Mémoire de fin d'étude

Présenté pour l'obtention du diplôme « Ingénieur en Horticulture »

Spécialité « Amélioration des plantes et ingénierie végétale méditerranéennes et tropicales »

# Evaluation d'une stratégie de sélection génomique dans 3 dispositifs expérimentaux chez la tomate

**Par Coralie Picard**

**Organisme d'accueil : INRA Avignon – Unité GAFL**

**Année de soutenance : 2015**

Mémoire de fin d'étude

Présenté pour l'obtention du diplôme « Ingénieur en Horticulture »

Spécialité « Amélioration des plantes et ingénierie végétale méditerranéennes et tropicales »

# Evaluation d'une stratégie de sélection génomique dans 3 dispositifs expérimentaux chez la tomate

**Par Coralie Picard**

**Mémoire préparé sous la direction de :**

- Christopher Sauvage
- Mathilde Causse
- Jacques David

**Présenté le : 17/09/2015**

**Devant le jury :**

- Jacques David
- Laurence Moreau
- Muriel Tavaud

**Organisme d'accueil : INRA Avignon – unité GAFL, UR1052**

Génétique et Amélioration des Fruits et Légumes

Domaine Saint Maurice

67, allée des chênes

CS60094, 84143

Montfavet cedex

# Remerciements

J'adresse mes remerciements aux personnes qui m'ont aidée dans la réalisation de ce mémoire ainsi que celles qui m'ont soutenue durant mon cursus scolaire.

En premier lieu, je souhaite remercier Christopher Sauvage pour son encadrement et sa disponibilité tout au long de ce stage, ainsi que pour la confiance qu'il m'a accordée en me laissant travailler en autonomie. De plus, la version finale de ce mémoire a bénéficié de sa lecture très attentive et de ses remarques précieuses.

Je remercie également Mathilde Causse et toute l'équipe QualiTom qui m'ont accueillie chaleureusement et qui m'ont permis de travailler dans un cadre agréable.

Merci également à tous les stagiaires de l'unité GAFL pour leur participation à une bonne ambiance de travail ainsi que pour les nombreux bons moments passés ensemble. Entre autres Nawel, Mariem, Kévin, Margaux, Maha, François-Xavier etc.

Je remercie les membres du jury de prendre le temps de lire ce mémoire et de se déplacer pour la soutenance.

Un grand merci à tous les enseignants et responsables de la formation APIMET de Montpellier SupAgro et d'Agrocampus Ouest centre d'Angers, qui m'ont encadrée et aidée à progresser dans la voie de la recherche. Je pense notamment à Jacques David, Jean-Luc Regnard, Isabel Martin Grande et Marie-Claude Asseray.

Je tiens également à remercier mes camarades d'Angers et de Montpellier sans qui, ces 6 années d'études n'auraient pas été les mêmes. Merci notamment à Alizée, Marie, Laurène, Pauline, Valentin, Clément, Jonathan, Agathe, Margaux, Benjamin etc.

Enfin, je remercie les membres de ma famille pour leur soutien dans mes efforts et plus personnellement, je remercie William pour son soutien sans failles et ses encouragements.

# Table des matières

# Glossaire

**BLUP :** *Best linear unbiased prediction* – Meilleur prédicteur linéaire non-biaisé

**BV :** *Breeding value* – Valeur génétique additive

**DL :** Déséquilibre de liaison

**EBV :** *Estimated BV* – BV estimée

**G-BLUP :** Méthode BLUP utilisant la matrice d'apparentement G

**GEBV :** *Genomic estimated BV* – Estimation génomique de la BV

**GWAS :** *Genome Wide Association Study*

**$H^2$ :** Héritabilité phénotypique

**HSD :** *Honest significant difference*

**IBD :** *Identity by descent* – Identité par descendance.

**Interaction G×E :** Interaction entre la génétique et l'environnement

**QTL :** *Quantitative trait locus.*

**MAGIC :** *Multi-parent Advanced Generation InterCross*

**MAF :** *Minor Allele Frequency*

**PE :** Population d'entrainement

**PV :** Population de validation

**REML :** *Restricted maximum likelihood* – Maximum de vraisemblance restreint

**RIL :** *Recombinant Inbred Line*

**SAM :** Sélection assistée par marqueurs

**SG :** Sélection génomique

**SNP :** Marqueur « *Single-nucleotide polymorphism* »

**TBV :** *True breeding value* – BV réelle

# Liste des illustrations

## Figures :

## Tableaux :

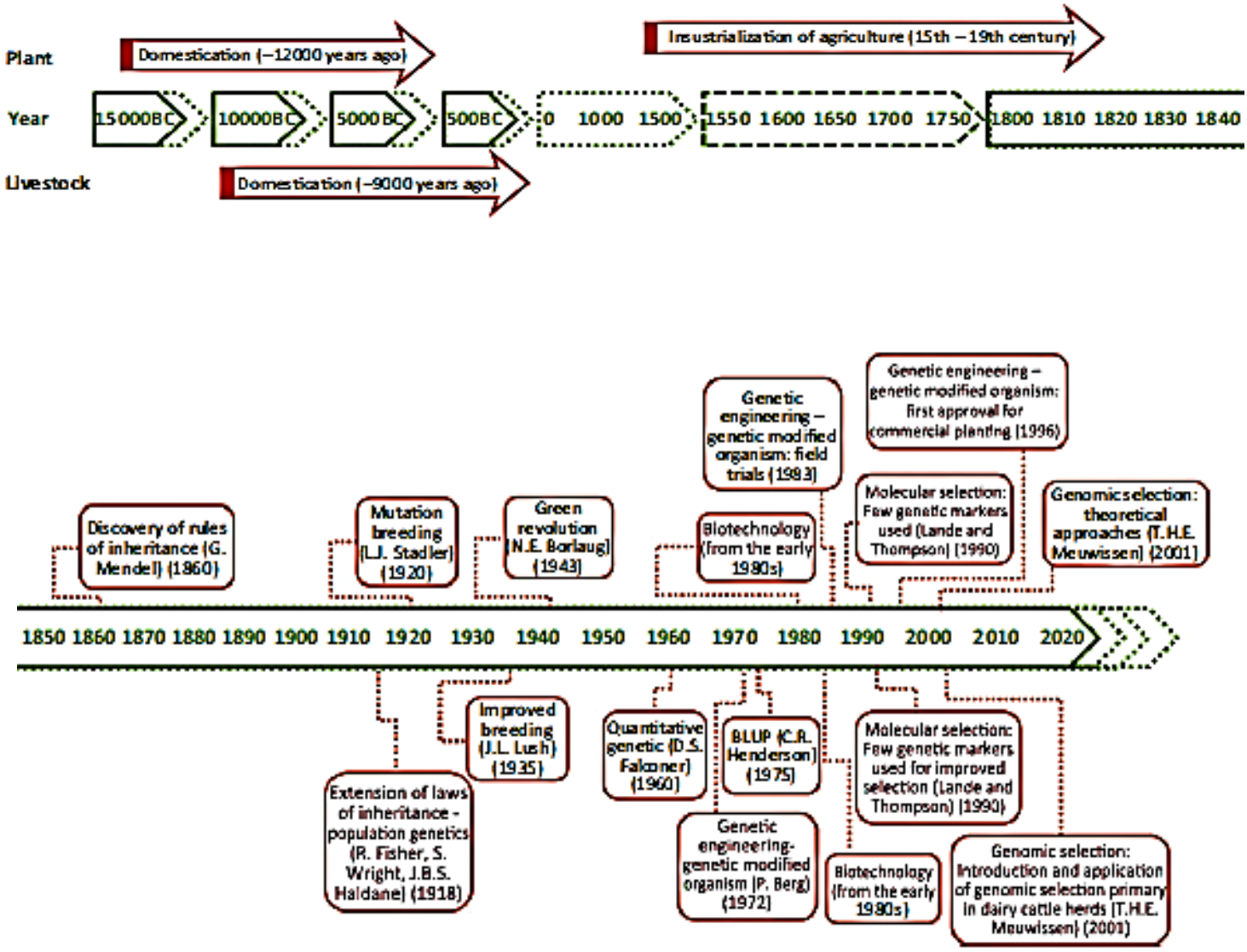



**Figure 1 : Etapes importantes de l'évolution des techniques de sélection qui ont permis l'augmentation de la productivité de l'agriculture (Jonas, E. & de Koning, D.-J., 2013)**

# Introduction

L'Homme participe à la création de nouvelles variétés de plantes depuis leur domestication. Aujourd'hui, l'amélioration des plantes intègre les progrès de la connaissance afin d'orienter la sélection.

Dans le contexte du réchauffement climatique, du changement de la règlementation (ex : plan Ecophyto) et d'une augmentation de la population mondiale, la création de variétés de plus en plus productives, résistantes aux maladies et s'adaptant au changement climatique devient une nécessité. Pour cela, les techniques de création variétale évoluent sans cesse (figure 1). De plus, la sélection récurrente classique, qui consiste à sélectionner de générations en générations les meilleurs individus, demande des étapes de phénotypage très coûteuses. De nos jours, la détection de 'quantitative trait loci' (QTLs) et la génétique d'association (GWA) sont très utilisées pour identifier des régions du génome qui participent à la variation d'un caractère phénotypique d'intérêt agronomique de la plante. Cela permet d'orienter les cycles de sélection grâce à la sélection assistée par marqueurs (SAM). Bien que ces techniques aient déjà fait leurs preuves, elles restent néanmoins coûteuses et sont difficilement utilisables pour des caractères polygéniques.

Dans l'objectif de poursuivre l'amélioration via l'augmentation du gain génétique, une nouvelle approche est apparue depuis quelques années : la sélection génomique (SG). Elle consiste à prédire le phénotype d'un individu en fonction de son génotype. La SG permet d'une part de diminuer le coût et le temps de la sélection en réduisant les étapes de phénotypage et de travailler sur des caractères polygéniques tels que la résistance à la sécheresse ou à certaines maladies d'autre part.

La sélection génomique est aujourd'hui largement utilisée chez les bovins laitiers où elle a démontré son intérêt avec l'augmentation du gain génétique pour des caractères de qualité du lait par exemple. Désormais, elle commence à faire son apparition dans le domaine du végétal avec des études qui sont menées sur des espèces économiquement importantes notamment des céréales telles que le blé, le riz, le seigle, le maïs, la pomme, etc. Cependant, chez d'autres espèces telles que les potagères, dont la tomate, aucune étude n'a encore été publiée.

Dans ce contexte, l'objectif de ce travail est d'étudier l'effet d'un ensemble de paramètres mathématiques sur la précision de prédiction de la sélection génomique pour des phénotypes liés à la qualité chez la tomate. Pour cela, une synthèse bibliographique explicitant les avancées de la sélection génomique sera présentée en première partie ; dans un second temps, nous présenterons le matériel et les méthodes utilisées pour répondre à la problématique. Une troisième partie présentera les résultats de notre étude portant sur l'effet des paramètres étudiés. Enfin, dans une dernière partie, une discussion des résultats permettra de positionner cette étude parmi différents travaux de recherche et proposer des perspectives quant à la place de la sélection génomique dans un schéma de sélection.

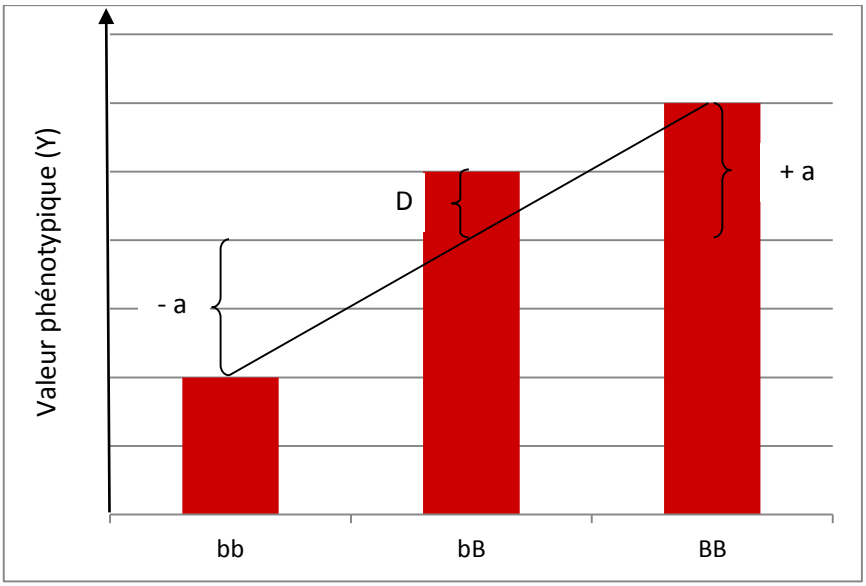



**Figure 2 : Valeurs biologiques moyennes des phénotypes ($Y_{BB}$, $Y_{Bb}$ et $Y_{bb}$) en fonction du génotype (bb, bB ou BB), a : la demi-différence entre $Y_{bb}$ et $Y_{BB}$, a= ($Y_{BB}$-$Y_{bb}$)/2, D : paramètre lié à la dominance de l'allèle B sur l'allèle b**

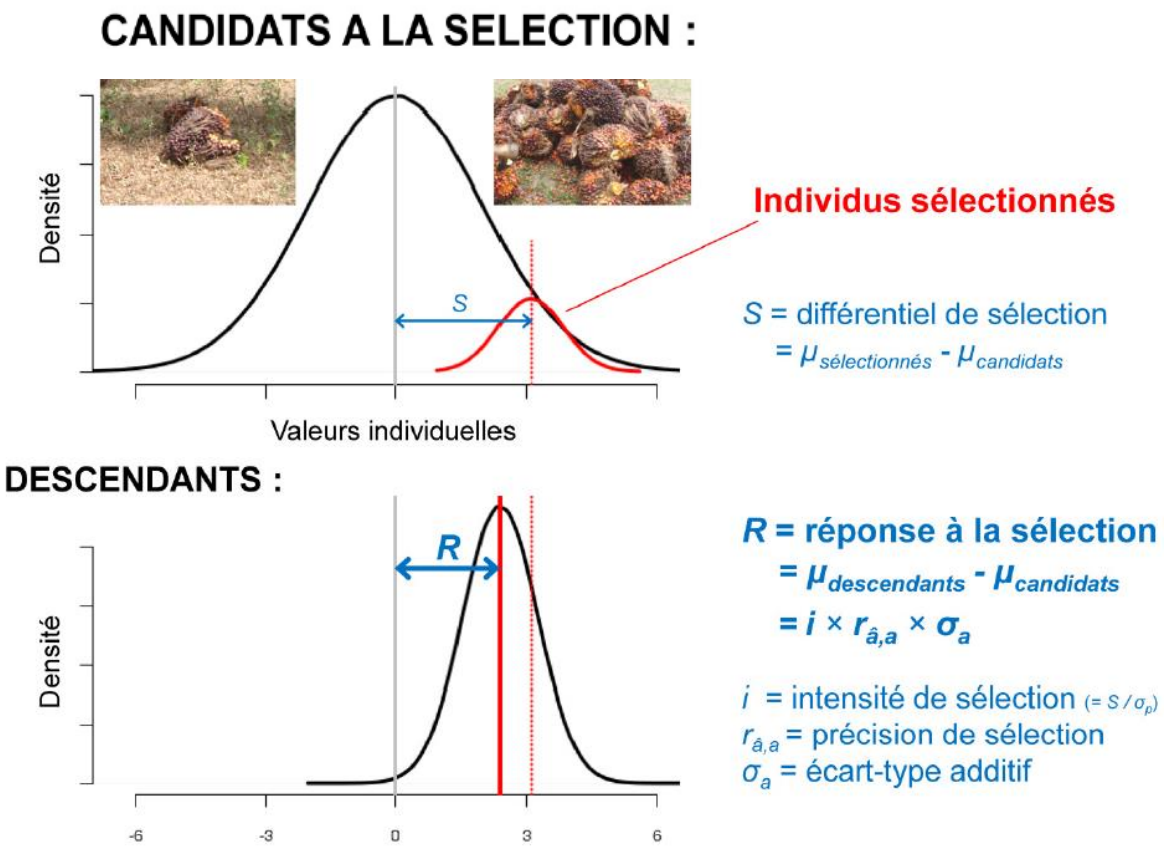



**Figure 3 : Réponse à la sélection R et différentiel de sélection S ($S = \mu_{sélectionnés} - \mu_{candidats}$) (Cros D, 2014)**

# 1 Synthèse bibliographique

## 1.1 Bases de la génétique quantitative

### 1.1.1 Notion de valeur génétique

Un individu possède une valeur phénotypique P selon l'équation P = $\mu + G + E$ avec $\mu$ la moyenne phénotypique, G sa valeur génotypique et E la déviation environnementale (Falconer et Mackay, 1996), calculée dans une population. La valeur génotypique peut être décomposée selon l'équation G = A + D + I avec A la somme des effets génétiques additifs (ou Breeding Value), D l'effet de dominance qui représente l'interaction entre les allèles d'un même locus et I les interactions entre les allèles à différents loci (également appelé épistasie). La nature de ces effets est représentée sur la figure 2.

La breeding value (BV) peut être définie comme la somme des BV à chacun des loci qui influent sur un caractère phénotypique et se définit selon la formule suivante : $A = \sum_{locus}(\alpha_i + \alpha_j)$ avec $\alpha_i$et $\alpha_j$ les effets des allèles i et j d'un même locus. La BV est utilisée pour sélectionner les individus avec une forte valeur génétique dans un schéma de sélection. Ils sont alors utilisés comme parents pour obtenir une descendance avec les caractères phénotypiques améliorés.

Cette BV fait partie de l'équation qui estime la réponse à la sélection de la manière suivante :

$R = i.r_{A\hat{A}}.\sigma_A$ où i représente l'intensité de la sélection (i = S/$\sigma_P$ : S représente le différentiel de sélection), $r_{A\hat{A}}$ la précision de la sélection et $\sigma_A$ la variance additive.

La précision de la sélection est la corrélation entre la valeur additive et son estimateur. Dans le cas où l'estimateur est le phénotype, la précision de la sélection ($r_{A,P}$) correspond à l'héritabilité au sens strict ($h^2$). Cette héritabilité correspond à la part de variance additive ($\sigma_A{}^2$) de la variance phénotypique totale ($\sigma_P{}^2$). En d'autres termes, l'héritabilité au sens strict mesure la part de la variation génétique transmissible par la reproduction sexuée dans la variation phénotypique.

$$r_{A,P} = \frac{cov(A,P)}{\sigma_A\sigma_P} = \frac{\sigma_A^2}{\sigma_A\sigma_P} = \frac{\sigma_A}{\sigma_P} = h_{SS}$$

La réponse à la sélection est alors : R = $h_{SS}^2$ × S, elle est illustrée sur la figure 3.

La réponse à la sélection correspond au progrès génétique, l'estimation de la précision de la sélection a donc un intérêt majeur : plus elle est élevée, plus le progrès génétique augmente.

Une autre valeur d'héritabilité peut être calculée pour chaque individu : l'héritabilité au sens large mesure la part de variation d'origine génétique ($\sigma_G$) dans la variation phénotypique totale ($\sigma_G + \sigma_E$ avec $\sigma_E$ la variance environnementale) :

$$h_{SL}^2 = \frac{\sigma_G{}^2}{\sigma_G{}^2+\sigma_E{}^2}$$

### 1.1.2 Apparentement et déséquilibre de liaison

Le déséquilibre de liaison (DL) est défini comme l'association non aléatoire entre les allèles de deux ou plusieurs loci. Plus deux loci sont situés prêts l'un de l'autre, moins leurs allèles ont de chances d'être transmis de façon indépendante (Jomphe, 2006). Ce DL peut être dû à la dérive

génétique, la sélection, la liaison physique entre deux locus ou à l'existence de groupes différenciés (les flux de gènes sont limités entre les groupes). Egalement, deux allèles portés par deux chromosomes différents peuvent être en DL.

Au cours des générations, le DL peut augmenter sous l'effet de la sélection et de la réduction de la diversité génétique (Robins et al., 2011). Cependant, le DL peut aussi décroitre au cours des générations avec les recombinaisons, et ce d'autant plus vite que les gènes ne sont pas liés. En effet, les individus ayant un ancêtre commun héritent par exemple d'un même fragment ponctuel du chromosome ancestral. Les gènes de ce fragment sont en DL. La région chromosomique autour de de fragment en commun est dite identique par descendance (IBD pour Identical By Descent). La taille de ces régions identiques tend à diminuer au fil des générations, ce qui entraine un DL plus faible. Plus la longueur totale des régions chromosomiques identiques par descendance entre deux individus est grande, plus leur apparentement est fort. Le DL est donc lié à l'apparentement des individus.

En génétique quantitative, il est possible de donner une valeur du taux d'apparentement entre deux individus. Les valeurs de l'apparentement peuvent être estimées à partir du pedigree (elles correspondent à la probabilité d'avoir deux allèles identiques par descendance) ou à partir des marqueurs moléculaires : l'apparentement est calculé à partir de la fréquence des allèles suivant plusieurs formules mathématiques (Habier et al., 2007 ; Vanraden, 2008). La matrice contenant les valeurs de l'apparentement entre chaque individu est couramment appelée «matrice kinship».

La SG utilise des marqueurs moléculaires censés représenter l'ensemble du génome, en espérant qu'au moins un marqueur se trouve dans chaque zone du génome en DL. Le nombre de marqueurs nécessaire dépend alors du DL, et donc de l'apparentement des individus.

## 1.2 La sélection génomique

### 1.2.1 L'intérêt de la sélection génomique

La SG est une pratique mise au point chez les bovins laitiers : les principales races bovines laitières sont sélectionnées depuis 2009 via cette méthode, et ce sera bientôt le cas sur les races à viande (« INRA - sélection génomique bovins » 2015). Cette technique peut permettre l'augmentation du taux de progrès génétique de 40 à 80% par rapport à la sélection classique (« SEMEX France - Sélection Génomique » 2015). La SG s'étend aujourd'hui à d'autres espèces animales telles que les canards et les moutons (Huang et al., 2008 ; Van Der Werf, 2009).

La SG prend également une ampleur considérable dans le monde du végétal car cette technique est très prometteuse. Elle permet d'augmenter le gain génétique par unité de temps en améliorant notamment l'efficacité de la sélection.

Elle permet dans un premier temps de diminuer le coût de la sélection en évaluant les candidats plus précocement dans un schéma de sélection. En effet, l'évaluation phénotypique des plantes est souvent coûteuse à cause de l'utilisation des terres et le coût de la main d'œuvre. De plus, le phénotypage peut demander beaucoup de temps suivant l'espèce considérée. A l'inverse, les coûts de séquençage diminuent fortement et deviennent abordables. La SG pourrait remplacer certaines étapes de phénotypage.

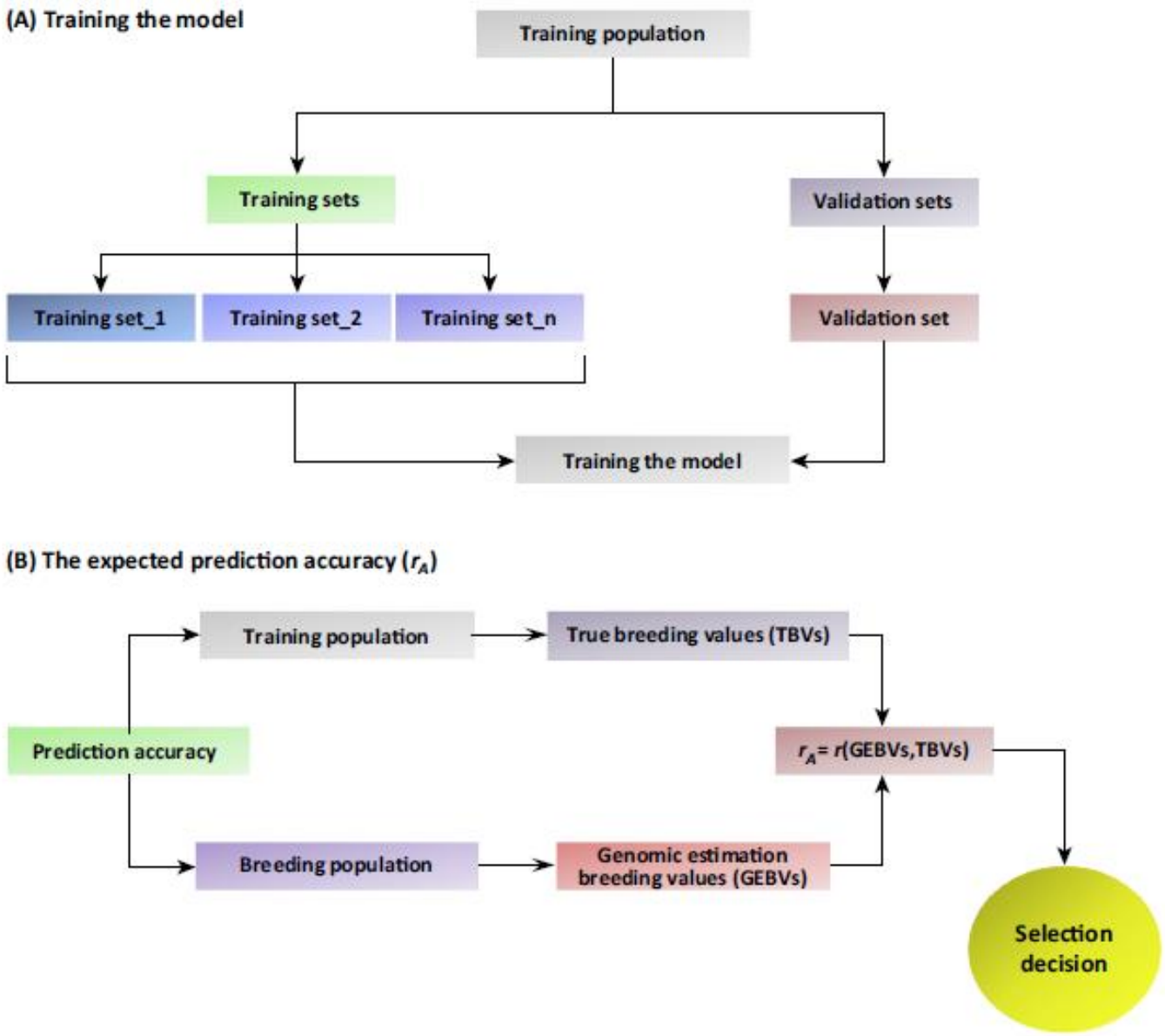



**Figure 4 : Mise en place du modèle de prédiction destiné à la sélection génomique. (A) La population d'entrainement permet d'élaborer le modèle de prédiction (le meilleur modèle, c'est à dire celui qui prédira le mieux les valeurs phénotypiques de la population test, sera retenu) (B) Le modèle est ensuite appliqué à la population candidate pour prédire les GEBV (genomic estimated breeding values). Les individus avec les GEBV les plus élevées seront sélectionnés (Desta et al, 2014).**

La SG est également jugée supérieure à la sélection assistée par marqueurs (SAM) pour certains caractères dans plusieurs publications : Heffner et al., 2011a (sur du blé); Massman et al., 2013 (sur du maïs); Wang et al., 2014 (sur des hybrides de seigle). Par exemple, la SG a permis d'augmenter la précision de la prédiction de 0,12 à 0,59 pour le rendement en grain pour les hybrides de seigle de Wang et al. (2014). Elle permet d'améliorer la sélection avec une plus forte précision des BV. Cette méthode permet de capturer les effets mineurs des gènes contrairement à la SAM qui ne détecte que les associations « marqueurs-QTL» significatives, car la sélection se fait à partir du génome entier. De plus, la variation des caractères quantitatifs expliqués par des marqueurs individuels est généralement faible, elle n'atteint que rarement 5% d'après Pelgas (2011). Un schéma comparant le mode de fonctionnement de la SG et la SAM est présenté en annexe 1.

La SAM est efficace pour des caractères peu polygéniques gouvernés par peu de régions génomiques et à larges effets. La SG est capable de sélectionner sur plusieurs caractères à la fois et sur des caractères qui peuvent être dus à plusieurs loci car les méthodes statistiques permettent d'estimer simultanément l'effet de tous les marqueurs, même ceux avec un effet faible.

### 1.2.2 Principe de la sélection génomique

La sélection génomique est une méthode qui permet de prédire le phénotype d'un individu en fonction de son génotype. L'objectif de cette méthode est donc de prédire la valeur génétique d'un individu pour un caractère d'intérêt agronomique, par exemple. Cette technique estime les effets des marqueurs sur l'ensemble du génome via un modèle de régression.

Les valeurs prédites sont appelées les GEBVs (Genetic Estimated Breeding Values) (Meuwissen, 2001). En d'autres termes, la breeding value mesure le nombre d'allèles ou de gènes et leurs effets qui sont intéressants à transférer à la descendance.

La première étape de la SG consiste à établir un modèle statistique qui assigne une valeur à chaque marqueur. Ce modèle de prédiction comprend des paramètres qui sont estimés sur une population dite « d'entrainement » qui est génotypée pour un ensemble de marqueurs moléculaires et phénotypée pour le(s) caractère(s) que l'on souhaite prédire.

Cette estimation du modèle de prédiction est faite à partir d'une validation croisée : la population d'entrainement est divisée en deux sous-groupes : le premier est utilisé pour prédire un premier modèle avec certains paramètres (le phénotype et le génotype sont donc utilisés), le deuxième permet de tester le modèle de prédiction pour en estimer sa précision (seul le génotype est utilisé). De nombreux paramètres sont testés et le modèle ayant la meilleure prédiction est retenu.

L'efficacité de la prédiction peut se mesurer par la précision, qui correspond au coefficient de corrélation de Pearson entre les GEBVs et les TBV (True Breeding Value). Cette précision intervient dans l'équation du sélectionneur suivante : $R = i \times r_{TBV,GEBV} \times \sigma_A$. Les étapes de la mise en place du modèle sont schématisées sur la figure 4.

Lors d'une seconde étape, il est possible de prédire la GEBV d'un individu dans une population candidate, dont seul le génotype est connu, sur la base du modèle mis au point avec la population d'entrainement. Les individus avec les plus fortes estimations de la breeding value sont sélectionnés. Par la suite, ils peuvent être utilisés dans d'autres schémas de sélection : la sélection

assistée par marqueurs peut par exemple être utile pour introgresser un caractère particulier (Desta, 2014).

La performance de la SG dépend donc de cette précision de la prédiction des valeurs additives, qui peut être affectée par différents paramètres.

## 1.3 Les limites de la SG

Une des limites principales de la sélection génomique porte sur l'établissement du modèle de prédiction. L'inférence, qui consiste à estimer les paramètres, est une étape clé : l'objectif étant de maximiser la vraisemblance du modèle, qui peut être défini comme « la probabilité des données sachant les paramètres ».

De nombreux paramètres entrent en jeu et peuvent varier d'une population à l'autre comme l'héritabilité des caractères, la densité des marqueurs moléculaires, la composition de la population d'entrainement (taille et relations d'apparentement) et le modèle statistique pour estimer les GEBVs. Ces paramètres sont présentés ci-dessous.

### 1.3.1 Modèles de prédiction

Plusieurs modèles de prédiction peuvent être utilisés en SG. Le choix du modèle dépend du caractère phénotypique étudié car l'objectif est d'utiliser le modèle statistique qui modélise de la manière la plus efficace la distribution des effets des marqueurs tout au long du génome.

