Développement de méthodes statistiques pour l'identification de gènes d'intérêt en présence d'apparentement et de dominance, application à la génétique du maïs

Fabien Laporte

NNT : 2018SACLS066

THÈSE DE DOCTORAT

de

L'Université Paris-Saclay

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : Université Paris-Sud

Établissements d'accueil : AgroParisTech
Institut national de la recherche agronomique

Laboratoires d'accueil : Mathématiques et informatique appliquées, UMR 518 INRA
Génétique Quantitative et Évolution - Le Moulon, UMR 320 INRA-CNRS

Spécialité de doctorat : Mathématiques Appliquées

Fabien LAPORTE

Développement de méthodes statistiques pour l'identification de gènes d'intérêt en présence d'apparentement et de dominance, application à la génétique du maïs

Date de soutenance : 13 Mars 2018

Après avis des rapporteurs : Mathieu EMILY (Agrocampus Ouest)
Grégory NUEL (Université Pierre et Marie Curie)

Jury de soutenance :

Christophe AMBROISE	(Université Évry Val-d'Essonne) Examinateur
Alain CHARCOSSET	(INRA) Invité, Codirecteur de thèse
Mathieu EMILY	(Agrocampus Ouest) Rapporteur
Elisabeth GASSIAT	(Université Paris-Sud) Présidente
Brigitte MANGIN	(INRA) Examinateur
Tristan MARY-HUARD	(INRA) Directeur de thèse
Grégory NUEL	(Université Pierre et Marie Curie) Rapporteur

Remerciements

Je remercie Christophe Ambroise, Mathieu Emily, Elisabeth Gassiat, Brigitte Mangin et Grégory Nuel pour avoir accepté d'être les examinateurs de ma thèse. Je remercie particulièrement Mathieu Emily et Grégory Nuel pour avoir accepté de rapporter mes travaux de thèse.

Je souhaite remercier tout particulièrement mes directeurs de thèse, Tristan Mary-Huard et Alain Charcosset, qui m'ont permis de progresser dans mes travaux de recherche et qui m'ont fait confiance tout le long de mon doctorat. Je remercie Tristan pour ses enseignements tant en statistiques que pour la planification et la préparation d'exposés. Je le remercie aussi pour ses conseils pour la suite de ma carrière (même si elle n'est pas encore tracée). Merci à sa bonne humeur qui a permis de passer des moments agréables malgré la pression à certains moments. Je remercie Alain pour ses enseignements en génétique (malgré mes questions "bêtes" : "qu'est-ce qu'un QTL?"), sa disponibilité à chaque instant où j'avais une question, ses relectures (je n'ai pas compté le nombre de fautes d'orthographe corrigées mais on doit dépasser la centaine) et sa bonne humeur quotidienne. Je remercie aussi mes encadrants pour leur patience.

Je remercie aussi Julie Fievet sans qui l'étude des données n'aurait pas été possible. Mais je la remercie surtout pour sa disponibilité, son arbitrage entre mes encadrants (type I, type II et type III?), ses encouragements et son soutien en fin de doctorat (les pauses cigarettes malgré ses multiples "arrêts"). Et bien entendu, je la remercie pour sa bonne humeur (les jours où il n'y a pas de problème de transport).
Je remercie Laurence Moreau pour ses discussions clairvoyantes (cette fois j'ai réussi à le placer) sur la GWAS et la décomposition en aptitudes générales à la combinaison et spécifiques à la combinaison. Je la remercie aussi pour ses connaissances sur l'estimation des matrices d'apparentement entre les individus (malgré le fait que je n'ai pas réussi à montrer quelle matrice est la meilleure en GWAS).
Je remercie Stéphane Nicolas pour ses explications sur les matrices de génotypage et le déséquilibre de liaison (même si je ne l'ai pas étudié au cours de mon doctorat, ça permet de réfléchir à de potentielles études). Je le remercie aussi pour les cartes génétiques que j'ai utilisées pour les analyses.
Je remercie Valérie Combes et Delphine Madur pour les données de génotypage, les explications et les vérifications sur ces dernières. Je remercie Valérie aussi pour tous les bons gâteaux que j'ai pu manger durant mon doctorat.
Je remercie Cyril Bauland pour la coordination des plate-formes de phénotypage et pour m'avoir fait découvrir ce qu'est une feuille de maïs (un petit arrêt de 10 minutes lors

d'une réunion pour me faire un dessin). Je le remercie aussi pour m'avoir emmener dans les champs, le seul jour où je n'avais pas de chemise, et pour le bizutage de F7p.
Je remercie Clément Mabire, Simon Rio et Adama Seye, ma co-bureau Camille Clipet et Emma Forst pour avoir été les cobayes (enthousiastes) de mon algorithme d'estimation des modèles mixtes. Je remercie Antoine Allier pour m'avoir introduit d'autres algorithmes dont je ne connaissais pas l'existence.
Je remercie bien entendu toute l'équipe GQMS pour leur bonne humeur quotidienne qui a permis de passer d'agréables moments (même au bureau).

Je souhaite remercier tous les partenaires du projet Amaizing tant pour les données de phénotypage que pour les échanges scientifiques lors des réunions annuelles. Je remercie aussi les membres du méta-programme SelGen pour les échanges et les pistes de recherche.

Je remercie tous les membres de l'UMR Génétique Quantitative et Evolution pour leur bonne humeur et leur accueil chaleureux (bien qu'étant un statisticien). Je remercie particulièrement Christophe Lecarpentier pour son exemple de thèse (je suis en avance en fait) et les soirées jeux, Pierre Montalent et Yannick De Oliveira pour l'animation basket (et bien entendu tous ses membres) pour la détente du Jeudi midi, Gaëlle Van Frank et Julie Borg pour la reprise des commandes de pains (qui m'ont nourris pendant le dernier mois), Zeineb Achour pour sa bonne humeur et les ragots de notre immeuble, Natalia Martinez pour les repas du midi (bref mais relaxants).
Je remercie aussi certains qui ne sont plus présents au Moulon aujourd'hui mais qui m'ont bien aidé, Margaux-Alison Fustier pour ses gaffes et histoires palpitantes, Jean-Tristan Brandenburg pour son enseignement en C et les pauses cigarettes, Cyril Martin pour sa gestion militaire et Stéphanie Thépot pour toutes les après-midi/soirées jeux de société.

Je remercie également les membres de l'école doctorale Mathématiques Hadamard, et en particulier Stéphane Nonnenmacher et Christelle Pires pour avoir facilité les différentes tâches administratives.

Je remercie ma famille pour m'avoir soutenu pendant ces années, même si certaines fois il ne fallait tout simplement pas me parler de ma thèse.
Enfin je remercie Charlotte Urien pour son soutien moral au quotidien, ses petits plats (que j'attends toujours), son émerveillement devant chaque ligne de code R qui marche et tout simplement pour être à mes côtés (même si elle était à la Corogne ou à Bayonne) chaque jour.

Contents

Chapter 1

Introduction

1.1 Contexte Génétique

En génétique quantitative, les scientifiques cherchent à établir une relation entre le phénotype (un caractère observé tel que la taille, le rendement, le développement de maladie, etc.) des individus et leur information génétique. Les objectifs sont, dans ce cas, de prédire le phénotype en se basant sur l'information génétique des individus et de comprendre le rôle de l'information génétique dans l'expression de certains phénotypes. On pourrait ainsi caractériser, par exemple, le risque de développer des maladies, le rendement des plantes cultivées ou la quantité de lait produite pour les vaches.

L'information génétique (que ce soit les gènes ou les marqueurs moléculaires) est contenue dans la molécule d'ADN qui est organisée en chromosomes. Pour les espèces haploïdes, chaque chromosome est présent en une seule copie alors que chez les espèces diploïdes, les chromosomes sont présents par paire. Dans ce cas, pour chaque paire, un des deux chromosomes homologues provient du parent femelle et l'autre du parent mâle. Une position précise sur ces chromosomes est appelée locus. Chez les espèces diploïdes, à chaque locus, un individu peut porter soit deux allèles différents (i.e. deux versions différentes d'un même locus) et dans ce cas il est hétérozygote à ce locus soit deux allèles identiques et dans ce cas il est homozygote à ce locus.
Afin de pouvoir se repérer sur la molécule d'ADN, et donc de suivre les régions d'intérêt, on utilise des marqueurs moléculaires. Un marqueur moléculaire est une position sur la molécule d'ADN dont on connaît la position, la séquence d'ADN et dont la valeur ne dépend pas de l'environnement. Les marqueurs moléculaires permettent de connaître le génotype (la version de l'allèle) portée par un individu. Pour un diploïde, chaque observation d'un marqueur est donc composée de deux valeurs correspondant aux allèles des deux chromosomes homologues à cette position. Si on se place dans le cas d'un marqueur bi-allélique (seulement deux allèles possibles à ce marqueur), sa valeur observée peut se résumer aux nombres de présence de l'allèle de référence. Si tous les marqueurs sont bi-alléliques dans une population, on obtient donc, pour chaque individu, un vecteur comprenant des 0, 1 ou 2.

Il existe deux grandes stratégies pour évaluer les valeurs phénotypiques des individus. La première consiste à observer l'individu lui même, on parle d'observation *per se* ou en valeur propre. La seconde consiste à observer la descendance des individus considérés, on parlera alors de valeur phénotypique en croisement. Le choix entre ces deux stratégies est essentiellement dicté par le régime de reproduction de l'espèce étudiée et donc du type de variété qui est commercialisé.

La relation entre le phénotype et le génotype d'un individu peut se modéliser de la manière suivante :

$$\begin{aligned} &Y_i = \mu + G_i + E_i \\ &G_i \perp E_i \end{aligned} \tag{1.1}$$

où Y_i est le phénotype de l'individu i, μ est la moyenne de ce phénotype, G_i est l'effet dû à l'information génétique de l'individu i et E_i est l'effet dû à l'environnement.

Si on note $X_{i,\ell}$, la valeur observée au marqueur ℓ chez l'individu i alors on peut décomposer la valeur génétique de l'individu i comme étant la somme des effets de chaque marqueur, *i.e.* :

$$G_i = \sum_{\ell=1}^{L} X_{i,\ell}\beta_\ell$$

où L est le nombre de marqueurs observés et β_ℓ est l'effet associé au marqueur ℓ. Il existe plusieurs cas possibles pour les β_ℓ. Le premier cas est celui où toute l'influence du génotype est concentrée sur un seul marqueurs, il existe ℓ tel que $\beta_\ell \neq 0$ et $\beta_{\ell'} = 0$ pour tout $\ell' \neq \ell$. On parlera d'un cas monogénique. Si plusieurs β_ℓ sont non nuls, on parlera alors d'un cas polygénique. Les marqueurs ℓ qui ont une valeur non nulle sont appelés QTL (*Quantative Trait Locus*). Dans la suite, on considérera avec attention le cas polygénique.
Si on suppose que les observations des marqueurs sont indépendantes entre un marqueur ℓ et un marqueur ℓ' alors on peut ainsi réécrire la covariance génétique de deux individus i et j comme :

$$cov(G_i, G_j) = \sum_{\ell=1}^{L} cov(X_{i,\ell}, X_{j,\ell})\beta_\ell^2$$

Si on suppose qu'il y a indépendance entre le génome des parents d'un même individu et que les marqueurs sont bi-alléliques, en utilisant la corrélation entre les observations des marqueurs, on peut obtenir :

$$cov(G_i, G_j) = \sum_{\ell=1}^{L} cor(X_{i,\ell}, X_{j,\ell}) 4 p_\ell q_\ell a_\ell^2 \tag{1.2}$$

où p_ℓ est la fréquence de l'allèle majoritaire et $q_\ell = 1 - p_\ell$. En supposant que la corrélation est la même pour tous les marqueurs, avec $K_{i,j} = cor(X_{i,\ell}, X_{j,\ell})$, alors on obtient une formulation de la covariance génétique dépendante du paramètre de ressemblance génétique $K_{i,j}$:

$$cov(G_i, G_j) = K_{i,j}\sigma_g^2$$

avec $\sigma_g^2 = \sum_{\ell=1}^{L} 4 p_\ell q_\ell \beta_\ell^2$.
Si dans le modèle (1.1), on suppose que les effets environnementaux sont indépendants

entre deux individus (i.e. $E_i \perp E_j$ pour tout $i \neq j$) alors la covariance phénotypique entre individus ne dépend que de la covariance génétique entre individus :

$$cov(Y_i, Y_j) = cov(G_i, G_j)$$

En utilisant toutes les hypothèses précédemment citées et en approximant la somme des effets des marqueurs par une loi normale, on obtient le modèle matriciel suivant, qui relie le phénotype au génotype :

$$\begin{aligned} &Y = 1\mu + G + E \\ &G \perp E \\ &G \sim \mathcal{N}(0, K\sigma_g^2) \\ &E \sim \mathcal{N}(0, I\sigma_e^2) \end{aligned} \tag{1.3}$$

où Y est le vecteur des phénotypes de tous les individus, μ est l'intercepte du modèle, G est le vecteur des effets aléatoires dûs aux génotypes des individus et E est le vecteur des effets aléatoires dûs à l'environnement. En génétique bovine, ce modèle est utilisé pour caractériser la valeur génétique des taureaux prédite par le modèle, $\widehat{G}$, en utilisant des méthodes de prédiction telles que le *Best Linear Unbiased Predictor* (BLUP).

Historiquement, l'apparentement entre individus était estimé à partir du pedigree de ces derniers. La valeur estimée dans ce cas est l'apparentement attendu entre deux individus, il correspond à un niveau moyen d'apparentement. En génétique végétale, le recours au croisement dirigé permet de facilement suivre le pedigree. Par exemple, si on effectue des observations sur une population composée de demi-soeurs (plantes issues d'un même parent commun) alors le pedigree est simple et l'apparentement attendu estimé à l'aide de ce dernier entre deux individus est de $\frac{1}{4}$ pour tous les couples. Cependant, l'utilisation de l'apparentement dans le modèle (1.3) n'a pas été immédiate. Dans l'exemple précédent, l'estimation de ce coefficient à l'aide du pedigree ne permet pas de prendre en compte la variance autour de l'apparentement, et on peut se retrouver avec une population d'individu dont tous les couples ont un même niveau d'apparentement. Les marqueurs moléculaires permettent d'aller plus loin dans l'estimation de l'apparentement. L'utilisation de ces derniers permet de prendre en compte la variance autour de l'apparentement moyen et donc d'estimer un apparentement réalisé entre les individus. Il existe différentes manières de procéder à l'estimation, par exemple, on peut considérer le coefficient de similarité génétique entre deux individus i et j à l'aide de la formule suivante :

$$K_{i,j} = \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{1}_{X_{i,\ell}=X_{j,\ell}}$$

Si tous les marqueurs sont bi-alléliques au sein de la population observée, alors on peut aussi utiliser l'équation (1.2) pour obtenir un estimateur basé sur la corrélation entre les observations des marqueurs (Astle and Balding, 2009) :

$$K_{i,j} = \frac{1}{L} \sum_{\ell=1}^{L} \frac{(X_{i,\ell} - p_\ell)(X_{j,\ell} - p_\ell)}{4p_\ell(1 - p_\ell)}$$

Il existe d'autres façons de modéliser le coefficient $K_{i,j}$ (Thompson, 1975; Milligan, 2003) comme détaillé dans le chapitre 2. L'utilisation des marqueurs pour estimer la valeur d'apparentement permet donc de différencier des valeurs qui étaient égales sur la base du pedigree ou encore d'estimer l'apparentement lorsque le pedigree est incomplet ou inconnu. Suite au développement des marqueurs moléculaires, le modèle (1.3) et l'estimation de ses paramètres ont pu être effectués dans le contexte de la génétique végétale (Bernardo, 1994).

Il est en plus devenu possible de détecter les marqueurs associés à une forte variation phénotypique à l'aide du modèle proposé par Yu et al. (2006) :

$$\begin{aligned} &Y = 1\mu + X_\ell \beta_\ell + ZG + E \\ &G \perp E \\ &G \sim \mathcal{N}(0, K\sigma_g^2) \\ &E \sim \mathcal{N}(0, I\sigma_e^2) \end{aligned} \tag{1.4}$$

où X_ℓ est le génotype des individus au marqueur ℓ, β_ℓ est l'effet fixe dû à ce même marqueur, Z est la matrice d'incidence des individus (on peut avoir des individus répétés), G est l'effet aléatoire provenant des autres marqueurs (fond polygénique) et E est un effet résiduel. Une fois les paramètres du modèle (1.4) estimés, on effectue le test $\beta_\ell = 0$ contre $\beta_\ell \neq 0$ pour détecter si le marqueur ℓ joue un rôle sur la variation du caractère phénotypique étudié. Pour effectuer la détection de gènes d'intérêt, il faut ajuster ces modèles sur tous les marqueurs disponibles (actuellement, chez le maïs, nous avons accès à un million de marqueurs). Il est donc important d'avoir des algorithmes d'estimation de modèles mixtes performants.

Le modèle (1.4) s'applique très bien lorsqu'on s'intéresse aux individus provenant d'une seule population. En effet, l'intercepte est le même pour tous les individus. L'effet du marqueur est le même pour tous les individus et on considère que cet effet est additif, il s'additionne en fonction de la dose d'allèle présent à ce marqueur (0, 1 ou 2 pour les individus bi-alléliques).

