intérêt ont été densifiées.

Nous avons comparé le taux de polymorphisme détecté chez les trois groupes constituant la core collection : le type *S. l. cerasiforme*, le type *S. l. esculentum* et l'espèce *S. pimpinellifolium*. L'étendue du déséquilibre de liaison a été étudiée sur l'ensemble du chromosome et sur quatre régions physiques afin d'estimer la résolution d'une analyse d'association de type « Whole Genome Analysis » sur un échantillon de tomates composé majoritairement d'accessions de type cerise. Nous avons ensuite réalisé les études d'associations entre les polymorphismes identifiés et les caractères étudiés.

Sur les 92 accessions de la « core collection », seulement 90 accessions ont été phénotypées durant les deux années consécutives. L'article suivant se concentre uniquement sur trois phénotypes : le poids du fruit, le nombre de loges et le contenu en solides solubles.

Une analyse étendue sur les autres caractères suivra l'article.

5.2. Genome admixture of *Solanum lycopersicum* var. *cerasiforme* allows successful association mapping in tomato (*Solanum lycopersicum*), an inbred crop.

Nicolas Ranc*, Stéphane Muños*, Marie-Christine Le Paslier§, Aurélie Chauveau§, Rémi Bounon§, Sophie Rolland*, Jean-Paul Bouchet*, Dominique Brunel§, Mathilde Causse*.

*INRA, UR1052, Unité de Génétique et Amélioration des Fruits et Légumes, Avignon, 84143, France
§INRA, UR1279, Unité Etude du Polymorphisme des Génomes Végétaux, CEA-Institut de Génomique-CNG, Evry, 91057, France

Abstract

Linkage disequilibrium mapping is an efficient tool to dissect molecular bases of interesting phenotypes but suffers severe limits when dealing with inbred crops like tomato (*Solanum lycopersicum*). Cultivated tomato exhibits low molecular polymorphism and high linkage disequilibrium extent, reducing mapping resolution. Cherry type tomato (*S. lycopersicum var. cerasiforme*) genome has been described as admixture between cultivated tomato and its wild ancestor. We have harnessed this admixture to increase resolution of association mapping. We sequenced 81 DNA fragments, spread over the chromosome 2, on a tomato core collection (N=90), mainly composed of cherry type tomato accessions, that was also phenotyped for fruit weight, fruit locule number and fruit soluble solid content. We assessed the structure of molecular polymorphism and the extent of linkage disequilibrium over genetic and physical distances. A large set of polymorphisms (340 SNPs and Indels) was detected and *S. l. cerasiforme* showed a higher rate of polymorphism than the cultivated or wild groups. Linkage disequilibrium decreased under r^2=0.3 within 1 cM and minimal estimated values (r^2=0.13) were reached over 20Kb over the four physical regions studied. Associations of polymorphism with phenotypes were detected with structured association methods. We validated previously identified candidate genes and QTL, and we found associations with new QTL and new candidate genes. We thus showed the efficiency of genome admixture to overcome the low-resolution limitation of association mapping for an inbred crop.

Introduction

Linkage mapping has proven its usefulness in detecting important qualitative and quantitative loci in crops (Doebley, Stec et al. 1997; Frary, Nesbitt et al. 2000). Linkage mapping strategies are limited in detecting genes underlying quantitative traits (QTL) because only two extreme parents are usually used for generating the segregating population and because of modest degree of recombination within the population (Flint-Garcia, Thuillet et al. 2005). Furthermore, discovery of new genes seems to be limited to those having a large effect on the phenotype variation (Buckler and Thornsberry 2002).

Most of the natural occurring diversity selected during domestication of crops is conserved in seed-bank collection. These resources present a large number of accessions with different histories, accumulating mutation and recombination events. Even if they represent treasure of phenotypic and molecular diversity, germplasm collections are poorly used for breeding. Association mapping strategy offers possibility to identify polymorphisms implied in phenotype variations using natural populations and could provide efficient valorization of genetic resources. A highly valuable literature is available today for plant geneticists about association mapping (Gupta, Rustgi et al. 2005; Zhu, Gore et al. 2008). New statistical methods have been developed to deal with structured samples (Pritchard, Stephens et al. 2000; Price, Patterson et al. 2006; Yu, Pressoir et al. 2006) and these methods have been efficiently applied to plants (Thornsberry, Goodman et al. 2001; Flint-Garcia, Thuillet et al. 2005; Zhao, Aranzana et al. 2007).

One of the most important parameter in association mapping is the intensity of linkage disequilibrium (LD) over the genome. LD is defined as non random association of alleles and determines the resolution of association mapping (Rafalski 2002). If LD extends within several hundreds of base-pairs (bp), a high number of markers are needed to cover the whole genome and alleles at selected candidate genes may be tested for association. If LD extends over higher distances, whole genome may be scanned with a lower density of markers, in order to identify regions that are associated with phenotype variation. The extent of LD over the genome is expected to be variable according to the species, the genome region and the population studied. The variation is attributable to both the history of recombination and to the history of mutations (Nordborg and Tavare 2002). LD is created by mutation, is increased by selection, drift and admixture between heterogeneous populations. Only recombination can break intra-chromosomal LD whereas inter-chromosomal LD is reduced by random assortments. LD is expected to be stronger for inbred than for outbred species, as recombination is less effective in selfing species where individuals are more likely to be

homozygous at a given locus than in outcrossing species (Flint-Garcia, Thornsberry et al. 2003). Demography acts on LD at the whole genome scale. For example, reduction in population size (bottleneck) increases drift effect and increase LD within and between chromosomes. Thus, inbred crops are theoretically less suitable for fine mapping because of a low level of molecular diversity and high overall genomic LD due bottlenecks.

Cultivated tomato (*Solanum lycopersicum* var. *esculentum*, formerly *Lycopersicon esculentum*) is a perennial plant, diploid, predominantly selfing and highly inbred. Tomato was domesticated from its wild relative *S. pimpinellifolium* with the first domesticated form presumably represented by *S. lycopersicum* var. *cerasiforme* (i.e. the cherry tomato). The modern cultivated tomato accessions exhibit a low amount of genetic diversity compared to wild relatives due to several bottlenecks undergone during domestication and exacerbated by its autogamous nature (Yang, Bai et al. 2004; van Deynze, Stoffel et al. 2007). As expected, LD extends along high genetic distances in the cultivated accessions (van Berloo, Zhu et al. 2008). Hopefully, part of the *S. lycopersicum* var. *cerasiforme* (*S. l. cerasiforme*) accessions display genetic admixture pattern between cultivated and wild tomato accessions (Ranc, Munos et al. 2008). Recent admixture of population with different allele frequencies tends to increase LD but consecutive inter-mating cycles have increased genetic diversity and decreased LD. Thus, admixture population could be compared to advanced intercross lines, i.e. populations derived from two inbred strains that were randomly intercrossed for several generations. As a consequence, cherry type tomatoes have higher level of genetic diversity than *S. l. esculentum* and higher phenotypic diversity than *S. pimpinellifolium* offering interesting properties for association mapping. Admixture mapping concept was previously used in human genetics to increase LD (Darvasi and Shifman 2005). In tomato, such strategy draws advantage for LD mapping.

Association mapping have been used to identify molecular bases of QTLs in tomato, in regions encompassing map-based cloned gene. Nesbitt and Tanksley (2002) failed to find any association between fruit size and genomic sequence of the *fw2.2* region in a collection of 39 cherry tomato accessions. This sequence had been previously cloned by positional cloning and its implication in fruit weight variation had been validated (Frary, Nesbitt et al. 2000). Nesbitt and Tanksley (2002) did not succeed in identifying the causal polymorphism for *fw2.2* but they demonstrated that the locus of cherry tomato accessions was a mosaic between *S. pimpinellifolium* polymorphisms and *S. lycopersicum* polymorphisms. Recently, association

mapping showed relevance, in tomato, to identify quantitative trait nucleotide (QTN) responsible of locule number difference between *S. l. cerasiforme* and *S. l. esculentum* (Muños, Ranc et al. *in prep.*). A sequence of 1800 bp containing the QTL lcn2.1 has been identified by map-based cloning. LD mapping detected two SNPs, within this sequence, perfectly associated with the phenotype variation.

Our objective was to identify the optimal conditions for whole genome association in tomato by (i) testing admixture LD mapping on tomato and (ii) assessing the marker density needed to perform association mapping in this crop. In this preliminary study, we focused on chromosome 2 because clusters of QTLs for fruit morphology and quality traits were previously mapped on this chromosome using cherry tomato (Causse, Saliba-Colombani et al. 2002) and other distantly related species (Tanksley, Grandillo et al. 1996). Several genes underlying these QTLs have been cloned as, for example, *fw2.2* responsible for fruit weight variation (Frary, Nesbitt et al. 2000), *Ovate* causing pear-shaped tomato fruit (Liu, Van Eck et al. 2002), *Cnr* causing non ripening fruit (Manning, Tor et al. 2006), and *lcn2.1* responsible of locule number (Muños, Ranc et al. *in prep.*). Ninety accessions from a core collection mainly composed of *S. l. cerasiforme* accessions, were genotyped by direct sequencing of DNA fragments. We sequenced 80 fragments mapped on chromosome 2 and spread over three different mapping densities: (i) a whole chromosome density (1 fragment/5cM), (ii) a fine mapping density (1 fragment/cM) and (iii) a physical mapping density (1 fragment/100Kb) in four regions where QTL were previously fine mapped (Lecomte, Saliba-Colombani et al. 2004). We thus described the amount of molecular polymorphism detected, assessed the extent of linkage disequilibrium over the entire chromosome and over physical distances and performed preliminary association tests with fruit weight, locule number and soluble solid content.

Material & Methods

Plant material: Tomato accessions were sampled from a germplasm collection maintained and characterized at INRA Avignon (France). These accessions are part of a core collection and maximize both genetic and phenotypic diversity (Ranc, Munos et al. 2008). This sample is composed of 63 cherry type tomato accessions (i.e.: *S. lycopersicum* var. *cerasiforme*, hereafter named *S. l. cerasiforme*), 17 large fruited accessions (*S. lycopersicum* var. *esculentum*, hereafter named *S. l. esculentum*) and 10 *S. pimpinellifolium* accessions. Accessions come from French researchers' prospecting, from breeders' collections, from the Tomato Genetics Resource Center (Davis, California, USA), the Centre for Genetic

Resources (Wageningen, Netherlands), the North Central Regional Plant Introduction Station (Ames, Iowa, USA) and from the N.I. Vavilov Research Institute of Plant Industry (St Petersburg, Russia).

DNA fragments sequenced: We designed pairs of primers for 120 fragments based on sequence data from genes and markers mapped on chromosome 2 (http://solgenomics.net/). These fragments were chosen to cover the entire chromosome with three different densities: (i) fragments every 5 cM chosen to cover the whole chromosome, (ii) fragments every cM chosen to cover the fine map of the middle of the chromosome and (iii) fragments every 100 Kb chosen to cover four major physical contigs mapped on candidate regions for fruit quality QTLs identified on a *S. l. cerasiforme* x *S. l. esculentum* cross (Lecomte, Saliba-Colombani et al. 2004). We used preliminary results of whole genome sequencing to draw physical contigs. Contig1 (614,386 bp) mapped in a sugar content QTL (sugs2.1) and was built with clones C02SLe0031D11, C02HBa0013N18, C02HBa0060J03, C02SLe0127J16, C02HBa0030D08, C02HBa0009K06 and C02HBa0056D15. Contig2 (434,250 bp) mapped in a locule number QTL (lcn2.1) and was built with clones C02SLm0132H19, C02HBa0044J01, C02HBa0130B04 and Le_HBa0139K19. Contig3 (492,698 bp) mapped in a soluble solid content (ssc2.2) and was built with clones C02HBa0164H08, C02SLm0097L01, C02SLm0128E12, C02HBa0074A14, C02HBa0030A21 and C02HBa0213A01. Contig4 (187,646 bp) mapped in a fruit weight QTL (fw2.2) and was built with BAC clones Fw2.2, CO2Hba0208N01 and CO2Hba0073P13 (a gap of unknown size localize between the latter and the former). Because a low amount of polymorphism was previously described in *S. lycopersicum*, we focused on intronic or intergenic regions for sequencing. For a specific unigene, intron localization was predicted with TblastX on *Arabidopsis thaliana* genomic sequence and primers were designed on exonic sequence surrounding introns. A specific bio-informatic program was designed for this purpose and is available upon request (Bres, Bouchet et al. 2005).

Fragment sequencing and analysis: DNA was isolated from 100 mg frozen leaves using to the DNeasy Plant Mini Kit (Qiagen, Valencia, California, USA). Amplification reactions were performed in a final volume of 5 µL in the presence of 2.5 ng of template DNA, 0.4 pmol of each primer, 0.05 mM concentration of each deoxynucleotide, 2 mM MgSO4, 1 X *Taq* polymerase buffer Pand 0.03 unit of Platinum *Taq* HiFi (Invitrogen, Carlsbad, CA). After 5 minutes at 94°C, 30 cycles were performed of 20 s at 94°C, 20 s at 55°C, 2 min at 68°C, followed by a final extension step of 5 min at 68°C. Couples of primer with single-band PCR product were chosen for sequencing. PCR products were purified by ExoSAP method with

Exonuclease I (NEB, Beverly, MA) and Shrimp Alkaline Phosphate (USB, Cleveland, Ohio), sequenced with SP6 universal primer in an adapted 5 µL reaction volume method using BigDye terminator V3.1 and analyzed on an ABI 3730 xl sequencer (Applied Biosystem, Foster City, CA). Only 86 pairs of primers gave unique band for PCR profile and were chosen for sequencing in forward sense on 90 tomato accessions. Sequence alignment and SNP detection as performed using Genalys software available at http://software.cng.fr/ (Takahashi, Matsuda et al. 2003). Five fragments were not readable because of heterozygous signal probably due to amplification of paralogous sequences. SNPs information is available in supplemental Table S1. Sequences of *lcn2.1* for this core collection were added in this study.

Tomato phenotyping: The 90 accessions were grown during 2007 and 2008 summers in Avignon (South of France). Four plants per accession were bred in plastic greenhouse. Three harvests of ten ripe fruits were done for each accession. The ten fruits were phenotyped for fruit weight (FW), locule number (LCN) and soluble solid content (SSC). For phenotypic traits, year and accession effect were analyzed using Anova implemented in R software. Heritabilities were calculated as $h_F^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2 / 2)$ with σ_g^2 and σ_e^2 the genetic and residual variance, respectively. σ_g^2 and σ_e^2 were estimated by (MSc-MSe)/2 and MSe, respectively, with MSc and MSe the mean square of cultivars and residual effects. Because accessions' effect over the two years was significantly much higher than year's effect, we used accessions adjusted mean for association studies calculated by the "all.effects" procedure implemented in the R package 'effects'.

***Linkage disequilibrium*:** The program Tassel (Bradbury, Zhang et al. 2007) was used to estimate the LD parameter r² among loci and the comparison-wise significance was computed by 1000 permutations. The decrease of LD over genetic distance was fitted by the equation: $y = a + b\mathrm{e}^{-c/x}$ using non-linear regression as implemented by the "nls" procedure in R (R Development Core Team 2005) where y is r² and x the genetic distance in cM or Kb (Tenesa, Wright et al. 2004).

Association analysis: Several statistical models were tested: (i) Simple Linear Model, (ii) Structured Association Model (Q model) and (iii) Mixed linear model (K+Q model) described by Yu et al. (2006) and implemented in Tassel software. The model with the smaller type I error rate was chosen for further association studies. Population assignment of individuals was inferred by STRUCTURE2.1 software (Pritchard, Stephens et al. 2000) based on 20 Simple Sequence Repeat (SSR) markers spread over the genome (Ranc, Munos et al. 2008). For inferring the most likelihood number of population, we used Evanno (2005) transformation

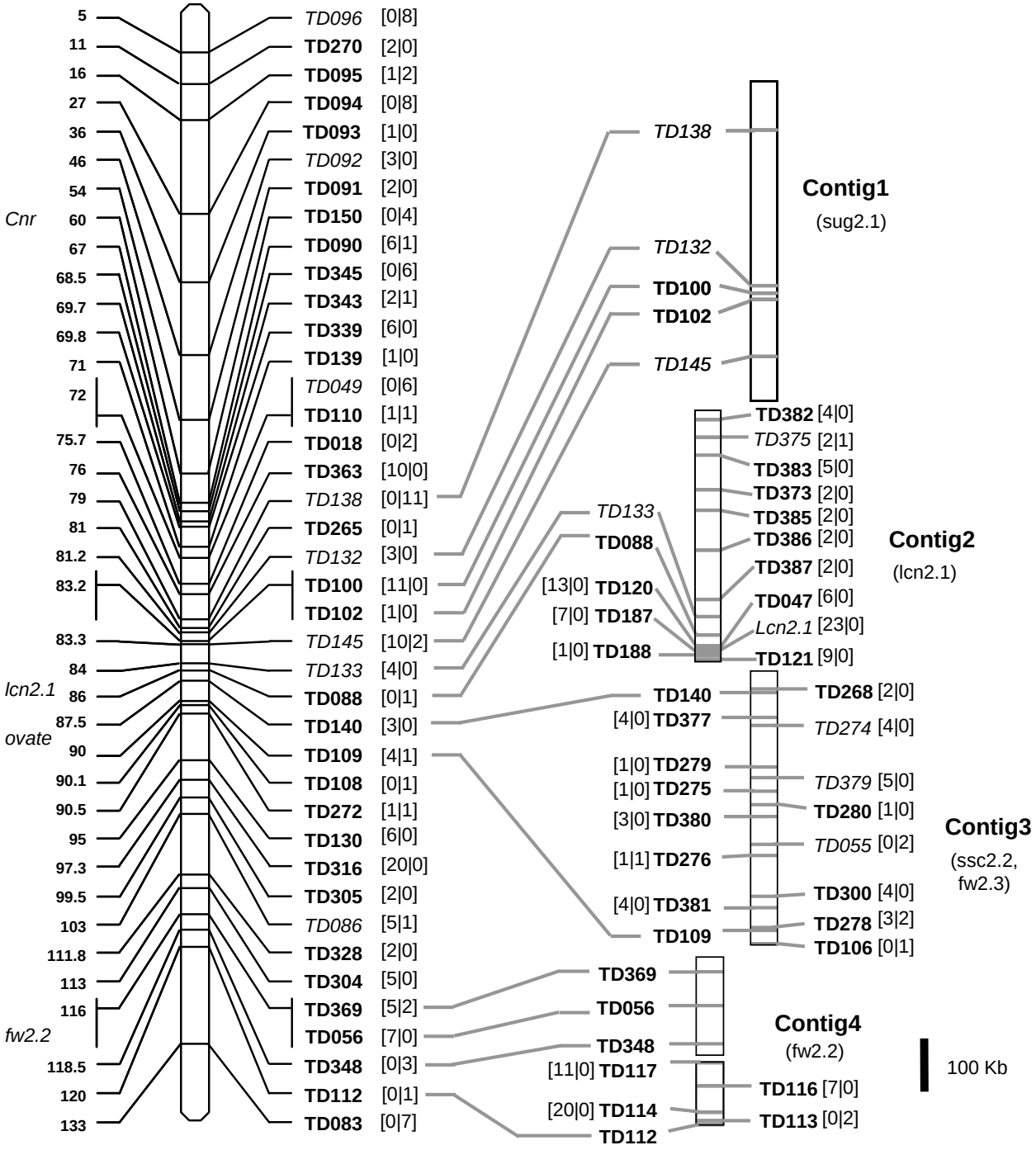


FIGURE 1. Genetic and physical location of the polymorphic fragments sequenced on chromosome 2. Genetic distances on the expen2000 reference map are indicated on the left of the chromosome. Physical contigs are drawn on the right of the scheme with their name and the name of the QTL targeted within parenthesis. Numbers of polymorphisms (SNPs and Indels) found in non-coding and coding regions are indicated within bracket in the first and second position, respectively. Markers in italic showed high LD when compared together. Cloned QTL are indicated on the left of the chromosome.

method. The matrix of relative kinship coefficients of Ritland (1996) implemented in the mixed linear model was estimated using SPAGeDi (Hardy and Vekemans 2002) based on the same set of SSR markers. To deal with multiple testing, we computed adjusted *p*-values using Benjamini & Hochberg (2000) procedures to control for the false discovery rate. Significant association was detected with adjusted *p*-value lower than 0.005. We choose stringent adjusted *p*-value threshold to show only the strongest association. For markers that were significantly associated with a trait, a general linear model with all fixed-effects terms was used to estimate the amount of phenotypic variation explained by each of the candidate markers, as measured by R^2. The standardized effect of each marker was also calculated by dividing the difference between the two homozygous classes by the phenotypic standard deviation of that trait (Weber, Briggs et al. 2008). Accession Heinz 1706, the cultivar used for tomato genome sequencing was used as reference for allele effect calculation.

Results

Polymorphism identification: Single Nucleotide Polymorphism (SNP) and Insertion-Deletion (InDel) were detected among 90 accessions. Only polymorphisms with Minimum Allele Frequency (MAF) higher than 5% were taken into account. The average size of sequences was 542 bp. The sequencing primarily focused on non-coding regions which represented more than 69% (30,396 bp) of the total length sequenced (44,223 bp). Characteristics of the 81 readable fragments are given in Table S2. Eleven fragments (13%) were not polymorphic. Figure 1 shows the location and polymorphism content of the 70 polymorphic fragments. A total of 300 SNPs and 52 Indels were detected. Polymorphisms were analyzed according to species membership of accessions (Table 1).

Table 1. Distribution and frequencies of polymorphisms (SNP and Indel) across species and ration of polymorphism in coding and non coding region. All fragments (81) are taken into account.

	number of access.	number of shared polymorph.[a]			number of total polymorphic site	Polymorph. frequency for 1000 bp		non-coding / coding polymorphisms ration
		esc	*cera*	*pimpi*		coding	non-coding	
esc	17	0			157	1.66	4.27	2.57
cera	63	11	5		349	5.42	8.61	1.59
pimpi	10	0	187	3	336	5.27	8.25	1.57

[a] Numbers in diagonal indicate species specific polymorphisms

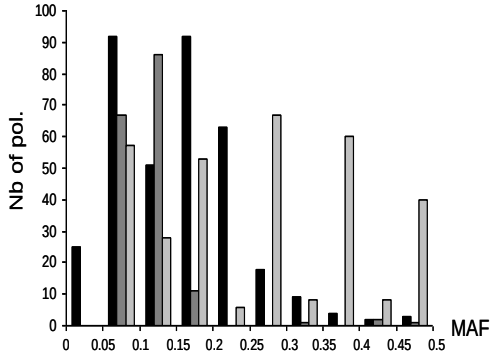


FIGURE 2. Distribution of polymorphism minimum allele frequencies (MAF) among tomato species. *S. l. cerasiforme* (N=63) is represented in black, *S. l. esculentum* (N=17) in dark gray and *S. pimpinellifolium* (N=10) in light gray. Polymorphisms with overall species MAF lower than 0.05 were previously discarded (see Method).

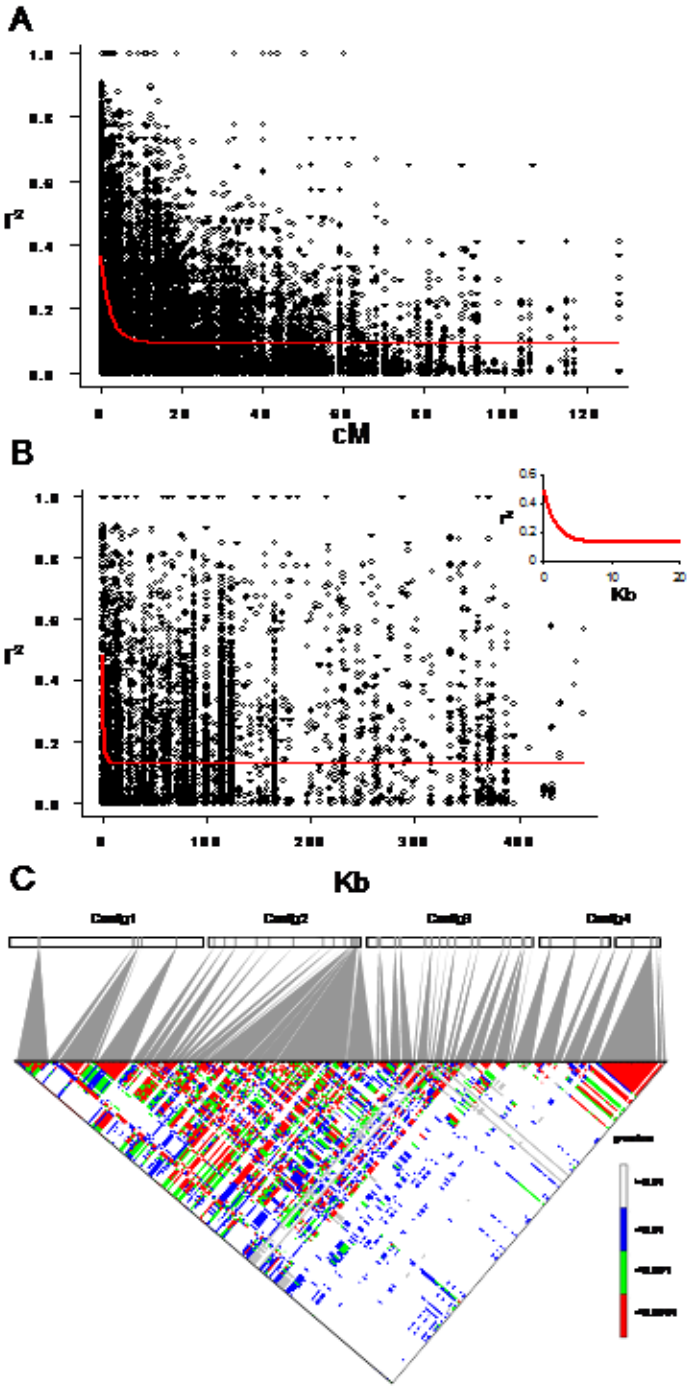


FIGURE 3. Estimates of r² *vs.* genetic and physical distance on chromosome 2 for the 64 *S. l. cerasiforme* accessions. Only polymorphic sites having MAF lower than 5% are indicated (see Method). A. Decay of r² over genetic distance on chromosome 2. Plot of r² over distance was fitted by non linear regression (red curve). B. Decay of r² over physical distance on the 4 major contigs. Plot of r² over distance is fitted by non linear regression (red curve).The inset shows a more detailed view of the LD decay curve for markers located less than 20 Kb apart. C. Matrix of pairwise LD p-value between and within physical contigs. *p*-values were calculated with 1000 permutations

SNPs and Indels were more frequent in non-coding regions, with an average of 8.7 polymorphisms every 1000 bp than in exonic part of genes (average of 5.4 polymorphisms every 1000 bp). *S. l. cerasiforme* accessions were more polymorphic than both *S. pimpinellifolium* and *S. l. esculentum* but deviations in observed diversity could be biased by higher individual number for the former. *S. l. cerasiforme* shared polymorphisms with both cultivated and wild accessions having only 5 specific polymorphisms and 344 polymorphisms shared with the two other species (187 with *S. pimpinellifolium,* 11 with *S. lycopersicum* and 146 with both species). Fifty four percent of polymorphism identified in *S. l. esculentum* represented singletons within this group. Only two accessions (LA0409 and StupickePolniRane) are responsible of most of these singletons.

The ratio of polymorphism in non coding regions to coding regions is similar for *S. pimpinellifolium* and *S. l. cerasiforme* but is strikingly higher for *S. lycopersicum* (Table 1). *S. l. esculentum* also showed an excess of low frequency polymorphisms as *S. l. cerasiforme*, but to a lesser extent (Figure 2). *S. pimpinellifolium* had a well balanced distribution of allele frequencies.

Linkage disequilibrium: We compared different strategies for analysing LD decay over genetic distances. We examined pairwise LD values for all polymorphisms with MAF higher than 5% or for only one polymorphism by fragment with the highest heterozygosity index. We also compared pairwise LD decay between polymorphisms assessed in the whole population (N=90) or in the *cerasiforme* subset (N=63). Pairwise r^2 were plotted according to genetic distance between two loci and non linear regression fitted the decay of LD over genetic distance. LD decreased over lower genetic distances when all polymorphisms per sequence were taken into account and when only *cerasiforme* subset was analyzed (Figure S1). LD is probably over-estimated in the whole sample because of genetic structure with both cultivated and wild accessions added to the *cerasiforme* subset. Linkage disequilibrium decayed under r^2=0.3 within 1 cM and minimal value of r^2=0.09 was obtained for distances higher than 13 cM (Figure 3A). Extend of LD over 20 cM were obtained by VAN BERLOO *et al.* (2008) who examined mainly commercial tomato accessions. Nevertheless, high pairwise r^2 (r^2=1) still remained even within 60 cM, but only 28 sites (corresponding to 13 fragments) among 340 were responsible of these high pairwise LD values. These fragments are spread over the chromosome 2 (Figure 1) and extreme LD between them could be the consequence of selection over several loci.

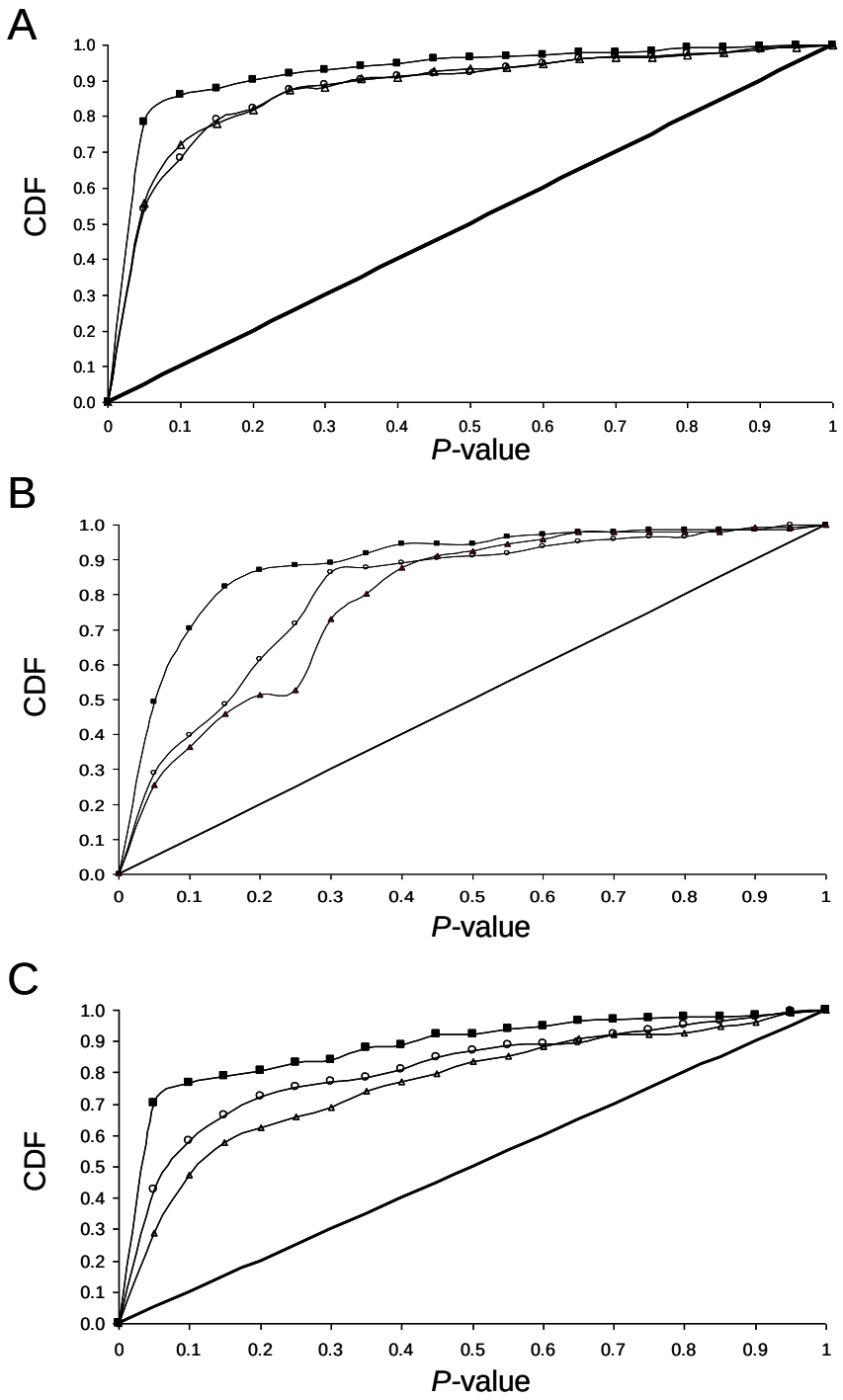


Figure 4. Cumulative density functions (CDF) using several alternative models of association. Model comparisons are performed for fruit weight (FW) (A), locule number (LCN) (B) and soluble solid content (SSC) (C). Associations are tested for all polymorphic sites with MAF>5% on 90 individuals. Naive (black square), Q (white circle) and Q+K (white triangle) models were tested. The diagonal indicates uniform distribution of *p*-values under the expectation that random SNPs are unlinked to the polymorphisms controlling these traits (H0: no SNP effect).

The extent of LD over physical distances was evaluated within the four physical contigs covering 1.86 Mb (Figure 3B). The minimal r^2 fitted value of 0.13 was obtained within 20 Kb but high pairwise LD persisted within 400 Kb. A matrix of *p*-values for LD between polymorphic sites of the four major contigs is shown on Figure 3C. Three different patterns of LD intensity over physical distances can be discriminated. First, in Contig1, polymorphisms within fragments formed blocks with high LD. LD between fragments of Contig1 and Contig2 was higher than LD between fragments of Contig1. Then, in Contig2 and the first part of the Contig3, fragments did not form haplotype blocks. High LD between and within fragments was interrupted by polymorphisms with low LD with other polymorphisms. Finally, a striking break in the LD pattern over physical distance appeared in the middle of Contig3 and subsequent Contig4, where intra-fragment blocks of LD and weak LD between fragments was observed. The genetic versus physical distance ratios in Contig3 was unevenly distributed with 136Kb/cM between TD140 and TD055 and 20 Kb/cM between TD109 and TD106. The difference in LD behavior between and within contigs is also clear on graphical haplotypes (Figure S2).

Association mapping: The genetic structure of 90 tomato accessions was estimated using 20 SSR markers spread over the genome. The most probable number of sub-populations in the sample was two with differentiation between cultivated *vs.* wild nature of accessions (Figure S3). Weak signal was also detected for subdivision in four populations. A larger sample of 318 accessions gave similar results with the same SSR markers (Ranc, Munos et al. 2008). Twenty six cherry tomato accessions were found in admixture between the two groups with maximal membership coefficient to one group lower than 0.8 (Table S3).

FW and LCN were log-transformed to fit graphically a normal distribution but LCN still fitted Poisson distribution. The three traits are correlated (Figure S4). Broad-sense heritabilities were high (0.96 for SSC, 0.99 for FW and 0.98 for LCN). This should imply good estimation of SNP-traits association with two year replicates. FW and SSC were correlated with the genetic structure, with R^2 values of 0.24 and 0.12, respectively, whereas population structure accounted only for 5% of the LCN variation. For association mapping, several models were compared for type I error rate control, as in Yu et al. (2006). The mixed model taking into account both genetic structure and coancestry matrices (K+Q model) resulted in the best approximation to the expected cumulative distribution of *p*-values, followed by the structured association model (Q model) and the simple model (t-test) for both SSC and LCN traits (Figure 4). For FW, no difference was detected between Q and K+Q models which both.

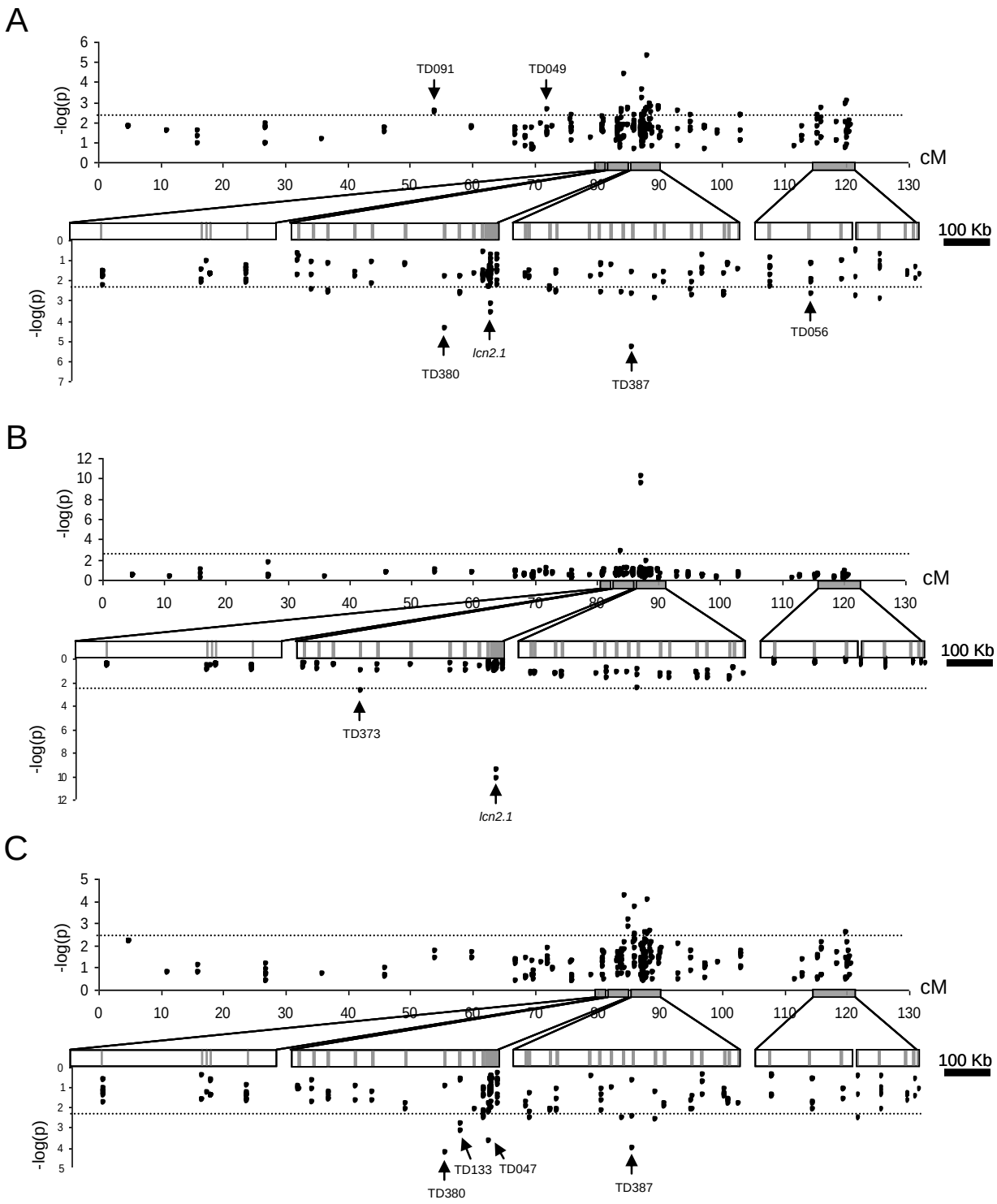


Figure 5. Plot of association p-values over the chromosome 2. Associations are estimated for 90 accessions. Q+K model was used to screen for association between polymorphisms and (A) fruit weight (FW), (B) locule number (LCN) and (C) soluble solid content (SSC). The upper part of each graph represents associations along genetic distance over the entire chromosome 2. The lower part shows associations for each physical contig. Arrows indicate the marker name of the most significant associations. Adjusted *p*-values for multiple testing (see Method) are shown.

performed better than the naive model. Significant associations (adjusted *p*-value lower than 0.005) are described in Table 2. For FW, LCN and SSC, polymorphisms with the highest association *p*-value explained a large part of the trait variation (22%, 44% and 21%, respectively) with R^2 calculated based on a linear model with all fixed effects. Allele effects were almost all positive for FW whereas allele effects for SSC were all negative.

Table 2. Significant associations for fruit weight (FW), locule number (LCN) and soluble solid content (SSC) estimated with Q+K model on 90 accessions. Only *p*-values lower than 0.005 are shown with indication on location and polymorphism effect.

trait	Locus	status	location[a]	*P*-value	corrected *P*-value[b]	R^2 [c]	a [d]	MAF[e]
log(FW)	TD091-415	non coding	54cM	0.0012	0.004	0.10	10.0	0.18
log(FW)	TD091-607	non coding	54cM	8.12E-04	0.003	0.10	9.2	0.24
log(FW)	TD049-528	non coding	72cM	6.04E-04	0.002	0.11	9.5	0.48
log(FW)	TD363-213	non coding	76cM	0.0019	0.005	0.07	9.6	0.39
log(FW)	TD383-419	non coding	84cM-c2.13	7.56E-04	0.003	0.12	12.1	0.11
log(FW)	TD383-558	non coding	84cM-c2.13	6.36E-04	0.002	0.13	11.3	0.13
log(FW)	TD383-60	non coding	84cM-c2.13	6.36E-04	0.002	0.13	11.3	0.13
log(FW)	TD375-573	coding	84cM-c2.14	0.0011	0.003	0.10	9.0	0.25
log(FW)	TD133-115	non coding	84cM-c2.8	3.34E-04	0.002	0.09	7.2	0.33
log(FW)	TD133-395	non coding	84cM-c2.8	5.57E-04	0.002	0.09	7.3	0.33
log(FW)	TD387-452	non coding	84cM-c2.9	9.40E-07	4.14E-05	0.19	11.6	0.27
log(FW)	lcn2.1-686	non coding	86cM-c2.3	2.86E-05	0.001	0.12	-11.7	0.38
log(FW)	lcn2.1-692	non coding	86cM-c2.3	8.95E-06	2.63E-04	0.15	-12.7	0.37
log(FW)	TD274-17	non coding	87.5cM-c3.13	9.32E-04	0.003	0.08	8.9	0.26
log(FW)	TD274-325	non coding	87.5cM-c3.13	4.76E-04	0.002	0.10	9.8	0.23
log(FW)	TD377-96	non coding	87.5cM-c3.14	0.0014	0.004	0.09	8.3	0.17
log(FW)	TD377-97	non coding	87.5cM-c3.14	0.0023	0.005	0.08	8.5	0.16
log(FW)	TD377-98	non coding	87.5cM-c3.14	0.0014	0.004	0.09	8.3	0.17
log(FW)	TD377-91	non coding	87.5cM-c3.14	0.0013	0.004	0.09	8.2	0.17
log(FW)	TD379-326	non coding	88cM-c3.11	4.42E-04	0.002	0.12	14.4	0.15
log(FW)	TD380-256	non coding	89cM-c3.8	3.04E-04	0.002	0.11	9.5	0.21
log(FW)	TD380-526	non coding	89cM-c3.8	6.13E-08	5.39E-06	0.22	13.2	0.36
log(FW)	TD280-328	non coding	89cM-c3.9	4.54E-04	0.002	0.10	10.5	0.48
log(FW)	TD055-469	coding	89.5cM-c3.7	9.46E-05	0.001	0.13	8.3	0.26
log(FW)	TD278-267	non coding	90cM-c3.3	1.73E-04	0.002	0.11	12.0	0.21
log(FW)	TD278-39	coding	90cM-c3.3	5.23E-04	0.002	0.10	15.0	0.15
log(FW)	TD278-444	non coding	90cM-c3.3	2.30E-04	0.002	0.12	12.4	0.22
log(FW)	TD278-524	non coding	90cM-c3.3	3.81E-04	0.002	0.12	11.9	0.20
log(FW)	TD300-257	non coding	90cM-c3.5	1.95E-04	0.002	0.12	11.6	0.20

trait	Locus	status	location[a]	*P*-value	corrected *P*-value[b]	R^2 [c]	a [d]	MAF[e]
log(FW)	TD300-41	non coding	90cM-c3.5	0.0011	0.003	0.11	9.2	0.33
log(FW)	TD108-347	coding	90.1cM	8.29E-04	0.003	0.10	7.4	0.27
log(FW)	TD056-134	non coding	116cM-c4.7	3.49E-04	0.002	0.12	10.8	0.35
log(FW)	TD369-493	coding	116cM-c4.8	0.0025	0.005	0.09	11.1	0.26
log(FW)	TD116-707	non coding	120cM-c4.3	4.90E-05	0.001	0.16	8.1	0.45
log(FW)	TD117-164	non coding	120cM-c4.4	1.16E-04	0.001	0.15	10.1	0.33
log(FW)	TD117-176	non coding	120cM-c4.4	1.16E-04	0.001	0.15	10.1	0.33
log(FW)	TD083-246	coding	133cM	0.0013	0.004	0.09	10.3	0.48
log(LCN)	TD373-391	non coding	86cM-c2.12	2.14E-05	0.002	0.21	-0.68	0.49
log(LCN)	lcn2.1-692	non coding	86cM-c2.3	5.93E-13	1.85E-10	0.44	-1.16	0.37
log(LCN)	lcn2.1-686	non coding	86cM-c2.3	5.32E-12	8.30E-10	0.44	-1.21	0.38
SSC	TD133-115	non coding	84cM-c2.8	1.87E-05	7.12E-04	0.16	-0.63	0.33
SSC	TD133-395	non coding	84cM-c2.8	4.90E-05	0.002	0.15	-0.58	0.33
SSC	TD387-452	non coding	84cM-c2.9	3.88E-07	5.89E-05	0.24	-0.86	0.27
SSC	TD047-274	non coding	86cM-c2.5	3.96E-06	2.01E-04	0.19	-1.00	0.12
SSC	TD120-212	non coding	86cM-c2.6	3.10E-04	0.004	0.13	-0.58	0.33
SSC	TD120-88	non coding	86cM-c2.6	2.22E-04	0.003	0.13	-0.59	0.32
SSC	TD140-180	non coding	87.5cM-c3.15	1.90E-04	0.003	0.14	-0.73	0.21
SSC	TD380-256	non coding	89cM-c3.8	2.57E-04	0.003	0.13	-0.65	0.21
SSC	TD380-526	non coding	89cM-c3.8	1.27E-06	9.68E-05	0.21	-0.70	0.36
SSC	TD280-328	non coding	89cM-c3.9	1.64E-04	0.003	0.14	-0.55	0.48
SSC	TD055-469	coding	89.5cM-c3.7	8.93E-05	0.002	0.15	-0.67	0.26
SSC	TD117-164	non coding	120cM-c4.4	1.52E-04	0.003	0.14	-0.70	0.33
SSC	TD117-176	non coding	120cM-c4.4	1.52E-04	0.003	0.14	-0.70	0.33

[a] Nomenclature for the location is as follows: "genetic distance on expen2000 reference map"-"the contig number"."the fragment number on this contig".
[b] *p*-values are corrected following the Benjamini & Hochberg (2000) procedure (see Material & Methods).
[c] R^2 were calculated using Q model.
[d] Allele effects are indicated in gram for FW, mean number of locule for LCN and °brix for SSC.
[f] Minimal allele frequencies (MAF) are shown for each polymorphism.