Le but de la SG est d'estimer l'effet de tous les marqueurs dans un modèle de régression. En SG, le nombre de marqueurs (variables explicatives) est beaucoup plus élevé que le nombre d'individus (observations) : p >> n, il n'est donc pas possible d'utiliser un modèle de régression linéaire multiple directement. La solution la plus courante est de « rétrécir » les estimations des effets additifs ; on parle aussi d'utiliser un « skrinkage » sur les effets estimés. Dans ce cas, les effets attribués aux marqueurs sont issus d'une loi que l'on appelle « prior ». Les modèles utilisés sont donc paramétriques : certains modèles autorisent des effets forts, faibles voire nuls des marqueurs suivant les lois qu'ils appliquent.

Le modèle de base, s'apparentant à un modèle linéaire mixte, estime les effets additifs de chaque marqueur ainsi :

$$y = \mu + X\beta + e$$ avec :

- p : nombre de marqueurs
- n : nombre d'individus
- y : vecteur des observations phénotypiques (p × 1)
- $\mu$ : moyenne globale des observations
- X : matrice des allèles (p × n), qui peut prendre les valeurs 1,0 ou -1 pour les génotypes AA, AB et BB respectivement
- $\beta$ : vecteur des effets des marqueurs (p × 1)
- e : $e \sim N(0, \sigma_e^2)$ (p × 1) avec e le vecteur des effets résiduels et $\sigma_e^2$ la variance résiduelle

**Tableau 1 : Détail des lois de prior de chaque modèle utilisé dans l'étude**

| *Modèles* | *Loi appliquée aux effets des marqueurs et aux effets résiduels* | *Estimation des variances* | *Hyper paramètres* |
|---|---|---|---|
| ***Flat*** ***(Fixé)*** | $\beta \sim N(0, \sigma_\beta^2)$ $e \sim N(0, I\sigma_e^2)$ | $\sigma_\beta^2 \sim X^{-2}$ $\sigma_e^2 \sim X^{-2}$ | |
| ***RR BLUP*** ***(Gaussien)*** | $\beta \sim N(0, \sigma_\beta^2)$ $e \sim N(0, I\sigma_e^2)$ | *REML* | |
| ***BRR*** ***(Gaussien)*** | $\beta \sim N(0, \sigma_\beta^2)$ $e \sim N(0, I\sigma_e^2)$ | $\sigma_\beta^2 \sim X^{-2}$ $\sigma_e^2 \sim X^{-2}$ | |
| ***BL*** ***(Double exponentiel)*** | $\beta \sim N(0, \sigma_{\beta(i)}^2)$ $e \sim N(0, I\sigma_e^2)$ | $\sigma_e^2 \sim X^{-2}$ $\sigma_{\beta(i)}^2 = \tau_i^2 \times \sigma_e^2$ | $\tau_i^2 \sim E\ (\lambda^2/2)$ $\lambda \sim Gamma$ |
| ***Bayes A*** ***(Scaled-t)*** | $\beta \sim Scale - t(df_b, S_b)$ ou $\beta \sim N(0, \sigma_{\beta(i)}^2)$ $e \sim N(0, I\sigma_e^2)$ | $\sigma_e^2 \sim X^{-2}$ $\sigma_\beta^2 \sim X^{-2}(S_b, df_b)$ | $S_b \sim Gamma$ *Par défaut,* $df_b = 5$ |
| ***Bayes B*** ***(Scaled-t)*** | $\beta \sim Scale - t(df_b, S_b)$ ou $\beta \sim N(0, \sigma_{\beta(i)}^2)$ $e \sim N(0, I\sigma_e^2)$ | $\sigma_e^2 \sim X^{-2}$ $\sigma_\beta^2 \sim X^{-2}(S_b, df_b)$ $\pi \sim Bêta$ : proportion des marqueurs avec un effet nul | $S_b \sim Gamma$ *Par défaut,* $df_b = 5$ |
| ***Bayes*** $C_\pi$ | $\beta \sim N(0, \sigma_\beta^2)$ $e \sim N(0, I\sigma_e^2)$ | $\sigma_\beta^2 \sim X^{-2}$ $\sigma_e^2 \sim X^{-2}$ $\pi \sim Bêta$ : proportion des marqueurs avec un effet nul | |

La population d'entrainement permet d'obtenir une estimation des $\beta$ : $\hat{\beta}$. L'estimation des breeding value de la population candidate se calcule comme suit :

$$\mathrm{GEBV_i} = \mu + \sum_{j=1}^{n} \left(\mathrm{X_{ij}}\, \hat{\beta}_{j.}\right)$$ avec :

- i : les individus
- j : les marqueurs

Les modèles **Bayes A** et **Bayes B** proposent une variance spécifique à chaque marqueur. La distribution a priori de cette variance est une loi de Chi² inverse, ce qui revient à dire que les effets associés aux allèles suivent une distribution t. Le modèle Bayes B propose en plus la possibilité d'avoir une proportion de marqueurs avec un effet nul.

Le modèle **RR-BLUP** (ridge regression BLUP) (Endelman et al., 2011) (Bayesian Ridge Regression) (Pérez et al., 2010) utilise une distribution normale des effets des marqueurs, ce qui permet d'autoriser des effets faibles et moyens aux marqueurs mais peu d'effets forts. Les variances sont estimées par REML (restricted maximum likelihood, Gilmour et al., 1995). Tous les marqueurs ont ici la même variance. Le modèle **BRR** est la version bayésienne du RR-BLUP : il utilise les informations fournies a priori. Ici, ces informations sont les variances des effets des marqueurs et des effets résiduels. Elles suivent une loi inverse du Chi². Le modèle **Bayes $C_\pi$** possède les mêmes caractéristiques que le BRR mais propose en plus la possibilité d'avoir une proportion de marqueurs avec un effet nul.

Le modèle **LASSO** (least absolute shrinkage selection operator) propose une variance spécifique à chaque marqueur. Il peut être implémenté d'une approche Bayesienne : Bayesian LASSO Regression (**BLR**) (de los Campos et al., 2009). Le modèle BLR suit une distribution de type double exponentielle.

Le détail des lois de prior de chaque modèle est présenté dans le tableau 1 et des exemples de distributions a priori des effets aux marqueurs ($\beta$) pour différentes méthodes bayésiennes de sélection génomique sont illustrés sur la figure 5.

Une autre méthode souvent utilisée est le **G-BLUP**, qui se base sur une matrice d'apparentement moléculaire (calculée avec l'hypothèse de la ségrégation mendélienne des allèles lors de la méiose). Elle permet d'estimer directement les GEBVs avec le modèle suivant :

$$\mathrm{y} = \mu + \mathrm{g} + \mathrm{e}$$ avec

- p : nombre de marqueurs
- y : vecteur des observations phénotypiques (p × 1)
- μ : moyenne globale des observations
- g : vecteur des GEBV (p × 1) associé à une matrice d'apparentement moléculaire
- e : e $\sim N(0, \sigma_e^2)$ (p × 1) avec e le vecteur des effets résiduels et $\sigma_e^2$ la variance résiduelle

Habier et al. (2013) préconisent l'utilisation d'une méthode Bayésienne plutôt que le G-BLUP car les méthodes Bayésiennes exploitent mieux le DL. Néanmoins, Spindel et al. (2015) ont comparé une méthode bayésienne (RR-BLUP) à la méthode G-BLUP et ont constaté que le premier modèle était plus performant pour le rendement en grains (où peu de QTL à effets forts ont été détectés). A l'inverse, le modèle G-BLUP donne de meilleurs résultats que le RR-BLUP pour la période de floraison (où de nombreux QTL à forts effets ont été détectés).

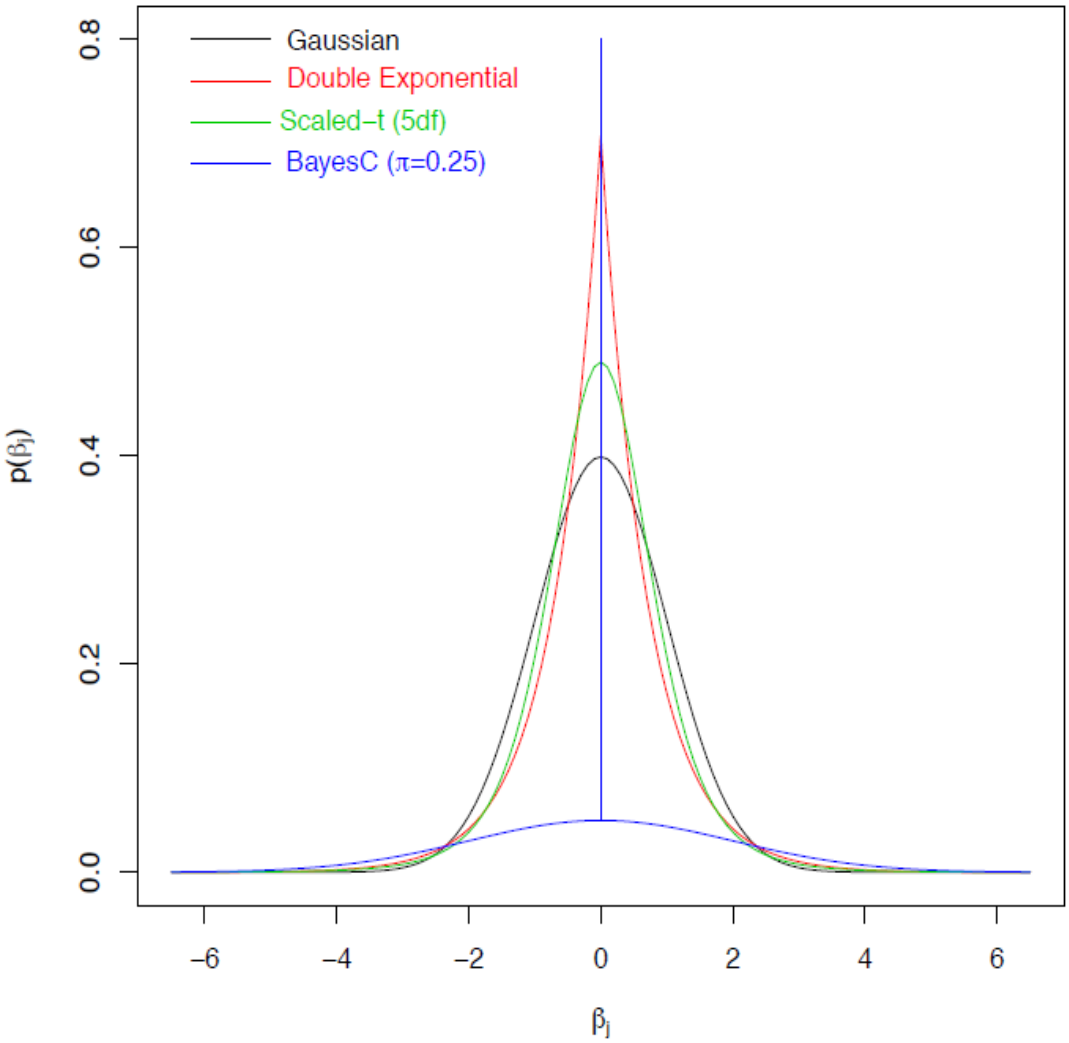



**Figure 5 : Exemple de distributions a priori des effets aux marqueurs (βj) pour différentes méthodes bayésiennes de sélection génomique (Pérez et al, 2013). La courbe Gaussian (en noir) peut correpondre aux modèles BRR ou RR-BLUP, la courbe Double Exponential (en rouge) aux modèle BL, la courbe Scaled-t (en vert) aux modèles BayesA et Bayes B et la dernière courbe (en bleu) au modèle BayesCπ.**

Des modèles non paramétriques existent également (Nadaraya-Watson estimator, support vector machine regression and neural networks). Ces modèles se sont révélés plus efficaces que les modèles paramétriques lorsque les phénotypes sont basés sur l'épistasie (Howard et al., 2014).

Gianola et al. (2006) ont proposé un modèle semi paramétrique nommé **RKHS** (Reproducing Kernel Hilbert Spaces) dans lequel ils ont combiné les fonctionnalités intéressantes de ces modèles avec un cadre de modèle mixte :

$$y_i = w_{i'}\beta + z_{i'}\mathrm{u} + \mathrm{g}(X_i) + e_i \quad \text{avec :}$$

- i = 1, 2, …, n
- $\beta$ : vecteur d'effets fixes inconnus (par exemple, l'emplacement physique d'un individu)
- u : vecteur des effets génétiques additifs $u \sim N(0, A\sigma_u^2)$ avec $\sigma_u^2$ la variance génétique additive et A la matrice de relation additive
- $w_{i'}$ et $z_{i'}$ : vecteur d'incidence inconnus
- $\mathrm{g}(X_i)$ : fonction inconnue des données SNP et vecteur des résidus
- e : e $\sim N(0, I\sigma_e^2)$ avec e le vecteur des effets résiduels et $\sigma_e^2$ la variance résiduelle

Ce modèle a déjà prouvé son efficacité en sélection génomique lorsque les caractères étudiés sont faiblement additifs : **RKHS** (Reproducing Kernel Hilbert Spaces). Ce modèle semi-paramétrique permet d'incorporer implicitement les effets de dominance et d'épistasie entre les marqueurs, en plus de la partie additive (Gianola et al. 2010).

### 1.3.2 Composition de la population d'entrainement et de la population de validation

L'apparentement entre les individus est un facteur très important à considérer en SG. En effet, lors de la validation croisée qui va permettre d'estimer les paramètres d'un modèle de prédiction, si la relation entre les deux groupes est très étroite, il va être courant d'obtenir une forte précision.

Dans l'étude de Beaulieu et al. (2014), chez l'épinette blanche, la précision des GEBVs obtenus après une validation croisée avec des individus n'ayant pas d'apparentement était deux fois plus faible que lorsque des demi-frères étaient présents dans les populations.

Cependant, si nous souhaitons utiliser ce modèle sur une population candidate, cette valeur de la précision ne sera valable que si la population en question présente des taux d'apparentement similaires à la population ayant servie pour le calibrage du modèle (population d'entrainement).

L'étude de Nakaya et al. (2012) a démontré que dans la plupart des études sur les espèces cultivées, la validation croisée s'effectue à l'intérieur d'une même génération alors que la population candidate provient d'une ou plusieurs générations en dessous. Par conséquent, la précision élevée de prédiction est potentiellement trompeuse.

De plus, le nombre de générations représentées par les individus de la population d'entrainement peut faire varier la précision du modèle. Ainsi, Kizilkaya et al. (2010) ont démontré que le fait d'avoir quatre générations représentées dans la population d'entrainement permettrait de maintenir plus longtemps la valeur de la précision dans les générations suivant la calibration du modèle qu'avec seulement deux générations représentées.

Il est donc indispensable de se projeter dans un schéma de sélection afin de choisir une population d'entrainement adéquate avant de se lancer dans un schéma de SG, afin que la

prédiction de GEBV imite les stratégies de validation croisée. Wang et al. (2014) obtiennent une précision de prédiction élevée seulement pour des candidats étroitement liés à la population d'entrainement, ce qui laisse paraître un pronostic plutôt faible pour une prédiction sur du matériel génétiquement éloigné.

Egalement, si la population de test présente un fort taux d'apparentement, la population d'entrainement adéquate n'est pas forcément celle qui est seulement composée d'individus très apparentés avec la population test. Il peut être important que la population d'entrainement soit diversifiée. Par exemple, les travaux de Technow et al. (2013) en SG chez des lignées de maïs indiquent que la prédiction des lignées dentées et cornées indépendamment est plus importante lorsque les deux types de lignées font partie de la population d'entrainement.

En ce qui concerne le nombre d'individus de la population d'entrainement, plusieurs études ont montré que la précision de la prédiction de la SG augmente avec sa taille (Varaden et al., 2010 ; Technow et al., 2013 ; Grattapaglia et al., 2014). Cela est dû à l'accumulation des informations phénotypiques et génotypiques (plus d'allèles rares) qui rend plus robuste l'estimation des effets des marqueurs (Hayes et al., 2009).

### 1.3.3 Héritabilité

De nombreuses études de simulation montrent que plus la précision de la prédiction est évaluée avec des caractères ayant une forte héritabilité, plus elle sera élevée (Legarra et al., 2008 ; Hayes et al., 2009 ; Lorenz et al., 2011 ; Grattapaglia et al., 2014). Ainsi, l'étude de Hayes et al. (2009) montre que pour deux caractères phénotypiques avec des héritabilités de 0.2 et 0.7, la précision de la prédiction est de 0.6 et 0.8 respectivement. De plus, la SG est plus performante que la SAM en ce qui concerne les caractères complexes à faible héritabilité (Heffner, 2011b).

Néanmoins, l'étude de Kumar et al. (2014) sur la pomme, qui porte également sur la prédiction de la qualité du fruit via la sélection génomique obtient une précision de la prédiction de ces caractères comprise entre 0.68 (pour l'astringence) et 0.89 (pour la teneur en matières solubles). Pourtant, ces phénotypes ont une héritabilité au sens strict de 0.26 et 0.19 respectivement. De plus, les caractères « fermeté du fruit » ($h^2_{ss} = 0.43$) et « acidité » ($h^2_{ss} = 0.16$) sont prédits avec une précision de 0.83 et 0.81.

### 1.3.4 Densité de marquage

La SG estime les effets des marqueurs sur l'ensemble du génome. La densité idéale de marquage dépend donc du niveau de DL autour des QTLs et dans tout le génome. Le but étant d'avoir un marqueur en DL avec le plus de gènes possibles influant sur un caractère.

La densité de marquage peut être un paramètre important à prendre en compte suivant l'apparentement de la population d'entrainement et de la population candidate. En effet, plus l'apparentement entre deux populations est faible, plus il existe de la diversité nucléotidique entre les deux groupes, plus la portée du DL diminue. Des études de prédiction entre populations éloignées ont montré que la précision de la prédiction peut être améliorée en augmentant la densité de marquage. Ainsi, Toosi et al. (2010) ont simulé une population d'animaux à partir de quatre lignées pures. En multipliant le nombre de marqueurs par quatre, la précision de la prédiction peut être augmentée jusqu'à 35%. Néanmoins, cette précision atteint un plateau à partir

d'environ 20 marqueurs par cM (centiMorgans). Egalement, la précision de la prédiction augmente jusqu'à 800 marqueurs puis reste stable dans la population élite de maïs de Zhao et al. (2012).

Une forte densité de marqueurs est nécessaire si le déséquilibre de liaison est faible. Ce facteur peut néanmoins être fixé à cause des technologies de génotypage (par exemple l'utilisation d'une puce à ADN fixe le nombre de marqueurs moléculaires).

### 1.3.5 Interaction G × E

Plus particulièrement chez les plantes, l'interaction G × E (Génotype × Environnement) est également un paramètre à prendre en compte dans le modèle si l'on souhaite utiliser la population d'entrainement pour prédire le phénotype d'une population qui évolue dans un environnement différent (ce qui est souvent le cas). Cette interaction dépend de la valeur de l'héritabilité du caractère.

Cependant, peu d'études ont encore porté sur cette interaction. Wang et al. (2014) ont néanmoins étudié cet aspect et ont montré que chez le seigle, la précision de la prédiction peut diminuer si le modèle est établi avec une population d'entrainement l'année n et testé sur une population l'année n+1. Cependant, la précision peut aussi rester stable suivant la variabilité du caractère étudié entre les différentes années (plus la moyenne et la variance du caractère varient entre les différentes années, plus la précision de la prédiction est faible).

Pour conclure, bien que la sélection génomique soit porteuse d'espoirs dans le domaine de la sélection végétale, sa mise en œuvre peut être difficile vu la complexité des paramètres du modèle de prédiction. Nous essayons par la suite d'estimer ces paramètres afin de prédire des caractères liés à la qualité chez la tomate. Une étude similaire sur la qualité de la pomme a déjà été effectuée (Kumar et al., 2014) avec des résultats très positifs.

## 1.4 L'étude de la tomate et ses ressources génétiques et génomiques

La tomate est une espèce originaire d'Amérique du Sud. Elle est aujourd'hui cultivée partout dans le monde pour être consommée fraiche ou transformée (concentré, sauces, conserves etc.). La consommation moyenne européenne de tomate est de 21 kg/habitant/an (« France Agrimer », 2014). En plus de son importance économique, elle est considérée comme une plante modèle en génétique. Le Consortium international du Génome de la Tomate lancé en 2003 a achevé en mai 2012 le séquençage des génomes de la tomate cultivée (*Solanum lycopersicum*) et de son ancêtre sauvage (*Solanum pimpinellifolium*), (Tomato Genome Consortium, 2012). Des outils s'y sont associés telle que la plateforme SGN (Sol Genomics Network, www.*solgenomics.net*) qui regroupe des données et des outils d'analyse pour les solanacées (Fernandez-Pozo et al., 2015). Une puce à ADN a également pu être développée dans le cadre du projet SolCAP (« SolCAP Solanaceae Coordinated Agricultural Project » 2013), basée sur plus de 7700 marqueurs SNP et ayant été utilisée dans de nombreuses applications telles que la cartographie génétiques (Sim et al., 2012) ou l'analyse de la diversité nucléotidique (Blanca et al., 2015). Ces outils sont aujourd'hui disponibles pour entreprendre de la SG. De plus, de nombreuses ressources génétiques sont disponibles dans le monde, notamment dans l'unité de recherche GAFL (Génétique et Amélioration des Fruits et Légumes) qui conserve plus de 3000 accessions de tomates différentes.

Actuellement, les études d'amélioration génétique sur la tomate portent principalement sur les résistances et l'adaptation de la plante (maladies, sècheresse, salinité etc.), son rendement et la qualité du fruit (sucres, fermeté, teneur en vitamine C, etc...). De récents travaux sur la qualité du fruit ont permis de révéler l'existence de gènes à l'origine de ces caractères via la détection de QTL (Causse et al., 2004 ; Pascual et al., 2015) ou la GWAS (Genome Wide Association Study, Sauvage et al., 2014). En effet, la première étude a localisé plus d'une centaine de gènes impliqués dans la production de sucres et d'acides ainsi que dans le poids du fruit. Lors de l'approche GWAS, les fruits ont été analysés en ce qui concerne plusieurs métabolites tels que les acides aminés, les sucres et l'ascorbate, ce qui a permis de mettre en évidence l'association de 44 loci avec 19 caractères phénotypiques.

## 1.5 Objectifs des travaux

La sélection pour la qualité de la tomate est aujourd'hui complexe car elle porte sur plusieurs caractères à la fois (forme, poids, couleur du fruit etc.), l'interaction GxE est souvent présente et les étapes de phénotypage sont très coûteuses. Dans le contexte de l'augmentation de la disponibilité de données génomiques et de l'essor de la SG chez les animaux, nous étudions ici la pertinence de l'utilisation de la SG dans le but de prédire un ensemble de phénotypes liés la qualité du fruit chez la tomate. L'objectif de ces travaux est d'étudier la faisabilité de cette pratique dans un schéma de sélection chez la tomate. Pour ce faire, les objectifs sont les suivants :

1) Tester le poids de paramètres sur la précision de la SG :
    - Les modèles de prédiction
    - La composition de la population d'entrainement, qui se subdivise en deux :
        - le nombre d'individus attribués à la population d'entrainement et celui attribué à la population de validation
        - l'optimisation ou non de cette répartition
    - L'héritabilité des caractères phénotypiques étudiés
    - La densité des marqueurs SNPs
    - La présence d'interaction GxE

2) Tester l'influence du type de population étudié. Afin de répondre à cet objectif, les populations utilisées seront de 3 types :
    - GWAS (Genome Wide Association Study)
    - RIL (Recombinant Inbred Line)
    - MAGIC (Multi-parent Advanced Generation InterCross)

3) Analyser les effets attribués aux marqueurs en comparant l'étude de la SG avec les SNPs identifiés dans une étude de GWAS pour les mêmes caractères phénotypiques.

4) Evaluer les effets positifs ou négatifs de ces paramètres et des populations sur la SG pour proposer une méthode optimisée d'utilisation de la SG chez la tomate dans un schéma de sélection

**Tableau 2 : Jeux de données utilisés dans l'étude (GWAS, GWAS 2, MAGIC, RIL et GBS) ainsi que les phénotypes étudiés, l'héritabilité des caractères (les valeurs suivies d'une * sont des héritabilités au sens scrict, les autres correspondent à des héritabilités au sens large), les caractéristiques des jeux de données, leur utilisation dans cette étude**

| Jeu de données | Phénotypes étudiés | $h^2$ | Caractéristiques | Etudes |
|---|---|---|---|---|
| **GWAS** | Threonate | 0,17* | **13 phénotypes**<br>**7608 marqueurs**<br>**164 génotypes** | • Choix du modèle de prédiction<br>• Comparaison des modèles de prédiction : RKHS et Bayes C<br>• Composition des PE et PV<br>• Héritabilité<br>• Densité de marquage<br>• Interaction GxE (2007 – 2008)<br>• Analyse des effets attribués aux marqueurs : comparaison avec l'étude de Sauvage et al (2014) |
| | Aspartate | 0,28* | | |
| | ASA | 0,55* | | |
| | Fructose | 0,57* | | |
| | Acidité | 0,75 | | |
| | Brix | 0,60* | | |
| | Teneur en matière soluble | 0,73 | | |
| | Malate | 0,64* | | |
| | Sucres | 0,63 | | |
| | Fermeté | 0,72 | | |
| | Poids du fruit | 0,83 | | |
| | Nombre de loges | 0,85 | | |
| | pH | 0,57 | | |
| **GWAS 2** | Brix | | **5 phénotypes**<br>**6768 marqueurs**<br>**231 génotypes** (dont 164 en commun avec GWAS) | • Comparaison des modèles de prédiction : Bayes A, Bayes B, Bayes C, BL, BRR, RKHS et G-BLUP |
| | Fructose | | | |
| | Poids du fruit | | | |
| | pH | | | |
| | Sucres | | | |
| **MAGIC** | Brix | 0,46 | **6 phénotypes**<br>**1345 marqueurs**<br>**397 génotypes** | • Comparaison des modèles de prédiction : RKHS et Bayes C<br>• Composition des PE et PV<br>• Héritabilité<br>• Densité de marquage |
| | pH | 0,35 | | |
| | Fermeté | 0,60 | | |
| | Poids du fruit | 0,27 | | |
| | Hauteur de la plante | 0,70 | | |
| | Acidité | 0,16 | | |
| **RIL** | Sucres | 0,61 | **4 phénotypes**<br>**501 marqueurs**<br>**124 génotypes** | • Comparaison des modèles de prédiction : RKHS et Bayes C<br>• Composition des PE et PV<br>• Héritabilité |
| | Poids du fruit | 0,75 | | |
| | pH | 0,51 | | |
| | Fermeté | 0,63 | | |
| **GBS** | Brix | | **5 phénotypes** (issus de GWAS)<br>**59 079 marqueurs**<br>**63 génotypes** (inclus dans GWAS) | • Densité de marquage |
| | Fructose | | | |
| | Poids du fruit | | | |
| | pH | | | |
| | Sucres | | | |

# 2 Matériel et méthodes

## 2.1 Matériel biologique

Dans notre étude, plusieurs jeux de données qui ont été produits par des travaux de recherche antérieurs, portant sur l'amélioration génétique de la tomate, sont utilisés, le but étant d'exploiter des données diverses afin d'évaluer la pertinence de la SG pour le plus de cas possible. Les phénotypes étudiés sont en relation avec la qualité de la tomate et plus particulièrement celle du fruit. Ces phénotypes sont détaillés dans les publications se rapportant à chaque population (Xu et al. (2013) et Sauvage et al. (2014) pour les données GWAS ; Pascual et al. (2015) pour les données MAGIC, Saliba-Colombani et al. (2001) pour les données RIL et Lin et al. (2014) pour le jeu de données GBS).

Les caractères phénotypiques étudiés portent sur la hauteur de la plante, le poids du fruit, sa forme (nombre de loges), son contenu en sucres (contenus en sucres solubles : mesuré en °brix, sucres totaux, fructose) et en acides, sa fermeté, son pH et quatre métabolites secondaires (acide ascorbique, malate, threonate, aspartate). Les jeux de données étudiés sont répertoriés dans le tableau 2, ainsi que leur utilisation dans l'étude.

### 2.1.1 Jeu de données GWAS

Une population de 164 accessions de tomate a été utilisée en génétique d'association (Sauvage et al., 2014), ce qui a permis de révéler 44 loci candidats pour des traits métaboliques du fruit. Ces 44 loci ont été significativement associés avec 19 caractères phénotypiques tels que les taux de sucrose, d'ascorbate et de malate par exemple. Le panel des individus étudiés provient de la core collection décrite par Xu et al. (2013). Il est composé de *S. lycopersicum* (SL), *S. lycopersicum* cv. *cerasiforme* (SLC) et *S. pimpinellifolium* (SP). En moyenne, le déséquilibre de liaison est fort chez les SL ($r^2$=0.57), moyen chez les SLC ($r^2$=0.54) et plus faible chez les SP ($r^2$=0.34). De plus, le taux d'apparentement moyen entre les individus est très faible (K=0.0738). L'étude a également révélé la présence de structure populationnelle entre les SL et SP et entre les SLC et SP ($F_{ST}$ = 0.2132 et $F_{ST}$=0.1583, respectivement). Les individus ont été génotypés sur la puce à ADN du projet SolCAP (« SolCAP Solanaceae Coordinated Agricultural Project » 2013).

Ces différentes données ont permis de réaliser 2 populations d'étude différentes (tableau 2). La première population (GWAS) est composée des 164 individus (31 SL, 115 SLC et 18 SP) utilisés en génétique d'association (Sauvage et al., 2014), qui sont génotypés avec 7608 marqueurs et phénotypés pour 13 caractères (threonate, aspartate, ASA (teneur en ascorbate), fructose, acidité, °brix, teneur en matières solubles, malate, sucres, fermeté, poids du fruit, nombre de loges et pH). La seconde population (GWAS 2) contient les individus du jeu de données GWAS ainsi que de nouveaux individus phénotypés à Avignon. Elle se compose de 231 individus au total (50 SL, 149 SLC et 32 SP) génotypés avec 6768 marqueurs et phénotypés pour 5 caractères (°brix, fructose, poids du fruit, pH et sucres).

Les données phénotypiques du jeu de données GWAS sont disponibles pour l'année 2007 et l'année 2008 pour les phénotypes aspartate, threonate, ASA, °brix poids du fruit et malate. Ces informations permettent de tester l'influence de l'interaction entre le génotype et l'environnement sur la prédiction.