1.2 Cas spécifique : plan de croisement chez les plantes

Chez les plantes, les individus étudiés peuvent être des lignées issues d'autofécondations successives, ce qui confère (i) un état homozygote des individus pour l'ensemble des locus du génome et (ii) une identité génétique entre individus d'une même lignée. L'observation de nombreux individus identiques génétiquement permet d'évaluer la valeur moyenne d'une lignée pour un caractère, qui sera nettement plus précise qu'une observation sur une plante unique. Le modèle (1.4) est adapté à la détection de gènes d'intérêt sur une population composée de lignées. Toutefois, chez un certain nombre d'espèces comme le maïs, l'état homozygote peut conduire à une perte de vigueur que l'on appelle dépression de consanguinité. Pour ces espèces, le matériel cultivé est donc des variétés hybrides (Shull, 1908) issues du croisement entre deux lignées homozygotes. On a alors accès à des individus plus performants que leurs parents : c'est l'effet d'hétérosis (Shull, 1914). Cela implique donc l'existence d'effet de dominance en plus des effets additifs.

Ces effets sont représentés schématiquement dans la figure 1.1.

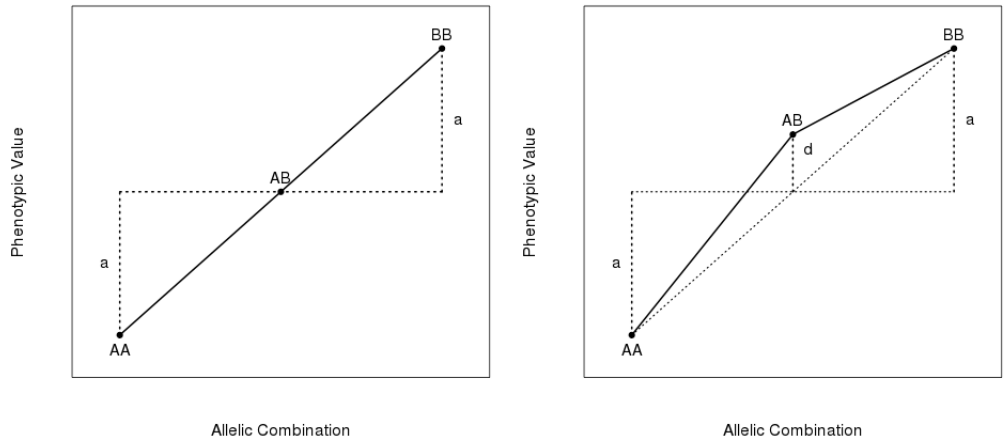


Figure 1.1: Gauche : effet des combinaisons alléliques avec un effet additif (a). Droite : effet des combinaisons alléliques avec un effet additif (a) et un effet de dominance (d)

Si les plantes mères dérivent de deux populations différentes, en plus de l'effet de dominance, on peut aussi s'intéresser aux effets polygéniques (l'effet G du modèle (1.4)) provenant de chacune des populations. Dans ces cas là, le modèle de Yu et al. (2006) ne peut pas s'appliquer. Lorsqu'on étudie des hybrides provenant d'un croisement entre deux populations (une population Mère et une population Père) alors le modèle (1.4) se généralise comme suit :

$$\begin{aligned} &Y = \mu + X_{\ell,M}\beta_{\ell,M} + X_{\ell,P}\beta_{\ell,P} + X_{\ell,M\times P}\beta_{\ell,M\times P} + Z_M G_M + Z_P G_P + Z_H G_H + E \\ &G_M \sim \mathcal{N}(0, K^{(M)}\sigma^2_M) \\ &G_P \sim \mathcal{N}(0, K^{(P)}\sigma^2_P) \\ &G_H \sim \mathcal{N}(0, \Phi\sigma^2_H) \\ &E \sim \mathcal{N}(0, I\sigma^2_e) \end{aligned} \tag{1.5}$$

où $X_{\ell,M}$ (resp. $X_{\ell,P}$, $X_{\ell,M\times P}$) est le génotype des mères (resp. des pères, des hybrides) au marqueur ℓ, $\beta_{\ell,M}$ (resp. $\beta_{\ell,P}$) est l'effet dû aux marqueurs présents chez les mères (resp. les pères), $\beta_{\ell,M\times P}$ est l'effet dû à l'interaction des génotypes mères et pères, Z_M (resp. Z_P) est la matrice d'incidence reliant les hybrides à leur mère (resp. père), Z_H est la matrice d'incidence des hybrides, G_M (resp. G_P) représente l'effet du fond polygénique dû à la population maternelle (resp. paternelle) avec $K^{(M)}$ (resp. $K^{(P)}$) l'apparentement entre les lignées mères (resp. pères), G_H est l'effet dû aux interactions entre le fond polygénique maternel et paternel avec Φ un indice de double ressemblance et E est un vecteur résiduel.
La ressemblance génétique entre les individus du croisement précédemment cité peut intervenir de trois manières. La première (resp. la seconde) est une ressemblance de l'information génétique apportée par les mères (resp. les pères) et est représentée par la matrice $K^{(M)}$ (resp. $K^{(P)}$). La troisième manière est une ressemblance double, *i.e.*

à la fois les informations maternelles et paternelles sont identiques et est représentée par la matrice Φ. Dans les deux premiers cas, on parlera de simple apparentement entre les parents et dans le troisième cas on parlera de double apparentement entre les hybrides.

D'autres études peuvent faire intervenir une modélisation plus complexe de la covariance génétique donnée en (1.2), comme par exemple dans Gallais (1990). Elles font intervenir d'autres coefficients d'apparentement tel que le coefficient de consanguinité (ressemblance génétique entre les chromosomes homologues d'un individu), en particuliers lorsqu'on croise des lignées maternelles d'une même population (un croisement mère $\times$ mère dans l'exemple précédent). Il faut donc s'intéresser à la modélisation de l'apparentement entre individus (Thompson, 1975; Milligan, 2003).
De plus, de nouveaux effets aléatoires apparaissent dans ce modèle, ce qui engendre un temps de calcul plus long pour l'estimation des modèles mixtes. On peut donc voir l'importance d'avoir des algorithmes performants pour ces estimations.

1.3 Objectifs de la thèse et plan général du manuscrit

Les différents modèles mettant en relation le phénotype et le génotype des individus cités précédemment font intervenir de nombreux coefficients de relation génétique entre les individus. Il est donc important de s'intéresser à l'estimation de ces coefficients dans un premier temps.
Nous nous sommes intéressés à la modélisation des coefficients de l'apparentement à partir de l'observation de marqueurs bi-alléliques en utilisant un modèle de mélange (Thompson, 1975; Milligan, 2003). Une fois le modèle écrit, nous avons étudié l'identifiabilité de ce modèle dans un cadre général et aussi dans le cadre de plan de croisement d'intérêt pour la sélection végétale. En effet ces plans de croisement nous apportent des informations supplémentaires telles que l'origine parentale des allèles (maternelle ou paternelle), la classification en populations et des parents communs *i.e.* partagés par plusieurs individus. Ces nouvelles informations jouent un rôle notable sur l'identifiabilité du modèle. Enfin nous avons développé un algorithme permettant l'estimation des paramètres d'apparentement entre individus à partir de marqueurs bi-alléliques qui peut prendre en compte ces informations. Les résultats de cette partie ont été publiés dans Laporte et al. (2017).

Les paramètres des modèles mixtes (1.4) et (1.5) doivent être estimés pour chaque marqueur étudié. Le nombre croissant de marqueurs disponibles allonge considérablement les temps de calcul des algorithmes d'inférence. De plus la prise en compte de nouveaux effets aléatoires dans le modèle (1.5) nous a amené à réfléchir à la performance des algorithmes d'inférence des modèles mixtes à composantes de la variance.
Les algorithmes d'inférence de modèles mixtes peuvent se classer en trois catégories. La première catégorie est celle des algorithmes d'optimisation directe de la vraisemblance. Ceux-ci ne peuvent être appliqués qu'à des modèles comportant deux effets aléatoires comme dans le modèle (1.4). En effet, ils se basent sur la diagonalisation simultanée des matrices de corrélation et des algorithmes d'optimisation à une variable, comme l'exemple de *FastLmm* développé par Lippert et al. (2011). A l'opposé, la seconde catégorie est celle des algorithmes de second ordre (type Newton-Rapthon). Ils se basent

sur une méthode itérative prenant en compte la matrice hessienne de la vraisemblance ou une matrice dérivée (comme l'espérance de cette dernière pour le Fisher Scoring). Un algorithme de second ordre largement utilisé en génétique des plantes est celui de la matrice d'information moyenne (AI) proposé par Gilmour et al. (1995) et utilisé dans l'algorithme *ASReml*. Durant mon doctorat, je me suis particulièrement intéressé à la dernière catégorie, celle des algorithmes de premier ordre. Ces derniers se basent aussi sur une méthode itérative mais font intervenir seulement le gradient de la vraisemblance. L'algorithme de MinMax (MM) présenté par Hunter and Lange (2004) et appliqué aux modèles mixtes par Zhou et al. (2015) en fait partie. Ces formules itératives simples nous ont permis de penser que ce dernier serait plus rapide que les algorithmes des autres classes. De plus, les astuces de calcul, telle que la diagonalisation simultanée des matrices, peuvent y être adaptées.
Nous avons codé un algorithme basé sur la méthode MM, en utilisant des astuces d'accélération dont celle proposée par Varadhan and Roland (2008). Nous avons ensuite comparé les performances des différentes méthodes et algorithmes du point de vue de la détection de gènes d'intérêt, à la fois en précision et en temps de calcul, en utilisant des jeux de données issus de la génétique végétale et présents dans la bibliographie.

Nous avons enfin effectué la détection de QTL sur un panel d'hybrides issus d'un croisement de deux populations distinctes (apparentement nul entre deux individus issus de populations différentes) en utilisant les algorithmes développés précédemment. Notre première réflexion a concerné la modélisation des phénotypes en fonction du génotype des individus. Nous avons été amenés à utiliser le modèle (1.5) dans le cadre de cette étude. Le second défi a été l'utilisation de différentes méthodes pour estimer les coefficients de l'apparentement dans le cadre de ce panel. Nous pouvons mettre en concurrence deux méthodes, une issue de la bibliographie Astle and Balding (2009) et celle que nous avons développée. Nous souhaitons voir l'impact de l'utilisation de ces différentes méthodes d'estimation de l'apparentement sur le détection de QTL. Enfin nous avons détaillé les hypothèses de test possibles dans le cadre de ce modèle et nous nous sommes penchés sur la pertinence de ces tests pour la détection de gènes d'intérêt.
Les hybrides ayant été phénotypés dans plusieurs lieux, nous avons également réalisé une étude en prenant en compte l'interaction du génotype avec l'environnement. Le défi ici a été la mise en place du modèle et l'optimisation des algorithmes pour permettre d'estimer les paramètres nécessaires.

Ma thèse est construite en trois chapitres. Le premier traite l'estimation de l'apparentement à partir de l'observation de marqueurs bi-alléliques. Le deuxième étudie l'estimation des paramètres des modèles mixtes à composantes de la variance. Et le dernier est une application des outils précédemment développés à la détection de gènes d'intérêt dans un plan de croisement hybride. Une dernière partie est consacrée aux conclusions générales et perspectives de mon doctorat.

Chapter 2

Estimation of relatedness coefficients from biallelic markers, application in plant mating designs

The relatedness between two individuals is the distribution (over all loci) of the number of alleles inherited from one (or several) common ancestor(s). The concept of relatedness (and its decomposition) was introduced by Wright (1922) back in the 1920s, and has been extensively investigated since then (Crow and Kimura, 1970). This is indeed an important concept in quantitative and population genetics. Relatedness can be useful on its own, for instance to quantify the level of consanguinity in a population (Crow and Kimura, 1970). It has also proven to be an important component of association genetics models, where the so-called kinship matrix is now widely used to account for the effect of genetic background on the phenotypic response (Yu et al., 2006).
Two approaches have been considered to infer relatedness, depending on the data at hand. Relatedness was first inferred from pedigree (Crow and Kimura, 1970). Nowadays, thanks to new genotyping technologies, the relatedness can be inferred from markers. The inference of relatedness from markers was first introduced by Thompson (1975), and was followed by numerous contributions (see for instance McPeek and Sun (2000), Milligan (2003), Hepler (2005), Bink et al. (2008) and Astle and Balding (2009)). Several statistical strategies have been considered to infer relatedness between individuals derived from a single population from genotypic data, many of these methods being currently available as R packages or algorithms (Coancestry, Wang (2011), ML-Relate, Kalinowski et al. (2006)). Some of the aforementioned methods actually aim at estimating only some specific components of the relatedness distribution (e.g. coefficients k_0, k_1 and k_2 in Thompson (1975), or the simple relatedness coefficient in Astle and Balding (2009)). These methods can be used whatever the genotypic information available, including SNP data. Other methods aim at estimating the whole relatedness distribution (Coancestry, ML-Relate). It has been recently shown that in this case, the information provided by biallelic marker data may be too poor for the full recovery of the relatedness

distribution (Csuros, 2014). This may be a reason why some of the available softwares (e.g. Coancestry) require the markers to be multi-allelic to perform inference. With the current development of (chip or sequencing) SNP data, this appears to be a severe limitation.

In this chapter we consider the plant genetics framework, where additional information about the relationship between individuals is usually available. This additional information may be of two kinds. First, when considering panels of hybrids derived from crosses between lines, one can partially retrieve relatedness information from the crossing design. Second, the lines considered in the crossing design may belong to different populations or heterotic groups. This membership is informative since lines belonging to different populations cannot be related. Therefore the crossing design and population membership should be accounted for in relatedness inference procedures. To this end, the contribution of the present chapter is double. From a theoretical point of view, we consider the problem of estimating relatedness from SNP data in a hybrid panel through the framework developed by Thompson (1975), Milligan (2003) and Csuros (2014). In this framework, inferring the (unknown) relatedness status at each marker from the (observed) genotypic information can be cast into a mixture model where the goal is to estimate the proportion of loci associated with each relatedness status over the genome. We investigate in which cases this model is identifiable or not depending on the additional information at hand. In particular we show that important quantities such as the double relatedness coefficient may or may not be estimable, depending on the crossing design. From a practical point of view, we provide an R package for the Maximum Likelihood Estimation (MLE) of the relatedness coefficients that handle both the classical case (unphased data, individuals belonging to a single population) and the case where estimation is performed on a panel of hybrids for which the crossing design is available.

The chapter is organized as follows. The classical case is presented in Section 2.1. The modification of the classical case to the case of line crossing designs is detailed in Section 2.2, with a detailed characterization of identifiability and estimation for each configuration (number of common parental lines between hybrids, membership of the lines to a same or to different populations). Applications of the proposed methodology is presented in Section 2.3 on both simulated and real data.

2.1 Statistical Framework

The purpose of this section is to provide a statistical framework for the estimation of relatedness. We rely on the model initially proposed by Thompson (1975) and also considered by Milligan (2003), where it is assumed that the genotypes of two individuals sampled from a same population are observed, and that these genotypes are unphased. Marker allelic frequencies are supposed to be known. We also briefly recall the results of Csuros (2014) regarding model identifiability. In Section 2.2, this framework will be adapted to the case where individuals are hybrids obtained by crossing of parental lines.

2.1.1 Statistical Modeling

We first consider the observation of the genotypes of two individuals at a given marker ℓ. Four alleles are observed (2 alleles per individual), with value 0 or 1 since

markers are all assumed to be biallelic. These four alleles together define the Identity By State (IBS) configuration of the marker, that can be one of the nine possible configurations described in Table 2.1. These four alleles can be inherited from one or several ancestors according to one of the 9 possible IBD (Identity by Descent) configurations of Table 2.2.

o_1	$\{00,00\}$	o_1^p
o_2	$\{00,01\}$	o_2^p
	$\{00,10\}$	o_3^p
o_3	$\{00,11\}$	o_4^p
o_4	$\{01,00\}$	o_5^p
	$\{10,00\}$	o_6^p
o_5	$\{01,01\}$	o_7^p
	$\{01,10\}$	o_8^p
	$\{10,01\}$	o_9^p
	$\{10,10\}$	o_{10}^p

o_6	$\{01,11\}$	o_{11}^p
	$\{10,11\}$	o_{12}^p
o_7	$\{11,00\}$	o_{13}^p
o_8	$\{11,01\}$	o_{14}^p
	$\{11,10\}$	o_{15}^p
o_9	$\{11,11\}$	o_{16}^p

Table 2.1: IBS configurations between two individuals. The first two numbers correpond to the alleles of individual 1 and the last two numbers correspond to the alleles of individual 2. Configurations $o_1, ..., o_9$ (respectively $o_1^p, ..., o_{16}^p$) correspond to distinguishable IBS configurations when the data are unphased (respectively phased).

c_1	$\{AA,AA\}$	c_1^p
c_2	$\{AA,BB\}$	c_2^p
c_3	$\{AA,AB\}$	c_3^p
	$\{AA,BA\}$	c_4^p
c_4	$\{AA,BC\}$	c_5^p
c_5	$\{AB,AA\}$	c_6^p
	$\{AB,BB\}$	c_7^p
c_6	$\{AB,CC\}$	c_8^p

c_7	$\{AB,AB\}$	c_9^p
	$\{AB,BA\}$	c_{10}^p
c_8	$\{AB,AC\}$	c_{11}^p
	$\{AB,CA\}$	c_{12}^p
	$\{AB,BC\}$	c_{13}^p
	$\{AB,CB\}$	c_{14}^p
c_9	$\{AB,CD\}$	c_{15}^p

Table 2.2: IBD configurations between two individuals. The first (respectively the last) two letters correspond to the alleles of individual 1 (respectively individual 2). Letters A, B, C and D refer to ancestor alleles. Two alleles have the same letter if they are both descended from a single allele in a common ancestor. Configurations $c_1, ..., c_9$ (respectively $c_1^p, ..., c_{15}^p$) correspond to distinguishable IBD configurations when the data are unphased (respectively phased).