Most of the polymorphisms found in association with one of these traits were part of a physical contig (Figure 5). For FW, the two strongest associations involved TD380-526 (fragment TD380 polymorphic site at the position 526) on Contig3 and TD387-452 on Contig2 (adjusted *p*-value = 5.39E-06 and 4.14E-05, respectively). The r^2 value for LD estimation between these two SNPs is 0.41 in the whole accession sample (Figure S5). Because other equivalent pairwise LD did not implied significant association, the two associations could correspond to two linked QTLs on adjacent contigs. TD387-452 was no more associated with FW when only the 63 *S. l. cerasiforme* accessions were used for association analysis but TD380-526 still was. TD049 (Figure 5A) was also associated with FW and colocated with FW2.1, a QTL for FW variation fine mapped in a *S. l. esculentum* x *S. l. cerasiforme* cross (Lecomte, Saliba-Colombani et al. 2004). Significant associations with *lcn2.1*-686 and *lcn2.1*-692 were detected. These two SNPs are responsible of locule number variation between two-locule accessions and more than two-locule accessions (Muños, Ranc et al. *in prep.*). We also detected significant associations for FW with coding polymorphism in TD055 fragment which corresponds to *OVATE* gene and TD056 which corresponds to the 5' region of *fw2.2* gene.

For LCN, only three associations were found (Figure 5B). The highest associations engaged the two SNPs cited above as *lcn2.1* QTL. LD between these two SNPs was almost total with r^2 of 0.95. Resolution was not sufficient to discriminate the effect of each SNP. The other significant association implied TD373-391 which is also on the same contig. TD373-391 showed the highest r^2 with *lcn2.1* haplotype (r^2=0.47) and this association may result of LD (Figure S5). For SSC, 13 polymorphic sites showed association. Strongest associations were found with TD380-526 and TD387-452, but these results could be a consequence of the high negative correlation between FW and SSC (Spearman correlation coefficient of -0.66). The other highly significant association implied TD047-274, an indel of 20bp located in the same region as TD133-115, TD133-395, TD120-212 and TD120-88, also in association with SSC. LD between polymorphisms of TD047 and TD120 or TD133 is weak ($r^2 < 0.25$, Figure S5) whereas LD between TD133 and TD120 is high ($r^2 > 0.7$). TD380-256 and TD055-469 were also significantly associated with SSC and are contiguous on Contig3. TD380-256 suffered high LD with fragments on the Contig3 ($r^2 = 0.53$ with TD380-526 and $r^2 = 0.5$ with TD055-469) but also with a fragment on Contig2 ($r^2 = 0.45$ with TD133-395). TD140-180 associated with SSC also mapped on the Contig3 but was in significant LD with TD055 and TD380 (r^2=0.5), with TD047-274 (r^2=0.44) and TD133 (r^2=0.3). Thus, such association could be due to LD.

Discussion

Power of *S. l. cerasiforme* for polymorphism discovery: We sequenced 81 DNA fragments in 90 accessions of wild and cultivated tomato and detected 352 polymorphisms. The 63 *S. l. cerasiforme* capture 98% of SSR alleles from a largest sample of 144 cherry type accessions (Ranc, Munos et al. 2008). We showed that *S. l. cerasiforme* contained a large level of molecular variability, almost identical to its wild progenitor *S. pimpinellifolium*. In tomato, several studies were interested in discovering SNPs and Indels. Nesbitt and Tanksley (2002) searched for molecular polymorphisms, in the *fw2.2* QTL region, explicative for tomato fruit weight variation within a germplasm of *S. l. esculentum* (N=4) *and S. l. cerasiforme* (N=39) accessions. Authors found only one SNP for 7Kb of sequences among *S. l. esculentum* accessions and 8 polymorphisms for 2.7Kb within the *S. l. cerasiforme* sample. Because of sequencing costs and low polymorphism available in tomato, a strategy consisting in mining polymorphic EST *in silico* and testing a part of them by genotyping was also developed (Yang, Bai et al. 2004). Authors detected the same amount of polymorphism with one SNP every 8,500 bp of coding sequence analyzed. More recently, Jimenez-Gomez and Maloof (2009) used this method to detect polymorphism within and between cultivated and wild species of tomato. They succeeded in finding more than 15,000 intraspecific polymorphisms in a set of 223,000 EST sequences for *S. lycopersicum*. cDNA library for cultivated tomato were obtained from only few individuals and most of the polymorphisms detected could have MAF lower than 5%. Another study resequencing polymorphic ESTs reported a higher amount of polymorphism but the individuals used for sequencing were described as highly variable compared to other *S. lycopersicum* accessions because of introgressions from wild relatives (Labate and Baldo 2005). Van Deynze, Stoffel et al. (2007) discovered a higher frequency of SNPs and indels than the previous studies, but authors focused on gene intron and found only 302 polymorphic fragments over 967 sequenced. Among 1,487 SNPs detected by Labate *et al.* (2009), only 162 were polymorphic in *S. lycopersicum* breeding germplasm and most of them had minor allele frequencies below 10% . *S. lycopersicum* presents the highest degree of morphological diversity but seems not to be suitable for detecting a large panel of polymorphisms. *S. l. cerasiforme* allowed us to screen a large set of polymorphisms (352 SNP and indels) by sequencing only 81 DNA fragments. Four of the 11 monomorphic fragments contained only coding region (TD085, TD098, TD111 and TD384), known to be poorer in polymorphism than non coding region, but the others were mainly constituted by non coding region. The difference in polymorphism content between species for non-coding region is consequence of (i) hitch-hiking of region surrounding a selected polymorphism or

(ii) demographic bottleneck during domestication with reduction of population effective size. *S. pimpinellifolium* undergone bottleneck only recently with a drastic reduction of its natural habitats. Accessions present in our collection preserved the genetic diversity of the ancestor prospected in the middle of the twentieth century. *S. l. cerasiforme* suffered a decrease of its population effective size during domestication from *S. pimpinellifolium* (Bai and Lindhout 2007). No differences between polymorphism amounts of *S. pimpinellifolium* and *S. l. cerasiforme* are highlighted in this study but could be due in part to a higher number of accessions sequenced for the latter. During domestication of tomato, human selection focused mainly in increasing fruit size of *S. pimpinellifolium* which was yet edible and quite tasty. Tomato domestication process may have implied several loci evolving from *S. pimpinellifolium* small-fruits alleles to *S. l. cerasiforme* medium-fruit alleles. Neither iconography nor archaeological evidence on the appearance of large fruit tomato is available. First description of tomato accessions in Mexico depicted globular fruit and tomato is first mentioned in European literature as "flattened" and "segmented" (Daunay, Laterrot et al. 2007). The only insurance we thus have is that few accessions were taken from Central or South America to Europe in the sixteenth century and the progeny of accessions were then the starting material of selection for tomato heirloom and parents of modern varieties. Molecular polymorphism is linked to the population size by the Watterson's estimate of the scaled mutation rate (per site) $\theta = 4Ne\mu$ where *Ne* is the population effective size and μ the mutation rate. Because, this second bottleneck highly reduced the effective population size of tomato, it also decreased the amount of molecular diversity found in *S. lycopersicum*. Selection pressure which targeted coding regions could explain the higher ratio between non-coding and coding polymorphisms for *S. lycopersicum* than for *S. pimpinellifolium* or *S. l. cerasiforme*. The reduction could arise on the fragment targeted by selection but also on region suffering genetic hitchhiking or background selection Innan and Stephan 2003). Thus, less drastic reduction in population size and continuous inter-mating with *S. pimpinellifolium* shaped a higher level of molecular variability for *S. l. cerasiforme*.

LD decay over genetic and physical distances: Ancient admixture increased polymorphism level of cherry type and limited the overall LD. We estimated minimal LD values ($r^2<0.09$) with distances higher than 13 cM, but extreme LD values until 60 cM. Our results are close to van Berloo *et al.* results (2008) who described LD extent to 15 to 20 cM using AFLP markers in a cherry sample (N=18). Nesbitt and Tanksley (2002) showed that LD in *S. l. cerasiforme* could be broken within 150 Kb comparing three loci linked around *fw2.2*. With an estimated

ratio of 750 kb/cM on the whole tomato genome (Tanksley, Ganal et al. 1992), results of LD decay over physical and genetic distances are not consistent. In our *S. l. cerasiforme* sample, some r^2 values were still extreme over hundreds Kb but the drop estimated by non linear regression indicated that minimal LD is reached from 20 Kb. *Arabidopsis thaliana* showed large extent of LD over the FRI region where LD extended to 200 Kb corresponding to one cM (Nordborg, Borevitz et al. 2002). This estimate is locus specific and when studies are performed on whole genome loci, LD decayed within 10 Kb in average (Kim, Plagnol et al. 2007). Nordborg (2000) estimated from simulations that LD should vanish over a scale of 10 Kb for inbred species. This is verified in our sample. Results of LD decay over genetic distances on tomato are similar to LD pattern assessed in barley, another highly inbreeding crop (Zhang, Marchand et al. 2009). In barley large differences are observed for LD decay pattern among cultivated, landraces and wild accessions (Caldwell, Russell et al. 2006). The higher LD extent for crop compared to wild ancestor or to *A. thaliana* could be due to important bottleneck undergone during domestication. The LD pattern observed on the physical contigs is relevant with haplotypic blocks described on soybean landraces, another inbred crop (Hyten, Choi et al. 2007) but also on *A. thaliana* (Kim, Plagnol et al. 2007) and human (Daly, Rioux et al. 2001). In tomato, as in soybean, all the haplotypic blocks are interleaved by polymorphisms that are in complete linkage equilibrium with other SNPs. For tomato, this LD pattern should not be due to recombination because, for two adjacent polymorphisms with weak LD, the four haplotypes are rarely met. This pattern could thus be due to recent mutation having low frequencies (more than 50% of polymorphism had MAF<0.2). Frequencies of these recent polymorphisms evolved by lineage effect rather than by recombination. For mapping purposes, it is the recombination history that is important. The mutations are of interest only to the extent that they reveal something about this unobservable history (Nordborg and Tavare 2002). The break in LD pattern described in the contig3 could be due to presence of hotspot of recombination. Mapping data offered direct confirmation of uneven distribution of recombination over Contig3 but the high density of polymorphism detected in this study should be mapped on a large F2 population to confirm the presence of a recombination hotspot (Drouaud 2006). The LD pattern described in the first part of contig3 and contig2 could have been shaped by selection. Clusters of QTL have been mapped in this region, with, for example locule number QTL, fruit shape QTL, fruit weight QTL, soluble solids content QTL and sugar QTL (Lecomte, Saliba-Colombani et al. 2004). Selection of new advantageous mutation during domestication should have increased LD (Nordborg and Tavare 2002) as in *A. thaliana*, where LD blocks surrounding selected

polymorphism were longer than blocks surrounding non-selected alleles (Kim, Plagnol et al. 2007).

Association mapping allows validation of candidate genes: Our core collection was efficient to detect association in candidate gene regions. *lcn2.1* was previously identified by map-based cloning approach as a QTN responsible for variation in tomato locule number (Muños, Ranc et al. *in prep.*). We used information on *lcn2.1* to highlight possible effect of these two SNPs on FW and SSC. The significant association found between these two SNPs and FW was the only one with negative allelic effect. The reference genotype Heinz 1706 has large fruit with only two locules. Almost all other two-locule genotypes had small fruit and the large number of these small-fruit accessions in the reference group induced negative effect for FW. Nesbitt and Tanksley (2002) could not detect any association between the *fw2.2* region previously cloned and FW in a *S. l. cerasiforme* sample. Authors concluded that other genes than *fw2.2* were responsible of variation in FW for cherry tomatoes. The number of accessions (39 *S. l. cerasiforme*, 4 *S. l. esculentum* and 3 *S. pimpinellifolium*) was the principal limitation of this study. Using 90 accessions selected to represent diversity of a larger collection, we found association with a polymorphic site located in the promoter of the gene. This polymorphism could be responsible for the phenotype variation or could be in LD with the responsible one. The entire cloned region should be sequenced and tested for association before concluding.

Association mapping for the discovery of new candidate genes: The screening of polymorphism on chromosome 2 with high density markers allowed the detection of many new associations and identification of some putative new candidate genes. All QTL fine-mapped in the mapping population *S. l. cerasiforme* x S. *l. esculentum* for FW, LCN and SSC (Lecomte, Saliba-Colombani et al. 2004) were also identified by association mapping. Association with LCN was found with TD373-391 tagged on the contig2. This polymorphism showed linkage disequilibrium with the two SNPs of *lcn2.1*. Thirteen accessions having a different haplotype between TD373 and *lcn2.1* polymorphisms reduced LD and thus the significance of association with TD373 was lower. The number of significant associations found with FW can ensue from whole chromosome LD caused by strong selection on this phenotype (Bai and Lindhout 2007). TD380-526 showed the most significant association with FW but no tomato unigene mapped within a 25Kb delimited region surrounding this fragment. However, EuGene'Hom software (Foissac, Bardou et al. 2003) predicted a gene

homologous to *Populus trichocarpa*'s gene coding for chromatin remodeling complex subunit (GENE ID: 7458812 CHR905) and *Arabidopsis thaliana*'s gene coding for SWI2/SNF2-like protein (GENE ID: 836808 CHR1). The associated polymorphism tagged in the first intron of the gene. Chromatin remodeling proteins are actors of reconfiguration of protein–DNA interactions that accompanies or induces changes in genomic activity like gene expression (Kaya, Shibahara et al. 2001; Verbsky and Richards 2001). The other highly significant associated fragment, TD387, had homology with a *S. lycopersicum* unigene (SGN-U596069) but had no homology with annotated sequence. Another association was detected for FW with TD049-528. TD049 tagged in the 3' region of a gene coding for glyoxalase I. This gene co-localise with a QTL for FW variation in a mapping population derived from a *S. l. esculentum* x *S. l. cerasiforme* cross (Saliba-Colombani, Causse et al. 2001). The corresponding protein showed differences in its amino acid sequence between the mapping population parents (data not shown). Because of putative impact of glyoxalase I protein on plant cell proliferation (Paulus, Köllner et al. 1993), this gene represents a good candidate gene for FW variation. The two most significantly associated polymorphisms with FW were also associated with SSC. This could be simply explained by dilution effect of soluble sugars and acids according to the fruit size. The two polymorphisms were no more statistically associated to SSC when we added FW effect as covariate in the K+Q-model (data not shown). We observed the same result for TD117, genetically close to *fw2.2*. The two other best associations TD047 and TD133 are both located on the same physical as TD120. Because TD047 and TD133 are separated by 2 cM, this region should be enriched in SNPs to locate precisely one or more responsible polymorphisms. TD055 was designed in *Ovate* gene and showed association with SSC. *Ovate* is implied in modification of fruit shape and no effect on SSC was previously reported. TD055 mapped in a SSC QTL (brix2.2) described in the mapping population involving cherry tomato (Saliba-Colombani, Causse et al. 2001; Lecomte, Saliba-Colombani et al. 2004). This polymorphism could be in LD with the responsible polymorphism. SSC also showed significant association with TD140 defined on a gene coding for an aldose-1-epimerase (EC 5.1.3.3). This enzyme catalyzes the transformation of alpha-D-glucose into beta-D-glucose and participates in glycolysis and gluconeogenesis. The aldose-1-epimerase thus represents a new candidate for SSC variation.

Optimal conditions for whole genome association studies (WGA): We highlighted the higher efficiency of K+Q-model in dealing with type I error rate for association mapping in tomato. Information on estimated familial relatedness in our sample did not influence results

for association with FW because most of the false positives are corrected only with genetic structure information. K+Q-model will prove all its power in sample of increased size as well as broader allelic diversity (Yu, Pressoir et al. 2006). The core collection could be efficient to detect gene with large effect on trait variation but will suffer a decrease of statistical power when dealing with low effect variants. Larger collection will be useful to map genes with low effect because greater power is achieved by increasing the sample size than by increasing the number of polymorphisms (LONG and LANGLEY 1999). Density of markers needed for association analysis is estimated by LD decay over genetic or physical distance (Rafalski 2002). An r^2 value of 0.3 indicates sufficiently strong LD to be useful for association mapping in human studies (Ardlie, Kruglyak et al. 2002). In *S. l. cerasiforme* accessions, LD estimated values decayed below $r^2=0.3$ within 1 cM. One SNP every cM could be highly valuable for medium resolution WGA. For physical LD mapping, a higher density is needed because, even if extreme LD is still found over hundreds Kb, estimate of LD decay indicates that LD is minimal after 20 Kb. With a genome length of 950 Mb for tomato, a set of 48 000 markers will be necessary to have high resolution for WGA. To validate these estimations based on LD, we looked at the number of significant associations for different mapping strategies. The number of fragments found in association with traits increased with densification of polymorphisms on genetic map and spread over distance. Significant associations (*p*-value <0.005) were found with large mapping strategy (1 marker every 5cM) for FW, but not for LCN and SSC. With this density of marker, the only way to find association with LCN and SSC was to use the less stringent Q-model. The K+Q-model may have eliminated true positive association in LD with the responsible polymorphism. The density of markers needed, will thus depend on the trait, the locus targeted and the population studied. For example, it would not have been possible to physically map *lcn2.1* QTN using only LD because the two responsible SNPs are in complete equilibrium with surrounding polymorphisms.

Our results suggest that genomic admixture of *S. l. cerasiforme* inflated the level of molecular diversity of domesticated tomato. Our core collection was efficient to detect associations in candidate regions where QTL have been previously mapped. We highlighted the higher efficiency of K+Q-model in dealing with type I error rate even in a relatively small sample. The screening of polymorphisms along chromosome 2 with high marker density allowed the detection of many new associations and the identification of some putative new candidate genes that are targets for further validation in larger collections. If results on chromosome 2

are extrapolated to the whole genome, around 50,000 SNPs will be sufficient for high-resolution mapping in such collection. With next generation sequencing facilities, a large amount of polymorphisms will be soon available for this purpose.

Literature cited

Ardlie, K. G., L. Kruglyak, et al. (2002). "Patterns of linkage disequilibrium in the human genome." **3**(4): 299-309.

Bai, Y. and P. Lindhout (2007). "Domestication and Breeding of Tomatoes: What have We Gained and What Can We Gain in the Future?" Ann Bot: mcm150.

Benjamini, Y. and Y. Hochberg (2000). "On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics." J. Edu. Behav. Stat. **25**(1): 60-83.

Bradbury, P. J., Z. Zhang, et al. (2007). "TASSEL: software for association mapping of complex traits in diverse samples." Bioinformatics **23**(19): 2633-2635.

Bres, C., J. P. Bouchet, et al. (2005). CGIS: an information system used for designing primers of candidate genes using Arabidopsis whole genome and Solanaceae EST databases. 16. Triennal Conference of the EAPR. Bilbao (ESP).

Buckler, I., Edward S. and J. M. Thornsberry (2002). "Plant molecular diversity and applications to genomics." Curr. Opin. Plant Bio. **5**(2): 107-111.

Caldwell, K. S., J. Russell, et al. (2006). "Extreme Population-Dependent Linkage Disequilibrium Detected in an Inbreeding Plant Species, Hordeum vulgare." Genetics **172**(1): 557-567.

Camus-Kulandaivelu, L., J.-B. Veyrieras, et al. (2006). "Maize Adaptation to Temperate Climate: Relationship Between Population Structure and Polymorphism in the Dwarf8 Gene." Genetics **172**(4): 2449-2463.

Causse, M., V. Saliba-Colombani, et al. (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**(377): 2089-2098.

Daly, M. J., J. D. Rioux, et al. (2001). "High-resolution haplotype structure in the human genome." Nat. Genet. **29**(2): 229-232.

Darvasi, A. and S. Shifman (2005). "The beauty of admixture." Nat. Genet. **37**(2): 118-119.

Daunay, M.-C., H. Laterrot, et al. (2007). Iconography and History of Solanaceae: Antiquity to the 17th Century. Hort. Rev. J. Janick. **34:** 1-119.

Doebley, J., A. Stec, et al. (1997). "The evolution of apical dominance in maize." Nature **386**(6624): 485-488.

Drouaud, J. (2006). "Variation in crossing-over rates across chromosome4 of *Arabidopsis thaliana* reveals the presence of meiotic recombination hot spots." Genome Res. **16**: 106-114.
Evanno, G., S. Regnaut, et al. (2005). "Detecting the number of clusters of individuals using the software structure: a simulation study." Mol. Ecol. **14**(8): 2611-2620.

Flint-Garcia, S. A., J. M. Thornsberry, et al. (2003). "Structure of linkage disequilibrium in plants." Annu. Rev. Plant Biol. **54**(1): 357-374.

Flint-Garcia, S. A., A. Thuillet, et al. (2005). "Maize association population: a high-resolution platform for quantitative trait locus dissection." The Plant Journal **44**(6): 1054-1064.

Foissac, S., P. Bardou, et al. (2003). "EUGENE'HOM: a generic similarity-based gene finder using multiple homologous sequences." Nucl. Acids Res. **31**(13): 3742-3745.

Frary, A., T. C. Nesbitt, et al. (2000). "*fw2.2*: A Quantitative Trait Locus Key to the Evolution of Tomato Fruit Size." Science **289**(5476): 85-88.

Gupta, P. K., S. Rustgi, et al. (2005). "Linkage disequilibrium and association studies in higher plants: Present status and future prospects." Plant Mol. Biol. **57**(4): 461-485.

Hardy, O. J. and X. Vekemans (2002). "SPAGeDI: a versatile computer program to analyse spatial genetic structure at the individual or population levels." Molecular Ecology Notes **2**: 618-620.

Hyten, D. L., I.-Y. Choi, et al. (2007). "Highly Variable Patterns of Linkage Disequilibrium in Multiple Soybean Populations." Genetics **175**(4): 1937-1944.

Innan, H. and W. Stephan (2003). "Distinguishing the Hitchhiking and Background Selection Models." Genetics **165**(4): 2307-2312.

Jimenez-Gomez, J. and J. Maloof (2009). "Sequence diversity in three tomato species: SNPs, markers, and molecular evolution." BMC Plant Biology **9**(1): 85.

Kaya, H., K.-i. Shibahara, et al. (2001). "*FASCIATA* Genes for Chromatin Assembly Factor-1 in *Arabidopsis* Maintain the Cellular Organization of Apical Meristems." Cell **104**(1): 131-142.

Kim, S., V. Plagnol, et al. (2007). "Recombination and linkage disequilibrium in *Arabidopsis thaliana*." Nat Genet **39**(9): 1151-1155.

Labate, J., L. Robertson, et al. (2009). "EST, COSII, and arbitrary gene markers give similar estimates of nucleotide diversity in cultivated tomato (*Solanum lycopersicum* L.)." Theo. App. Genet. **118**(5): 1005-1014.

Labate, J. A. and A. Baldo (2005). "Tomato SNP Discovery by EST Mining and Resequencing." Mol. Breed. **16**(4): 343-349.

Lecomte, L., V. Saliba-Colombani, et al. (2004). "Fine mapping of QTLs of chromosome 2 affecting the fruit architecture and composition of tomato." Mol. Breed. **V13**(1): 1-14.

Liu, J., J. Van Eck, et al. (2002). "A new class of regulatory genes underlying the cause of pear-shaped tomato fruit." Proc. Nat. Acad. Sci. USA **99**(20): 13302-13306.
Long, A. D. and C. H. Langley (1999). "The Power of Association Studies to Detect the Contribution of Candidate Genetic Loci to Variation in Complex Traits." Genome Res. **9**(8): 720-731.

Manning, K., M. Tor, et al. (2006). "A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening." Nat. Genet. **38**(8): 948-952.

Muños, S., N. Ranc, et al. (*in prep.*). "Increases in tomato fruit size and locule number is controlled by two key SNP located near Wuschel."

Nesbitt, T. C. and S. D. Tanksley (2002). "Comparative Sequencing in the Genus *Lycopersicon*: Implications for the Evolution of Fruit Size in the Domestication of Cultivated Tomatoes." Genetics **162**(1): 365-379.

Nordborg, M. (2000). "Linkage Disequilibrium, Gene Trees and Selfing: An Ancestral Recombination Graph With Partial Self-Fertilization." Genetics **154**(2): 923-929.

Nordborg, M., J. O. Borevitz, et al. (2002). "The extent of linkage disequilibrium in Arabidopsis thaliana." Nat. Genet. **30**(2): 190-193.

Nordborg, M. and S. Tavare (2002). "Linkage disequilibrium: what history has to tell us." Trends Genet. **18**: 83-90.

Paulus, C., B. Köllner, et al. (1993). "Physiological and biochemical characterization of glyoxalase I, a general marker for cell proliferation, from a soybean cell suspension." Planta **189**(4): 561-566.

Price, A. L., N. J. Patterson, et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nat. Genet. **38**(8): 904-909.

Pritchard, J. K., M. Stephens, et al. (2000). "Inference of Population Structure Using Multilocus Genotype Data." Genetics **155**(2): 945-959.

Pritchard, J. K., M. Stephens, et al. (2000). "Association Mapping in Structured Populations." Am. J. Hum. Genet. **67**: 170-181.

R Development Core Team (2005). R: A language and environment for statistical computing, reference index version 2.2.1. F. f. S. Computing. Vienna, Austria.

Rafalski, A. (2002). "Applications of single nucleotide polymorphisms in crop genetics." Curr. Opin. Plant Biol. **5**: 94-100.

Ranc, N., S. Munos, et al. (2008). "A clarified position for *solanum lycopersicum* var. *cerasiforme* in the evolutionary history of tomatoes (solanaceae)." BMC Plant Biol. **8**(1): 130.

Ritland, K. (1996). "Estimators for pairwise relatedness and individual inbreeding coefficients." Genetics Research **67**(02): 175-185.

Saliba-Colombani, V., M. Causse, et al. (2001). "Genetic analysis of organoleptic quality in fresh market tomato. 1. Mapping QTLs for physical and chemical traits." Theo. App. Genet. **V102**(2): 259-272.

Takahashi, M., F. Matsuda, et al. (2003). "Automated identification of single nucleotide polymorphisms from sequencing data." J. Bioinform. Comput. Biol. **1**: 253-265.

Tanksley, S. D., M. W. Ganal, et al. (1992). "High Density Molecular Linkage Maps of the Tomato and Potato Genomes." Genetics **132**(4): 1141-1160.

Tanksley, S. D., S. Grandillo, et al. (1996). "Advanced backcross QTL analysis in a cross between an elite processing line of tomato and its wild relative *L. pimpinellifolium*." Theo. App. Genet. **92**(2): 213-224.

Tenesa, A., A. F. Wright, et al. (2004). "Extent of linkage disequilibrium in a Sardinian sub-isolate: sampling and methodological considerations." Hum. Mol. Genet. **13**(1): 25-33.

Thornsberry, J. M., M. M. Goodman, et al. (2001). "Dwarf8 polymorphisms associate with variation in flowering time." Nat. Genet. **28**: 286 - 289.

van Berloo, R., A. Zhu, et al. (2008). "Diversity and linkage disequilibrium analysis within a selected set of cultivated tomatoes." Theo. App. Genet. **117**(1): 89-101.

van Deynze, A., K. Stoffel, et al. (2007). "Diversity in conserved genes in tomato." BMC Genomics **8**(1): 465.

Verbsky, M. L. and E. J. Richards (2001). "Chromatin remodeling in plants." Curr. Opin. Plant Biol. **4**(6): 494-500.

Weber, A. L., W. H. Briggs, et al. (2008). "The Genetic Architecture of Complex Traits in Teosinte (*Zea mays ssp. parviglumis*): New Evidence From Association Mapping." Genetics **180**(2): 1221-1232.

Yang, W., X. Bai, et al. (2004). "Discovery of single nucleotide polymorphisms in Lycopersicon esculentum by computer aided analysis of expressed sequence tags." Mol. Breed. **V14**(1): 21-34.

Yu, J., G. Pressoir, et al. (2006). "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness." Nat. Genet. **38**(2): 203-208.

Zhang, L., S. Marchand, et al. (2009). "Population structure and linkage disequilibrium in barley assessed by DArT markers." Theo. App. Genet. **119**(1): 43-52.

Zhao, K., M. J. Aranzana, et al. (2007). "An *Arabidopsis* Example of Association Mapping in Structured Samples." PLoS Genet **3**(1): e4.

Les données supplémentaires sont présentées dans les annexes 6 à 13.

5.3. Complément d'analyse

En résumé, nous avons vu que la position intermédiaire des accessions *S. l. cerasiforme*, permet de détecter rapidement du polymorphisme intéressant pour réaliser des études d'associations. Nous avons défini le meilleur modèle pour réaliser les tests d'association sur la « core collection » de 90 accessions et il est fort probable que ce modèle soit aussi le plus efficace sur des échantillons beaucoup plus grands. Les analyses d'association, focalisées sur le chromosome 2, permettent de retrouver les QTL identifiés dans la population issue du croisement entre Cervil, accession de type cerise, et Levovil, accession de type élite. L'utilisation de la diversité présente dans les ressources génétiques de tomate, a aussi permis d'identifier de nouvelles régions d'intérêt. Les analyses d'association permettent de donner des arguments supplémentaires en faveur de gènes candidats mais aussi de détecter de nouveaux polymorphismes qu'il sera intéressant de valider par la suite.

5.3.1. Validation des associations sur un plus grand échantillon

Nous avons complété les analyses réalisées dans l'article précédent. Tout d'abord, nous avons cherché à valider les associations les plus significatives identifiées sur 90 accessions.

Pour cela nous avons re-séquencé les TD047, TD380 et TD387 qui donnaient les associations les plus significatives avec le poids du fruit (FW) et la teneur en solides solubles (SSC). Un marqueur CAPS (Cleaved Amplified Polymorphic Sequence) a pu être dessiné sur le SNP du TD049 (glyoxalate) qui montrait une association avec le poids du fruit. Le génotypage de ces polymorphismes a été réalisé sur la collection totale mais nous nous sommes intéressés uniquement aux 201 lignées phénotypées durant les deux années. La structure de cet échantillon a été calculée en utilisant le logiciel STRUCTURE avec les mêmes paramètres que dans le chapitre 3. Deux sous-populations (domestiquée *vs.* sauvage) représentent, ici aussi, la structure la plus vraisemblable. La matrice d'apparentement (kinship) a été calculée de la même manière que dans l'article. L'association a été réalisée en utilisant le modèle Q+K. Aucune correction n'a été apportée étant donné que les trois polymorphismes ont été testés de façon indépendante.

Les SNP TD380-526 et TD387-452 restent fortement associés au poids du fruit. La *p*-value du marqueur TD380-526 diminue de 5,39e-6 pour 90 accessions à 1,04e-7 pour 201 accessions. La *p*-value du marqueur TD387-452 augmente de 9,4e-7 à 13,87e-5. Le SNP TD047-274 qui

est associé à la teneur en solides solubles montre un gain conséquent dans la significativité de l'association (*p*-value de 3,96e-6 sur 90 accessions et de 5,31e-12 sur 201 accessions). La figure 5-6 représente la distribution des différentes valeurs des phénotypes moyens de chaque accession en fonction des allèles, à chaque site polymorphe associé. Le polymorphisme TD049-528 montre une forte association avec le poids du fruit lorsqu'on teste les 201 accessions avec une *p*-value = 9,98e-5 (*p*-value = 6,04e-4 sur 90 accessions). Le fait de retrouver ces associations sur un panel d'accessions beaucoup plus grand valide l'échantillonnage de la « core collection ». Celle-ci représente bien la diversité observée dans la collection initiale. On peut donc espérer qu'une majorité des associations identifiées sur la « core collection » seront retrouvées en testant la collection totale. Il sera important par la suite de se concentrer sur ces régions en augmentant encore la densité des marqueurs.

5.3.2. Analyse d'association sur un échantillon composé de tomate de type cerise uniquement (N=63).

Nous avons voulu savoir si on pouvait retrouver ces associations en travaillant sur un échantillon contenant uniquement 63 accessions de *S. l. cerasiforme*. Le tableau 5-3 montre les associations significatives identifiées avec ce sous-échantillon. Etant donné le nombre d'individus utilisés, nous avons choisi le seuil $\alpha < 0.05$ plus conservatif.

trait	Locus	*P*-value	corrected *p*-value[a]
log(FW)	TD380-526	0.0012	0.0159
log(FW)	TD056-134	8.12E-04	0.0366
log(FW)	TD116-707	6.04E-04	0.0366
log(FW)	TD117-219	0.0019	0.0366
log(FW)	TD138-61	7.56E-04	0.0435
log(LCN	lcn2.1-692	9.31E-10	3.43E-07
log(LCN	lcn2.1-686	8.77E-09	1.61E-06

[a] Les *p*-value sont corrigées en suivant la procédure de Benjamini & Hochberg (2000).

Tableau 5-3. Associations significatives pour le poids du fruit (FW) et le nombre de loges (LCN), estimées avec le modèle Q+K sur 63 accessions de *S. l. cerasiforme*. Seules les associations présentant une *p*-value corrigée inférieure à 0,05 sont présentées.

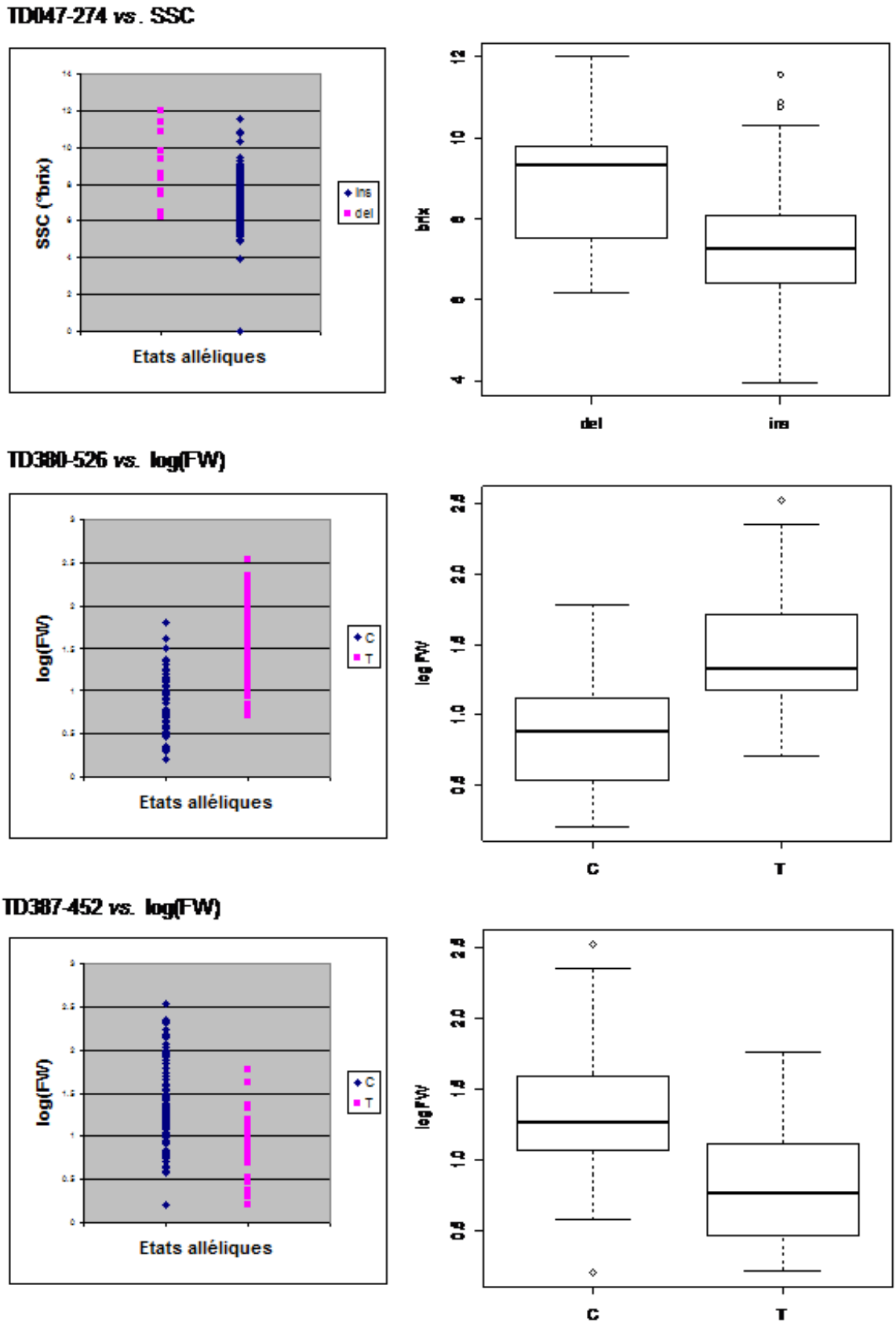


Figure 5-6. Effet des différents SNP sur le phénotype associé. Les SNP ont été re-séquencés sur 201 accessions. Les données manquantes ne sont pas indiquées. Les graphes de gauche représentent les nuages de point pour chaque allèle et les graphes de droites représentent les boîtes de dispersions de Tukey des phénotypes en fonction de l'état allélique.

Plus aucun polymorphisme n'est associé avec la teneur en solides solubles. On remarque aussi que le fragment TD387 n'est plus associé avec le poids du fruit. Le marqueur TD056, dessiné dans le promoteur de *fw2.2*, apparaît comme le deuxième locus significativement associé avec le poids du fruit mais le polymorphisme impliqué n'est pas le même que celui identifié sur les 90 accessions. Les polymorphismes associés avec le caractère SSC, dans les tests avec la « core collection » totale, ont des effets plus forts entre les différents groupes cultivés, sauvage et intermédiaire par rapport aux effets intra-groupes, identifié sur les accessions *S. l. cerasiforme* uniquement. Par exemple, les individus présentant la délétion au niveau du fragment TD047, ayant un effet positif sur SSC, sont principalement des individus issus de l'espèce sauvage. Il en est de même pour les accessions portant l'allèle T au marqueur TD387-452. En éliminant les accessions sauvages, on gomme les différences entre individus ce qui réduit énormément la puissance d'association. La diminution de l'effectif peut aussi expliquer la diminution de la significativité des marqueurs.

Il est aussi important de noter que la prise en compte de la structure et des apparentements comme variables dans les modèles utilisés peuvent augmenter le taux à des faux négatifs, notamment par une correction trop importante des effets de structure génétique et d'apparentements. Ceci peut être observé dans le cas d'un caractère phénotypique dont la variation serait liée à la structure des populations, comme c'est par exemple le cas pour la précocité de floraison chez le maïs (Flint-Garcia, Thuillet et al. 2005; Camus-Kulandaivelu, Veyrieras et al. 2006).

5.3.3. Analyse d'association sur d'autres caractères liés à la qualité du fruit

Nous avons aussi complété ces analyses avec d'autres caractères mesurés : l'acidité titrable (TA), les paramètres de couleurs (L*, a*, b*) et la fermeté (FIR).

L'acidité titrable montre un effet cultivar et un effet année fortement significatifs (*p*-value < 2,2e-16) mais présente une héritabilité élevée (h^2 = 0,94). Le caractère est significativement corrélé à la teneur en solides solubles (r = 0,51), au poids du fruit (r =-0,69) et au pH (r = -0,73).

Seuls deux fragments sont associés à l'acidité titrable avec des *p*-values inférieures au seuil $\alpha < 0.005$. Il s'agit du fragment TD278 (avec quatre polymorphismes associés) et du fragment TD086 (avec deux polymorphismes associés). Les quatre polymorphismes du fragment TD278 montrent aussi une association avec le poids du fruit ce qui n'est pas le cas de TD086. De plus, ce dernier co-localise avec un QTL détecté (*TA2*.2) sur la population de

cartographie Cervil x Levovil. Ce fragment est dessiné dans un gène qui code pour la sous-unité béta d'une peptidase mitochondriale. Le marqueur TD278 a été dessiné dans un gène codant pour une protéine de la famille des flavodoxines ayant un rôle dans le transport des électrons et une activité d'oxydo-reductase. Ces deux gènes ont un rôle potentiel dans l'activité respiratoire de la mitochondrie. Les acides malique et citrique interviennent dans le cycle de Krebs qui a lieu dans la matrice mitochondriale.

Les composantes de la couleur L*, a* et b* montrent des effets cultivar fortement significatifs (*p*-value < 2,2e-16). Seuls a* et b* montrent des effets années significatifs (*p*-value < 2,2e-16 et *p*-value = 2,2e-12 respectivement pour a et b). Les héritabilités de ces caractères sont toujours élevées : $h^2 = 0,95$ pour a*, $h^2 = 0,89$ pour L* et $h^2 = 0.92$ pour b*. Le caractère a* est significativement corrélé à la teneur en solides solubles (r = 0,51) au poids du fruit (r =-0,69) et au pH (r = -0,73). L* est significativement corrélé à b* (r = 0,57) c'est-à-dire que le contraste est fortement corrélé à la teinte jaune des fruits, plus claire que les teintes rouges. L* est aussi corrélé au poids du fruit (r = 0,50) ce qui peut être expliqué par le fait que la quasi-totalité des fruits de gros calibre analysés sont de couleur plus orangée alors que les fruits de taille plus petite présente des teintes plus foncées. Le seul marqueur associé au caractère b* (au seuil $\alpha < 0.005$) est le fragment TD120 (p-value =6,8e-4). Cette association ne correspond à aucun QTL identifié précédemment. Deux fragments sont associés avec le caractère L* : TD380 (un seul site également associé au poids du fruit) et TD316 (un seul site). Ces deux fragments, même s'ils sont situés à une distance importante, co-localisent avec le QTL *L2.1* identifié dans la population de cartographie. Le faible déséquilibre de liaison ($r^2 = 0,24$) entre ces deux marqueurs ne peut pas expliquer le fait qu'on identifie ces deux polymorphismes associés avec le même caractère. Aucun marqueur n'est associé avec le caractère a*.

La fermeté présente des effets cultivar et année significatifs (*p*-value < 2,2e-16). Ce caractère possède lui aussi une héritabilité très forte ($h^2 = 0,94$). Il n'est corrélé significativement à aucun caractère, cependant on le retrouve associé avec le fragment TD056 (*p*-value < 0,0049). Le polymorphisme associé est identique à celui associé au poids du fruit dans l'échantillon de 63 accessions de type cerise. Il apparaît qu'un QTL d'élasticité du fruit (résistance à la déformation) a été localisé près du QTL *FW2.2* dans la population Cervil x Levovil. Il est possible que cette association soit le fruit du déséquilibre de liaison entre le polymorphisme causal et le promoteur de *fw2.2*.

5.3.4. Etude des co-associations entre SSC, TA et les teneurs en acides et sucres

Une analyse de certains composés (sucres et acides) issus des mesures du métabolome réalisées au Max-Planck-Institut für molekulare Pflanzenphysiologie à Golm (Potsdam, Allemagne) a permis de vérifier s'il existait des co-associations entre la composition en sucres et acides avec les caractères acidité titrable et teneur en solides solubles. Etant donné que les mesures ont été réalisées de façon relative par rapport à un témoin interne, aucune information n'est disponible quant à la concentration réelle des composés par rapport au poids de matière fraîche du fruit.

Tout d'abord, aucune association n'est identifiée avec la teneur en malate. Les fragments TD086 et TD278, identifiés lors des études d'association avec l'acidité titrable, montrent des associations avec la quantité de citrate mais les *p*-values ne résistent pas à la correction pour les tests multiples. Des résultats équivalents sont obtenus pour le fragment TD094 qui se localise au niveau du QTL d'acidité titrable *TA2.1*, identifié dans la population Cervil x Levovil mais pas dans notre analyse.

Certains marqueurs entourant *lcn2.1* (TD047, TD133 et TD120) sont associés statistiquement à la quantité de glucose (*p*-value < 0.05). Les marqueurs TD387 et TD380 sont, eux aussi, associés à la quantité de glucose mais ces résultats sont sans doute dus à un effet de dilution de la matière sèche (et donc des sucres) dans des fruits de calibre important (corrélation négative entre SSC et FW). Dans les différentes populations de cartographie analysées, les QTL de teneur en sucres co-localisent souvent avec des QTL du poids du fruit (Prudent, Causse et al. 2009). Le polymorphisme TD047-274, associé à la quantité de glucose, est le même que celui associé à la teneur en solides solubles. Ce polymorphisme n'est plus associé à la teneur en glucose lorsqu'on étudie 201 accessions.

Les concentrations en glucose sont fortement corrélées aux teneurs en fructose, il n'est donc pas aberrant de retrouver les mêmes marqueurs associés avec les deux caractères. On retrouve des associations (*p*-value < 0.05) avec les fragments TD047, TD133 et TD120 qui se localisent dans la même région physique (environs 100 Kb). On retrouve aussi les fragments TD380 et TD387 qui témoignent, là aussi, d'un effet antagoniste entre la teneur en sucre et le poids du fruit. Ces trois marqueurs restent associés à la teneur en fructose lorsqu'on analyse 201 accessions. Le fragment TD140, identifié dans l'article (aldose-1-épimérase), présente lui aussi une association avec la teneur en fructose.

Normalement très peu de saccharose est retrouvé dans les fruits mûrs car ce sucre complexe est dégradé en glucose et fructose pour alimenter les voies énergétiques du fruit en croissance (Prudent, 2008). Nous avons quand même testé l'association entre les quantités relatives de saccharose chez les 90 accessions et les polymorphismes. De nombreux polymorphismes sont associés à la teneur en saccharose (153 polymorphismes sur 345 présentent une *p*-value corrigée inférieure à 0.005). Ces résultats aberrants témoignent d'un biais dans la méthode de dosage du saccharose, étant donné que de tels résultats n'ont été observés que pour ce caractère.

La figure 5-7 présente une comparaison des QTL détectés sur la population Cervil x Levovil et par génétique d'association sur la « core collection ». La majorité des QTL sont identifiés par les deux approches. Il est fort probable qu'en densifiant encore le chromosome en polymorphisme, les études d'association permettront d'identifier un plus grand nombre de QTL étant donné que la diversité utilisée est beaucoup plus grande.

5.4. Conclusion

Le développement et l'utilisation de « core collection » est un choix stratégique efficace dans la recherche de polymorphisme en vue d'identifier des locus impliqués dans la variation de caractères quantitatifs. Dans le cas de populations cultivées présentant très peu de polymorphismes, il semble intéressant de caractériser des accessions présentant un patron de diversité en mélange entre l'espèce cultivée et l'espèce apparentée sauvage. De telles accessions peuvent être rencontrées dans des zones sympatriques où il y a recouvrement de l'écosystème naturel des espèces sauvages et des zones de culture de l'espèce. Cette étude d'association est encourageante car, la plupart des QTL identifiés dans la population issue du croisement intra-spécifique, ont été retrouvés. Cette étude a notamment permis d'identifier de nouveaux polymorphismes candidats expliquant une part de la variation de caractères liés à la qualité du fruit. Cette étude a aussi permis de valider la méthode statistique d'association la plus pertinente : modèle Q+K sachant que cette méthode peut être utilisée sur d'autres échantillons. Si l'étendue du déséquilibre de liaison identifié sur le chromosome 2 peut être inféré à tout le génome, alors une analyse de type Genome Wide Analysis pourra être réalisée avec une résolution relativement fine. Jusqu'à maintenant les freins à de telles analyses étaient le coût d'identification des polymorphismes moléculaires et le coût de génotypage des polymorphismes identifiés, sur de grandes populations. L'arrivée sur le marché des

technologies de re-séquençage ou « Next-Generation Sequencing Technologies » permettent déjà d'obtenir, en peu de temps et pour un coût tout à fait raisonnable, une quantité conséquente de polymorphismes qui pourront être testé en association.