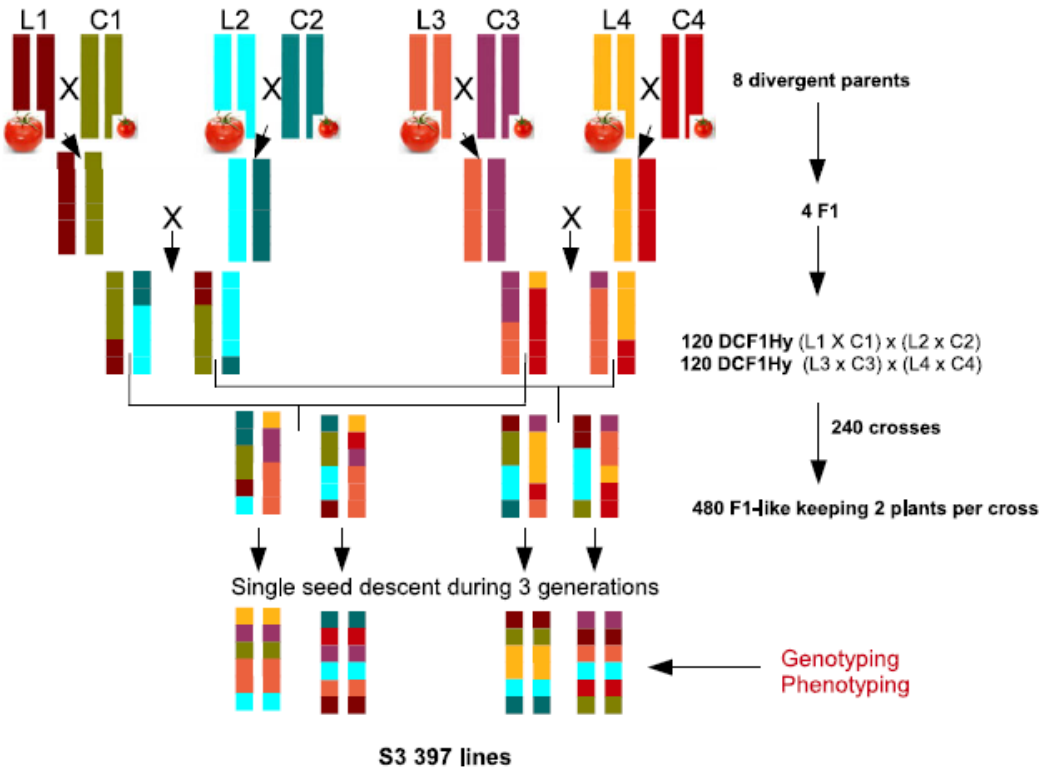



**Figure 6 : Construction de la population MAGIC à partir de 8 parents : 4 à gros fruits (L1 : Levovil, L2 : Stupicke PR, L3 : LA0147 et L4 : Ferum) et 4 à petits fruits ( C1 : Cervil, C2 : Criollo, C3 : Plovdiv24A, C4 : LA1420), Pascual et al (2015)**

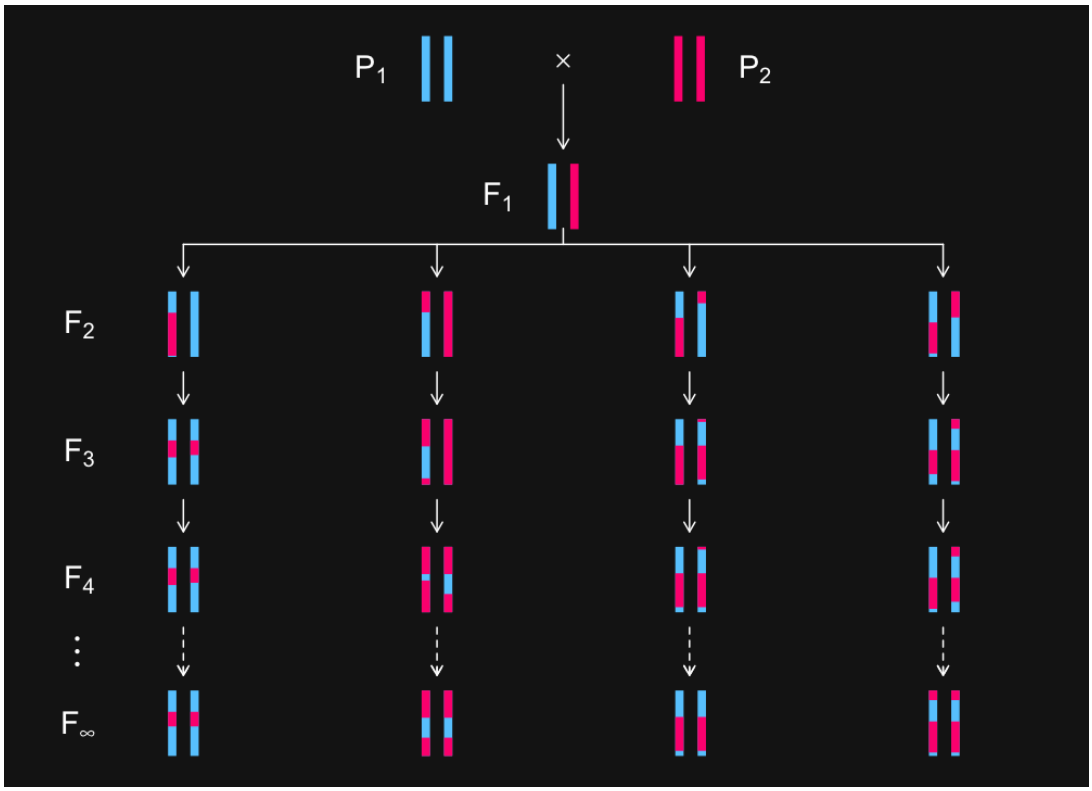



**Figure 7 : Schéma de création d'une population RILs (recombinant inbred lines). («QTL mapping for phenotypes measured over time » 2015)**

### 2.1.2 Jeu de données MAGIC

La population MAGIC a été développée à partir de 8 accessions de tomates ayant une forte diversité génétique. Ces variétés font également partie du jeu de données GWAS. Elle est composée de 397 lignées qui proviennent de 4 générations de croisements et 3 autofécondations (figure 6). La particularité de ce type de croisement est le fait qu'il permet d'augmenter le nombre d'événements de recombinaisons. Cela entraine l'absence de structure populationnelle, aucun sous-groupe n'est détecté. De plus, le DL est très faible (< 0.7 pour 5 cM, and < 0.3 pour 25 cM). La matrice d'apparentement contient donc des valeurs très faibles (75% des valeurs sont inférieures à 0.042) même si un faible nombre d'individus reste fortement apparenté (0.01% des valeurs sont environ égales à 0.8). Ces lignées ont été décrites par Pascual et al. (2015) à l'INRA d'Avignon. Elles ont été génotypées avec 1345 marqueurs SNPs et phénotypées pour 6 caractères (°brix, pH, fermeté, poids du fruit, hauteur de la plante et acidité titrable (TA)) dans des conditions environnementales aussi proches que possible que celle appliquées aux deux autres types de populations.

### 2.1.3 Jeu de données RIL

La population RIL (Recombinant Inbred Lines) étudiée est une population biparentale utilisée pour une analyse génétique de la qualité organoleptique de la tomate (Saliba-Colombani et al., 2001). Les RILs ont été obtenues après croisement de l'accession Cervil (C) et de l'accession Levovil (L) (appartenant également au jeu de données GWAS) suivi de 6 générations d'auto croisements (figure 7) puis la population a été phénotypée à Châteaurenard (13) de février à juin 1996. Cervil est une tomate cerise (*S. lycopersicum* cv. *cerasiforme*) à petits fruits (6–10 g) qui a été choisie pour son bon goût et sa forte intensité d'arômes. Levovil (*S. lycopersicum* Mill.) est une lignée à plus gros fruits (90–160 g) avec un goût plus commun. Le fait que cette population provienne de 2 parents seulement entraine un taux de recombinaison efficace plus faible que pour les populations GWAS et MAGIC. Il y a donc un fort taux d'apparentement entrainant un fort DL. Le jeu de données RIL est composée de 124 individus, génotypés avec 501 marqueurs et phénotypés pour 4 caractères (sucres, poids du fruit, pH et fermeté).

### 2.1.4 Jeu de données GBS

Cette population est composée de 63 accessions reséquencées (7 accessions SP, 44 SLC et 12 SL). A l'origine, elle est issue d'une population constituée de 360 accessions (Lin et al., 2014) provenant de plusieurs organismes : INRA, TGRC (Tomato Genetics Resource Center), USDA (US Department of Agriculture), EU-SOL (European Union Solanaceae project) et IVF-CAAS (Institute of Vegetables and Flowers, Chinese Academy of Agricultural Science). A partir des 11 millions de SNPs polymorphes, identifiés par séquencage à haut-débit parmi ces 360 accessions, un jeu de données composé de 59 079 marqueurs SNPs a été utilisé (filtrés avec une MAF supérieure à 5% et un minimum de données manquantes inférieur à 30%). Pour ce jeu de données, les phénotypes disponibles sont les mêmes que ceux du jeu de données GWAS 2 (°brix, fructose, poids du fruit, pH et sucres) car les 63 accessions font partie de cette population.

Pour chaque jeu de données (GWAS, MAGIC, RIL et GBS), des analyses en composantes principales (ACP) permettent de visualiser la structure et les corrélations entre les caractères phénotypiques étudiés (annexe 2), un spectre de fréquence des allèles est réalisé pour évaluer la

part d'allèles rares présents dans la population (AFS pour Allele Frequency Spectrum, annexe 3) et une matrice représente l'apparentement entre les individus (annexe 4).

## 2.2 Méthodes

Afin d'évaluer la précision de la prédiction du phénotype pour chaque jeu de données, le protocole d'utilisation des modèles de prédiction est étudié. Ensuite, l'impact des paramètres suivants sont testés : les modèles de prédiction, la composition des populations d'entrainement et de validation, l'héritabilité des caractères phénotypiques, la densité de marquage et l'interaction entre le génotype et l'environnement. Pour finir, les effets attribués aux marqueurs en SG sont comparés à une étude effectuée en GWAS (Sauvage et al., 2014) qui, avec des approches statistiques, propose un seuil de détection des marqueurs impliqués dans l'expression d'un phénotype. L'objectif est ici de comparer les marqueurs mis en évidence en SG (qui capte les allèles à petits effets) et en GWAS.

### 2.2.1 Protocole d'utilisation des modèles de prédiction

Le package BGLR (Gustavo de los Campos et al., 2014) implémenté dans le logiciel R (R Development Core Team, 2008) est utilisé via le serveur de l'unité GAFL de l'INRA d'Avignon pour utiliser les modèles de prédiction. Les modèles produisent une valeur de la précision de la prédiction ($r^2$ : corrélation entre le phénotype prédit et le phénotype mesuré). Pour un jeu de données avec des paramètres fixes, un modèle de prédiction fonctionne plusieurs fois afin d'obtenir plusieurs valeurs de précision de la prédiction. A chaque fois, les populations d'entrainement et de validation sont composées d'individus différents. Cela permet de construire des diagrammes en boîtes et d'évaluer la variabilité des résultats entre phénotypes ou entre conditions par exemple.

En premier lieu, le nombre de cycles (nombre de fois où le modèle fonctionne) doit être optimisé, c'est à dire réduit le plus possible pour que l'étude soit rapide et ce, sans que cela influe sur les résultats de prédiction. Afin de choisir le nombre de cycles nécessaires, les individus issus de du jeu de données GWAS sont utilisés pour les caractères phénotypiques suivant : « pH » et « nombre de loges ». Le modèle BL est utilisé pour prédire ces 2 caractères 10 fois indépendamment avec 100, 200, 300, 400, 500, 600, 700, 800, 900 et 1000 cycles (75% des individus font partie de la population d'entrainement et 25% de la population de validation). Dix groupes de données sont donc obtenus avec le nombre de valeurs correspondant au nombre de cycles utilisé. Le test de Student (Student, 1908) permet ensuite d'évaluer la différence entre le fait de faire fonctionner le modèle 100 fois et de le faire fonctionner 200, 300, 400, 500, 600, 700, 800, 900 ou 1000 fois. Egalement le test Tukey HSD permet d'estimer la différence entre tous les essais deux à deux.

Par la suite, les paramètres (modèles de prédiction, composition de la population d'entrainement, héritabilité des caractères phénotypiques, densité des marqueurs SNPs, présence d'interaction GxE) sont modifiés pour estimer un $r^2$ moyen (précision de la prédiction moyenne) avec le nombre de cycles qui parait le plus pertinent.

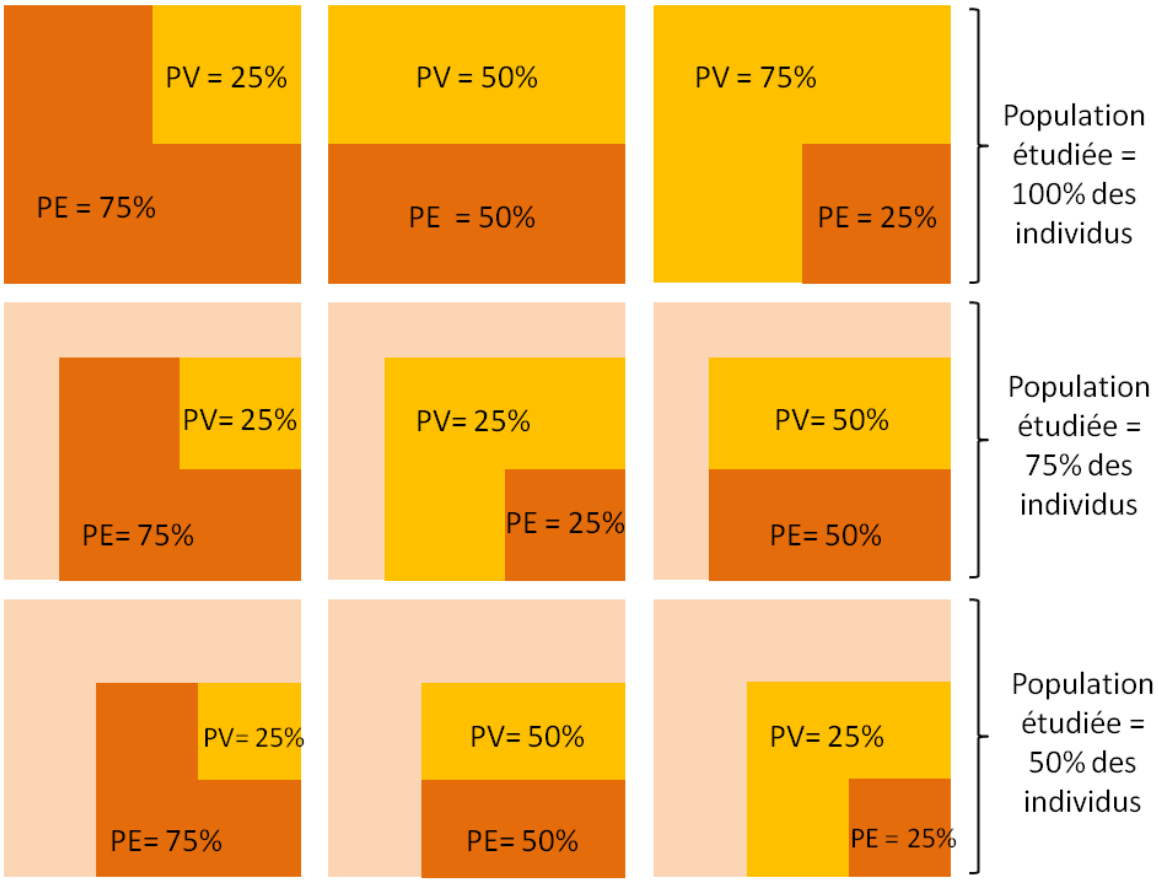



**Figure 8 : Schéma du choix de la répartition des individus dans les populations d'entrainement (PE) et les populations de validation (PV). Sur la première ligne, la totalité des individus est utilisé, sur les deux autres seulement un pourcentage.**

### 2.2.2 Modèles de prédiction

Afin de comparer les modèles statistiques de prédiction (Bayes A, Bayes B, Bayes C, BL, BRR, RKHS et G-BLUP), le jeu de données GWAS 2 est utilisée (pour représenter un panel de diversité le plus large possible). La prédiction des caractères phénotypiques du jeu de données est testée pour différents paramètres de composition des populations d'entrainement et de validation explicités dans les parties suivantes. Dans chaque cas, les effets des modèles sur la précision de la prédiction du phénotype sont comparés à l'aide du test Tukey HSD (Lane, 2010).

A la suite de cette comparaison, seuls les modèles donnant des résultats significativement différents sont choisis pour optimiser le temps de l'étude. Ils sont employés pour l'analyse des jeux de données GWAS, MAGIC et RIL.

Des tests ANOVA sont ensuite appliqués afin d'évaluer pour chaque jeu de données l'influence des modèles, des paramètres (composition de la population d'entrainement et nombre de marqueurs pris en compte dans le modèle) et leurs interactions. Pour les paramètres donnant la meilleure précision de la prédiction, les deux meilleurs modèles sont comparés avec un test *t* de Student. De plus, le pourcentage sur tous les cas testés (avec les différents paramètres) où un modèle permet une meilleure prédiction par rapport à un autre est calculé pour chaque jeu de données et chacun des phénotypes.

### 2.2.3 Composition de la population d'entrainement et de la population de validation

Dans un premier temps, l'impact du pourcentage d'individus répartis dans les populations d'entrainement et de validation sur les jeux de données GWAS, MAGIC et RIL est étudié. Nous testons l'utilisation de 75%, 50% et 25% d'individus dans la population d'entrainement choisis au hasard dans la population totale, le reste des individus faisant partie de la population de validation. A chaque cycle, les individus sont répartis au hasard dans les deux populations. Pour chacun des trois essais, le nombre d'individus total est diminué progressivement. Les choix de la composition des différentes populations sont schématisés figure 8.

Egalement, le jeu de données GWAS 2 comprenant les individus de GWAS ainsi que les individus d'une core collection est étudié pour visualiser l'influence de l'augmentation du nombre d'individus sur la précision de la prédiction (N=164 *vs* N= 231).

De plus, une méthode nommée « CDmean » est testée. Elle permet d'optimiser le choix de la population d'entrainement en répartissant les individus dans les différentes populations afin qu'elles soient, génétiquement, le plus diversifiées possible (Rincent et al., 2012). Cette méthode est aussi expérimentée pour une population d'entrainement de 75%, 50% et 25% du nombre d'individus de la population totale. Le test de Student permet de comparer l'effet de la méthode d'optimisation sur la précision de la prédiction.

Egalement, les valeurs des effets attribuées aux marqueurs SNPs par les modèles sont relevées à chaque cycle pour le modèle Bayes C. Lorsque le marqueur n'a pas d'effet sa valeur est nulle ou négative. Lorsqu'il a un effet, sa valeur est comprise entre 0 et 1. La moyenne des effets attribués à chaque marqueur sur l'ensemble des cycles est calculée, ce qui permet d'analyser la répartition de ces effets sur le génome (peu de marqueurs à forts effets, beaucoup de marqueurs à moyens effets etc.).

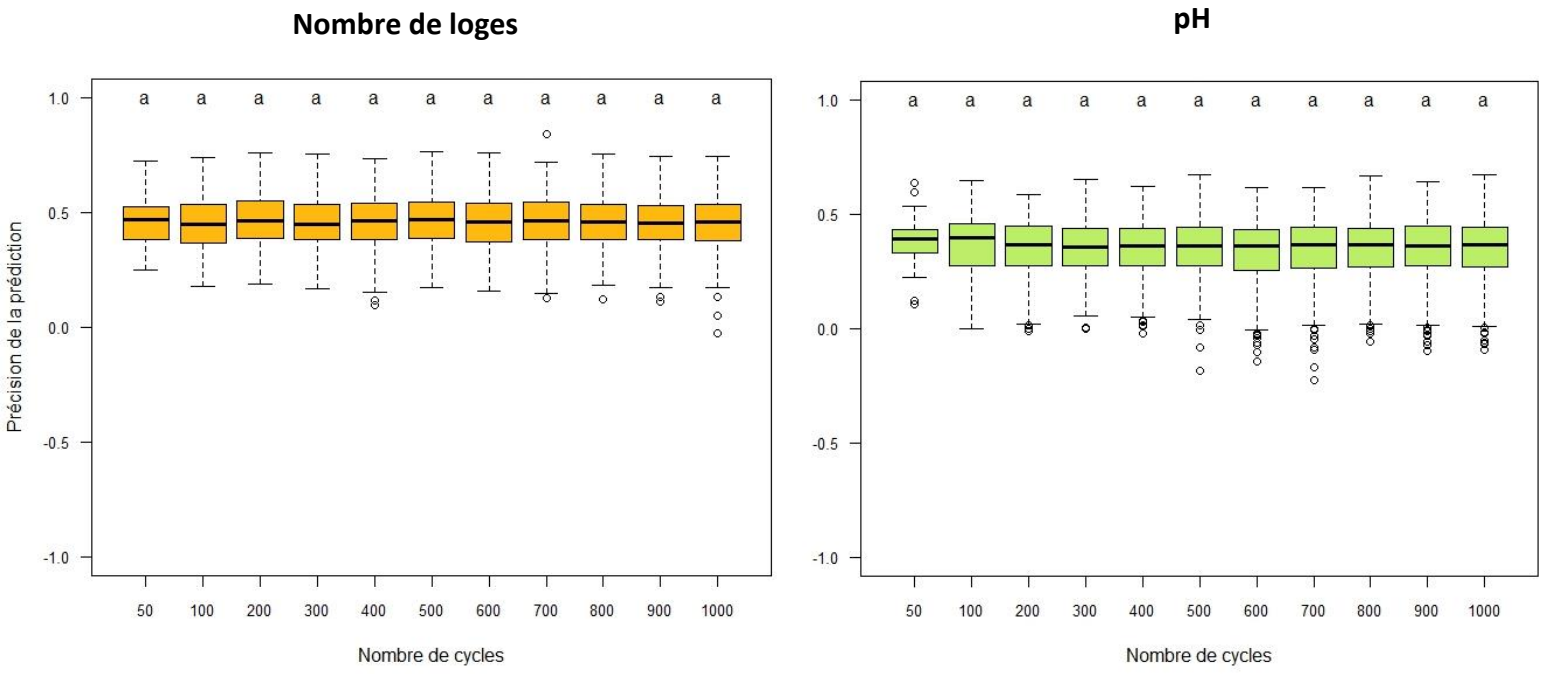



**Figure 9 : Précision de la prédiction du modèle Bayes C pour le jeu de données GWAS (75% des individus font partie la population d'entrainement 25% de la population de validation), en fonction du nombre de cycles effectués par le modèle. Les résultats pour le caractère phénotypique « Nombre de loges » se situe à gauche (en jaune) et ceux pour le caractère phénotypique « pH » à droite (en vert). Les lettres représentent le résultat du test statistique Tukey HSD.**

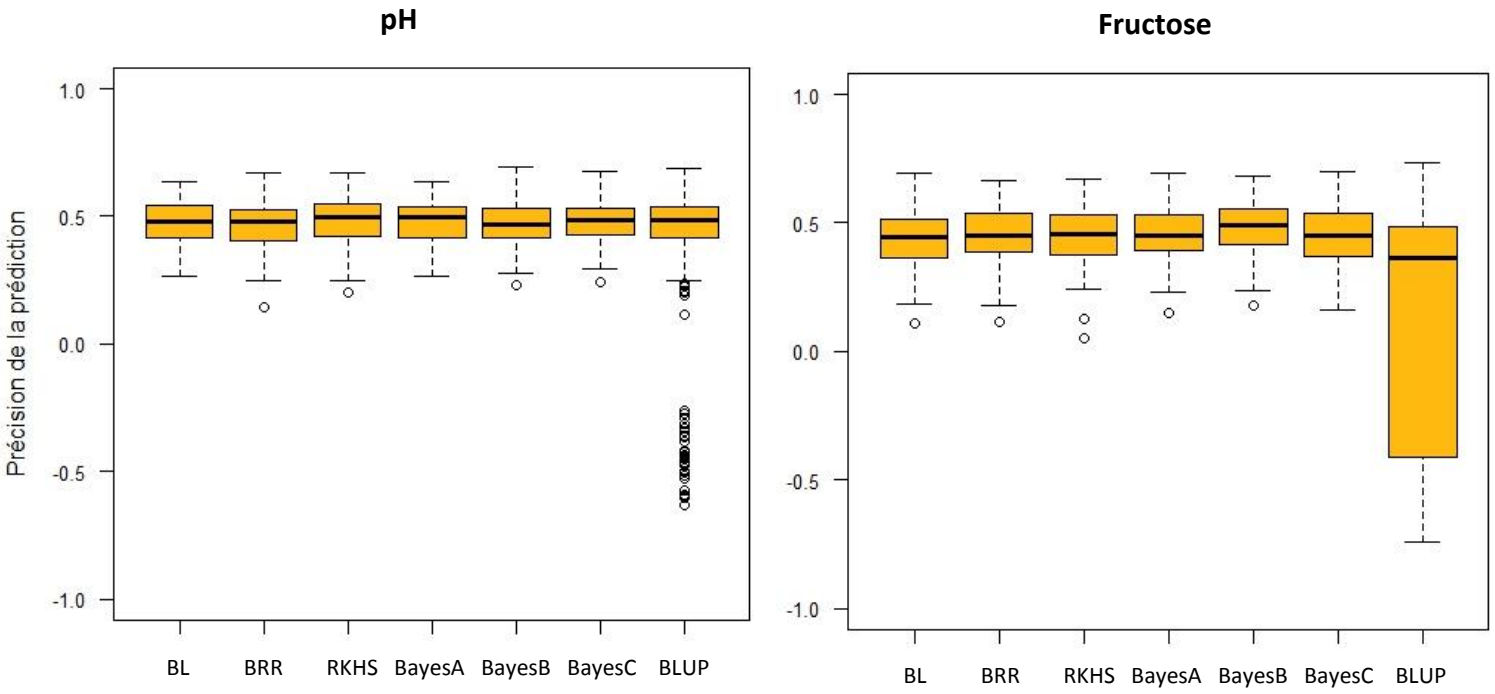



**Figure 10 : Précision de la prédiction du modèle Bayes C pour la population GWAS 2 (75% des individus font partie la population d'entrainement 25% de la population de validation), en fonction de différents modèles de prédiction (à gauche le caractère phénotypique étudié est le pH, à droite le fructose).**

### 2.2.4 Héritabilité phénotypique

Pour les jeux de données GWAS, MAGIC et RIL, la précision de la prédiction est comparée en fonction des phénotypes afin de visualiser les caractères les mieux prédits. De plus, la corrélation entre la précision de la prédiction et les héritabilités phénotypiques est étudiée lorsque les héritabilités au sens large ont été calculées (tableau 2) : le coefficient de Pearson est calculé afin de mesurer le lien entre les deux variables et un test de significativité de ce coefficient est effectué.

Egalement, un des objectifs de l'étude est d'estimer la variabilité de la précision de la prédiction pour un même phénotype dans différents jeux de données.

### 2.2.5 Densité de Marquage

Pour les jeux de données GWAS, MAGIC et RIL, la précision de la prédiction est testée avec une population d'entrainement correspondant à 75% de la population totale, en changeant le nombre de marqueurs pris en compte dans l'étude (de 10% à 100% de la totalité des marqueurs disponibles par incrément de 10). Les marqueurs sont choisis aléatoirement.

Egalement, le jeu de données GBS avec 59 079 marqueurs et le jeu de données GWAS2 avec 6 768 marqueurs sont comparés avec 63 individus en commun (75% des individus font partie de la population d'entrainement et la méthode CDmean est utilisée pour répartir les individus). L'objectif est de savoir si le fait d'augmenter le nombre de marqueurs permet d'obtenir une meilleure précision de la prédiction, dans l'optique d'analyser la pertinence de l'utilisation du séquençage haut débit en sélection génomique.

De plus, les effets attribués aux marqueurs par le modèle Bayes C sont observés dans chaque étude.

### 2.2.6 Interactions G × E

L'interaction entre le génotype et l'environnement est étudiée à partir du jeu de données GWAS car les phénotypes obtenus en tunnel plastiques à Avignon pour les années 2007 et 2008 sont disponibles. Pour les phénotypes aspartate, threonate, ASA (teneur en ascorbate), °brix, poids du fruit et malate, le modèle Bayes C est utilisé avec 164 individus dont 75% dans la population d'entrainement et 25% dans la population de validation.

Dans un premier temps, la PE est composée des individus d'une année et la PV des individus de la même année (prédiction des individus de l'année 2007 à partir des individus de l'année 2007 et prédiction des individus de l'année 2008 à partir des individus de l'année 2008). Dans un deuxième temps, la PE est composée des individus d'une année et la PV des individus de l'autre année (prédiction des individus de l'année 2008 à partir des individus de l'année 2007 et prédiction des individus de l'année 2007 à partir des individus de l'année 2008).

De plus, la variance et la moyenne des données phénotypiques de 2007 et 2008 sont comparées avec les tests de Bartlett et le test de Student respectivement.

**Tableau 3 : Précision de la prédiction obtenue pour chaque jeux de données et chacun des phénotypes disponibles avec la méthode CDmean (75% de la totalité des individus sont utilisés dans la population d'entrainement et 25% dans la population de validation). En rouge : résultats les plus élevés lorsqu'ils sont significativement différents avec le test de Student.**

| | GWAS | | MAGIC | | RIL | |
|---|---|---|---|---|---|---|
| | RKHS | BayesC | RKHS | BayesC | RKHS | BayesC |
| **Aspartate** | 0,38 | 0,31 | | | | |
| **Threonate** | 0,38 | 0,34 | | | | |
| **Malate** | 0,70 | 0,66 | | | | |
| **ASA** | 0,74 | 0,73 | | | | |
| **Poids du fruit** | 0,93 | 0,93 | 0,65 | 0,65 | 0,88 | 0,88 |
| **Nombre de loges** | 0,73 | 0,61 | | | | |
| **Fermeté** | 0,81 | 0,80 | 0,56 | 0,57 | 0,64 | 0,60 |
| **Hauteur de la plante** | | | 0,74 | 0,74 | | |
| **Fructose** | 0,46 | 0,53 | | | | |
| **Sucres** | 0,70 | 0,77 | | | 0,59 | 0,59 |
| **Brix** | 0,73 | 0,77 | 0,52 | 0,51 | | |
| **Acidité** | 0,74 | 0,78 | 0,56 | 0,57 | | |
| **Teneur en matière solubles** | 0,78 | 0,80 | | | | |
| **pH** | 0,52 | 0,49 | 0,61 | 0,64 | 0,41 | 0,45 |

**Tableau 4 : Pourcentage des cas (parmi les différentes compositions des populations d'entrainement et de validation et la densité de marquage) où le modèle RKHS conduit à une meilleure prédiction que le modèle Bayes C (en rouge : cas où dans plus de 50% des cas, le modèle RKHS est meilleur)**

| | GWAS | MAGIC | RIL |
|---|---|---|---|
| **Aspartate** | 96% | | |
| **Threonate** | 77% | | |
| **Malate** | 50% | | |
| **ASA** | 62% | | |
| **Poids du fruit** | 19% | 44% | 35% |
| **Nombre de loges** | 76% | | |
| **Fermeté** | 46% | 20% | 74% |
| **Hauteur de la plante** | | 36% | |
| **Fructose** | 42% | | |
| **Sucres** | 23% | | 58% |
| **Brix** | 65% | 80% | |
| **Acidité** | 8% | 40% | |
| **Teneur en matière solubles** | 42% | | |
| **pH** | 31% | 28% | 32% |

### 2.2.7 Analyse des effets attribués aux marqueurs : comparaison avec une étude de GWAS

Afin d'évaluer la pertinence des résultats obtenus, les effets attribués aux marqueurs avec le modèle Bayes C en ce qui concerne le jeu de données GWAS (CDmean, 75% des individus dans la population d'entrainement) est comparée à l'étude de génétique d'association de Sauvage et al. (2014), car les populations étudiées sont identiques. Pour cela, 7 phénotypes sont étudiés (aspartate, threonate, malate, ASA, fructose, °brix et sucrose).