Denote p_ℓ and $q_\ell = 1 - p_\ell$ the allelic frequencies of alleles 0 and 1 at marker ℓ, respectively. Furthermore, define IBS_ℓ the (observed) random variable corresponding to the IBS configuration at locus ℓ, and IBD_ℓ the (hidden) random variable corresponding to the IBD configuration at this same locus. The likelihood of IBS_ℓ can be written as

follows:

$$\begin{aligned} P_\Delta(\text{IBS}_\ell = o_i) &= \sum_{j=1}^{9} P(\text{IBS}_\ell = o_i | \text{IBD}_\ell = c_j) P(\text{IBD}_\ell = c_j) \\ &= \sum_{j=1}^{9} P(\text{IBS}_\ell = o_i | \text{IBD}_\ell = c_j) \Delta_j \ , \end{aligned} \tag{2.1}$$

where $\Delta_j = P(\text{IBD}^\ell = c_j)$. Note that the proportion parameters $\Delta_1, ..., \Delta_9$ do not depend on the particular marker that is considered. The vector $\Delta = (\Delta_1, ..., \Delta_9)^T$, where T denote the transpose function, is called the IBD distribution or the relatedness distribution, and Δ satisfies

$$\Delta \in \mathcal{S}_+^9 \ , \text{where } \mathcal{S}_+^J = \left\{ x \in \mathbb{R}^J : \ x_i \geq 0, \ \sum_{i=1}^{J} x_i = 1 \right\}$$

Likelihood (2.1) can be straightforwardly generalized to the case where several independent markers $\ell = 1, ..., L$ are considered:

$$\prod_{\ell=1}^{L} \left(\sum_{j=1}^{9} \Delta_j P(\text{IBS}_\ell = o_i | \text{IBD}_\ell = c_j) \right) \ .$$

Note that in the previous expression the conditional probabilities only depend on the (known) marker allelic frequencies. The model is then a mixture model with known emission distributions that differ for one marker to another, where the unknown parameters are the proportions $\Delta_1, ..., \Delta_9$ that have to be estimated. In the following we will also consider the matrix representation for this model. For locus ℓ one has

$$P_\Delta^v(\text{IBS}_\ell) = M_\ell \Delta \tag{2.2}$$

where

- $P_\Delta^v(\text{IBS}_\ell) = (P_\Delta(\text{IBS}_\ell = o_1), ..., P_\Delta(\text{IBS}_\ell = o_9))^T$
- $M_\ell = (m_{ij}^\ell)_{1 \leq i \leq 9, 1 \leq j \leq 9}$ with $m_{ij}^\ell = P(\text{IBS}_\ell = o_i | \text{IBD}_\ell = c_j)$,
- $\Delta = (\Delta_1, ..., \Delta_9)^T$.

Note that matrix M_ℓ only depends on allelic frequencies p_ℓ and q_ℓ. For locus ℓ, M_ℓ can be expressed as followed:

$$M_\ell = \begin{pmatrix} q_\ell & q_\ell^2 & q_\ell^2 & q_\ell^3 & q_\ell^2 & q_\ell^3 & q_\ell^2 & q_\ell^3 & q_\ell^4 \\ 0 & 0 & q_\ell p_\ell & 2q_\ell^2 p_\ell & 0 & 0 & 0 & q_\ell^2 p_\ell & 2q_\ell^3 p_\ell \\ 0 & q_\ell p_\ell & 0 & q_\ell p_\ell^2 & 0 & q_\ell^2 p_\ell & 0 & 0 & q_\ell^2 p_\ell^2 \\ 0 & 0 & 0 & 0 & q_\ell p_\ell & 2q_\ell^2 p_\ell & 0 & q_\ell^2 p_\ell & 2q_\ell^3 p_\ell \\ 0 & 0 & 0 & 0 & 0 & 0 & 2q_\ell p_\ell & q_\ell p_\ell & 4q_\ell^2 p_\ell^2 \\ 0 & 0 & 0 & 0 & q_\ell p_\ell & 2q_\ell p_\ell^2 & 0 & q_\ell p_\ell^2 & 2q_\ell p_\ell^3 \\ 0 & q_\ell p_\ell & 0 & q_\ell^2 p_\ell & 0 & q_\ell p_\ell^2 & 0 & 0 & q_\ell^2 p_\ell^2 \\ 0 & 0 & q_\ell p_\ell & 2q_\ell p_\ell^2 & 0 & 0 & 0 & q_\ell p_\ell^2 & 2q_\ell p_\ell^3 \\ p_\ell & p_\ell^2 & p_\ell^2 & p_\ell^3 & p_\ell^2 & p_\ell^3 & p_\ell^2 & p_\ell^3 & p_\ell^4 \end{pmatrix}$$

2.1.2 Model Identifiability

Although Model (2.2) has been largely used to estimate IBD from genotypic data (Milligan, 2003), little attention has been devoted to the identifiability of this model. A parametric model P_θ is identifiable if

$$P_\theta = P_{\theta'} \Rightarrow \theta = \theta'.$$

It is only recently that this question has been addressed in Csuros (2014), where the author proved that when the data are unphased and the markers biallelic, there exist two IBD parameter sets Δ and Δ' such that

$$\begin{cases} \Delta \neq \Delta', \\ \Delta, \Delta' \in \mathcal{S}_+^9, \\ P_\Delta^v(\text{IBS}_\ell) = P_{\Delta'}^v(\text{IBS}_\ell),\ \forall \ell \in \{1, ..., L\} \end{cases}.$$

The model is not always identifiable because in some cases one can build Δ' from Δ using $\Delta' = \Delta + \Delta_{KER}$, where $\Delta_{KER} = (0, x, 0, -x, 0, -x, -x, 2x, 0)^T$ is a vector that belongs to the kernel of matrix M_ℓ whatever ℓ. Note that even when the model is not identifiable, some linear combinations of $\Delta_1, ..., \Delta_9$ are estimable. In Csuros (2014) it is shown that some classical quantities such as the simple relatedness coefficient are always estimable.

2.1.3 Maximum Likelihood Estimation

When the model is identifiable, inference can be performed using Maximum Likelihood Estimation (MLE). In Milligan (2003), the author suggested to maximize the likelihood using a Hill Climbing algorithm. Alternatively, following McPeek and Sun (2000); Bink et al. (2008), the likelihood maximization may be performed using the EM algorithm (Dempster et al., 1977) (see Appendix A for an overview of the EM algorithm). Note that, in this case, the E step and the M step have an explicit (and simple) expression, since only the proportion parameters have to be estimated. One has

- E-step:

$$\tau_{\ell,j}^{(n)} = \frac{P(\text{IBS}_\ell = o_i | \text{IBD}_\ell = c_j)\Delta_j^{(n-1)}}{\sum_{k=1}^9 P(\text{IBS}_\ell = o_i | \text{IBD}_\ell = c_k)\Delta_k^{(n-1)}}$$

  where $\tau_{\ell,j}^{(n)}$ is the posterior probability for locus ℓ to have the j^{th} IBD configuration, computed at step n,

- M-step:

$$\Delta_j^{(n)} = \frac{1}{L}\sum_{\ell=1}^{L} \tau_{\ell,j}^{(n)}$$

This EM algorithm has been accelerated using the method presented by Varadhan and Roland (2008). The case where the model is not identifiable will be addressed in Section 2.2.1.

2.2 Inference of Relatedness for Hybrids

In plant genetics, it is of common practice to cross inbred lines - possibly belonging to several populations or heterotic groups - to obtain hybrids, that will constitute the basis of the association or genomic selection study to come (Bernardo, 1994; Technow et al., 2012). In such a configuration, extra information is available for each individual (hybrid), that may be highly valuable for the inference of relatedness. First, the crossing design provides an explicit information about the phase of the genotypic data. In Table 2.2, IBD configuration c_8 (respectively c_3, c_5 and c_7) can be decomposed into four (respectively two) different configurations $c^p_{11}, ..., c^p_{14}$ that can now be distinguished. IBS configurations can also be decomposed the same way (Table 2.1). Consequently there are now 15 distinguishable IBD configurations and 16 IBS configurations. Second, the crossing design indicates whether two hybrids share one or more common parental lines. Lastly, information about the membership of the parental lines to different populations may be available. Accounting for all this information in the inference process can greatly help the estimation of the relatedness coefficients. Additionally, one may expect this additional information to solve the identifiability issue identified by Csuros (2014), at least in some crossing configurations. These two points are investigated in the next 3 sections. The following proposition will be useful for the study of model identifiability for the mixture models to come:

Proposition 1. *Assume that the model can be written as*

$$P_\Delta(Y_\ell) = M_\ell \Delta, \ \forall \ell \in \{1, ..., L\} \ , \tag{2.3}$$

where Y_ℓ is the ℓ^{th} observed data, M_ℓ is a matrix of known coefficients that possibly depends on ℓ, and Δ is the vector of unknown parameters. Then Model (2.3) *is identifiable as soon as*

$$K = \bigcap_{\ell=1}^{L} Ker(M_\ell) = \{0\}$$

2.2.1 Phased Genotypic Data

We first consider the general case where one aims at estimating the relatedness between two hybrids H_1 and H_2 resulting respectively from crossings $L_1 \times L_2$ and $L_3 \times L_4$. We assume the four parental lines to be different, to come from a same population, and to be potentially related to each others.

Model There are 15 possible IBD status (noted c^p_j, $j = 1, ..., 15$, where subscript p stands for "phased") and 16 IBS status possible (noted o^p_i, $i = 1, ..., 16$). For locus ℓ the model is

$$P_\Delta(\text{IBS}_\ell = o^p_i) = \sum_{j=1}^{15} P(\text{IBS}_\ell = o^p_i | \text{IBD}_\ell = c^p_j)\Delta_j \ . \tag{2.4}$$

As in Section 2.1.1, one can write

$$P^v_\Delta(\text{IBS}_\ell) = M_\ell \Delta, \quad 1 \leq \ell \leq L \tag{2.5}$$

where M_ℓ is now a 16×15 matrix and Δ is a vector with 15 elements.

Identifiability One can state the following proposition:

Proposition 2. *The intersection of the kernels of matrices M_ℓ, $\ell = 1, ..., L$ is*

$$K = \left\{(0, -x, 0, 0, x, 0, 0, x, -y, x+y, y, -x-y, -x-y, y, 0)^T,\ x \in \mathbb{R},\ y \in \mathbb{R}\right\}.$$

Model (2.5) *is not identifiable if there exists x, y satisfying*

$$\begin{aligned}
\max(\Delta_2 - 1, -\Delta_5, -\Delta_8) \leq & x \leq \min(\Delta_2, 1 - \Delta_5, 1 - \Delta_8) \\
\max(\Delta_9 - 1, -\Delta_{11}, -\Delta_{14}) \leq & y \leq \min(\Delta_9, 1 - \Delta_{11}, 1 - \Delta_{14}) \\
\max(-\Delta_{10}, \Delta_{12} - 1, \Delta_{13} - 1) \leq & x + y \leq \min(1 - \Delta_{10}, \Delta_{12}, \Delta_{13})
\end{aligned}$$

Furthermore, linear combinations $C^T\Delta$ such that $C \in K^\perp$ are always estimable.

Proof. In what follows we skip the subscript ℓ corresponding to the locus index. For a given locus, we note p the allelic frequency of allele 1 and $q = 1 - p$. For a given vector x, the set of equations defined by $Mx = 0$ (with M the matrix of conditional probability as defined in section 2) is:

$$\left\{\begin{array}{ll}
q^4x_{15} + q^3(x_5 + x_8 + x_{11} + x_{12} + x_{13} + x_{14}) + q^2(x_2 + x_3 + x_4 + x_6 + x_7 + x_9 + x_{10}) + qx_1 = 0 & L_1 \\
q^3px_{15} + q^2p(x_5 + x_{11} + x_{13}) + qpx_3 = 0 & L_2 \\
q^3px_{15} + q^2p(x_5 + x_{12} + x_{14}) + qpx_4 = 0 & L_3 \\
q^2p^2x_{15} + qp^2x_5 + q^2px_8 + qpx_2 = 0 & L_4 \\
q^3px_{15} + q^2p(x_8 + x_{11} + x_{12}) + qpx_6 = 0 & L_5 \\
q^2p^2x_{15} + qp^2x_{11} + q^2px_{14} + qpx_9 = 0 & L_6 \\
q^2p^2x_{15} + qp^2x_{12} + q^2px_{13} + qpx_{10} = 0 & L_7 \\
qp^3x_{15} + qp^2(x_8 + x_{13} + x_{14}) + qpx_7 = 0 & L_8 \\
q^3px_{15} + q^2p(x_8 + x_{13} + x_{14}) + qpx_7 = 0 & L_9 \\
q^2p^2x_{15} + q^2px_{12} + qp^2x_{13} + qpx_{10} = 0 & L_{10} \\
q^2p^2x_{15} + q^2px_{11} + qp^2x_{14} + qpx_9 = 0 & L_{11} \\
qp^3x_{15} + qp^2(x_8 + x_{11} + x_{12}) + qpx_6 = 0 & L_{12} \\
q^2p^2x_{15} + q^2px_5 + qp^2x_8 + qpx_2 = 0 & L_{13} \\
qp^3x_{15} + qp^2(x_5 + x_{12} + x_{14}) + qpx_4 = 0 & L_{14} \\
qp^3x_{15} + qp^2(x_5 + x_{11} + x_{13}) + qpx_3 = 0 & L_{15} \\
p^4x_{15} + p^3(x_5 + x_8 + x_{11} + x_{12} + x_{13} + x_{14}) + p^2(x_2 + x_3 + x_4 + x_6 + x_7 + x_9 + x_{10}) + px_1 = 0 & L_{16}
\end{array}\right.$$

This is equivalent to:

$$\left\{\begin{array}{ll}
\text{No change} & L_1 \\
x_3 = -q^2x_{15} - q(x_5 + x_{11} + x_{13}) & L_2 \\
x_3 = -p^2x_{15} - p(x_5 + x_{11} + x_{13}) & L_{15} \\
x_4 = -q^2x_{15} - q(x_5 + x_{12} + x_{14}) & L_3 \\
x_4 = -p^2x_{15} - p(x_5 + x_{12} + x_{14}) & L_{14} \\
x_2 = -qpx_{15} - px_5 - qx_8 & L_4 \\
x_2 = -qpx_{15} - qx_5 - px_8 & L_{13} \\
x_6 = -q^2x_{15} - q(x_8 + x_{11} + x_{12}) & L_5 \\
x_6 = -p^2x_{15} - p(x_8 + x_{11} + x_{12}) & L_{12} \\
x_9 = -qpx_{15} - px_{11} - qx_{14} & L_6 \\
x_9 = -qpx_{15} - qx_{11} - px_{14} & L_{11} \\
x_{10} = -qpx_{15} - px_{12} - qx_{13} & L_7 \\
x_{10} = -qpx_{15} - qx_{12} - px_{13} & L_{10} \\
x_7 = -p^2x_{15} - p(x_8 + x_{13} + x_{14}) & L_8 \\
x_7 = -q^2x_{15} - q(x_8 + x_{13} + x_{14}) & L_9 \\
\sum_{i=1}^{15} x_i = 0 & \sum_{i=1}^{16} L_i
\end{array}\right.$$

After some maths:

$$\begin{cases} \text{No change} & L_{15} \\ x_{15} = -(x_5 + x_{11} + x_{13}) & L_2 - L_{15} \\ x_{15} = -(x_5 + x_{12} + x_{14}) & L_3 - L_{14} \\ x_5 = x_8 & L_4 - L_{13} \\ x_{15} = -(x_8 + x_{11} + x_{12}) & L_5 - L_{12} \\ x_{11} = x_{14} & L_6 - L_{11} \\ x_{12} = x_{13} & L_7 - L_{10} \\ x_{15} = -(x_8 + x_{13} + x_{14}) & L_8 - L_9 \\ x_2 = -qpx_{15} - px_5 - qx_8 & L_4 \\ x_9 = -qpx_{15} - px_{11} - qx_{14} & L_6 \\ x_{10} = -qpx_{15} - px_{12} - qx_{13} & L_7 \\ x_3 = -q^2x_{15} - q(x_5 + x_{11} + x_{13}) & L_2 \\ x_4 = -q^2x_{15} - q(x_5 + x_{12} + x_{14}) & L_3 \\ x_6 = -q^2x_{15} - q(x_8 + x_{11} + x_{12}) & L_5 \\ x_7 = -p^2x_{15} - p(x_8 + x_{13} + x_{14}) & L_8 \\ \sum_{i=1}^{15} x_i = 0 & \sum_{i=1}^{16} L_i \end{cases}$$

If we change what is known:

$$\begin{cases} \text{No change} & L_1 \\ x_{15} = -(x_5 + x_{11} + x_{12}) & \\ x_8 = x_5 & \\ x_{13} = x_{12} & \\ x_{14} = x_{11} & \\ x_2 = (qp-1)x_5 + qpx_{11} + qpx_{12} & L_4 \\ x_9 = qpx_5 + (qp-1)x_{11} + qpx_{12} & L_6 \\ x_{10} = qpx_5 + qpx_{11} + (qp-1)x_{12} & L_7 \\ x_3 = -qp(x_5 + x_{11} + x_{12}) & L_2 \\ x_4 = -qp(x_5 + x_{11} + x_{12}) & L_3 \\ x_6 = -qp(x_5 + x_{11} + x_{12}) & L_6 \\ x_7 = -qp(x_5 + x_{11} + x_{12}) & L_8 \\ x_1 = qp(x_5 + x_{11} + x_{12}) & \sum_{i=1}^{16} L_i \end{cases}$$

The kernel of matrix M is

$$Ker(M_\ell) = vect\left\{\begin{pmatrix} qp \\ qp-1 \\ -qp \\ -qp \\ 1 \\ -qp \\ -qp \\ 1 \\ qp \\ qp \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} qp \\ qp \\ -qp \\ -qp \\ 0 \\ -qp \\ -qp \\ 0 \\ qp-1 \\ qp \\ 1 \\ 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} qp \\ qp \\ -qp \\ -qp \\ 0 \\ -qp \\ -qp \\ 0 \\ qp \\ qp-1 \\ 0 \\ 1 \\ 1 \\ 0 \\ -1 \end{pmatrix}\right\}$$

Note that this kernel is obtained for a given marker, and depends on the allelic frequencies of this marker. To check wether the model is identifiable, the intersection of all kernels corresponding to the different markers has to be $\{0\}$. From the previous expression, it is easy to prove that each kernel contains the following space:

$$vect\left\{\begin{pmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ -1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ 1 \\ 1 \\ -1 \\ -1 \\ 1 \\ 0 \end{pmatrix}\right\}$$

which shows that the intersection of all kernels has dimension 2. □

Inference If Model (2.5) is not fully identifiable, the classical EM algorithm will converge to one of the possible (local) solutions. Consequently, the same algorithm can be applied whatever the identifiability status of Model (2.5). The only difference lies in the convergence criterion, that can be adapted to account only for the estimable part of Δ. More precisely, the by-default stopping criterion of the EM algorithm is $||\Delta^{(n)} - \Delta^{(n+1)}||_2 < \varepsilon$ with $\Delta^{(n)}$ the estimation of Δ at iteration n, and ε the required precision. Since Δ is not fully identifiable in the general case, the criterion may be

adapted as follows:

$$||\Pi_{K^\perp}(\Delta^{(n)} - \Delta^{(n+1)})||_2 \leq \varepsilon \ , \tag{2.6}$$

where $\Pi_{K^\perp}$ is the projection matrix on the orthogonal complement of K. Since $\Pi_{K^\perp}$ is a projection, criterion (2.6) is actually less stringent that the previous one, i.e. the number of iterations for (2.6) to be satisfied (and convergence achieved) should be smaller. In practice the improvement of the EM computational cost observed in our experiments was marginal (not shown).