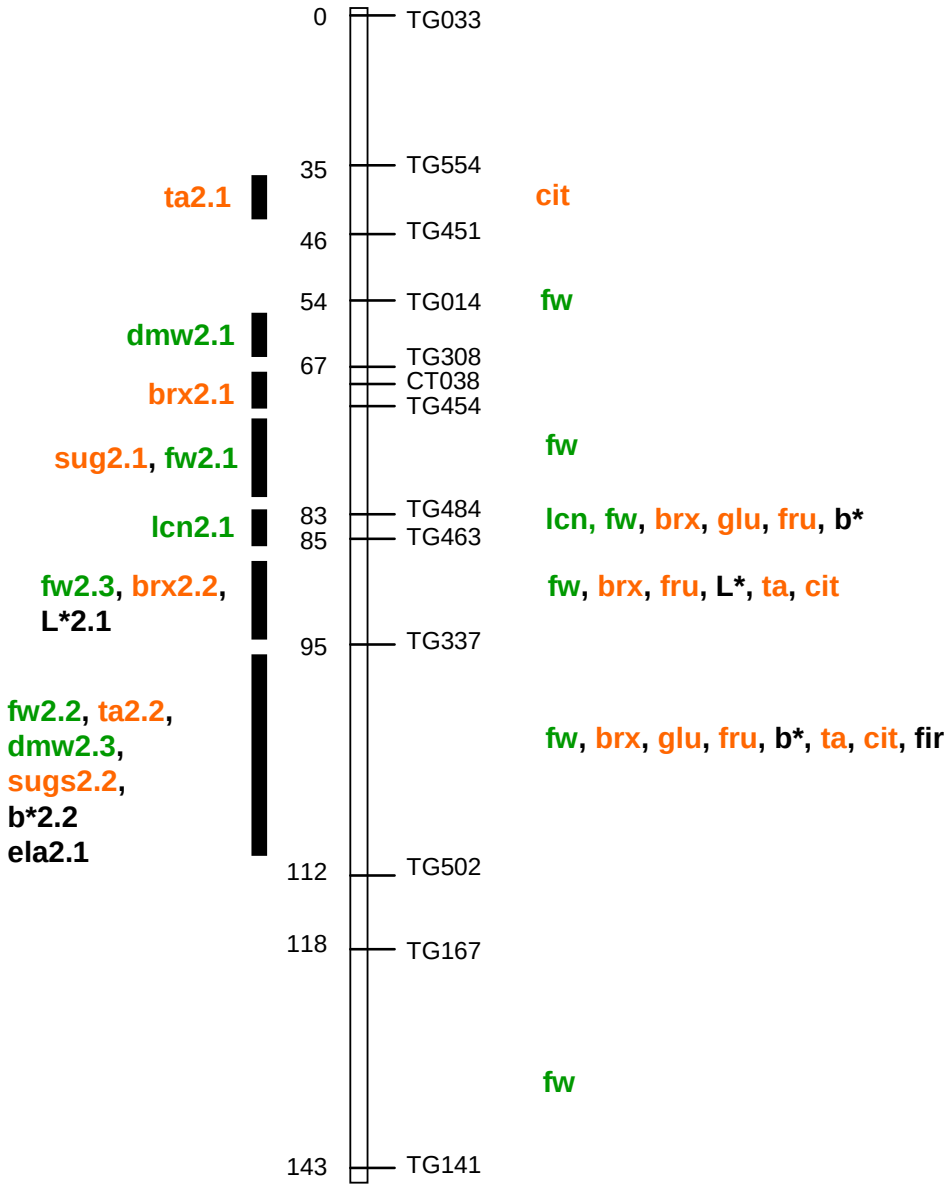


Figure 5-7. Comparaison des QTL identifiés par cartographie génétique et par études d'association sur le chromosome 2. Les QTL identifiés par cartographie génétique dans la population issue du croisement Cervil x Levovil sont indiqués à gauche du chromosome. Les QTL détectés par association sur la collection de 90 accessions de tomate sont indiqués à droite du chromosome. Les QTL sont nommés de la façon suivante : ta = acidité titrable, dmw = poids de matière sèche, brx = teneur en solides solubles, sug = teneur en sucres totaux, lcn = nombre de loges, L* a* b* = composantes colorimétriques, ela = élasticité, cit = teneur en citrate, glu = teneur en glucose, fru = teneur en fructose, fir = fermeté. Les QTL liés au citrate ne sont plus significatifs après correction pour tests multiples. Les distances indiquées pour les marqueurs sont celles de la carte génétique de référence (Expen2000).

Chapitre 6 : Modélisation de l'histoire évolutive de la tomate cultivée par simulation de coalescents

6.1. Introduction

La tomate est un fruit qui a subi d'importantes modifications morphologiques au cours de la domestication. Le changement majeur réside en l'augmentation, parfois extraordinaire, du poids du fruit. Nous avons vu que l'augmentation du nombre de loges semble avoir accompagné l'utilisation croissante de la tomate par l'homme. Etant donné que l'organe consommé ne se conserve pas, aucune trace archéologique témoignant d'une culture (ou d'une utilisation) précoce n'a pu être identifiée dans les régions d'Amérique Centrale ou de la cordillère des Andes. Les études génétiques n'ont pas pu identifier la zone de domestication et deux hypothèses continuent à s'opposer quant à l'origine de la tomate cultivée.

Un des apports de la biologie moléculaire est l'identification de mutations ponctuelles, le long d'un fragment génomique. Ces mutations portent la signature des évènements passés. De par la sélection exercée par l'homme durant la domestication et la réduction drastique de l'effectif efficace de la population originale, on s'attend à une fluctuation du polymorphisme entre le compartiment sauvage et le compartiment cultivé. Lorsqu'on s'intéresse à un gène potentiellement sélectionné pendant la domestication, il est relativement difficile d'identifier la part de la fluctuation du polymorphisme due à la sélection de celle due aux évènements démographiques.

Plusieurs méthodes permettent d'analyser la diversité moléculaire de séquences génomiques. Seules les plus courantes sont présentées ici.

L'indice π prend en compte le nombre de sites polymorphes et leurs fréquences respectives. Cet indice est calculé comme le nombre moyen de différences entre les séquences prises deux à deux, divisé par le nombre de nucléotides séquencés. Il se calcule de la façon suivante :

$\pi = \left(C_n^2\right)^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{i,j}$ avec n le nombre total de sites dans la séquence et $d_{i,j}$ la fréquence du nucléotide *j* à la position *i*.

L'indice θ est appelé paramètre mutationnel. L'estimateur le plus connu de cet indice a été défini par Watterson (1975) de la façon suivante $\theta_W = S / a_n$ où S désigne la proportion de sites polymorphe, a_n vaut $(1 + ½ + \ldots + {}^{1}/_{n-1})$ où n représente ici le nombre de séquences étudiées.

D'après les deux formules, on comprend que les allèles de fréquences faibles influent de façon importante sur θ_W mais par sur π. C'est sur cette propriété que le test D de Tajima (1989) est fondé. Celui-ci examine la relation entre les deux statistiques afin d'établir s'il y a excès ou défaut d'allèles rares par rapport au modèle neutre afin de détecter les traces de la sélection. La statistique D se calcule de la façon suivante : $D = (\pi - \theta_W)/[\mathrm{Var}(\pi - \theta_W)]^{1/2}$. Les pressions de sélection ont des effets différents sur le D de Tajima qui sont représentés dans le tableau 6-1.

Processus démographique ou génétique	**Signe attendu du D de Tajima**
Expansion démographique	-
Sélection d'un allèle favorable	-
Réduction de l'effectif démographique	+
Effet Wahlund (population structurée)	+
Sélection balancée (fréquence-dépendante)	+

Tableau 6-1. Interprétations du signe de D de Tajima en fonction des pressions évolutives démographiques ou génétiques.

La seule différence dans les patrons de polymorphisme laissés par l'une ou l'autre de ces pressions évolutives est la localisation de l'effet. Une pression de sélection s'exercera sur un locus (ou plusieurs locus dans le cas d'un réseau de régulation ou d'interaction) et modifiera le polymorphisme du locus ainsi que des régions directement adjacentes, par autostop génétique (Figure 6-1). Les évènements démographiques vont, quant à eux, avoir un impact sur tout le génome. Il est donc important de connaître les évènements qui ont modelé le polymorphisme moléculaire des populations étudiées afin d'éviter de confondre les deux pressions évolutives lors de la détection de traces de sélection d'un gène d'intérêt (Wright and Gaut 2005).

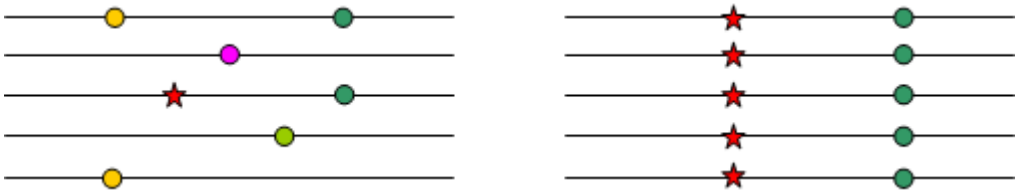


Figure 6-1. Schéma de l'effet de l'autostop génétique sur la diversité des locus encadrant un polymorphisme soumis à sélection positive. Les cercles représentent des allèles différents pour un locus neutre donné et l'étoile représente la mutation soumise à sélection positive. L'augmentation en fréquence de l'étoile rouge dans la population entraîne une augmentation de la fréquence du cercle vert associé. On assiste donc à une diminution de la diversité des locus adjacents.

Une méthode permettant d'étudier la vraisemblance d'une histoire évolutive consiste à confronter les données de polymorphisme moléculaire de marqueurs neutres, obtenues sur la population d'intérêt à des données obtenues par simulations d'échantillons suivant un modèle évolutif sans sélection. Le modèle évolutif le plus simple est le modèle standard neutre (Standard Neutral Model, SNM) décrit par Tajima (1989). Celui-ci suppose que l'échantillon a été collecté dans une population non structurée et de taille constante, ce qui est rarement le cas lorsqu'on travaille avec des populations naturelles ou avec des accessions issues de ressources génétiques.

La méthode permettant de simuler des échantillons suivant un modèle évolutif donné est basée sur la théorie de la coalescence (Weiss and von Haeseler 1998). La coalescence décrit la réunion de deux lignées phylogénétiques de séquences orthologues dans une population. On dit que deux séquences coalescent lorsqu'on identifie l'ancêtre commun le plus récent (Most Recent Common Ancester, MRCA). Le temps, en génération, qui sépare le MRCA des deux séquences est appelé temps de coalescence (Figure 6-2).

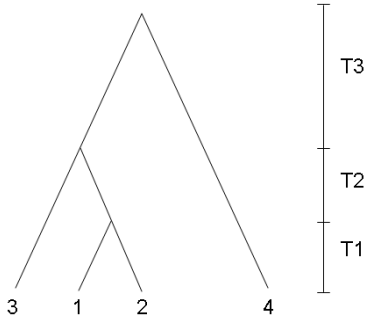


Figure 6-2. Arbre phylogénétique de quatre séquences (notées 1 à 4) et temps de coalescence de ces séquences. Le temps T1 sépare les séquences 1 et 2 de leur ancêtre commun, le temps T2 sépare les séquences 3 et 1-2 de leur ancêtre commun et le temps T3 sépare les séquences 4 et 1-2-3 de leur ancêtre commun.

Dans une population à l'équilibre Wright-Fisher (taille constante, générations disjointes, pas de sélection, accouplement aléatoire), chaque individu peut être vu comme « prélevant » ses parents de façon aléatoire dans la génération précédente. Il en découle que la généalogie d'un groupe d'individus peut être générée simplement en retraçant les évènements de coalescence entre lignées, en remontant le temps génération après génération, jusqu'à éventuellement atteindre un ancêtre commun.

Dans les méthodes de simulations d'échantillon par coalescence, une généalogie est simulée suivant un modèle évolutif puis les mutations sont distribuées sur chaque branche en fonction du modèle mutationnel (taux de mutation, taux de transition, taux de transversion) prédéfini par l'utilisateur. Le modèle prend en compte la taille de l'effectif efficace *Ne* de la population (effectif d'une population idéale à l'équilibre de Wright-Fisher pour laquelle on aurait une fluctuation de polymorphisme équivalente à celle de la population naturelle) et les évènements démographiques (goulet, expansion, migration). Certains programmes de simulations prennent maintenant en compte le taux de recombinaison entre sites.

6.2. Matériel et méthodes

Le logiciel BayeSSC, utilisé pour les simulations, est construit sur la base de Simcoal 1.0 (Excoffier, Novembre et al. 2000) mais présente une méthode de simulation plus flexible. Il réalise automatiquement une sortie analysée des simulations demandées avec un résumé de statistiques telles que le nombre de sites en ségrégation, le nombre d'haplotypes, l'indice π de diversité moléculaire ainsi que le D de Tajima. Trois modèles différents vont être testés sur les deux populations de tomates domestiquées et sauvages : un modèle de taille de population constante, un modèle incluant un goulet d'étranglement 500 générations dans le passé et un modèle plus compliqué incluant une séparation des deux populations, un goulet d'étranglement, une croissance démographique 500 générations dans le passé et des évènements de migration (Figure 6-3).

Les séquences étant de longueur relativement faible, on néglige l'effet de la recombinaison intra-fragment, ce qui va simplifier la méthode de simulation de coalescence utilisée. BayeSSC va permettre de simuler l'histoire évolutive des populations de tomate en s'appuyant sur les résultats de diversité des séquences obtenus sur le pool sauvage et cultivé. Ces simulations permettront de paramétrer l'importance du goulet d'étranglement et la

période la plus probable où ce goulet a pu avoir lieu. BayeSSC va générer des généalogies qui seront fonction du modèle indiqué puis les mutations seront placées le long de cette généalogie en utilisant le modèle de site fini avec deux états alléliques potentiels pour chaque site.

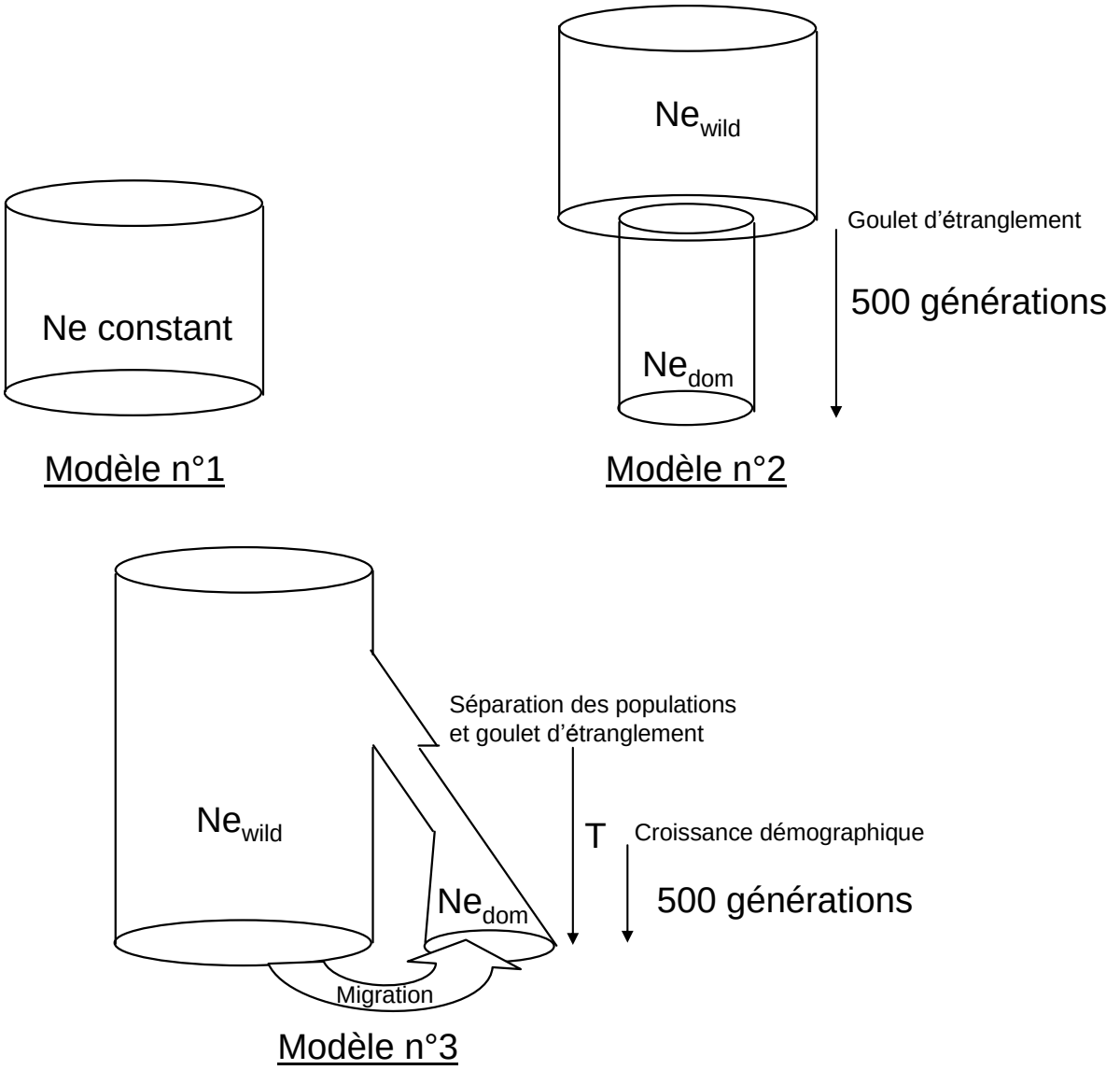


Figure 6-3. Modèles évolutifs utilisés dans l'analyse.
Le modèle n°1 est un modèle de taille constante, le modèle n°2 implique un goulet d'étranglement 500 générations dans le passé et le modèle n°3 intègre une division des deux populations, un goulet d'étranglement, une croissance démographique et la migration d'accessions sauvages vers le groupe cultivé. Ne_{dom} et Ne_{wild} représentent respectivement la taille efficace de la population domestiquée et de la population sauvage.

Vingt séquences de 550 bp seront simulées pour les deux populations « sauvage » et « domestiquée » et quatre statistiques seront définies sur ces deux jeux de séquences : le nombre de sites polymorphes, le nombre d'haplotypes, la diversité nucléotidique π et le D de Tajima. Pour chaque modèle, 1000 généalogies sont simulées où un seul paramètre est déclaré comme variable suivant une distribution définie à l'avance. Cette distribution est ensuite découpée en classes et nous calculons un score relatif à la vraisemblance par rapport aux données observées, pour chacune des classes. La méthode décrite par Belle, Ramakrishnan et al. (2006) a été adaptée afin de prendre en compte la variation d'un paramètre dans le calcul du score. Ensuite un script a été développé pour R afin d'analyser les sorties de BayeSSC et directement étudier la vraisemblance des modèles par rapport aux données réelles (http://www.stanford.edu/group/hadlylab/ssc/index.html). Un plus grand nombre de simulations peut être analysé en même temps.

Pour chaque statistique obtenue, la vraisemblance empirique P est calculée de la façon suivante : supposons que la statistique observée est x, qui se classe dans le $k^{\text{ième}}$ rang parmi S valeurs simulées de moyenne m. Si $x > m$, la vraisemblance empirique se calcule comme le nombre de simulations donnant des valeurs supérieures à x divisé par S. Donc on compte le nombre de simulations donnant une valeur supérieure à x dans la queue droite de la distribution puis on double ce nombre pour obtenir un test bilatéral. On réalise l'analyse symétrique si $x < m$. Quand la statistique observée tombe hors de la gamme de variation des valeurs simulées, on fixe P=0.0005 comme une estimation conservative de la vraisemblance. Afin de combiner les probabilités sur plusieurs statistiques, on utilise la méthode de combinaison de probabilité de Fisher. Ce test suppose que les probabilités sont indépendantes. Ce n'est pas le cas ici mais nous négligerons les effets de dépendances entre π et le D de Tajima. De cette manière, le test statistique, qui suit une loi du χ^2, a été estimé à partir des quatre statistiques analysées.

Pour tous les modèles, l'estimation du taux de substitution par site et par an chez la tomate est fixée à la valeur donnée par Gaut (1998) pour des gènes nucléaires chez les végétaux supérieurs et validé chez la tomate (Ku, Vision et al. 2000; Nesbitt and Tanksley 2002). Cette valeur est fixée à 6.03e-9 substitutions silencieuses par site et par an. Même si la tomate sauvage est une espèce potentiellement pérenne, on fixe le taux de génération à une génération par an. Le taux de mutations (probabilité que la séquence subisse un évènement de mutation par génération) est alors égal à 3.3e-6 ($6.03e-9 \times 1 \times 550\,\text{sites}$).

Les séquences de certains fragments décrits dans l'article présenté dans le chapitre 5 ont été utilisées pour étudier la diversité du chromosome 2. Le fragment *lcn2.1* n'a pas été retenu car il apparaît avoir évolué différemment par rapport au chromosome. On utilise toutes les séquences disponibles obtenues à partir de 92 accessions sauvages et cultivées (*S. l. esculentum*, *S. l. cerasiforme* et *S. pimpinellifolium*) Les accessions sont classées en fonction de leur appartenance aux groupes « sauvage » (Wild) ou « domestiqué » (Domesticated) identifiés par STRUCTURE dans le chapitre 3.

Vingt trois fragments, ciblant majoritairement des régions non codantes, sont considérés comme neutres. Toutes les séquences ne sont pas disponibles pour tous les individus (données manquantes) mais, en moyenne, 60 séquences par fragment dans le groupe « domestiqué » et 18 séquences par fragment dans le groupe « sauvage » ont été utilisées. La diversité moléculaire des séquences pour les 23 fragments a été analysée à l'aide du programme DNAsp5.0.

6.3. Résultats et Discussion.

Le tableau 6-2 résume la diveristé moléculaire, identifiée pour chaque fragment, pour les deux groupes d'accessions.

On remarque que les nombres moyens de polymorphismes détectés par fragment ainsi que les nombres moyens d'haplotypes sont sensiblement les mêmes entre le pool sauvage et le pool domestiqué. La diversité nucléotidique est deux fois plus faible chez le pool domestiqué que chez le pool sauvage. Le D de Tajima moyen pour le pool domestiqué est négatif alors que celui du pool sauvage est proche de zéro, mais avec un écart-type relativement élevé. Le premier modèle testé (modèle n°1, Figure 6-3) est un modèle où la taille effective de la population sauvage est constante au cours du temps. Ce modèle a été utilisé pour simuler la population sauvage. Ces simulations permettent de définir les effectifs efficaces les plus vraisemblables pour cette population. La taille effective a été déclarée comme variable suivant une distribution uniforme entre 100 000 et 500 000 individus. Les simulations montrent qu'un effectif efficace compris entre 300 000 et 450 000 individus semble le plus vraisemblable (Figure 6-4). Un plus grand nombre de simulations sont réalisées en faisant varier le nombre d'individus entre 300 000 et 450 000. Cette nouvelle analyse permet de redéfinir à 330 000 l'effectif efficace le plus vraisemblable pour la population sauvage.

	L [a]	S [b]		H [c]		π [d]		D Tajima	
TD		*dom*	*wild*	*dom*	*wild*	*dom*	*wild*	*dom*	*wild*
TD096	662	8	8	2	2	0.0020	0.0044	-0.543	0.900
TD095	448	4	5	5	6	0.0011	0.0029	-0.842	-0.313
TD094	674	6	7	5	6	0.0015	0.0026	-0.515	-0.587
TD093	438	1	1	2	2	0.0005	0.0012	0.108	1.547
TD092	570	5	4	6	5	0.0014	0.0022	-0.597	0.220
TD091	609	4	4	5	5	0.0006	0.0014	-1.250	-0.778
TD090	676	5	7	4	6	0.0021	0.0032	0.788	0.177
TD139	524	2	1	3	2	0.0008	0.0009	-0.120	1.321
TD100	570	12	12	5	5	0.0032	0.0053	-0.829	-0.344
TD133	575	4	6	4	5	0.0012	0.0022	-0.409	-0.875
TD088	382	1	0	2	1	0.0001	0.0000	-0.903	NA
TD120	451	10	9	6	6	0.0041	0.0071	-0.379	0.534
TD047	577	7	6	4	6	0.0022	0.0033	-0.455	0.314
TD187	554	12	9	8	6	0.0023	0.0037	-1.399	-0.671
TD188	250	1	1	2	2	0.0008	0.0004	-0.056	-1.165
TD121	485	11	17	7	8	0.0017	0.0065	-1.803	-1.302
TD107	433	3	2	4	2	0.0013	0.0019	-0.252	1.030
TD130	555	7	8	7	8	0.0021	0.0047	-0.559	0.441
TD086	729	5	7	3	4	0.0007	0.0030	-1.195	0.346
TD085	807	0	5	1	5	0.0000	0.0007	NA	-1.966
TD114	647	21	19	6	3	0.0060	0.0151	-0.480	2.895
TD129	244	1	3	2	4	0.0003	0.0019	-0.890	-1.377
TD083	775	9	8	8	8	0.0014	0.0037	-1.253	0.763
Moyenne		6.04	6.48	4.39	4.65	0.0016	0.0034	-0.629	0.051
Ecart type		4.9	4.77	2.06	2.1	0.0014	0.0031	0.568	1.132

[a] Longueur du fragment en nucléotides sans compter les indels.
[b] Nombre de sites polymorphes.
[c] Nombre d'haplotypes différents.
[d] Diversité nucléotidique.

Tableau 6-2. Analyse de la diversité des fragments localisés sur le chromosome 2.
Le groupe domestiqué est indiqué par dom et le groupe sauvage est indiqué par wild.

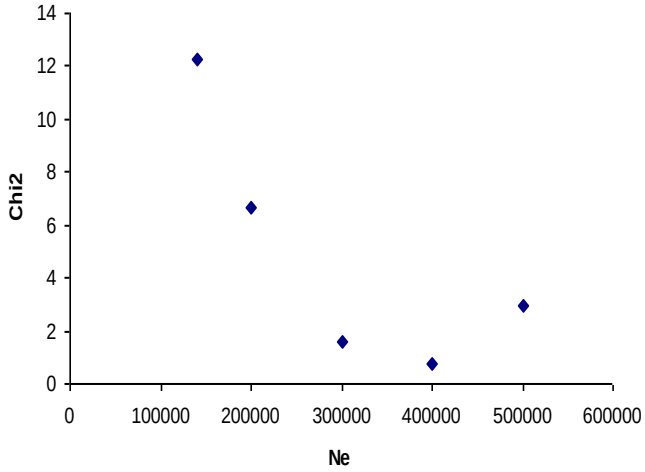


Figure 6-4. Variation de la vraisemblance du modèle par rapport aux données réelles en fonction de Ne_{wild}.

Compte tenu que la domestication a impliqué un fort goulet d'étranglement, le modèle a été complété par une restriction de la taille efficace de la population (Figure 6-3). Aucune information sur le moment où a eu lieu ce goulet n'est disponible. Rick (1990) propose que la tomate a été domestiquée récemment car des habitants du Pérou indiquent n'utiliser cette plante que depuis quelques générations. Le goulet d'étranglement est donc fixé arbitrairement à 500 générations dans le passé. Les simulations permettent de vérifier quel effectif de l'espèce cultivée est le plus vraisemblable en fonction de l'importance du goulet d'étranglement. Seules des divisions de l'effectif efficace par 2, 4, 6 ou 8 sont testées. Pour calculer la vraisemblance, nous nous sommes intéressés uniquement à la diversité nucléotidique π car c'est la statistique qui décrit le mieux la diminution de la diversité due au goulet d'étranglement. Le tableau 6-3 montre une corrélation entre la taille efficace de la population domestiquée et l'importance du goulet d'étranglement.

Ne	div/2	div/4	div/6	div/8
10000	8.24	5.07	3.73	2.19
20000	5.90	3.20	1.97	**1.16**
30000	4.25	1.90	**1.23**	1.77
40000	3.03	**1.09**	1.87	3.30
50000	2.49	1.52	2.80	4.10
60000	1.92	2.16	3.63	4.87
70000	**0.85**	2.40	4.65	5.91
80000	0.93	2.95	5.28	7.00
90000	1.07	3.64	5.34	7.82
100000	1.56	4.37	5.84	7.38

Tableau 6-3. Tableau représentant les valeurs de χ^2 en fonction de la force du goulet d'étranglement (division de Ne par 2, 4, 6 ou 8) et de l'effectif efficace final (Ne). Les valeurs en gras sont les valeurs les plus vraisemblables par rapport aux données observées.

La vraisemblance la plus forte indique une taille efficace de la population domestiquée inférieure à 70 000 individus. Cependant, le goulet d'étranglement a dû être très important lors de la domestication de la tomate. Il est fort probable que la taille efficace de la population originale ait été divisée par un facteur supérieur à 8. De nouvelles simulations devront être réalisées en prenant en compte une plus grande gamme de variation pour l'importance du goulet.

Le D de Tajima moyen du groupe domestiqué est négatif ce qui implique une expansion récente de la population. Le modèle final doit donc inclure une phase de croissance de la population domestiquée après le goulet d'étranglement. La croissance d'une population

est modélisée par l'équation $N(t) = N(0)e^{rt}$ ou $N(0)$ représente l'effectif efficace au temps 0, $N(t)$ représente l'effectif efficace au temps t et r représente le taux de croissance. Etant donné que les simulations de coalescence fonctionnent en remontant le temps, un taux de croissance négatif indique que la population au temps t sera de plus grande taille qu'au temps t=0. Le modèle a été complété en créant une subdivision entre la population sauvage et la population domestiquée au moment du goulet d'étranglement (Figure 6-3). Dans ce nouveau modèle, la séparation entre les deux populations et le goulet d'étranglement sont contemporains et ont eu lieu T générations dans le passé. Le début de la phase d'expansion est fixé à 500 générations dans le passé. On utilise ce dernier modèle pour rétablir différents paramètres variables, suivant une loi uniforme:

- Ne_{wild} varie entre 260 000 et 360 000 individus

- Ne_{dom} varie entre 10 000 et 70 000 individus

- Le taux de croissance t varie entre -5 et -0.5

- T, le nombre de générations suivant la séparation entre les populations et le goulet d'étranglement, varie entre 500 et 10 000 générations.

Ce nouveau modèle n'est pas valide car les populations domestiquées simulées ne présentent aucun polymorphisme. Des inter-croisements récurrents ont eu lieu entre *S. pimpinellifolium* et *S. lycopersicum* (Rick and Holle 1990; Ranc, Munos et al. 2008). Des flux de gènes entre les deux populations ont donc été rajoutés au modèle (évènement de migration), afin d'augmenter le niveau de polymorphisme des populations domestiquées simulées. Le taux de migration représente la probabilité qu'à chaque génération, toutes les lignées migrent d'une population vers l'autre. L'importance des flux de gènes est définie comme variable. Les flux sont unilatéraux, du pool sauvage vers le pool cultivé.

Les valeurs les plus vraisemblables pour les différents paramètres testés sont les suivantes :

- Ne_{wild} est compris entre 300 000 et 330 000 individus

- Ne_{dom} est compris entre 12 000 et 15 000 individus

- Le taux de croissance t est égal à -3 à -1

- Le taux de migration de la population sauvage vers la population domestiquée est compris entre 0.1 et 0.3

- T varie entre 3000 et 5000

Un nombre plus important de simulations a été réalisé avec ce même modèle et les résultats ont été analysés grâce au script développé sous R. L'approximation bayesienne développée permet de vérifier la probabilité relative d'obtenir des statistiques équivalentes à celles obtenues sur le vrai jeu de données, pour différentes valeurs des paramètres (priors). Ces simulations montrent que les paramètres ne peuvent pas être estimés avec beaucoup plus de précisions. De plus, l'écart type des différentes statistiques observées n'a pas été pris en compte lors du calcul de la vraisemblance. Ceci implique une grande simplification du système.

6.4. Conclusion.

Ces simulations montrent cependant plusieurs points importants dans la domestication de la tomate. Tout d'abord, la restriction dans l'effectif efficace a été très importante, avec une diminution vraisemblable de 300 000 à 15 000 individus. Les échanges entre les deux compartiments sauvage et domestiqué ont permis de maintenir de la diversité dans le pool domestiqué. La domestication de la tomate semble beaucoup plus ancienne que ce qui était admis jusqu'à maintenant. La comparaison de statistiques de diversité observées à des statistiques obtenues par simulations de coalescents a déjà été utilisée auparavant chez le riz, afin d'établir la force du goulet d'étranglement subit par cette espèce lors de la domestication (Zhu, Zheng et al. 2007). Ce type d'analyse peut aussi être appliqué à des marqueurs microsatellites. Par exemple, Gao et Innan (2008) utilisent des informations de génotypage avec des marqueurs SSR combinées à des simulations de coalescents pour démontrer que la domestication des deux espèces de riz cultivés ne sont pas indépendantes.

Cette étude a notamment permis de constituer un modèle qui permettra de simuler des jeux de données auxquels pourront être confrontées des données réelles. On pourra alors confronter des gènes d'intérêt à un modèle neutre impliquant seulement des évènements démographiques afin de vérifier si les fluctuations de la diversité sont dues à la sélection ou seulement aux évènements démographiques.

Ce modèle pourra être complété en séparant la population domestiquée, majoritairement composées d'accessions *S. l. cerasiforme*, et la population cultivée composée d'accessions *S. l. esculentum*. Un des problèmes majeurs qui sera rencontré est le faible taux de polymorphisme du groupe cultivé qui ne permettra pas d'estimer avec précision les paramètres suivants la domestication.

Chapitre 7 : Discussion et perspectives

L'objectif de cette thèse était de valider l'utilisation des ressources génétiques chez la tomate en vue de disséquer les bases génétiques des composantes de la qualité du fruit. De nombreuses études antérieures se sont focalisées sur l'utilisation de la génétique d'association comme méthode d'identification de locus ayant un impact sur la variation de caractères d'intérêt (Zhu, Gore et al. 2008). Toutes ces études sont unanimes, il est primordial de bien connaître la structure génétique de l'échantillon et d'intégrer cette information dans les modèles d'analyses afin de limiter le nombre de faux positifs détectés. Une meilleure connaissance des ressources génétiques qui sont à notre disposition permettra une valorisation plus efficace de celles-ci.

7.1. Structuration des ressources génétiques de tomate, intérêts et limites des SSR.

7.1.1. Aujourd'hui, les marqueurs SSR…

Un des premiers freins à l'utilisation des ressources génétiques de tomates cultivées et sauvages proches était le manque de connaissance sur la structuration génétique de ces ressources génétiques. Les marqueurs SSR représentaient jusqu'à maintenant des marqueurs de choix pour l'étude de la diversité génétiques de populations naturelles ou de ressources génétiques. En effet, ils peuvent être multiplexés facilement, sont peu coûteux et peuvent être analysés automatiquement grâce à l'utilisation de fluorochromes. Leur nature multi-allélique offre un pouvoir discriminant des individus beaucoup plus élevé que des marqueurs de type bi-allélique. Cependant, il semble nécessaire de bien prendre en compte la nature des motifs microsatellites dans le choix des différents marqueurs car elle peut influer sur la diversité potentielle que peut offrir le marqueur en termes de nombre d'allèles. Par exemple nous avons montré que des motifs répétés riches en A et en T présentent des taux de mutations plus élevés que les marqueurs riches en G et en C ou les marqueurs mixtes. Ces marqueurs seront donc très intéressants pour étudier les relations entre individus au sein de populations présentant une faible diversité génétique. Ce très fort taux de mutation peut néanmoins présenter certains inconvénients. Le plus important est sans doute l'homoplasie, cas où deux alléles de deux individus peuvent être identiques par état mais pas par descendance (Curtu, Finkeldey et al.

2004). En effet deux allèles ancestraux différents ont plus de chance de donner dans leur généalogie des allèles identiques si leur taux de mutation est élevé. Plusieurs modèles évolutifs ont été décrits pour les marqueurs microsatellites mais deux d'entre eux semblent avoir reçu plus d'attention. Le plus simple est le modèle dit d'allèle infini (Infinite Allele Model) qui décrit chaque allèle différent comme issu d'un seul évènement d'addition ou d'élimination d'une ou plusieurs répétitions. L'autre modèle, le Step-Wise Mutation Model (SMM) décrit comment un allèle de longueur *n* répétitions peut être issue de l'allèle *n-x* ou *n+x* avec *x* le nombre de répétitions ajoutées ou éliminées. La prise en compte d'un tel modèle permet de dresser non plus une étude de la diversité des individus mais une étude phylogénétique de ceux-ci car on peut retracer l'histoire évolutive de chaque allèle et caller une notion de temps sur les branches séparant deux individus. Ici, la nature des allèles identifiés n'a pas permis d'utiliser le modèle SMM. En effet, les pas de mutation n'ont pas été respectés pour aucun des marqueurs. Une part de la variation de la longueur des allèles est provoquée par des mutations dans la séquence flanquant le motif. Ces marqueurs n'ont donc pas pu retracer la phylogénie des accessions.

Les marqueurs microsatellites les plus utiles pour une espèce peu polymorphe sont les marqueurs possédant de nombreuses répétitions (supérieures à 30). Ces marqueurs offrent un niveau de variabilité supérieur mais se regroupent en clusters près des centromères (Areshchenkova and Ganal 1999). Cette propriété empêche leur utilisation si l'étude vise à échantillonner une grande part du génome.

Les marqueurs SSR ont été des marqueurs de choix pour les études de diversité car parce qu'ils sont multi-alléliques, basés sur la PCR, reproductibles et sélectivement neutres. Cependant, avec une densité sur le génome beaucoup plus grande, un taux de mutations plus faible et le développement de techniques « haut-débit », les SNP sont en train de devenir des marqueurs de choix pour de telles études.

7.1.2. Demain, les SNP…

Un set de marqueurs SNP a été défini au cours de la thèse mais les résultats de génotypage ne sont pas encore disponibles. Une « core collection » de 24 accessions a été utilisée pour identifier du polymorphisme dans les introns de gènes candidats. Les polymorphismes détectés, ainsi que les SNP récupérés dans la bibliographie (van Deynze, Stoffel et al. 2007) sont venus incrémenter la collection de SNP identifiés sur les 96

accessions afin de construire quatre panels SNPlex® (Gut 2001), constitués de 48 SNP ou Indel chacun. Tous ces polymorphismes sont en cours de génotypage sur la collection totale d'Avignon. Ces polymorphismes forment un panel d'intérêt pour ré-analyser la structure de l'échantillon. Avec un taux de données manquantes d'environ 20% déjà observé, nous aurons plus de 150 marqueurs bi-alléliques à notre disposition, répartis sur le génome. Ces marqueurs sont potentiellement neutres car ils ont été détectés dans les introns et les régions intergéniques. Les résultats pourront être comparés à ceux déterminés par le génotypage des marqueurs SSR. De plus, tous ces SNP ont été choisis sur la base de leurs fréquences alléliques dans l'échantillon de 24 accessions. Très peu d'allèles rares, moins informatifs, ne devraient se trouver dans ce panel.

7.1.3. La structure de la collection de 340 tomates cultivées et sauvages proches.

Le génotypage de 20 marqueurs SSR a permis d'analyser la structuration de la diversité présente dans l'échantillon de 340 accessions. Le logiciel STRUCTURE (Pritchard, Stephens et al. 2000), couplé à la méthode d'Evanno (2005), a permis de déterminer la structuration la plus probable de la collection. La structuration identifiée avec cette méthode était cohérente à celle identifiée par l'analyse en coordonnées principales (ACoP) qui se base sur la dissimilarité entre individus pris deux à deux. Ces deux méthodes sont complémentaires car, alors que l'une donne des résultats de probabilité d'appartenance de chacune des accessions à une population, l'autre montre bien l'écart existant entre les populations et la diversité génétique intrinsèque à chacune des populations. D'autres méthodes d'analyses bayesiennes sont disponibles pour la communauté scientifique pour étudier la structure génétique à partir de marqueurs moléculaires (Holsinger, Lewis et al. 2002; Corander, Waldmann et al. 2003; Corander, Waldmann et al. 2004; Holsinger and E. 2004). Ces méthodes n'ont pas été testées ici mais peuvent présenter certains avantages, comme le fait de ne pas forcer de structure lorsque celles-ci n'existent pas (Corander, Waldmann et al. 2003), ce qui n'est pas le cas de STRUCTURE.

Le premier niveau de structuration sépare les accessions domestiquées des accessions sauvages. Le deuxième niveau de structuration permet de séparer les accessions portant de gros fruits des accessions portant de petits fruits parmi le pool domestiqué. La structure est ici fortement corrélée avec un caractère d'intérêt pour la domestication et l'agriculture moderne. La sélection exercée par l'homme explique sûrement cette structure.

7.1.3.1. La structuration chez S. pimpinellifolium semble être expliquée par des différences du taux d'autogamie des accessions.

Le deuxième niveau de structuration différencie deux populations d'accessions sauvages. La seule différence trouvée entre ces deux populations est le taux de fleurs possédant des styles introrses dans le cône d'étamines. Ce caractère témoigne du taux potentiel d'intercroisement. Le groupe A qui présente 100% de fleurs brévistyles (donc fortement autogames) possède une diversité génétique plus faible que la population B, où 50% des accessions ont des fleurs longistyles. Un des locus contrôlant l'insertion du style chez la tomate a été cloné (Chen, Cong et al. 2007). *STYLE2.1* code pour un facteur de transcription qui régule l'élongation des cellules dans le style en développement. Il serait intéressant d'étudier les traces de sélection autour de ce gène pour les deux groupes d'accessions sauvages afin de vérifier s'il y a eu ou pas sélection pour le passage à l'autogamie d'une partie des accessions sauvages. Etant donné que le phénotype est fortement lié à l'environnement ces informations sont à prendre néanmoins avec beaucoup de précautions.

Il semble aussi intéressant d'augmenter le nombre d'accessions *S. pimpinellifolium* dans les analyses futures, étant donné sa proximité génétique avec la tomate cultivée. Cette espèce a sans doute été le point de départ de la domestication de la tomate, contrairement à ce qui est admis par Rick (1976), qui postule qu'il s'agit d'une espèce indépendante.

7.1.3.2. Les accessions de type cerise présentent une position « admixture » originale.

L'étude de la structure génétique de la collection permet aussi d'observer une différence entre les accessions de type cerise qui se regroupent dans une population proche des accessions de type gros fruits et les accessions de même type qui dessinent un *continuum* entre le compartiment sauvage et le compartiment cultivé. Ceci apporte la preuve qu'il y a eu deux histoires complètement différentes dans l'évolution de la tomate cerise. Une hypothèse probable est que la domestication à partir de *S. pimpinellifolium* a entraîné une réduction drastique de la diversité génétique de la tomate, au début de sa culture par l'homme. Les premières accessions utilisées devaient présenter une taille de fruit relativement réduite équivalente aux fruits de tomate de type cerise. L'homme a continué à sélectionner la tomate jusqu'à ce que la taille du fruit augmente. Lorsque les conquistadors envahissent le Mexique, ils rapportent avec eux quelques accessions avec des fruits moyens à gros. Les premières représentations, dessinées en Europe, montrent des fruits fasciés relativement gros. Ceux-ci

seront encore sélectionnés en Europe puis aux Etats-Unis jusqu'à donner des tailles de fruits extrêmes. Les accessions de type cerise qui existaient à l'état sauvage au Mexique et dans la cordillère des Andes ont continué à s'intercroiser librement avec des accessions cultivées et des accessions sauvages dans des zones de sympatrie. D'autres accessions de type cerise ont continué à être utilisées comme variétés locales. L'histoire de la tomate cerise n'est donc pas simplement celle d'une forme transitoire entre les formes cultivées et les formes sauvages.
Des intercroisements récurrents entre les différentes formes ont eu lieu tout au long de son histoire.

Le *continuum* formé par les accessions *S. l. cerasiforme* apporte une diversité génétique qu'on croyait présente uniquement chez l'espèce sauvage. En plus d'une diversité génétique importante, ces accessions possèdent une grande variabilité au niveau phénotypique. Les accessions sauvages ne présentent pas de diversité morphologique pour la taille du fruit, ce caractère, qui sépare l'espèce sauvage de l'espèce cultivée et fait partie du syndrome de domestication.

7.1.3.3. Le nombre de marqueurs SSR est insuffisant pour mettre en évidence une structuration des accessions cultivées.

Aucune structure génétique n'a pu être identifiée au sein du groupe cultivé moderne. Un plus grand nombre de marqueurs SSR hautement polymorphes serait nécessaire. Chez le maïs, qui est l'espèce modèle pour les études sur la diversité génétique des plantes cultivées, pas moins de 99 marqueurs SSR sont utilisés pour détecter la structure génétique de l'échantillon. Une autre étude de diversité chez la tomate, basée sur l'utilisation de SNP, montre qu'il existe une différenciation génétique entre des accessions cultivées de frais, des accessions cultivées d'industries, des variétés anciennes, des variétés mexicaines et des accessions sauvages (*S. pimpinellifolium*) (Sim, Robbins et al. 2009). Cette étude est réalisée en considérant la structure génétique avec *a priori*. Il serait intéressant d'utiliser ce panel de SNP sur la collection d'accessions cultivées d'Avignon afin de valider cette structure à partir du modèle Bayesien implémenté dans STRUCTURE.

7.1.4. Construction de « core collections » emboîtées.

L'échantillonnage des « core collections » a été volontairement focalisé sur les accessions appartenant au groupe botanique des *S. lycopersicum* var. *cerasiforme*. Certains individus de type sauvage *S. pimpinellifolium*, ainsi que du type cultivé, *S. lycopersicum* var

esculentum ont été rajoutés afin de pouvoir déterminer l'origine de chaque allèle. L'échantillonnage a été réalisé séparément sur chaque espèce car, lorsque la collection totale est utilisée, une grande part de la « core collection » est constituée par des accessions sauvages. Lorsque cette « core collection » de 96 individus a été échantillonnée, les résultats de STRUCTURE n'étaient pas encore disponibles. Il serait intéressant de ré-échantillonner une « core collection » à partir des 340 accessions en indiquant en co-variable, le degré d'appartenance à chacune des sous-populations, indiqué par STRUCTURE, pour que ce nouvel échantillon représente au mieux la structuration initiale.

Etant donné le faible polymorphisme identifié chez la tomate, on s'attendait à ce que l'échantillonnage d'individus représentant la collection totale à l'aide de MSTRAT présente un gain substantiel par rapport à un échantillonnage aléatoire (Bataillon, David et al. 1996). Ce gain dans la procédure d'échantillonnage est apporté majoritairement par les marqueurs moléculaires et beaucoup moins avec les caractères phénotypiques.

Les « core collections » emboîtées représentent une nouvelle ressource disponible pour toute la communauté scientifique s'intéressant à l'étude de la diversité chez la tomate avec des objectifs divers et variés : détection de polymorphisme, histoire évolutive, domestication, détection de QTL, diversité de gènes d'intérêt.