Pour chacun des caractères, Sauvage et al. (2014) ont réalisé des Manhattan plots (-log10(p-value) vs position chromosomique) permettant de visualiser les marqueurs significatifs (qui ont une forte influence sur le phénotype). Les marqueurs contribuant à 10% des effets totaux attribués à l'ensemble des marqueurs en SG sont positionnés sur ces Manhattan plots. L'objectif de cette étude est de distinguer si les marqueurs estimés en SG (c'est à dire pour lesquels un effet a été attribué dans le modèle de prédiction) ont été considérés comme significatifs en GWAS ou non, afin de conforter les résultats obtenus en ce qui concerne les effets attribués aux marqueurs en SG.

De plus, les valeurs de déséquilibre de liaison existant entre les marqueurs de la puce ont été calculées. Afin de visualiser le DL le long du chromosome, la moyenne des valeurs du DL entre les marqueurs est effectuée toutes les 100 000 paires de bases.

**Tableau 5 : Résultats des tests d'analyse de la variance (ANOVA) pour les jeux de données GWAS, MAGIC et RIL utilisés avec les modèles RKHS et Bayes C pour différent paramètres de composition des populations d'entrainement et de validation. Signification des codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

| | | Paramètres | Modèles | Interaction |
|---|---|---|---|---|
| **GWAS** | Threonate | *** | *** | *** |
| | Aspartate | *** | *** | *** |
| | ASA | *** | | *** |
| | Fructose | *** | | *** |
| | Acidité | *** | *** | *** |
| | Brix | *** | | *** |
| | Teneur en matière solubles | *** | *** | *** |
| | Malate | *** | | *** |
| | Sucres | *** | *** | *** |
| | Fermeté | *** | | |
| | Poids du fruit | *** | *** | *** |
| | Nombre de loges | *** | *** | *** |
| | pH | *** | | |
| **MAGIC** | Brix | *** | ** | *** |
| | PH | *** | *** | *** |
| | Fermeté | *** | *** | *** |
| | Poids du fruit | *** | *** | *** |
| | Hauteur de la plante | *** | *** | *** |
| | Acidité | *** | *** | *** |
| **RIL** | Sucres | *** | | * |
| | Poids du fruit | *** | | |
| | pH | *** | | *** |
| | Fermeté | *** | *** | *** |

# 3 Résultats

Suivant les paramètres utilisés, les jeux de données et les phénotypes étudiés, la précision de la prédiction peut être très mauvaise (proche de 0) ou très bonne (proche de 1). Les résultats obtenus sont présentés dans un ordre similaire à celui des parties « Matériel et Méthodes ».

### 3.1.1 Protocole d'utilisation des modèles de prédiction

Pour tester le nombre de cycles, le test de Student est utilisé pour estimer l'effet du nombre de cycles (entre 100 et 1000 fois par incrément de 100) sur la précision de la prédiction. Les analyses montrent qu'il n'y a pas de différences significatives (les p-values sont comprises entre 0.49 et 0.98 pour le pH et entre 0.30 et 0.69 pour le nombre de loges) (figure 9). De plus, le test de Tukey HSD indique qu'il n'y a pas de différence significative entre tous ces essais pris deux à deux.

Le nombre de cycles utilisé dans la suite de l'étude est donc de 100 pour permettre de multiplier les analyses.

### 3.1.2 Modèles de prédiction

Sur les 6 modèles testés, le modèle RKHS est significativement meilleur que les autres (Bayes A, Bayes B, Bayes C, BL et BRR) pour 3 phénotypes sur 5 : poids du fruit, pH et sucrose (les p-values sont de $2.6\times10^{-06}$, $3.1\times10^{-06}$ et $2.3\times10^{-04}$ respectivement). Le modèle G-BLUP est significativement moins bon pour 4 phénotypes sur 5 : °brix, fructose, poids du fruit et sucrose. Néanmoins, pour tous les phénotypes, quelques simulations ont une précision de la prédiction négative ou proche de 0 (figure 10). Les autres modèles ne sont pas significativement différents. Pour la suite de l'étude les modèle RKHS et Bayes C ont donc été retenus pour leur vitesse de calcul supérieure aux autres modèles.

Pour les jeux de données GWAS, MAGIC et RIL, les tests ANOVA indiquent que les paramètres utilisés influent sur les résultats de prédiction pour tous les phénotypes et que le choix du modèle (RKHS et Bayes C) influe sur 14 phénotypes sur 23 (60.8%). De plus, une interaction entre les paramètres et les modèles est mise en évidence pour 20 phénotypes sur 23 (86.9%), ce qui signifie que le modèle n'a pas la même efficacité suivant les paramètres utilisés (tableau 5).

Le paramètre qui conduit à la meilleure précision de la prédiction quel que soit le jeu de données est la méthode du CDmean lorsque l'on utilise 75% des individus dans la population d'entrainement. Le test de Student est donc appliqué sur ces résultats de précision de la prédiction pour les 3 jeux de données afin de comparer les modèles RKHS et Bayes C (tableau 3). De plus, les pourcentages sur tous les cas testés (avec les différents paramètres) où un modèle permet une meilleure prédiction sont présentés dans le tableau 4 et détaillés dans les paragraphes suivants. En ce qui concerne les jeux de données GWAS, lorsque la méthode du CDmean est utilisée, le modèle RKHS conduit à une précision de la prédiction plus élevée pour les deux phénotypes à faible héritabilité (les précisions des prédictions de l'aspartate et du threonate sont respectivement de 0.38 et 0.38 pour le modèle RKHS et 0.34 et 0.31 pour le modèle Bayes C). Le modèle RKHS est également meilleur pour les phénotypes teneur en malate, nombre de loges, fermeté, et pH. A l'inverse, le modèle Bayes C donne de meilleures prédictions pour les phénotypes teneur en

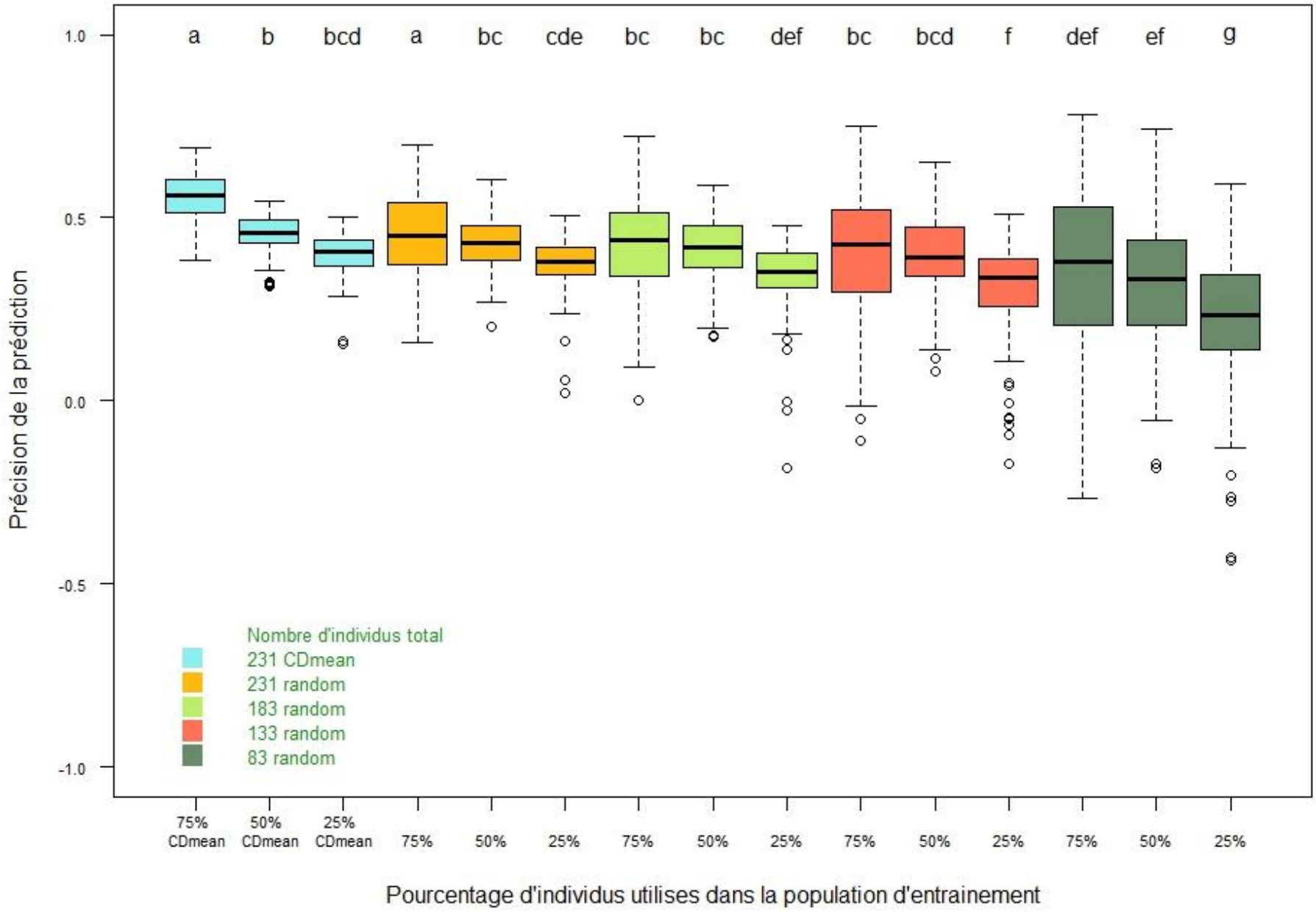



**Figure 11 : Précision de la prédiction du caractère phénotypique « Fructose » du modèle Bayes C pour la population GWAS 2 en fonction du nombre d'individus total utilisé dans l'étude et de la proportion d'individus utilisés dans les populations d'entrainement et de validation. Les lettres en haut sont le résultat du test HSD : pour une même lettre, les paramètres conduisent à des résultats qui ne sont pas significativement différents.**

ascorbate (ASA), fructose, sucres, °brix, acidité et teneur en matières solubles. Les deux modèles ne sont pas différents en ce qui concerne le poids du fruit.

Si l'on compare l'efficacité de RKHS et Bayes C en fonction de tous les paramètres (tableau 4), le modèle RKHS est meilleur que le modèle Bayes C pour la prédiction de l'aspartate, du threonate, du malate, de l'ASA, du nombre de loges et du °brix dans 96%, 77%, 50%, 62%, 76% et 65% des cas respectivement. A l'inverse, plus de la moitié des paramètres conduisent à une meilleure prédiction avec Bayes C pour les phénotypes poids du fruit, fermeté, fructose, sucres, acidité, teneur en matières solubles et pH. Globalement, le modèle Bayes C devient meilleur que RKHS dans les cas où le nombre d'individus total diminue ou lorsque l'on diminue le pourcentage d'individus de la population d'entrainement. En effet, lorsque 164 individus font partie de l'étude et 75% sont destinés à la population d'entrainement, 9 phénotypes sur 13 sont mieux prédits avec le modèle RKHS. A l'inverse, lorsque 164 individus font partie de l'étude et 25% sont destinés à la population d'entrainement, seulement 4 phénotypes sur 13 sont mieux prédits avec le modèle RKHS. De plus, lorsque 44 individus font partie de l'étude et 75% sont destinés à la population d'entrainement, 4 phénotypes sur 13 sont mieux prédits avec le modèle RKHS.

Pour le jeu de données MAGIC, la méthode CDmean ne permet pas d'obtenir des résultats significativement différents entre les deux modèles sauf pour le pH où le modèle Bayes C est légèrement plus efficace ($r^2$=0.64 contre 0.61 avec RKHS, p-value = $2.08x10^{-06}$). De plus, pour tous les cas testés avec différents paramètres, le modèle Bayes C est plus efficace dans plus de la moitié des cas pour tous les phénotypes hormis le °brix. Néanmoins, le modèle RKHS devient plus efficace lorsque le nombre de marqueurs utilisé diminue. En effet le modèle RKHS propose de meilleures prédictions par rapport au modèle Bayes C pour 1, 3, 5 et 6 phénotypes sur 6 lorsque 100%, 90% ou 80%, 70 ou 60% et moins de 50% des marqueurs sont pris en compte respectivement.

La méthode CDmean différencie l'efficacité des modèles Bayes C et RKHS pour la fermeté ($r^2$=0.6 contre $r^2$=0.64, p-value = $1x10^{-04}$) et le pH ($r^2$=0.41 contre $r^2$=0.45, p-value = $1.43x10^{-02}$) en ce qui concerne le jeu de données RIL. De plus, pour tous les cas testés avec différents paramètres, Bayes C est plus efficace dans plus de la moitié des cas pour le poids du fruit et le pH et RKHS pour la fermeté et les sucres totaux. De même que pour le jeu de données MAGIC, le modèle RKHS devient plus efficace lorsque le nombre de marqueurs utilisés diminue.

### 3.1.3 Composition de la population d'entrainement et de la population de validation

Pour les jeux de données GWAS, GWAS 2, MAGIC et RIL, quel que soit le phénotype étudié, plus le pourcentage d'individus composant la population d'entrainement est élevé et plus le nombre d'individus total est grand, plus la précision de la prédiction est forte. Par exemple, en ce qui concerne l'étude du contenu en fructose avec le jeu de données GWAS 2 avec le modèle Bayes C, lorsque la totalité des individus est prise en compte dans l'étude (231 individus), avec la population d'entrainement formée de 75%, 50% et 25% des individus, la précision de la prédiction est en moyenne de 0.45, 0.43 et 0.37 respectivement (figure 11). De plus, pour une population d'entrainement formée de 75% des individus et un nombre d'individus total de 231, 183, 133 et 83, la précision de la prédiction est en moyenne de 0.45, 0.44, 0.41 et 0.34 respectivement. La méthode du CDmean, utilisée sur les jeux de données GWAS, GWAS 2, MAGIC et RIL conduit à la meilleure précision de la prédiction. Lorsque la totalité des individus est utilisée dans le modèle avec 75% dans la population d'entrainement et 25% dans la population de validation, le CDmean

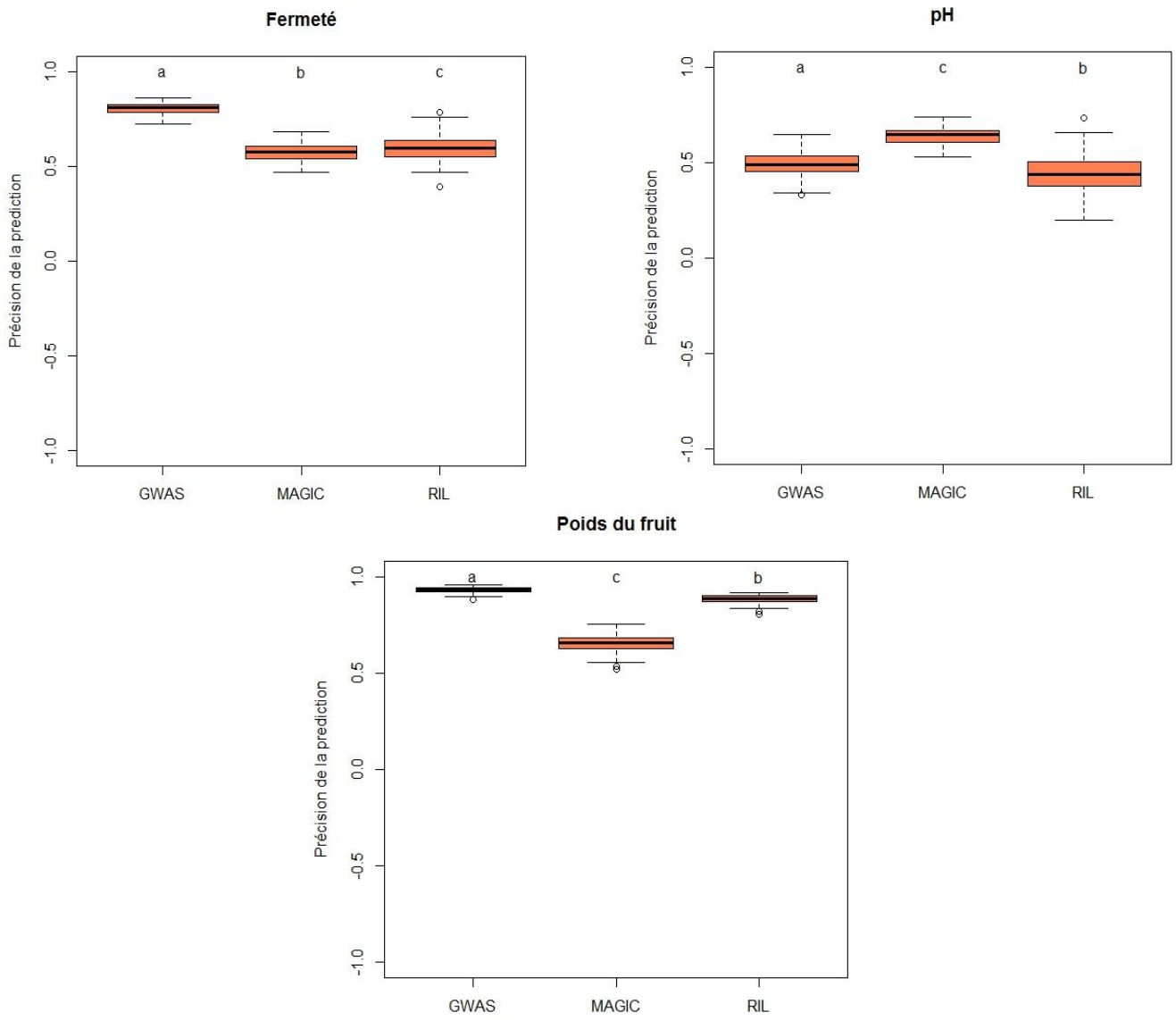



**Figure 12 : Diagrammes en boîte indiquant la précision de la prédiction du modèle Bayes C pour l'ensemble les jeux de données GWAS, MAGIC et RIL sur les caractères phénotypiques suivants : fermeté, pH, poids du fruit (75% des individus font partie de la population d'entrainement et 25% de la population de validation, en utilisant la méthode CD mean)**

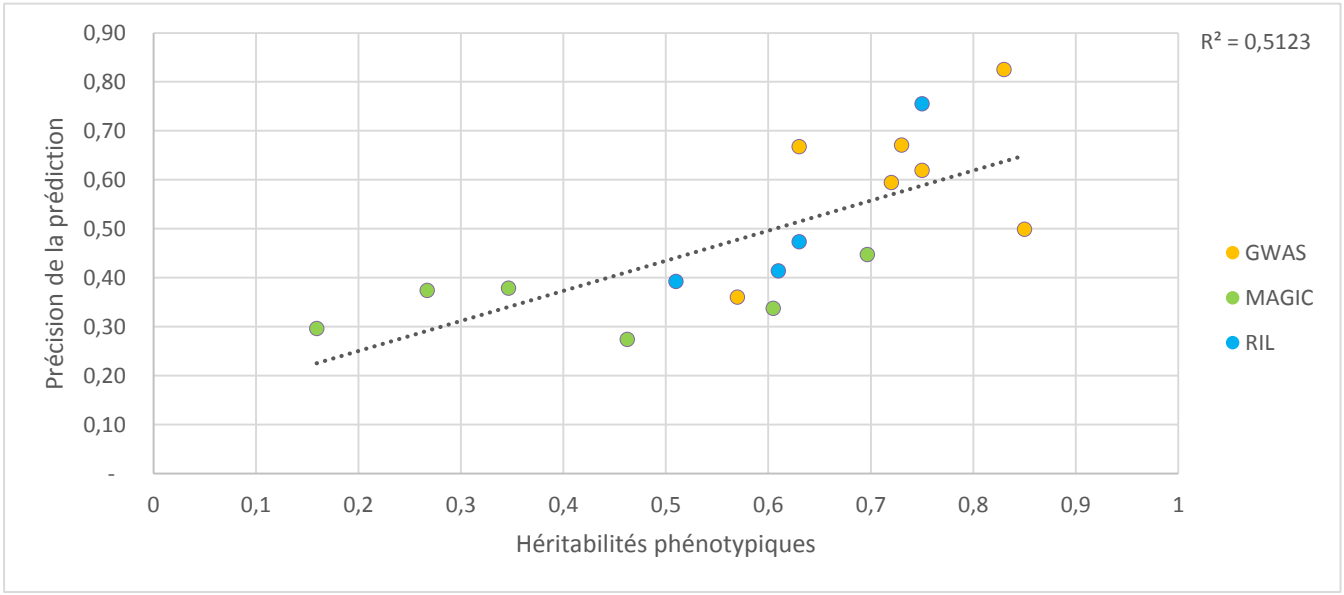



**Figure 13 : Précision de la prédiction du modèle Bayes C pour les phénotypes acidité, teneur en matières solubles, sucres, fermeté, poids du fruit, nombre de loges et pH pour le jeu de données GWAS ; brix, pH, fermeté, poids du fruit, hauteur de la plante et acidité pour le jeu de données MAGIC ; sucres, poids du fruit, pH et fermeté pour le jeu de données RIL, en fonction des héritabilités phénotypiques (75% des individus font partie de la population d'entrainement et 25% de la population de validation)**

conduit en moyenne à une augmentation de la précision de la prédiction de 9% (pour les modèles RKHS et Bayes C) avec un maximum de 21% pour la prédiction du contenu en Aspartate (pour les modèles RKHS et Bayes C également) dans le jeu de données GWAS. Dans tous les cas (modèle RKHS et modèle Bayes C, pour tous les jeux de données) l'utilisation de la méthode CDmean conduit à une prédiction plus précise (p-values comprises entre $2.2x10^{-04}$ et $5.25x10^{-45}$) sauf pour le fructose du jeu de données GWAS avec le modèle RKHS et la fermeté du jeu de données RIL avec le modèle Bayes C où la différence n'est pas significative avec le test de Student (les p-values sont de 0,31 et 0,29).

Par exemple, pour le jeu de données GWAS 2, lorsque la totalité des individus est prise en compte dans l'étude (231 individus), avec la population d'entrainement formée de 75%, 50% et 25% des individus en utilisant la méthode CDmean, la précision de la prédiction est en moyenne de 0.57, 0.46 et 0.40 respectivement contre 0.45, 0.43 et 0.35 sans cette méthode d'optimisation de la population d'entrainement.

Enfin, la précision de la prédiction des 5 phénotypes du jeu de données GWAS 2 est en moyenne supérieure de 2% avec le modèle RKHS par rapport aux mêmes phénotypes du jeu de données GWAS, et inférieure de 3% avec le modèle Bayes C.

En ce qui concerne les effets attribués aux marqueurs avec le modèle Bayes C, pour le jeu de données GWAS, plus le nombre d'individus total étudié est faible, plus les effets attribués aux marqueurs sont répartis de façon homogène sur le génome (annexe 5 a,b). Egalement, pour tous les phénotypes, la répartition des effets des marqueurs est plus hétérogène lorsque l'on considère le jeu de données avec un choix optimisé des individus (CDmean) avec 75% des individus pris en compte dans la population d'entrainement que lorsque l'on n'optimise pas le choix des individus dans les populations d'entrainement et de validation. Enfin, lorsque le nombre de marqueurs pris en compte dans le modèle est diminué, les effets attribués aux marqueurs sont également répartis de façon plus homogène sur le génome (annexe 5 a,c).

Pour les données RIL et MAGIC, cette variation des effets des marqueurs en fonction des paramètres n'est pas aussi marquée, hormis lorsque l'on diminue le nombre de marqueurs pris en compte dans le modèle. Ainsi, pour le phénotype pH, 2 marqueurs se voient attribuer un fort effet ($5.3x10^{-05}$ et $6x10^{-05}$) et 6 un effet moyen (de $2.2x10^{-05}$ à $3.7x10^{-05}$) en ce qui concerne le jeu de données MAGIC avec le paramètre CDmean et 75% d'individus dans la population d'entrainement. Cette répartition des effets ne varie pas en fonction des paramètres sauf lorsque l'on diminue le nombre de marqueurs. Les mêmes observations sont faites sur les jeux de données RIL (annexe 5 d, e, f, g)

### 3.1.4 Héritabilité phénotypique

Les phénotypes étudiés possèdent des héritabilités différentes (tableau 2). La précision de la prédiction varie en fonction de ces phénotypes, et ce, en utilisant des paramètres d'analyses identiques. Par exemple, avec le modèle Bayes C et la méthode CDmean lorsque 75% des individus font partie la population d'entrainement et 25% de la population de validation, la précision de la prédiction du jeu de données GWAS varie de 0.31 pour l'aspartate et 0.93 pour le poids du fruit. L'annexe 6 illustre ces propos pour les jeux de données GWAS, MAGIC et RIL.

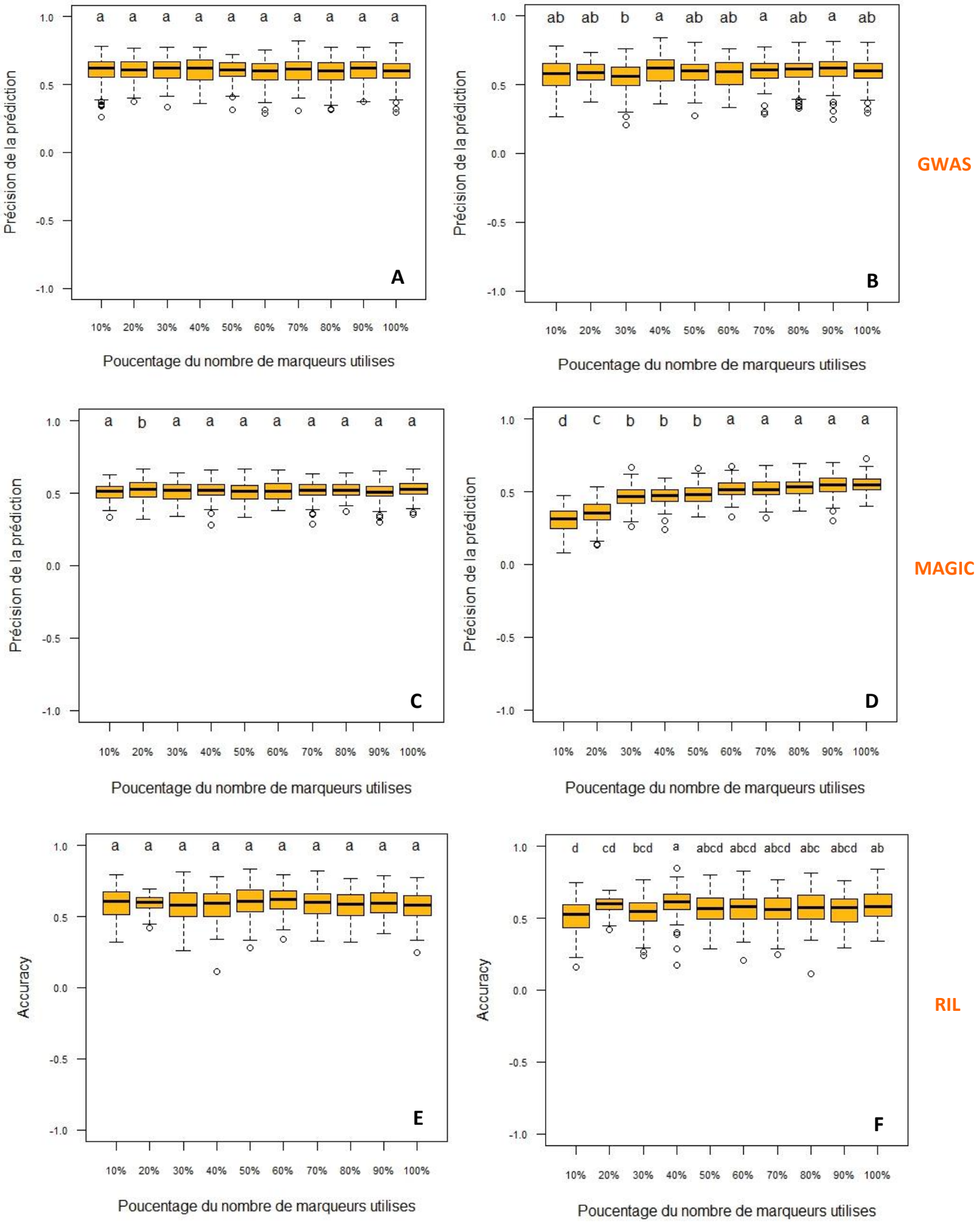



**Figure 14 : Diagrammes en boîte indiquant la précision de la prédiction des modèles RKHS (A, C, E) et Bayes C (B, D, F) pour les jeux de données GWAS (A et B), MAGIC (C et D) et RIL (E et F) sur le caractère phénotypique « Fermeté » (75% des individus font partie de la population d'entrainement et 25% de la population de validation) en fonction du nombre de marqueurs utilisés pour la prédiction**

De plus, pour un même caractère phénotypique, la précision de la prédiction dépend du jeu de données utilisé (figure 12). En effet, la fermeté est mieux prédite dans le jeu de données GWAS ($r^2$=0.80) que dans le jeu de données MAGIC ($r^2$=0.57) ou dans le jeu de données RIL ($r^2$= 0.60). A l'inverse, le jeu de données MAGIC permet une meilleure prédiction du pH ($r^2$=0.65) que les deux autres populations ($r^2$=0.49 pour GWAS et 0.45 pour RIL). En ce qui concerne le poids du fruit, les jeux de données GWAS et RIL proposent une meilleure prédiction que le jeu de données MAGIC ($r^2$= 0.93, $r^2$=0.88 et $r^2$=0.65 respectivement).

Enfin, la précision de la prédiction a tendance à augmenter lorsque l'héritabilité (au sens large) du phénotype est élevée (figure 13), et ce quel que soit le jeu de données utilisé (GWAS, MAGIC ou RIL). Le coefficient de régression linéaire entre l'héritabilité phénotypique et la précision de la prédiction est de 0.51. De plus, le coefficient de Pearson entre ces deux variables est de 0.72 (p-value = $1x10^{-0.3}$). Globalement, les phénotypes les mieux prédits sont donc ceux avec une forte héritabilité.

### 3.1.5 Densité de Marquage

L'étude de l'influence de la densité de marquage sur la précision de la prédiction a été en premier lieu réalisée en visualisant l'effet de la diminution du nombre de marqueurs sur la précision dans les jeux de données GWAS, MAGIC, et RIL pour tous les phénotypes.

Lorsque l'on utilise le modèle Bayes C avec le jeu de données MAGIC, la diminution du nombre de marqueurs entraine la baisse de la précision de la prédiction. Cette tendance est observée pour tous les phénotypes du jeu de données MAGIC avec le modèle Bayes C. Par exemple, lorsque l'on utilise 100% des marqueurs disponibles pour la prédiction de la fermeté, la précision de la prédiction est de 0.55. Lorsque l'on en utilise moins de 50%, la précision diminue progressivement jusqu'à une moyenne de 0.31 (figure 14D). Toutefois, pour un phénotype, plus les effets associés aux marqueurs sont répartis sur l'ensemble des marqueurs (beaucoup de marqueurs à moyens effets), moins la précision de la prédiction va chuter rapidement lorsque l'on diminue le nombre de marqueurs pris en compte dans l'étude.

En ce qui concerne les jeux de données GWAS et RIL (avec les modèles RKHS et Bayes C) et le jeu de données MAGIC (avec le modèle RKHS), la diminution du nombre de marqueurs n'entraine pas la baisse de la précision de la prédiction (figure 14).