2.2.2 Accounting for Common Parents

We now assume that the two hybrids share at least one common parental line. We note $L_1, ..., L_m$ the set of m parental lines involved in the crossing design, and $L_k \times L_{k'}$ the hybrid obtained by crossing lines L_k and $L_{k'}$. All lines are assumed to belong to a same population.

Identifiability The different configurations of two hybrids sharing at least one common parent are provided in Table 2.3, along with the relatedness coefficients that are potentially non null. All other coefficient are automatically set at 0 since their associated IBD configuration is impossible. For instance, consider hybrid $H_1 = L_1 \times L_2$ and hybrid $H_2 = L_1 \times L_3$. Then all IBD configurations stating the absence of IBD link between the first allele of each hybrid (such as Δ_2 for instance) are impossible. This leads to a fully identifiable model, as stated in the next proposition:

Proposition 3. *If there is at least one common parental line between the four parental lines crossed to obtain the two hybrids, then Model* (2.5) *is identifiable.*

The proof requires the study of each case detailed in Table 2.3 separately, but one can already notice that since some IBD coefficients are set to 0, some columns and rows of matrix M_ℓ can be discarded. One can show that the resulting reduced matrix is full rank, ensuring the identifiability of the model.

Proof. We consider the case where hybrids H_1 and H_2 share a common parent. Suppose that $H_1 = L_1 \times L_2$ and $H_2 = L_1 \times L_3$. Under this assumption the conditional probability matrix at a given locus boils down to:

$$M = \begin{array}{c} \begin{array}{ccccc} c_1^p & c_3^p & c_6^p & c_9^p & c_{11}^p \end{array} \\ \begin{pmatrix} q & q^2 & q^2 & q^2 & q^3 \\ 0 & qp & 0 & 0 & q^2p \\ 0 & 0 & qp & 0 & q^2p \\ 0 & 0 & 0 & qp & qp^2 \\ 0 & 0 & 0 & qp & q^2p \\ 0 & 0 & qp & 0 & qp^2 \\ 0 & qp & 0 & 0 & qp^2 \\ p & p^2 & p^2 & p^2 & p^3 \end{pmatrix} \end{array} \begin{array}{c} o_1^p \\ o_2^p \\ o_5^p \\ o_6^p \\ o_{11}^p \\ o_{12}^p \\ o_{15}^p \\ o_{16}^p \end{array}$$

with p the frequency of allele 1 at the locus and $q = 1-p$. If vector $X = (x_1, x_2, x_3, x_4, x_5)^T$ belongs to the kernel of matrix M, it satisfies:

$$\begin{cases} q^3x_5 + q^2x_4 + q^2x_2 + q^2x_3 + qx_1 = 0 & L_1 \\ q^2px_5 + qpx_2 = 0 & L_2 \\ q^2px_5 + qpx_3 = 0 & L_3 \\ qp^2x_5 + qpx_4 = 0 & L_4 \\ q^2px_5 + qpx_2 = 0 & L_5 \\ qp^2x_5 + qpx_3 = 0 & L_6 \\ qp^2x_5 + qpx_2 = 0 & L_7 \\ p^3x_5 + p^2x_4 + p^2x_2 + p^2x_3 + px_1 = 0 & L_8 \end{cases}$$

After some maths this is equivalent to:

$$\begin{cases} qx_5 + x_2 = 0 & \frac{L_2}{qp} \\ px_5 + x_2 = 0 & \frac{L_7}{qp} \\ qx_5 + x_3 = 0 & \frac{L_3}{qp} \\ px_5 + x_3 = 0 & \frac{L_6}{qp} \\ px_5 + x_4 = 0 & \frac{L_4}{qp} \\ qx_5 + x_4 = 0 & \frac{L_5}{qp} \\ q^3x_5 + q^2x_4 + q^2x_2 + q^2x_3 + qx_1 = 0 & L_1 \\ p^3x_5 + p^2x_4 + p^2x_2 + p^2x_3 + px_1 = 0 & L_8 \end{cases}$$

Assuming $p \neq \frac{1}{2}$, it follows:

$$\begin{cases} x_5 = 0 & \frac{L_2}{qp} - \frac{L_7}{qp} \\ x_5 = -2x_2 & \frac{L_2}{qp} + \frac{L_7}{qp} \\ x_5 = -2x_3 & \frac{L_3}{qp} + \frac{L_6}{qp} \\ x_5 = -2x_4 & \frac{L_4}{qp} + \frac{L_5}{qp} \\ q^3x_5 + q^2x_4 + q^2x_2 + q^2x_3 + qx_1 = 0 & L_1 \\ p^3x_5 + p^2x_4 + p^2x_2 + p^2x_3 + px_1 = 0 & L_8 \end{cases}$$

This equation leads to:

$$x_1 = x_2 = x_3 = x_4 = x_5 = 0$$

Then for all locus l satisfying $p \neq \frac{1}{2}$ we have $Ker(M) = \{0\}$, which is a sufficient condition for the model to be identifiable. The other cases (more than one common parent) can be dealt with using the same proof lines. □

Inference One would like the inference algorithm to ensure that the estimated values of null coefficient are actually equal to 0. This can be done straightforwardly via the initialization step of the EM algorithm. Indeed, according to the expression of the estimators appearing in the E and M steps provided in Section 2.1.3, one can observe that if a coefficient $\Delta_j^{(0)}$ is initialized at 0 then for all n, $\Delta_j^{(n)} = 0$. Consequently no additional constraint is required for the EM algorithm once the null coefficients are initialized at 0.

Parental crossing	Non null Coefficients
$L_1 \times L_1$, $L_1 \times L_1$	Δ_1
$L_1 \times L_1$, $L_1 \times L_2$	Δ_1, Δ_3
$L_1 \times L_1$, $L_2 \times L_3$	Δ_1, Δ_2, Δ_3, Δ_4, Δ_5
$L_1 \times L_2$, $L_1 \times L_2$	Δ_1, Δ_9
$L_1 \times L_2$, $L_1 \times L_3$	Δ_1, Δ_3, Δ_6, Δ_9, Δ_{11}

Table 2.3: Crossing design allowing for at least one common parental line, along with the list of positive relatedness coefficients. Here the Δs correspond to the set of 15 relatedness coefficients of Model (2.4) *(phased case).*

2.2.3 Accounting for Population Structure

The strategy that consists in crossing parental lines coming from different populations has been largely investigated, especially for species in which known heterotic groups exist (Bernardo, 1994). Regarding the estimation of relatedness, accounting for the population structure is crucial since allelic frequencies may vary from one population to another. Model (2.5) is unchanged, but matrix M_ℓ is different since the probabilities of the IBS configuration conditionally to the IBD configuration now depend on the allelic frequencies in the different populations the parental lines belong to.

Identifiability Since one can assume that individuals belonging to different populations do not share a common ancestor, some relatedness coefficients will be known to be null, as shown in table 2.4. As for the case of common parental lines, this will ensure the identifiability of Model (2.5):

Proposition 4. *If parental lines are derived from two populations or more then Model* (2.5) *is identifiable.*

Inference Similarly to the previous section, one only needs to ensure that the Δ coefficients that are known to be 0 thanks to the parental population structure information are set at 0. This can be done by initializing these parameters at 0. Note also that matrix M_ℓ has to be computed using the allelic frequencies of marker ℓ in the different populations.

2.2.4 Impact of Identifiability

In this part we will discuss about the identifiability problem for somes studies. The first one is a theoretical result regarding the inference of the double relatedness in one specific design. The second one is about the classification of individuals with respect to their relatedness.

Parental Population	Non null Coefficients
$P_1 \times P_2$, $P_3 \times P_4$	Δ_{15}
$P_1 \times P_1$, $P_2 \times P_3$	Δ_5, Δ_{15}
$P_1 \times P_2$, $P_1 \times P_3$	Δ_{11}, Δ_{15}
$P_1 \times P_2$, $P_1 \times P_2$	Δ_9, Δ_{11}, Δ_{14}, Δ_{15}
$P_1 \times P_1$, $P_1 \times P_2$	Δ_3, Δ_5, Δ_{11}, Δ_{13}, Δ_{15}

Table 2.4: Population structures allowing for at least two different populations for parental lines, along with the list of positive relatedness coefficients. Here the Δs correspond to the set of 15 relatedness coefficients of Model (2.4) *(phased case).*

Diallel crossing design To illustrate the potential consequences of non-identifiability on the analysis of plant genetic experiments, we consider here a diallel crossing design (Hayman, 1954; Lynch and Walsh, 1998). In this design, individuals are hybrids obtained by crossing lines coming from P populations. Usually only a (small) proportion of the possible crosses are performed, but the design is complete in the sense that for any pair of populations (p_1, p_2) - with possibly $p_1 = p_2$ - there exists at least one hybrid resulting from the crosse of two lines with respective population memberships p_1 and p_2. Such designs are commonly used in hybrid breeding (Bernardo, 1994).

In this context, we aim at estimating the double relatedness coefficient, defined as:

$$\Phi = \Delta_1 + \Delta_9 + \Delta_{10} \tag{2.7}$$

that can be understood as the probability that both alleles of the first individual are IBD to those of the second individual. This quantity is a key component of relatedness that naturally appears when computing the phenotypic covariance between two hybrids (see Crow and Kimura (1970) for instance). The two following situations may arise:

- if both hybrids are derived from four different lines belonging to a same population, then the double relatedness coefficient is non-estimable,
- if the two hybrids share a common parental line, or if at least two of the four parental lines come from different populations, then the double relatedness coefficient is estimable.

Note that the previous result does not definitely prevent the estimation of Φ in a diallel experiment: the result only states that Φ cannot be correctly inferred from *biallelic* marker data.

Class identification In some contexts the identification of different types of relatedness relationship may be at stake. While it is out of the scope of this chapter to propose a general strategy for the identification of relationship classes, we present here a very simple simulation study to illustrate the following intuitive idea: having access to the

full set of relatedness parameters rather that synthetic measures only should help to identify relationship classes, as long as the estimation task is possible, i.e. as long as the set of parameters is actually identifiable.
The different variants of Model (2.5) (but also of Model (2.2) corresponding to the unphased case) presented in the previous section have been implemented in the `Relatedness` R package (Laporte and Mary-Huard, 2017), and the method will be referred to as RelML furtherdown.

We simulated couples of hybrids corresponding to one of two possible relatedness distributions given in the following table:

	Δ_9	Δ_{11}	Δ_{14}	Δ_{15}	Other Δs
Class 1	0.275	0.3	0.2	0.225	0
Class 2	0.275	0.25	0.25	0.225	0

Genotypic data corresponding to 10000 SNPs were simulated for 50 couples of hybrids in each class.

One can observe that the two classes can only be distinguished based on coefficients Δ_{11} and Δ_{14}. Note that according to Proposition 2 the two set of relatedness parameters are fully identifiable. In the present context, the simple and double relatedness coefficients are:

$$\begin{aligned} K &= \frac{1}{2}\Delta_9 + \frac{1}{4}(\Delta_{11} + \Delta_{14}) = 0.2625 \\ \Phi &= \Delta_9 = 0.275 \end{aligned}$$

for any couple, whatever the class. Consequently, these two synthetic measures are non-informative for the classification task and any attempt to classify the couples according to these classical measures would yield a spurious classification. In comparison, we processed the genotypic data to obtain the estimated relatedness distribution parameters for each couple, then classified couples into two clusters using a K-means algorithm, based on the full set of estimated relatedness coefficients. Figure 2.1 (left) displays the boxplots for the two coefficient Δ_{11} and Δ_{14}, in each class. For a given coefficient, the two boxplots do not overlap meaning that the relevant information for clustering is correctly inferred. This is confirmed in Table 2.5 (left) where one can observe a perfect classification by the K-means algorithm.

The same analysis was rerun except that the two relatedness distributions were slightly modified as follows:

	Δ_9	Δ_{11}	Δ_{14}	Δ_{15}	Other Δs
Class 1	$0.275 - 4\epsilon$	$0.3 - 4\epsilon$	$0.2 - 4\epsilon$	$0.225 + \epsilon$	ϵ
Class 2	$0.275 - 4\epsilon$	$0.25 - 4\epsilon$	$0.25 - 4\epsilon$	$0.225 + \epsilon$	ϵ

Here we chose $\epsilon = 0.03$. While for a given coefficient Δ_j the gap $\|\Delta_j^1 - \Delta_j^2\|$ remains the same as in the previous analysis, the consequence of the modification is that coefficients Δ_{11} and Δ_{14} are not identifiable anymore, according to Proposition 2 in Section 2.2.1.

	Classified 1	Classified 2
Class 1	50	0
Class 2	0	50

Identifiable case

	Classified 1	Classified 2
Class 1	36	14
Class 2	33	17

Non-identifiable case

Table 2.5: Couple classification using the K-means clustering algorithm (true classes in lines, predicted clusters in columns).

As expected non-identifiability strongly impacts the inference, as illustrated in Figure 2.1 (right), and the classification performance is significantly degraded, see Table 2.5 (right).

The present simulation study is simplistic in many aspects: in practice the true number of classes (i.e. the different levels of relatedness between individuals) would be unknown, the class membership information would be spread across the different relatedness parameters, and even if non-identifiability may be expected to some level, it may be moderate in terms of impact on the precision of the estimators. The simulation shows that when identifiability is guaranteed working on the full set of relatedness parameters should help to retrieve the relatedness class membership between individuals, and that non-identifiability, if not accounted for, may blur information and impact the results of the statistical analysis based on the inferred relatedness coefficients.

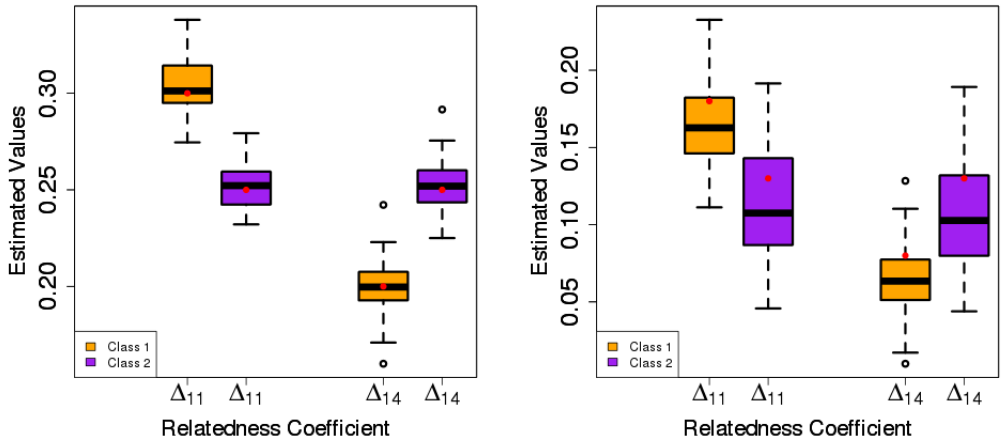


Figure 2.1: Left: Boxplots over 50 couples of hybrids of the estimates of Δ_{11} and Δ_{14} using RelML, obtained for the identifiable set of relatedness parameters. Red dots corresponding to the true values of these coefficients. Right: Same boxplots when the relatedness parameters are non-identifiable.