7.1.5. Les « core collections » dans la découverte de gènes d'intérêt.

Ces « core collections » sont nécessaires pour mieux utiliser la diversité génétique naturelle présente dans les collections de ressources génétiques. Elles représentent un échantillon de taille moyenne permettant d'assurer une puissance statistique confortable tout en se concentrant sur un nombre relativement restreint d'individus. De nombreuses « core collections » ont été établies auparavant chez *A. thaliana, M. truncatula* et la vigne (*vitis vinifera*) et les résultats d'intérêt qu'on leur doit ne sont plus à débattre (McKhann, Camilleri et al. 2004; Ronfort, Bataillon et al. 2006; Le Cunff, Fournier-Level et al. 2008). Des analyses de diversité réalisées sur la « core collection » échantillonnée pour *A. thaliana* ont permis de mettre en évidence une structure régionale au niveau de populations européennes (Ostrowski, David et al. 2006). Les auteurs montrent que cette structuration engendre un déséquilibre de liaison significatif entre marqueurs non liés, ce qui engendrera un fort taux de faux positifs dans le cas d'études d'associations, si elle n'est pas prise en compte. Sur la vigne, la « core collection » de 141 individus a permis d'identifier des polymorphismes dans des facteurs de

transcription de la famille des Myb associés à le teneur en anthocyane dans les baies de raisin qui est responsable de la coloration de celles-ci (Fournier-Level, Le Cunff et al. 2009).

7.2. Potentiels et limites de l'analyse d'association chez la tomate, une espèce cultivée hautement autogame.

7.2.1. Taille de l'échantillon.

Toutes les associations identifiées durant ce travail sont une nouvelle source de cibles intéressantes en vue de caractérisations futures de QTL. Cependant, il est important de noter que dans ces études d'associations, les « core collections » restent des échantillons de taille restreinte et qu'il est nécessaire de valider les résultats obtenus en testant un plus grand nombre d'accessions. Nous n'avons réalisé cette vérification que sur les SNP associés très significativement avec un caractère. Il semble que les 201 accessions présentent une taille d'échantillon encore trop faible pour identifier des polymorphismes ayant un effet faible sur les caractères étudiés. En effet, une étude utilisant des simulations de jeux de données a montré qu'un échantillon de 500 accessions permet de détecter avec suffisamment de puissance, des polymorphismes avec des effets relativement faibles sur le phénotype (Long and Langley 1999). Cette étude montre aussi qu'un gain de puissance supérieur est obtenu en augmentant la taille de l'échantillon plutôt qu'en augmentant le nombre de polymorphismes testés.

Pour cela, il est nécessaire que les 340 accessions de départ soient génotypées avec un grand nombre de marqueurs afin d'identifier la structure la plus vraisemblable de l'échantillon et qu'elles soient phénotypées pour les caractères étudiés de façon récurrente au laboratoire (poids du fruit, acidité titrable, teneur en solide soluble, colorimétrie, fermeté, etc.). Une fois que ces informations seront disponibles, il sera plus efficace de tester rapidement l'association entre un polymorphisme et un caractère. Par la suite, ces 340 accessions pourront être génotypées pour un grand nombre de SNP grâce aux méthodes haut-débit (ex : génotypage de 1536 SNP sur un millier d'échantillon par technologie Golden Gate, Illumina) afin de détecter directement des associations. Cet échantillon pourra être complété avec d'autres accessions, présentes dans la collection de ressources génétiques de l'unité GAFL, ainsi qu'avec des accessions provenant d'autres collections. Une collaboration a notamment été engagée avec le laboratoire de Dani Zamir, qui nous a fait parvenir plus de 300 accessions de type cerise, déjà caractérisées en Israël.

7.2.2. Structure du déséquilibre de liaison.

Les analyses préliminaires sur la structuration du DL chez la tomate ont été focalisées uniquement sur le chromosome 2. Le patron d'étendue du DL n'est pas homogène le long du chromosome. Nous avons identifié une zone ou même des polymorphismes distants montrent des corrélations entre eux relativement fortes. Cette zone est adjacente à une région où le DL est très faible, dans la région mais aussi avec la région précédente. Ce patron de DL peut être du à la sélection de certaines régions génomiques qui seraient brusquement contrebalancée par de la recombinaison. En effet, de nombreuses associations, pour le poids du fruit et la teneur en solides solubles, sont retrouvées dans la région présentant un fort DL. Le principal paramètre qui semble influer sur la distribution du DL est la différence de fréquences alléliques entre les marqueurs. La plupart du temps, lorsqu'on s'intéresse à deux polymorphismes, seulement trois haplotypes sont retrouvés, sur les quatre attendus en cas de recombinaison. Cette observation peut être due à des effets de « lineage » des mutations qui s'accumulent sans qu'il y ait recombinaison pour ré-équilibrer les allèles entre les locus (Figure 7-1). Elle peut aussi être expliquée par un effectif trop faible d'accessions qui empêche de retrouver le quatrième haplotype.

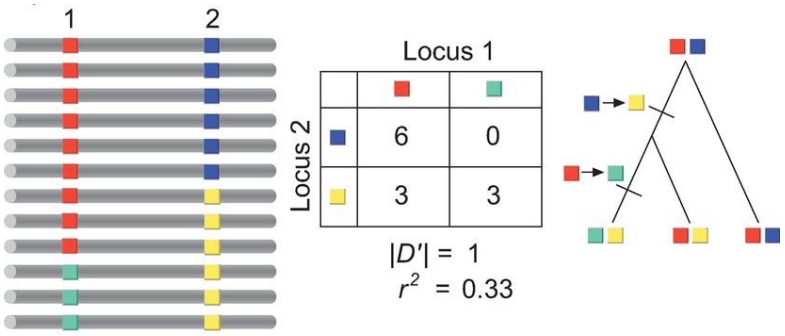


Figure 7-1. Scenario expliquant le DL entre deux locus polymorphes liés sans recombinaison. Modifié à partir de Flint-Garcia, Thornsberry et al. (2003)
Le DL est dû à deux évènements de mutation successifs ayant eu lieu sur deux branches différentes sans recombinaison entre les locus. Le r^2 et *D'* sont ici très différents

7.2.3. Inférence à partir d'une étude focalisée sur le chromosome 2.

Nous avons vu que l'étude du chromosome 2 permettait d'inférer un nombre d'environs 50 000 polymorphismes nécessaires pour réaliser une étude d'association avec une résolution relativement fine sur l'échantillon de 90 accessions. Cette information n'est donnée qu'à titre indicatif car il est nécessaire d'étudier l'étendue du DL sur d'autres régions chromosomiques. En effet, il a été reporté des différences dans l'étendue du DL sur deux

chromosomes différents chez le blé, une autre espèce fortement autogame (étendue du DL inférieure à 1 cM sur le chromosome 2D et jusqu'à 5cM sur le chromosome 5A) (Breseghello and Sorrells 2006). De plus le chromosome 2 porte de nombreux QTL, liés à la qualité du fruit, potentiellement sélectionnés lors de la domestication. Il est donc tout à fait possible que ce chromosome reflète une étendue du DL différente des autres chromosomes. Cette hypothèse pourra notamment être vérifiée en utilisant l'information de génotypage des quatre panels SNPlex sur les 340 accessions.

7.2.4. Validation des associations identifiées sur 90 accessions.

Les associations les plus fortes, identifiées sur la collection de 90 accessions, ont pu être validées sur 201 accessions. Pratiquement toutes les associations testées sont retrouvées (sauf celle entre le fragment TD047 et la teneur en glucose), ce qui révèle bien que la « core collection » a été échantillonnée avec succès. Ici, l'échantillon utilisé implique juste une augmentation du nombre d'accessions testées car il est majoritairement composé d'accessions. *S. l. cerasiforme* (144 accessions de type cerise sur 201 accessions au total). Des échantillons indépendants peuvent être utilisés mais il est possible que les différences de fréquences alléliques ne permettent pas de retrouver les associations. Par exemple, en travaillant uniquement sur les accessions de type cerise, on perd l'association entre le TD387 et le poids du fruit ainsi que toutes les associations relatives à la teneur en solides solubles.

De plus, les polymorphismes associés ne sont pas forcément les polymorphismes causaux expliquant la variation des phénotypes. Il est donc important de se concentrer sur ces régions en densifiant en marqueurs la région autour de ces gènes. La définition de nouveaux marqueurs pourra se concentrer sur une région de 50 Kb autour du polymorphisme associé. Une autre possibilité serait d'analyser uniquement le polymorphisme des gènes candidats dans la région d'intérêt. Cette stratégie pose cependant un problème. Cardon et Bell (2001) précisent qu'une source d'erreur dans les études d'association chez l'homme, est la déclaration qu'un gène est 'candidat' d'après sa fonction, lorsqu'on retrouve une association dans une région génétique arbitraire. L'exemple de *lcn2.1* avec *Wuschel* montre bien le danger d'utiliser une telle approche en génétique.

Les gènes potentiellement candidats, peuvent également être étudiés en modifiant leurs expressions. Ceci peut être réalisé en créant des lignées RNAi (RNA interference) ou en créant des lignées de surexpression. Il est aussi possible d'utiliser des approches de TILLING

pour identifier des mutations au niveau des gènes candidats liées à une modification du phénotype.

La seule façon de valider définitivement une région génomique est la construction de lignées quasi-isogéniques différentes uniquement pour la région en question et pour le phénotype d'intérêt.

7.2.5. Les limites de l'étude.

Cette étude d'association est la première réalisée chez la tomate (hormis l'étude de Mazzucato, Papa et al. (2008) qui s'intéresse à un panel de 61 accessions génotypées uniquement avec 20 SSR). C'est pour cela que nous nous sommes intéressés à un échantillon relativement limité mais représentant la diversité d'une collection beaucoup plus grande. La taille de cet échantillon ne permet pas de détecter des polymorphismes à effets faibles. Cet échantillon présente par conséquent une faiblesse statistique pour l'estimation des paramètres associés à chaque polymorphisme associé avec un phénotype : effet allélique, variance génétique, etc.

L'échantillonnage est aussi limitant car nous nous sommes focalisés sur les accessions de type cerise (*S. l. cerasiforme*). Cet échantillon semble être efficace pour identifier des locus sélectionnés pendant la domestication. Par contre, aucune information ne sera apportée par la « core collection » sur les locus sélectionné par la suite dans l'amélioration moderne, le nombre d'accessions cultivées modernes étant trop faible. L'utilisation des nouvelles technologies de re-séquençage (Next Generation Sequencing ou NGS) permettra de capturer le faible taux de polymorphisme moléculaire présent dans ce groupe. La structuration devra être prise en compte dès l'échantillonnage, en fonction de l'importance des gènes ciblés, dans l'histoire de la sélection (domestication *vs.* sélection pour le marché de frais *vs.* sélection pour l'industrie). Cette information n'est presque jamais connue à l'avance.

Le génotypage s'est fait directement par séquençage avec la technologie Sanger ce qui n'est pas envisageable sur un échantillon de taille plus grande. Ces fragments étaient de petite taille (300-700 bp) à cause de la technique utilisée. Il a donc fallu se concentrer sur les régions non codantes afin d'augmenter la probabilité d'identifier du polymorphisme chez une espèce très peu diversifiée génétiquement. Dans les régions codantes séquencées, peu de polymorphismes ont pu être identifiés. Il y a donc peu de chance que les polymorphismes

identifiés par association soient directement responsables de la variation du phénotype. Ces polymorphismes sont donc en DL avec les polymorphismes causaux. Il va être nécessaire de caractériser le déséquilibre de liaison avec les polymorphismes adjacents afin de rechercher les mutations causales. Des fragments plus longs vont devoir être étudiés autour des régions d'intérêt.

7.2.6. L'héritabilité disparue.

Malgré tous les résultats concluants obtenus en génétique humaine par les approches de génétique d'association et d'analyse de liaison, les effets cumulés des locus identifiés n'expliquent qu'une part infime de la variation du caractère, observée dans les populations (Maher 2008). Cette « héritabilité disparue » peut avoir plusieurs causes. Tout d'abord, une modification de structure du génome, qui n'est pas toujours prise en compte dans la recherche de polymorphismes causaux, peut expliquer une part de la variation. Par exemple, des variations du nombre de copies de certaines régions (Copy Number Variant ou CNV) expliquent le phénotype (Isaksson, Stenberg et al. 2007). Pour l'instant très peu d'études prennent en compte ce type de variation chez les plantes. Les technologies NGS permettront d'avoir accès rapidement à ce type de polymorphisme.

D'autre part, les gènes fonctionnent souvent en interaction avec d'autres partenaires et il est possible que l'effet d'un de ces gènes ne puisse être identifié sans connaître les effets des autres. La prise en compte des phénomènes épistatiques est donc nécessaire si on veut pouvoir expliquer la totalité de la variation d'un caractère.

Enfin, la modification d'un phénotype peut être due à des variations, non pas de la séquence en nucléotides d'une région, mais de la nature chimique des bases. Ainsi des modifications épigénétiques peuvent être transmises aux générations futures sans qu'il y ait eu mutation. De plus en plus d'exemples de phénotypes liés à des variations épigénétiques sont identifiés chez les plantes (Manning, Tor et al. 2006; Martin, Troadec et al. 2009). Toutes ces modifications ne sont pas prises en compte lors d'études d'association où seules les modifications de séquences nucléotidiques sont observées. Il est donc important de garder à l'esprit l'existence de tels phénomènes lors de la dissection du déterminisme génétique de caractères d'intérêt.

7.3. Analyse de la diversité du chromosome 2 – histoire évolutive.

L'étude de *lcn2.1* par rapport à d'autres séquences réparties sur le chromosome 2 a permis d'identifier un écart significatif à la neutralité mais surtout un écart au patron d'évolution du chromosome entier. Cet écart implique un D de Tajima significativement positif qui est cohérent avec l'hypothèse de sélection par l'homme, d'une forte diversité morphologique de l'organe consommé décrite par Frary and Doganlar (2003). Cette diversité morphologique est notamment retrouvée au niveau des épis de maïs des cultivars anciens, des fruits de piment et d'aubergine.

Le modèle d'évolution de la tomate établi au chapitre 6 ne concerne que la domestication de cette plante. Il implique un goulet d'étranglement relativement ancien, une croissance de la population importante mais surtout des flux de gènes entre les compartiments sauvages et domestiqués qui expliquent le taux de diversité retrouvé dans ce dernier groupe. Les fragments séquencés ont permis l'estimation des paramètres du modèle. Les quatre fragments n'ont pas été testés car leur taille était trop faible (Yamasaki, Schroeder et al. 2008). Il serait maintenant intéressant de séquencer complètement un gène impliqué dans la domestication de la tomate comme *fw2.2*, afin de voir si le patron de diversité qu'il présente s'éloigne significativement du modèle neutre. Cela confirmerait le modèle.

Un défaut, lors de la construction du modèle, est l'utilisation de deux populations avec tous les individus affectés à l'une ou à l'autre de ces populations. Seul les individus présentant au moins 80% de probabilité d'appartenance à une population (sortie Structure) auraient du être utilisés, pour éviter le bruit de fond apporté par l'« admixture ».

Le modèle établi doit être complété par le passage de l'espèce domestiquée vers l'espèce cultivée. Ce passage implique un deuxième goulet d'étranglement. Ceci permettra de comparer des gènes sélectionnés lors de la domestication mais aussi les gènes sélectionnés lors de l'amélioration moderne. Un point faible de la tomate cultivée pour ce type d'approche est le faible taux de polymorphisme général qui ne permettra pas de faire la distinction entre effet démographique et sélection positive. Seul des mutations en fréquence balancée s'éloigneront significativement du modèle neutre.

7.4. Perspectives

7.4.1. Construction de nouvelles populations de cartographie à partir d'accessions maximisant la diversité.

Un moyen de combiner les apports de la cartographie génétique et de la génétique d'association est la création de populations de ségrégation, basées sur des croisements entre plusieurs parents qui représentent un maximum de la diversité génétique. L'échantillonnage de « core collections » peut alors servir à constituer de nouvelles populations en ségrégation qui maximisent la diversité identifiée chez une espèce. La « core collection » constituée chez *A. thaliana* a été utilisée afin de construire des populations de RIL (Recombinant Inbred Lines) à partir de plusieurs parents qui maximisent la diversité génétique identifiée sur une collection de 300 accessions (Simon, Loudet et al. 2008). Cette étude a été réalisée en combinant les apports de la cartographie de QTL qui utilise des populations synthétiques et ceux de la génétique d'association qui se focalise sur la diversité naturelle présente dans les ressources génétiques.

Cette stratégie a été adaptée au maïs où l'accession de référence a été croisée avec différents individus représentant la diversité du maïs (Yu, Holland et al. 2008). Les descendants sont ensuite autofécondés jusqu'à obtention d'une une seule population de plusieurs milliers de lignées d'introgression (Figure 7-2). Ces populations NAM (Nested Association Mapping) présentent une forte puissance statistique car le phénotypage des RIL peut être répété facilement, la population totale présente une grande richesse allélique, une bonne résolution de cartographie et une faible sensibilité à l'hétérogénéité génétique par rapport à des populations naturelles (Yu, Holland et al. 2008). Cette population a pu être testée pour détecter des QTL lié à la précocité de floraison du maïs ce qui implique le phénotypage de près d'un million de plantes (Buckler, Holland et al. 2009).

La théorie développée pour les AIL (Advanced Intercrossed Line) a été adaptée à des populations multi-parentales (Darvasi and Soller 1995; Cavanagh, Morell et al. 2008). Ces populations, appelées populations MAGIC pour Multi-parent Advanced Generation Inter-Cross (Figure 7-3), s'appuient à la fois sur des méthodologies de cartographie génétiques et de génétique d'association sans les difficultés associées aux populations fortement structurées (Valdar, Flint et al. 2006). Ces populations permettent de capturer une part plus importante de la variation génétique naturelle et les ressources développées peuvent être utilisées

directement pour réaliser la cartographie fine de locus d'intérêt. Le tableau 7-1 présente une comparaison des stratégies de cartographie génétique, d'association et d'utilisation de population MAGIC. Une population MAGIC est en cours de construction au laboratoire et les huit parents constituant les fondateurs ont été choisis parmi la core collection de 96 accessions. Cette population sera phénotypée pour les caractères de qualité du fruit, et pourra être utilisée comme ressource pour pouvoir identifier des locus liés à d'autres caractères d'intérêt chez la tomate. Les technologies de re-séquençage vont être utilisées pour détecter une grande quantité de polymorphismes entre les lignées parentales en se focalisant sur du séquençage de cDNA.

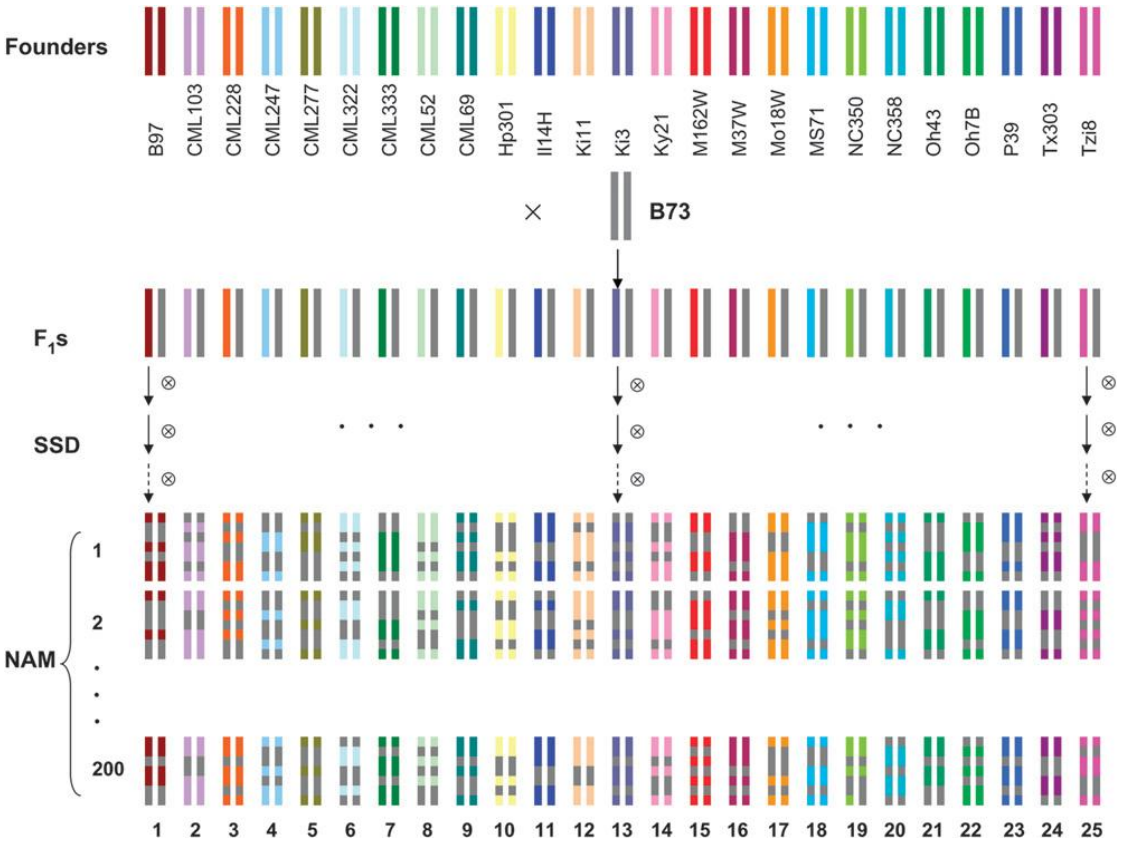


Figure 7-2. Schéma de construction d'une population NAM (Nested Association Mapping) issue d'un croisement entre un parent de référence (B73) et 25 accessions diversifiées (fondateurs). D'après Yu, Holland et al. (2008).
A cause de la diminution du taux de recombinaison lorsque la distance génétique diminue avec un nombre limité de génération, les génomes de ces RIL correspondent à des mosaïques des génomes fondateurs. Les différentes F1 sont autofécondées afin de fixer les recombinaisons par Single Seed Descend (SSD). La population finale est composée de 25 populations, correspondant aux 25 fondateurs, contenant chacune 200 lignées recombinantes.

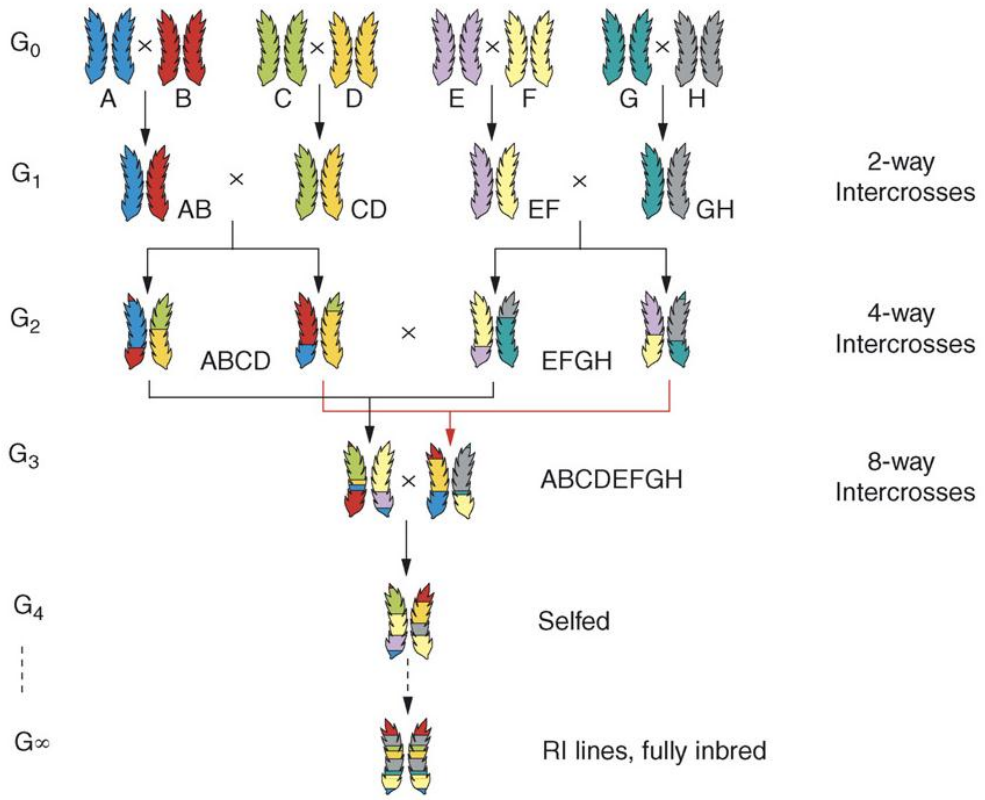


Figure 7-3. Dispositif expérimental utilisé dans la création d'une population MAGIC (Multi-parent Advanced Generation Inter-Cross). D'après Cavanagh, Morell et al. (2008). Les *n* fondateurs sont inter-croisés *n*/2 fois, jusqu'à ce que chacun d'entre eux soit combiné dans des proportions égales dans la descendance. Ensuite, les RIL peuvent être produites directement ou après plusieurs évènements d'inter-croisements.

Application	Liaison	Association	MAGIC
Cartographie large	+	-	+
Cartographie fine	-	+	+
Faible besoin en génotypage	+	-	-
Faible besoin en phénotypage	+	-	-
Résistant à la structure des populations	+	-	+
Pertinence pour les sélectionneurs	-	+	+
Pertinence dans le temps	-	+	+
Temps de mise en place de l'expérimentation	-	+	-

Tableau 7-1. Avantages et inconvénients des trois méthodes utilisées dans l'identification de QTL chez les végétaux : analyse de liaison bi-parentale (liaison), génétique d'association (association) et Multi-parent Advanced Generation Inter-Cross (MAGIC). D'après Cavanagh, Morell et al. (2008).

7.4.2. La recherche de traces de sélection : une autre approche pour identifier des gènes candidats.

Toutes les méthodes vues jusqu'à présent vont dans le sens de la variation phénotypique vers l'identification du polymorphisme moléculaire. Un certain nombre d'études présentent un schéma d'analyse inverse et contribuent elles aussi à valoriser les ressources génétiques. Ces études s'intéressent à la détection de traces de sélection dans des gènes candidats en vue d'identifier des cibles potentielles de la domestication ou de la sélection des plantes cultivées. Ces gènes concernent potentiellement des caractères agronomiques d'intérêt. La figure 7-4 montre la comparaison entre les approches dites de « Top-Down » et les approches dites de « Bottom-up ».

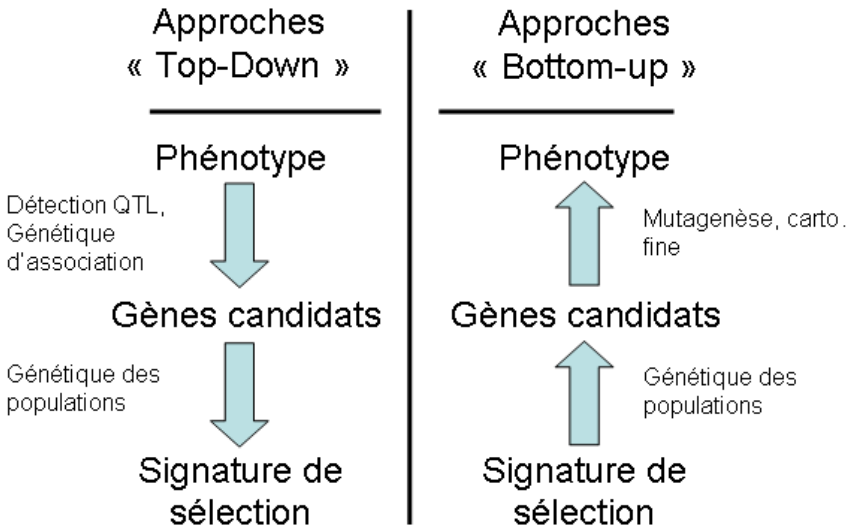


Figure 7-4. Schéma de la hiérarchie phénotype-génotype représentée par les approches dites « Top-Down » et « Bottom-up ». D'après Ross-Ibarra, Morrell et al. (2007)

Dans les approches de type « Top-Down », l'identification des polymorphismes candidats se fait sur la base de la variation d'un caractère phénotypique. Une fois la région causale identifiée, on réalise une étude de diversité de la région afin d'essayer d'inférer son histoire évolutive. C'est notamment ce qui a été fait sur le locus *lcn2.1* au cours de cette thèse. La sélection balancée qu'a subie ce locus au cours de l'évolution est cohérente avec l'hypothèse d'une sélection par l'homme, de la diversité morphologique des organes consommés.

Si plusieurs candidats restent présents dans la région isolée après cartographie fine, il est possible d'étudier les traces de sélection de ces gènes. Etant donné que la plupart des QTL

clonés chez les plantes cultivées sont responsables de la variation de caractères agronomiques, il est fort probable qu'ils aient été soumis à sélection durant la domestication et qu'ils continuent à porter les traces de cette sélection. Dans le cas de régions régulatrices, un nouveau modèle permet maintenant d'analyser les régions de l'ADN où se fixent des protéines de régulation (Moses 2009). Ce test prend en compte la modification de l'affinité d'une séquence d'ADN à une protéine s'il y a substitution.

L'approche de type « Bottom-Up » propose d'identifier des gènes en fonction de la signature de sélection qu'ils portent, puis une batterie d'outils génétiques est mise en œuvre afin d'identifier les phénotypes auxquels ces gènes contribuent. Cette approche nécessite de bien connaître l'histoire évolutive de l'espèce considérée car il est nécessaire de différencier les patrons de diversité dus à l'histoire démographique de la population des patrons de diversité induits par une sélection réelle. C'est pour cette raison que des modèles démographiques sont comparés afin de choisir le plus vraisemblable par rapport aux données réelles, récoltées sur les populations étudiées. Ces modèles servent ensuite à générer des distributions pour plusieurs statistiques de diversité pour des séquences simulées qui répondent uniquement à l'évolution démographique. Si les statistiques identifiées sur un gène candidat sortent significativement de cette distribution, alors il y a de forte chance pour qu'il ait subi une pression évolutive supplémentaire. Ainsi chez *Medicago truncatula*, le canal d'ion *DM11* a été identifié comme le seul gène soumis à sélection dans la voie de transduction du signal permettant la nodulation (De Mita, Ronfort et al. 2007). Chez le maïs, les modèles évolutifs sont bien établis grâce à de nombreuses études génétiques telles que celles réalisées par Eyre-Walker, Gaut et al. (1998) et Tenaillon, U'Ren et al. (2004). Ces études ont permis d'établir des schémas de réduction de la diversité différents pour la domestication et la sélection moderne (Figure 7-5).

Des gènes montrant des traces de sélection uniquement chez les cultivars anciens ont été sélectionnés pendant la domestication et les gènes montrant des traces de sélection chez les cultivars anciens et chez les variétés modernes ont été sélectionnés pendant la sélection moderne.

Une analyse de la diversité de plus d'un millier de gènes chez le maïs a été conduite et 35 gènes candidats ont été identifiés car ils montraient des traces de sélection dans le groupe des cultivars anciens ou dans le groupe des variétés modernes (Yamasaki, Tenaillon et al. 2005). Ces gènes correspondent à des facteurs de méthylation de l'ADN, des facteurs de transcription, des facteurs contrôlant l'horloge circadienne, des facteurs de transduction de

signal, et des gènes impliqués dans le métabolisme des acides aminés. Cette étude présentait le défaut de travailler sur des alignements de séquences trop courtes (< 300 bp). Les gènes identifiés ont été re-séquencés et ré-analysés en utilisant la coalescence pour confirmer ou non leurs positions de candidats (Yamasaki, Schroeder et al. 2008).

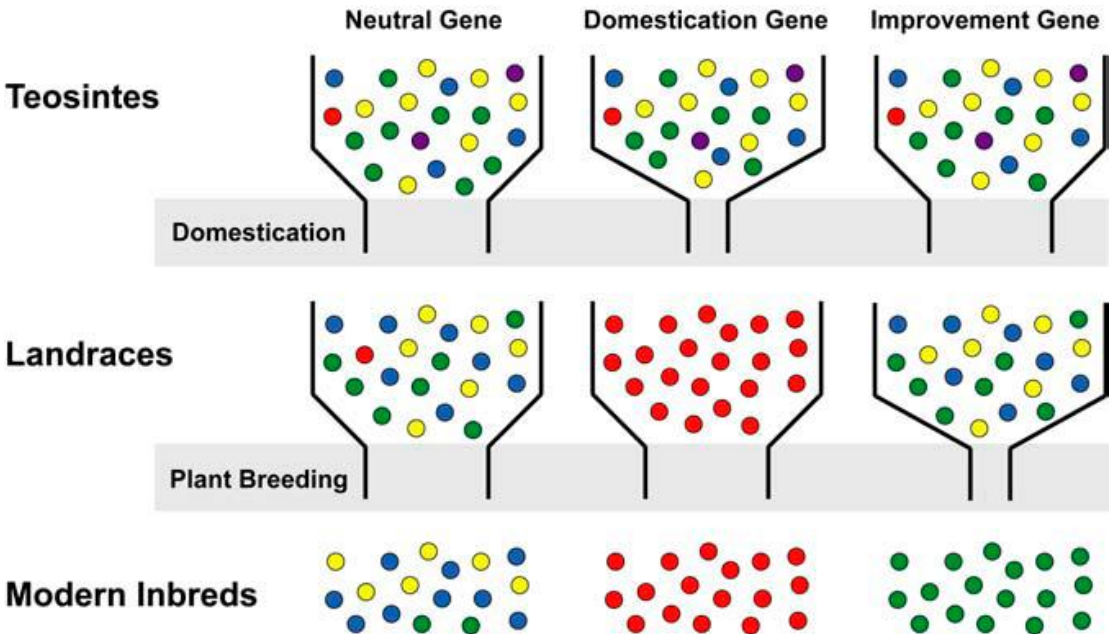


Figure 7-5. Effet de la domestication et de la sélection moderne (Breeding) sur la diversité des gènes chez le maïs. D'après Yamasaki, Tenaillon et al. (2005).
Les cercles colorés représentent les différents allèles. Les zones grisées indiquent l'effet des goulets d'étranglement de la domestication et de l'amélioration variétale (Plant Breeding) sur des gènes neutres (neutral gene), les gènes sélectionnés durant la domestication ou l'amélioration moderne. Les effets sont considérés sur l'espèce sauvage (Teosintes), les cultivars anciens (Landraces) ainsi que les variétés modernes (Modern Inbred).

7.4.3. Les techniques de re-séquençage nouvelle génération (NGS) représentent un nouvel essor pour la génétique d'association.

Ces analyses deviennent de plus en plus intéressantes avec l'avancée des technologies de re-séquençage. Avec la technologie Solexa®, par exemple, il sera bientôt possible de re-séquencer un génome entier d'une accession de tomate en moins d'une semaine et avec une précision importante (profondeur de 90x). De nouveaux protocoles de détection de SNP utilisant ces technologies sont publiés régulièrement, ce qui montre bien la concentration des efforts autour de ces méthodes. Par exemple, une seule expérience permet de révéler plus de 125 000 SNP entre deux lignées de maïs, en utilisant des alignements de séquence de 100 à 200 bp (Gore, Wright et al. 2009). Le seul inconvénient qui est imputé à ces nouvelles

technologies est le besoin bio-informatique nécessaire pour traiter les données (Pop and Salzberg 2008; McPherson 2009). Un exemple remarquable d'utilisation de ces technologies pour identifier des gènes candidats est le travail réalisé par Xia, Guo et al. (2009). Cette étude porte sur le re-séquençage de 40 génomes de vers à soie (*Bombyx*). Cette espèce présente un génome mesurant la moitié de celui de la tomate (432 Mb). En comparant la diversité trouvée dans une population sauvage et une population domestiquée, les auteurs définissent 1041 régions présentant un signal de sélection. Dans ces régions, ils identifient 354 gènes codant pour des protéines ayant un rôle potentiel dans l'amélioration de la quantité et de la qualité de la soie produite.

La diminution du coût de séquençage (projet de séquençage de génomes humains pour 1000$) ainsi que l'amélioration des technologies (longueur de fragments séquencés) et des méthodes bioinformatiques de traitement des données, rendent une telle approche envisageable pour des espèces avec des génomes plus importants et notamment des espèces végétales comme la tomate.

Ces approches vont être utilisées pour identifier du polymorphisme. Deux applications sont envisageables :

- une détection par séquençage sur peu d'individus et un génotypage par technologie haut-débit.
- directement en re-séquençant tous les individus avec des techniques d'étiquetage des individus.

Pour que la dernière méthode soit envisageable sur de gros échantillon il est nécessaire que le prix de l'étiquetage (qui permet de retrouver tous les individus) diminue. Il est aussi possible de réaliser des simplifications du génome par capture de séquences d'intérêt, par séquençage de banques d'ADNc normalisées ou par séquençage de fragments digérés et amplifiés.

L'arrivée de séquenceurs de troisième génération qui permettent un séquençage de brins beaucoup plus longs, en temps réel (Eid, Fehr et al. 2009), ponctuera sans aucun doute la révolution que connaît la génétique moderne. Il est fort probable que dans les années qui viennent plus aucun génome ne soit orphelin (sans séquence génomique de référence).

Certains envisagent même de génotyper des accessions, et non plus juste identifier du polymorphisme, directement par re-séquençage du génome de celles-ci.

Il y a 20 ans, les marqueurs moléculaires ont fait dialoguer la génétique moléculaire avec la génétique quantitative. Aujourd'hui, avec le génotypage haut-débit et la technologie de re-séquençage, ce sont la génomique évolutive et génétique d'association qui sont concernées par ces révolutions. Il est important que des domaines tels que la génétique évolutive, qui est restée pendant longtemps l'apanage de la recherche fondamentale, se voit appropriées par l'amélioration génétique moderne. De nouvelles méthodologies doivent être mises en place pour permettre une seconde révolution verte nécessaire pour répondre aux nouveaux enjeux agricoles : sécurité alimentaire, changement global et durabilité. Il est aussi important pour les générations futures, que nous maintenions de façon optimale la diversité génétique des espèces cultivées. La sensibilisation autour de la conservation de la biodiversité ne pourra se faire qu'avec une connaissance approfondie de celle –ci.

La valorisation de la diversité contenue dans les ressources génétiques de plantes cultivées pourrait passer par de nouveaux évènements de domestication qui permettraient de réintégrer de la variabilité génétique dans le compartiment cultivé moderne afin de préparer plus efficacement les réponses aux changements globaux potentiels.

Références bibliographiques

Abdurakhmonov, I. Y. and A. Abdukarimov (2008). "Application of association mapping to understanding the genetic diversity of plant germplasm ressources." Int. J. Plant. Geno. **2008**: 18 pages.

Adams, M. D. and J. J. Sekelsky (2002). "From sequence to phenotype: reverse genetics in *Drosophila melanogaster*." Nat Rev Genet **3**(3): 189-198.

Aguayo, E., V. Escalona, et al. (2004). "Quality of fresh-cut tomato as affected by type of cut, packaging, temperature and storage time." European Food Research and Technology **219**(5): 492-499.

Alvarez, A. E., C. C. M. v. d. Wiel, et al. (2001). "Use of microsatellites to evaluate genetic diversity and species relationships in the genus *Lycopersicon*." TAG Theoretical and Applied Genetics **103**(8): 1283-1292.

Aranzana, M. i. a. J. e., S. Kim, et al. (2005). "Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes." PLoS Genetics **1**(5): e60.

Ardlie, K. G., L. Kruglyak, et al. (2002). "Patterns of linkage disequilibrium in the human genome." **3**(4): 299-309.

Areshchenkova, T. and M. W. Ganal (1999). "Long tomato microsatellites are predominantly associated with centromeric regions." Genome **42**: 536-544.

Avery, O. T., C. M. McLeod, et al. (1944). "Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* Type III." J. Exp. Med. **79**: 137-158.

Bai, Y. and P. Lindhout (2007). "Domestication and breeding of tomatoes: What have we gained and what can we gain in the future?" Ann Bot: mcm150.

Baldwin, E. A., K. Goodner, et al. (2004). "Effect of volatiles and their concentration on perception of tomato descriptors." Journal of food science **69**(8): 310-318.

Baldwin, E. A., M. O. Nisperos-Carriedo, et al. (2002). "Quantitative analysis of flavor parameters in six Florida tomato cultivars (*Lycopersicon esculentum* Mill)." Journal of Agricultural and Food Chemistry **39**(6): 1135-1140.

Bandelt, H.-J. and A. W. M. Dress (1992). "Split decomposition: A new and useful approach to phylogenetic analysis of distance data." Molecular Phylogenetics and Evolution **1**(3): 242-252.

Bao, J., H. Corke, et al. (2006). "Microsatellites, single nucleotide polymorphisms and a sequence tagged site in starch-synthesizing genes in relation to starch physicochemical properties in nonwaxy rice (*Oryza sativa* L.)." TAG Theoretical and Applied Genetics **113**(7): 1185-1196.

Barnaud, A., T. Lacombe, et al. (2006). Linkage disequilibrium in cultivated grapevine, *Vitis vinifera* L. **112:** 708 - 716.

Barrero, L. S., B. Cong, et al. (2006). "Developmental characterization of the *fasciated* locus and mapping of *Arabidopsis* candidate genes involved in the control of floral meristem size and carpel number in tomato." Genome **49**(8): 991-1006.

Barrero, L. S. and S. D. Tanksley (2004). "Evaluating the genetic basis of multiple-locule fruit in a broad cross section of tomato cultivars." TAG Theoretical and Applied Genetics **109**(3): 669-679.

Barry, C. and J. Giovannoni (2007). "Ethylene and fruit ripening." Journal of Plant Growth Regulation **26**(2): 143-159.

Barry, C. S. and J. J. Giovannoni (2006). "Ripening in the tomato *Green-ripe* mutant is inhibited by ectopic expression of a protein that disrupts ethylene signaling." **103**(20): 7923-7928.

Bataillon, T. M., J. L. David, et al. (1996). "Neutral genetic markers and conservation genetics: Simulated germplasm collections." Genetics **144**(1): 409-417.

Belkir , K., P. Borsa , et al. (2004). GENETIX, logiciel sous WindowsTM pour la génétique des populations. Laboratoire Génome, Populations, Interactions CNRS UMR 5000, Montpellier.

Belle, E. M. S., U. Ramakrishnan, et al. (2006). "Serial coalescent simulations suggest a weak genealogical relationship between Etruscans and modern Tuscans." **103**(21): 8012-8017.

Benjamini, Y. and Y. Hochberg (2000). "On the adaptive control of the false discovery rate in multiple testing with independent statistics." J. Edu. Behav. Stat. **25**(1): 60-83.

Bermudez, L., U. Urias, et al. (2008). "A candidate gene survey of quantitative trait loci affecting chemical composition in tomato fruit." J. Exp. Bot.: ern146.

Bernacchi, D., T. Beck-Bunn, et al. (1998). "Advanced backcross QTL analysis in tomato. I. Identification of QTLs for traits of agronomic importance from *Lycopersicon hirsutum*." TAG Theoretical and Applied Genetics **V97**(3): 381-397.

Bohs, L. and R. G. Olmstead (1997). "Phylogenetic relationships in *Solanum* (*Solanaceae*) based on ndhF sequences." Systematic Botany **22**(1): 5-17.

Bradbury, P. J., Z. Zhang, et al. (2007). "TASSEL: software for association mapping of complex traits in diverse samples." Bioinformatics **23**(19): 2633-2635.

Brand, U., J. C. Fletcher, et al. (2000). "Dependence of stem cell fate in *Arabidopsis* on a feedback loop regulated by *CLV3* activity." Science **289**(5479): 617-619.

Bredemeijer, G., R. Cooke, et al. (2002). "Construction and testing of a microsatellite database containing more than 500 tomato varieties." TAG Theoretical and Applied Genetics **105**(6): 1019-1026.

Bres, C., J. P. Bouchet, et al. (2005). CGIS: an information system used for designing primers of candidate genes using *Arabidopsis* whole genome and *Solanaceae* EST databases. 16. Triennal Conference of the EAPR. Bilbao (ESP).

Breseghello, F. and M. E. Sorrells (2006). "Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars." Genetics **172**(2): 1165-1177.

Brücher, H. (1989). Useful plants of neotropical origin and their wild relatives. Berlin, Springer-Verlag.

Brummell, D. A. and M. H. Harpster (2001). "Cell wall metabolism in fruit softening and quality and its manipulation in transgenic plants." Plant Molecular Biology **47**(1): 311-339.

Buckler, E. S., J. B. Holland, et al. (2009). "The genetic architecture of maize flowering time." Science **325**(5941): 714-718.

Buckler, I., Edward S. and J. M. Thornsberry (2002). "Plant molecular diversity and applications to genomics." Curr. Opin. Plant Bio. **5**(2): 107-111.

Budiman, M. A., L. Mao, et al. (2000). "A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing." Genome Research **10**(1): 129-136.

Butler, L. (1952). "The linkage map of the tomato." Jour. Heredity **43**: 25-36.

Caldwell, K. S., J. Russell, et al. (2006). "Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*." Genetics **172**(1): 557-567.

Camus-Kulandaivelu, L., L.-M. Chevin, et al. (2008). "Patterns of molecular evolution associated with two selective sweeps in the *Tb1-Dwarf8* region in maize. 10.1534/genetics.108.088849." Genetics **180**(2): 1107-1121.

Camus-Kulandaivelu, L., J.-B. Veyrieras, et al. (2006). "Maize adaptation to temperate climate: Relationship between population structure and polymorphism in the *Dwarf8* Gene." Genetics **172**(4): 2449-2463.

Cardon, L. R. and J. I. Bell (2001). "Association study designs for complex diseases." Nat Rev Genet **2**(2): 91-99.

Carreiro, F., A. Petrozza, et al. (2004). Tilling in tomato: production of EMS mutant collections. Proceedings of the XLVIII Italian Society of Agricultural Genetics. Lecce, Italy.

Carroll, S. B. (2000). "Endless forms: The evolution of gene regulation and morphological diversity." Cell **101**(6): 577-580.

Casa, A. M., G. Pressoir, et al. (2008). "Community resources and strategies for association mapping in *Sorghum*." Crop Sci **48**(1): 30-40.

Causse, M., M. Buret, et al. (2003). "Inheritance of nutritional and sensory quality traits in fresh market tomato and relation to consumer preferences." Journal of Food Science **68**(7): 2342-2350.

Causse, M., P. Duffe, et al. (2004). "A genetic map of candidate genes and QTLs involved in tomato fruit size and composition." Journal of Experimental Botany **55**(403): 1671-1685.

Causse, M., C. Friguet, et al. (*in prep.*). "Consumer preferences for tomato cultivars: a european comparison."

Causse, M., V. Saliba-Colombani, et al. (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**(377): 2089-2098.

Causse, M., V. Saliba-Colombani, et al. (2001). "Genetic analysis of organoleptic quality in fresh market tomato. 2. Mapping QTLs for sensory attributes." TAG Theoretical and Applied Genetics **V102**(2): 273-283.

Cavanagh, C., M. Morell, et al. (2008). "From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants." Current Opinion in Plant Biology **11**(2): 215-221.

Chaib, J., M.-F. Devaux, et al. (2007). "Physiological relationships among physical, sensory, and morphological attributes of texture in tomato fruits." J. Exp. Bot. **58**(8): 1915-1925.

Chaïb, J., L. Lecomte, et al. (2006). "Stability over genetic backgrounds, generations and years of quantitative trait locus (QTLs) for organoleptic quality in tomato." TAG Theoretical and Applied Genetics **V112**(5): 934-944.