La répartition des effets des marqueurs varie suivant la population étudiée : ils sont mieux répartis pour les jeux de données GWAS que pour le jeu de données RIL. Par exemple, pour le jeu de données GWAS, plus de 55% des marqueurs ont un effet inférieur à 0.01% des effets totaux attribués à l'ensemble des marqueurs pour tous les phénotypes. Les phénotypes sont donc prédits à partir de peu de marqueurs à forts effets et de nombreux marqueurs à plus faibles effets. A l'inverse, pour le jeu de données RIL, plus de 35% des marqueurs ont un effet supérieur à 0.19% des effets totaux attribués à l'ensemble des marqueurs pour tous les phénotypes. Les phénotypes sont donc prédits à partir d'une grande proportion de marqueurs à forts effets et d'une faible proportion de marqueurs à effets plus faibles. En ce qui concerne le jeu de données MAGIC, moins de 10% des marqueurs ont un effet supérieur à 0.19% des effets totaux attribués à l'ensemble des marqueurs et moins de 20% un effet inférieur à 0.01% (figure 15).

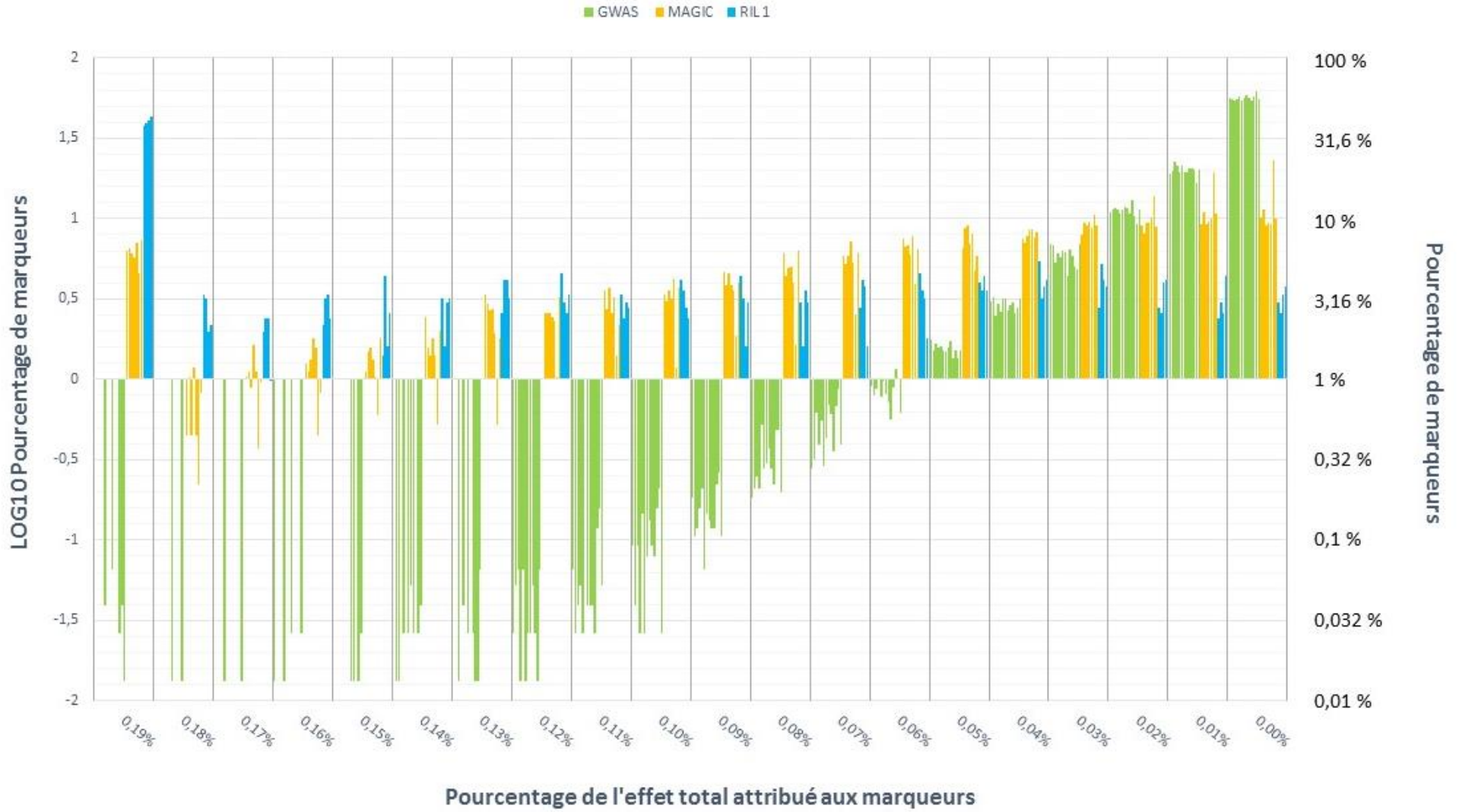



**Figure 15 : $Log_{10}$ du pourcentage de marqueurs (multiplié par 100) parmi la totalité des marqueurs pour chaque jeu de données (en ordonnée), en fonction du pourcentage de l'effet total attribué aux marqueurs (en abscisse) pour les phénotypes brix, fructose, poids du fruit, pH et sucres pour les jeux de données GWAS, MAGIC et RIL. L'unité en ordonnée a été choisie afin de comparer les 3 populations de manière équitable (pourcentage du nombre de marqueurs parmi la totalité des marqueurs pour chaque jeu de données) et de façon à mettre en évidence le pourcentage de marqueurs, qu'il soit fort ou faible (les valeurs négatives correspondant à un pourcentage de marqueurs compris entre 0.01% à 1%). En ordonnée à droite, les correspondances des pourcentages des marqueurs avec le $Log_{10}$ du pourcentage de marqueurs multiplié par 100 sont indiquées. Ce graphique permet de visualiser la répartition des effets des marqueurs : lorsqu'une grande proportion des marqueurs (en ordonnée) correspond à un faible pourcentage de l'effet total attribué aux marqueurs (en abscisse), et une faible proportion des marqueurs correspond à un fort pourcentage de l'effet total attribué aux marqueurs, le phénotype est caractérisé par de nombreux marqueurs à faibles effets et peu de marqueurs à effets forts, et inversement. Les effets des marqueurs correspondent à une étude effectuée avec la méthode d'optimisation CDmean avec 75% des individus dans la population d'entrainement**

D'autre part, en ce qui concerne la comparaison de la prédiction entre les jeux de données GWAS 2 (6 768 marqueurs) et GBS (59 079 marqueurs), qui comporte les mêmes 63 individus, les gains et des pertes de la précision de la prédiction sont observés en fonction des phénotypes (figure 16). En effet, la précision de la prédiction augmente pour le °brix (0.74 en moyenne pour GWAS 2 et 0.86 pour GBS), le poids du fruit (0.80 pour GWAS 2 et 0.82 pour GBS) et le pH (0.55 pour GWAS 2 et 0.80 pour GBS). Néanmoins, elle diminue pour le fructose (0.65 pour GWAS 2 et 0.57 pour GBS) et les sucres (0.65 pour GWAS 2 et 0.48 pour GBS).

Le spectre des fréquences allélique du jeu de données GBS indique que plus de 70% des marqueurs ont une MAF inférieure à 0.2 et plus de 4.4% une MAF inférieure à 0.1, ce qui est dû aux paramètres de sélection des SNPs (filtrés avec une MAF à 0.05). Egalement, le spectre de fréquence allélique du jeu de données GWAS 2 (avec 63 individus) indique que plus de 35% des marqueurs ont un allèle dont la fréquence est inférieure à 0.05 (annexe 3E). De plus, 242 marqueurs et 2,747 marqueurs ont un allèle dont la fréquence est inférieure à 0.05 pour les jeux de données GBS et GWAS 2 (63 individus) respectivement.

Egalement, le résultat de la répartition des effets des marqueurs sur le génome est illustré figure 17 (75% des individus font partie de la population d'entrainement et la méthode CDmean est utilisée). Elle indique que pour le jeu de données GBS, les effets attribués aux marqueurs sont mieux répartis que pour le jeu de données GWAS 2. Ainsi, pour le °brix, 24.78% des marqueurs ont un effet supérieur à 0.02% des effets attribués à l'ensemble des marqueurs pour le jeu de données GWAS. Pour le même phénotype en ce qui concerne le jeu de données GBS, aucun marqueur n'a un effet supérieur à 0.02% des effets totaux attribués à l'ensemble des marqueurs. Egalement, pour tous les phénotypes, plus de la moitié des marqueurs ont un effet inférieur à 0.001% des effets totaux attribués à l'ensemble des marqueurs pour le jeu de données GBS et plus de 20% en ce qui concerne le jeu de données GWAS.

### 3.1.6 Interaction G × E

L'interaction GxE est étudiée en utilisant une population GWAS cultivée en 2007 et en 2008. Les précisions de la prédiction évaluées à partir des individus de la même année ou d'une année différente sont comparées. Les résultats sont présentés sur la figure 18. Egalement, les comparaisons des moyennes et des variances des différents phénotypes se trouvent en annexe 7. Les phénotypes « contenu en aspartate et en threonate » sont mieux prédits lorsque la PE et la PV sont constituées des individus de la même année que lorsque qu'elles sont constituées d'individus issus de 2 années différentes. Lorsque les deux jeux de données des phénotypes sont comparés (2007 et 2008), le test de comparaison de moyenne (test de Student) indique que les moyennes des phénotypes sont significativement très différentes (p-value = $5.13x10^{-06}$ pour l'aspartate et p-value = $2.03^{e}$-05 pour le threonate) et le test de comparaison des variances (test de Barlett) indique que les variances ne sont pas significativement différentes (p-value = $6.78x10^{-02}$ pour l'aspartate et p-value = $6.02^{e}$-01 pour le threonate).

La précision de la prédiction pour le contenu en ASA est plus élevée avec les individus de 2007 dans la PE et la PV qu'avec les individus de 2008. Néanmoins, la prédiction des individus de 2007 à partir des individus de 2008 (et inversement) est meilleure que la prédiction des individus de 2008 à partir des individus de 2008 et plus faible que la prédiction des individus de 2007 à partir des individus de 2007. En ce qui concerne le °brix, la précision de la prédiction est plus élevée avec les

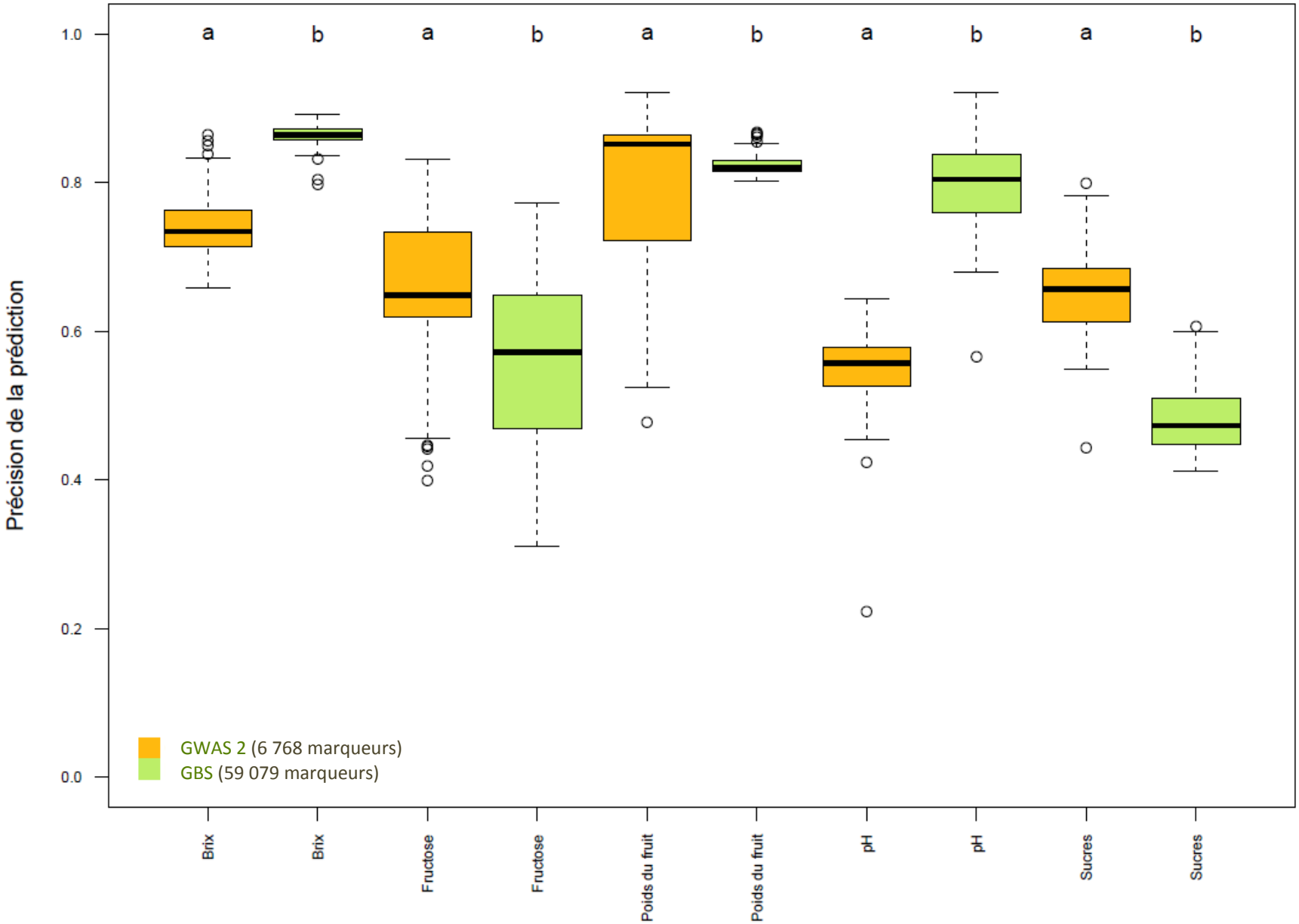



**Figure 16 : Précision de la prédiction des caractères phénotypiques Brix, Fructose, Poids du fruit, pH et Sucres du modèle Bayes C pour les populations GWAS 2 avec 6 768 marqueurs (en jaune) et GBS avec 59 079 marqueurs (en vert) avec l'utilisation de la méthode CDmean et 75% des individus dans la population d'entrainement**

individus de 2008 dans la PE et la PV qu'avec les individus de 2007. Néanmoins, la prédiction des individus de 2007 à partir des individus de 2008 (et inversement) est meilleure que la prédiction des individus de 2007 à partir des individus de 2007 et plus faible que la prédiction des individus de 2008 à partir des individus de 2008. Pour les deux caractères, les moyennes entre les deux années sont significativement différentes (p-value = 6.52 x10 $^{-01}$ pour l'ASA et p-value = 3.76 x 10 $^{-04}$ pour le °brix) mais les variances sont égales (p-value = 0.65 pour l'ASA et p-value = 0.56 pour le °brix).

Quelle que soit la composition des PE et PV, la précision de la prédiction du poids du fruit est équivalente. La moyenne et la variance entre les deux années ne sont pas significativement différentes (p-value = 6.78 x 10 $^{-02}$ et p-value = 8.52 x 10 $^{-01}$ respectivement).

Enfin, la précision de la prédiction pour le contenu en malate est plus élevée avec les individus de 2008 dans la PE et la PV qu'avec les individus de 2007. De plus, la prédiction des individus de 2007 à partir des individus de 2008 est équivalente à la prédiction des individus de 2007 à partir de 2007 et la prédiction des individus de 2008 à partir des individus de 2007 est équivalente à la prédiction des individus de 2008 à partir de 2008. La moyenne entre les deux années n'est pas significativement différente (p-value = 1.92 x 10 $^{-01}$) mais les variances ne sont égales (p-value = 9.56 x 10 $^{-08}$).

### 3.1.7 Analyse des effets attribués aux marqueurs : comparaison avec une étude de GWAS

La comparaison du jeu de données GWAS avec l'étude de génétique d'association est présentée en figure 19. Pour chacun des sept phénotypes étudiés, les Manhattan plot ont été réalisés par Sauvage et al. (2014) (-log10(p-value) vs position chromosomique). Des marqueurs ont été mis en évidence avec la GWA (en rouge sur les graphiques). Les marqueurs correspondant à 10% des effets totaux attribués à l'ensemble des marqueurs en SG sont positionnés sur ces Manhattan plots (en jaune sur les graphiques).

Les marqueurs mis en évidence en génétique d'association (21 marqueurs au total sur les 7 phénotypes étudiés) correspondent aux marqueurs dont les effets sont les plus élevés en SG, ou à des marqueurs très proches. En effet, parmi les marqueurs correspondant à 10% des effets totaux attribués à l'ensemble des marqueurs en SG, 5 marqueurs ont été identifiés comme significatifs en GWAS, ils correspondent aux phénotypes ASA, fructose, °brix, sucrose et malate. De plus, les marqueurs dont les effets sont élevés en SG qui sont proches des marqueurs mis en évidence par la GWAS se situent dans des régions du génome où le déséquilibre de liaison est élevé. En effet, parmi les marqueurs mis en évidence par la GWAS, 4 sont en déséquilibre de liaison avec au moins un marqueur ayant un fort effet en SG (faisant partie du groupe de marqueurs correspondant à 10% des effets totaux attribués à l'ensemble des marqueurs).

Ainsi, le marqueur mis en évidence pour le threonate par la GWAS est corrélé avec des valeurs de 0.78 et 0.84 avec deux marqueurs ayant des forts effets en SG. Le marqueur mis en évidence pour l'aspartate est corrélé avec trois marqueurs (les valeurs de corrélation du DL sont de 0.78, 0.82 et 0.84). Un des marqueurs mis en évidence par la GWAS pour la teneur en ascorbate est corrélé avec un marqueur (la valeur de corrélation du DL est de 0.28). Enfin, le marqueur mis en évidence pour le contenu malate par l'approche de GWAS est corrélé avec une valeur d'effet de 0.76 avec un marqueur ayant un fort effet en SG.

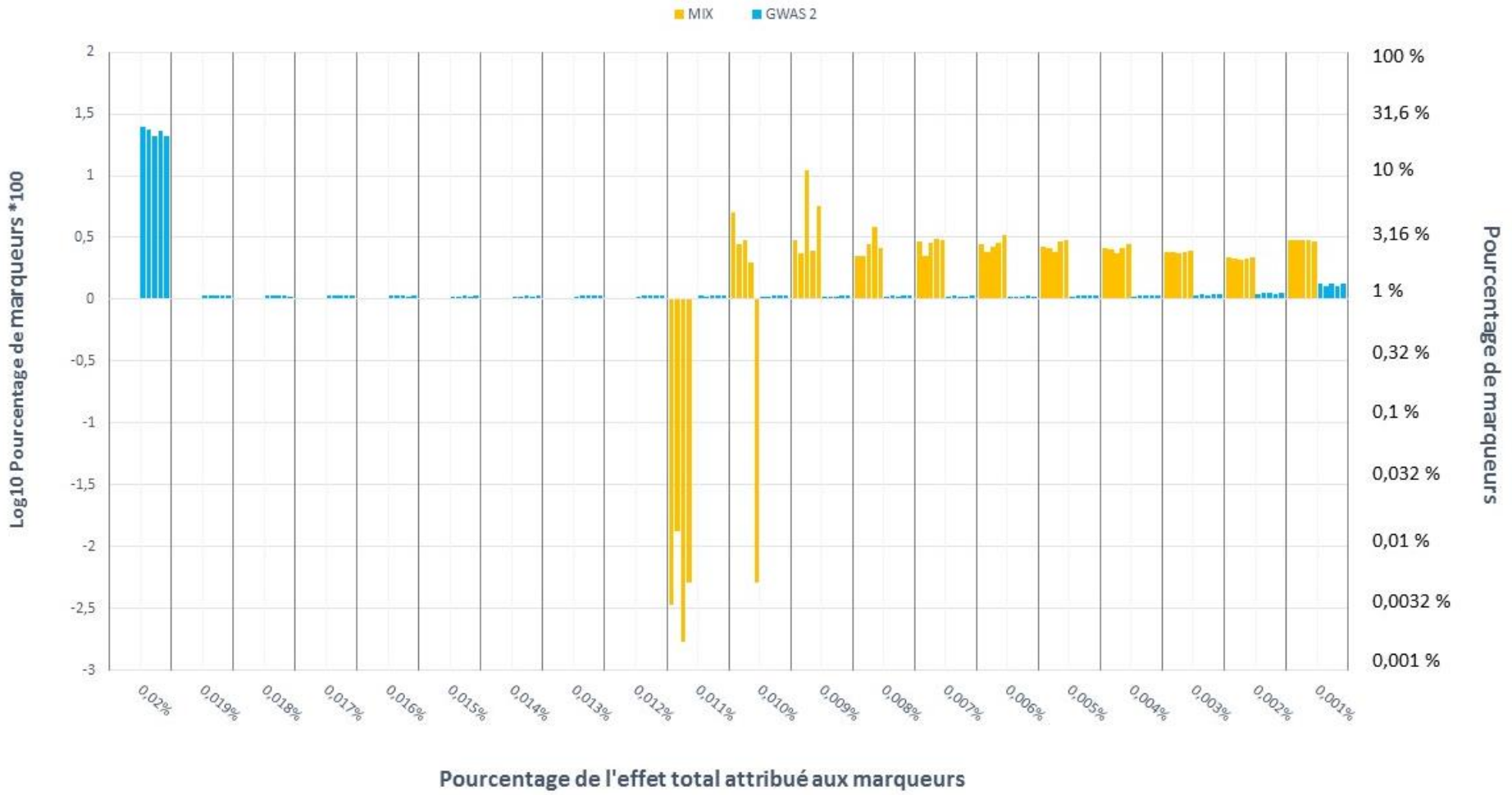



**Figure 17 : $Log_{10}$ du pourcentage de marqueurs (multiplié par 100) parmi la totalité des marqueurs pour chaque jeu de données (en ordonnée), en fonction du pourcentage de l'effet total attribué aux marqueurs (en abscisse) brix, fructose, poids du fruit, pH et sucres pour les jeux de données GWAS 2 et GBS avec 63 individus en commun. L'unité en ordonnée a été choisie afin de comparer les 3 populations de manière équitable (pourcentage du nombre de marqueurs parmi la totalité des marqueurs pour chaque jeu de données) et de façon à mettre en évidence le pourcentage de marqueurs, qu'il soit fort ou faible (les valeurs négatives correspondant à un pourcentage de marqueurs compris entre 0.01% à 1%). En ordonnée à droite, les correspondances des pourcentages des marqueurs avec le Log10 du pourcentage de marqueurs multiplié par 100 sont indiquées. Ce graphique permet de visualiser la répartition des effets des marqueurs : lorsqu'une grande proportion des marqueurs (en ordonnée) correspond à un faible pourcentage de l'effet total attribué aux marqueurs (en abscisse), et une faible proportion des marqueurs correspond à un fort pourcentage de l'effet total attribué aux marqueurs, le phénotype est caractérisé par de nombreux marqueurs à faibles effets et peu de marqueurs à effets forts, et inversement. Les effets des marqueurs correspondent à une étude effectuée avec la méthode d'optimisation CDmean avec 75% des individus dans la population d'entrainement**

# 4 Discussion

L'objectif de cette étude est d'étudier l'effet de paramètres sur la précision de la prédiction de phénotypes liés à la qualité du fruit chez la tomate, dans le but d'étudier la faisabilité d'introduire la sélection génomique dans un schéma de sélection. Les résultats obtenus vont être discutés dans cette partie dans un ordre similaire à celui des parties « Matériel et Méthodes » et « Résultats ».

### 4.1.1 Protocole d'utilisation des modèles de prédiction

Le choix d'optimisation du nombre de cycles des modèles a été fait à partir du modèle BL et dans des conditions précises (75% des individus font partie de la population d'entrainement et 25% de la population de validation). La suite de l'étude considère que ces résultats sont également valables pour les autres modèles (et notamment les modèles Bayes C et RKHS) ainsi qu'avec les autres paramètres.

Néanmoins, dans la première partie du schéma de sélection en SG (c'est-à-dire lors de l'inférence), le nombre de cycles nécessaire doit être testé en fonction de tous les paramètres et en fonction du modèle utilisé afin de ne pas biaiser certaines moyennes de précision de la prédiction.

Dans la littérature, le nombre de cycles utilisés pour la prédiction est plus faible, généralement autour de 10. Par exemple, le nombre de cycles de l'étude de Fodor et al. (2014) sur la prédiction de caractères concernant la vigne est seulement de 10. Les résultats montrent que la précision de la prédiction peut varier de +/- 0.15. De même, dans l'étude de Kumar et al. (2014), 10 cycles ont permis de prédire des caractères concernant la pomme avec une variation de +/- 0.5. Néanmoins, le nombre d'individus utilisés dans ces deux publications est de 3000 et 1200 respectivement. La précision de la prédiction augmente avec le nombre d'individus considérés dans le modèle, l'optimisation du nombre de cycles en dépend donc aussi.

Dans un schéma de sélection, il parait donc pertinent de faire le choix du nombre de cycles en testant le modèle utilisé sur la population en question.

### 4.1.2 Différences entre les modèles de prédiction

Chez la population GWA et pour les phénotypes testés, les modèles de prédiction n'ont pas produit de résultats significativement différents hormis en ce qui concerne les modèles RKHS et G-BLUP (qui conduit à une grande variabilité des résultats). Ces conclusions sont confortées par le fait que des résultats similaires aient été trouvés dans la bibliographie. En effet, Daetwyler et al. (2013) ont testé les modèles Bayes A, Bayes B, Bayes C, BRR, BL et G-BLUP sur des jeux de données chez le blé et le pin et aucune différence n'a été observée quant à la précision de la prédiction. Egalement, Howard et al. (2014) ont testé l'influence de l'utilisation de modèles paramétriques et de modèles non ou semi-paramétriques sur la précision de la prédiction. Ils démontrent que les méthodes paramétriques ne sont pas efficaces que lorsqu'il s'agit de prédire des caractères basés entièrement sur l'épistasie, contrairement aux modèles non ou semi-paramétriques. A l'inverse, les modèles paramétriques sont plus efficaces lorsque le caractère étudié porte sur des effets additifs.

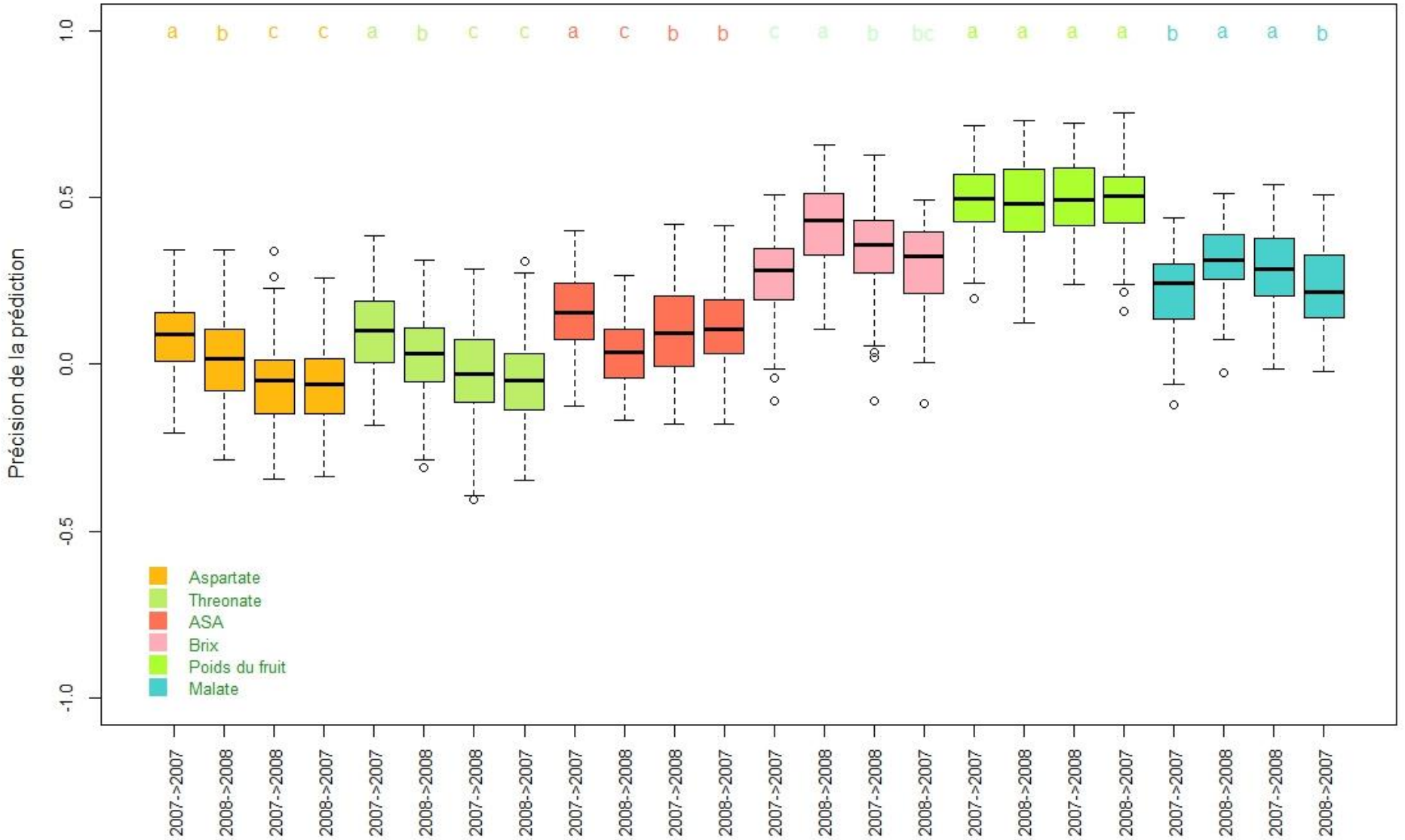



**Figure 18 : Précision de la prédiction avec le modèle Bayes C (75% des individus font partie de la population d'entrainement) des caractères phénotypiques « Aspartate, Threonate, ASA, Brix, Poids du Fruit et Malate » du modèle Bayes C pour la population GWAS en fonction de la composition des populations d'entrainement (PE) et de validation (PV) : 2007->2007 : prédiction des individus de 2007 à partir des individus de 2007 ; 2008->2008 : prédiction des individus de 2008 à partir des individus de 2008 ; 2007->2008 : prédiction des individus de 2008 à partir des individus de 2007 ; 2008->2007 : prédiction des individus de 2007 à partir des individus de 2008 . Les lettres en haut sont le résultat du test HSD : pour une même lettre, les paramètres conduisent à des résultats qui ne sont pas significativement différents.**

Ce résultat a également été démontré par Gianola et al. (2010) qui explique que le modèle RKHS prend en compte les effets de dominance et d'épistasie.

Cette étude conforte également le fait que lorsque l'on compare le modèle RKHS (semi paramétrique) et le modèle Bayes C (paramétrique) pour des phénotypes à faible héritabilité phénotypique, c'est-à-dire avec une faible part de variation d'origine génétique dans la variation phénotypique totale, le modèle RKHS permet d'obtenir de meilleures prédictions pour les phénotypes testés dans cette étude. De même, lorsque l'on diminue le nombre de marqueurs des jeux de données MAGIC ou RIL, la prise en compte des effets épistasie joue un rôle majeur dans la prédiction puisque le modèle RKHS devient plus efficace. Cette observation n'est pas faite sur le jeu de données GWAS car la diminution du nombre de marqueurs n'influence pas la précision de la prédiction.