Merging markers When the marker data at hand are biallelic, several strategies have been proposed to "merge" adjacent markers, i.e. to synthesize multi-allelic marker data based on the initial biallelic information. These strategies may be useful but require a supplementary pre-processing step of the data, and their efficiency depends on the

relevant tuning of the parameters of the merging strategy (should the merging be performed within a sliding window along the genome ? If so what should be the size of the window ? Etc.). The results presented here provide guidelines on when (and why) such alternatives should be considered.

2.3 Comparison with Classical Inference Method

We compare ML estimators to the classical moment estimator proposed by Astle and Balding (2009), noted A&B in the following. Note that the A&B method only estimates the simple relatedness coefficient, defined for two individuals i and j as

$$K_{i,j} = \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_4 + \Delta_6 + \Delta_7 + \Delta_9 + \Delta_{10}) + \frac{1}{4}(\Delta_{11} + \Delta_{12} + \Delta_{13} + \Delta_{14}) \quad (2.8)$$

This coefficient can be interpreted as the probability that two alleles picked randomly, one in individual i and the other in individual j, are inherited from a common ancestor. The A&B estimator for this coefficient is

$$\widehat{K}_{i,j}^{A\&B} = \frac{1}{L}\sum_{\ell=1}^{L} \frac{(G_{i,\ell} - p_\ell)(G_{j,\ell} - p_\ell)}{p_\ell(1-p_\ell)}$$

where $G_{i,\ell}$ is the genotype of individual i at marker ℓ (coded with 0, 0.5 or 1 for a hybrid, and 0 or 1 for a line).

We will also consider the double relatedness coefficient, defined in Section 2.2.4. When dealing with two hybrids $H_1 = L_1 \times L_2$ and $H_2 = L_3 \times L_4$, the simple and double relatedness coefficients are obtained from the simple relatedness coefficients of the parental lines using the following formulas (Bernardo, 1994):

$$K_{H_1,H_2} = \frac{1}{4}(K_{L_1,L_3} + K_{L_1,L_4} + K_{L_2,L_3} + K_{L_2,L_4}) \quad (2.9a)$$

$$\Phi_{H_1,H_2} = \frac{1}{2}(K_{L_1,L_3}K_{L_2,L_4} + K_{L_1,L_4}K_{L_2,L_3}) \quad (2.9b)$$

As discussed in Section 2.3.1, note that Equality (2.9b) actually requires an extra assumption to be verified. Combining these formulas with the A&B estimator for $K_{L_u,L_v},\ u,v \in [1,4]$ yields estimators for K_{H_1,H_2} and Φ_{H_1,H_2}, hereafter referred to as the A&B estimators. We compare the A&B estimators with the ones provided by RelML (obtained by applying formulas (2.8) and (2.7) to the estimated relatedness coefficients) on both simulated and real data.

2.3.1 Simulated Data

We simulated genotypic data corresponding to (i) pairs of hybrids (settings 1 and 2) or (ii) a complete hybrid crossing design (setting 3). The first two settings will be used for the comparison of methods A&B and RelML regarding the precision when estimating the simple relatedness coefficient (setting 1) or the double relatedness coefficient (setting 2). Setting 3 will be used to investigate the impact of allelic frequency estimation on simple relatedness estimation. 10,000 biallelic markers were simulated for settings 1 and 2, and 8,000 for setting 3.

Simulation setting 1 In this setting, all hybrids are assumed to belong to a same population. Four different values for the simple relatedness coefficient are considered: $K = 0.05,\ 0.10,\ 0.15$ and 0.20. For each of these values, 20 configurations are simulated, each configuration corresponding to one relatedness distribution Δ and one allelic frequency vector F. Each relatedness distribution Δ is generated under the following constraints:

- $\Delta \in \mathcal{S}_+^{15}$,
- K is set at the chosen value,
- the double relatedness Φ is set at 0.05.

Each of the 10,000 allelic frequencies is generated using a uniform distribution over $[0.1, 0.9]$. For a given combination (Δ, F), 20 pairs of hybrids are generated. For a given pair, the IBD and IBS status of locus ℓ are simulated as follows. First, the IBD status is drawn in the multinomial distribution $\mathcal{M}(\Delta)$. Then, to obtain the IBS status (i.e. the genotype of the two hybrids), the 15×16 matrix of conditional probabilities $m_{ij}^{\ell} = P(\mathrm{IBS}_\ell = o_i | \mathrm{IBD}_\ell = c_j^p)$ is computed using allelic frequency F_ℓ. Assuming $\mathrm{IBD}_\ell = j$, the IBS status at locus ℓ is drawn in the multinomial distribution $\mathcal{M}(m_{1j}^{\ell}, ..., m_{15j}^{\ell})$. The IBS status directly provides the genotypic data. To summarize, the data at hand consist in $4 \times 20 \times 20 = 1,600$ pairs of hybrids, each genotyped at 10,000 markers.

Simulation setting 2 In this setting the generated data correspond to a factorial panel, i.e. each hybrid is assumed to be derived from a cross between two lines belonging to two different populations. This ensures the identifiability of the double relatedness coefficient, and also fixes $\Delta_i = 0$ for $i \in \{1, 2, 3, 4, 5, 6, 7, 8, 10, 12, 13\}$. Four different values of double relatedness coefficient are considered: $\Phi = 0.01,\ 0.05,\ 0.10$ and 0.15. Importantly, an extra condition can be imposed:

$$\Delta_9 = (\Delta_9 + \Delta_{11}) \times (\Delta_9 + \Delta_{14}) \ . \tag{2.10}$$

Condition (2.10) states that the double relatedness coefficient is equal to the product of the simple relatedness coefficient computed for each couple of parental lines belonging to the same population. While there is no guarantee that this condition is satisfied in real experiments, it is implicitly assumed to be satisfied in formula (2.9b). The rest of the simulation is similar to the previous setting (with one allelic frequency vector per population). To sum up, the simulation of the relatedness distribution Δ is performed with the constraints

- $\Delta \in \mathcal{S}_+^{15}$,
- Φ is set at the chosen value,
- the simple relatedness K is set at 0.18,
- $\Delta_i = 0$ for $i \in \{1, 2, 3, 4, 5, 6, 7, 8, 10, 12, 13\}$,
- with or without satisfying condition (2.10).

We experimentally observed in this setting that, when condition (2.10) is not satisfied, the difference $\Delta_9 - (\Delta_9 + \Delta_{11}) \times (\Delta_9 + \Delta_{14})$ decreases with respect to Φ (see Figure 2.3, center). Also note that in settings 1 and 2 the considered values for K and Φ were selected based on the value range observed for these two quantities in the dataset presented in Section 2.3.2.

Simulation setting 3 In this setting, a complete hybrid crossing design is simulated as follow. First, a set of allelic frequencies is generated for the 8,000 markers. These frequencies are drawn independently in the [0.1,0.9] uniform distribution. These allelic frequencies characterize the founder population. Based on these founder allelic frequencies, 5 founder lines are independently drawn. A family is then derived from each founder line: each family consists of 20 lines independently drawn from each others such that their expected relatedness level to the founder line is fixed, and differs from a family to another. The relatedness levels between a line and its associated founder are $\sqrt{0.1}$, $\sqrt{0.3}$, $\sqrt{0.5}$, $\sqrt{0.7}$ and $\sqrt{0.9}$ for the 5 families respectively, resulting in relatedness levels of 0.1, 0.3, 0.5, 0.7 and 0.9 between lines within a family. Families A and B - corresponding to those with relatedness level 0.3 and 0.7, respectively - are then selected, and only the crossings between distinct lines are made. This results in a set of $20 * 19/2 = 190$ hybrids exhibiting complex patterns of relatedness.

Comparison with A&B Figure 2.2 displays the boxplots of the raw differences between the true and estimated values of K. One can observe that the two estimators perform equally well for the estimation of the simple relatedness coefficient. Figure 2.3 displays the boxplots of the raw differences between the true and estimated values of Φ. One can observe that the two estimators perform equally for the estimation of the double relatedness coefficient if condition (2.10) is verified. When the condition is not verified the A&B procedure yields highly biased estimates whereas the RelML estimators are quite robust in this context. Note that the apparent bias reduction of the A&B procedure when Φ increases is mostly due to the fact that the difference $\Delta_9 - (\Delta_9 + \Delta_{11}) \times (\Delta_9 + \Delta_{14})$ decreases with respect to Φ in this setting, as illustred in Figure 2.3 (center).

Impact of allelic frequency estimation We consider the data generated using simulation setting 3. The goal is to estimate coefficient K for each pair of hybrids. The estimation is performed with RelML, using either the true allelic frequencies or the frequencies estimated from the 100 parental lines. Figure 2.4 displays the differences between the true and estimated K in these two situations. One observes a slight overestimation of K when the theoretical frequencies are used. This is mainly explained by the fact that neither the crossing design nor the population structure was accounted for RelML, therefore simple relatedness coefficients for hybrids belonging to different families (i.e. a hybrid AA and a hybrid BB) are not estimated at exactly 0. One can also observe an underestimation of K when the allelic frequencies are estimated using the parental lines. This underestimation can also be observed for the A&B procedure, and can be explained as follows. Whatever the method, the estimation of relatedness is based on the observation that, over all loci, individuals share more alleles than expected if they were not related. This expectation is related to the genetic diversity $H_e \propto \sum_\ell p_\ell(1-p_\ell)$.

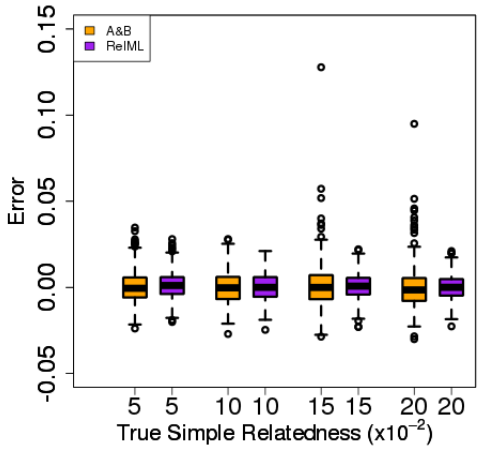


Figure 2.2: Boxplot of estimation errors for the simple relatedness coefficient, computed on 400 couples. Numerical results are available in the Appendix B.

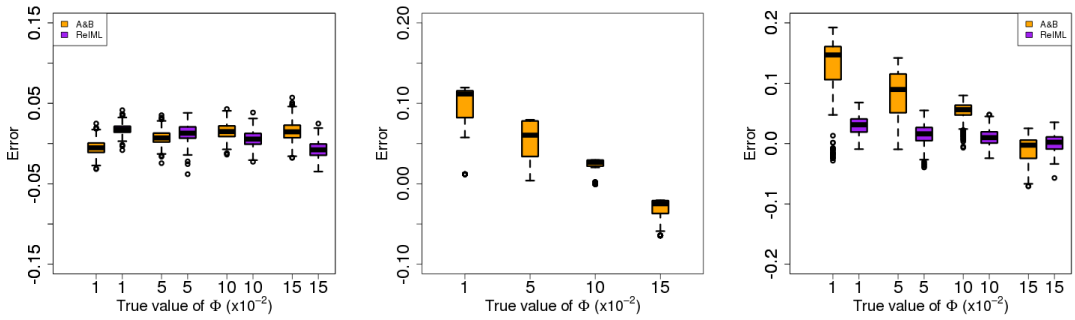


Figure 2.3: Left: Boxplot of estimation errors for coefficient Φ where the condition $\Delta_9 = (\Delta_9 + \Delta_{11}) \times (\Delta_9 + \Delta_{14})$ is verified, computed on 400 couples. Center: Boxplot of differences between the true value of Δ_9 and $\Delta_9 = (\Delta_9 + \Delta_{11}) \times (\Delta_9 + \Delta_{14})$ when the condition is not satisfied, represented as a function of the true value of coefficient Φ. Right: Boxplot of estimation errors for coefficient Φ where the previous condition is not verified, computed on 400 couples. Not that the scale of the y-axis differs between the 3 figures. Numerical results are available in the Appendix B.

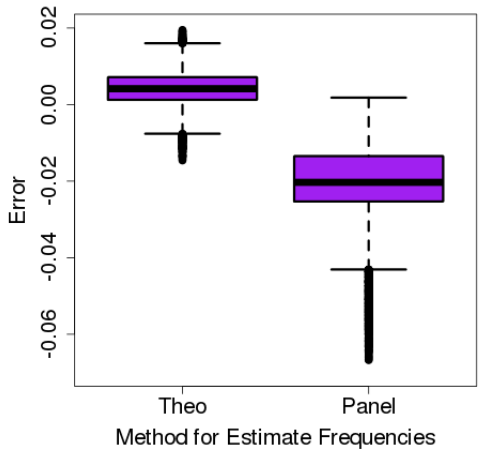


Figure 2.4: Difference between true and RelML-estimated values of the simple relatedness coefficient, using the true (left) or panel-estimated (right) allelic frequencies.

When the allelic frequencies are estimated on a sample of closely related individuals, H_e will be underestimated, and consequently the probability that two unrelated individuals share a same allele will be overestimated. Figure 2.5 illustrates the impact of allelic frequency estimation on the estimated diversity. This will result in an underestimation of K for related individuals. We also observed that removing markers having a low MAF in the reference population reduces this underestimation. Figure 2.6 illustrates the impact of removing markers having a (supposedly known) low MAF in the population. Here markers are filtered based on their MAF with a threshold t at 0 (no filter), 0.2 and 0.4. One can see that the bias significantly decreases with respect to t.

The choice of the reference panel used for the estimation of allelic frequencies is therefore crucial, and a poor estimation may lower the performance of RelML - and probably of most relatedness estimation methods (see Bink et al. (2008) for a comprehensive study on this topic).

2.3.2 Real Data

We present the application of the two estimation procedures to a panel of 346 maize hybrids obtained by crossing lines according to a diallel design. More precisely, 120 (respectively 126) lines were collected in the Flint (respectively Dent) population, and genotyped for 49,574 SNPs. Hybrids were derived from crossings with the following distribution: 90 Flint×Flint hybrids, 92 Dent×Dent hybrids, 76 Flint×Dent hybrids and 88 Dent×Flint hybrids.
In this multi-population design there are 3 possible ways to perform an A&B estimation. All strategies are based on Equation (2.9), but the ways to obtain the simple relatedness coefficient between parental lines are different:

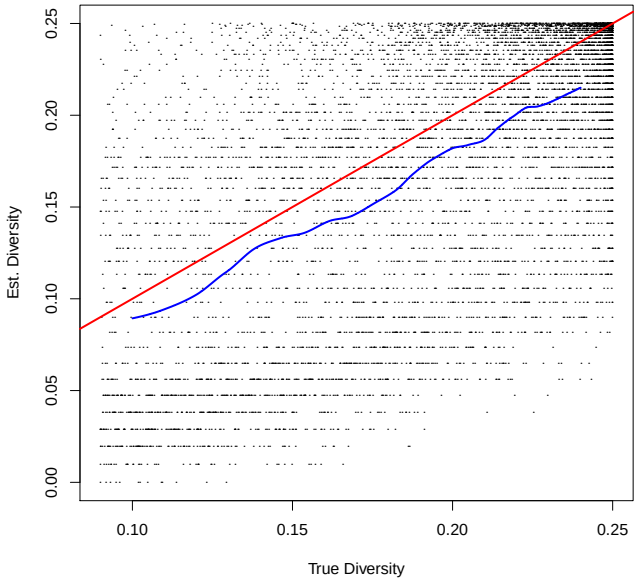


Figure 2.5: Estimated versus True diversity, computed on data simulated using setting 3. Each point corresponds to a locus. The red (respectively blue) curve corresponds to the first bisector (respectively a loess estimation of the relationship between estimated and true diversity).

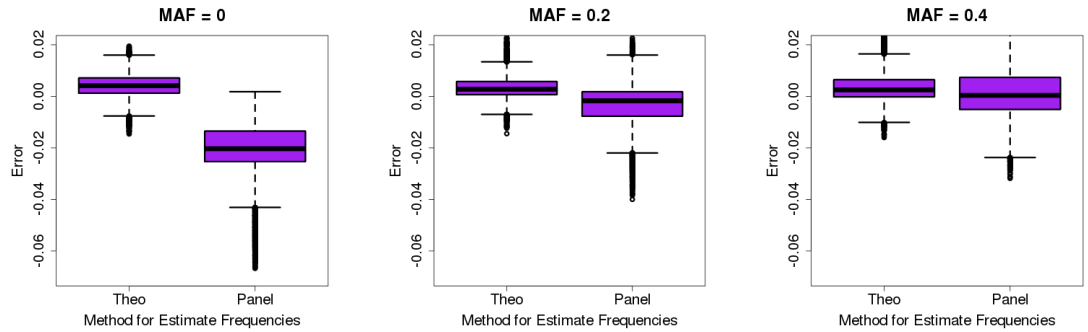


Figure 2.6: Difference between true and RelML-estimated values of the simple relatedness coefficient, using the true (left) or panel-estimated (right) allelic frequencies with a filter on the minor allele frequency (MAF).

- A&B V1: The simple relatedness coefficient is inferred between lines using:

$$K_{L_1^k,L_2^{k'}} = \frac{1}{L}\sum_{l=1}^{L} \frac{(G_{l,L_1^k} - p_l^k)(G_{l,L_2^{k'}} - p_l^{k'})}{\sqrt{p_l^k(1-p_l^k) \times p_l^{k'}(1-p_l^{k'})}} \mathbf{1}_{\{k=k'\}} \ ,$$

  where G_{l,L_1^k} is the genotype of line L_1 in population k, and p_l^k the allelic frequency for marker ℓ in population k. Note that the simple relatedness coefficient between two lines belonging to two different populations is set to 0.