Chen, F. Q., M. R. Foolad, et al. (1999). "Mapping of QTLs for lycopene and other fruit traits in a *Lycopersicon esculentum* x *L. pimpinellifolium* cross and comparison of QTLs across tomato species." Molecular Breeding **V5**(3): 283-299.

Chen, K.-Y., B. Cong, et al. (2007). "Changes in regulation of a transcription factor lead to autogamy in cultivated tomatoes." Science **318**(5850): 643-645.

Clark, R. M., E. Linton, et al. (2004). "Pattern of diversity in the genomic region near the maize domestication gene *tb1*." Proc. Nat. Acad. Sci. USA **101**(3): 700-707.

Cong, B., L. S. Barrero, et al. (2008). "Regulatory change in *YABBY-like* transcription factor led to evolution of extreme fruit size during tomato domestication." Nature Genetics **40**(6): 800-804.

Cong, B., J. Liu, et al. (2002). "Natural alleles at a tomato fruit size quantitative trait locus differ by heterochronic regulatory mutations." PNAS **99**(21): 13606-13611.

Corander, J., P. Waldmann, et al. (2004). "BAPS 2: enhanced possibilities for the analysis of genetic population structure." Bioinformatics **20**(15): 2363-2369.

Corander, J., P. Waldmann, et al. (2003). "Bayesian analysis of genetic differentiation between populations." Genetics **163**(1): 367-374.

Cox, S. (2000). "I Say Tomayto, You Say Tomahto..."
from http://lamar.colostate.edu/~samcox/Tomato.html.

Crossa, J., J. Burgueno, et al. (2007). "Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure." Genetics: genetics.107.078659.

Curtu, A.-L., R. Finkeldey, et al. (2004). "Comparative sequencing of a microsatellite locus reveals size homoplasy within and between european oak species (*Quercus* spp.)." Plant Molecular Biology Reporter **22**(4): 339-346.

D'Agostino, N., M. Aversano, et al. (2006). "TomatEST database: in silico exploitation of EST data to explore expression patterns in tomato species." Nucl. Acids Res.: gkl921.

D'hoop, B. B., M. João Paulo, et al. (2008). "Association mapping of quality traits in potato (*Solanum tuberosum* L.)." Euphytica **161**(1): 47-60.

Daly, M. J., J. D. Rioux, et al. (2001). "High-resolution haplotype structure in the human genome." Nat. Genet. **29**(2): 229-232.

Dan, Y., Z. Fei, et al. (2007). "MicroTom - A new model plant for genomics." Genes Genomes Genomics **1**(2): 167-179.

Darvasi, A. and S. Shifman (2005). "The beauty of admixture." Nat. Genet. **37**(2): 118-119.

Darvasi, A. and M. Soller (1995). "Advanced intercross lines, an experimental population for fine genetic mapping." Genetics **141**(3): 1199-1207.

Darwin, C. (1859). On the Origin of Species by Means of Natural Selection. London, J. Murray.

Darwin, C. (1868). The Variation of Plants and Animals Under Domestication. London, J. Murray.

Daunay, M.-C., H. Laterrot, et al. (2007). Iconography and history of *Solanaceae*: Antiquity to the 17th Century. Hort. Rev. J. Janick. **34:** 1-119.

David-Schwartz, R., H. Badani, et al. (2001). "Identification of a novel genetically controlled step in mycorrhizal colonization: plant resistance to infection by fungal spores but not extra-radical hyphae." The Plant Journal **27**(6): 561-569.

Davies, J. N. and G. E. Hobson (1981). "The constituents of tomato fruit - the influence of environment, nutrition, and genotype." Crit Rev Food Sci Nutr. **15**(3): 205-280.

de Candolle, A. P. (1882). L'Origine des Plantes Cultivées. Paris.

De Mita, S., J. Ronfort, et al. (2007). "Investigation of the demographic and selective forces shaping the nucleotide diversity of genes involved in *Nod* factor signaling in *Medicago truncatula*." Genetics **177**(4): 2123-2133.

de Vetten, N., F. Quattrocchio, et al. (1997). "The *an11* locus controlling flower pigmentation in petunia encodes a novel WD-repeat protein conserved in yeast, plants, and animals." Genes & Development **11**(11): 1422-1434.

Deu, M. and J. C. Glaszmann (2004). Linkage disequilibrium in *Sorghum*. Plant and Animal Genomes Conference. San Diego, Etats Unis.

Diamond, J. (2002). "Evolution, consequences and future of plant and animal domestication." Nature **418**(6898): 700-707.

Djè, Y., M. Heuertz, et al. (2000). "Assessment of genetic diversity within and among germplasm accessions in cultivated sorghum using microsatellite markers." TAG Theoretical and Applied Genetics **100**(6): 918-925.

Doebley, J., A. Stec, et al. (1997). "The evolution of apical dominance in maize." Nature **386**(6624): 485-488.

Doganlar, S., A. Frary, et al. (2002). "Mapping quantitative trait loci in inbred backcross lines of *Lycopersicon pimpinellifolium* (LA1589)." Genome **45**(6): 1189-1202.

Doré, C. and F. Varoquaux (2006). Histoire et amélioration de cinquante plantes cultivées. Paris.

Drouaud, J. (2006). "Variation in crossing-over rates across chromosome 4 of *Arabidopsis thaliana* reveals the presence of meiotic recombination hot spots." Genome Res. **16**: 106-114.

East, E. M. (1910). "A Mendelian interpretation of variation that is napparently continuous." The American Naturalist **44**: 65-82.

Edwards, M. D., T. Helentjaris, et al. (1992). "Molecular-marker-facilitated investigations of quantitative trait loci in maize." TAG Theoretical and Applied Genetics **83**(6): 765-774.

Ehrenreich, I. M. and M. D. Purugganan (2006). "The molecular genetic basis of plant adaptation." Am J Bot **93**(7): 953 - 962.

Eid, J., A. Fehr, et al. (2009). "Real-time DNA sequencing from single polymerase molecules." Science **323**(5910): 133-138.

El-Din El-Assal, S., C. Alonso-Blanco, et al. (2001). "A QTL for flowering time in *Arabidopsis* reveals a novel allele of *CRY2*." Nat Genet **29**(4): 435-440.

Eshed, Y. and D. Zamir (1995). "An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL." Genetics **141**(3): 1147-1162.

Eshed, Y. and D. Zamir (1996). "Less-than-additive epistatic interactions of quantitative trait loci in tomato." Genetics **143**(4): 1807-1817.

Estoup, A., P. Jarne, et al. (2002). "Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis." Molecular Ecology **11**(9): 1591-1604.

Evanno, G., S. Regnaut, et al. (2005). "Detecting the number of clusters of individuals using the software structure: a simulation study." Mol. Ecol. **14**(8): 2611-2620.

Excoffier, L., J. Novembre, et al. (2000). "Computer note. SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography." J Hered **91**(6): 506-509.

Excoffier, L. and S. Schneider (2005). "Arlequin ver. 3.0: An integrated software package for population genetics data analysis." Evolutionary Bioinformatics Online **1**: 47-50.

Eyal, E. and A. A. Levy (2002). "Tomato mutants as tools for functional genomics." Current Opinion in Plant Biology **5**(2): 112-117.

Eyre-Walker, A., R. L. Gaut, et al. (1998). "Investigation of the bottleneck leading to the domestication of maize." PNAS **95**(8): 4441-4446.

Feldman, R. M. R., C. C. Correll, et al. (1997). "A complex of Cdc4p, Skp1p, and Cdc53p/Cullin catalyzes ubiquitination of the phosphorylated CDK Inhibitor Sic1p." Cell **91**(2): 221-230.

Flint-Garcia, S. A., J. M. Thornsberry, et al. (2003). "Structure of linkage disequilibrium in plants." Annu. Rev. Plant Biol. **54**(1): 357-374.

Flint-Garcia, S. A., A. Thuillet, et al. (2005). "Maize association population: a high-resolution platform for quantitative trait locus dissection." The Plant Journal **44**(6): 1054-1064.

Foissac, S., P. Bardou, et al. (2003). "EUGENE'HOM: a generic similarity-based gene finder using multiple homologous sequences." Nucl. Acids Res. **31**(13): 3742-3745.

Fournier-Level, A., L. Le Cunff, et al. (2009). "Quantitative genetic bases of anthocyanin variation in grape (*Vitis vinifera* L. ssp. *sativa*) berry: A quantitative trait locus to quantitative trait nucleotide integrated study." Genetics **183**(3): 1127-1139.

Frary, A. and S. Doganlar (2003). "Comparative genetics of crop plant domestication and evolution." Turk J Agric For **27**: 59-69.

Frary, A., T. C. Nesbitt, et al. (2000). "*fw2.2*: A quantitative trait locus key to the evolution of tomato fruit size." Science **289**(5476): 85-88.

Frary, A., Y. Xu, et al. (2005). "Development of a set of PCR-based anchor markers encompassing the tomato genome and evaluation of their usefulness for genetics and breeding experiments." Theor Appl Genet **V111**(2): 291-312.

Fray, R. G. and D. Grierson (1993). "Identification and genetic analysis of normal and mutant phytoene synthase genes of tomato by sequencing, complementation and co-suppression." Plant Molecular Biology **22**(4): 589-602.

Freedman, M. L., D. Reich, et al. (2004). "Assessing the impact of population stratification on genetic association studies." **36**(4): 388-393.

Fridman, E., F. Carrari, et al. (2004). "Zooming in on a quantitative trait for tomato yield using interspecific introgressions." Science **305**(5691): 1786-1789.

Fridman, E., T. Pleban, et al. (2000). "A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene." Proc. Nat. Acad. Sci. USA **97**(9): 4718-4723.

Fulton, T. M., T. Beck-Bunn, et al. (1997). "QTL analysis of an advanced backcross of *Lycopersicon peruvianum* to the cultivated tomato and comparisons with QTLs found in other wild species." TAG Theoretical and Applied Genetics **V95**(5): 881-894.

Fulton, T. M., P. Bucheli, et al. (2002). "Quantitative trait loci (QTL) affecting sugars, organic acids and other biochemical properties possibly contributing to flavor, identified in four advanced backcross populations of tomato." Euphytica **127**(2): 163-177.

Fulton, T. M., S. Grandillo, et al. (2000). "Advanced backcross QTL analysis of a *Lycopersicon esculentum* × *Lycopersicon parviflorum* cross." TAG Theoretical and Applied Genetics **100**(7): 1025-1042.

Fulton, T. M., R. Van der Hoeven, et al. (2002). "Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants." Plant Cell **14**(7): 1457-1467.

Galtier, N. and L. Duret (2007). "Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution." Trends in Genetics **23**(6): 273-277.

Gao, H., S. Williamson, et al. (2007). "An MCMC approach for joint inference of population structure and inbreeding rates from multi-locus genotype data." Genetics: genetics.107.072371.

Gao, L.-z. and H. Innan (2008). "Non-independent domestication of the two rice subspecies, *Oryza sativa* ssp. *indica* and ssp. *japonica*, Demonstrated by multilocus microsatellites." Genetics **179**(2): 965-976.

Gaut, B. S. (1998). "Molecular clocks and nucleotide substitution rates in higher plants." Evolutionary Biology **30**: 93-120.

Genlou, S. and S. Björn (2003). "Microsatellite variability and heterozygote deficiency in the arctic-alpine Alaskan wheatgrass (*Elymus alaskanus*) complex." Genome **46**: 729-737.

Giovannoni, J. J. (2004). "Genetic regulation of fruit development and ripening." Plant Cell **16**(suppl_1): S170-180.

Giovannucci, E. (1999). "Tomatoes, tomato-based products, lycopene, and cancer: Review of the epidemiologic literature." J. Natl. Cancer Inst. **91**(4): 317-331.

Goldman, I. L., I. Paran, et al. (1995). "Quantitative trait locus analysis of a recombinant inbred line population derived from a *Lycopersicon esculentum* x *Lycopersicon cheesmanii* cross." TAG Theoretical and Applied Genetics **90**(7): 925-932.

Gonzalo, M. and E. van der Knaap (2008). "A comparative analysis into the genetic bases of morphology in tomato varieties exhibiting elongated fruit shape." TAG Theoretical and Applied Genetics **116**(5): 647-656.

Gore, M. A., M. H. Wright, et al. (2009). "Large-scale discovery of gene-enriched SNPs." The Plant Genome **2**(2): 121-133.

Gouesnard, B., T. M. Bataillon, et al. (2001). "MSTRAT: An algorithm for building germplasm core collections by maximizing allelic or phenotypic richness." J Hered **92**(1): 93-94.

Grandillo, S., H. M. Ku, et al. (1999). "Identifying the loci responsible for natural variation in fruit size and shape in tomato." TAG Theoretical and Applied Genetics **V99**(6): 978-987.

Grandillo, S. and S. D. Tanksley (1996). "QTL analysis of horticultural traits differentiating the cultivated tomato from the closely related species *Lycopersicon pimpinellifolium*." TAG Theoretical and Applied Genetics **92**(8): 935-951.

Green , J. M., J. H. A. Barker , et al. (2001). "Microsatellite analysis of the inbreeding grass weed Barren Brome (*Anisantha sterilis*) reveals genetic diversity at the within- and between-farm scales." Molecular Ecology **10**(4): 1035-1045.

Griffith, F. (1928). "The significance of pneumococcal types." J. Hyg **27**: 113-159.

Guichard, S., N. Bertin, et al. (2001). "Tomato fruit quality in relation to water and carbon fluxes." Agronomie **21**(4): 385-392.

Gupta, P. K., S. Rustgi, et al. (2005). "Linkage disequilibrium and association studies in higher plants: Present status and future prospects." Plant Mol. Biol. **57**(4): 461-485.

Gut, I. G. (2001). "Automation in genotyping of single nucleotide polymorphisms." Human Mutation **17**(6): 475-492.

Hall, T. A. (1999). "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT." Nucl. Acids Symp. Ser. **41**: 95-98.

Hardy, O. J. and X. Vekemans (2002). "SPAGeDI: a versatile computer program to analyse spatial genetic structure at the individual or population levels." Molecular Ecology Notes **2**: 618-620.

Hershey, A. D. and M. Chase (1952). "Independant functions of viral proteins and nucleic acid in growth of bacteriophage." J. Gen. Physiol. **36**: 39-56.

Hirano, H. Y., M. Eiguchi, et al. (1998). "A single base change altered the regulation of the *Waxy* gene at the posttranscriptional level during the domestication of rice." Mol Biol Evol **15**(8): 978-987.

Holsinger, K. E. and W. L. E. (2004). "Bayesian approaches for the analysis of population genetic structure: an example from *Platanthera leucophaea* (*Orchidaceae*)." Mol. Ecol. **13**(4): 887-894.

Holsinger, K. E., P. O. Lewis, et al. (2002). "A Bayesian approach to inferring population structure from dominant markers." Mol. Ecol. **11**(7): 1157-1164.

Huson, D. H. (1998). "SplitsTree: analyzing and visualizing evolutionary data." Bioinformatics **14**(1): 68-73.

Hyten, D. L., I.-Y. Choi, et al. (2007). "Highly variable patterns of linkage disequilibrium in multiple soybean populations." Genetics **175**(4): 1937-1944.

Innan, H. and W. Stephan (2003). "Distinguishing the hitchhiking and background selection models." Genetics **165**(4): 2307-2312.

Isaacson, T., G. Ronen, et al. (2002). "Cloning of tangerine from tomato reveals a carotenoid isomerase essential for the production of {beta}-carotene and xanthophylls in plants." Plant Cell **14**(2): 333-342.

Isaksson, M., J. Stenberg, et al. (2007). "MLGA a rapid and cost-efficient assay for gene copy-number analysis." Nucl. Acids Res. **35**(17): e115-.

Ivandic, V., W. T. B. Thomas, et al. (2003). "Associations of simple sequence repeats with quantitative trait variation including biotic and abiotic stress tolerance in *Hordeum spontaneum*." Plant Breeding **122**(4): 300-304.

Iwata, H., Y. Uga, et al. (2007). "Bayesian association mapping of multiple quantitative trait loci and its application to the analysis of genetic variation among *Oryza sativa* L. germplasms." TAG Theoretical and Applied Genetics **114**(8): 1437-1449.

Jarvis, D. I. and T. Hodgkin (1999). "Wild relatives and crop cultivars: detecting natural introgression and farmer selection of new genetic combinations in agroecosystems doi:10.1046/j.1365-294X.1999.00799.x." Molecular Ecology **8**(s1): S159-S173.

Jenkins, J. (1948). "The origin of the cultivated tomato." Economic Botany **2**(4): 379-392.

Jimenez-Gomez, J. and J. Maloof (2009). "Sequence diversity in three tomato species: SNPs, markers, and molecular evolution." BMC Plant Biology **9**(1): 85.

Johanson, U., J. West, et al. (2000). "Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time." Science **290**(5490): 344-347.

Kang, H. M., N. A. Zaitlen, et al. (2008). "Efficient control of population structure in model organism association mapping." Genetics **178**(3): 1709-1723.

Kaya, H., K.-i. Shibahara, et al. (2001). "*FASCIATA* genes for chromatin assembly factor-1 in *Arabidopsis* maintain the cellular organization of apical meristems." Cell **104**(1): 131-142.

Kim, S., V. Plagnol, et al. (2007). "Recombination and linkage disequilibrium in *Arabidopsis thaliana*." Nat Genet **39**(9): 1151-1155.

Kraakman, A. T. W., R. E. Niks, et al. (2004). "Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars." Genetics **168**(1): 435-446.

Ku, H.-M., T. Vision, et al. (2000). "Comparing sequenced segments of the tomato and *Arabidopsis* genomes: Large-scale duplication followed by selective gene loss creates a network of synteny" Proceedings of the National Academy of Sciences of the United States of America **97**(16): 9121-9126.

Ku, H. M., S. Grandillo, et al. (2000). "fs8.1 , a major QTL, sets the pattern of tomato carpel shape well before anthesis." TAG Theoretical and Applied Genetics **101**(5): 873-878.

Kurlovich, B. S., S. I. Rep'ev, et al. (2000). "The significance of Vavilov's scientific expeditions and ideas for development and use of legume genetic resources." Plant Genetic Resources(124): 23-32.

Labate, J., L. Robertson, et al. (2009). "EST, COSII, and arbitrary gene markers give similar estimates of nucleotide diversity in cultivated tomato (*Solanum lycopersicum* L.)." Theo. App. Genet. **118**(5): 1005-1014.

Labate, J. A. and A. Baldo (2005). "Tomato SNP discovery by EST mining and resequencing." Mol. Breed. **16**(4): 343-349.

Labate, J. A., S. Grandillo, et al. (2007). Tomato. Vegetables**:** 1-125.

Labate, J. A., S. Grandillo, et al. (2007). Tomato. Genome mapping and molecular breeding in plants. C. Kole. NY, Springer Publishing. **5:** 1-125.

Lander, E. S. and N. J. Schork (2006). "Genetic Dissection of Complex Traits." Focus **4**(3): 442-458.

Laux, T., K. F. Mayer, et al. (1996). "The *WUSCHEL* gene is required for shoot and floral meristem integrity in *Arabidopsis.*" Development **122**(1): 87-96.

Le Cunff, L., A. Fournier-Level, et al. (2008). "Construction of nested genetic core collections to optimize the exploitation of natural diversity in *Vitis vinifera* L. subsp. *sativa*." BMC Plant Biology **8**(1): 31.

Lecomte, L., P. Duffé, et al. (2004). "Marker-assisted introgression of five QTLs controlling fruit quality traits into three tomato lines revealed interactions between QTLs and genetic backgrounds." TAG Theoretical and Applied Genetics **V109**(3): 658-668.

Lecomte, L., V. Saliba-Colombani, et al. (2004). "Fine mapping of QTLs of chromosome 2 affecting the fruit architecture and composition of tomato." Mol. Breed. **V13**(1): 1-14.

Lee, C. E. (2002). "Evolutionary genetics of invasive species." Trends in Ecology & Evolution **17**(8): 386-391.

Lexer, C., C. A. Buerkle, et al. (2006). "Admixture in European *Populus* hybrid zones makes feasible the mapping of loci that contribute to reproductive isolation and trait differences." Heredity **98**(2): 74-84.

Librado, P. and J. Rozas (2009). "DnaSP v5: A software for comprehensive analysis of DNA polymorphism data." Bioinformatics **25**: 1451-1452.

Lippman, Z. and S. D. Tanksley (2001). "Dissecting the genetic pathway to extreme fruit size in tomato using a cross between the small-fruited wild species *Lycopersicon pimpinellifolium* and *L. esculentum* var. Giant Heirloom." Genetics **158**(1): 413-422.

Lippman, Z. B., O. Cohen, et al. (2008). "The making of a compound inflorescence in tomato and related nightshades." PLoS Biol **6**(11): e288.

Lippman, Z. B., Y. Semel, et al. (2007). "An integrated view of quantitative trait variation using tomato interspecific introgression lines." Current Opinion in Genetics & Development **17**(6): 545-552.

Liu, J., J. Van Eck, et al. (2002). "A new class of regulatory genes underlying the cause of pear-shaped tomato fruit." Proc. Nat. Acad. Sci. USA **99**(20): 13302-13306.

Long, A. D. and C. H. Langley (1999). "The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits." Genome Res. **9**(8): 720-731.

Maher, B. (2008). "Personnal genomes: The case of missing heritability." Nature **456**: 18-21.

Manning, K., M. Tor, et al. (2006). "A naturally occurring epigenetic mutation in a gene encoding an *SBP-box* transcription factor inhibits tomato fruit ripening." Nat. Genet. **38**(8): 948-952.

Martin, A., C. Troadec, et al. (2009). "A transposon-induced epigenetic change leads to sex determination in melon." Nature **461**(7267): 1135-1138.

Mather, K. A., A. L. Caicedo, et al. (2007). "The extent of linkage disequilibrium in rice (*Oryza sativa* L.)." Genetics **177**(4): 2223-2232.

Matsuoka, Y., S. E. Mitchell, et al. (2002). "Microsatellites in zea - variability, patterns of mutations, and use for evolutionary studies." TAG Theoretical and Applied Genetics **104**(2): 436-450.

Matthiolus, P. A. (1544). Di Pedacio Diascoride Anazarbeo libri einque della historia, et materia medicinale tradotti in lingua volgare Italiana. Venetia.

Mauricio, R. (2001). "Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology." Nat Rev Genet **2**(5): 370-381.

Mayer, K. F. X., H. Schoof, et al. (1998). "Role of *WUSCHEL* in regulating stem cell fate in the *Arabidopsis* shoot meristem." Cell **95**: 805–815.

Mazzucato, A., R. Papa, et al. (2008). "Genetic diversity, structure and marker-trait associations in a collection of Italian tomato (*Solanum lycopersicum* L.) landraces." TAG Theoretical and Applied Genetics **116**(5): 657-669.

McCallum, C. M., L. Comai, et al. (2000). "Targeting Induced Local Lesions IN Genomes (TILLING) for plant functional genomics." Plant Physiol. **123**(2): 439-442.

McKhann, H. I., C. Camilleri, et al. (2004). "Nested core collections maximizing genetic diversity in *Arabidopsis thaliana*." The Plant Journal **38**(1): 193-202.

McMeekin, D. (1992). "Representations on pre-columbian spindle whorls of the floral and fruit structure of economic plants." Economic Botany **46**(2): 171-180.

McPherson, J. D. (2009). "Next-generation gap." Nat Meth **6**(11s): S2-S5.

Meissner, R., V. Chague, et al. (2000). "A high throughput system for transposon tagging and promoter trapping in tomato." The Plant Journal **22**(3): 265-274.

Menda, N., Y. Semel, et al. (2004). "*In silico* screening of a saturated mutation library of tomato." The Plant Journal **38**(5): 861-872.

Mendel, G. (1866). Experiments in plant hybridization: traduction originale de *Versuche über Plflanzenhybriden* par W. Bateson complétée par R. Blumberg, Electronic Scholarly Publishing Project.

Menz, M. A., R. R. Klein, et al. (2004). "Genetic diversity of public inbreds of *Sorghum* determined by mapped AFLP and SSR markers." Crop Science **44**(4): 1236-1244.

Miller, J. C. and S. D. Tanksley (1990). "RFLP analysis of phylogenetic relationships and genetic variation in the genus *Lycopersicon*." TAG Theoretical and Applied Genetics **80**(4): 437-448.

Monforte, A. J. and S. D. Tanksley (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." TAG Theoretical and Applied Genetics **100**(3): 471-479.

Moore, S., J. Vrebalov, et al. (2002). "Use of genomics tools to isolate key ripening genes and analyse fruit maturation in tomato." J. Exp. Bot. **53**(377): 2023-2030.

Morgan, T. H., A. H. Sturtevant, et al. (1915). The Mechanism of Mendelian Heredity. New York, Holt.

Moses, A. (2009). "Statistical tests for natural selection on regulatory regions based on the strength of transcription factor binding sites." BMC Evolutionary Biology **9**(1): 286.

Mueller, L. A., S. D. Tanksley, et al. (2005). "The Tomato Sequencing Project, the first cornerstone of the International Solanaceae Project (SOL)." Comp Funct Genom **6**(3): 153-158.

Muños, S., N. Ranc, et al. (*in prep.*). "Increases in tomato fruit size and locule number is controlled by two key SNP located near *Wuschel.*"

Nesbitt, T. C. and S. D. Tanksley (2002). "Comparative sequencing in the genus *Lycopersicon*: Implications for the evolution of fruit size in the domestication of cultivated tomatoes." Genetics **162**(1): 365 - 379.

Nguyen, M. L. and S. J. Schwartz (1998). "Lycopene stability during food processing." Proc-Soc-Exp-Biol-Med **218**(2): 101-105

Nordborg, M. (2000). "Linkage disequilibrium, gene trees and selfing: An ancestral recombination graph with partial self-fertilization." Genetics **154**(2): 923-929.

Nordborg, M., J. O. Borevitz, et al. (2002). "The extent of linkage disequilibrium in *Arabidopsis thaliana*." Nat. Genet. **30**(2): 190-193.

Nordborg, M., T. T. Hu, et al. (2005). "The pattern of polymorphism in *Arabidopsis thaliana*." PLoS Biol **3**(7): e196.

Nordborg, M. and S. Tavare (2002). "Linkage disequilibrium: what history has to tell us." Trends Genet. **18**: 83-90.

Ostrowski, M.-F., J. David, et al. (2006). "Evidence for a large-scale population structure among accessions of *Arabidopsis thaliana*: possible causes and consequences for the distribution of linkage disequilibrium." Molecular Ecology **15**(6): 1507-1517.

Pajerowska-Mukhtar, K., B. Stich, et al. (2009). "Single nucleotide polymorphisms in the *Allene oxide synthase 2* gene are associated with field resistance to late blight in populations of tetraploid potato cultivars." Genetics **181**(3): 1115-1127.

Palaisa, K., M. Morgante, et al. (2004). "Long-range patterns of diversity and linkage disequilibrium surrounding the maize *Y1* gene are indicative of an asymmetric selective sweep." PNAS **101**(26): 9885-9890.

Paran, I., I. Goldman, et al. (1995). "Recombinant inbred lines for genetic mapping in tomato." TAG Theoretical and Applied Genetics **90**(3): 542-548.

Paran, I. and E. van der Knaap (2007). "Genetic and molecular regulation of fruit and plant domestication traits in tomato and pepper." J. Exp. Bot. **58**(14): 3841-3852.

Parker, S. C. J., L. Hansen, et al. (2009). "Local DNA topography correlates with functional noncoding regions of the Human genome." Science **324**(5925): 389-392.

Paterson, A. H., S. Damon, et al. (1991). "Mendelian factors underlying quantitative traits in tomato: Comparison across species, generations, and environments." Genetics **127**(1): 181-197.

Paterson, A. H., J. W. DeVerna, et al. (1990). "Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes, in an interspecies cross of tomato." Genetics **124**(3): 735-742.

Paterson, A. H., E. S. Lander, et al. (1988). "Resolution of quantitative traits into mendelian factors by using a complete linkage map of restriction fragment length polymorphisms." Nature **335**: 721-726.

Paulus, C., B. Köllner, et al. (1993). "Physiological and biochemical characterization of glyoxalase I, a general marker for cell proliferation, from a soybean cell suspension." Planta **189**(4): 561-566.

Peralta, I. E. and D. Spooner (2007). History, origin and early cultivation of tomato (*Solanaceae*). Genetic improvement of Solanaceous crops. M. K. Razdan and A. K. Mattoo. Enfield (NH), Science Publisher. **2:** 1-24.

Peralta, I. E., D. Spooner, et al. (2007). Taxonomy of wild tomatoes and their relatives (*Solanum* sect. *Lycopersicoides*, sect. *Juglandifolia*, sect. *Lycopersicon*; *Solanaceae*).

Peralta, I. E. and D. M. Spooner (2005). Morphological characterization and relationships of wild tomatoes (*Solanum L.* Section *Lycopersicon*). A Festschrift for William G. D'Arcy. T. B. Croat, V. C. Hollowell and R. C. Keating, Missouri Botanical Garden Press. **104:** 227-257.

Perrier , X., A. Flori , et al. (2003). Data analysis methods. Genetic diversity of cultivated tropical plants. P. Hamon, Seguin, M., Perrier, X. ,Glaszmann, J. C. Ed. Montpellier, Enfield, Science Publishers**:** 43 - 76.

Perrier, X. and J. P. Jacquemoud-Collet (2006). DARwin software http://darwin.cirad.fr/darwin.

Pop, M. and S. L. Salzberg (2008). "Bioinformatics challenges of new sequencing technology." Trends in Genetics **24**(3): 142-149.

Price, A. L., N. J. Patterson, et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nat. Genet. **38**(8): 904-909.

Pritchard, J. K. and P. Donnelly (2001). "Case-control studies of association in structured or admixed populations." Theoretical Population Biology **60**(3): 227-237.

Pritchard, J. K. and M. Przeworski (2001). "Linkage disequilibrium in humans: models and data." The American Journal of Human Genetics **69**(1): 1-14.

Pritchard, J. K., M. Stephens, et al. (2000). "Inference of population structure using multilocus genotype data." Genetics **155**(2): 945-959.

Pritchard, J. K., M. Stephens, et al. (2000). "Association mapping in structured populations." Am. J. Hum. Genet. **67**: 170-181.

Prudent, M. (2008). Analyse des variations de poids et de teneur en sucres du fruit de tomate par une approche intégrative combinant des études écophysiologique, génétique et moléculaire. Avignon, Université d'Avignon et des Pays de Vaucluse. **Thèse de doctorat:** 152p.

Prudent, M., M. Causse, et al. (2009). "Genetic and physiological analysis of tomato fruit weight and composition: influence of carbon availability on QTL detection." J. Exp. Bot. **60**(3): 923-937.

R Development Core Team (2005). R: A language and environment for statistical computing, reference index version 2.2.1. F. f. S. Computing. Vienna, Austria.

Rafalski, A. (2002). "Applications of single nucleotide polymorphisms in crop genetics." Curr. Opin. Plant Biol. **5**: 94-100.

Rafalski, A. and M. Morgante (2004). "Corn and humans: recombination and linkage disequilibrium in two genomes of similar size." Trends in Genetics **20**(2): 103-111.

Ranc, N., S. Munos, et al. (2008). "A clarified position for *Solanum lycopersicum* var. *cerasiforme* in the evolutionary history of tomatoes (*Solanaceae*)." BMC Plant Biol. **8**(1): 130.

Ray, J., P. Moureau, et al. (1992). "Cloning and characterization of a gene involved in phytoene synthesis from tomato." Plant Molecular Biology **19**(3): 401-404.

Reeves, A. F. (1973). An observation on natural outcrossing in the tomato (*Lycopersicon esculentum* L.) in Northwest Arkansas. Arkansas Academy of Science Proceedings XXVII.

Reich, D. E., M. Cargill, et al. (2001). "Linkage disequilibrium in the human genome." Nature **411**(6834): 199-204.

Reinhardt, D., M. Frenz, et al. (2003). "Microsurgical and laser ablation analysis of interactions between the zones and layers of the tomato shoot apical meristem." Development **130**(17): 4073-4083.

Remington, D. L., J. M. Thornsberry, et al. (2001). "Structure of linkage disequilibrium and phenotypic associations in the maize genome. 10.1073/pnas.201394398." Proceedings of the National Academy of Sciences of the United States of America **98**(20): 11479-11484.

Rick , C. M. (1976). Tomato *Lycopersicon esculentum* (*Solanaceae*). Evolution of Crop Plants. N. W. Simmonds, Longman**:** 268-273.

Rick, C. M. and R. T. Chetelat (1995). "Utilization of related wild species for tomato improvement." Acta Horticulturae **412**: 21-38.

Rick , C. M. and F. Fobes (1975). "Allozyme variation in the cultivated tomato and closely related species." Bulletin of the Torrey Botanical Club **102**(6): 376-384.

Rick, C. M., J. F. Fobes, et al. (1977). "Genetic variation in *Lycopersicon pimpinellifolium*: Evidence of evolutionary change in mating systems." Plant Systematics and Evolution **127**(2): 139-170.

Rick , C. M. and M. Holle (1990). "Andean *Lycopersicum esculentum* var. *cerasiforme*: Genetic Varaition and Its Evolutionary Significance." *Economic Botany* **44**: 69-78.

Rick, C. M., H. Latterot, et al. (1990). "A revised key for the *Lycopersicum* and related *Solanum* species." Tomato Genetics Cooperative Report **40**: 31.

Rieseberg, L. H. and C. A. Buerkle (2002). "Genetic mapping in hybrid zones." The American Naturalist **159**(s3): S36-S50.

Ritland, K. (1996). "Estimators for pairwise relatedness and individual inbreeding coefficients." Genet. Res. **67**(02): 175-185.

Ronen, G., L. Carmel-Goren, et al. (2000). "An alternative pathway to beta-carotene formation in plant chromoplasts discovered by map-based cloning of *Beta* and *old-gold color* mutations in tomato." Proceedings of the National Academy of Sciences **97**(20): 11102-11107.

Ronen, G., M. Cohen, et al. (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*." The Plant Journal **17**(4): 341-351.

Ronfort, J., T. Bataillon, et al. (2006). "Microsatellite diversity and broad scale geographic structure in a model legume: building a set of nested core collection for studying naturally occurring variation in *Medicago truncatula*." BMC Plant Biology **6**(1): 28.

Rosenberg, N. A. (2007). *Distruct*: a program for the graphical display of population structure. Ann Arbor.

Rosenberg, N. A., J. K. Pritchard, et al. (2002). "Genetic structure of human populations." Science **298**(5602): 2381-2385.

Ross-Ibarra, J., P. L. Morrell, et al. (2007). "Plant domestication, a unique opportunity to identify the genetic basis of adaptation." PNAS **104**(suppl_1): 8641-8648.

Rostoks, N., L. Ramsay, et al. (2006). "Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties." Proceedings of the National Academy of Sciences **103**(49): 18656-18661.

Rozen, S. and H. Skaletsky (2000). Primer3 on the WWW for general users and for biologist programmers. Bioinformatics Methods and Protocols: Methods in Molecular Biology. S. Krawetz and S. Misener. Totowa, NJ, USA, Humana Press. **132:** 365 - 386.

Saliba-Colombani, V., M. Causse, et al. (2000). "Efficiency of AFLP, RAPD, and RFLP markers for the construction of an intraspecific map of the tomato genome." Genome **43**(1): 29-40.

Saliba-Colombani, V., M. Causse, et al. (2001). "Genetic analysis of organoleptic quality in fresh market tomato. 1. Mapping QTLs for physical and chemical traits." Theo. App. Genet. **V102**(2): 259-272.

Schoen, D. and A. Brown (1993). "Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers." PNAS **90**(22): 10623-10627.

Schoof, H., M. Lenhard, et al. (2000). "The stem cell population of *Arabidopsis* shoot meristems is maintained by a regulatory loop between the *CLAVATA* and *WUSCHEL* genes." Cell **100**(6): 635-644.

Seldin, M. F. (2007). "Admixture mapping as a tool in gene discovery." Current Opinion in Genetics & Development **17**(3): 177-181.

Sim, S.-C., M. Robbins, et al. (2009). "Oligonucleotide array discovery of polymorphisms in cultivated tomato (*Solanum lycopersicum* L.) reveals patterns of SNP variation associated with breeding." BMC Genomics **10**(1): 466.

Simko, I., S. Costanzo, et al. (2004). "Linkage disequilibrium mapping of a *Verticillium dahliae* resistance quantitative trait locus in tetraploid potato (*Solanum tuberosum*) through a candidate gene approach." TAG Theoretical and Applied Genetics **V108**(2): 217-224.

Simko, I., K. G. Haynes, et al. (2006). "Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers." Genetics **173**(4): 2237-2245.

Simon, M., O. Loudet, et al. (2008). "Quantitative trait loci mapping in five new large recombinant inbred line populations of *Arabidopsis thaliana* genotyped with consensus single-nucleotide polymorphism markers." Genetics **178**(4): 2253-2264.

Skøt, L., J. Humphreys, et al. (2007). "Association of candidate genes with flowering time and water-soluble carbohydrate content in *Lolium perenne* (L.)." Genetics **177**(1): 535-547.

Skøt, L., M. O. Humphreys, et al. (2005). "An association mapping approach to identify flowering time genes in natural populations of *Lolium perenne* (L.)." Molecular Breeding **15**(3): 233-245.

Smith, T. F., C. Gaitatzes, et al. (1999). "The WD repeat: a common architecture for diverse functions." Trends in Biochemical Sciences **24**(5): 181-185.

Smulders, M. J. M., G. Bredemeijer, et al. (1997). "Use of short microsatellites from database sequences to generate polymorphisms among *Lycopersicon esculentum* cultivars and accessions of other *Lycopersicon* species." TAG Theoretical and Applied Genetics **94**(2): 264-272.

Sneath, P. H. A. and R. R. Sokal (1973). Numerical Taxonomy. The Principles and Practice of Numerical Classification. San Francisco, W.H. Freeman an Co.

Somers, D. J., T. D. Banks, R., et al. (2007). "Genome-wide linkage disequilibrium analysis in bread wheat and durum wheat." Genome **50**: 557-567.

Spielman, R. S., R. E. McGinnis, et al. (1993). "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)." Am J Hum Genet. **52**(3): 506-516.

Spooner, D. M., I. E. Peralta, et al. (2005). "Comparison of AFLPs with other markers for phylogenetic inference in wild tomatoes [*Solanum L*. section *Lycopersicon* (Mill.) Wettst.]." Taxon **54**(1): 43-61.

Stevens, R. (2007). Tomato: A model plant for *Solanaceae* genomics. Functional Plant Genomics. J. F. Morot-Gaudry, P. Lea and J. F. Briat, Science Publisher.

Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." Genetics **123**(3): 585-595.

Takahashi, M., F. Matsuda, et al. (2003). "Automated identification of single nucleotide polymorphisms from sequencing data." J. Bioinform. Comput. Biol. **1**: 253-265.

Takezaki, N. and M. Nei (1996). "Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA." Genetics **144**(1): 389-399.

Tam, S. M., M. Causse, et al. (2007). "The distribution of copia-type retrotransposons and the evolutionary history of tomato and related wild species." Journal of Evolutionary Biology **20**(3): 1056-1072.

Tam, S. M., C. Mhiri, et al. (2005). "Comparative analyses of genetic diversities within tomato and pepper collections detected by retrotransposon-based SSAP, AFLP and SSR." TAG Theoretical and Applied Genetics **V110**(5): 819-831.

Tanksley, S. D. (2004). "The genetic, developmental, and molecular bases of fruit size and shape variation in tomato." Plant Cell **16**(1): 181-189.

Tanksley, S. D., M. W. Ganal, et al. (1992). "High density molecular linkage maps of the tomato and potato genomes." Genetics **132**(4): 1141-1160.

Tanksley, S. D., S. Grandillo, et al. (1996). "Advanced backcross QTL analysis in a cross between an elite processing line of tomato and its wild relative *L. pimpinellifolium*." Theo. App. Genet. **92**(2): 213-224.

Tenaillon, M. I., J. U'Ren, et al. (2004). "Selection *versus* demography: A multilocus investigation of the domestication process in maize." Mol Biol Evol **21**(7): 1214-1225.

Tenesa, A., A. F. Wright, et al. (2004). "Extent of linkage disequilibrium in a sardinian sub-isolate: sampling and methodological considerations." Hum. Mol. Genet. **13**(1): 25-33.

Tezozomoc, F. A. (1531). Cronica mexicayotl. México.

This, P., T. Lacombe, et al. (2007). "Wine grape (*Vitis vinifera* L.) color associates with allelic variation in the domestication gene *VvmybA1*." Theor Appl Genet **114**(4): 723-30.

Thompson, A. J., M. Tor, et al. (1999). "Molecular and genetic characterization of a novel pleiotropic tomato-ripening mutant." Plant Physiol. **120**(2): 383-390.

Thornsberry, J. M., M. M. Goodman, et al. (2001). "*Dwarf8* polymorphisms associate with variation in flowering time." Nat. Genet. **28**: 286 - 289.

Tommasini, L., T. Schnurbusch, et al. (2007). "Association mapping of *Stagonospora nodorum* blotch resistance in modern european winter wheat varieties." TAG Theoretical and Applied Genetics **115**(5): 697-708.

Tricon, D. (2005). Le nombre de loges chez la tomate : vers le clonage d'un QTL. Sciences de l'Environnement. Marseille, Faculté des Sciences de Luminy. **Master 2ème année:** 29.

Valdar, W., J. Flint, et al. (2006). "Simulating the collaborative cross: Power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice." Genetics **172**(3): 1783-1797.

van Berloo, R., A. Zhu, et al. (2008). "Diversity and linkage disequilibrium analysis within a selected set of cultivated tomatoes." Theo. App. Genet. **117**(1): 89-101.

van der Hoeven, R., C. Ronning, et al. (2002). "Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing." Plant Cell **14**(7): 1441-1456.

van der Knaap, E. and S. D. Tanksley (2003). "The making of a bell pepper-shaped tomato fruit: identification of loci controlling fruit morphology in Yellow Stuffer tomato." TAG Theoretical and Applied Genetics **107**(1): 139-147.

van Deynze, A., K. Stoffel, et al. (2007). "Diversity in conserved genes in tomato." BMC Genomics **8**(1): 465.

Verbsky, M. L. and E. J. Richards (2001). "Chromatin remodeling in plants." Curr. Opin. Plant Biol. **4**(6): 494-500.

Vitaliano-Prunier, A., A. Menant, et al. (2008). "Ubiquitylation of the COMPASS component Swd2 links H2B ubiquitylation to H3K4 trimethylation." Nat Cell Biol **10**(11): 1365-1371.

Vrebalov, J., D. Ruezinsky, et al. (2002). "A *MADS-Box* gene necessary for fruit ripening at the tomato *Ripening-Inhibitor* (*Rin*) locus." Science **296**(5566): 343-346.

Wall, J. D. and J. K. Pritchard (2003). "Haplotype blocks and linkage disequilibrium in the human genome." Nat Rev Genet **4**(8): 587-597.

Watson, J. D. and F. H. C. Crick (1953). "A structure for desoxyribose nucleic acid." Nature **171**: 737-738.

Watterson, G. (1975). "On the number of segregating sites in genetical models without recombination." Theor. Popul. Biol. **7**(2): 256-276.

Weber, A. L., W. H. Briggs, et al. (2008). "The genetic architecture of complex traits in teosinte (*Zea mays ssp. parviglumis*): New evidence from association mapping." Genetics **180**(2): 1221-1232.

Weigel, D. and M. Nordborg (2005). "Natural variation in *Arabidopsis*. How do we find the causal genes?" Plant Physiol. **138**(2): 567-568.

Weiss, G. and A. von Haeseler (1998). "Inference of population history using a likelihood approach." Genetics **149**(3): 1539-1546.

Whitt, S. R. and E. S. Buckler (2003). "Using natural allelic diversity to evaluate gene function." Methods in Molecular Biology **236**: 123-139.

Wilkinson, J. Q., M. B. Lanahan, et al. (1995). "An ethylene-inducible component of signal transduction encoded by never-ripe." Science **270**(5243): 1807-1809.

Wright, S. I. and B. S. Gaut (2005). "Molecular population genetics and the search for adaptive evolution in plants." Mol Biol Evol **22**(3): 506-519.

Xia, Q., Y. Guo, et al. (2009). "Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*)." Science **326**(5951): 433-436.

Yamasaki, M., S. G. Schroeder, et al. (2008). "Empirical analysis of selection screens for domestication and improvement loci in maize by extended DNA sequencing." The Plant Genome **1**(1): 33-43.

Yamasaki, M., M. I. Tenaillon, et al. (2005). "A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement." Plant Cell **17**(11): 2859-2872.

Yang, W., X. Bai, et al. (2004). "Discovery of single nucleotide polymorphisms in *Lycopersicon esculentum* by computer aided analysis of expressed sequence tags." Molecular Breeding **14**(1): 21 - 34.

Yeager, A. (1937). "Studies on the inheritance and development of fruit size and shape in the tomato." J Agric Res **55**: 141-152.

Yong-Sheng, L., G. Amit, et al. (2003). "There is more to tomato fruit colour than candidate carotenoid genes." Plant Biotechnology Journal **1**(3): 195-207.

Yousef, G. G. and J. A. Juvik (2001). "Evaluation of breeding utility of a chromosomal segment from *Lycopersicon chmielewskii* that enhances cultivated tomato soluble solids." TAG Theoretical and Applied Genetics **103**(6): 1022-1027.

Yu, J. and E. S. Buckler (2006). "Genetic association mapping and genome organization of maize." Current Opinion in Biotechnology **17**(2): 155-160.

Yu, J., J. B. Holland, et al. (2008). "Genetic design and statistical power of nested association mapping in maize." Genetics **178**(1): 539-551.

Yu, J., G. Pressoir, et al. (2006). "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness." Nat. Genet. **38**(2): 203-208.

Zhang, L., S. Marchand, et al. (2009). "Population structure and linkage disequilibrium in barley assessed by DArT markers." Theo. App. Genet. **119**(1): 43-52.

Zhao, K., M. J. Aranzana, et al. (2007). "An *Arabidopsis* example of association mapping in structured samples." PLoS Genet **3**(1): e4.

Zhu, C., M. Gore, et al. (2008). "Status and prospects of association mapping in plants." The Plant Genome **1**(1): 5-20.

Zhu, Q., X. Zheng, et al. (2007). "Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice." Mol Biol Evol **24**(3): 875-888.

Zhu, X., A. Luke, et al. (2005). "Admixture mapping for hypertension loci with genome-scan markers." **37**(2): 177-181.

Annexes

Annexe 1. Détail des 360 accessions utilisées durant les travaux de thèse

Les collections d'origines des accessions sont indiquées de la façon suivante :
INRA: Institut National de Recherche Agronomique
VIR: Vavilov Research Institute of Plant Industry
USDA: United State Department of Agriculture (North Central Regional Plant Introduction Station)
La « core collection » de 24 accessions (cc24) est emboîtée dans la collection de 96 accessions.