Pour des phénotypes avec des héritabilités plus élevées ($h^2$>0.5), le modèle Bayes C est plus efficace que le modèle RKHS hormis lorsque le nombre d'individus dans les populations d'entrainement et de validation devient trop faible (124 individus dans la population d'entrainement - jeu de données GWAS).

Dans un schéma de sélection, le choix du modèle de prédiction peut donc être important, notamment entre les modèles paramétriques (à privilégier lorsque le phénotype étudié dispose d'une forte héritabilité au sens strict) et semi/non paramétriques (à privilégier lorsque le caractère étudié est contrôlé en grande partie par des phénomènes d'épistasie ou de dominance). Néanmoins, en ce qui concerne les modèles paramétriques, il parait pertinent d'utiliser un seul modèle afin d'optimiser le temps de calcul.

### 4.1.3 Composition de la population d'entrainement et de la population de validation

Dans notre étude, les résultats obtenus démontrent que la précision de la prédiction augmente avec la taille de la population étudiée car l'accumulation des informations phénotypiques et génotypiques rend plus robuste l'estimation des effets des marqueurs (Hayes et al., 2009), ce qui est en adéquation avec les travaux présentés en SG dans la littérature (Varaden et al., 2010 ; Technow et al., 2013 ; Grattapaglia et al., 2014). Ce résultat peut être mis en relation avec l'observation de la répartition des effets des marqueurs le long du génome. En effet, plus le nombre d'individus étudiés est élevé et plus le nombre de marqueurs pris en compte est grand, plus les effets des marqueurs sont répartis de manière hétérogène. En d'autres termes, les marqueurs sont plus différenciés les uns des autres par le modèle, ce qui rend la précision de la prédiction plus élevée.

De plus, nous avons cherché à étudier l'effet du choix non-aléatoire des individus dans la population d'entrainement par le biais d'une méthode d'optimisation (CDmean, Rincent et al., 2012) qui permet de répartir les individus dans les populations d'entrainement et de validation afin qu'elles soient, génétiquement, le plus diversifiées possible. La méthode du CDmean conduit ici aux meilleurs résultats de prédiction ce qui est également le cas dans l'étude de Cros (2014b) chez le palmier à huile. Il est donc important de répartir les individus de manière à ce que les deux populations soient les plus diversifiées possibles pour optimiser les résultats en SG. Cela permet en effet de répartir les allèles rares dans les populations d'entrainement et de validation. Néanmoins, il existe d'autres algorithmes d'optimisation du choix des individus d'une population et certains

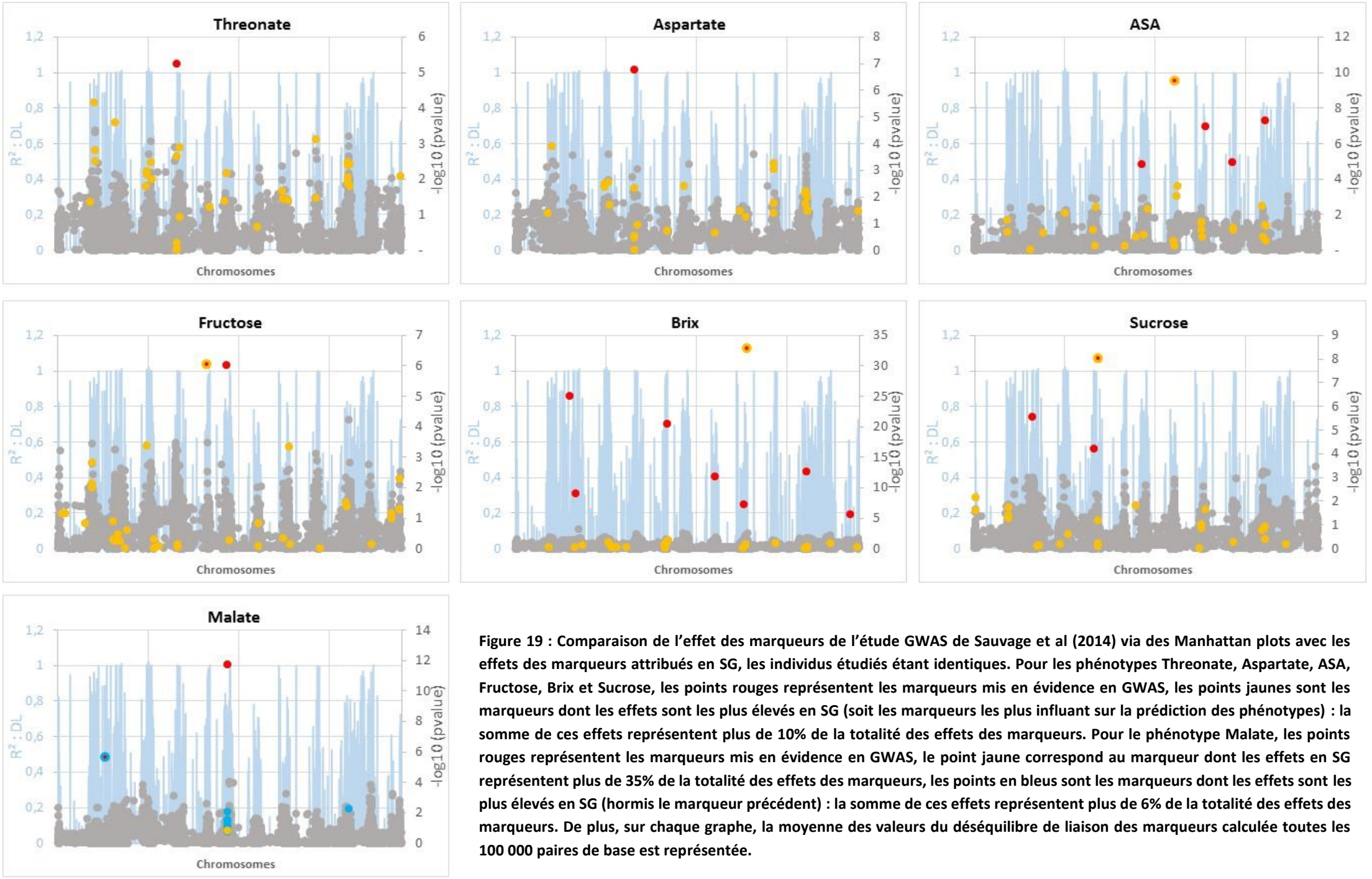



**Figure 19 : Comparaison de l'effet des marqueurs de l'étude GWAS de Sauvage et al (2014) via des Manhattan plots avec les effets des marqueurs attribués en SG, les individus étudiés étant identiques. Pour les phénotypes Threonate, Aspartate, ASA, Fructose, Brix et Sucrose, les points rouges représentent les marqueurs mis en évidence en GWAS, les points jaunes sont les marqueurs dont les effets sont les plus élevés en SG (soit les marqueurs les plus influant sur la prédiction des phénotypes) : la somme de ces effets représentent plus de 10% de la totalité des effets des marqueurs. Pour le phénotype Malate, les points rouges représentent les marqueurs mis en évidence en GWAS, le point jaune correspond au marqueur dont les effets en SG représentent plus de 35% de la totalité des effets des marqueurs, les points en bleus sont les marqueurs dont les effets sont les plus élevés en SG (hormis le marqueur précédent) : la somme de ces effets représentent plus de 6% de la totalité des effets des marqueurs. De plus, sur chaque graphe, la moyenne des valeurs du déséquilibre de liaison des marqueurs calculée toutes les 100 000 paires de base est représentée.**

sont plus adaptés suivant la structure de la population (Isidro, 2014). Les méthodes CDmean et StratCDmean seraient adaptés à des populations présentant une structure populationnelle peu marquée et l'algorithme « stratified sampling » à des populations avec une forte structure. Il pourrait être intéressant de comparer les deux méthodes avec nos jeux de données, et notamment pour le jeu de données GWAS qui présente de la structure.

La valeur de la précision de la prédiction est évaluée à partir d'une population de validation. Si la population test n'appartient pas à la même génération ou est trop éloignée génétiquement de la population de validation, la précision de la prédiction sera sûrement différente (Nakaya, 2012 ; Wang et al., 2014). Il serait donc intéressant ici de tester la variation de la précision de la prédiction des phénotypes chez la tomate en fonction de la génération de la population de validation comme l'ont déjà effectué Kizilkaya et al. (2010) sur le blé. Ainsi, la précision de la prédiction pourrait être évaluée en fonction de l'écart entre les populations d'entrainement et de validation (prédiction d'une génération x en fonction des générations x-1, x-2, x-3 etc.) ou du nombre de générations d'individus se trouvant dans la population d'entrainement (prédiction de la génération x en fonction des générations x-1 et x-2, en fonction des générations x-1, x-2 et x-3 etc.).

Pour finir, il aurait été intéressant d'évaluer la précision de la prédiction des phénotypes d'un jeu de données en fonction d'un jeu ou plusieurs jeux de données différents (par exemple, GWAS vs MAGIC ou GWAS vs MAGIC + RIL). Cela permettrait d'augmenter le nombre d'individus de la PE ainsi que sa diversité, semblablement à ce qui est fait chez les bovins. Néanmoins, des marqueurs identiques entre les deux jeux de données auraient été nécessaires.

Dans un schéma de sélection, il est donc important d'optimiser le choix de la répartition des individus dans les populations d'entrainement et de validation. Egalement, le nombre d'individus de la population d'entrainement doit être suffisamment élevé pour obtenir une forte valeur de précision de la prédiction. Cela signifie que dans la deuxième partie du schéma de sélection, lorsque la population d'entrainement est utilisée pour estimer les BV d'une population de test dont on ne connait pas le phénotype, plus la population d'entrainement est grande, plus la valeur de prédiction du phénotype se rapprochera la vraie valeur moyenne du caractère.

### 4.1.4 Héritabilité phénotypique

Les résultats de cette partie nous permettent de dire que la sélection génomique est un outil puissant qui doit être utilisé avec prudence selon les phénotypes étudiés. En effet, la SG ne sera que peu ou pas efficace sur des phénotypes à faible héritabilité, cette technique n'est donc pas forcément pertinente dans ce cas. Nos résultats sont confortés par les conclusions de certains auteurs comme la partie 1.3.3 le détaille (Legarra et al., 2008 ; Hayes et al., 2009 ; Lorenz et al., 2011 ; Grattapaglia et al., 2014 ).

Les travaux de Kumar et al. (2014) portant sur la prédiction de caractères concernant la qualité de la pomme (avec 1120 individus et 8000 SNPs) obtiennent une précision de la prédiction de 0.83 pour la fermeté, 0.89 pour la teneur en sucres solubles et 0.81 pour l'acidité. Les résultats de précision de la prédiction chez la tomate sont de 0.80, 0.80 et 0.78 pour ces caractères respectivement, avec le jeu de données GWAS avec le modèle Bayes C, ce qui laisse penser que quelle que soit l'espèce étudiée, les caractères phénotypiques ont la même architecture génétique.

De plus, la précision de la prédiction varie pour un même phénotype mais pour des jeux de données différents. Le jeu de données GWAS conduit le plus souvent à des prédictions plus fortes, ce qui peut être dû à sa composition génétique. En effet, ce jeu de données présente de nombreux allèles à faible fréquence (annexe 3A) qui jouent un rôle dans le modèle de prédiction. Un des intérêts majeurs de la SG et donc de capter les faibles effets des marqueurs.

De plus, la différence de résultats observée entre les populations GWAS, MAGIC et RIL peut provenir du fait que le nombre de marqueurs et leur position sur le génome ne soit pas identiques. Ainsi, les gènes représentés par les marqueurs SNPs utilisés dans le jeu de données MAGIC sont peut-être moins impliqués dans la détermination du poids du fruit que les marqueurs des jeux de données GWAS et RIL.

Dans un schéma de sélection, la phase d'élaboration du modèle avec la population d'entrainement et la population de validation est donc importante car suivant le phénotype, la précision de la prédiction va varier. En fonction de cette valeur, le sélectionneur pourra décider de continuer en SG en appliquant le modèle sur une population test ou non.

### 4.1.5 Densité de Marquage

Globalement, la diminution du nombre de marqueurs dans le jeu de données MAGIC avec le modèle Bayes C engendre la baisse de la précision de la prédiction (à partir d'une diminution de 50% des marqueurs, c'est-à-dire environ 670 marqueurs). Le fait que le même phénomène ne soit pas observé sur le jeu de données GWAS vient peut-être du nombre de marqueurs qui est nettement supérieur. Ainsi, pour ce jeu de données, 10% des marqueurs correspondent à 771 marqueurs. Néanmoins, le jeu de données RIL ne possède que 501 marqueurs et cette diminution de la précision n'est pas autant observée que pour le jeu de données MAGIC.

Le déséquilibre de liaison des jeux de données GWAS et RIL est plus fort que pour le jeu de données MAGIC, il y a donc plus de marqueurs liés dans ces jeu de données. Les marqueurs jouant un rôle dans l'expression d'un phénotype sont donc plus nombreux. En conséquence, la probabilité d'enlever un marqueur jouant un rôle sur le phénotype est donc moins grande pour les jeux de données GWAS et RIL que pour le jeu de données MAGIC. La diminution du nombre de marqueurs pris en compte dans le modèle affecte donc moins la prédiction des jeux de données GWAS et RIL. Pour valider cette hypothèse, il est nécessaire d'entreprendre une étude en choisissant des marqueurs de façon pertinente sur le génome (c'est-à-dire en limitant le DL entre les marqueurs d'une part et en sélectionnant ceux ayant le plus d'effet sur les phénotypes d'autre part). De plus, un jeu de données séquencé avec les techniques NGS permettrait de mieux estimer le DL avec un plus grand nombre de marqueurs.

Egalement, le spectre de fréquence allélique du jeu de données RIL (annexe 3C) nous indique que très peu d'allèles rares se trouvent dans cette population. En effet, aucun allèle n'a une fréquence inférieure à 30%. A l'inverse, plus de la moitié des allèles du jeu de données MAGIC ont une fréquence inférieure à 25%. Si ces allèles à faible fréquence sont impliqués dans l'expression phénotypique, il est important qu'ils soient associés à des marqueurs pour obtenir des prédictions plus fortes.

En résumé, l'effet de la baisse du nombre de marqueurs utilisés dans le modèle sur la précision de la prédiction peut être influencé par la structure du DL dans la population, par la répartition des

effets des marqueurs sur le génome, et donc par l'architecture du caractère étudié. Par exemple, pour un phénotype, si peu de marqueurs se voient attribuer un effet fort, la diminution du nombre de marqueurs pris en compte dans le modèle n'influencera pas la prédiction si ces derniers sont choisis parmi ceux utilisés dans le modèle. A l'inverse, si ces marqueurs ne sont pas utilisés par le modèle lors de la prédiction, alors sa précision va chuter. Néanmoins, si de nombreux marqueurs se voient attribuer un effet moyen, la précision de la prédiction baisse progressivement si l'on diminue le nombre de marqueurs pris en compte dans l'étude.

D'autre part, la précision de la prédiction est en moyenne plus élevée lorsque 59 079 marqueurs issus du séquençage haut débit (GBS) sont utilisés plutôt que 6 768 issus d'une puce pour les phénotypes °brix, poids du fruit et pH, et inversement pour le fructose et les sucres (GWAS 2).

Le génotypage avec des techniques de séquençage haut débit permet de capter des SNPs situés dans de nombreux blocs en DL sur le génome, qui ont un effet mineur sur les phénotypes. C'est pourquoi les caractères °brix, poids du fruit et pH sont mieux prédits aves le jeu de données GBS qu'avec le jeu de données GWAS 2.

Néanmoins les caractères concernant les sucres (fructose et sucres totaux) sont mieux prédits par le jeu de données GWAS 2 qu'avec le jeu de données GBS. Les spectres de fréquence alléliques (annexe 3D et 3E) indiquent que très peu d'allèles rares se trouvent dans le jeu de données GBS contrairement au jeu de données GWAS 2, ce qui est du à la MAF appliquée sur les données NGS. Nous pouvons donc penser que ces allèles rares correspondent en partie à ceux liés aux QTL des sucres du fruit. De plus, une autre limite de cette étude concerne le nombre d'individus utilisés pour la prédiction. En effet, seulement 63 individus sont pris en compte par le modèle dans cette étude, ce qui est encore plus faible que le nombre d'individus pris en compte dans la partie de l'étude où l'influence du nombre d'individus sur la précision de la prédiction est étudiée (partie 2.2.3).

Il serait pertinent de comptabiliser le nombre de SNPs GBS qui font également partie des SNPs de la puce (ou qui se trouvent en DL) afin de visualiser si certains marqueurs mis en évidence avec le jeu de données GWAS 2 en ce qui concerne les sucres se situent également dans le jeu de données GBS. De plus, il serait intéressant d'évaluer la différence de la précision de la prédiction pour les jeux de données RIL et MAGIC entre l'utilisation de peu de marqueurs issus d'une puces et de nombreux marqueurs issus d'un séquençage haut débit.

Par ailleurs, une étude de l'impact du nombre de marqueurs utilisés dans le modèle en fonction des générations dans un schéma de sélection serait pertinente. En effet, au fur et à mesure des générations, le nombre de recombinaisons augmentent et le DL diminue. Le nombre de marqueurs nécessaires pour obtenir une forte précision de la prédiction tend donc à augmenter.

Dans un schéma de sélection, il est donc préférable de visualiser la structure et le DL de la population ainsi que les effets attribués aux marqueurs par le modèle pour chaque phénotype pour déterminer le nombre de marqueurs à utiliser. Cet aspect est économiquement important car le choix de la méthode de génotypage (choix de l'utilisation ou non des techniques de séquençage haut débit) et le temps passé pour la prédiction en dépendent. De plus, le choix du nombre de marqueurs à utiliser en SG dépendra de l'avancée de la recherche sur les phénotypes à améliorer. L'utilisation de données NGS pourra être efficace en SG si les marqueurs disponibles sur une puce ne recouvrent pas assez de QTLs. Afin de confirmer cette hypothèse, il pourrait être intéressant de

réitérer cette étude en choisissant des marqueurs sur une puce correspondant à des QTLs pour le °brix, le poids du fruit et le pH pour voir si la précision de la prédiction reste meilleure avec la puce qu'avec les données NGS.

### 4.1.6 Interaction G × E

Les phénotypes ayant une faible héritabilité phénotypique sont mieux prédits lorsque les populations d'entrainement et de validation sont constituées des individus de la même année. Ce résultat est dû au changement d'environnement. En effet, les moyennes des phénotypes des individus de 2007 sont significativement très différents des moyennes des individus de 2008.

A l'inverse, le poids du fruit possède une forte héritabilité phénotypique ($h^2 = 0.76$) et les valeurs moyennes de ce phénotype entre les deux années ne sont pas significativement différentes. La précision de la prédiction ne varie donc pas malgré l'implication d'individus de différentes années dans la prédiction.

Les phénotypes ASA, °brix et contenu en malate possèdent une héritabilité plus faible que le poids du fruit (0.55, 0.60 et 0.64 respectivement). Les moyennes entre les individus de 2007 et 2008 sont différentes pour l'ASA et le °brix et les variances sont différentes pour le malate. Cela explique pourquoi des différences sont observées en ce qui concerne la précision de la prédiction des individus de 2007 ou de 2008 à partir des individus de la même année.

Pour intégrer la SG dans un schéma de sélection, il est donc indispensable de regarder la répartition des données en fonction de l'interaction entre le génotype et l'environnement. Si les données ne sont pas significativement différentes bien qu'elles soient d'origines diverses (si l'environnement n'influe pas sur la moyenne et la variance des données phénotypiques), il est intéressant de les utiliser dans la population d'entrainement pour augmenter le nombre d'individus et donc la précision de la prédiction. Dans le domaine de la sélection des bovins, le projet « 1000 génomes » (Daetwyler et al., 2014) a pour but d'augmenter le plus possible le nombre d'individus de la population d'entrainement (l'interaction GxE est toutefois très limité chez les animaux). Néanmoins, lorsque l'interaction GxE influe fortement sur les phénotypes (faible héritabilité), le mélange des informations provoque la baisse de la précision de la prédiction.

Cette étude de l'influence de l'interaction GxE reste néanmoins superficielle. Il faudrait par exemple évaluer l'effet de cette interaction sur la précision de la prédiction en utilisant des individus du même génotype provenant de lieux et d'années différentes à la fois.

### 4.1.7 Analyse des effets attribués aux marqueurs : comparaison avec une étude de GWAS

La comparaison de l'approche en SG avec une approche GWAS permet de conforter l'idée que les effets attribués aux marqueurs par le modèle Bayes C est pertinent. En effet, pour les phénotypes étudiés, 9 marqueurs sur les 21 considérés comme significatifs en GWAS font partie du groupe de marqueurs correspondant à 10% des effets totaux attribués à l'ensemble des marqueurs en SG, ou sont en DL avec au moins un de ces marqueurs.

Néanmoins, les marqueurs significatifs en GWAS sont peu nombreux par rapport au nombre de marqueurs détectés en SG. La SG met en valeur des loci à faibles effets que la GWAS n'identifie pas. Cela démontre que les études GWAS sont très stringentes et ne prennent en compte que les

marqueurs très influents sur le phénotype étudié. En effet, la GWAS est limitée par des effets de seuils statistiques liés aux tests multiples (FDR : false discovery rate), utilisés pour contrôler la présence des faux négatifs causés par les tests multiples (Visscher et al., 2012).

La SG et la GWAS sont donc deux techniques qui s'accordent sur les effets à attribuer aux marqueurs quant à leur participation à la variation d'un phénotype. Néanmoins, la GWAS (qui étudie l'association des gènes avec des traits phénotypiques) permet de déterminer les marqueurs les plus influents dans le but de les caractériser ou de les utiliser dans un schéma de sélection (avec la sélection assistée par marqueurs par exemple). La SG propose une alternative en considérant tous les marqueurs et donc tous leurs effets (des plus faibles aux plus forts) dans le schéma de sélection en proposant une prédiction du caractère étudié, ce qui permet de sélectionner les parents de la génération suivante.

# 5 Conclusions et Perspectives

Bien que la SG soit une méthode qui n'en est qu'à ses débuts en ce qui concerne le monde du végétal (Jonas, 2013), les résultats obtenus dans le cadre de cette étude sont positifs et laissent penser que la SG chez la tomate pourrait être applicable à l'échelle commerciale, à l'instar de la SG animale. En effet, la précision de la prédiction est supérieure à 0.5 pour la majorité des phénotypes, ce qui revient à dire que les phénotypes prédits sont corrélés positivement aux phénotypes vrais. Cela permettrait de réduire le temps et le coût de la sélection en remplaçant certaines étapes de phénotypage par la SG, et de sélectionner des caractères induits par plusieurs locus.

Cependant, la composition des populations d'entrainement et de test doivent être élaborées de manière réfléchie, pour ensuite appliquer la SG à plus grande échelle. En effet, plus la population d'entrainement est grande et plus elle contient des individus proches génétiquement et ayant été phénotypés dans des environnements similaires par rapport à la population de validation, plus la précision de la prédiction est élevée. De plus, il est souhaitable que la population d'entrainement soit optimisée de manière à ce qu'elle contienne une grande diversité génétique (avec la méthode CDmean par exemple). Egalement, la précision de la prédiction va dépendre de la valeur de l'héritabilité du caractère étudié. Plus elle est élevée, meilleure sera la prédiction. Enfin, le nombre de marqueurs nécessaires dépend du DL de la population étudiée. Les nouvelles technologies de séquençage peuvent aujourd'hui répondre à un besoin d'augmentation du nombre de marqueurs néanmoins, une puce à ADN peut suffire pour certains caractères si de nombreux QTLs sont par exemple représentés par des marqueurs implantés sur cette même puce.

L'étude a été réalisée avec trois populations (GWAS, MAGIC et RIL) de la même génération. Il pourrait être intéressant de la confronter à des travaux portants sur des populations commerciales (variétés à plus gros fruits, avec moins de diversité génétique, hybrides F1 etc.) ainsi que sur d'autres phénotypes. En effet, l'objectif de cette étude étant de soumettre l'idée que la SG puisse fonctionner pour les entreprises travaillant sur la tomate, une étude comme celle-ci sur les variétés obtenues et commercialisées par des entreprises privées pourrait être pertinente. De plus, un autre aspect qui n'a pas été abordé ici serait intéressant à développer : une approche de sélection sur plusieurs caractères à la fois, avec des phénotypes corrélés entre eux en utilisant des index comme phénotype, par exemple. Enfin, l'évaluation de la transférabilité de ces travaux sur des espèces proches telles que les autres solanacées permettrait de faire évoluer cette méthode en ce qui concerne le domaine du végétal.

Egalement, le choix de la place de la SG dans le schéma de sélection est un point sur lequel il faut continuer l'investigation en se posant les questions suivantes :

- Combien et quelles étapes de phénotypage faut-il remplacer par la SG et à quel moment dans le schéma de sélection ?
- Pour une même génération, faut-il remplacer totalement le phénotypage par la SG ou sélectionner seulement une partie des individus avec cette technique ?

Pour finir, une étude économique valorisant les aspects de la sélection génomique pourrait mettre en valeur ces résultats (origine des financements, impact sur la structuration de la filière, impact sur le consommateur etc.). De plus, une étude d'impact sur l'érosion et la conservation de la diversité génétique est indispensable.

# 6 L'organisation de l'étude

Au cours de cette période de stage, le premier mois a été consacré à l'étude bibliographique du sujet et au formatage des données de génotypage et de phénotypage. En effet, les données traitées proviennent de plusieurs études antérieures publiées, il m'a donc fallu me les approprier pour pouvoir les traiter sous le logiciel R afin de les rendre comparables.

En parallèle, j'ai réfléchi à la méthode que j'allais employer pour répondre à la problématique. Les modèles de prédiction demandant du temps pour obtenir des résultats, j'ai choisi d'une part les paramètres qui me paraissaient les plus pertinents à étudier (modèle statistique, composition de la population d'entrainement, densité des marqueurs moléculaires, héritabilité des caractères) et la façon dont j'allais tester ces paramètres. Pour cela, je me suis appuyée sur les études bibliographiques dans le but de pouvoir comparer mes travaux et interagi avec David Cros, ingenieur au CIRAD, qui a travaillé sur la sélection génomique chez le palmier à huile.

Par la suite, j'ai développé des scripts sur le logiciel R permettant de faire fonctionner les modèles statistiques de prédiction avec toutes les modalités que j'avais choisies, et ce, en les optimisant de manière à ce qu'ils soient exécutés le plus rapidement possible par le serveur de l'unité GAFL de l'INRA.

Dès l'obtention des premiers résultats de mes modèles, j'ai exploré mes données visuellement (via des diagrammes en boîte en particulier) et effectué des tests statistiques afin de comparer les différents modèles et paramètres utilisés. Pour finir, j'ai comparé mes résultats aux études précédentes pour évaluer leur pertinence.

# 7 Bibliographie


Beaulieu, J., Doerksen, T., Clément, S., MacKay, J., & Bousquet, J. (2014). Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. *Heredity*, *113*(4), 343-352.

Blanca, J., Montero-Pau, J., Sauvage, C., Bauchet, G., Illa, E., Díez, M. J., ... & Cañizares, J. (2015). Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. *BMC genomics*, *16*(1), 257.

Causse, M., Saliba-Colombani, V., Lecomte, L., Duffe, 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. *Journal of experimental botany*, *53*(377), 2089-2098.

Causse, M., Duffe, P., Gomez, M. C., Buret, M., Damidaux, R., Zamir, D., ... & Rothan, C. (2004). A genetic map of candidate genes and QTLs involved in tomato fruit size and composition. *Journal of Experimental Botany*, *55*(403), 1671-1685.

Cros, D. (2014a). *Etude des facteurs contrôlant l'efficacité de la sélection génomique chez le palmier à huile (Elaeis guineensis Jacq.)* (Doctoral dissertation, Montpellier SupAgro).

Cros, D., Denis, M., Sánchez, L., Cochard, B., Flori, A., Durand-Gasselin, T., ... & Bouvet, J. M. (2014b). Genomic selection prediction accuracy in a perennial crop: case study of oil palm (*Elaeis guineensis* Jacq.). *Theoretical and Applied Genetics*, *128*(3), 397-410.

Daetwyler, H. D., Calus, M. P., Pong-Wong, R., de los Campos, G., & Hickey, J. M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics*, *193*(2), 347-365.

Daetwyler, H., Capitan, A., Pausch, H., Stothard, P., Van Binsbergen, R., Brandum, R., ... & Eggen, A. (2014). The 1000 bull genome project.

De Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., ... & Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, *182*(1), 375-385.

Desta, Z. A., & Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends in plant science*, *19*(9), 592-601.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome*, *4*(3), 250-255.

Falconer D. et Mackay T. (1996) Introduction to quantitative genetics. Longman, Harlow,Essex, UK, 464p.

Fernandez-Pozo, N., Menda, N., Edwards, J. D., Saha, S., Tecle, I. Y., Strickler, S. R., ... & Mueller, L. A. (2015). The Sol Genomics Network (SGN)—from genotype to phenotype to breeding. *Nucleic acids research*, *43*(D1), D1036-D1041.

Fodor, A., Segura, V., Denis, M., Neuenschwander, S., Fournier-Level, A., Chatelet, P., ... & Le Cunff, L. (2014). Genome-Wide prediction methods in highly diverse and heterozygous species: proof-of-concept through simulation in grapevine.

Gallais, A. (2011). *Méthodes de création de variétés en amélioration des plantes*. Editions Quae, 280 pages

Gianola, D., Fernando, R. L., & Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, *173*(3), 1761-1776.

Gianola, D., Wu, X. L., Manfredi, E., & Simianer, H. (2010). A non-parametric mixture model for genome-enabled prediction of genetic value for a quantitative trait. *Genetica*, *138*(9-10), 959-977.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 1440-1450.

Grattapaglia, D. (2014). Breeding forest trees by genomic selection: Current progress and the way forward. In *Genomics of Plant Genetic Resources* (pp. 651-682). Springer Netherlands.

De los Campos, G., & Pérez-Rodríguez, P. (2013). BGLR: Bayesian Generalized Linear Regression. *R package version*, *1*(3).

Habier, D., Fernando, R. L., & Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, *177*(4), 2389-2397.

Habier, D., Fernando, R. L., & Garrick, D. J. (2013). Genomic BLUP decoded: a look into the black box of genomic prediction. *Genetics*, *194*(3), 597-607.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science*, *92*(2), 433-443.

Heffner, E. L., Jannink, J. L., Iwata, H., Souza, E., & Sorrells, M. E. (2011a). Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Science*, *51*(6), 2597-2606.

Heffner, E. L., Jannink, J. L., & Sorrells, M. E. (2011b). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *The Plant Genome*, *4*(1), 65-75.

Howard, R., Carriquiry, A. L., & Beavis, W. D. (2014). Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3: Genes| Genomes| Genetics*, *4*(6), 1027-1046.

Huang, Y. H., Li, N., Burt, D. W., & Wu, F. (2008). Genomic research and applications in the duck (*Anas platyrhynchos*). *World's Poultry Science Journal*, *64*(03), 329-341.

Jonas, E., & de Koning, D. J. (2013). Does genomic selection have a future in plant breeding? *Trends in biotechnology*, *31*(9), 497-504.

Jomphe, V. (2006). Comparaison de la puissance de tests de déséquilibre de liaison dans les études génétiques.

Kizilkaya, K., Fernando, R. L., & Garrick, D. J. (2010). Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *Journal of animal science*, *88*(2), 544-551.