- A&B V2: The simple relatedness coefficient between two lines is inferred using:

$$K_{L_1^k,L_2^{k'}} = \frac{1}{L}\sum_{l=1}^{L} \frac{(G_{l,L_1^k} - p_l^k)(G_{l,L_2^{k'}} - p_l^{k'})}{\sqrt{p_l^k(1-p_l^k) \times p_l^{k'}(1-p_l^{k'})}} \ .$$

- A&B V3: The simple relatedness coefficient between two lines is inferred using a single allelic frequency vector obtained as the weighted mean of the allelic frequency vector of each population.

$$K_{L_1,L_2} = \frac{1}{L}\sum_{l=1}^{L} \frac{(G_{l,L_1} - p_l)(G_{l,L_2} - p_l)}{\sqrt{p_l(1-p_l) \times p_l(1-p_l)}}$$

Applied to the present dataset, the results with the procedure A&B V2 and those with the procedure A&B V3 are quite similar.

Consequently we omitted procedure A&B V3, and only comparisons between RelML, A&B V1 and A&B V2 are presented here. The relatedness distribution has been inferred for each couple of hybrids, requiring a computational time of 13 hours (using 8 cores). Note that this time can be significantly reduced using less markers, possibly selected based on mapping information, see Appendix C and Appendix D for details. Here we focus on the estimation of the simple relatedness coefficient.

Figure 2.7 (left) displays the boxplot of the estimated simple relatedness coefficient K obtained with the A&B V1 or A&B V2 (orange) or the RelML method (purple), with respect to the number of common parent(s) between the two hybrids. The same data are represented in the right panel of Figure 2.7, where each point represents a pair of hybrids, represented by its estimated K value using A&B (x-axis) and RelML (y-axis). Colors correspond to the number of common parent(s). Comparing A&B V1 and RelML (on top panel of Figure 2.7), one can first observe that the two sets of estimation are quite concordant. The major difference is the fact that the A&B estimator yields estimated values that can be lower than one could expect from theory. It is already well-known that A&B can produce negative kinship values. In the case where one or two parents are shared by the hybrids, the simple relatedness coefficient should be equal or greater than 0.25 (1 common parent) or 0.5 (2 common parents). As observed here, the smallest estimations obtained with the A&B estimator are 0.15 and 0.39, respectively. In comparison, the RelML estimates are always equal or larger than the expected lower bound. When the number of common parent is 0 (i.e. the lower bound for K is also 0), it comes with no surprise since by construction $\widehat{\Delta} \in \mathcal{S}_+^{15}$. It turns out that this simplex constraint, associated with the fact that RelML accounts

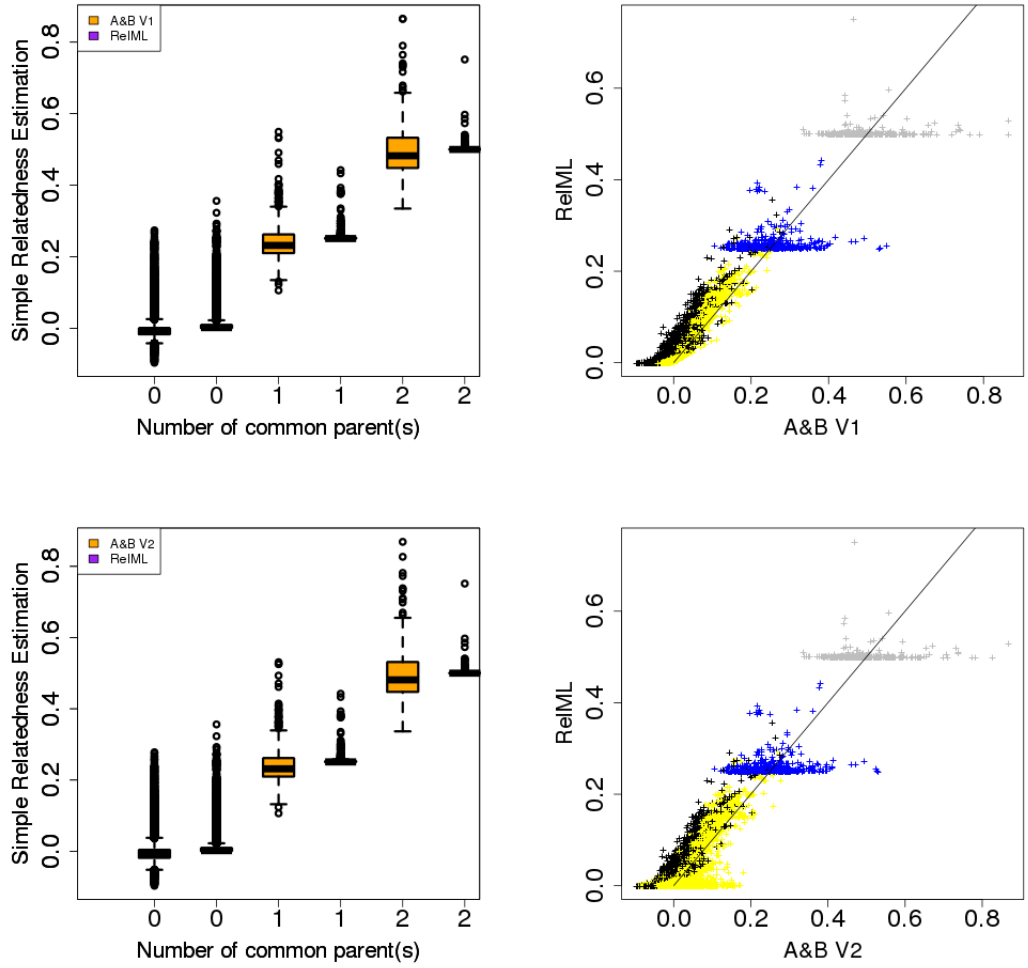


Figure 2.7: Left: Boxplot of simple relatedness coefficient with respect to the number of common parent(s). Right: Colors correspond to different parental configurations (black: no common parent between hybrids, all parents come from a same population, yellow: no common parent, parents come from two different populations, blue: one common parent, grey: two common parents). Top: Comparison between RelML and A&B V1. Bottom: Comparison between RelML and A&B V2.

for the crossing design, enforces the estimated simple relatedness coefficient to be equal or higher than its theoretical lower bound in the case where one or two parents are shared by the hybrids. Indeed, if two hybrids have a common parent then only $\widehat{\Delta}_i$ for $i \in I = \{1, 3, 4, 6, 7, 9, 10, 11, 12, 13, 14\}$ will be non zero (thanks to the crossing design), and will sum to 1 (thanks to the simplex constraint). Combining this with the formula of K given in Equation (2.8) leads to $\widehat{K} \geq \frac{1}{4} \sum_{i \in I} \widehat{\Delta}_i = \frac{1}{4}$. A similar reasoning shows that $\widehat{K} \geq \frac{1}{2}$ when the hybrids share two common parents.

Comparing RelML with A&B V2 (Figure 2.7), one can observe a higher difference than with A&B V1. This difference comes from hybrids without common parent but with parental lines derived from two populations. The values of K obtained with A&B V2 are higher than the ones obtained with RelML or A&B V1, because the simple relatedness coefficient between lines belonging to different populations (hence known to be 0) is not inferred at 0 in A&B V2. This illustrates that one should care about which version of A&B to use according to different situations. In comparison, RelML provides a unique framework adapted to all situations.

2.4 Conclusions

The `Relatedness` R package presented in this article provides MLE for the complete set of relatedness coefficients from SNP data. While the results have been illustrated on "hybrid data", i.e. assuming that the individuals are hybrids derived from crossing between lines, `Relatedness` also handles the classical case where the genotypic information corresponds to unphased data collected on individuals in a single population. We illustrated that the non-indentifiability issue raised by Csuros (2014) in the case of unphased data also occurs in phased data, but may be solved as soon as extra information about the crossing design is taken into account in the inference step. The specific case of crossing design is particularly relevant in plant genetics, but the principle of accounting for population membership could be extended to animal breeding where intra and inter population mating also occur. More generally, it illustrates how accounting for the full information available may be of importance to avoid statistical identifiability issues. It also illustrates how a naive inference of the relatedness distribution may lead to misleading conclusions when applied to experimental design such as diallel designs, where a same relatedness component (the double relatedness coefficient) may be identifiable or not, depending on the couple of individuals considered.

Infering relatedness is highly related to some classical tasks in quantitative genetics such as QTL detection or genomic prediction. Regarding the QTL detection task, association studies nowadays routinely include a kinship matrix K (with K_{ij} the simple relatedness coefficient between individuals i and j) to account for background additive effects. Several attempts have been proposed to generalize this modeling by including a second matrix Φ (with Φ_{ij} the double relatedness coefficient between individuals i and j) to account for possible background dominance effects (Technow et al., 2012). The

performance of such strategies may be impacted by the estimation precision of the relatedness parameters. Similarly, matrix Φ could also be accounted for in genomic selection to improve the prediction accuracy of the GBlup procedure (Technow et al., 2012) in presence of dominance. Before studying marker detection, we had been interested in performance of algorithm to infer mixed model parameters.

2.5 Appendix

Appendix A: An overview of EM algorithm

The EM algorithm is largely used to obtain Maximum Likelihood Estimator (MLE) of unknown parameters in mixture models. Quoting X a variable with a density function $f(x, \theta)$ where θ is an unknown parameter. The log-likelihood of the data $X = (X_1, ..., X_n)$, where X_i are independent and identically distributed with the density function $f(x, \theta)$, is:

$$\mathcal{L}(\theta; X) = \sum_{i=1}^{n} \log[f(X_i, \theta)] \tag{2.11}$$

The goal of MLE is to maximize the function defined in (2.11) with respect to θ. The EM algorithm is based on the use of hidden variables $Z = (Z_1, ..., Z_n)$. Quoting $h(Z_i|(X_i, \theta))$ the density function of these variables knowing X_i and θ. Then the log-likelihood of the couple (X, Z) is:

$$\mathcal{L}(\theta; (X, Z)) = \sum_{i=1}^{n} \log[h(Z_i|(X_i, \theta))] + \log[f(X_i, \theta)] \tag{2.12}$$

Combining equations (2.11) and (2.12), and using the expectation with respect to variables Z, one obtains the following equation:

$$E\big[\mathcal{L}(\theta; X)|\theta^{(t)}\big] = E\big[\mathcal{L}(\theta; (X, Z))|\theta^{(t)}\big] - E\big[\sum_{i=1}^{n} \log[h(Z_i|(X_i, \theta))]|\theta^{(t)}\big]$$

$$\mathcal{L}(\theta; X) = Q(\theta|\theta^{(t)}) - H(\theta|\theta^{(t)}) \tag{2.13}$$

Maximize $Q(\theta|\theta^{(t)})$ in (2.13) with respect to θ increases the log-likelihood of the complete data. For all θ, one has:

$$\mathcal{L}(\theta; X) - \mathcal{L}(\theta^{(t)}; X) = Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) - (H(\theta|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)})) \tag{2.14}$$

And the Jensen Inequalities ensure that:

$$\forall \theta,\ H(\theta|\theta^{(t)}) \leq H(\theta^{(t)}|\theta^{(t)})$$

So equation (2.14) can be written as an inequality:

$$\mathcal{L}(\theta; X) - \mathcal{L}(\theta^{(t)}; X) \geq Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) \tag{2.15}$$

The inequality (2.15) prooves the improvement of the log-likelihood if one use the iteration update $\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)})$.
The EM algorithm is composed of two steps:

- Expectation (E-step): Calculate $Q(\theta|\theta^{(t)})$
- Maximization (M-step): Calculate $\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)})$

In the model (2.3) the observed vector X corresponds to the vector of IBS configurations through all locus, parameter θ corresponds to the vector of unknown distributions Δ and vector Z corresponds to the vector of IBD configurations through all locus.

Appendix B: Numeric values for the A&B and RelML estimators

The numerical results displayed in Suppl. Table 2.6 (respectively Suppl. Table 2.7 and 2.8) correspond to the ones presented in Figure 2.2 (respectively 2.3), Section 2.3.1 of the chapter 2. The displayed values are the mean, bias, standard deviation and MSE of the A&B and RelML estimates.

K	A&B			
	Mean Est	Bias	Sd	MSE
0.05	0.05	4.21×10^{-4}	9.51×10^{-3}	9.03×10^{-5}
0.10	0.10	-2.84×10^{-4}	9.23×10^{-3}	8.51×10^{-5}
0.15	0.15	15.01×10^{-4}	15.71×10^{-3}	24.84×10^{-5}
0.20	0.20	-0.89×10^{-4}	12.35×10^{-3}	15.23×10^{-5}

K	RelML			
	Mean Est	Bias	Sd	MSE
0.05	0.05	15.68×10^{-4}	7.98×10^{-3}	6.59×10^{-5}
0.10	0.10	4.01×10^{-4}	8.05×10^{-3}	6.48×10^{-5}
0.15	0.15	4.02×10^{-4}	7.63×10^{-3}	5.82×10^{-5}
0.20	0.20	1.47×10^{-4}	7.20×10^{-3}	5.18×10^{-5}

Suppl. Table 2.6: Mean, bias, standard deviation and MSE of the A&B and RelML estimates for the simple relatedness coefficient.

Φ	A&B			
	Mean Est	Bias	Sd	MSE
0.01	0.01	-4.95×10^{-3}	8.89×10^{-3}	1.03×10^{-4}
0.05	0.06	7.37×10^{-3}	8.41×10^{-3}	1.25×10^{-4}
0.10	0.11	15.42×10^{-3}	9.85×10^{-3}	3.35×10^{-4}
0.15	0.16	14.89×10^{-3}	12.88×10^{-3}	3.87×10^{-4}

Φ	RelML			
	Mean Est	Bias	Sd	MSE
0.01	0.03	17.98×10^{-3}	6.18×10^{-3}	3.61×10^{-4}
0.05	0.06	13.12×10^{-3}	10.43×10^{-3}	2.80×10^{-4}
0.10	0.11	5.81×10^{-3}	10.42×10^{-3}	1.42×10^{-4}
0.15	0.14	-7.89×10^{-3}	10.38×10^{-3}	1.70×10^{-4}

Suppl. Table 2.7: Mean, bias, standard deviation and MSE of the A&B and RelML estimates for the double relatedness coefficient, when condition $\Delta_9 = (\Delta_4 + \Delta_9)(\Delta_7 + \Delta_9)$ is satisfied.

Appendix C: Computational time

While obtaining the relatedness coefficients for a couple of individuals is quite fast, in practice one usually needs to compute the set of coefficients for all couples of individuals in a panel. The computational burden associated to this task is linear in the number of markers and quadratic in the number of individuals. To speed up the

Φ	A&B			
	Mean Est	Bias	Sd	MSE
0.01	0.14	0.13	45.73×10^{-3}	18.70×10^{-3}
0.05	0.13	0.08	39.62×10^{-3}	8.15×10^{-3}
0.10	0.15	0.05	16.55×10^{-3}	3.04×10^{-3}
0.15	0.14	-0.01	23.36×10^{-3}	0.68×10^{-3}

Φ	RelML			
	Mean Est	Bias	Sd	MSE
0.01	0.04	29.38×10^{-3}	15.61×10^{-3}	1.11×10^{-3}
0.05	0.06	14.47×10^{-3}	16.57×10^{-3}	0.48×10^{-3}
0.10	0.11	10.28×10^{-3}	13.65×10^{-3}	0.29×10^{-3}
0.15	0.15	0.77×10^{-3}	14.76×10^{-3}	0.22×10^{-3}

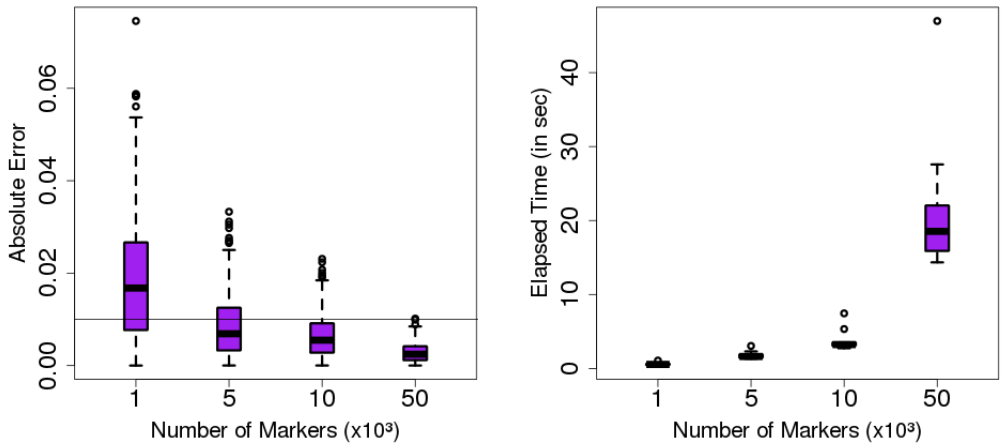
Suppl. Table 2.8: Mean, bias, standard deviation and MSE of the A&B and RelML estimates for the double relatedness coefficient, when condition $\Delta_9 = (\Delta_4 + \Delta_9)(\Delta_7 + \Delta_9)$ *is not satisfied.*

computation the EM algorithm has been coded in C with the acceleration proposed by Varadhan and Roland (2008) and parallelized in the `Relatedness` package. Another way to reduce the computational time is to reduce the number of markers taken into account for the computation. The number of markers can be selected according to the required precision. Top Figures in Suppl. Figure 2.8 displays the error $|K_{H_1,H_2} - \widehat{K}^R_{H_1,H_2}|$ over 50 simulations as a function of the number of markers, for the estimation of the simple relatedness coefficients. One can observe that only 10,000 markers are required to guarantee an error range as low as 0.01. Applying RelML to all pairs of hybrids described by 10,000 biallelic markers in a panel of size 100 (i.e. 5050 couples) results in a computational time of 7 minutes (on a personal laptop with 8 cores). See also Web Appendix D for a study about marker selection on real data.