Code	Nom	Collection d'origine	Espèces	core collection	Phénotypage tunnel
CR001	Cervil	INRA	*S. lycopersicum* var *cerasiforme*	cc24	x
CR002	Levovil	INRA	*S. lycopersicum*	cc24	x
CR003	Ferum	INRA	*S. lycopersicum*	cc24	x
CR004	M-82	INRA	*S. lycopersicum*	cc24	x
CR005	Mospomorist	INRA	*S. lycopersicum*		x
CR006	Apeline	INRA	*S. lycopersicum*		x
CR007	Apedice	INRA	*S. lycopersicum*		
CR008	Gardener's Delight	INRA	*S. lycopersicum*		x
CR009	Sekaiichi	INRA	*S. lycopersicum*		
CR010	Coeur de Boeuf	INRA	*S. lycopersicum*		
CR011	Rodade	INRA	*S. lycopersicum*		x
CR012	Saint Pierre	INRA	*S. lycopersicum*		
CR013	Claudine	INRA	*S. lycopersicum*		
CR014	Clémentine	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR015	Flora Dade	INRA	*S. lycopersicum*		
CR016	Leningradskij skorospelij	INRA	*S. lycopersicum*		
CR017	Philippino n°2	INRA	*S. lycopersicum*		x
CR018	Poncette	INRA	*S. lycopersicum*		
CR019	Platense	INRA	*S. lycopersicum*		
CR020	San Marzano	INRA	*S. lycopersicum*	cc96	x
CR021	Rio Grande	INRA	*S. lycopersicum*		x
CR022	Marmande	INRA	*S. lycopersicum*		x
CR023	Supermarmande	INRA	*S. lycopersicum*		
CR024	Vendor	INRA	*S. lycopersicum*		
CR025	VF145-7879	INRA	*S. lycopersicum*		
CR026	Wva63	INRA	*S. lycopersicum*		
CR027	Plovdivska Konserva	INRA	*S. lycopersicum*		x
CR028	Plovdiv XXIVa	INRA	*S. lycopersicum*	cc96	x
CR029	Justar	INRA	*S. lycopersicum*		
CR030	Severianin	INRA	*S. lycopersicum*		x

Code	Nom	Collection d'origine	Espèces	core collection	Phénotypage tunnel
CR031	Microtom	INRA	*S. lycopersicum* var *cerasiforme*	cc24	x
CR032	Moneymaker	INRA	*S. lycopersicum*	cc24	x
CR033	Monalbo	INRA	*S. lycopersicum*		
CR034	Motelle	INRA	*S. lycopersicum*		
CR035	Moboglan	INRA	*S. lycopersicum*		
CR036	Momor	INRA	*S. lycopersicum*		
CR037	Mogéor	INRA	*S. lycopersicum*		
CR038	Pistou	INRA	*S. lycopersicum*		
CR039	Cigalou	INRA	*S. lycopersicum*		
CR040	Coudoulet	INRA	*S. lycopersicum*		
CR041	Caraïbo	INRA	*S. lycopersicum*		
CR042	Blanche	INRA	*S. lycopersicum*		
CR043	Gold Nugget	INRA	*S. lycopersicum*		
CR044	High Crimson	INRA	*S. lycopersicum*		
CR045	Jaune Demi Lisse	INRA	*S. lycopersicum*		
CR046	Nagcarlan	INRA	*S. lycopersicum*		x
CR047	Pêche	INRA	*S. lycopersicum*		
CR048	Yellow Pear	INRA	*S. lycopersicum* var *cerasiforme*		x
CR049	Red Pear	INRA	*S. lycopersicum* var *cerasiforme*		x
CR050	Sucrée à gros fruits	INRA	*S. lycopersicum*		
CR051	Vesuvio	INRA	*S. lycopersicum*		x
CR052	Sweetie	INRA	*S. lycopersicum* var *cerasiforme*		x
CR053	Poivron des Andes	INRA	*S. lycopersicum*		
CR054	Purple Kalabash	INRA	*S. lycopersicum*		
CR055	Ailsa Craig	INRA	*S. lycopersicum*		x
CR056	Wva700	INRA	*S. pimpinellifolium*	cc24	x
CR057	L3708	INRA	*S. pimpinellifolium*		x
CR058	Wva106	INRA	*S. pimpinellifolium*	cc96	x
CR059	Hirsute	INRA	*S. pimpinellifolium*		
CR060	LA0121	TGRC	*S. pimpinellifolium*		x
CR061	LA1582	TGRC	*S. pimpinellifolium*		
CR062	LA1478	TGRC	*S. pimpinellifolium*	cc96	x
CR063	63280	INRA	*S. pimpinellifolium*		
CR064	66083	INRA	*S. pimpinellifolium*		x
CR065	64316	INRA	*S. pimpinellifolium*		x
CR066	Racemigerum		*S. pimpinellifolium*		
CR067	VIR1018	VIR	*S. pimpinellifolium*		x
CR068	VIR108	VIR	*S. pimpinellifolium*	cc96	x
CR069	VIR2863	VIR	*S. pimpinellifolium*		x
CR070	VIR2909	VIR	*S. lycopersicum* var *cerasiforme*	cc96	x
CR071	VIR2920	VIR	*S. pimpinellifolium*		x
CR072	VIR2921	VIR	*S. pimpinellifolium*	cc96	x
CR073	VIR3101	VIR	*S. lycopersicum* var *cerasiforme*		x

Code	Nom	Collection d'origine	Espèces	core collection	Phénotypage tunnel
CR074	VIR4053	VIR	*S. lycopersicum* var *cerasiforme*		x
CR075	VIR4156	VIR	*S. pimpinellifolium*	cc96	x
CR076	VIR135	VIR	*S. lycopersicum* var *cerasiforme*	cc96	x
CR077	VIR1565	VIR	*S. lycopersicum* var *cerasiforme*	cc96	x
CR078	VIR2759	VIR	*S. lycopersicum* var *cerasiforme*	cc96	x
CR079	VIR933	VIR	*S. lycopersicum* var *cerasiforme*	cc96	x
CR080	VIR914	VIR	*S. lycopersicum* var *cerasiforme*		x
CR081	VIR342	VIR	*S. lycopersicum* var *cerasiforme*		x
CR082	VIR362	VIR	*S. lycopersicum* var *cerasiforme*		x
CR083	VIR749	VIR	*S. lycopersicum* var *cerasiforme*		x
CR084	VIR939	VIR	*S. lycopersicum* var *cerasiforme*		x
CR085	VIR996	VIR	*S. lycopersicum* var *cerasiforme*		x
CR086	VIR364	VIR	*S. lycopersicum* var *cerasiforme*		x
CR087	VIR350	VIR	*S. lycopersicum* var *cerasiforme*		x
CR088	VIR697	VIR	*S. lycopersicum* var *cerasiforme*		x
CR089	VIR859	VIR	*S. lycopersicum* var *cerasiforme*		x
CR090	VIR2448	VIR	*S. lycopersicum* var *cerasiforme*		x
CR091	VIR416	VIR	*S. lycopersicum* var *cerasiforme*		x
CR092	VIR2277	VIR	*S. lycopersicum* var *cerasiforme*		x
CR093	VIR2257	VIR	*S. lycopersicum* var *cerasiforme*	cc96	x
CR094	VIR1011	VIR	*S. lycopersicum*	cc24	x
CR095	VIR746	VIR	*S. lycopersicum* var *cerasiforme*		x
CR096	VIR797	VIR	*S. lycopersicum* var *cerasiforme*		x
CR097	VIR347	VIR	*S. lycopersicum* var *cerasiforme*	cc96	x
CR098	VIR795	VIR	*S. lycopersicum* var *cerasiforme*	cc96	x
CR099	VIR744	VIR	*S. lycopersicum* var *cerasiforme*		x
CR100	VIR341	VIR	*S. lycopersicum* var *cerasiforme*		x
CR101	VIR884	VIR	*S. lycopersicum* var *cerasiforme*	cc96	x
CR102	VIR739	VIR	*S. lycopersicum* var *cerasiforme*	cc96	x
CR103	LA0148	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR104	LA0292	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR105	LA1025	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR106	LA1204	TGRC	*S. lycopersicum* var *cerasiforme*	cc96	x
CR107	LA1228	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR108	LA1231	TGRC	*S. lycopersicum* var *cerasiforme*	cc96	x
CR109	LA1268	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR110	LA1307	TGRC	*S. lycopersicum* var *cerasiforme*	cc96	x
CR111	LA1312	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR112	LA1314	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR113	LA1320	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR114	LA1323	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR115	LA1338	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR116	LA1385	TGRC	*S. lycopersicum* var *cerasiforme*		x

Code	Nom	Collection d'origine	Espèces	core collection	Phénotypage tunnel
CR117	LA1388	TGRC	*S. lycopersicum* var *cerasiforme*	cc96	x
CR118	LA1420	TGRC	*S. lycopersicum* var *cerasiforme*	cc96	x
CR119	LA1425	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR120	LA1429	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR121	LA1453	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR122	LA1456	TGRC	*S. lycopersicum* var *cerasiforme*	cc96	x
CR123	LA1461	TGRC	*S. lycopersicum* var *cerasiforme*	cc96	x
CR124	LA1464	TGRC	*S. lycopersicum* var *cerasiforme*	cc24	x
CR125	LA1482	TGRC	*S. lycopersicum* var *cerasiforme*	cc96	x
CR126	LA0126	TGRC	*S. lycopersicum*		x
CR127	LA0134C	TGRC	*S. lycopersicum*		
CR128	LA0146	TGRC	*S. lycopersicum*		
CR129	LA0147	TGRC	*S. lycopersicum*	cc24	x
CR130	LA0172	TGRC	*S. lycopersicum* var *cerasiforme*	cc96	x
CR131	LA0358	TGRC	*S. lycopersicum*		
CR133	LA0409	TGRC	*S. lycopersicum*	cc24	x
CR134	LA0466	TGRC	*S. lycopersicum*	cc96	x
CR135	LA0468	TGRC	*S. lycopersicum*		x
CR136	LA0473	TGRC	*S. lycopersicum*	cc96	x
CR137	LA0477	TGRC	*S. lycopersicum*		x
CR138	LA1021	TGRC	*S. lycopersicum*		x
CR139	LA1162	TGRC	*S. lycopersicum*		x
CR140	LA1251	TGRC	*S. lycopersicum*		x
CR141	LA1286	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR142	LA1509	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR143	LA1511	TGRC	*S. lycopersicum* var *cerasiforme*		
CR144	LA1542	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR145	LA1543	TGRC	*S. lycopersicum* var *cerasiforme*	cc96	x
CR146	LA1620	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR147	LA1622	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR148	LA2078	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR149	LA2095	TGRC	*S. lycopersicum* var *cerasiforme*	cc96	x
CR150	LA2131	TGRC	*S. lycopersicum* var *cerasiforme*	cc96	x
CR151	LA2137	TGRC	*S. lycopersicum* var *cerasiforme*	cc24	
CR152	LA2307	TGRC	*S. lycopersicum* var *cerasiforme*	cc96	x
CR153	LA2308	TGRC	*S. lycopersicum* var *cerasiforme*	cc96	x
CR154	LA2392	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR155	LA2402	TGRC	*S. lycopersicum* var *cerasiforme*	cc96	x
CR156	LA2619	TGRC	*S. lycopersicum* var *cerasiforme*	cc96	x
CR157	LA2670	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR158	LA2675	TGRC	*S. lycopersicum* var *cerasiforme*	cc24	x
CR159	LA2688	TGRC	*S. lycopersicum* var *cerasiforme*	cc96	x
CR160	LA2703	TGRC	*S. lycopersicum* var *cerasiforme*		

Code	Nom	Collection d'origine	Espèces	core collection	Phénotypage tunnel
CR161	LA2709	TGRC	*S. lycopersicum* var *cerasiforme*		x
CR162	LA0373	TGRC	*S. pimpinellifolium*		
CR163	LA0400	TGRC	*S. pimpinellifolium*	cc96	x
CR164	LA0411	TGRC	*S. pimpinellifolium*	cc96	x
CR165	LA1237	TGRC	*S. pimpinellifolium*		
CR166	LA1245	TGRC	*S. pimpinellifolium*		
CR167	LA1246	TGRC	*S. pimpinellifolium*		
CR168	LA1261	TGRC	*S. pimpinellifolium*		
CR169	LA1371	TGRC	*S. pimpinellifolium*	cc96	x
CR170	LA1375	TGRC	*S. pimpinellifolium*		
CR171	LA1478	TGRC	*S. pimpinellifolium*		
CR172	LA1521	TGRC	*S. pimpinellifolium*		
CR173	LA1547	TGRC	*S. pimpinellifolium*	cc96	x
CR174	LA1576	TGRC	*S. pimpinellifolium*		
CR175	LA1578	TGRC	*S. pimpinellifolium*		
CR176	LA1582	TGRC	*S. pimpinellifolium*		
CR177	LA1584	TGRC	*S. pimpinellifolium*		
CR178	LA1590	TGRC	*S. pimpinellifolium*		
CR179	LA1593	TGRC	*S. pimpinellifolium*		
CR180	LA1599	TGRC	*S. pimpinellifolium*		
CR181	LA1602	TGRC	*S. pimpinellifolium*	cc24	
CR182	LA1606	TGRC	*S. pimpinellifolium*		
CR183	LA1617	TGRC	*S. pimpinellifolium*		
CR184	LA1659	TGRC	*S. pimpinellifolium*		
CR185	LA1683	TGRC	*S. pimpinellifolium*		
CR186	LA1689	TGRC	*S. pimpinellifolium*	cc96	x
CR187	LA1729	TGRC	*S. pimpinellifolium*		
CR188	LA1923	TGRC	*S. pimpinellifolium*		
CR189	LA1950	TGRC	*S. pimpinellifolium*		
CR190	LA2102	TGRC	*S. pimpinellifolium*		
CR191	LA2173	TGRC	*S. pimpinellifolium*		
CR192	LA2401	TGRC	*S. pimpinellifolium*		
CR193	LA2181	TGRC	*S. pimpinellifolium*		
CR194	LA2183	TGRC	*S. pimpinellifolium*		
CR195	LA2533	TGRC	*S. pimpinellifolium*		
CR196	LA2852	TGRC	*S. pimpinellifolium*		
CR197	CGN18401	CGN	*S. pimpinellifolium*		
CR198	Pimpinellifolium hirsute	INRA	*S. pimpinellifolium*		
CR199	tomate Richter's	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR200	CGN15809	CGN	*S. pimpinellifolium*		
CR201	CGN15811	CGN	*S. pimpinellifolium*		
CR202	CGN18399	CGN	*S. lycopersicum* var *cerasiforme*	cc96	x
CR203	LA1589	TGRC	*S. pimpinellifolium*	cc24	x

Code	Nom	Collection d'origine	Espèces	core collection	Phénotypage tunnel
CR204	L. pimpinellifolium typique, site 10 (F300044)	INRA	*S. pimpinellifolium*		
CR205	L. pimpinellifolium atypique, site 10 (F300045)	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR206	PI247087	USDA	*S. habrochaites*	cc24	
CR207	LA0716	TGRC	*S. pennellii*	cc24	
CR208	721404	INRA	*S. pimpinellifolium*		
CR209	732292	INRA	*S. pimpinellifolium*		
CR210	PI126410	USDA	*S. pimpinellifolium*		
CR211	LA1842	TGRC	*S. pimpinellifolium*		
CR212	LA1843	TGRC	*S. pimpinellifolium*		
CR213	LA1845	TGRC	*S. pimpinellifolium*		
CR214	groseille jaune	INRA	*S. pimpinellifolium*		
CR215	groseille rouge	INRA	*S. pimpinellifolium*		
CR216	PI134417	USDA	*S. habrochaites*		
CR217	B	INRA	*S. habrochaites*		
CR218	G1290	INRA	*S. habrochaites*		
CR219	H 2	INRA	*S. habrochaites*		
CR220	LA1777	TGRC	*S. habrochaites*		
CR221	PI 390660	USDA	*S. habrochaites*		
CR222	87610012	INRA	*S. habrochaites*		
CR223	10496	INRA	*S. habrochaites*		
CR224	LA1317	TGRC	*S. chmielewskii*		
CR225	LA1318	TGRC	*S. chmielewskii*		
CR226	LA1969	TGRC	*S. chilense*		
CR227	LA1971	TGRC	*S. chilense*		
CR228	LA1401	TGRC	*S. galapagense*	cc24	
CR229	LA1450	TGRC	*S. cheesmaniae*		
CR230	CMV sél INRA	INRA	*S. peruvianum*		
CR231	PI126435	USDA	*S. peruvianum*		
CR232	Clayberg	INRA	*S. pennellii*		
CR233	LA1321	TGRC	*S. neorickii*		
CR234	Atom	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR235	Blondköpchen	INRA	*S. lycopersicum* var *cerasiforme*		x
CR236	PI365923	USDA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR237	PI365925	USDA	*S. lycopersicum* var *cerasiforme*		x
CR238	PI129088	USDA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR239	Chello yellow Cherry	INRA	*S. lycopersicum* var *cerasiforme*		x
CR240	L285	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR241	SVS n° 1	INRA	*S. lycopersicum* var *cerasiforme*		x
CR242	cerise (-)	INRA	*S. lycopersicum* var *cerasiforme*		x
CR243	Yellow Pico	INRA	*S. lycopersicum* var *cerasiforme*		x
CR244	Yellow Pear	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR245	Cherita	INRA	*S. lycopersicum* var *cerasiforme*		x

Code	Nom	Collection d'origine	Espèces	core collection	Phénotypage tunnel
CR246	Zucker Kleinfrüchtige Rote	INRA	*S. lycopersicum* var *cerasiforme*		x
CR247	Cerise Jaune	INRA	*S. lycopersicum* var *cerasiforme*		x
CR248	Cerise Rouge	INRA	*S. lycopersicum* var *cerasiforme*		x
CR249	Cherry Gold	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR250	Cherry VFNT	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR251	Cherry VFNT sp	INRA	*S. lycopersicum* var *cerasiforme*		
CR252	Droplet	INRA	*S. lycopersicum*	cc96	x
CR253	Monplaisir	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR254	Farthest North	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR255	Florida Petite	INRA	*S. lycopersicum* var *cerasiforme*		x
CR256	Minibel	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR257	NC 2 C Cherry	INRA	*S. lycopersicum* var *cerasiforme*		x
CR258	Ohmiya Suncherry	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR259	Oregon Cherry	INRA	*S. lycopersicum* var *cerasiforme*		x
CR260	Osu 4014-4	INRA	*S. lycopersicum* var *cerasiforme*		x
CR261	Parteno (Italie)	INRA	*S. lycopersicum* var *cerasiforme*		x
CR262	Prachtlow	INRA	*S. lycopersicum* var *cerasiforme*		x
CR263	Red Cherry Small	INRA	*S. lycopersicum* var *cerasiforme*		x
CR264	Red Plum	INRA	*S. lycopersicum* var *cerasiforme*		x
CR265	Red robin	INRA	*S. lycopersicum* var *cerasiforme*		x
CR266	Sub arctic Cherry	INRA	*S. lycopersicum* var *cerasiforme*		x
CR267	Tiny tim	INRA	*S. lycopersicum* var *cerasiforme*	cc24	x
CR268	Willamette Cherry	INRA	*S. lycopersicum* var *cerasiforme*		x
CR269	Yellow Plum	INRA	*S. lycopersicum* var *cerasiforme*		x
CR270	Monita	INRA	*S. lycopersicum* var *cerasiforme*		x
CR271	Celsior	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR272	Peruvianum	INRA	*S. lycopersicum*		x
CR273	Orange Cocktail	INRA	*S. lycopersicum*	cc24	x
CR274	Marpha n°2	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR275	Cerise Ildi	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR276	Banjul 2	INRA	*S. lycopersicum* var *cerasiforme*		x
CR277	Mirabella	INRA	*S. lycopersicum* var *cerasiforme*		x
CR278	Uovetto di Grossetto	INRA	*S. lycopersicum* var *cerasiforme*		x
CR279	Cerise Orange d'Uzès	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR280	Cerise du sud ouest n° 2	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR281	Cerise du sud ouest n° 1	INRA	*S. lycopersicum* var *cerasiforme*		x
CR282	Cherry Belle	INRA	*S. lycopersicum* var *cerasiforme*		x
CR283	Cerise Brillante	INRA	*S. lycopersicum* var *cerasiforme*		x
CR284	cerise rose	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR285	Tondo Rossa Grappoli	INRA	*S. lycopersicum* var *cerasiforme*		x
CR286	Poc-poc	INRA	*S. lycopersicum* var *cerasiforme*		x
CR287	Cisterno	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR288	Criollo	INRA	*S. lycopersicum* var *cerasiforme*	cc24	x

Code	Nom	Collection d'origine	Espèces	core collection	Phénotypage tunnel
CR289	Piguti Tamagoshi Nadi	INRA	*S. lycopersicum* var *cerasiforme*		x
CR290	Luteum	INRA	*S. lycopersicum* var *cerasiforme*		x
CR291	Pyriforme	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR292	8 bis	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR293	Costa Rica	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR294	Phyra	INRA	*S. lycopersicum* var *cerasiforme*	cc24	x
CR295	Poire rouge	INRA	*S. lycopersicum* var *cerasiforme*		x
CR296	Poire jaune	INRA	*S. lycopersicum* var *cerasiforme*	cc96	x
CR297	Montfavet 133-5	INRA	*S. lycopersicum*		
CR298	College Abundant	INRA	S. lycopersicum		
CR299	Scorpio	INRA	*S. lycopersicum*		
CR300	Santa Cruz Gigante Resistente	INRA	*S. lycopersicum*		
CR301	Angela Gigante I-5100	INRA	*S. lycopersicum*		
CR302	Santa Cruz Samano	INRA	*S. lycopersicum*		x
CR303	Placero Lobulado	INRA	*S. lycopersicum*		
CR304	Huando	INRA	*S. lycopersicum*		
CR305	Coldset outdoor Seeder	INRA	*S. lycopersicum*		
CR306	Ontario 798	INRA	*S. lycopersicum*		
CR307	Pasionato	INRA	*S. lycopersicum*		
CR308	Ping Pong	INRA	*S. lycopersicum* var *cerasiforme*		x
CR309	Summerdawn	INRA	*S. lycopersicum*		
CR310	Vantage	INRA	*S. lycopersicum*		
CR311	Veecrop	INRA	*S. lycopersicum*		
CR312	Wisconsin A 55 VR	INRA	*S. lycopersicum*		
CR313	Alexander 630818	INRA	*S. lycopersicum*		
CR314	Beltsville 67B 833-1	INRA	*S. lycopersicum*		
CR315	Burpee's Longkeeper	INRA	*S. lycopersicum*		
CR316	Florida MH1	INRA	*S. lycopersicum*		
CR317	Heinz 1706	INRA	*S. lycopersicum*	cc24	x
CR318	Purdue 135	INRA	*S. lycopersicum*		
CR319	Rutgers	INRA	*S. lycopersicum*		x
CR320	Roodeplaat Albesto	INRA	*S. lycopersicum*		
CR321	Edkawy	INRA	*S. lycopersicum*	cc96	x
CR322	Xeewel I navet	INRA	*S. lycopersicum*		
CR323	Xina	INRA	*S. lycopersicum*		
CR324	Fenhong Tianrou	INRA	*S. lycopersicum*		x
CR325	Heiyuanxuan 2-2	INRA	*S. lycopersicum*		
CR326	Heiyuanxuan 2-3	INRA	*S. lycopersicum*		
CR327	Hongmanao 144	INRA	*S. lycopersicum*		
CR328	Qiangfeng	INRA	*S. lycopersicum*		
CR329	Qianglimishou	INRA	*S. lycopersicum*		
CR330	Guangyong	INRA	*S. lycopersicum*		
CR331	Heimeiren	INRA	*S. lycopersicum*		

Code	Nom	Collection d'origine	Espèces	core collection	Phénotypage tunnel
CR332	Wei 12	INRA	*S. lycopersicum*		
CR333	Pinguan 7	INRA	*S. lycopersicum*		
CR334	Zaofen n°1	INRA	*S. lycopersicum*		
CR335	Zongshu n° 5	INRA	*S. lycopersicum*		
CR336	Lichun	INRA	*S. lycopersicum*		
CR337	Kagome 6	INRA	*S. lycopersicum*		
CR338	Kikko 413	INRA	*S. lycopersicum*		
CR339	Ohmiya 163	INRA	*S. lycopersicum*		
CR340	Kurihara	INRA	*S. lycopersicum*		
CR341	Cra 66	INRA	*S. lycopersicum*	cc96	x
CR342	Taes n° 9	INRA	*S. lycopersicum*		
CR343	Hisar Aun	INRA	*S. lycopersicum*		
CR344	Pusa Ruby	INRA	*S. lycopersicum*		
CR345	Buzau 22	INRA	*S. lycopersicum*		
CR346	Buzau 1600	INRA	*S. lycopersicum*		
CR347	Somesan	INRA	*S. lycopersicum*		
CR348	Kecskemeti 476	INRA	*S. lycopersicum*		
CR349	Hebros	INRA	*S. lycopersicum*		
CR350	Marti	INRA	*S. lycopersicum*		
CR351	Nesthäckchen	INRA	*S. lycopersicum*		x
CR352	Opus	INRA	*S. lycopersicum*		
CR353	Orlowski	INRA	*S. lycopersicum*		
CR354	Stupicke Polni Rane	INRA	*S. lycopersicum*	cc96	x
CR355	Doneckij 3/2-1	INRA	*S. lycopersicum*		
CR356	Pridneprovskij Grasevidnyj	INRA	*S. lycopersicum*		
CR357	De Colgar or. Espagne	INRA	*S. lycopersicum*		
CR358	Montserrat	INRA	*S. lycopersicum*		
CR359	Muchamiel	INRA	*S. lycopersicum*	cc96	x
CR360	Allongée Cornichon	INRA	*S. lycopersicum*		x
CR361	Cannery Row	INRA	*S. lycopersicum*		

Annexe 2. Détails des multiplex utilisés pour les marqueurs microsatellites.

Triplex1

	Nom d'origine	Motif	Espèce d'origine	Chr.	Séquence F et R	Ref biblio
MS130	SSR599	[TCATTA]2[TCA]6	Tomate	9	GGATTTCTCATGGAGAATCAGTC 6-fam CCTTGATCTTGATGATGTTG	SNG
MS116	SSR111	$[TC]_6[TCTG]_6$	Tomate	3	TTCTTCCCTTCCATCAGTTCT hex TTTGCTGCTATACTGCTGACA	Frary et al. 2005
MS117	SSR14	$[ATA]_9$	Tomate	3	TCTGCATCTGGTGAAGCAAG ned CTGGATTGCCTGGTTGATTT	Frary et al. 2005

Frary A, Xu Y, Liu J, Mitchell S, Tedeschi E, Tanksley S (2005) Development of a set of PCR-based anchor markers encompassing the tomato genome and evaluation of their usefulness for genetics and breeding experiments. Theor Appl Genet 111:291312

PCR

Eau	qsp 20µl
Tampon 10X	1X
dNTP	0.8mM
Primers MS130 F	2pmole
Primers MS130 R	1pmole
Primers MS116 F	1.5pmole
Primers MS116 R	2pmole
Primers MS117 F	3pmole
Primers MS117 R	1.4pmole
DNA	50ng
Taq Pol	1u

Programme PCR

	5min	
94°C		
94°C	30sec	
55°C	45sec	} 35cycles
72°C	1min30	
72°C	7min	
15°C		

Dilution : 1/50
3µl de PCR diluée + 8µl de formamide/marqueur de taille GS400HD (7.85/0.15)

Electrophorèse : Séquenceur capillaires ABI3130xl, 50cm, POP7

Genotypage sur 8 témoins: GeneMapper

	MS130				MS116				MS117			
	6-fam - bleu				hex - vert				ned - jaune			
	?				$[TC]_6[TCTG]_6$				$[ATA]_9$			
nom	Allele 1	Allele 2	Size 1	Size 2	Allele 1	Allele 2	Size 1	Size 2	Allele 1	Allele 2	Size 1	Size 2
CR031	287		287.68		185		185.79		167		167.34	
CR001	287		287.72		177		177.73		167		167.73	
CR002	287		287.72		185		186.2		167		167.65	
CR003	287				177		178.61		167		167.6	
CR004	287		287.82		177		178.34		177		177.25	
CR203	287		287.78		177		177.71		164		164.18	
CR206	278		278.5		173		172.59		167		167.61	
CR207	278		278.67		163		162.94		158		158.1	

Triplex2

	nom d'origine	Motif	Espèce d'origine	Chr.	Séquence F et R	Ref biblio
MS131	SSR248	$[TA]_{21}$	Tomate	10	GCATTCGCTGTAGCTCGTTT 6-fam GGAGCTTCATCATAGTAACG	Frary et al. 2005
MS052	SSR52	$[AAC]_9$	Tomate	7	TGATGGCAGCATCGTAGAAG hex GGTGCGAAGGGATTTACAGA	SGN
MS114	SSR150	$[CTT]_7$	Tomate	1	ATGCCTCGCTACCTCCTCTT ned AATCGTTCGTTCACAAACCC	Frary et al. 2005

Frary A, Xu Y, Liu J, Mitchell S, Tedeschi E, Tanksley S (2005) Development of a set of PCR-based anchor markers encompassing the tomato genome and evaluation of their usefulness for genetics and breeding experiments. Theor Appl Genet 111:291312

PCR

Eau	qsp 20µl
Tampon 10X	1X
dNTP	0.8mM
Primers MS131 F	2pmole
Primers MS131 S	1pmole
Primers MS052 F	1pmole
Primers MS052 S	0.6pmole
Primers MS114 F	1pmole
Primers MS114 S	0.6pmole
DNA	50ng
Taq Pol	1u

Programme PCR

	5min	
94°C		
94°C	30sec	
55°C	45sec	} 35cycles
72°C	1min30	
72°C	7min	
15°C		

Dilution : 1/50
3µl de PCR diluée + 8µl de formamide/marqueur de taille GS400HD (7.85/0.15)

Electrophorèse : Séquenceur capillaires ABI3130xl, 50cm, POP7

Genotypage 8 témoins : GeneMapper

	MS131				**MS052**				**MS114**			
	6-fam - bleu				**hex - vert**				**ned - jaune**			
	$[TA]_{21}$				$[AAC]_9$				$[CTT]_7$			
nom	Allele 1	Allele 2	Size 1	Size 2	Allele 1	Allele 2	Size 1	Size 2	Allele 1	Allele 2	Size 1	Size 2
CR031	248		248.42		191		191.42		214		213.92	
CR001	244		244.35		191		191.43		214		213.98	
CR002	240		240.24		191		191.42		214		213.95	
CR003	244		244.31		191		191.43		214		213.95	
CR004	240		240.39		191		190.84		214		214.23	
CR203	228	230	227.9	229.92	191		191.36		212		212.11	
CR206	222		221.87		185		185.01		209		209.27	
CR207	218		217.99		202		201.22		208		208.25	

Triplex3

	nom d'origine	Motif	Espèce d'origine	Chr.	Séquence F et R	Ref biblio
MS112	SSR117	$[TC]_{11}$	Tomate	1	AATTCACCTTTCTTCCGTCG 6-fam GCCCTCGAATCTGGTAGCTT	Frary et al. 2005
MS115	SSR66	$[ATA]_8$	Tomate	2	TGCAACAACTGGATAGGTCG hex TGGATGAAACGGATGTTGAA	Frary et al. 2005
MS122	SSR13	$[CAG]_7$	Tomate	5	GGGTCACATACACTCATACTAAGGA ned CAAATCGCGACATGTGTAAGA	Frary et al. 2005

Frary A, Xu Y, Liu J, Mitchell S, Tedeschi E, Tanksley S (2005) Development of a set of PCR-based anchor markers encompassing the tomato genome and evaluation of their usefulness for genetics and breeding experiments. Theor Appl Genet 111:291312

PCR

Eau	qsp 20µl
Tampon 10X	1X
MgCl2	2mM
dNTP	0.8mM
Primers MS112 F	1.5pmole
Primers MS112 R	0.8pmole
Primers MS115 F	2pmole
Primers MS115 R	1pmole
Primers MS122 F	2.5pmole
Primers MS122 R	1.2pmole
DNA	50ng
Taq Pol	1u

Programme PCR

	5min	
94°C		
94°C	30sec	
55°C	45sec	} 35cycles
72°C	1min30	
72°C	7min	
15°C		

Dilution : 1/50
3µl de PCR diluée + 8µl de formamide/marqueur de taille GS400HD (7.85/0.15)

Electrophorèse : Séquenceur capillaires ABI3130xl, 50cm, POP7

Genotypage 8 témoins : GeneMapper

	MS112				MS115				MS122			
	6-fam - bleu				hex - vert				ned - jaune			
	$[TC]_{11}$				$[ATA]_8$				$[CAG]_7$			
nom	Allele 1	Allele 2	Size 1	Size 2	Allele 1	Allele 2	Size 1	Size 2	Allele 1	Allele 2	Size 1	Size 2
CR031	240		239.68		183		183.42		103		102.44	
CR001	236		235.69		180		180.32		103		102.7	
CR002	240		239.68		183		183.42		103		102.81	
CR003	240		239.76		183		183.45		103		102.58	
CR004	240		239.68		183		183.42		103		102.63	
CR203	234		233.63		180		180.32		96	101	95.89	100.51
CR206	250		249.47		168		167.96		99		99.42	
CR207	232		231.65		170		170.54		97		96.90	

Triplex4

	nom d'origine	Motif	Espèce d'origine	Chr.	Séquence F et R	Ref biblio
MS123	SSR578	$[AAC]_6[ATC]_5$	Tomate	6	ATTCCCAGCACAACCAGACT 6-fam GTTGGTGGATGAAATTTGTG	Frary et al. 2005
MS124	SSR47	$[AT]_{14}$	Tomate	6	TCCTCAAGAAATGAAGCTCTGA hex CCTTGGAGATAACAACCACAA	Frary et al. 2005
MS125	SSR594	$[TCT]_8$	Tomate	8	TTCGTTGAAGAAGATGATGGTC ned CAAAGAGAACAAGCATCCAAGA	Frary et al. 2005

Frary A, Xu Y, Liu J, Mitchell S, Tedeschi E, Tanksley S (2005) Development of a set of PCR-based anchor markers encompassing the tomato genome and evaluation of their usefulness for genetics and breeding experiments. Theor Appl Genet 111:291312

PCR

Eau	qsp 20 µl
Tampon 10X	1X
MgCl2	2 mM
dNTP	0.8mM
Primer MS124 F	4 pmole
Primer MS124 R	2 pmole
Primer MS125 F	2pmole
Primer MS125 R	1pmole
DNA	50ng
Taq Pol	1u

Eau	qsp 20 µl
Tampon 10X	1X
MgCl2	2 mM
dNTP	0.8mM
Primer MS123 F	1pmole
Primer MS123 R	0.6pmole
DNA	50ng
Taq Pol	1u

Programme PCR

94°C	5min	
94°C	30sec	} 35cycles
55°C	45sec	} 35cycles
72°C	1min30	} 35cycles
72°C	7min	
15°C		

94°C	5min	
94°C	30sec	} 35cycles
60°C	45sec	} 35cycles
72°C	1min30	} 35cycles
72°C	7min	
15°C		

Mélange : 4µl PCR 55°C + 1µl PCR 60°C

Dilution : 1/50
3µl de PCR diluée + 8µl de formamide/marqueur de taille GS400HD (7.85/0.15)

Electrophorèse : Séquenceur capillaires ABI3130xl, 50cm, POP7

Genotypage 8 témoins : GeneMapper

	MS123				MS124				MS125			
	6-fam - bleu				hex - vert				ned - jaune			
	$[AAC]_6[ATC]_5$				$[AT]_{14}$				$[TCT]_8$			
nom	Allele 1	Allele 2	Size 1	Size 2	Allele 1	Allele 2	Size 1	Size 2	Allele 1	Allele 2	Size 1	Size 2
CR031	299		298.73		194		194 .29		296		295.68	
CR001	296		295.77		192		191.17		293		292.78	
CR002	296		295.82		194		193.19		293		292.79	
CR003	299		298.81		192		191.06		293		292.71	
CR004	299		298.84		194		193.19		293		292.71	
CR203	296		295.72		224		225.29		287		286.96	
CR206	290		289.6						282		281.49	
CR207	293		292.75		173		172.82		272		271.41	

Triplex5

	nom d'origine	Motif	Espèce d'origine	Chr.	Séquence F et R	Ref biblio
MS118	SSR22	$[AT]_{11}$	Tomate	3	GATCGGCAGTAGGTGCTCTC 6-fam CAAGAAACACCCATATCCGC	Frary et al. 2005
MS127	SSR66327	$[AAT]_7$	Tomate	8	TCAGGATCAGGAGCAGGAGT hex TGGACTTGTTCCATGAACCC	Frary et al. 2005
MS120	SSR593	$[TAC]_7$	Tomate	4	TGGCATGAACAACAACCAAT ned AGGAAGTTGCATTAGGCCAT	Frary et al. 2005

Frary A, Xu Y, Liu J, Mitchell S, Tedeschi E, Tanksley S (2005) Development of a set of PCR-based anchor markers encompassing the tomato genome and evaluation of their usefulness for genetics and breeding experiments. Theor Appl Genet 111:291312

PCR

Eau	qsp 20µl
Tampon 10X	1X
MgCl2	2 mM
dNTP	0.8mM
Primers MS118 F	4 pmole
Primers MS118 R	2 pmole
Primers MS127 F	4 pmole
Primers MS127 R	2 pmole
Primers MS120 F	2.5 pmole
Primers MS120 R	1.2 pmole
DNA	50ng
Taq Pol	1u

Programme PCR

	5min	
94°C		
94°C	30sec	
55°C	45sec	35cycles
72°C	1min30	
72°C	7min	
15°C		

Diution : 1/50
3µl de PCR diluée + 8µl de formamide/marqueur de taille GS400HD (7.85/0.15)

Electrophorèse : Séquenceur capillaires ABI3130xl, 50cm, POP7

Genotypage 8 témoins : GeneMapper

	MS118				**MS127**				**MS120**			
	6-fam - bleu				**hex - vert**				**ned - jaune**			
	$[AT]_{11}$				$[AAT]_7$				$[TAC]_7$			
nom	Allele 1	Allele 2	Size 1	Size 2	Allele 1	Allele 2	Size 1	Size 2	Allele 1	Allele 2	Size 1	Size 2
CR031	211		211.01		148		147.97		299		299.27	
CR001	218		218.01		145		144.88		302		302.22	
CR002	211		210.93		148		147.87		302		302.32	
CR003	218		218.00		148		147.95		299		299.01	
CR004	211		210.99		148		147.86		299		299.01	
CR203	213	216	212.93	215.93	145	148	144.78	147.86	299		299.18	
CR206	206		205.96		170	169.77			287		286.76	
CR207	207		207.13		159	159.13			290		290.24	

Triplex6

	nom d'origine	Motif	Espèce d'origine	Chr.	Séquence F	Ref biblio
MS045	SSR26	$[CGG]_7$	Tomate	2	CGCCTATCGATACCACCACT 6-fam ATTGATCCGTTTGGTTCTGC	SGN
MS053	SSR45	$[AAT]_{14}$	Tomate	7	TGTATCCTGGTGGACCAATG hex TCCAAGTATCAGGCACACCA	SGN
MS062	SSR20	$[GAA]_8$	Tomate	12	GAGGACGACAACAACAACGA ned GACATGCCACTTAGATCCACAA	SGN

Frary A, Xu Y, Liu J, Mitchell S, Tedeschi E, Tanksley S (2005) Development of a set of PCR-based anchor markers encompassing the tomato genome and evaluation of their usefulness for genetics and breeding experiments. Theor Appl Genet 111:291312

PCR

Eau	qsp 20µl
Tampon 10X	1X
MgCl2	2 mM
dNTP	0.8mM
Primers MS045 F	1.25pmole
Primers MS045 R	0.7pmole
Primers MS053 F	2pmole
Primers MS053 R	1pmole
Primers MS062 F	1.5pmole
Primers MS062 R	0.8pmole
DNA	50ng
Taq Pol	1u

Programme PCR

	5min	
94°C		
94°C	30sec	
60°C	45sec	} 35cycles
72°C	1min30	
72°C	7min	
15°C		

Dilution : 1/50
3µl de PCR diluée + 8µl de formamide/marqueur de taille GS400HD (7.85/0.15)

Electrophorèse : Séquenceur capillaires ABI3130xl, 50cm, POP7

Genotypage 8 témoins : GeneMapper

	MS045				MS053				MS062			
	6-fam - bleu				hex - vert				ned - jaune			
	$[CGG]_6$				$[AAT]_{14}$				$[GAA]_8$			
nom	Allele 1	Allele 2	Size 1	Size 2	Allele 1	Allele 2	Size 1	Size 2	Allele 1	Allele 2	Size 1	Size 2
CR031	177		177.21		254		253.60		151		151.41	
CR001	171		170.98		251		250.57		151		151.41	
CR002	177		177.02		251		250.47		157		157.61	
CR003	177		177.23		251		250.52		151		150.98	
CR004	177		177.15		248		247.45		157		157.68	
CR203	171		171.23		254		253.58		148		147.89	
CR206	171		170.89		263		262.47		148		148.43	
CR207	171		170.87		278		277.71		148		148.35	

Triplex7

	nom d'origine	Motif	Espèce d'origine	Chr.	Séquence F	Ref biblio
MS129	SSR70	$[AT]_{20}$	Tomate	9	TTTAGGGTGTCTGTGGGTCC 6-fam GGAGTGCGCAGAGGATAGAG	SGN
MS060	SSR136	$[CAG]_7$	Tomate	11	GAAACCGCCTCTTTCACTTG hex CAGCAATGATTCCAGCGATA	SGN
MS049	SSR188	$[AT]_{11}$	Tomate	4	TGCAGTGAGTCTCGATTTGC ned GGTCTCATTGCAGATAGGGC	SGN

Frary A, Xu Y, Liu J, Mitchell S, Tedeschi E, Tanksley S (2005) Development of a set of PCR-based anchor markers encompassing the tomato genome and evaluation of their usefulness for genetics and breeding experiments. Theor Appl Genet 111:291312

PCR

Eau	qsp 20µl
Tampon 10X	1X
MgCl2	2 mM
dNTP	0.8mM
Primers MS129 F	1.5pmole
Primers MS129 R	0.8pmole
Primers MS060 F	0.75pmole
Primers MS060 R	0.4pmole
Primers MS049 F	3 pmole
Primers MS049 R	1.4 pmole
DNA	50ng
Taq Pol	1u

Programme PCR

94°C	5min	
94°C	30sec	
50°C	45sec	35cycles
72°C	1min30	
72°C	7min	
15°C		

Dilution : 1/50
3µl de PCR diluée + 8µl de formamide/marqueur de taille GS400HD (7.85/0.15)

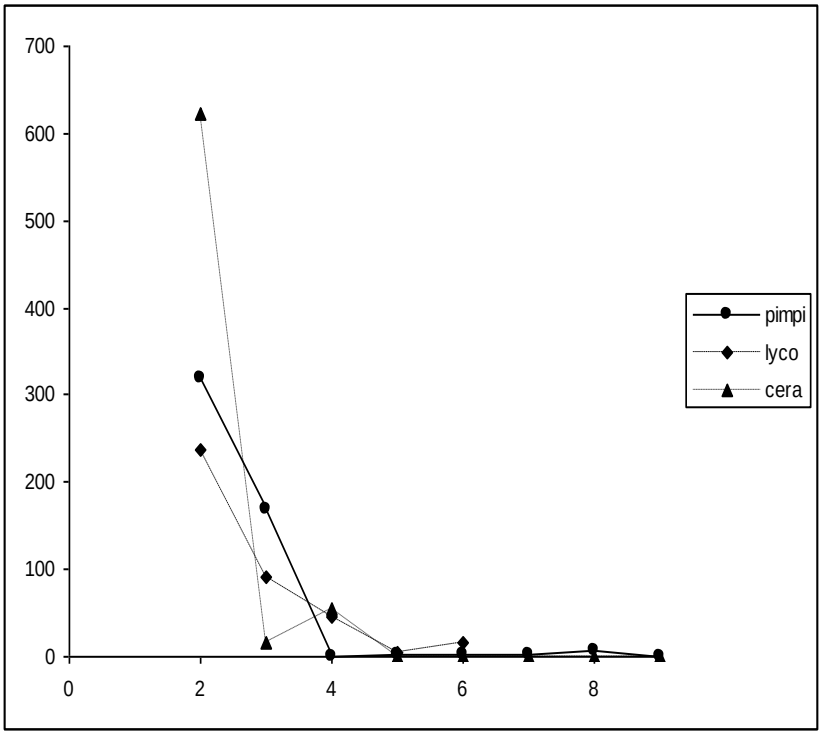
Electrophorèse : Séquenceur capillaires ABI3130xl, 50cm, POP7

Genotypage : GeneMapper

	MS129				MS060				MS049			
	6-fam - bleu				hex - vert				ned - jaune			
	$[AT]_{20}$				$[CAG]_7$				$[AT]_{11}$			
Nom	Allele 1	Allele 2	Size 1	Size 2	Allele 1	Allele 2	Size 1	Size 2	Allele 1	Allele 2	Size 1	Size 2
CR031	117		117.2		148		148.05		136		136.23	
CR001	113		112.92		148		148.44		138		138.12	
CR002	117		117.18		148		148.45		136		136.29	
CR003	117		117.21		148		148.44		134		134.03	
CR004	117		117.22		148		148.44		136		136.21	
CR203	102		102.29		148		148.47		136		136.22	
CR206	123		123.47		151		151.06		290		290.34	
CR207	89		88.37		151		151.15		141	143	142.41	143.46

Annexe 3. Données supplémentaires de l'article : A clarified position for *Solanum lycopersicum* var. *cerasiforme* in the evolutionary history of tomatoes (*solanaceae*)

S1. Détermination du nombre de sous-population optimal pour chaque espèce à partir des résultat de Structure 2.1 corrigés par la méthode d'Evanno et al. (2000).



pimpi : *S. pimpinellifolium*; esc : *S. lycopersicum* var. *esculentum*; cera : *S. lycopersicum var cerasiforme*

S3. Détails de l'échantillonnage en vue de constituer les différentes « core collections ». Les probabilités d'apparatenance à chacune des quatre sous-populations sont définies par Structure 2.1

Accessions *S. lycopersicum* var. *cerasiforme*

individual code	Name	core collection size	Individual memberships for K_{opt}=4			
			popA	popB	popC	popD
CR288	Criollo	8	0.004	0.322	0.48	0.194
CR001	Cervil	8	0.457	0.008	0.032	0.503
CR158	LA 2675	8	0.005	0.003	0.242	0.75
CR056	Wva 700	8	0.062	0.008	0.919	0.011
CR294	Phyra	8	0.398	0.139	0.14	0.324
CR151	LA 2137	8	0.013	0.004	0.972	0.01
CR267	Tiny tim	8	0.013	0.228	0.127	0.632
CR124	LA 1464	8	0.003	0.003	0.024	0.97
CR097	N° 347 Yablochnyi	24	0.006	0.006	0.97	0.018
CR130	LA 0172	24	0.004	0.364	0.526	0.107
CR250	Cherry VFNT	24	0.022	0.276	0.008	0.694

individual code	Name	core collection size	Individual memberships for K_{opt}=4			
			popA	popB	popC	popD
CR249	Cherry Gold	24	0.108	0.518	0.03	0.344
CR118	LA 1420	24	0.005	0.953	0.031	0.01
CR238	PI 129088	24	0.42	0.059	0.418	0.102
CR202	CGN 18399= PI 187002-1	24	0.023	0.961	0.007	0.009
CR199	tomate Richter's	24	0.99	0.004	0.003	0.003
CR106	LA 1204	24	0.148	0.292	0.534	0.026
CR076	N° 135 Green Gage	24	0.018	0.012	0.964	0.006
CR14	Clémentine	24	0.003	0.014	0.01	0.972
CR205	L. pimpinellifolium atypique, site 10 (F300045)	24	0.002	0.992	0.002	0.003
CR028	Plovdiv XXIVa	24	0.004	0.003	0.01	0.984
CR070	N° 2909 Lycopersicon sp.	24	0.98	0.014	0.003	0.003
CR271	Celsior	24	0.085	0.004	0.836	0.076
CR101	N° 884 Alagabotskii	24	0.003	0.004	0.978	0.014
CR287	Cisterno	32	0.003	0.004	0.835	0.158
CR279	Cerise Orange d'Uzès	32	0.24	0.019	0.02	0.721
CR274	Marpha n°2	32	0.003	0.003	0.733	0.261
CR058	Wva 106	32	0.726	0.262	0.005	0.007
CR152	LA 2307	32	0.003	0.016	0.973	0.008
CR122	LA 1456	32	0.004	0.004	0.987	0.005
CR149	LA 2095	32	0.004	0.504	0.387	0.105
CR244	Yellow Pear	32	0.003	0.007	0.944	0.046
CR123	LA 1461	64	0.004	0.138	0.849	0.01
CR117	LA 1388	64	0.003	0.003	0.971	0.023
CR110	LA 1307	64	0.342	0.006	0.603	0.049
CR108	LA 1231	64	0.371	0.016	0.237	0.375
CR098	N° 795 Pescio	64	0.003	0.003	0.981	0.012
CR093	N° 2257 Dikorastushii	64	0.15	0.007	0.084	0.759
CR078	N° 2759 Enano	64	0.004	0.097	0.86	0.04
CR077	N°1565	64	0.003	0.003	0.984	0.009
CR102	N° 739	64	0.006	0.442	0.182	0.37
CR254	Farthest North	64	0.426	0.004	0.176	0.394
CR159	LA 2688	64	0.009	0.38	0.593	0.017
CR079	N° 933	64	0.004	0.004	0.529	0.463
CR296	Poire jaune	64	0.003	0.003	0.344	0.651
CR293	Costa Rica	64	0.006	0.082	0.23	0.682
CR292	8 bis	64	0.223	0.006	0.574	0.196
CR291	Pyriforme	64	0.003	0.003	0.775	0.218
CR284	cerise rose	64	0.15	0.006	0.021	0.823
CR280	Cerise du sud ouest n° 2	64	0.003	0.003	0.964	0.03
CR275	Cerise Ildi	64	0.003	0.003	0.796	0.198
CR153	LA 2308	64	0.004	0.003	0.984	0.009
CR256	Minibel	64	0.151	0.005	0.034	0.81
CR125	LA 1482	64	0.022	0.01	0.85	0.118
CR252	Droplet	64	0.261	0.003	0.012	0.724
CR240	L 285	64	0.196	0.008	0.425	0.371
CR236	PI 365923	64	0.213	0.01	0.084	0.694
CR234	Atom	64	0.251	0.006	0.01	0.734
CR253	Monplaisir	64	0.421	0.006	0.015	0.558
CR155	LA 2402	64	0.003	0.003	0.983	0.01

individual code	Name	core collection size	Individual memberships for K_{opt}=4			
			popA	popB	popC	popD
CR150	LA 2131	64	0.113	0.006	0.873	0.009
CR146	LA 1620	64	0.023	0.008	0.953	0.016
CR145	LA 1543	64	0.004	0.052	0.936	0.008
CR258	Ohmiya Suncherry	64	0.56	0.005	0.018	0.416

Accessions *S. lycopersicum* var.. *esculentum*

individual code	Name	Individual memberships for K_{opt}=4			
		popA	popB	popC	popD
CR002	Levovil	0.231	0.006	0.032	0.731
CR004	M-82	0.064	0.004	0.068	0.864
CR031	Microtom	0.070	0.008	0.337	0.585
CR032	Moneymaker	0.004	0.003	0.012	0.982
CR003	Ferum	0.009	0.010	0.007	0.974
CR317	Heinz 1706	0.003	0.003	0.010	0.983
CR129	LA0147	0.004	0.012	0.971	0.013
CR273	Orange cocktail	0.009	0.281	0.039	0.673
CR094	Vir 1011	0.004	0.003	0.984	0.009
CR133	LA 0409	0.016	0.005	0.932	0.048
CR020	San Marzano	0.003	0.004	0.375	0.618
CR134	LA0466	0.164	0.004	0.711	0.121
CR136	LA0473	0.019	0.003	0.956	0.022
CR321	Edkawy	0.004	0.004	0.020	0.972
CR341	Cra66	0.117	0.098	0.122	0.662
CR354	Stupicke Polni Rane	0.004	0.006	0.223	0.768
CR359	Muchamiel	0.003	0.003	0.019	0.975

Accessions *S. pimpinellifolium*

individual code	Name	Individual memberships for K_{opt}=4			
		popA	popB	popC	popD
CR062	LA1478	0.004	0.535	0.057	0.404
CR068	VIR108	0.247	0.102	0.629	0.022
CR072	VIR2921	0.458	0.008	0.055	0.479
CR075	VIR4156	0.789	0.005	0.077	0.129
CR163	LA0400	0.003	0.959	0.015	0.023
CR164	LA0411	0.156	0.705	0.035	0.104
CR169	LA1371	0.007	0.985	0.004	0.004
CR173	LA1547	0.051	0.943	0.003	0.003
CR186	LA1689	0.456	0.536	0.005	0.003
CR203	LA1589	0.018	0.974	0.004	0.004
CR181	LA1602	0.064	0.874	0.042	0.021

Accessions Wild relatives

individual code	Name	species
CR00X	LA1840	*S. chmielewskii*
CR206	PI247087	*S. habrochaites*
CR207	LA716	*S.pennelii*
CR228	LA1401	*S.chesmaniae*

Annexe 4. Données supplémentaires de l'article : Increases in tomato fruit size and locule number is controlled by two key SNP located near Wuschel.