Kumar, S., Chagné, D., Bink, M. C. A. M., Volz, R. K., Whitworth, C., & Carlisle, C. (2012). Genomic selection for fruit quality traits in apple (Malus × domestica Borkh.). *PloS one*, *7*(5), e36674.

Lane, D. (2010). Tukey's honestly significant difference (HSD). In N. Salkind (Ed.), Encyclopedia of research design. (pp. 1566-1571). Thousand Oaks, CA: SAGE Publications, Inc. doi: http://dx.doi.org/10.4135/9781412961288.n478

Legarra, A., Robert-Granié, C., Manfredi, E., & Elsen, J. M. (2008). Performance of genomic selection in mice. *Genetics*, *180*(1), 611-618.

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., ... & Huang, S. (2014). Genomic analyses provide insights into the history of tomato breeding. *Nature genetics*.

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., ... & Jannink, J. L. (2011). 2 Genomic Selection in Plant Breeding: Knowledge and Prospects. *Advances in agronomy*, *110*, 77.

Massman, J. M., Jung, H. J. G., & Bernardo, R. (2013). Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. *Crop Science*, *53*(1), 58-66.

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

Meuwissen, T. H. (2009). Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. *Genetics Selection and Evolution*, *41*(1), 35.

Nakaya, A., & Isobe, S. N. (2012). Will genomic selection be a practical method for plant breeding? *Annals of botany*, mcs109.

Pascual, L., Desplat, N., Huang, B. E., Desgroux, A., Bruguier, L., Bouchet, J. P., ... & Causse, M. (2015). Potential of a tomato MAGIC population to decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. *Plant biotechnology journal*, *13*(4), 565-577.

Pelgas, B., Bousquet, J., Meirmans, P. G., Ritland, K., & Isabel, N. (2011). QTL mapping in white spruce: gene maps and genomic regions underlying adaptive traits across pedigrees, years and environments. *BMC genomics*, *12*(1), 145.

Pérez, P., de los Campos, G., Crossa, J., & Gianola, D. (2010). Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. *The Plant Genome*, *3*(2), 106-116.

Pérez P. et de los Campos G. (2013) BGLR: A Statistical package for whole genome regression and prediction. R package version 1.0.2.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., ... & Moreau, L. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics*, *192*(2), 715-728.

Saliba-Colombani, V., Causse, M., Langlois, D., Philouze, J., & Buret, M. (2001). Genetic analysis of organoleptic quality in fresh market tomato. 1. Mapping QTLs for physical and chemical traits. *Theoretical and Applied Genetics*, *102*(2-3), 259-272.

Sauvage, C., Segura, V., Bauchet, G., Stevens, R., Do, P. T., Nikoloski, Z., ... & Causse, M. (2014). Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant physiology*, *165*(3), 1120-1132.

Sim, S. C., Durstewitz, G., Plieske, J., Wieseke, R., Ganal, M. W., Van Deynze, A., ... & Francis, D. M. (2012). Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. *PLoS One*, *7*(7), e40563.

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., ... & McCouch, S. R. (2015). Genomic Selection and Association Mapping in rice (*Oryza sativa*): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genet*, *11*(2), e1004982.

Student (1908). Biometrika, Volume 6, Issue 1 (Mar, 1908), 1-25.

Technow, F., Bürger, A., & Melchinger, A. E. (2013). Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. *G3: Genes| Genomes| Genetics*, *3*(2), 197-203.

Toosi, A., Fernando, R. L., Dekkers, J. C. M., & Quaas, R. L. (2010). Genomic selection in admixed and crossbred populations. *Journal of Animal Science*, *88*(1), 32.

Tomato Genome Consortium. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, *485*(7400), 635-641.

Van der Werf, J. H. J. (2009). Potential benefit of genomic selection in sheep. In *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* (Vol. 18, pp. 38-41).

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of dairy science*, *91*(11), 4414-4423.

VanRaden, P. M., & Sullivan, P. G. (2010). International genomic evaluation methods for dairy cattle. *Genetic, Selection and Evolution*, *42*(7).

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics*, *90*(1), 7-24.

Wang, Y., Mette, M. F., Miedaner, T., Gottwald, M., Wilde, P., Reif, J. C., & Zhao, Y. (2014). The accuracy of prediction of genomic selection in elite hybrid rye populations surpasses the accuracy of marker-assisted selection and is equally augmented by multiple field evaluation locations and test years. *BMC genomics*, *15*(1), 556.

Xu, J., Ranc, N., Muños, S., Rolland, S., Bouchet, J. P., Desplat, N., ... & Causse, M. (2013). Phenotypic diversity and association mapping for fruit quality traits in cultivated tomato and related species. *Theoretical and applied genetics*, *126*(3), 567-581.

Zhao, Y., Gowda, M., Liu, W., Würschum, T., Maurer, H. P., Longin, F. H., ... & Reif, J. C. (2012). Accuracy of genomic selection in European maize elite breeding populations. *Theoretical and Applied Genetics*, *124*(4), 769-776.

# 8 Sitographie

« France Agrimer ». 2014. Consulté le avril 23. http://www.franceagrimer.fr/content/download/33795/306715/file/SYN-FEL-2014%20Tomate-Transf%202013.pdf.

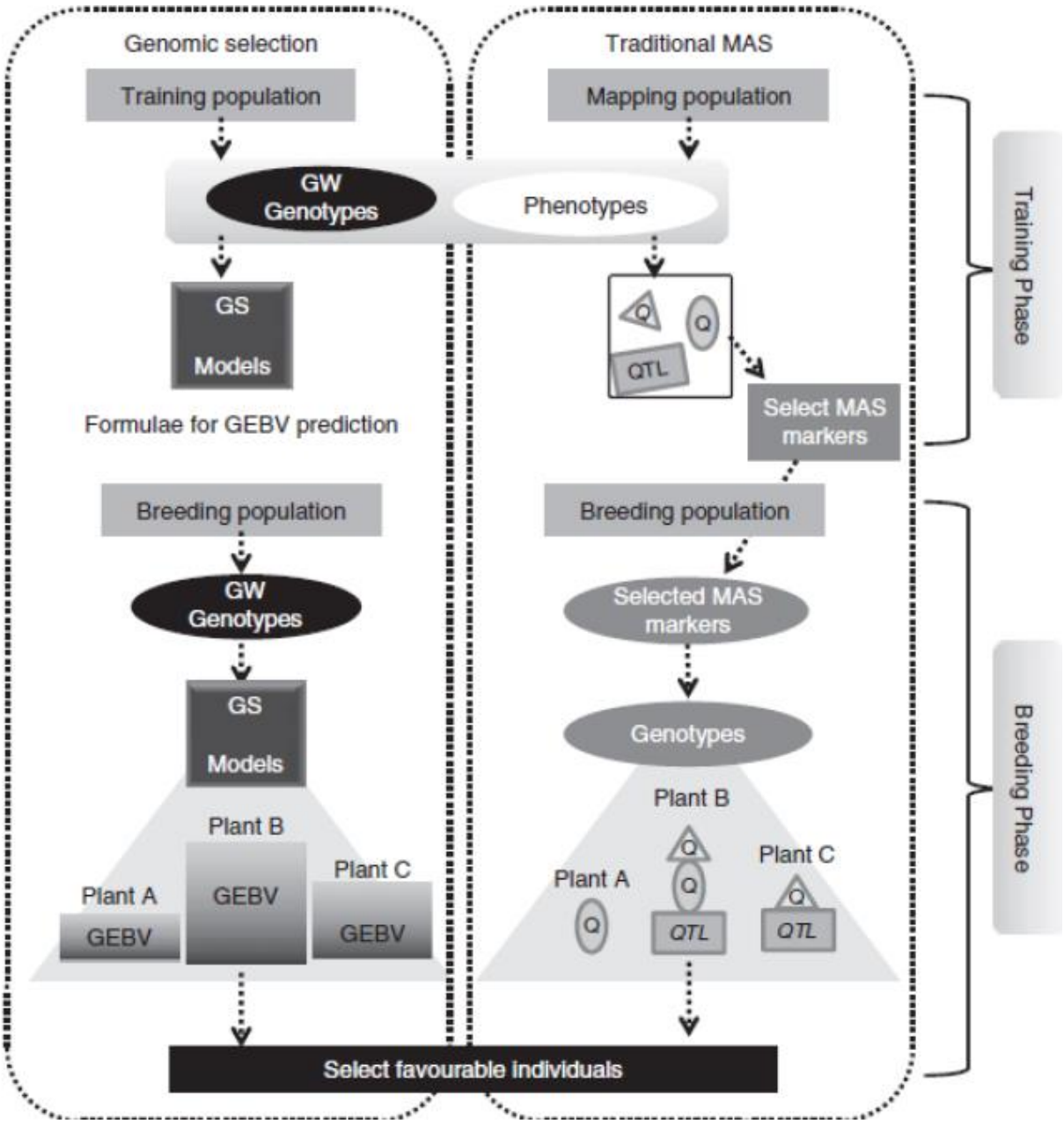
« INRA - sélection génomique bovins » (2015) Consulté le mars 25. http://www.inra.fr/Entreprises-Monde agricole/Resultats-innovation-transfert/Toutes-les-actualites/selection-genomique-bovins.

« SEMEX France - Sélection Génomique » (2015) Consulté le mars 25. http://www.semex.fr/semex-france/selection_genomique.php.

« SolCAP Solanaceae Coordinated Agricultural Project ». 2015. Consulté le avril 20. http://solcap.msu.edu/index.shtml.

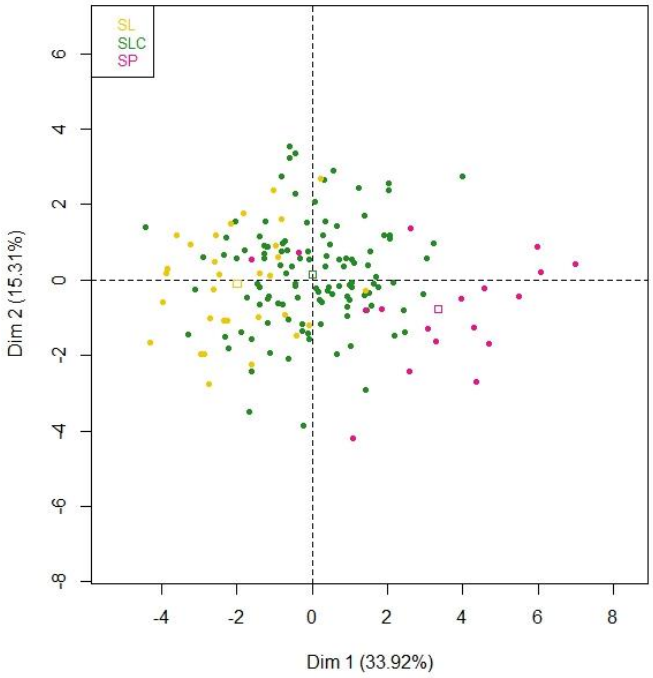
# 9 Annexes

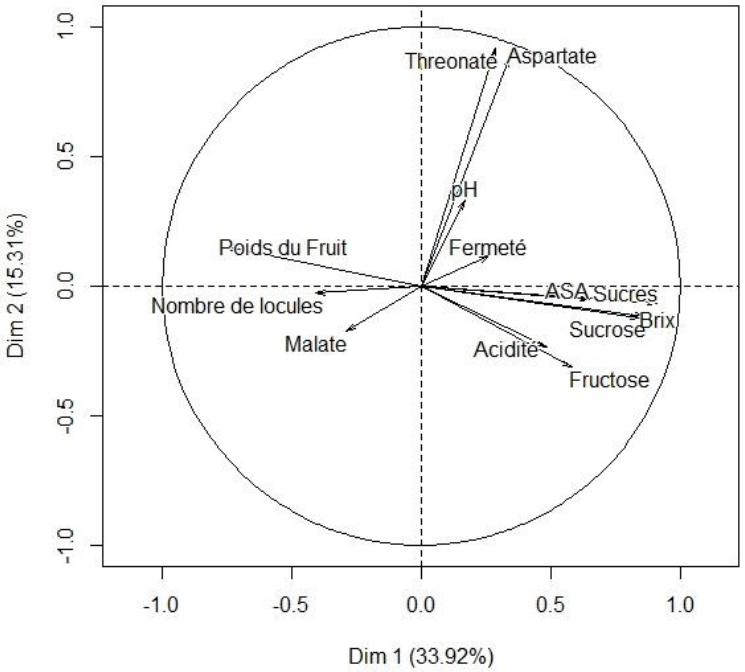
**Annexe 1 : Schéma de sélection génomique à gauche (SG) et de sélection assistée par marqueurs à droite (SAM) (Nakaya et al, 2012). Les deux méthodes sont composées d'une partie « training » (entrainement) et d'une partie « breeding » (amélioration). Dans la SAM classique, la partie « training » comprend l'identification des associations (QTLs) et dans la SG elle comprend la définition d'un modèle statistique permettant de prédire les GEBVs (genomic estimated breeding values). La partie « breeding » représente la sélection des individus sur la base de leurs génotypes aux marqueurs identifiés pour la SAM classique ou selon leurs GEBVs pour la SG.**



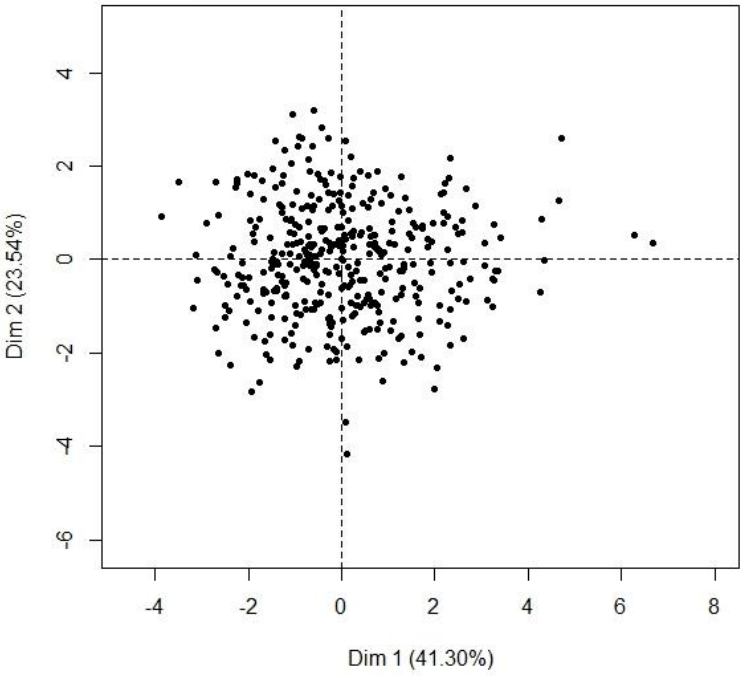
**Annexe 2 : Analyse en composantes Principales des jeux de données GWAS (A), MAGIC (B), RIL (C) et GWAS (2007 et 2008) (D)**

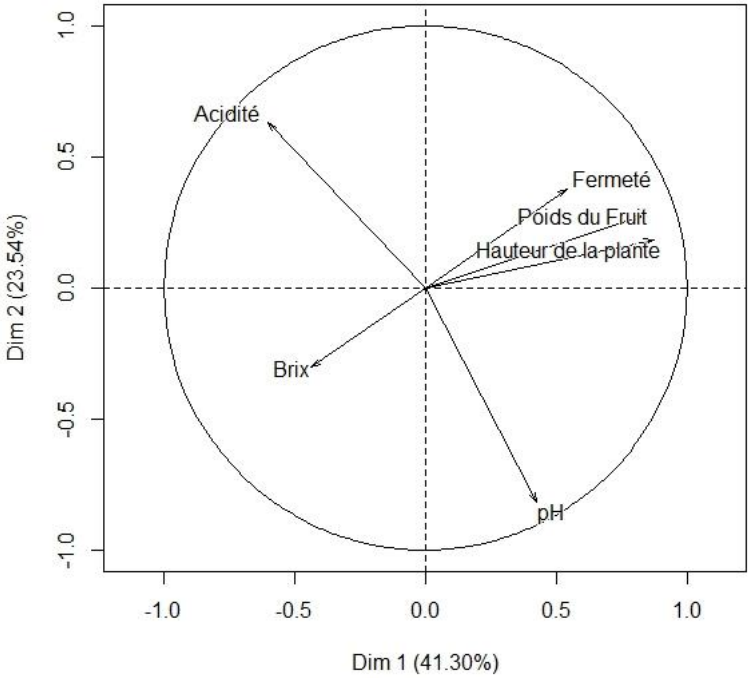
**A**



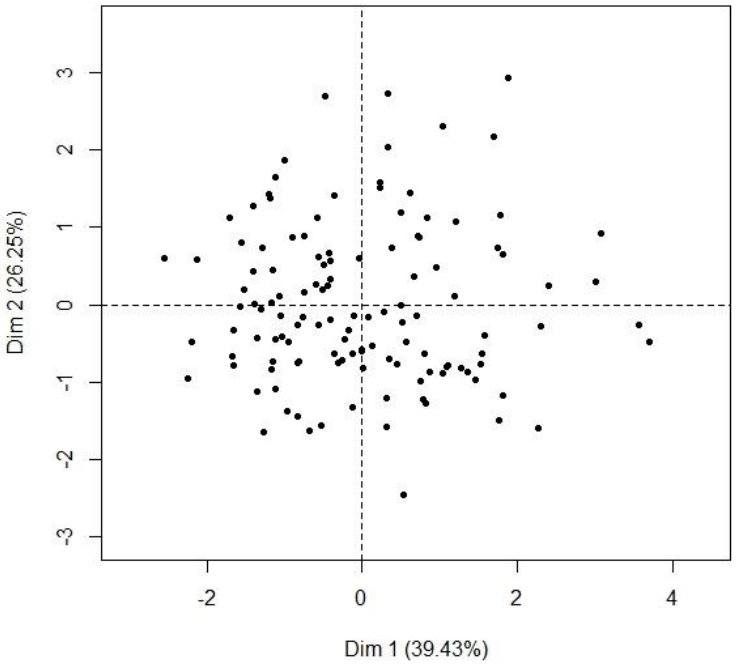



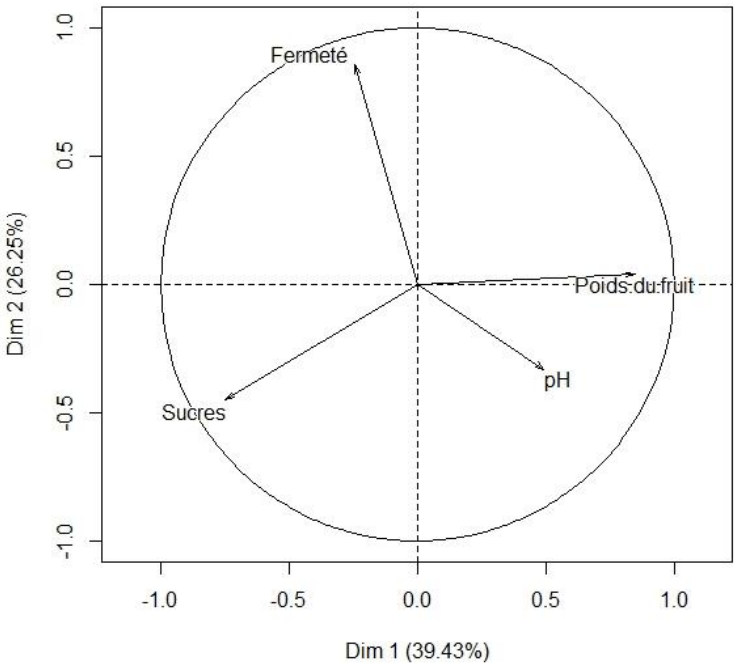

**B**



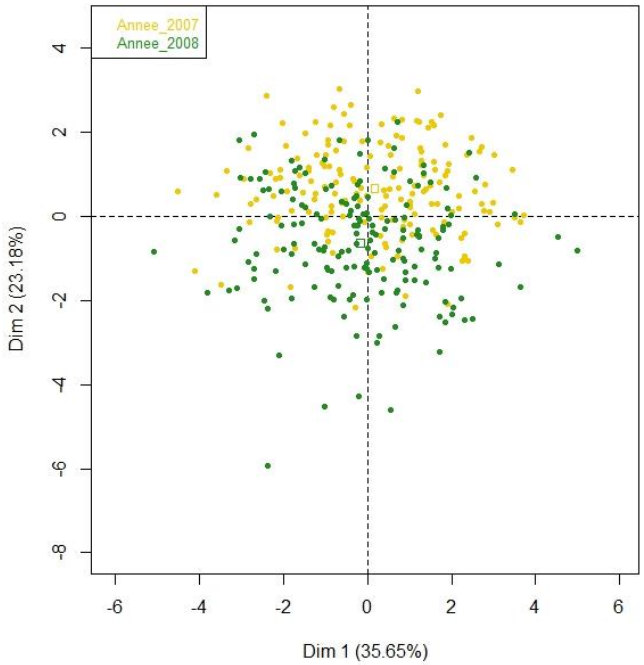



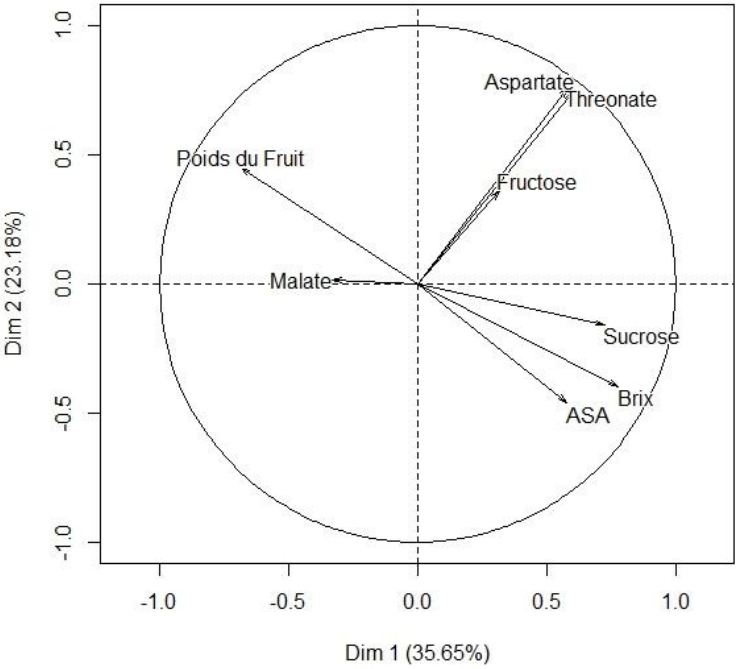
**C**


Dim 2 (26.25%)
Dim 1 (39.43%)

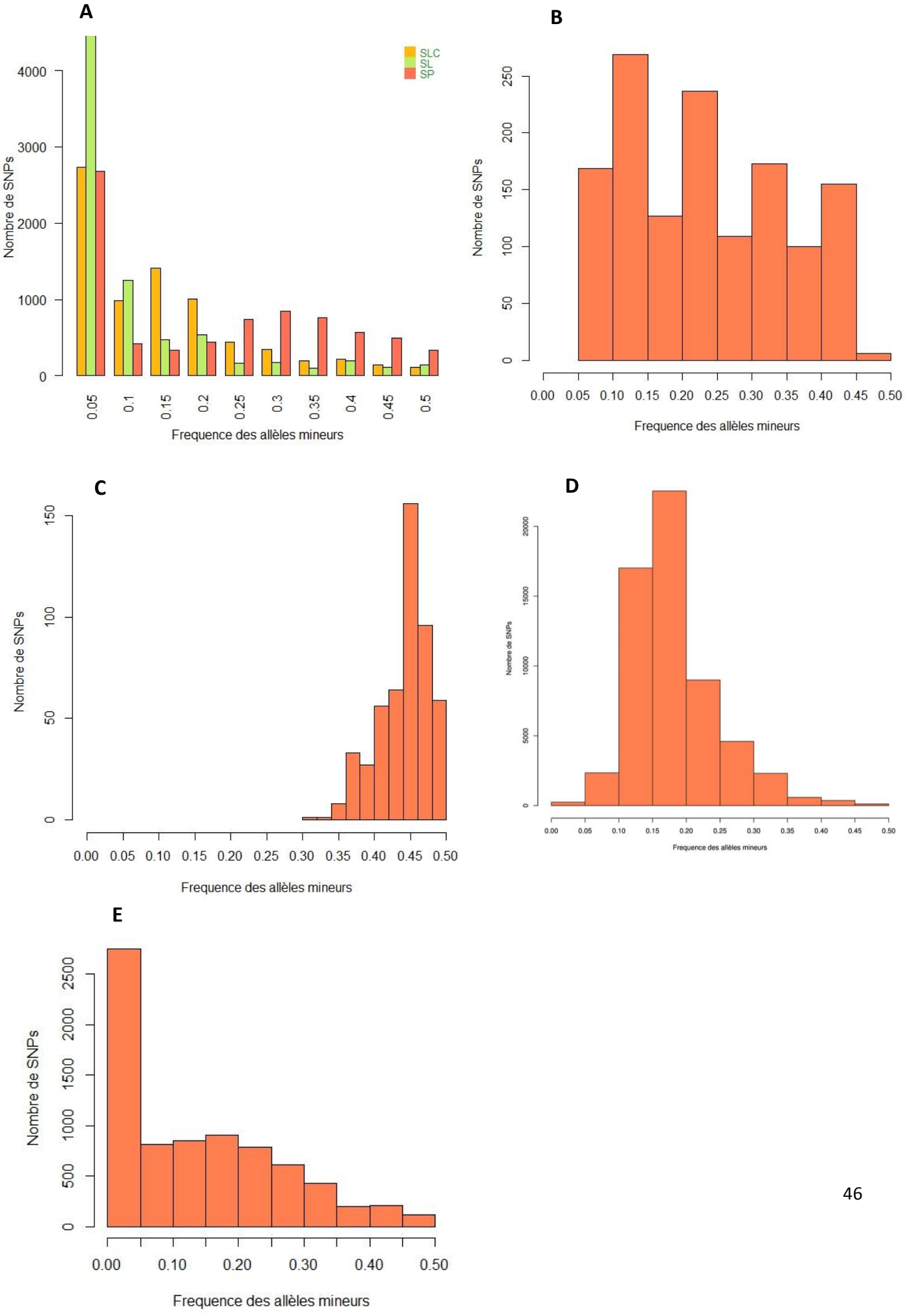

Fermeté
Poids du fruit
pH
Sucres
Dim 2 (26.25%)
Dim 1 (39.43%)

**D**


Annee_2007
Annee_2008
Dim 2 (23.18%)
Dim 1 (35.65%)

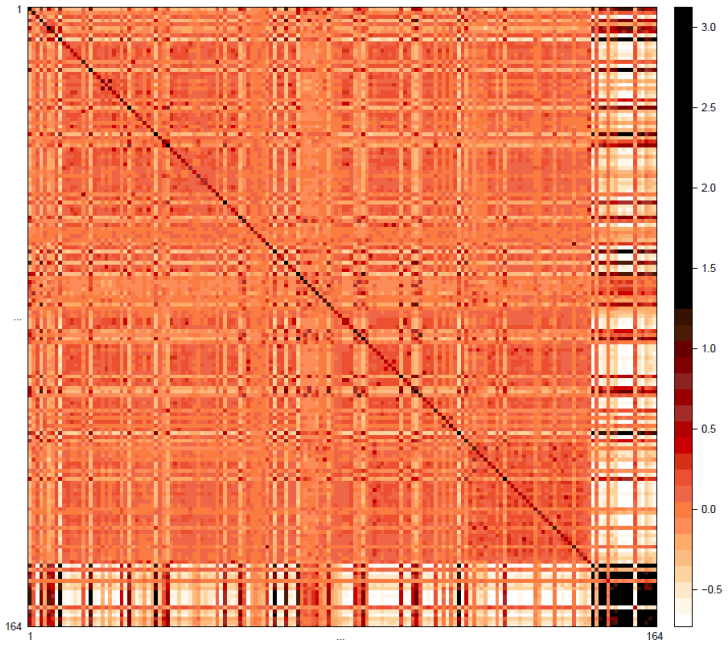

Aspartate
Threonate
Poids du Fruit
Fructose
Malate
Sucrose
Brix
ASA
Dim 2 (23.18%)
Dim 1 (35.65%)

**Annexe 3 : Fréquence des allèles mineurs des jeux de données GWAS (A), MAGIC (B), RIL (C), GBS (D) et GWAS 2 (E)**

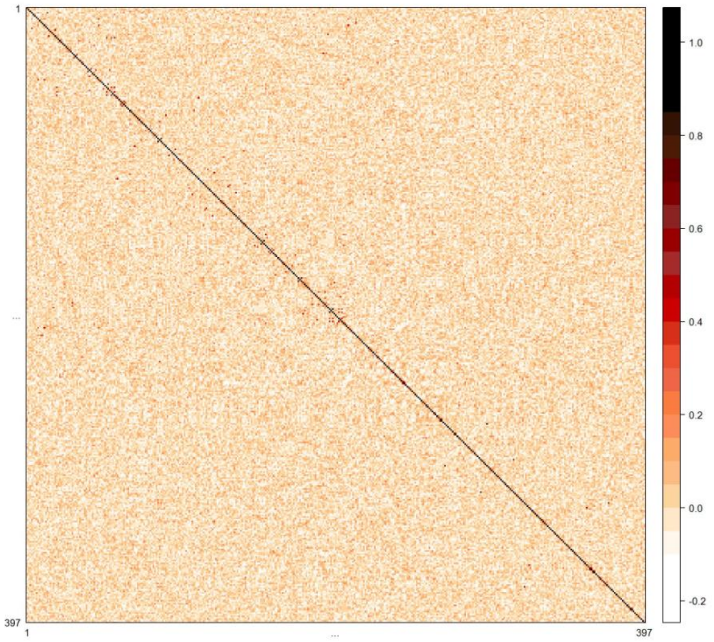


**Annexe 4 : Matrices d'apparentement GWAS (A), MAGIC (B) RIL (C) et GBS (D) basées sur l'apparentement entre les marqueurs. Le lien de parenté entre les individus est calculé selon les formules dans Habier et al. (2007) ZZ'/(2∑ pi(1-pi)) où Z=W-P, W est la matrice des marqueurs, P contient les fréquences alléliques multipliées par 2, pi est la fréquence de l'allèle du marqueur i, et la somme est faite sur tous les loci.**