Appendix D: Impact of Linkage Disequilibrium and SNP selection on relatedness coefficient estimation

In Appendix C the problem of reducing the computational cost by reducing the number of markers was considered. The numerical study showed that marker selection has minor impact on the precision of the simple relatedness coefficient estimator. However, this conclusion was drawn on the basis of simulated data, where markers were generated independently from each others. It is not clear how these results can be extended to real data experiments where the level of Linkage Disequilibrium (LD) may be high. We present here some additional results obtained on the maize dataset used in Section 2.3.2 of the manuscript. Note that for an equivalent panel, Rincent et al. (2014b) recently evaluated the actual number of independent markers to be ≈ 4000.

To study the impact on inference of i) LD and ii) working with a reduced number of markers, we applied RelML to different subsets of selected markers. Two different strategies were considered for the selection. The first strategy consisted in selecting markers such that the minimum genetic distance between two consecutive markers along the chromosome was higher than d_{min}. Different values of d_{min} were considered, corresponding to the division of the total genetic length of the map by 1000, 5000, 10000



Suppl. Figure 2.8: Left: Boxplot of the absolute error for the estimation of simple relatedness coefficients with respect to the number of marker, over 500 simulated couples. The solid line indicates a threshold of 0.01 for the absolute error. Right: Boxplot of computational time for 50 couples, averaged over 10 trials. Note that the scale of this axis is not linear.

	Minimal Genetic Distance (cM)			
quantile	2.29	0.46	0.23	0.11
75%	7.2×10^{-3}	3.8×10^{-3}	3.2×10^{-3}	2.9×10^{-3}
90%	13.7×10^{-3}	7.6×10^{-3}	6.7×10^{-3}	5.9×10^{-3}

Suppl. Table 2.9: Quantiles of the absolute error between simple relatedness coefficients estimated from the whole set of markers or from a set markers selected based on their genetic distances.

and 20000, respectively. For each value of d_{min}, 10 marker selections were performed, corresponding to different "starting" markers along the chromosome. On average, the number of selected markers corresponding to each value of d_{min} were 897, 3370, 5393 and 7938, respectively. The second strategy consisted in selecting a same number of markers randomly. For each subset, the absolute errors between simple coefficients computed on the whole set and on the subset of markers were recorded.

Suppl. Table 2.9 displays the quantiles of the absolute error for the first strategy (subsetting wih a minimal genetic distance). One can observe that for most couples of hybrids the absolute difference is lower than 0.01. One can conclude that accounting for linkage disequilibrium has limited impact on the estimation of relatedness using RelML. The same results are displayed in Suppl. Table 2.10 for the second strategy (subsetting

	Number of markers			
quantile	897	3370	5393	7938
75%	6.6×10^{-3}	3.0×10^{-3}	2.2×10^{-3}	1.8×10^{-3}
90%	12.6×10^{-3}	5.8×10^{-3}	4.4×10^{-3}	3.4×10^{-3}

Suppl. Table 2.10: Quantiles of the absolute error between simple relatedness coefficients estimated from the whole set of markers or from a set of randomly selected markers.

at random). One concludes that using ≈ 8000 markers instead of 46474 (i.e. dividing the number of used markers by 5.8) does not substantially affect the estimation of simple relatedness. Note that in a similar study (Bink et al. (2008)) the authors reached a similar conclusion.

In both tables, the difference between the simple relatedness coefficient inferred with (i) all the markers and (ii) a reduced number of markers increases as the number of markers decreases. This difference appears larger with the markers selected according to the genetic map, especially when short intervals are considered. Still, one can note that with this method the "error" decreases as the inverse of the square-root of the number of markers, which is expected with independent markers. This expectation of variation is find with the simulated data too.

Chapter 3

Evaluating the performance of different mixed model inference procedures in the context of statistical genetics

Since their formal introduction in the early 50's (Henderson, 1953; Scheffe, 1956), mixed models have become an indispensable tool of modern statistics. It has been successfully applied in many application fields (Gibbons and Hedeker, 2000) to model data with multiple sources of (biological or technical) variations. Starting with Griffing (1956) and Henderson (1973), mixed models have been a flavored methodology in quantitative genetics, and are still widely used in the context of Genome-Wide Association Studies (GWAS) and Genomic Selection (GS).
With the development of high throughout technologies, a special care has been dedicated to the development of efficient algorithmic procedures for the inference of mixed models (Kang et al., 2008; Lippert et al., 2011; Zhou and Stephens, 2012; Perdry and Dandine-Roulland, 2017). This is illustrated by the availability of many tools/softwares that either perform inference in a mixed model including many (fixed and random) effects on large datasets, or alternatively that efficiently fit hundreds of thousands of mixed models with a limited number of variance components. Due to the numerous optimization algorithms available and the many technical shortcuts involved in each of them, it may become difficult for the user to identify which algorithm would lead to the most efficient inference on a particular dataset.

The goal of this chapter is to compare different algorithms used in plant breeding GWAS and GS studies. In the context of plant breeding panels are often of moderate size, including a few hundreds/thousands of individuals only due to experimental and cost constraints. At the same time, the variance component mixed models used for the statistical analysis may be more complex that the ones used in human genetics, since the modeling should account for the specificities of the experimental design (e.g. when a multi-site experiment involving the same varieties in different environments has been

carried out) or of the crossing design (e.g. when considering hybrids obtained from parental lines belonging to different populations). One then aims at choosing the algorithmic procedure that is best suited to cope with these specific constraints. In the present paper we focus on the set of algorithmic procedures that provide the ReML estimates of the fixed effects and the variance components. This set includes procedures that directly optimize the restricted log-likelihood over a grid of values (available when the number of components is equal to 2), a first order method based on the Min-Max (MM) algorithm described by Hunter and Lange (2004) and applied to mixed models by Zhou et al. (2015), and a second order (Newton like) method called the Average Information (AI) algorithm described in Johnson and Thompson (1995). Note that this last algorithm has been widely used in animal genetics and is the core procedure of the ASReml software for mixed model inference (Gilmour et al., 2009). We first provide an algorithmic overview of the different methods, highlighting the different technical tricks that lead to their efficient implementation. We then present a benchmark analysis where three popular R packages (gaston, FaSTLMM and ASReml) and the more recently proposed MM procedures are compared on different scenarios. This empirical study reveals that while the differences in terms of precision may be marginal between procedures, the differences in terms of computational efficiency may be huge. Moreover, one can easily exhibit configurations where a given algorithm will perform poorly compared with its competitors. The observed performances are related to the respective algorithmic properties of the different procedures in order to provide some general guidelines to users. Importantly, we also investigated the impact that the algorithmic procedure may have on the resulting list of detected QTLs. The impact may be important, especially in a context where the list of detected QTLs is usually small due to the lack of power inherent to the moderate size of the available panels.

The chapter is organized as follows: Section 3.1 introduces the statistical framework of variance component mixed models, the different algorithmic procedures are presented in Section 3.2, and the results of the benchmark study are displayed in Section 3.3. Lastly some discussion is developed in Section 3.4.

3.1 Mixed Model

In this chapter we focus on variance component models of the form:

$$Y \sim \mathcal{N}(X\beta, \sum_{k=1}^{K} \sigma_k^2 V_k) \tag{3.1}$$

where Y is the vector of observations, X is an incidence matrix, β is the vector of fixed effects, V_k is the (known) covariance matrix associated to the k^{th} random effect and $\gamma = (\sigma_1^2, ..., \sigma_K^2)$ is the vector of variances associated to the K random effects. A special

case of Model (3.1) is the following mixed model:

$$\begin{aligned} Y &= X\beta + \sum_{k=1}^{K-1} Z_k u_k + e \qquad (3.2) \\ \text{with} \quad & \begin{cases} u_k \sim \mathcal{N}(0, \sigma_k^2 R_k), \quad k = 1, ..., K-1, \\ e \sim \mathcal{N}(0, \sigma_K^2 V_K) \\ u_1 \perp ... \perp u_{K-1} \perp e \end{cases} , \end{aligned}$$

where u_k is the k^{th} random effect vector, Z_k (resp. R_k) is the incidence matrix (resp. the correlation matrix) associated with random effect u_k, e is an error vector, and notation $A \perp B$ stands for "A and B are independent". Model (3.2) boils down to Model (3.1) where $V_k = Z_k R_k Z_k^T$. Lastly, we introduce $\Sigma_\gamma = \sum_{k=1}^{K} \sigma_k^2 V_k$, the covariance matrix of vector y. In the following we will use Σ in place of Σ_γ when no confusion is possible.

The goal is then to infer the unknown mean and variance parameters β and γ. In this chapter we consider the Restricted Maximum Likelihood (ReML) estimation procedure (Harville, 1977).
Let $\Pi_{X^\perp} = I - X(X^TX)^{-1}X^T$ be the projection matrix on $span(X)^\perp$, and M be any matrix built from the columns of $\Pi_{X^\perp}$ such that M is of full rank and $rank(M) = rank(\Pi_{X^\perp})$. Applying M to the initial data vector y allows one to get rid of the fixed effects. The restricted (log-) likelihood corresponds to the (log-) likelihood of the transformed data My, and has the following expression

$$\begin{aligned} \mathcal{L}_R(\gamma|y) = -\tfrac{1}{2}\big[\quad & \log(|X^T\Sigma_\gamma^{-1}X|) + \log(|\Sigma_\gamma|) \\ & + y^T P_\gamma y \big] \end{aligned} \qquad (3.3)$$

where $|H|$ stands for the determinant of matrix H and

$$\begin{aligned} P_\gamma &= M^T(M\Sigma_\gamma M^T)^{-1}M \\ &= \Sigma_\gamma^{-1} - \Sigma_\gamma^{-1}X(X^T\Sigma_\gamma^{-1}X)^{-1}X^T\Sigma_\gamma^{-1}. \end{aligned}$$

Note that $\mathcal{L}_R(\gamma|y)$ does not depend on β (since $MX = 0$ by construction), nor on the specific choice of M thanks to the second expression of P_γ (see also Searle et al. (1992)). Variance parameters $\widehat{\gamma}$ can be estimated by applying the classical Maximum Likelihood procedure to $\mathcal{L}_R$, then fixed effects can be obtained using the following formula:

$$\widehat{\beta} = (X^T\Sigma_{\widehat{\gamma}}^{-1}X)^{-1}X^T\Sigma_{\widehat{\gamma}}^{-1}y.$$

Although quite popular, the ReML procedure may be quite challenging from a computational point of view, the bottleneck being the maximization of the log-likelihood (3.3) w.r.t. γ. Although the first derivative of $\mathcal{L}_R$ with respect to σ_k^2 has a simple expression:

$$\frac{\partial \mathcal{L}_R}{\partial \sigma_k^2} = -\frac{1}{2}\left[tr(P_\gamma V_k) - y^T P_\gamma V_k P_\gamma y\right],$$

solving the system of K equations $\frac{\partial \mathcal{L}_R}{\partial \sigma_k^2} = 0$, $k = 1, ..., K$ does not lead to a closed form expression for $\widehat{\gamma}$. Consequently likelihood maximization has to be performed numerically. The next section present some classical optimization algorithms to obtain the ReML variance estimates.

3.2 Algorithm Overview

In this section we consider two algorithms for the maximization of the restricted log-likelihood. The first one is an adaptation of the Newton algorithm, where the classical Hessian matrix is replaced by the so-called "Average Information" matrix that can be efficiently computed (Johnson and Thompson, 1995; Gilmour et al., 1995). This algorithm is quite popular and is implemented both in the *gaston* R package (Perdry and Dandine-Roulland, 2017) and in the statistical software package *Asreml* (Gilmour et al., 2009). The second one is the iterative Minimization-Maximization algorithm for variance component models recently proposed in Zhou et al. (2015). The last paragraph of this section focuses on the particular case of the 2-variance component mixed model (i.e. $K = 2$) where efficient computational tricks exist to significantly speed up the optimization.

3.2.1 Newton-based algorithms

Newton and Fisher algorithms

Let first rewrite Model (3.1) as

$$Y = X\beta + Zu + e$$

where $Z = (Z_1|...|Z_{K-1})$, $u = (u_1|...|u_{K-1})^T$. The joint distribution of (u, e) is

$$\begin{aligned}
\begin{bmatrix} u \\ e \end{bmatrix} &\sim \mathcal{N}\left(0, \sigma_K^2 \begin{bmatrix} G(\delta) & 0 \\ 0 & V_K \end{bmatrix}\right) \\
\text{where } \delta &= (\frac{\sigma_1^2}{\sigma_K^2}, ...\frac{\sigma_{K-1}^2}{\sigma_K^2}) \\
\text{and } G_\delta &= \begin{pmatrix} \delta_1 R_1 & 0 & ... & 0 \\ 0 & \delta_2 R_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \delta_{K-1} R_{K-1} \end{pmatrix}.
\end{aligned}$$

Using these notations, $\mathcal{L}_R$ can be reformulated as:

$$\begin{aligned}
\mathcal{L}_R(\sigma_K^2, \delta|y) = \; & -\tfrac{1}{2}\big[(n-r)\log(\sigma_K^2) + \log(|X^T H^{-1} X|) \\
& + \log(|H|) + \tfrac{y^T P_\delta y}{\sigma_K^2}\big]
\end{aligned}$$

where $H = ZG_\delta Z^T + V_K$, $P_\delta = \sigma_K^2 P_\gamma$ and r is the rank of matrix X. Starting from this last expression, one can perform optimization of $\mathcal{L}_R$ using an iterative scheme like the Newton algorithm that requires the first and second derivatives of $\mathcal{L}_R$ w.r.t. both δ and σ_K^2. The first derivatives are

$$\begin{aligned}
\left[\nabla \mathcal{L}_R\left(\delta, \sigma_K^2\right)\right]_k &= -\frac{1}{2}\left(tr(P_\delta V_k) - \frac{y^T P_\delta V_k P_\delta y}{\sigma_K^2}\right) \quad 1 \le k \le K-1, \\
\left[\nabla \mathcal{L}_R\left(\delta, \sigma_K^2\right)\right]_K &= -\frac{1}{2}\left(\frac{n-r}{\sigma_K^2} - \frac{y^T P_\delta y}{\sigma_K^4}\right).
\end{aligned}$$

Similarly, the second derivatives are

$$\begin{aligned}
\left[\mathcal{HL}_R\left(\delta,\sigma_K^2\right)\right]_{kk'} &= \frac{1}{2}tr(P_\delta V_k P_\delta V_{k'}) - \frac{y^T P_\delta V_k P_\delta V_{k'} P_\delta y}{\sigma_K^2},\\
\left[\mathcal{HL}_R\left(\delta,\sigma_K^2\right)\right]_{kK} &= -\frac{1}{2}\frac{y^T P_\delta V_k P_\delta y}{\sigma_K^4},\\
\left[\mathcal{HL}_R\left(\delta,\sigma_K^2\right)\right]_{KK} &= \frac{n-r}{2\sigma_K^4} - \frac{y^T P_\delta y}{\sigma_K^6}\ .
\end{aligned}$$

Denoting $\nabla^{(t)}_{\mathcal{L}_R}$ and $\mathcal{H}^{(t)}_{\mathcal{L}_R}$ the gradient and Hessian matrix of $\mathcal{L}_R$ evaluated at point $\left(\delta^{(t)},\sigma_K^{2(t)}\right)$ respectively, the Newton method then iterates the following recursion:

$$\begin{pmatrix} \delta^{(t+1)} \\ \sigma_K^{2(t+1)} \end{pmatrix} = \begin{pmatrix} \delta^{(t)} \\ {\sigma_K}^2(t) \end{pmatrix} - \left[\mathcal{H}^{(t)}_{\mathcal{L}_R}\right]^{-1} \nabla^{(t)}_{\mathcal{L}_R}\ . \tag{3.4}$$

Replacing the Hessian matrix $\mathcal{H}^{(t)}_{\mathcal{L}_R}$ by its expected value in equation (3.4) leads to the Fisher Scoring (FS) algorithm. The expected moments of the Hessian matrix are

$$\begin{aligned}
\mathbb{E}\left[\mathcal{HL}_R\left(\delta,\sigma_K^2\right)\right]_{kk'} &= -\frac{1}{2}tr(P_\delta V_k P_\delta V_{k'}),\\
\mathbb{E}\left[\mathcal{HL}_R\left(\delta,\sigma_K^2\right)\right]_{kK} &= -\frac{1}{2}\frac{tr(P_\delta V_k)}{\sigma_K^4},\\
\mathbb{E}\left[\mathcal{HL}_R\left(\delta,\sigma_K^2\right)\right]_{KK} &= -\frac{n-r}{2\sigma_K^4}\ .
\end{aligned}$$

AI algorithm

A popular alternative to the Newton and FS algorithms is the use of the Average Information (AI) matrix in place of the Hessian matrix Johnson and Thompson (1995). The AI matrix is defined as the average of the Hessian and its expectation. The efficiency of this strategy leads in the general expression of the resulting matrix. One has

$$\begin{aligned}
AI_{kk'} &= \tfrac{y^T P_\delta V_k P_\delta V_{k'} P_\delta y}{2\sigma_K^2},\\
AI_{kK} &\approx \tfrac{y^T P_\delta V_k P_\delta y}{2\sigma_K^4},\\
AI_{KK} &= \tfrac{y^T P_\delta y}{2\sigma_K^6},
\end{aligned} \tag{3.5}$$

where for the second term the approximation $tr(V_k P_\delta) = y^T P_\delta V_k P_\delta y$ is used. Compared with the previous expressions obtained for the Newton and FS algorithms, computing the AI matrix does not involve any trace computation anymore. Note that P_δ is computed at each step using formula

$$P_\delta = \Sigma^{-1} - \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$$

where δ and σ_K^2 are fixed at their current value.