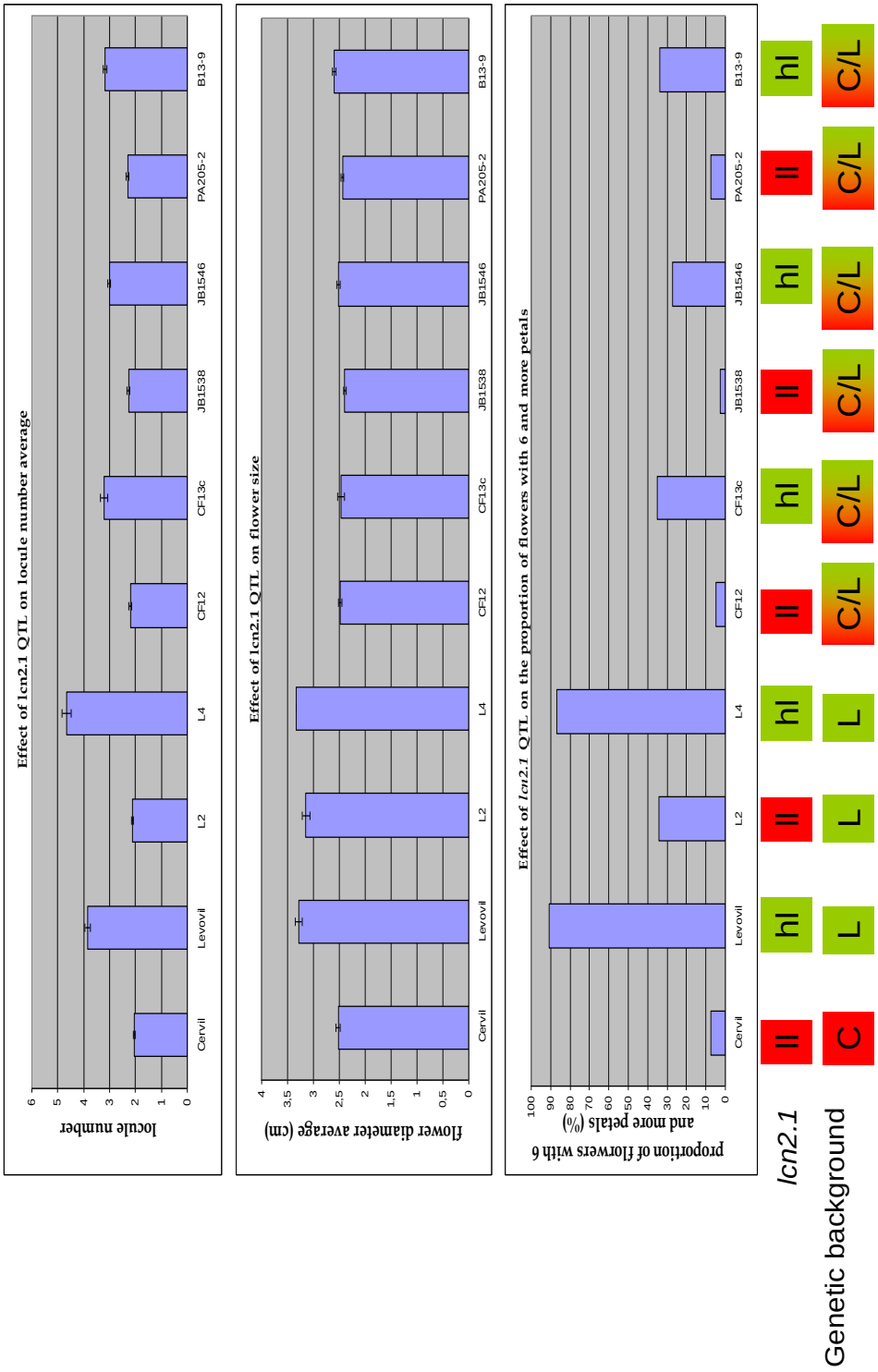


Figure S1: Effect of *lc* QTL on petal number

Cervil and Levovil are the parental lines. L2 and L4 have been obtained by SAM from Levovil; they are both identical to Levovil except for a region of chromosome 4 from Cervil (L4) and the region of the chromosome 2 from Cervil which contains *lcn2.1* (L2). L2 and L4 must be compared to Levovil. CF12 and CF13c are a couple of isogenic lines identical one from the other except for a 30 cM region which contains *lc*. JB1538 and JB1546 are identical to CF12 and CF13c respectivelly but differ for less than 30cM. PA205-2 and B13-9, on the same principle, differ one from the other to a 28kb region containing lcn2.1.
The lines are all homozygous and contain either the low locule allele of lc (ll) from Cervil (C) or the high locule allele of lc (hl) from Levovil (L).

Annexe 5. Données supplémentaires de l'article : Genome admixture of *Solanum lycopersicum* var. *cerasiforme* allows successful association mapping in tomato (*Solanum lycopersicum*), an inbred crop.

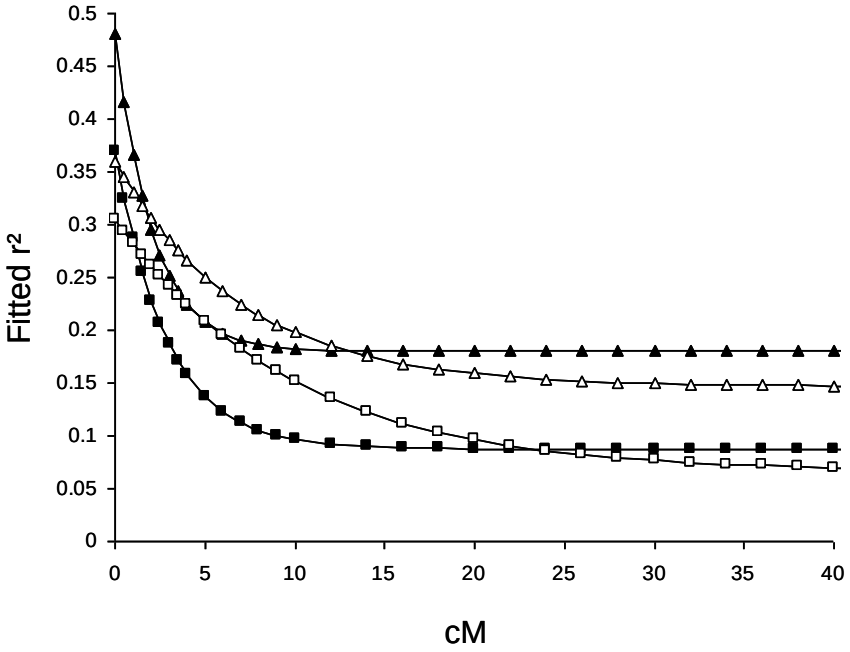


FIGURE S1.−Comparison of different models for analysis of linkage disequilibrium decay over genetic distances. Whole collection and *S. l. cerasiforme* survey strategies are represented by triangle and square, respectively. One polymorphic site per fragment and all sites with MAF>5% survey strategies are represented in white and black, respectively.

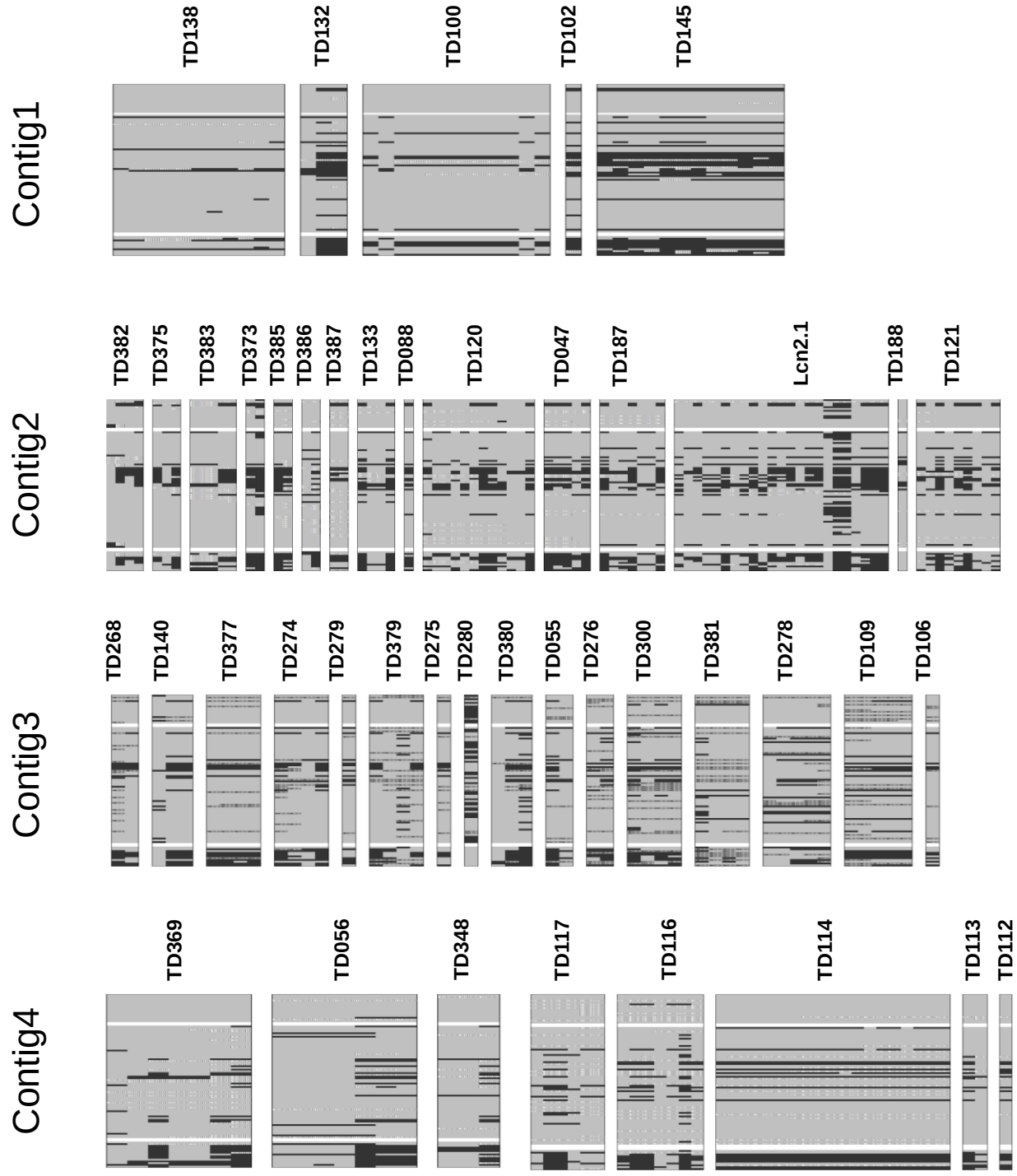


Figure S2.−Graphical haplotypes of 90 accessions for markers located on physical contigs. Rows represent accessions and columns represent polymorphic sites. Fragment are separated by white rows. For each polymorphic site, most frequent allele is represented in light gray and the other allele is represented in black. Data failed are represented in white. The three species *S. l. esculentum*, *S. l. cerasiforme* and *S. pimpinellifolium* are separated by continuous white lines.

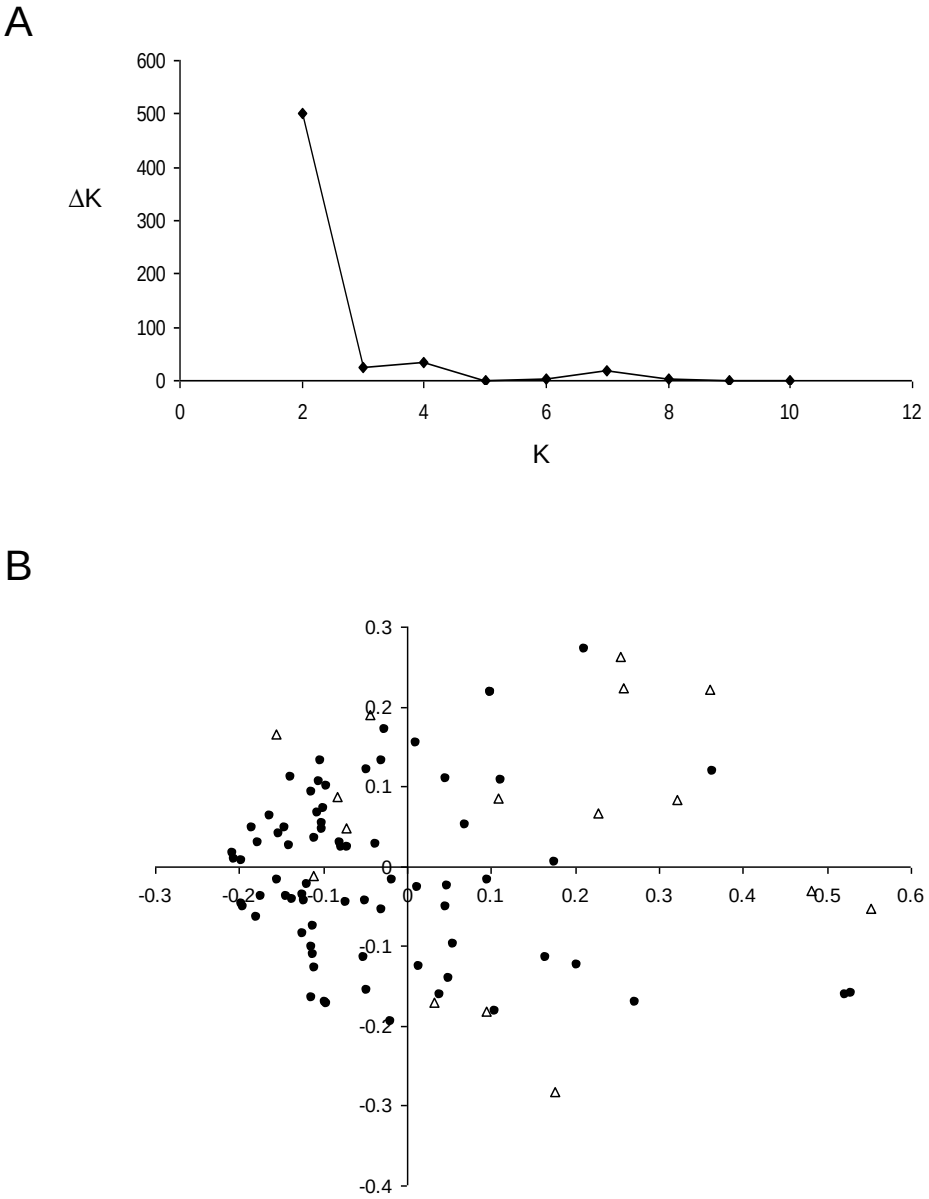


Figure S3.−Genetic structure determination of the 90 accessions of wild and cultivated tomato. (A) Determination of the optimal number of sub-population (K) for 90 accessions following the method of Evanno et al. (2005). The rate of change of the posterior probability of the data given the number of clusters is plotted against K, the number of clusters. ΔK was calculated as |L"(K)|/s[Pr(x|k)] (see Materials and Methods). The first peak (K = 2) corresponds to the optimum number of clusters. (B) Principal coordinate analysis of the 90 accessions based on 20 SSR markers. The two groups identified by Structure software are represented by black square and white triangle.

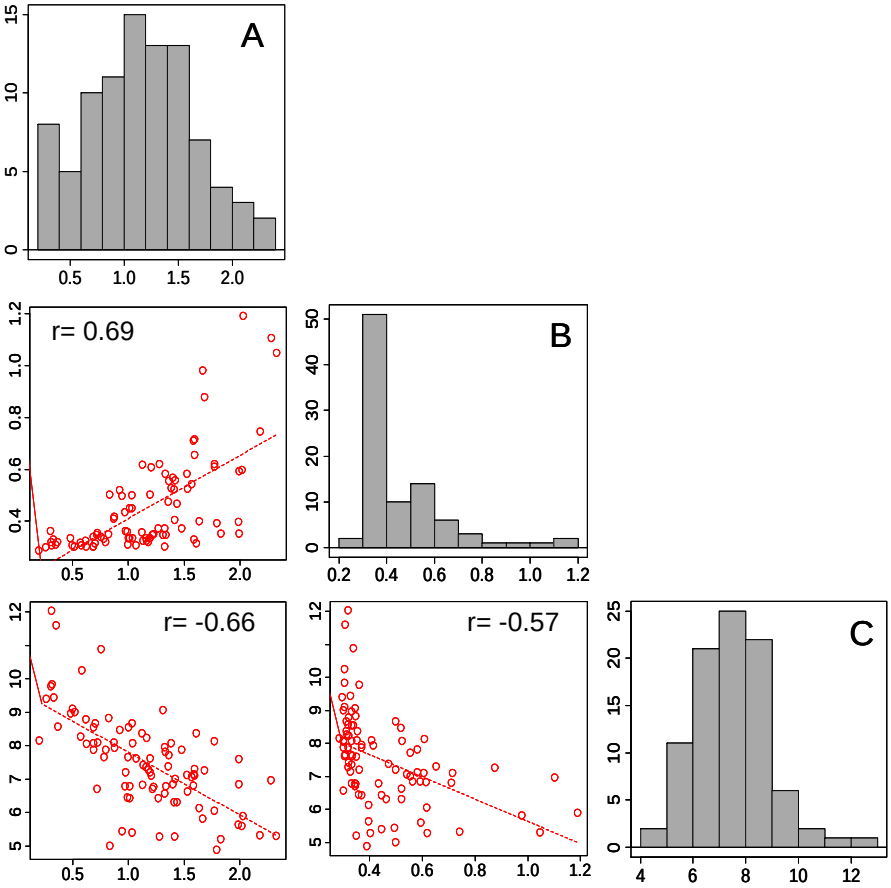


Figure S4.−Distribution and correlation of fruit weight (FW), locule number (LCN) and soluble solid content (SSC) for the 90 accessions. Adjusted mean over two year of experiment for each accessions are used to show histogram distribution of traits for logarithm of Fruit Weight in gram (A), logarithm of Locule Number (B) and Soluble Solid Content in °brix (C). Scatter plot diagrams show correlation between traits and Spearman's rank correlation coefficient are indicated.

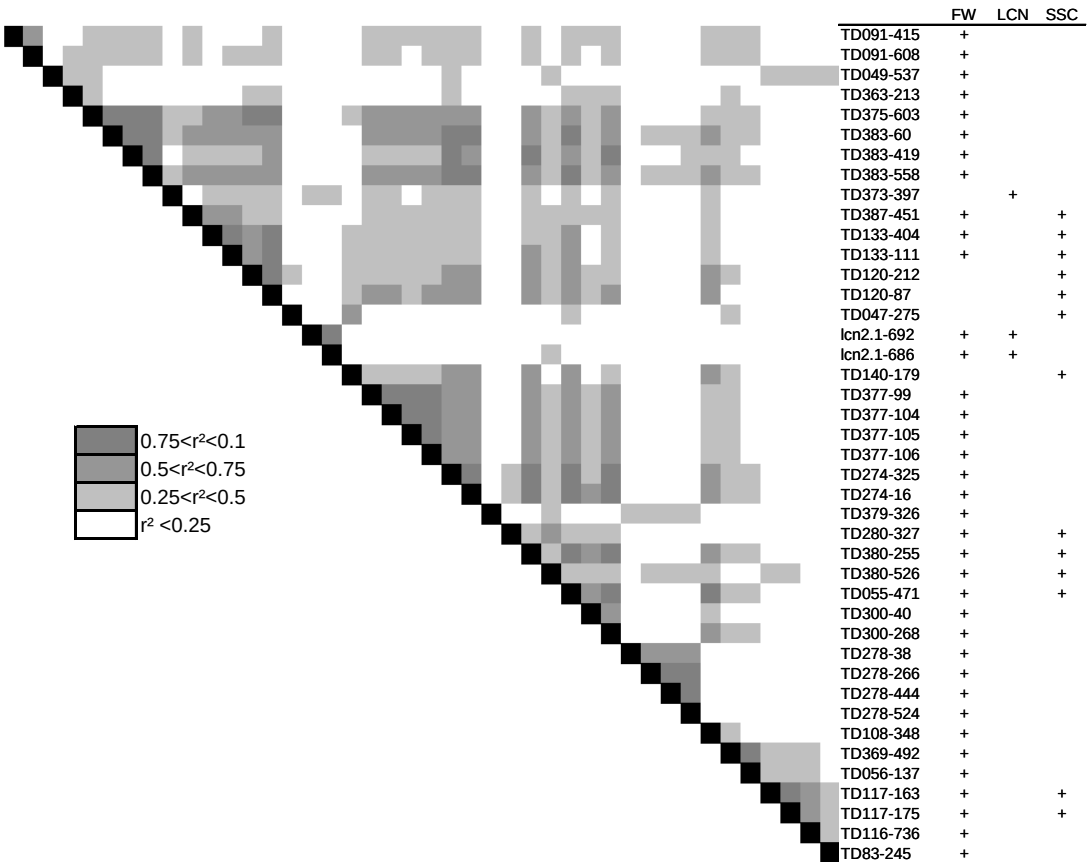


Figure S5.−Matrix of r² between significant polymorphisms and associated traits. The associations and r² are calculated on 90 accessions.

Table S1. Polymorphism information

polymorphism name	allele ref [a]	allele 2	3' seq	5'seq	Frequence of reference allele		
					S. l. cera (N=63)	*S. l. esc* (N=17)	*S. l. pimpi* (N=10)
TD096-184	T	C	CAAGAAATGT	GTTTTTGATA	0.89	0.94	0.40
TD096-205	G	T	TTTTTACAGG	AAGGGTCAAG	0.94	0.94	0.40
TD096-259	G	A	GGAAAATGAG	AGGATTATGA	0.89	0.94	0.40
TD096-313	G	T	AGTATAGAGG	GGGTGGGAGT	0.89	0.94	0.40
TD096-328	G	A	GGGAGTCTGA	GCTGCCAGTG	0.92	0.94	0.40
TD096-340	A	T	CTGCCAGTGT	GTTGAGGATG	0.92	0.94	0.40
TD096-376	A	G	AATCTGCTAC	AATGGGATTG	0.93	0.94	0.40
TD096-592	T	C	ACTTGAAGAT	TGTAAGGAGG	0.90	0.94	0.40
TD270-301	G	A	GCTGTCTACC	AAAATATTTC	0.93	1.00	0.70
TD270-312	A	T	AAAATATTTC	CTCTTGAGGT	0.93	1.00	0.70
TD095-15	G	A	GGTGAGCATC	TGCTCTGTTC	0.92	0.94	0.91
TD095-249	G	C	TGCTACGTAC	AAATGATAGG	0.84	0.88	0.18
TD095-329	G	A	GAATATGCTG	TAGAGGTGTT	0.97	1.00	0.73
TD094-128	C	G	TCCCGAACAT	AATTAACTCT	0.80	0.94	0.55
TD094-131	T	C	CGAACATCAA	TAACTCTCAA	0.80	0.94	0.55
TD094-306	T	C	GCGTTATCTG	ATTTTACTCG	0.94	0.94	1.00
TD094-433	G	A	CGCATTTCGG	TGTGTTGCTG	0.94	1.00	0.90
TD094-447	T	C	GTTGCTGATA	TGTATATGAG	0.86	0.93	0.50
TD094-456	A	G	ATTGTATATG	GAAAGGCGGA	0.95	1.00	0.90
TD094-623	ATA	---	AGGTCGATGG	ATAATAATAA	0.89	0.93	0.20
TD094-626	ATA	---	TCGATGGATA	ATAATAATAA	0.94	1.00	0.60
TD093-148	C	T	AATAGAATCT	TAATAACATT	0.88	0.94	0.09
TD092-450	C	G	ACATTGTATG	TAATTGCATG	0.80	1.00	0.36
TD092-484	A	G	CTTCTAAGGT	TTCAGCGAGA	0.91	1.00	1.00
TD092-523	G	A	TTGTATGGAC	TCTCTTTTGA	0.80	1.00	0.27
TD091-415	A	-	GGACAAAGGT	AAAAAAAAAA	0.86	1.00	0.30
TD091-607	T	C	TTTTGAGTTT	TAATCATTAC	0.80	1.00	0.10
TD150-16	----------	ACAGTTCTTATT	TTTCTGATGA	TTATCCGTGC	0.83	1.00	0.30
TD150-303	C	A	CTATGCATAT	ATTTATTTTA	0.85	1.00	0.30
TD150-337	G	C	AAAAGCAAAA	TGAGTCTTTT	0.85	1.00	0.30
TD150-531	---	TTC	TGATGATGAC	TTCTTAGAAT	0.71	1.00	0.10
TD090-14	A	C	ATTTTCTTGT	TTAATTTATA	0.81	1.00	0.56
TD090-36	-	A	TTTTTTTTAA	GTTGTATTAT	0.75	1.00	0.33
TD090-266	-	T	CTCATTTTTT	CTCTATAGGT	0.75	1.00	0.33

polymorphism name	allele ref [a]	allele 2	3' seq	5'seq	Frequence of reference allele		
					S. l. cera (N=63)	*S. l. esc* (N=17)	*S. l. pimpi* (N=10)
TD090-270	C	T	ATTTTTTCTC	ATAGGTTTGG	0.75	1.00	0.22
TD090-306	T	C	TATTTTGTCA	GTTCCTTAGA	0.92	0.92	0.89
TD090-357	C	T	TTTTTTGATT	AGGTGGTCTG	0.75	1.00	0.22
TD090-625	A	T	TACTTCTCTT	TATCATCCCA	0.75	1.00	0.33
TD345-138	A	C	TCGGCAAGGA	GTTACTCGCT	0.94	1.00	0.82
TD345-195	A	T	CCATTTTTTT	AAAAAAGATG	0.89	1.00	0.82
TD345-208	A	T	AAAAGATGTC	ACTTTATAAT	0.94	1.00	0.80
TD345-241	A	T	TTAATTGTTG	TTGTTGTTTA	0.70	0.93	0.09
TD345-253	T	C	TGTTGTTTAT	CTCTTCATGT	0.94	1.00	0.82
TD345-463	TTTTAA	------	AAAGAGATAC	TTTTAATTTT	0.94	1.00	0.80
TD343-102	-	T	TCTGTTTTTT	CCCTATGTAG	0.71	0.94	0.09
TD343-133	T	-	TATTTTTTTT	CCTGGAATTA	0.86	0.94	0.55
TD343-175	G	T	TTCCTCACTT	CTCTGATGAT	0.91	1.00	1.00
TD339-81	G	A	TGTTGAATTC	TATATTCGAC	0.87	1.00	0.82
TD339-293	A	C	GATTAATTAG	TAATTTCTCC	0.98	1.00	0.90
TD339-308	G	C	TTCTCCTTCA	TTCCAGTATA	0.98	1.00	0.90
TD339-321	C	T	CCAGTATATA	GACTGTTTGA	0.97	1.00	1.00
TD339-358	T	A	CAAAGTTAAT	TGAAGTGAAT	0.89	1.00	0.90
TD339-389	T	C	GGATTTAACA	AATTTTCAGA	0.89	1.00	0.90
TD139-547	C	T	TCTCATTTTT	GTGAGTGAAG	0.75	1.00	0.27
TD049-96	TCAAATTC	--------	GTTAGAGATG	ACATTGTAAC	0.78	1.00	0.45
TD049-339	C	A	GAGCTCATAT	CCATTGATCA	0.78	1.00	0.45
TD049-348	T	C	TCCCATTGAT	ACATACATGC	0.81	1.00	0.64
TD049-445	T	A	GAAATGAAGT	ATCTTGGTGT	0.81	1.00	0.64
TD049-457	A	G	TCTTGGTGTT	AGTCCAGGAA	0.78	1.00	0.36
TD049-528	T	C	TTGACAACCT	CGCGCTTTTG	0.51	1.00	0.09
TD110-8	T	C	TTGTTCT	GAAGATTTGC	0.84	1.00	0.45
TD110-49	T	G	CTGGAGTTGT	GCCTTGCTAG	0.83	1.00	0.45
TD363-126	T	G	GAGTAAATAT	TCATTTTATA	0.83	0.93	0.90
TD363-170	A	T	GATAGACTAA	GGAATAATTG	0.79	1.00	0.30
TD363-213	C	T	TTTGTTAACG	ATAGAATCAA	0.61	0.93	0.11
TD363-241	A	T	TAGGACATGT	AAATGGGAGT	0.82	1.00	0.89
TD363-486	G	A	CATAGGAGTG	GGTTTTTACC	0.83	0.93	0.89
TD363-498	C	A	GTTTTTACCT	GTGCGCACTC	0.83	1.00	0.89
TD363-499	G	T	TTTTTACCTC	TGCGCACTCA	0.83	0.93	0.89
TD363-517	G	C	TCAAAGGGTA	CAGCTGTGGA	0.96	1.00	0.78
TD363-531	C	T	CTGTGGATTT	CCTTGATGTA	0.83	0.93	0.89

polymorphism name	allele ref [a]	allele 2	3' seq	5'seq	Frequence of reference allele		
					S. l. cera (N=63)	*S. l. esc* (N=17)	*S. l. pimpi* (N=10)
TD363-542	T	A	CCTTGATGTA	AAAAA	0.64	0.93	0.11
TD138-28	G	A	TCGACTTTCC	CTGCAGTATT	0.95	1.00	0.82
TD138-39	C	T	CTGCAGTATT	GGACCCATCC	0.95	1.00	0.82
TD138-48	C	A	TCGGACCCAT	CCACATTCAA	0.94	1.00	0.73
TD138-49	C	T	CGGACCCATC	CACATTCAAA	0.94	1.00	0.73
TD138-50	C	T	GGACCCATCC	ACATTCAAAT	0.94	1.00	0.73
TD138-59	A	G	CCACATTCAA	TTGACCGTTA	0.94	1.00	0.73
TD138-61	T	C	ACATTCAAAT	GACCGTTAAT	0.92	1.00	0.73
TD138-62	G	C	CATTCAAATT	ACCGTTAATG	0.94	1.00	0.73
TD138-114	T	C	TATTAGCTGC	GTATCTTTAG	0.92	1.00	0.73
TD138-121	T	G	TGCTGTATCT	TAGGGGATGG	0.92	1.00	0.64
TD138-123	A	T	CTGTATCTTT	GGGGATGGCG	0.92	1.00	0.73
TD265-456	C	T	AATTACAGAT	GACTACTTCC	0.84	1.00	0.82
TD132-122	C	A	GCATCTAATT	CACTCATGAG	0.90	1.00	1.00
TD132-167	G	T	AGGCAAAGAT	TTTGTGGATA	0.65	0.88	0.09
TD132-436	G	A	ACCCTAAAAA	GGGGGAAAGT	0.64	0.87	0.09
TD100-68	A	T	ACTTAATCCT	TGGGAATTAT	0.91	1.00	0.64
TD100-115	T	C	GTTTGCATTT	TTTTTACAAA	0.94	1.00	0.82
TD100-121	C	-	TTTTTTTTTA	AAAAAAGAAT	0.91	1.00	0.64
TD100-172	A	T	ATTCTGAAGT	GATGGTCTGA	0.92	1.00	0.64
TD100-176	---------	TATGACACC	AAGTAGATGG	TCTGATGGAA	0.92	1.00	0.64
TD100-179	C	T	AGTAGATGGT	TGATGGAAAA	0.92	1.00	0.64
TD100-212	A	T	CCATAAATGT	ATATGTGGTG	0.92	1.00	0.64
TD100-213	A	T	CATAAATGTA	TATGTGGTGA	0.92	1.00	0.64
TD100-387	A	G	GCGGCAGGTG	GGAGAGGGGG	0.92	1.00	0.64
TD100-388	G	A	CGGCAGGTGA	GAGAGGGGGG	0.94	1.00	0.82
TD100-569	G	T	TAAATGGGGC	TGCTGATGGC	0.92	1.00	0.64
TD102-263	G	A	CTGGTCTGGC	TTTTCCCAAT	0.72	0.88	0.18
TD145-9	C	G	ATTCATCA	TCCTTTGTTA	0.73	0.88	0.27
TD145-34	T	C	CGCTCTTCTA	TCTACATCTT	0.66	0.88	0.09
TD145-45	A	G	TCTACATCTT	CCTCTGCGAA	0.73	0.88	0.27
TD145-59	T	C	CTGCGAAAGA	TCATTCTCAG	0.73	0.88	0.27
TD145-90	T	A	TAGTTTCTCA	ATGTACATTT	0.66	0.88	0.09
TD145-232	A	G	CTATTACTCA	ATTCGTAAAT	0.66	0.88	0.09
TD145-289	T	C	TGGATCTCAG	AAGGTAAGAT	0.66	0.88	0.09
TD145-317	T	A	AACAACTCTT	TGATGATACT	0.73	0.88	0.27
TD145-328	T	A	TGATGATACT	TTCGAAAGAG	0.72	0.88	0.27

polymorphism name	allele ref [a]	allele 2	3' seq	5'seq	Frequence of reference allele		
					S. l. cera (N=63)	*S. l. esc* (N=17)	*S. l. pimpi* (N=10)
TD145-389	A	G	GAGAGATTAT	GCATAGATGA	0.72	0.88	0.27
TD145-454	A	G	TCATTACTCA	AAGAGTGTTT	0.73	0.88	0.27
TD145-466	A	G	AGAGTGTTTT	ATTAGACCTG	0.72	0.88	0.27
TD382-102	C	G	AGACATGCAT	ACTATACTTG	0.93	0.88	1.00
TD382-172	T	C	ACGTGTTTTG	TGATCTTGTG	0.81	0.86	0.44
TD382-251	T	A	ATACATAGAT	TGCCCTTAAA	0.92	0.88	0.82
TD382-37	G	A	TGCATGGACA	ATATGTCCCT	0.79	0.81	0.64
TD375-98	G	T	CATGTAGTAC	AAGTGTCATA	0.86	0.94	0.70
TD375-386	TTGAGCTAAT	----------	TTTAGCTAAT	GGATTGAGTG	0.94	1.00	0.90
TD375-573	T	C	ACATCTCTGG	GCAATCCTCT	0.78	0.93	0.30
TD383-60	A	G	CCATGTGTTT	GCCATAAAAT	0.92	1.00	0.17
TD383-419	A	G	TACTAGAGAG	GTGTTTTTGT	0.92	1.00	0.33
TD383-558	A	G	AAATTGTATG	ACAAACATTA	0.92	1.00	0.17
TD383-679	T	C	TACAATTTTA	CTTTTAACTT	0.85	0.88	0.60
TD383-684	T	-	TTTTATCTTT	AACTTTAAAA	0.84	0.88	0.60
TD373-140	T	C	CCTACATTCT	AAACCTTTTA	0.77	0.86	0.18
TD373-391	G	T	TGTTCTAATT	GGTTGATTAA	0.42	0.50	0.91
TD385-209	T	A	TCAAACCAAG	AAAGCATCAA	0.73	0.88	0.09
TD385-613	T	C	GAGAATGTTA	ACATGGATGT	0.74	0.86	0.11
TD386-130	C	G	ATTTAATTAA	AAGTAATTTT	0.88	0.83	0.44
TD386-201	A	G	AGAGAAGCAC	GGTCCTCAAT	0.90	0.83	0.44
TD387-339	G	A	ATCTGCTTTG	TATTTCTTGT	0.93	1.00	0.82
TD387-452	C	T	AACTGTCAAA	CATGTGTAGA	0.78	1.00	0.10
TD133-395	A	C	TAGAGTATGC	GAATCCAGGA	0.73	0.87	0.09
TD133-390	T	C	TAAACTAGAG	ATGCAGAATC	0.92	1.00	0.91
TD133-305	T	C	TGGAAGTTAT	GTGTATGATT	0.92	1.00	0.91
TD133-115	----------	CTGCGATTTG	AAAGCCTTTG	TTGGAACATT	0.72	0.88	0.09
TD88-204	G	T	AAAGGAAGAC	CTCCTTATTG	0.78	0.88	0.18
TD120-444	-	T	ATTATGGTGT	TTTTTAAAAA	0.67	0.85	0.25
TD120-445	-	T	TTATGGTGTT	TTTTAAAAA	0.71	0.85	0.25
TD120-418	A	T	TTTTTTTTTT	AATGGTAATT	0.87	1.00	0.67
TD120-382	T	A	CCCCTAAATA	AATACGGATT	0.87	1.00	0.67
TD120-333	A	T	TTTGGTTTTA	TTTTAATCAA	0.93	1.00	0.56
TD120-310	C	T	TTTTAAAAAC	GACTAGATTG	0.87	1.00	0.78
TD120-309	C	T	ATTTTAAAAA	CGACTAGATT	0.89	0.87	0.67
TD120-283	-	T	TCGTTTGGTT	GGTTTGGTTT	0.68	0.88	0.09
TD120-212	T	C	TACTTTTAAA	CGGATAAACC	0.69	0.88	0.18

polymorphism name	allele ref [a]	allele 2	3' seq	5'seq	Frequence of reference allele		
					S. l. cera (N=63)	*S. l. esc* (N=17)	*S. l. pimpi* (N=10)
TD120-172	A	T	TGTTTTTGTC	ATCGATGTAA	0.86	0.94	0.64
TD120-93	G	A	CTTTGGTCGT	TAATGATAAC	0.97	1.00	0.73
TD120-90	C	T	ACACTTTGGT	GTGTAATGAT	0.97	1.00	0.82
TD120-88	G	T	AAACACTTTG	TCGTGTAATG	0.73	0.88	0.09
TD047-573	---	GAT	AACGAACGAT	TATTCAAGA	0.74	0.83	0.09
TD047-571	G	A	CATAACGAAC	ATTATTCAAG	0.74	0.85	0.09
TD047-505	C	T	TATAACGATA	TTATAAAGTT	0.80	0.85	0.27
TD047-435	A	G	TCATGTAAAT	TTTTAAATAA	0.78	0.85	0.18
TD047-274	ATTAAATTAATGAAAGATAA	----------	TTTTAAAATT	ATTAAATTAA	0.92	1.00	0.55
TD047-220	C	T	TTTTTTAATA	GAGGAAATTT	0.73	0.80	0.09
TD187-472	G	A	GTTACCACTT	CCACATATAA	0.70	0.86	0.09
TD187-448	--------	TATAGTAG	AGAATATAGA	TAATATAGTA	0.90	1.00	0.73
TD187-65	-	T	AAATTTACAC	TTTTTTTTAC	0.76	0.86	0.36
TD187-73	-	T	ACTTTTTTTT	ACCTCATTTA	0.76	0.86	0.36
TD187-51	A	C	AAATTTTCTT	ATGAAAATTT	0.92	1.00	0.82
TD187-48	C	T	TTTAAATTTT	TTAATGAAAA	0.90	1.00	0.82
TD187-41	A	G	TTTATATTTT	AATTTTCTTA	0.78	0.86	0.27
lcn2.1-53	A	G	TTAAATTAAT	ATTATTTTAA	0.69	0.88	0.09
lcn2.1-60	T	C	AATAATTATT	TAATTCAAAT	0.69	0.88	0.09
lcn2.1-324	C	T	GTACAAATTA	GTTAACCAGA	0.69	0.88	0.27
lcn2.1-434	T	-	GTTTTTTTTT	GTTTTAAAAA	0.89	1.00	0.64
lcn2.1-686	T	C	TGGCATGATG	TTACTAATTG	0.60	0.44	1.00
lcn2.1-692	A	G	GATGTTTACT	ATTGGACAAT	0.62	0.44	1.00
lcn2.1-744	-	T	ATTTTTTTTT	GGACATATTT	0.19	0.31	0.00
lcn2.1-904	G	A	GGTTTGAAAT	TTGATGTGTT	0.74	0.88	0.40
lcn2.1-939	G	A	ATGTTTTTCA	AATTTTTTTT	0.78	0.88	0.40
lcn2.1-942	T	A	TTTTTCAGAA	TTTTTTTCGT	0.90	1.00	0.70
lcn2.1-964	G	A	TTCCTTGCTT	TTTTATGTGT	0.76	0.88	0.40
lcn2.1-1023	C	T	TACGTATAAT	TAGACAAATA	0.86	0.88	0.70
lcn2.1-1073	A	T	AGTGTGATGG	GATAACGGAT	0.77	0.88	0.40
lcn2.1-1161	G	A	TGATGAAAAT	ACGGATGGAG	0.88	1.00	0.64
lcn2.1-1185	T	-	TGAGCATGAT	GAACGTTATT	0.74	0.88	0.10
lcn2.1-1362	A	G	CAGCCTCATA	TTAAATTACA	0.83	1.00	0.73
lcn2.1-1450	A	-	CAAAATAAAA	TTAGTTTTTC	0.91	1.00	0.90
lcn2.1-1480	A	G	AATTCAAATT	TGTTTAATGT	0.67	0.88	0.09
lcn2.1-1505	A	T	AAATATTTTT	AAAATTTTTT	0.83	1.00	0.82
lcn2.1-1536	G	A	CATATCACGA	AAATATCAGC	0.83	1.00	0.82