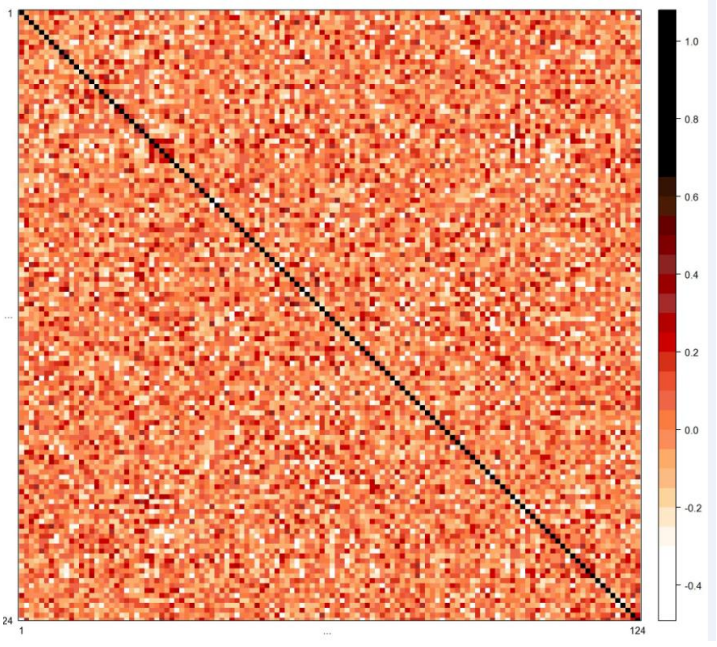
**A**



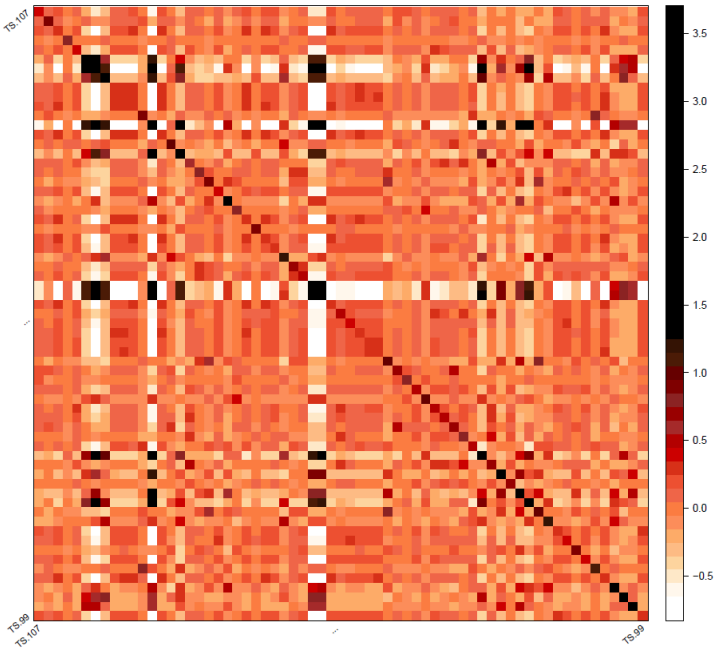

**B**




**C**



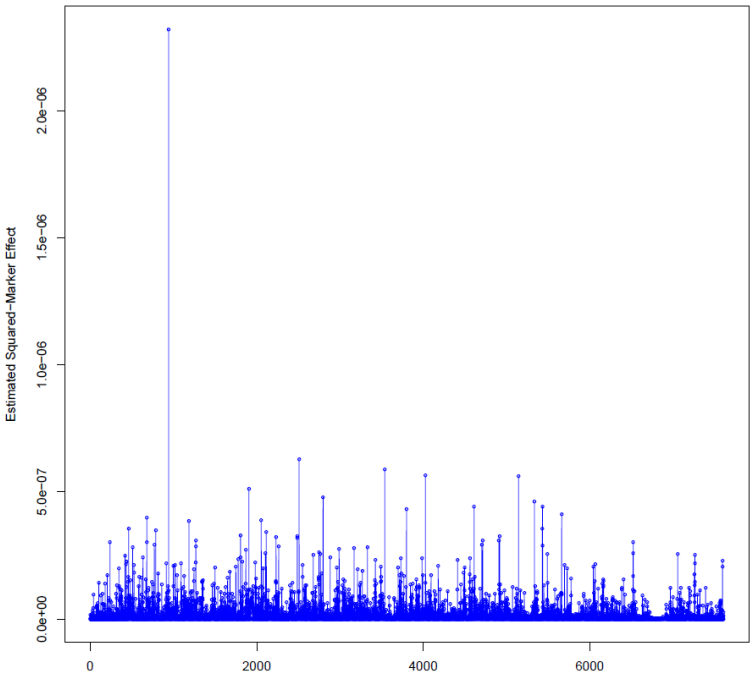

**D**



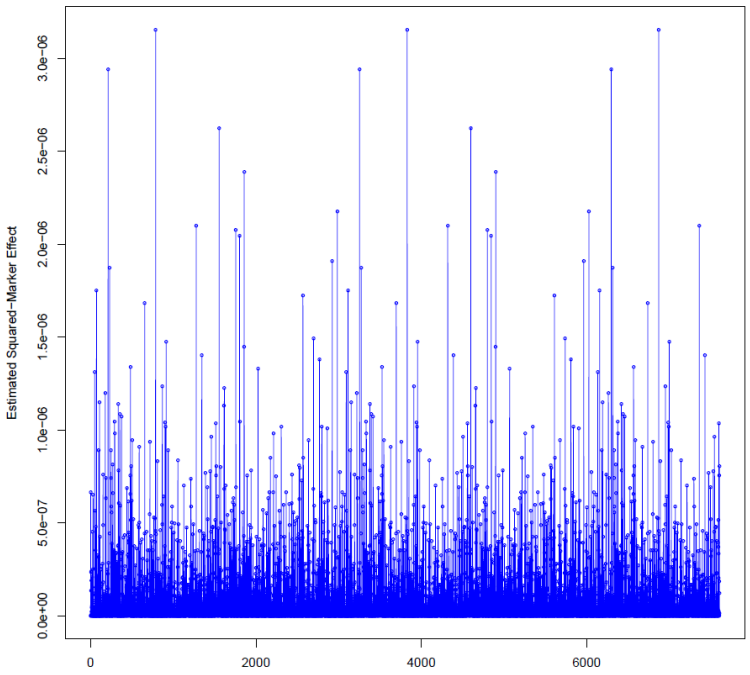
**Annexe 5 : Effets attribués aux marqueurs par le modèle Bayes C pour le phénotype « pH »:**

A. **Jeu de données GWAS avec 164 individus, méthode CDmean, PE = 75%, 100% des marqueurs sont utilisés**
B. **Jeu de données GWAS avec 44 individus, méthode CDmean, PE = 75%, 100% des marqueurs sont utilisés**
C. **Jeu de données GWAS avec 164 individus, PE = 75%, 40% des marqueurs sont utilisés**
D. **Jeu de données MAGIC avec 397 individus, méthode CDmean, PE = 75%, 100% des marqueurs sont utilisés**
E. **Jeu de données MAGIC avec 397 individus, PE = 75%, 40% des marqueurs sont utilisés**
F. **Jeu de données RIL avec 231 individus, méthode CDmean, PE = 75%, 100% des marqueurs sont utilisés**
G. **Jeu de données RIL avec 231 individus, PE = 75%, 40% des marqueurs sont utilisés**

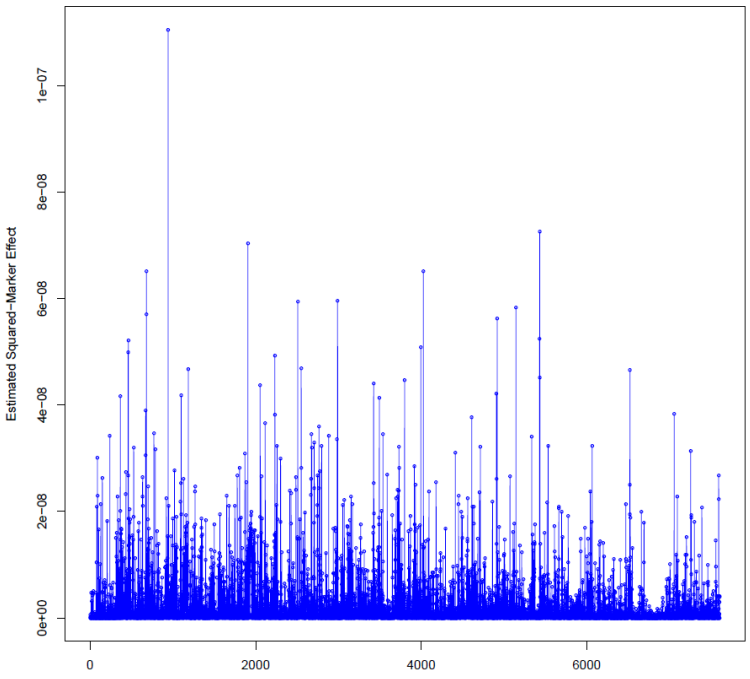
**A**



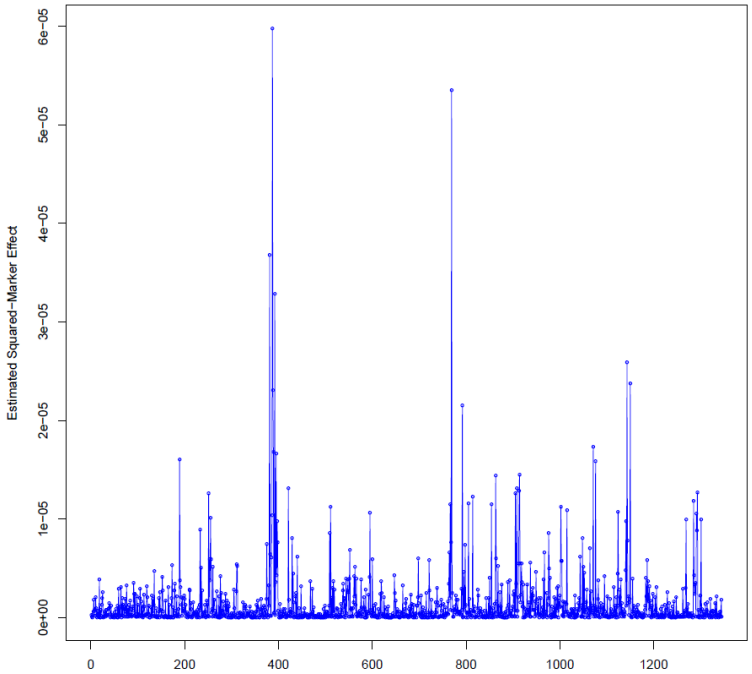

**B**



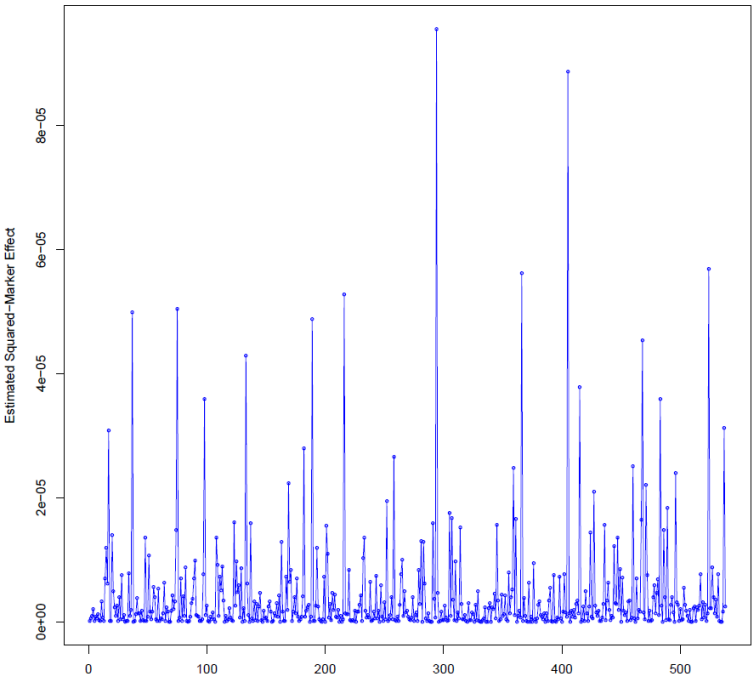

**C**



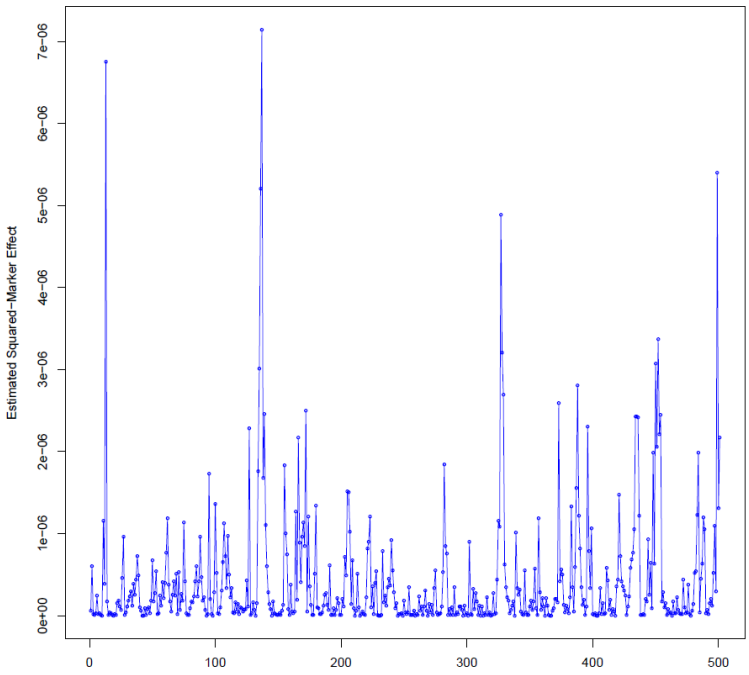
D


Estimated Squared−Marker Effect
0e+00
1e−05
2e−05
3e−05
4e−05
5e−05
6e−05
0
200
400
600
800
1000
1200

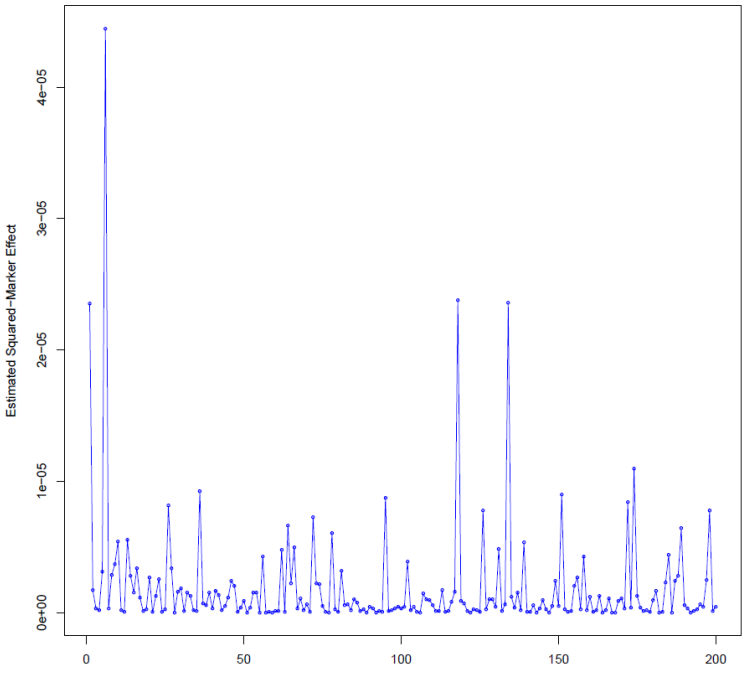
E


Estimated Squared−Marker Effect
0e+00
2e−05
4e−05
6e−05
8e−05
0
100
200
300
400
500

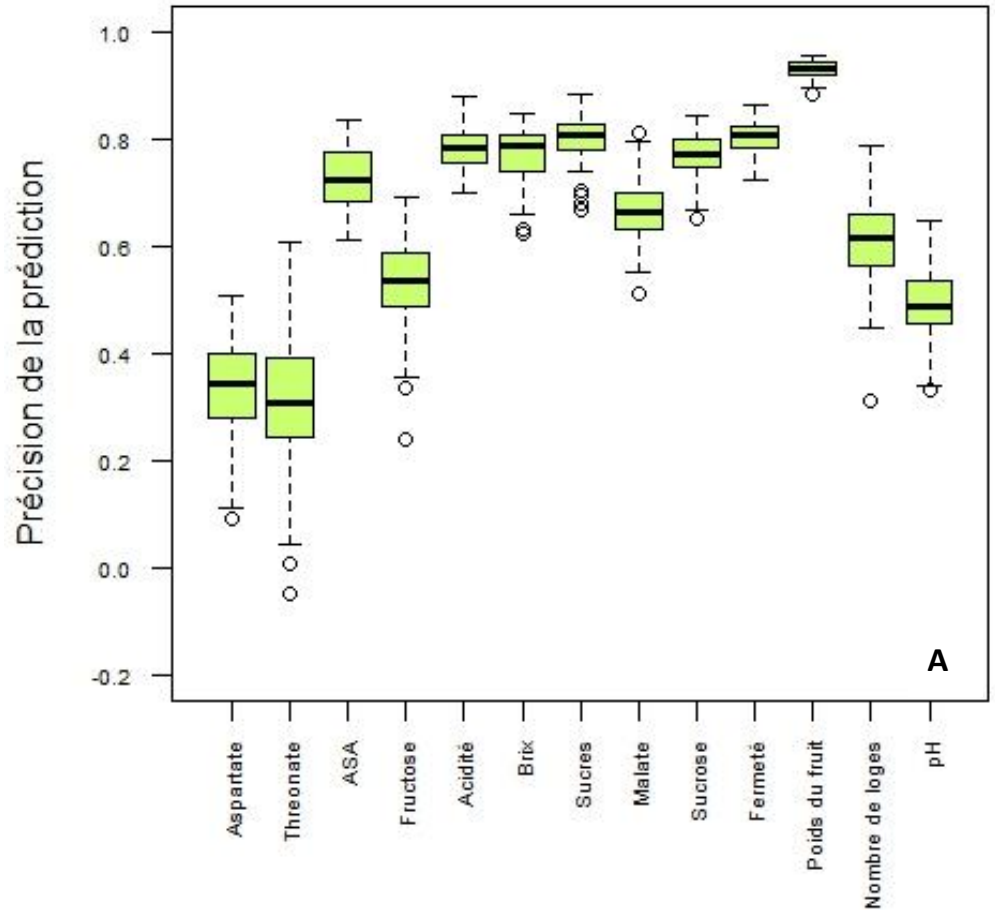
F

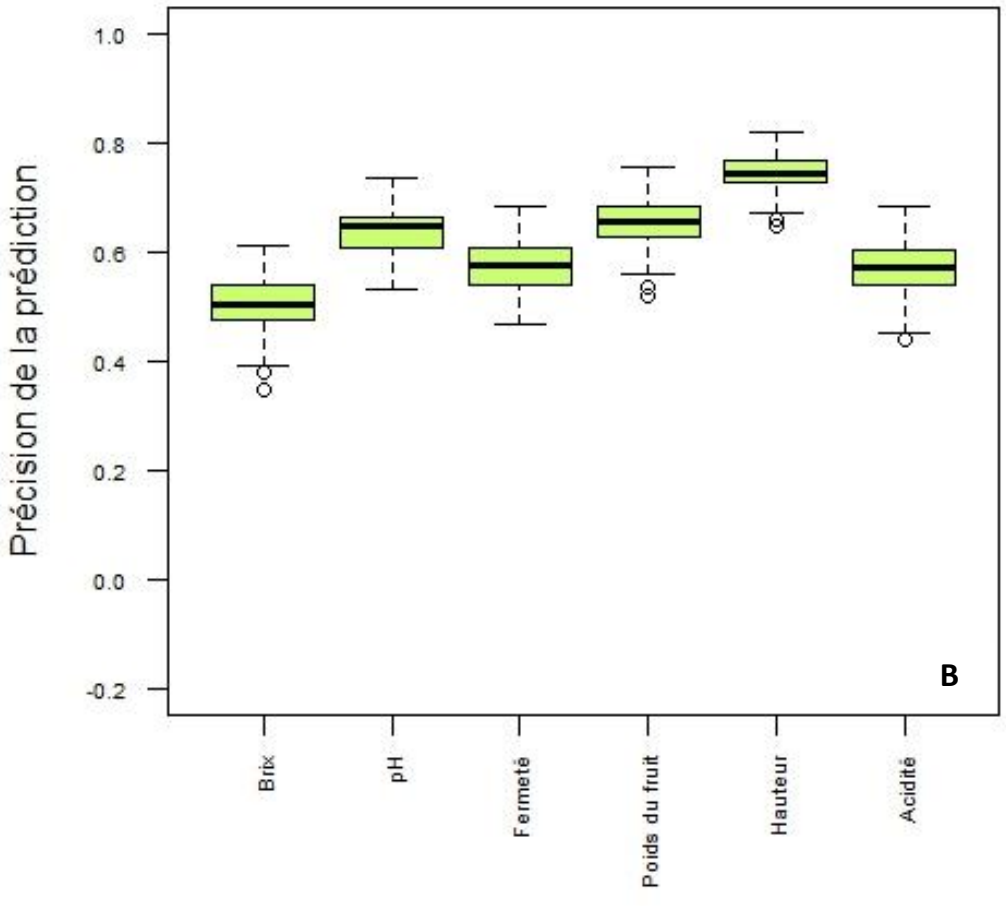

Estimated Squared−Marker Effect
0e+00
1e−06
2e−06
3e−06
4e−06
5e−06
6e−06
7e−06
0
100
200
300
400
500

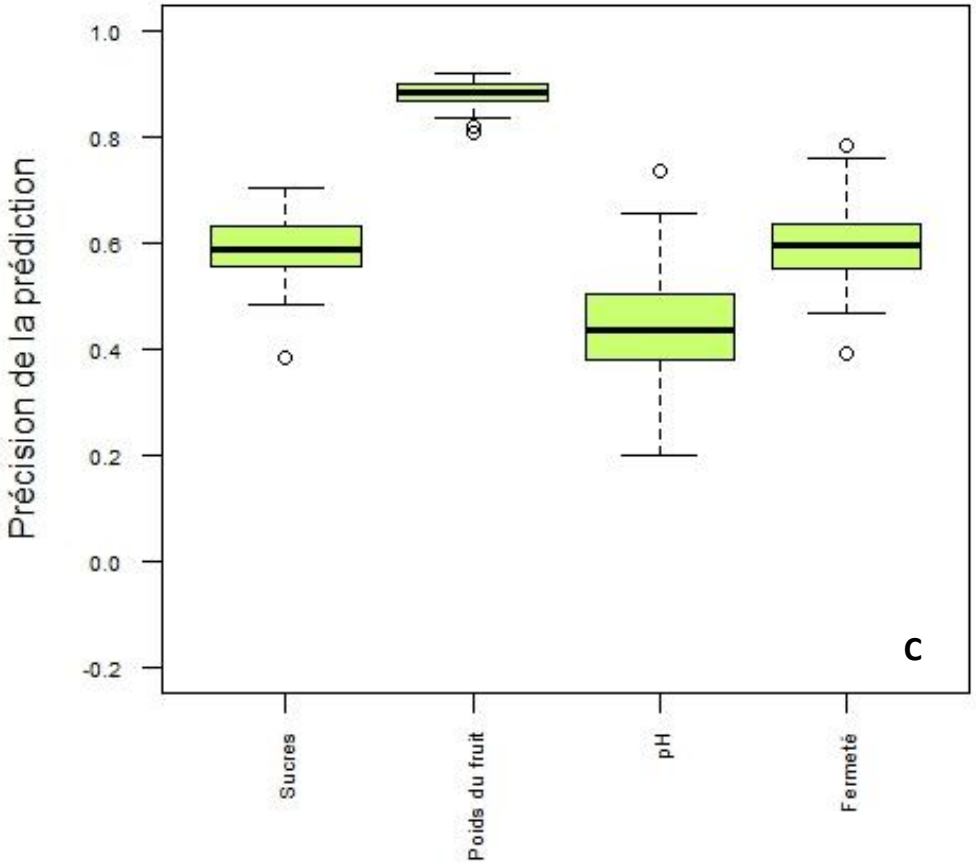
G


Estimated Squared−Marker Effect
0e+00
1e−05
2e−05
3e−05
4e−05
0
50
100
150
200

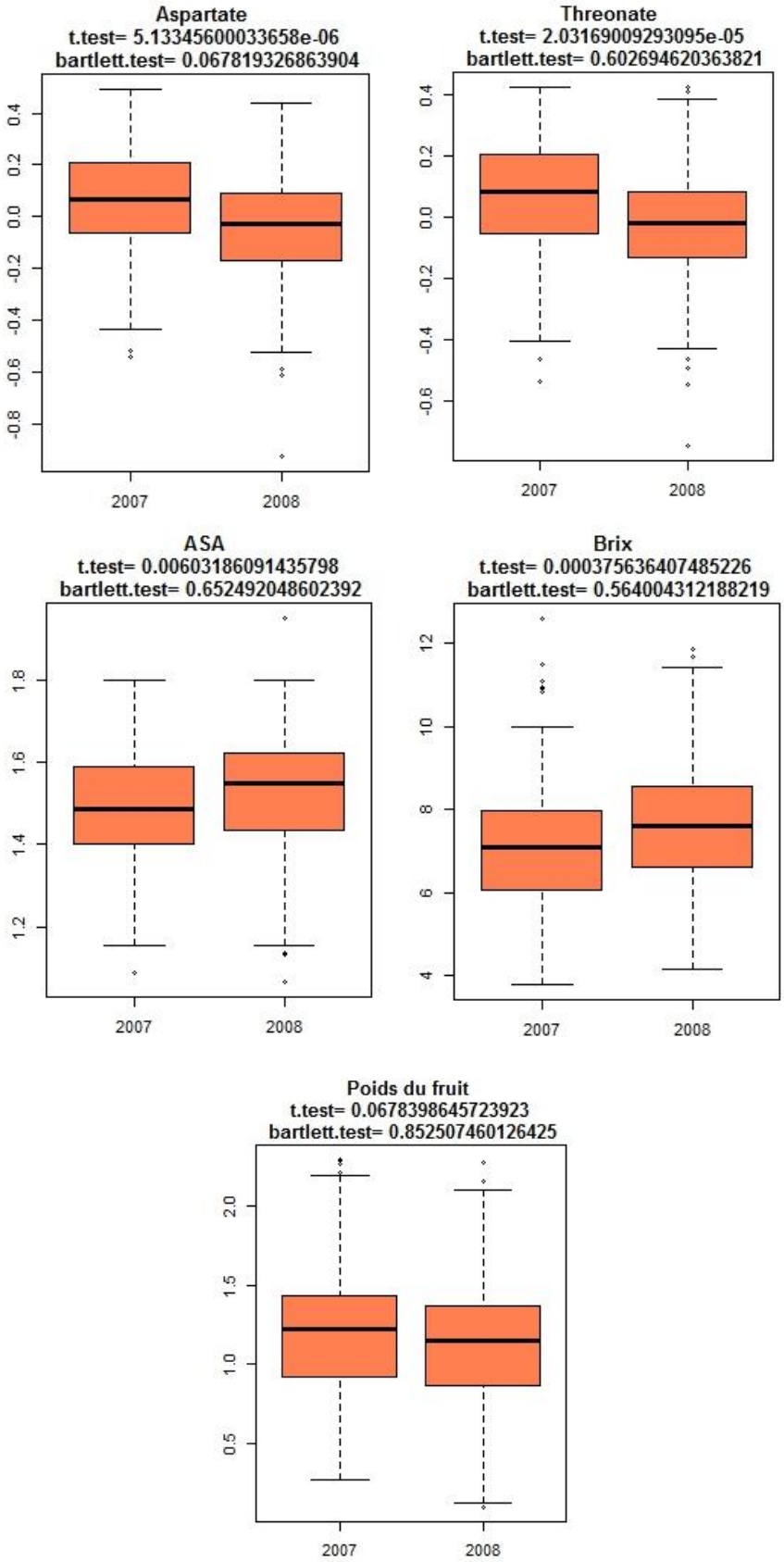
**Annexe 6 : Diagrammes en boîte de la précision de la prédiction des caractères phénotypiques en fonction des phénotypes avec l'utilisation du modèle Bayes C pour différents jeux de données (GWAS (A), MAGIC (B) et RIL (C)) (75% des individus font partie la population d'entrainement 25% de la population de validation).**




Précision de la prédiction
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
Sucres
Poids du fruit
pH
Fermeté
C

**Annexe 7 : Comparaison des phénotypes Aspartate, Threonate, ASA, Brix, Poids du fruit et Malate issus du jeu de données GWAS de 2007 et 2008.**



# Résumé

La sélection pour améliorer les caractères d'intérêts agronomiques est réalisée à l'aide de la SAM ou de la GWA qui capturent difficilement les loci à effets mineurs, et grâce à des étapes de phénotypage souvent longues et couteuses. La sélection génomique (SG) est une méthode de prédiction de la performance des caractères via des modèles statistiques pour sélectionner des plantes élites dans un schéma de sélection, en estimant l'effet de tous les marqueurs qui sous-tendent l'architecture génétique des caractères étudiés.

Nous avons évalué l'utilisation de la SG chez la tomate en estimant la précision de la prédiction pour 25 caractères liés à la qualité du fruit pour trois types de populations différentes (GWA, MAGIC et RIL). Une démarche de validation croisée a été menée en estimant le poids de paramètres (taille des populations d'entrainement (PE) et de validation (PV), interactions GxE, modèles statistiques prédictifs, densité de marquage, héritabilité des caractères, type de population) sur la précision de la prédiction.

Nos résultats démontrent que l'approche SG semble puissante pour prédire les valeurs phénotypiques (avec des précisions allant de 0.32 à 0.93 lorsque la PE est optimisée, par exemple). Nous observons que (1) plus la PE est grande et plus elle contient des individus proches génétiquement et ayant été phénotypés dans des environnements similaires par rapport à la PV, plus la précision de la prédiction est élevée (2) la précision de la prédiction est plus élevée avec l'optimisation de la PE avec une grande diversité génétique (+ 9% en moyenne) (3) plus l'héritabilité du caractère est élevée, meilleure est la prédiction (4) les différents modèles statistiques conduisent à des prédictions très similaires (5) la densité de marqueurs optimale dépend du DL de la population.

Pour conclure, bien que la SG soit une pratique qui n'en est qu'à ses débuts en ce qui concerne le monde du végétal l'application de la SG chez la tomate semble très prometteuse et pourrait être utilisée dans les programmes de sélection.



**Mots clés :** Sélection génomique, *Solanum lycopersicum,* validation croisée, interaction G×E, héritabilité



# Abstract

Selection to improve traits of economical interest in crop was achieved through MAS or GWA which capture with difficulty loci of minor effect, and through steps phenotyping often long and costly. Genomic selection (GS) is a new tool for selecting elite plants in a breeding program by predicting the performance of traits of interest in application of statistical model; GS has the potential to catch the effect of all markers underlying the genetic architecture of the studied traits.

We aimed at evaluating the use of GS into tomato to evaluate the prediction accuracy for 25 traits related to the fruit quality for three different type of population (GWA, MAGIC and RIL panel). We conducted a cross validation approach while estimating the relative weight of parameters (training and testing populations sizes, GxE interactions, predictive statistical models, markers density, heritability of the trait, type of population) onto the prediction accuracy.

Our results demonstrated that the GS approach seems powerful at predicting phenotypes values with accuracies ranging from 0.32 to 0.93 in the panel when training set is optimizing, for example. We observe that (1) if the population is the bigger and if it contains individuals the most close genetically, and that were phenotyped in similar environments compared to PV, prediction accuracy is better (2) optimizing the training set with the larger genetic diversity, the better the predictions are (+9% on average) (3) the larger the heritability of the trait value is, the better the predictions are (3) the statistical prediction models perform very similarly to estimate accuracies (4) statistical models lead to very similar predictions (5) the optimal markers density is depending of the population LD.

To conclude, although SG still a practice which is still in its infancy in plants, applying GS in tomato seems very promising and could be used in tomato breeding programs.



**Keys words:** Genomic selection, *Solanum lycopersicum,* cross validation, G×E interaction, heritability