polymorphism name	allele ref [a]	allele 2	3' seq	5'seq	Frequence of reference allele		
					S. l. cera (N=63)	*S. l. esc* (N=17)	*S. l. pimpi* (N=10)
lcn2.1-1555	G	A	GCTTAAAATA	TTAATTTCTC	0.78	0.88	0.64
lcn2.1-1565	C	A	GTTAATTTCT	TCAATTTCAA	0.83	1.00	0.73
lcn2.1-1579	A	G	ATTTCAATTT	TTTGTCTTTA	0.83	1.00	0.64
TD188-201	G	C	CTTTTTTCTC	GATGTAAAGA	0.85	1.00	1.00
TD121-384	A	G	TTTAAAAATT	AACTGTTTAG	0.88	0.87	0.90
TD121-286	C	T	GTGCGTCATT	ATTATTAAAT	0.92	1.00	0.60
TD121-278	G	A	CCTTGTTGGT	CGTCATTCAT	0.68	0.88	0.09
TD121-272	G	A	ACAAATCCTT	TTGGTGCGTC	0.93	1.00	0.64
TD121-267	T	A	TCACCACAAA	CCTTGTTGGT	0.77	0.88	0.55
TD121-236	A	G	TTTAGTTAAA	ATGAAAAATT	0.68	0.88	0.09
TD121-218	T	G	GGAAACGAAC	GAGACCTTTT	0.90	0.88	0.82
TD121-196	C	A	AAAGACGATT	TGTATTTAAG	0.92	1.00	0.91
TD121-124	C	G	ATTTTTATAA	TATATTATTG	0.82	0.88	0.64
TD268-431	G	C	TTCATCATAT	GTCGGAGGAC	0.86	1.00	0.27
TD268-161	T	G	TATTTGTTAC	GTTGCAGTTA	0.90	0.94	0.55
TD140-480	C	A	TGATCCCGAT	TGGACCAATA	0.08	0.31	0.00
TD140-180	T	A	TATTGAACCT	AATCTGATAT	0.84	0.93	0.27
TD140-123	C	A	TTGTTGGATA	CAAGTAGGAT	0.84	0.93	0.27
TD377-91	-	T	GGTGGACTTA	TTTTAAGTAT	0.90	1.00	0.20
TD377-96	A	T	GACTTATTTT	AGTATTTTTT	0.90	1.00	0.20
TD377-97	A	T	ACTTATTTTA	GTATTTTTTT	0.90	1.00	0.30
TD377-98	G	T	CTTATTTTAA	ATTTTTTTTT	0.90	1.00	0.20
TD274-325	A	C	GAAAACACTT	CCTTCCTACC	0.83	0.94	0.20
TD274-222	C	G	TAAGTATCTA	AGATTGATAA	0.91	0.94	0.50
TD274-38	C	T	ACATTTATTT	CGTTTTTTTC	0.95	1.00	0.73
TD274-17	A	T	AGGTATTTAT	TTGTCTTAGT	0.79	0.94	0.18
TD279-253	C	A	CATTACTAGG	CAAACAAGAA	0.84	0.94	0.55
TD379-180	-	A	AGAATTAAAA	TCATTATTTT	0.90	0.94	0.11
TD379-219	-	T	ATTATTATTA	TATATATACC	1.00	1.00	0.56
TD379-326	C	T	TATTGCAATT	GAAAAGGTTA	0.80	1.00	0.89
TD379-353	A	G	TTCCTATACT	TTCAAAAAGG	0.90	0.93	0.56
TD379-483	A	C	TTTTCCTAAC	CACTCAACCT	0.91	0.93	0.13
TD275-240	C	T	AAATGCGAGA	CTTTATCTAA	0.89	0.94	0.64
TD280-328	T	C	CTTTCGTTGG	ATCTAGCGTG	0.49	0.80	0.00
TD380-242	C	T	TAACCAACAA	ACCTACAAAC	1.00	1.00	0.82
TD380-256	A	G	TACAAACTAC	AAGTATACTT	0.85	0.94	0.20
TD380-526	T	C	AAAATATAGG	AACTCAGTAA	0.69	0.88	0.00

polymorphism name	allele ref [a]	allele 2	3' seq	5'seq	Frequence of reference allele		
					S. l. cera (N=63)	*S. l. esc* (N=17)	*S. l. pimpi* (N=10)
TD055-469	T	A	GGAAACTGAA	CTTTAGTTTC	0.80	0.93	0.10
TD055-418	A	G	GTACAGCGGG	TATTAAGCGG	0.93	1.00	0.80
TD276-20	C	T	ATTTTCTTTT	TTTTTCAGTT	0.89	1.00	0.38
TD276-97	T	C	GATGCCAAGA	GTGAGTTACA	0.80	1.00	0.13
TD300-41	T	C	CACTCATATT	TAGAAATTTT	0.70	0.92	0.11
TD300-47	A	G	TATTTTAGAA	TTTTAAATTC	0.91	1.00	0.63
TD300-175	----------	TAATAATAATAA	AGAGAAATAA	TAATAATAAT	0.81	0.92	0.40
TD300-257	G	A	GACAAAAGCA	AAAGAGAAAA	0.83	1.00	0.33
TD381-8	C	T	TTGTCAC	CTTTTGTCTT	0.76	0.93	0.20
TD381-516	G	A	TTCAAATCTT	AAAATAAAAA	0.95	0.93	1.00
TD381-548	C	T	GCATAAAGGA	ACAGAATTTT	0.94	0.93	0.50
TD381-568	C	T	TAGAATTTCA	GTATTAATTT	0.94	0.93	1.00
TD278-21	G	A	ATCTTTATGA	TACAATCAGA	0.86	1.00	0.90
TD278-39	A	G	AGAAGGAACG	AGAACTGTAC	0.80	1.00	0.90
TD278-267	G	A	GTTAGCCAAC	TTAACCCTGT	0.78	1.00	0.55
TD278-444	A	G	TAGTTAGGTA	CCAACACTAT	0.77	1.00	0.50
TD278-524	C	T	TGTAGACGCT	GATCTTTCTC	0.79	1.00	0.50
TD109-498	T	C	ACATTAGTGC	TTGCAGAATG	0.82	1.00	0.38
TD109-488	A	G	GAGAATAAGG	CATTAGTGCT	0.84	1.00	0.38
TD109-404	G	A	ATGCTTGAGA	TTGAGACTAC	0.82	0.91	0.40
TD109-329	T	C	AAATATTTCT	CGTTTAAGAT	0.82	0.91	0.40
TD109-216	T	C	CTCATCTATA	CAGCAGCTCT	0.82	0.92	0.40
TD106-219	C	T	TGAGAACCCA	AAAGGTGCTG	0.93	1.00	0.55
TD108-347	C	A	AGCTATGTAA	GAAAAATATG	0.80	0.92	0.18
TD272-104	A	G	ACAAGGCGAT	AGAGAAGTCC	0.92	1.00	0.82
TD272-277	C	T	AATGGTGGTT	GTATTTTAAC	0.77	0.93	0.56
TD130-44	A	T	GATCTTAAGC	TGTCATATGA	0.77	0.93	0.09
TD130-117	T	C	GACAAACATA	TGTAACGAGG	0.82	1.00	0.73
TD130-143	G	C	TTAAGATAGA	ATCAATTCCT	0.95	0.80	1.00
TD130-261	G	T	CATGTAACGA	ACTCAGGGAT	0.98	1.00	0.55
TD130-266	A	G	AACGAGACTC	GGGATTTAAA	0.98	1.00	0.55
TD130-430	C	A	GATCAACATC	ATAAATAAAT	0.77	0.93	0.09
TD316-112	G	C	ATTGATTGTT	TTGATTTTGC	0.75	0.88	0.11
TD316-143	C	T	ATGACCTAGA	AGATCGAGAT	0.75	0.88	0.11
TD316-197	T	A	TTTTTCTTGT	TCTTGTGTAG	0.77	0.88	0.11
TD316-206	A	G	TTTCTTGTGT	GCACAACTTC	0.77	0.88	0.11
TD316-23	A	T	AGTATGTAGT	GTTCTCTTTT	0.75	0.88	0.20

polymorphism name	allele ref [a]	allele 2	3' seq	5'seq	Frequence of reference allele		
					S. l. cera (N=63)	*S. l. esc* (N=17)	*S. l. pimpi* (N=10)
TD316-256	G	A	GTCTAATTTT	ACCTTGATAT	0.77	0.88	0.11
TD316-259	C	-	TAATTTTGAC	TTGATATGGA	0.77	0.88	0.11
TD316-268	G	A	CCTTGATATG	AGTGTTTAAA	0.77	0.88	0.11
TD316-269	A	C	CTTGATATGG	GTGTTTAAAA	0.77	0.88	0.11
TD316-270	G	T	TTGATATGGA	TGTTTAAAAG	0.77	0.88	0.11
TD316-28	T	C	GTAGTAGTTC	CTTTTTTCTA	0.75	0.88	0.20
TD316-272	G	T	GATATGGAGT	TTTAAAAGTA	0.77	0.88	0.11
TD316-274	T	A	TATGGAGTGT	TAAAAGTAAA	0.77	0.88	0.11
TD316-275	T	A	ATGGAGTGTT	AAAAGTAAAG	0.77	0.88	0.11
TD316-364	T	C	AAAAGTTGGA	TGAAGAGTGC	0.77	0.88	0.11
TD316-407	CAA	---	TTTGAAACAA	ACAAATTGAA	0.87	0.94	0.63
TD316-62	T	G	ATCTTTTACT	CGATTTGTTA	0.75	0.88	0.11
TD316-63	C	T	TCTTTTACTT	GATTTGTTAT	0.78	0.88	0.20
TD316-669	T	A	TATGGATGGA	TGAGATTATC	0.87	0.94	0.63
TD316-77	T	-	TTGTTATTTT	AGTCAACCCC	0.75	0.88	0.11
TD305-236	T	C	GTTTAACAGA	TTGTAATGAT	0.90	0.93	0.40
TD305-355	T	A	AGAAGTTATT	TAGTATTGAC	0.81	0.80	0.10
TD86_T7-673	G	A	AGCATCCAGG	AACTCAACAA	0.95	1.00	0.45
TD86_T7-671	G	A	TTAGCATCCA	GGAACTCAAC	0.95	1.00	0.45
TD86_T7-523	A	G	TAAATAATAC	ATTAGATTTA	0.96	1.00	0.55
TD86_T7-486	C	T	ACCTATGTTT	ATTGGATTCA	0.88	1.00	0.73
TD86_T7-434	A	G	ATAATCCTCA	TAAAATCTTA	0.88	1.00	0.73
TD86-312	A	T	AACTCGCCAC	CCCAACTTTA	1.00	1.00	0.64
TD328-95	G	T	CTCTATTTCT	ATTTCAATTG	0.90	0.94	0.50
TD328-318	T	-	AAATGAATGA	GAATTCTCAT	0.91	0.94	0.50
TD304-235	A	G	CTCTACTGTT	TTGGAAGCTT	0.85	1.00	0.50
TD304-322	T	C	TATGGATCAA	CCTTTCGGAT	0.13	0.07	0.91
TD304-453	A	T	ACATAGTATA	TGAAAACAAC	0.85	1.00	0.45
TD304-514	T	G	CAAACCATGT	TTTATTTAAC	0.85	1.00	0.45
TD304-524	C	T	TTTTATTTAA	TAGGGAAGTG	0.88	1.00	0.60
TD369-23	A	-	GAGTGCTTCC	AAAATCACTG	0.93	1.00	0.91
TD369-146	A	G	TGTAGGTACA	ATGAATATTG	0.97	1.00	1.00
TD369-328	T	C	TATGCCTTCC	TAGTGAAACT	0.86	1.00	0.20
TD369-340	A	G	AGTGAAACTG	AAAGTTTCAC	0.96	1.00	1.00
TD369-383	G	A	GAGATGATAG	TTTTTGTTAA	0.96	1.00	1.00
TD369-430	T	A	AAAAACTTCT	TATCTTTCAA	0.87	1.00	0.30
TD369-493	G	A	TTATAAGGAA	TCGTTGAGTA	0.77	1.00	0.11

polymorphism name	allele ref [a]	allele 2	3' seq	5'seq	Frequence of reference allele		
					S. l. cera (N=63)	*S. l. esc* (N=17)	*S. l. pimpi* (N=10)
TD056-28	T	C	TTCATTTGAT	TGAAAACGAA	0.94	1.00	0.82
TD056-64	-	A	AGAATATTGT	ACGATTATAA	0.94	1.00	0.82
TD056-78	-	T	TTATAATTAG	TAAGCTTAAT	0.94	1.00	0.82
TD056-80	A	T	TATAATTAGT	AGCTTAATTT	0.94	1.00	0.82
TD056-134	C	T	CTGACCCTCA	TCTCTTTTCT	0.68	0.93	0.09
TD056-155	A	T	TGGCATTATT	TCCAGAAAAG	0.74	0.93	0.27
TD056-252	AC	--	AAAAAACAAT	ACTCATTTAC	0.82	0.93	0.27
TD348-71	A	C	TCTTTCAAAG	CTTCCTCTGT	0.97	1.00	0.73
TD348-72	-	T	CTTTCAAAGA	CTTCCTCTGT	0.97	1.00	0.73
TD348-300	C	T	GTTAATTTTC	TTATTGAGTT	0.81	1.00	0.22
TD117-101	C	T	AAGATTTTTC	GTTGATCTAT	0.95	1.00	0.82
TD117-164	T	C	ACTGATTTAT	ATCCTAACCA	0.72	1.00	0.09
TD117-176	T	A	TCCTAACCAT	TTATGGTATG	0.72	1.00	0.09
TD117-219	G	A	GGAGATTCTT	AATTGCTTTT	0.93	1.00	0.91
TD117-399	C	T	AGTTGGGTCA	GTTATGTTTA	0.94	1.00	0.80
TD117-420	G	A	TTGACTCAAA	TCTTCTTTGA	0.94	1.00	0.80
TD117-422	C	A	GACTCAAAGT	TTCTTTGACT	0.94	1.00	0.80
TD117-570	A	C	GATGTACAGT	GCTTATATTC	0.83	1.00	0.13
TD117-623	AC	--	TTGATTTATG	ACACTTTCCA	0.87	1.00	0.30
TD117-672	G	A	AATATGATGT	CGTGTCAAA	0.79	1.00	0.10
TD117-699	C	A	GCCAGTGACG	AGCATACTTC	0.79	1.00	0.10
TD116-66	C	G	GTGATTTGAA	GTAGGAGGAG	0.89	1.00	0.40
TD116-260	G	T	TGGGGCTTTT	GCATCACAAT	0.81	0.91	0.10
TD116-283	C	A	TAATTTTCTA	TAAAATAATA	0.77	0.91	0.10
TD116-355	G	A	TTTTTTCCGT	TCTGTGATTT	0.98	1.00	0.90
TD116-393	G	C	TAAAGATCTC	TCTCTTTCGT	0.94	0.91	0.80
TD116-707	A	G	GTGTCTTACT	CAAGATTCCC	0.59	0.90	0.00
TD116-745	C	T	TCCTCGGATA	GCTACTACTT	0.88	1.00	0.50
TD114-102	TTAGGAGTAACAATTG	----------	TGATTGATTA	TTAGGAGTAA	0.85	1.00	0.36
TD114-151	G	A	AGCAAACAAT	CAAGGAAAGA	0.84	1.00	0.30
TD114-166	C	G	GAAAGAATTA	AGTTGTAATT	0.84	1.00	0.30
TD114-203	G	T	TTTTGCCAGT	CTATTGCCTC	0.84	1.00	0.30
TD114-207	T	A	GCCAGTGCTA	TGCCTCTTAA	0.84	1.00	0.30
TD114-217	----------	ATTCTTTAAGAGAC	TGCCTCTTAA	AGATCATGGG	0.85	1.00	0.30
TD114-259	G	T	GTTAATCTTT	GTATTTTAAT	0.84	1.00	0.30
TD114-359	G	A	GAACCCGTAT	TGCAATGCTA	0.83	1.00	0.30
TD114-373	A	G	AATGCTAGAT	CACCTATAAT	0.83	1.00	0.30

polymorphism name	allele ref [a]	allele 2	3' seq	5'seq	Frequence of reference allele		
					S. l. cera (N=63)	*S. l. esc* (N=17)	*S. l. pimpi* (N=10)
TD114-392	C	T	ATAGCATGAA	ATCTATATAA	0.83	1.00	0.30
TD114-439	C	T	AATGGTAAAC	TGTAGTTTAC	0.83	1.00	0.30
TD114-464	A	G	TGTAGTGTTA	TGTTGCTTAG	0.85	1.00	0.30
TD114-507	C	T	ATCTTAGGTT	GTTGATAATA	0.83	1.00	0.30
TD114-551	T	C	TGCACCTTAA	GCAGAAATAA	0.86	1.00	0.30
TD114-598	G	T	GCCAAAAGTG	GTCGCGTGCC	0.82	1.00	0.30
TD114-604	G	A	AGTGGGTCGC	TGCCTTTAAT	0.82	1.00	0.30
TD114-613	A	C	CGTGCCTTTA	TGATTTTGTT	0.86	1.00	0.30
TD114-635	T	A	ATTATAGATC	ATGTGGCTTC	0.82	1.00	0.30
TD114-638	G	C	ATAGATCTAT	TGGCTTCTGT	0.82	1.00	0.30
TD114-648	T	C	GTGGCTTCTG	TTCGTAGTGG	0.82	1.00	0.30
TD113-132	G	T	AGTTCCTCTG	TTTTTTTACT	0.84	1.00	0.18
TD113-230	AGTCTTTCC	---------	CTTTGCTTTC	TCTTCCTAGA	0.97	1.00	0.64
TD112-525	G	T	AATTGATTAG	TAGAGTTTTG	0.85	1.00	0.36
TD083-222	C	A	GCCGATTCGA	TCAGTCGAAA	0.82	1.00	0.64
TD083-246	G	T	CCTTTTCAGT	GCTGCTTCCA	0.48	0.81	0.00
TD083-277	---	CCA	TCCTCCACCG	CCACCACCAC	0.92	1.00	0.64
TD083-366	G	C	CCTGATTTTG	GCAAAACTCA	0.97	1.00	0.64
TD083-404	G	T	CAAGGTATTT	TGCCGTTTAG	0.81	0.92	0.09
TD083-666	G	A	TGAATGGGAG	AAGAAACCGC	0.85	1.00	0.27
TD083-685	C	T	GCCTTTGTTT	ATTCTCTCTT	0.81	0.92	0.09
TD018-103	C	T	GAAGCCCTTT	AAAGTCGTTG	0.67	0.92	0.14
TD018-611	T	C	CAAACGCCAA	TAAGGAGGAT	0.78	1.00	0.14

Table S2. Identification of the 81 fragment sequenced on 90 individuals and located on chromosome 2

fragment name	F-primer sequence R-primer sequence	Putative Function [a]	Unigene	genetic position on chr. 2 [d]
TD018	CCGCCGCTCTTTCTTTGCT TTCATGACTCCAGCTGGTC	pyrophosphate-fructose-6-phosphate 1-phosphotransferase beta subunit, putative	SGN-U583285	75.7cM (Expen2000)
TD047	GTTAACGTGGTGTAGGTGCA AAGGTTGCAGGTACCTCTTGA	promoter of homeodomain transcription factor (WUSCHEL), putative	SGN-U589573	88.0cM (Expen2000)
TD049	ACGGTCCTAATTGCTAACGCA CTTGGGCCATAATGTAATTGTCT	lactoylglutathione lyase / glyoxalase I, putative	SGN-U578074	72cM (Expen2000)
TD055	AGTTTGAAGCTTCGGTTCTCC TACCTACAGTAGGTGGGTGT	ovate protein	SGN-U582169	89.5cM (Expen2000)
TD056	GATTGCGCATTGAGATGCT CGGGGGCAGATACATAGTGA	5' region of fw2.2 gene (Frag 4 : Nesbitt & Tanksley 2002)	no	116cM (Expen2000)
TD083	TAGTGCCGGATCCGCTATG TATTAGTCCCAGCCTTTGCA	sucrose-responsive element binding factor, myb family transcription factor, putative	SGN-U569474	133cM (Expen2000)
TD085	CCAACTTCAACTTGTTTGGGA AGGGACTTCCATAATCATATC	homology with ubiquitin interaction motif-containing protein / LIM domain-containing protein	SGN-U584698	114cM (Expen2000)
TD086	GACCAGAGCGTGCTTCTTGA TACCATTCTGGGAGCGGTAT	mitochondrial processing peptidase beta subunit, putative	SGN-U580309	103cM (Expen2000)
TD088	TAGATGGAGTGGGTATTGTTGA CAAAGTCGGTCTTAACACGACA	-	no	86cM (Expen2000)
TD090	ATGGATATGGTAATTGGAGGA CATGAGTTGAGCTTCATGACT	transcription factor, myb-like protein, putative	SGN-U574375	67cM (Expen2000)
TD091	TCTTGGCATTCGCACAGGAA CCTGGGAGTTATGGGATCTT	expressed protein similar to UV-B and ozone similarly regulated protein	SGN-U580690	54cM (Expen2000)
TD092	GTAGAGTGTGGGAATATGGAG CCACTAAGCAAAGCTAACCA	-	no	46cM (Expen2000)
TD093	TTGGGAGAGGACGAAGGA CTCATCATAAGGCTCTTGCT	expressed protein similar to Glutathione S-transferase	SGN-U583311	36cM (Expen2000)
TD094	TTAACGTACTCGTTGCGTGC ATTGGAAATCCCAACAGCCA	CHP-rich zinc finger protein, putative	SGN-U583902	27cM (Expen2000)
TD095	GAAGAACATGAGAAGCAGCC AGTTCCTACCCACAAGTATCA	expressed protein	SGN-U583803	16cM (Expen2000)
TD096	CAAACTACCCCCAGGTCCA GTATCAATCTCGTCTCGGAGT	cinnamic acid 4-hydroxylase (C4H), putative	SGN-U581122	5cM (Expen2000)
TD098	CCAAGGCAGAGATAAACGTG CCATGAGGTTCCTACACATC	Rho-GTPase-activating protein-related, putative	SGN-U580990	82cM (Expen2000)

fragment name	F-primer sequence R-primer sequence	Putative Function [a]	Unigene	genetic position on chr. 2 [d]
TD100	ATCTCTCTGAGGGTTCAAGACAGG TATATCAGCTCCATACTTCTTTGC	expressed protein	SGN-U567423	83.2cM (Expen2000)
TD102	TCTGAAGAAGCTGAAGCAAGTAGAGC TGCCAACTGACGAGCATAAGCTGC	expressed protein	SGN-U565307	83.4cM (Expen2000)
TD106	CTGGCATGGGATGTAGTGC GATGCAGAAATGTTCGAAGGC	CER6; very-long-chain fatty acid condensing enzyme (CUT1 homology)	SGN-U566767	89cM (Expen2000)
TD108	GTGAATTGTCCGGTTCTCGT AATGCTCATCCCTTGTTTGC	expressed protein	SGN-U575206	90cM (Expen2000)
TD109	GTCTACCCTGCTGCACAAC CTGTTGAACTGTGGTCTCCA	RAD23; ubiquitin family protein	SGN-U569990	90.1cM (Expen2000)
TD110	CACCTCCCAGATCACAATG CATGTGTGGTACATGCTCTG	chlorophyll a/b binding protein CP24 10B, putative; homology to peroxidase 42	SGN-U577555	72.0cM (Expen2000)
TD111	CAGTGGTGCTGAAGATGTTG GTAAGCCCTTTAGAGCTCTC	asr; Adenylyl sulfate reductase (homology to APR1)	SGN-U580331	NA
TD112	GCGAGGATAACGGTGAGAAG TGCCTTTGGAGACTCCTTGT	protein kinase, putative	SGN-U563886	120cM (Expen2000)
TD113	TTCATTGATTTCTCCGCTGC TGAACCACCAAGACGGGA	fruc3, fructokinase 3	SGN-U570882	120cM (Expen2000)
TD114	CTGTAAAGAGGGTGCCTAC CGGTACTTGGTTCAAACCTG	expressed protein	SGN-U573852	120cM (Expen2000)
TD116	CCTGGAATTCGTGCCTTGC GATACCACTAAGTACAGCCTC	oligopeptidase A, putative	SGN-U565226	120cM (Expen2000)
TD117	GACTCTTCTTTGGTGCTGC CCAACTGCTTCGCTTCTCT	expressed protein	SGN-U598856	120cM (Expen2000)
TD125	TCATGAGCAACCTGCATATG TAGGCTCAATGTCATTGATCAC	cyclin family protein, similar to cyclin D3.1 protein	SGN-U583476	2-K(CL) (ILs pennellii)
TD120	ACCCAACACTCTAGCCCAACT ATGGCCGTAATTTTCGTAATCATCA	-	no	physical map Contig 2
TD121	GTAGTACGTATAGAATGGGTTGT TAATGATGCGGCACTTGCTCA	-	no	physical map Contig 2
TD129	TGAATTTTGGGAAGTCTGGTT TGCAACTCTACCTCTTTTTCAGC	nuclear matrix constituent protein, putative	SGN-U584279	126cM (Expen2000)
TD130	GGTTTTGACTTGACATGAAG CTTATTCTCCAAACAATTGC	-	no	95cM (Expen2000)
TD132	CGATCGTGCATACTCTCGTG GCTTCTACTGATGATCCTAC	-	no	81,2cM (Expen2000)
TD133	TTGGGCGACCACGCTGAATC TTACCCACATCAGGACCTTGCC	peptide methionine sulfoxide reductase, putative	SGN-U576565	83,1cM (Expen2000)

fragment name	F-primer sequence R-primer sequence	Putative Function [a]	Unigene	genetic position on chr. 2 [d]
TD137	ACCTAGAGAGGACCTTCCAGAGCCC AGGAATTCAGTGCCTTCAATGCAG	hydantoin utilization protein-related, putative	SGN-U562909	29cM (Expen2000)
TD138	TTTTCACCAGTAGGAACATTCAAGGTAC CAGGATAATAAACCATCATGCCACAA	Z15141; chitinase (endo-), acidic, 26 kD	SGN-U581507	79cM (Expen2000)
TD139	ATCCTCTGCCCTTTCTTTCC CAATTGCAGGGGATATGCTT	transport protein, ATPase, gamma chain, chloroplast, putative	SGN-U581255	71cM (Expen2000)
TD140	TTCAGGAATGGCATTTGCAAGTGTG ACCATTGAATACAGCATCTGGTCGAAC	aldose 1-epimerase family protein, similar to apospory-associated protein C	SGN-U585203	87.5cM (Expen2000)
TD145	GCCAATTTCTAGCGAACATG TCTGTTTCTTCCAACACTAG	histidine kinase family, putative	SGN-U603568	83.30cM (Expen2000)
TD150	GACTCCAGAAGCGATTCTGT CTTCGGCAACTCCTCTTAGC	Cnr; squamosa promoter-binding protein-like 3	SGN-U576708	60;0cM (Expen2000)
TD187	GCGGAGCTTGAAATCAGTAAC AAGGCACGCATGAGATGATAA	-	no	physical map Contig 2
TD188	AAACTCACTTACCACTTTCC CTCCTAGTAAGTCTCGATTC	-	no	physical map Contig 2
TD265	TGCTCCAAAGTGTGCTCATC AGCGGTACCTTTCTCCTGGT	er60; ethylene-responsive catalase	SGN-U578479	81cM (Expen2000)
TD268	ATTGGCAAATGAGCTTGCTT TTGCAAAAAGAACGGTTTCC	expressed protein	SGN-U585443	88.5cM (Expen2000)
TD270	AATGCACATGCGTCACAAAT TTGCACCCTCTCATCAACAG	chaperonin, putative (response to oxidative stress)	SGN-U576925	11.0cM (Expen2000)
TD272	CTACCTGGATCGCAATGGTT TTAGGTGCAACAGCATCTCG	peptidyl-prolyl cis-trans isomerase, putative (protein folding)	SGN-U573199	90.5cM (Expen2000)
TD274	GCTAAATCGAATGCCTGAGC GGAAACCGCCAAAACTATCA	integral membrane protein / sugar transporter family protein, putative	SGN-U575594	physical map Contig 3
TD275	GAGGGTGAGCGATTTATGGA GTCCAGGGATCACAGCATCT	succinate dehydrogenase flavoprotein, putative (nergy pathways; carbohydrate metabolism; citric acid cycle; highly expressed in fruit)	SGN-U580353	physical map Contig 3
TD276	AGTTGACGTGTGGCTTACCC CAGGCTTTTTCTCCTTGACG	SIN-like family protein, putative (transcription)	SGN-U572730	physical map Contig 3
TD278	GCCGAACATGAGAAGGAGAG CAGCTAACCATGACGAGCAA	flavodoxin family protein, putative	SGN-U576263	physical map Contig 3
TD279	AGCTCCTTACAGAGGCAGCA ACCCAAGGGACAGCCTAGTT	contig_ovate16	no	physical map Contig 3
TD280	ACATCCAAGCATGGGCTAAT TGGGCACACAATGCTTAGAA	contig_ovate22	no	physical map Contig 3

fragment name	F-primer sequence R-primer sequence	Putative Function [a]	Unigene	genetic position on chr. 2 [d]
TD300	TGGATAGCACGTGAAATGGT AATGGAAATCCAGGATCAGC	contig_ovate35	no	physical map Contig 3
TD304	GTTCATTCTGGGGATGGGT TGCAGCTATCCTTGCTTTG	expressed protein	SGN-U586574	113.0cM (Expen2000)
TD305	TGGTGAATGGAGAAATGCAG TGATGCCCACTTACACAAGC	expressed protein	SGN-U569081	99.5cM (Expen2000)
TD316	GATGCTGCCTTATTTGCTC CCATCTCAGGGTGTTTTGT	cellular repressor of E1A-stimulated genes (CREG) family, putative	SGN-U578677	97.30cM (Expen2000)
TD328	CCGTTGGTTGGATATTGCTT AAAAGGCACCCCAAAAGAGT	regulator of chromosome condensation (RCC1) family protein, putative	SGN-U565338	111.80cM (Expen2000)
TD339	CTCATCTTCAACTTCCCTTCC CATCAACCACTGAGCCAAC	inositol monophosphatase family protein, putative	SGN-U572028	69.80cM (Expen2000)
TD343	TCCGCCGTATCTAACCTATC CTGTCCAGTAGTAGCATCCC	transducin / WD-40 repeat protein family, putative	SGN-U565169	69.70cM (Expen2000)
TD345	GAAGTGGAAGACCCACAAA CCACTAGAGCCTCCATGTATC	NADPH quinone oxidoreductase-like protein, putative	SGN-U579777	68.50cM (Expen2000)
TD348	ATTCGCCAGAAATGGATCAG TGGTTGCAACACAATCATCA	expressed protein	SGN-U595227	118.50cM (Expen2000)
TD350	GAAAGGAAGCAACCCAATC GCTTAATCCTCGACCAGACA	expressed protein	SGN-U563682	78.50cM (Expen2000)
TD356	TATGTGGGCAACAAGTCAGC CAAAAAGGAGACCGAACCAA	pyridoxal kinase, putative	SGN-U580571	119.50cM (Expen2000)
TD363	ACCCGTTCAGTCTCACATTTC CCAATGCTATCCACCTTATCC	ribosomal protein L15 family protein, putative	SGN-U583446	76.00cM (Expen2000)
TD369	TCCTGAGGACATTGGACACA TGGCAGAAACCTCCATTCTT	weak homology with nodulation protein-related	SGN-U570126	physical map Contig 4
TD373	CAAGCAGCCAAGATCTGTCA TCCCATCTTCAAACCTGGTC	expressed protein	SGN-U581635	physical map Contig 2
TD374	AAGAGGAGAAGGCCCAGAAG CTTTCTGTGTCCGAGGAAGC	expressed protein	SGN-U563261	physical map Contig 2
TD375	CGCGGTACACCGTCTTTTAT TTCACATTTTCTGGCCTTCC	plastidic fructose-bisphosphate aldolase (photosynthesis; Calvin cycle; carbon fixation)	SGN-U580022	physical map Contig 3
TD376	AAGGGCCTTCAGATGAGGTT CCGATTGCCTCTCTTAGTGC	vesicle tethering family protein, putative	SGN-U582526	physical map Contig 3
TD377	CAAGACGATGCGAAAGATGA CAGCATTCATGGAATCATGC	-	no	physical map Contig 3
TD379	TTAAAAGATGGGGCATGAGG ACGTCAAACTGGACCAGACC	-	no	physical map Contig 3

fragment name	F-primer sequence R-primer sequence	Putative Function [a]	Unigene	genetic position on chr. 2 [d]
TD380	GCCTTGGAAACTCACGAAAG GCGACAATATTTCGGGCTTA	chromatin remodeling complex subunit	no	physical map Contig 3
TD381	TTTGTTCCCCTGCGTAAGAG GGGTATTTTAGGCCCTCGTC	-	no	physical map Contig 3
TD382	GCACGCCACGACAGTTACTA ACGTTTTCTGCGCGAGTTAT	homology with retrotransposon Tork11	SGN-U594026	physical map Contig 2
TD383	CTCCGTCCCTAGTTGTCCAC CAGGCCATAATCCAAATGGT	acs8; 1-aminocyclopropane-1-carboxylate synthase	SGN-U565888	physical map Contig 2
TD384	CTGCAAGGGCTAGTTCAAGG CGGGAGTGAGGTGTTTGAAT	putative receptor-like protein kinase gene	SGN-U603238	physical map Contig 2
TD385	AACAAAAGCACCACCAAAGG AAAGGAGAGGCTCCGAGTTC	-	no	physical map Contig 2
TD386	TTAACAAGGGCGTGCACATA CCCGTGCAATACCTTGATCT	-	no	physical map Contig 2
TD387	GAAAATGCAGGAGGAAACCA ATGTGAATCCCGATAGCAACA	-	no	physical map Contig 2

[a] Putative functions of genes are given according to annotation of unigene (http://solgenomics.net/) or manual annotation.
[b] BAC overgo indicates sequence identity with BAC sequences available on genbank.
[c] Name of marker located on the reference map are indicated when available.
[d] Genetic distances are available from the Expen2000 reference map (http://solgenomics.net/), from *S. pennellii* introgression lines (Ils) map.

Table S 3. Information on 90 accessions used in the association study

Accession Number	Accession Name	Species [a]	traits [c]			STRUCTURE results [c]	
			FW (g)	LCN	SSC (°brix)	pop1	pop2
CR001	Cervil	*S. l. cera*	5.80	2.18	10.88	0.393	0.607
CR002	Levovil	*S. l. esc*	109.13	3.91	6.83	0.188	0.812
CR003	Ferum	*S. l. esc*	109.66	2.24	7.59	0.092	0.908
CR004	M-82	*S. l. esc*	62.60	2.45	4.88	0.04	0.96
CR014	Clémentine	*S. l. cera*	5.40	2.25	8.08	0.012	0.988
CR020	San Marzano	*S. l. esc*	70.00	2.25	5.20	0.01	0.99
CR028	Plovdiv XXIVa	*S. l. cera*	41.58	2.05	8.35	0.012	0.988
CR031	Microtom	*S. l. esc*	6.91	3.17	5.00	0.017	0.983
CR032	Moneymaker	*S. l. esc*	99.54	2.49	5.63	0.008	0.992
CR056	Wva 700	*S. l. cera*	4.25	2.10	8.77	0.015	0.985
CR058	Wva 106	*S. l. cera*	9.69	2.72	6.78	0.99	0.01
CR062	LA 1478	*S. pimpi*	2.05	2.30	9.77	0.459	0.541
CR068	N° 108 Red Currant	*S. pimpi*	2.00	1.98	9.38	0.363	0.637
CR070	N° 2909 Lycopersicon sp.	*S. l. cera*	5.12	2.05	8.65	0.989	0.011
CR072	N° 2921 Lyc. Pimpinellifolium	*S. pimpi*	2.15	2.02	9.82	0.438	0.562
CR075	N° 4156 Blumen Strauss	*S. pimpi*	1.66	1.93	8.13	0.791	0.209
CR076	N° 135 Green Gage	*S. l. cera*	39.90	2.13	7.13	0.157	0.843
CR077	N°1565	*S. l. cera*	10.19	2.28	6.43	0.008	0.992
CR078	N° 2759 Enano	*S. l. cera*	34.84	3.33	6.62	0.135	0.865
CR079	N° 933	*S. l. cera*	33.64	3.82	7.12	0.009	0.991
CR093	N° 2257 Dikorastushii...	*S. l. cera*	23.55	3.58	7.70	0.058	0.942
CR094	N° 1011 Srednei Velichiny	*S. l. esc*	23.30	2.98	7.37	0.008	0.992
CR097	N° 347 Yablochnyi	*S. l. cera*	30.49	2.35	7.87	0.022	0.978
CR098	N° 795 Pescio	*S. l. cera*	22.11	3.82	7.80	0.015	0.985
CR101	N° 884 Alagabotskii	*S. l. cera*	24.49	3.35	8.07	0.02	0.98
CR102	N° 739	*S. l. cera*	54.13	7.52	7.25	0.412	0.588
CR106	LA 1025	*S. l. cera*	15.08	2.10	8.22	0.478	0.522
CR108	LA 1231	*S. l. cera*	4.99	2.18	8.53	0.346	0.654
CR110	LA 1307	*S. l. cera*	14.75	2.15	7.35	0.286	0.714
CR117	LA 1388	*S. l. cera*	16.89	4.05	7.10	0.009	0.991
CR118	LA 1420	*S. l. cera*	39.68	5.13	6.80	0.962	0.038

Accession Number	Accession Name	Species [a]	traits [c]			STRUCTURE results [c]	
			FW (g)	LCN	SSC (°brix)	pop1	pop2
CR122	LA 1456	*S. l. cera*	4.96	2.00	8.07	0.01	0.99
CR123	LA 1461	*S. l. cera*	3.93	2.02	10.23	0.332	0.668
CR124	LA 1464	*S. l. cera*	3.25	2.02	9.10	0.008	0.992
CR125	LA 1482	*S. l. cera*	9.66	2.30	7.20	0.023	0.977
CR129	LA 0147	*S. l. esc*	116.77	3.94	5.58	0.072	0.928
CR130	LA 0172	*S. l. cera*	37.68	3.48	7.05	0.398	0.602
CR133	LA 0409	*S. l. esc*	116.72	15.48	5.88	0.011	0.989
CR134	LA 0466	*S. l. esc*	208.89	12.71	6.95	0.039	0.961
CR136	LA 0473	*S. l. esc*	49.89	9.53	5.80	0.009	0.991
CR145	LA 1543	*S. l. cera*	11.27	2.15	8.07	0.053	0.947
CR149	LA 2095	*S. l. cera*	26.90	3.60	7.00	0.419	0.581
CR150	LA 2131	*S. l. cera*	40.36	4.51	7.28	0.025	0.975
CR152	LA 2307	*S. l. cera*	26.00	3.33	6.30	0.38	0.62
CR153	LA 2308	*S. l. cera*	27.70	2.92	6.30	0.01	0.99
CR155	LA 2402	*S. l. cera*	6.77	2.23	8.82	0.009	0.991
CR156	LA 2619	*S. l. cera*	13.77	4.13	6.82	0.93	0.07
CR158	LA 2675	*S. l. cera*	4.99	2.00	7.87	0.009	0.991
CR159	LA 2688	*S. l. cera*	4.34	2.00	8.03	0.467	0.533
CR163	LA 0400	*S. pimpi*	2.10	2.08	12.02	0.973	0.027
CR164	LA 0411	*S. pimpi*	3.14	2.15	8.95	0.884	0.116
CR169	LA 1371	*S. pimpi*	2.30	2.03	11.58	0.988	0.012
CR173	LA 1547	*S. pimpi*	3.42	2.00	8.98	0.992	0.008
CR186	LA 1689	*S. pimpi*	2.20	2.13	9.43	0.992	0.008
CR199	tomate Richter's	*S. l. cera*	3.82	2.07	8.27	0.988	0.012
CR202	CGN 18399	*S. l. cera*	6.45	2.08	7.87	0.985	0.015
CR203	LA 1589	*S. pimpi*	2.40	2.08	8.57	0.991	0.009
CR205	L. pimpinellifolium atypique, site 10 (F300045)	*S. l. cera*	10.42	2.15	6.78	0.994	0.006
CR234	Atom	*S. l. cera*	26.50	2.53	5.27	0.186	0.814
CR236	PI 365923	*S. l. cera*	15.32	2.08	7.27	0.057	0.943
CR238	PI 129088	*S. l. cera*	12.33	3.15	8.65	0.46	0.54
CR240	L 285	*S. l. cera*	15.96	2.16	7.60	0.158	0.842
CR244	Yellow Pear	*S. l. cera*	19.05	2.35	6.42	0.018	0.982

Accession Number	Accession Name	Species [a]	traits [c]			STRUCTURE results [c]	
			FW (g)	LCN	SSC (°brix)	pop1	pop2
CR249	Cherry Gold	*S. l. cera*	7.53	2.56	8.09	0.71	0.29
CR250	Cherry VFNT	*S. l. cera*	21.60	2.00	6.57	0.306	0.694
CR252	Droplet	*S. l. cera*	16.71	2.22	6.70	0.205	0.795
CR253	Monplaisir	*S. l. cera*	22.49	2.22	6.77	0.275	0.725
CR254	Farthest North	*S. l. cera*	8.91	3.13	5.43	0.375	0.625
CR256	Minibel	*S. l. cera*	19.45	4.17	5.27	0.029	0.971
CR258	Ohmiya Suncherry	*S. l. cera*	13.82	2.10	7.40	0.44	0.56
CR267	Tiny tim	*S. l. cera*	11.08	2.80	5.38	0.292	0.708
CR271	Celsior	*S. l. cera*	12.12	2.02	7.60	0.013	0.987
CR273	Orange Cocktail	*S. l. esc*	60.71	4.07	8.12	0.247	0.753
CR274	Marpha n°2	*S. l. cera*	8.54	3.30	8.47	0.009	0.991
CR275	Cerise Ildi	*S. l. cera*	7.65	2.60	7.92	0.008	0.992
CR279	Cerise Orange d'Uzès	*S. l. cera*	13.82	2.27	8.37	0.21	0.79
CR280	Cerise du sud ouest n° 2	*S. l. cera*	10.22	2.15	8.53	0.008	0.992
CR284	cerise rose	*S. l. cera*	10.45	2.80	6.42	0.104	0.896
CR287	Cisterno	*S. l. cera*	21.55	2.35	7.95	0.011	0.989
CR288	Criollo	*S. l. cera*	26.11	3.67	6.83	0.337	0.663
CR291	Pyriforme	*S. l. cera*	10.04	2.03	7.65	0.008	0.992
CR292	8 bis	*S. l. cera*	20.65	2.22	9.05	0.138	0.862
CR293	Costa Rica	*S. l. cera*	15.87	3.17	7.20	0.019	0.981
CR294	Phyra	*S. l. cera*	5.25	2.22	6.70	0.546	0.454
CR296	Poire jaune	*S. l. cera*	16.94	2.23	6.75	0.007	0.993
CR317	Heinz 1706	*S. l. esc*	43.70	2.50	6.12	0.008	0.992
CR321	Edkawy	*S. l. esc*	224.30	11.18	5.30	0.01	0.99
CR341	Cra 66	*S. l. esc*	40.45	5.18	7.10	0.186	0.814
CR354	Stupicke Polni Rane	*S. l. esc*	61.21	4.16	6.05	0.024	0.976
CR359	Muchamiel	*S. l. esc*	172.94	5.54	5.32	0.008	0.992

[a] Accessions are part of *S. l. cerasiforme* (*S. l. cera*), *S. l. esculentum* (*S. l. esc*) or *S. pimpinellifolium* (*S. pimpi*).

[b] Values for fruit weight (FW), locule number (LCN) and soluble solid content (SSC) are adjusted mean from two years of experiment.

[c] STRUCTURE software results are probability of membership in each subpopulation and are based on SSR markers used in Ranc *et al.* 2008

Résumé
Chez la tomate, l'amélioration pour la qualité du fruit est rendue difficile par la multiplicité et la complexité des caractères. La cartographie de QTL a permis la caractérisation génétique de ces caractères. L'objectif est maintenant d'identifier les gènes sous-jacents aux QTL. Nous avons utilisé la cartographie par déséquilibre de liaison (DL) dans ce but. Pour éviter les fausses associations entre les caractères et les polymorphismes moléculaires, la structure génétique a été prise en compte dans l'analyse. La tomate cultivée montre un faible niveau de diversité génétique, ce qui réduit la résolution de cartographie. Le génome de la tomate de type cerise (*S. lycopersicum* var. *cerasiforme*) est décrit comme une mosaïque entre celui de la tomate cultivée et de l'ancêtre sauvage. Ce mélange devrait augmenter la résolution des études d'association. Nous avons utilisé une « core collection » focalisée sur des accessions de type cerise pour valider la région génomique contenant un QTL pour le nombre de loges. Deux mutations sont associées avec le caractère. Ces deux SNP ont évolué différemment du reste du chromosome 2, en subissant une sélection balancée qui témoigne de l'augmentation de la diversité morphologique lors de la domestication. L'étude d'association, focalisée sur le chromosome 2, a permis d'analyser l'étendue du DL en fonction de la distance génétique et physique. Des associations entre des polymorphismes et les phénotypes étudiés ont été détectés avec des méthodes prenant en compte la structure génétique. Nous avons montré l'intérêt d'utiliser la structure en mosaïque du génome des accessions de type cerise pour surmonter les limitations de résolution dans les analyses d'associations chez une espèce cultivée autogame. Nous avons validé des QTL identifiés précédemment et nous avons trouvé des associations avec de nouveaux QTL et de nouveaux gènes candidats. Un modèle d'évolution incluant un goulet d'étranglement et des flux de gènes entre compartiment sauvage et cultivé de tomate est aussi présenté.

Mots clés : tomate, qualité du fruit, ressources génétiques, déséquilibre de liaison, génétique d'association, diversité moléculaire

Abstract
In Tomato (*Solanum lycopersicum*), breeding for fruit quality is difficult due to the multiplicity and complexity of the traits. QTL mapping has allowed the genetic characterization of these traits. One of the challenges is now to identify the genes underlying these QTLs. Following this aim, we used linkage-disequilibrium (LD) mapping. To avoid hazardous associations between traits and polymorphisms, the genetic structure has to be taken into account for LD mapping. Cultivated tomato showed low genetic diversity reducing mapping resolution. Cherry type tomato (*S. lycopersicum* var. *cerasiforme*) genome is described to be admixture between cultivated tomato and its wild ancestor. Such admixture may increase resolution of association mapping. We used a core collection focused on cherry type accessions to validate a candidate gene for a fruit locule-number QTL. We found that two single nucleotide polymorphisms (SNP) were highly associated with the trait. These two SNP evolved differently from the rest of the chromosome 2. They underwent a balanced selection which testifies a selection for fruit morphology diversity by human. Association mapping, focused on whole chromosome 2, allowed us to assess the extent of linkage disequilibrium over genetic and physical distances. Associations of polymorphisms with phenotypes were detected with structured association methods. We thus showed efficiency of genome admixture to overcome the low-resolution limitation of association mapping for an inbred crop. We validated previously identified QTLs and found associations with new QTLs and new candidate genes. An evolutionary model including bottleneck and gene flow between wild and domesticated forms of tomato is also presented.

Key words: tomato, fruit-quality traits, genetic resources, linkage disequilibrium, association mapping, molecular diversity