MME trick

One can avoid the explicit computation of matrix P_δ by obtaining the quantities appearing in equation set (3.5) through the solving of Henderson's mixed model equations (MME):

$$\begin{bmatrix} X^T V_K^{-1} X & X^T V_K^{-1} Z \\ Z^T V_K^{-1} X & Z^T V_K^{-1} Z + G(\delta)^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T V_K^{-1} y \\ Z^T V_K^{-1} y \end{bmatrix} \tag{3.6}$$

Let C be the coefficient matrix of the MME. One one hand, assuming C is sparse C^{-1} can be efficiently derived, giving access to $\hat{\beta}$, $\hat{u}$ and any submatrix of C^{-1}. On the other hand one can reexpress the AI matrix as

$$\begin{aligned} AI_{kk'} &= \frac{[y^T P_\delta Z_k] R_k [Z_k^T P_\delta Z_{k'}] R_{k'} [Z_{k'} P_\delta y]}{2\sigma_K^2} \\ AI_{kK} &= \frac{[y^T P_\delta Z_k] R_k [Z_k^T P_\delta y]}{2\sigma_K^4} \\ AI_{KK} &= \frac{y^T P_\delta y}{2\sigma_K^6} \end{aligned}$$

and provide the following expressions for the quantities involved in each term:

$$\begin{aligned} &Z^T P_\delta y = G^{-1} \hat{u} \\ &Z^T P_\delta Z = G^{-1} - G^{-1} [C^{-1}]_{uu} G^{-1} \\ &P_\delta y = V_K^{-1} \hat{e}. \end{aligned} \tag{3.7}$$

where $\hat{e} = y - X\widehat{\beta} - Z\hat{u}$ and $[C^{-1}]_{uu}$ corresponds to the submatrix of C^{-1} associated with the random component u. Assuming all matrices V_k are invertible (i.e. definite positive), all these expressions are easily obtained from $\hat{\beta}$, $\hat{u}$ and C^{-1}, leading to an efficient computation of the AI matrix.

Proof. In this proof, we will demonstrate equalities given in equation (3.7). First P_δ can be written using the matrix $S = V_K^{-1} - V_K^{-1} X (X^T V_K^{-1} X)^{-1} X^T V_K^{-1}$:

$$P_\delta = S - SZ(Z^T SZ + G^{-1})^{-1} Z^T S \tag{3.8}$$

Using the equation (3.8) and Petersen et al. (2008), one can obtain:

$$Z^T P_\delta Z = G^{-1} - G^{-1}(Z^T SZ + G^{-1})^{-1} G^{-1}$$

One can note that $(Z^T SZ + G^{-1})^{-1} = [C^{-1}]_{uu}$ using the Schur complement.
From the equation (3.6), one can deduce:

$$\widehat{u} = (Z^T SZ + G^{-1})^{-1} Z^T Sy \tag{3.9}$$

Derived from (3.8) and (3.9), one has:

$$Z^T P_\delta y = G^{-1} \widehat{u}$$

And derived from (3.8), (3.9) and (3.6), one has:

$$P_\delta y = Sy - SZ\tilde{u} = S(y - X\widehat{\beta} - Z\widehat{u}) = V_K^{-1} \tilde{e}$$

where $\tilde{e} = y - X\widehat{\beta} - Z\widehat{u}$.

□

3.2.2 MM method

MM algorithms represent another class of iterative schemes that have been thoroughly described in Hunter and Lange (2004). We provide here a brief overview of the MM principle based on the previous reference. Consider an optimization problem where the goal is to find the minimizer θ^* of a function $f(\theta)$ (in our setting $f = -\mathcal{L}_R$ and $\theta = \gamma$), one builds at each step t a new surrogate function $g^{(t)}$ that upper bounds f in the following sense:

$$g^{(t)}(\theta) \geq f(\theta) \text{ and } g^{(t)}(\theta^{(t)}) = f(\theta^{(t)}) ,$$

where $\theta^{(t)}$ is the current evaluation of θ^*. Assuming the surrogate function can be minimized easily, one defines

$$\theta^{(t+1)} = \arg\min_{\theta} g^{(t)}(\theta) .$$

One can show that the sequence $\left(\theta^{(t)}\right)_{t\geq 1}$ satisfies the descent property $f(\theta^{(t+1)}) \leq f(\theta^{(t)})$.

In the context of variance component mixed models, Zhou et al. (2015) thoroughly described the application of the MM method to the maximization of the likelihood. Here we illustrate the strategy applied to the restricted likelihood. The following surrogate function in place of the restricted log-likelihood:

$$g^{(t)}(\gamma) = \frac{1}{2}\sum_{k=1}^{K}\left[\sigma_k^2 tr(P_\gamma^{(t)}V_k) + \frac{\sigma_k^{4(t)}}{\sigma_k^2} y^T P_\gamma^{(t)} V_k P_\gamma^{(t)} y)\right]$$

where $\Sigma^{(-t)} = \Sigma(\gamma^{(t)})^{(-1)}$. Thank to the linearity of $g^{(t)}$ with respect to $\sigma_1^2, ..., \sigma_K^2$, the update formulas are easily obtained by setting the gradient of $g^{(t)}$ at 0:

$$\sigma_k^{2(t+1)} = \sigma_k^{2(t)} \sqrt{\frac{y^T P_\gamma^{(t)} V_k P_\gamma^{(t)} y}{tr(P_\gamma^{(t)} V_k)}}$$

Similar to EM algorithms, MM algorithms can benefit from accelerating strategies to achieve better rates of convergence (by reducing the number of iterations required to achieve a given precision). In the present chapter we combined the MM algorithm with the acceleration strategy of Varadhan and Roland (2008).

3.2.3 A special case: two variance component mixed models

Joint diagonalization

In the particular case where $K = 2$, the computational burden can be greatly alleviated by applying a simple transformation to the data. Assuming V_2 is symmetric positive definite, there exists a matrix U such that

$$UV_1U^T = D$$
$$UV_2U^T = I_n$$

Where D is a diagonal matrix and I_n is the identity matrix of size n. Applying the transformation $\tilde{Y} = UY$ (and $\tilde{X} = UX$) Model (3.1) is modified as follows:

$$\tilde{Y} \sim \mathcal{N}(\tilde{X}\beta, \sigma_1^2 D + \sigma_2^2 I_n) \ ,$$

Once matrix U is computed (which requires one diagonalization and one Cholesky decomposition) all traces and inverse matrix computations involved in the computation of $\mathcal{L}_R$ and its derivatives become straightforward. While this diagonalization trick is already useful when fitting one model, the situation where many two variance component MM with identical variance matrices (but different observation vectors Y and/or matrices X) have to be fitted will greatly benefit from it since the computation of U will be done only one time. This situation arises in GWAS analysis where the same trick has also been exploited Lippert et al. (2011). The joint diagonalization trick may be directly combined with any optimization procedure such as the AI and MM algorithms presented in the previous sections. Alternatively, one can first "profile" one of the two variance parameters out of the log-likelihood and then apply optimization techniques, as explained in the next paragraph.

Profiling trick

The profiling trick consists in two steps. First, one can reexpress the restricted log-likelihood associated with Model 3.2.3 as a function of σ_2^2 and ratio $\delta = \frac{\sigma_1^2}{\sigma_2^2}$ as follows:

$$\begin{aligned} \mathcal{L}_R(\sigma_2^2, \delta|\tilde{y}) = \ & -\tfrac{1}{2}\big[(n-r)\log(\sigma_2^2) + \log(|D\delta + I_n|) \\ & + \log(|\tilde{X}^T(D\delta + I_n)^{-1}\tilde{X}|) + \tfrac{1}{\sigma_2^2}\tilde{y}^T\tilde{P}_\delta\tilde{y}\big] \end{aligned} \tag{3.10}$$

where $\tilde{P}_\delta = (D\delta + I_n)^{-1} - (D\delta + I_n)^{-1}X(X^T(D\delta + I_n)^{-1}X)^{-1}X^T(D\delta + I_n)^{-1}$.
Second, one can observe that the optimization of (3.10) in σ_2^2 leads to a closed form expression:

$$\hat{\sigma}_2^2(\delta) = \frac{1}{n-r}\tilde{y}^T\tilde{P}_\delta\tilde{y} \ .$$

This expression can be plugged back into equation (3.10) to obtain a function that depends on δ only:

$$\begin{aligned} \mathcal{L}_R(\delta|\tilde{y}) = \ & -\tfrac{1}{2}\big[(n-r)\log(\hat{\sigma}_2^2(\delta)) + \log(|D\delta + I_n|) \\ & + \log(|\tilde{X}^T(D\delta + I_n)^{-1}\tilde{X}|) + (n-r)\big] \ . \end{aligned} \tag{3.11}$$

This last expression of the restricted log-likelihood can then be optimized in δ using one of the aforementioned procedures. This strategy is implemented in the R package *gaston* where the Newton algorithm is used for the optimization of (3.11), and also in FaSTLmm where the optimization is first performed on a grid then refined using the Brent algorithm (see Lippert et al. (2011) and Perdry and Dandine-Roulland (2017) for details). Next we describe the application of the MM algorithm to optimize (3.11) and show that at each step the update of δ has a closed form expression.

Fast profiling for the MM algorithm

Following the same line of proof as in Zhou et al. (2015), we consider the following surrogate function

$$\begin{aligned} g^{(t)}(\delta) = \; & \tfrac{1}{2}\big[(n-r)\log(\tfrac{1}{n-r}\tilde{y}^T\tilde{P}_\delta^{(t)}(\tfrac{\delta^{2(t)}}{\delta}D + I_n)\tilde{P}_\delta^{(t)}\tilde{y}) \\ & + tr(\tilde{P}_\delta^{(t)}(D\delta + I)) + c^{(t)}\big] \; . \end{aligned}$$

Introducing some additional notations $A = \tilde{y}^T\tilde{P}_\delta^{(t)}\tilde{y}$, $B = \tilde{y}^T\tilde{P}_\delta^{(t)}D\delta^{2(t)}\tilde{P}_\delta^{(t)}\tilde{y}$ and $C = tr(\tilde{P}_\delta^{(t)}D)$, the first derivative of the surrogate function 3.2.3 is:

$$\frac{\partial g(\delta|\delta^{(t)})}{\partial\delta} = -\frac{1}{2}\big[(n-r)\frac{-B}{\delta B + \delta^2 A} + C\big]$$

Solving $\frac{\partial g(\delta|\delta^{(t)})}{\partial\delta} = 0$ boils down to solving the following quadratic problem :

$$\frac{AC}{n-r}\delta^2 + \frac{BC}{n-r}\delta - B = 0$$

that admits a unique positive solution $\hat{\delta}$.

3.3 Results

3.3.1 Setting

Algorithms

We compared 7 methods that correspond to different implementations of the optimization algorithms described in Section 3.2.1.

AI-ASreml corresponds to the AI algorithm as implemented in the licensed software ASreml available on R. When computing the AI matrix the direct computation of P_δ is avoided, and the required quantities are obtained as by-products of the solutions of Henderson's equation, as explained in Section 3.2.1. Note that in order to apply this trick each variance matrix V_k must be definite positive which may represent a limitation of the method.

AI-gaston corresponds to the AI algorithm as implemented in the R package *gaston*. Here the AI matrix is obtained through the direct computation of P_δ, and consequently this version of the AI algorithm does not require the variance matrix to be definite positive. Note that this algorithm is only available for $K > 2$ in *gaston*. When $K = 2$ another optimization algorithm is applied (see below).

MM corresponds to the implementation of the MM algorithm presented in Section 3.2.2, combined with the acceleration of Varadhan and Roland (2008). We implemented the MM procedure in the MM4LMM R package. The core inference procedure was coded in C++. When applied to a 2-VC mixed model the closed form expression update presented in Section 3.2.3 is used.

FaST-LMMe is dedicated to the fitting of mixed models with two variance components only, and implements the diagonalization and profiling tricks presented in Sections 3.2.3 and 3.2.3. The "e" subscript stands for "exact", meaning that an exact computation of the variance ratio δ is required for each fitting of the model.

FaST-LMMa corresponds to the approximate FaST-LMM algorithm where parameter δ is computed at once on a null model, then plugged without further modification in each fitted model. This may significantly reduce the computational burden since given δ only one variance component needs to be estimated. While the comparison of computational times between FaST-LMMa and the other (exact) methods is somewhat unfair since the first one solves a simpler optimization problem, it provides some insight about the impact the approximation may have on the resulting list of markers found to be significant (see Section 3.3.1).

We also included the *optimizeLmer* function of the *lme*4 package in the present comparison setting, as a gold standard for variance estimation and p-value precision - *lme*4 being one of the most popular R packages for mixed model fitting. Lastly, in the experiment involving mixed models with only two variance components we included *gaston* in the comparison setting. In the $K = 2$ setting *gaston* performs ReML optimization using Newton method.

Criteria

We focus here on the application of mixed models to statistical genetics, with two specific application cases in mind: variance component analysis (VCA) and genome wide association study (GWAS). These two cases are described here, along with the specific qualities that are required for the mixed model fitting algorithms for each of them.

In VCA the goal is to quantify the contribution of each random component to the global variance of a given phenotype Y. Such an analysis may be performed jointly on data from different trials involving the same genotypes, which may result in i) a large number of measurements, and ii) a complex mixed model including a higher number of fixed and random effects. In this context, optimization algorithms may be compared according to i) the precision of the variance estimates, and i) their ability to scale with the size of the data and the number of variance components.

In GWAS, the goal is to identify markers that are significantly associated with phenotypic variation. The relationship between a given marker and the phenotype may be tested in a model that accounts for both the marker effect and background genetic effect specific to each individual. This background effect can be modeled as a random effect whose associated covariance matrix may be deduced from the kinship between individuals, leading to a VC mixed model.
Since GWAS requires the fitting of a mixed model for each tested marker, computational efficiency is of major importance. However, one should also be sure that the optimization algorithm yields accurate p-values. In the following we compare the different algorithms based on i) their computational time, ii) their associated list of markers that are found

to be significantly related to the phenotype, and iii) the ordered list of the first 100 markers, where the ranking of the markers is based on their associated p-values.

3.3.2 Algorithm comparison

All algorithms were run and compared on a server with a cpu of 2200 MHz and 8 cores, applying their by-default configuration settings.

Variance component analysis

In this section, we study the performance of algorithms AI-ASreml, AI-gaston and MM in terms of variance component estimation when applied on two different datasets. FaST-LMM is not considered here since it only handles the case $K = 2$.

Datasets The Technow dataset is a maize panel consisting in hybrids derived from an incomplete factorial crossing design between lines belonging to two heterotic groups: flint and dent. A total of $n_D = 123$ parental dent lines and $n_F = 86$ parental flint lines were crossed to obtain $n_H = 1,254$ hybrids. Parental lines were genotyped at 35,478 markers, and two phenotypic traits were quantified: grain yield (GY) and grain moisture (GM). We present here the results obtained on GY, the ones obtained with GM being similar.
The Giraud dataset consists is also a maize panel that consists in $n_H = 951$ maize hybrids derived from an incomplete factorial crossing design between $n_D = 875$ dent lines and $n_F = 883$ flint lines, and described in Giraud et al. (2017). Parental lines were genotyped at 9643 biallelic markers, and 4 phenotypic traits were quantified: dry matter content (DMC), dry matter yield (DMY), days to silking (DtSilk) and plant height (PH). Here we focus on DMY, the results obtained with the other traits being also quite similar. In this experiment hybrids were phenotyped in 8 trials, therefore different strategies can be considered for the data analysis. One can first compute least-square means for each hybrid (correcting for trial effects), then perform a VCA on the resulting dataset where each hybrid is characterized by a single measurement. Alternatively, one can directly perform the VCA on the initial dataset without averaging per hybrid. In the following we refer to the two datasets corresponding to these two strategies as the LSM (Least Square Means) dataset and the NAM (Non Averaged Measurement) dataset, respectively.

Statistical analysis For the Technow dataset and the LSM dataset, VCA was performed using the following model:

$$\begin{aligned} &Y = 1\mu + Z_F G_F + Z_D G_D + Z_H G_H + E \\ &G_F \sim \mathcal{N}(0, \sigma_F^2 K_F) \\ &G_D \sim \mathcal{N}(0, \sigma_D^2 K_D) \\ &G_H \sim \mathcal{N}(0, \sigma_H^2 \Phi) \\ &E \sim \mathcal{N}(0, \sigma_E^2 I) \\ &G_F \perp G_D \perp G_H \perp E \end{aligned} \tag{3.12}$$

where Y is the phenotypic vector, μ is the intercept, G_F, G_D and G_H are the polygenic effect derived from the flint parent, the dent parent and their specific interaction, respectively, with Z_F, Z_D and Z_H their associated incidence matrices. Correlation matrices

Study	VC	AI-gaston	AI-ASreml	MM	lme4
Technow	σ^2_D	32.14	32.14	32.14	32.14
	σ^2_F	25.34	25.34	25.34	25.34
	σ^2_H	3.80	3.80	3.80	3.80
	σ^2_E	14.84	14.84	14.84	14.84
LSM	σ^2_D	0.57	0.57	0.57	0.57
	σ^2_F	0.38	0.38	0.38	0.38
	σ^2_H	0.02	0.02	0.02	0.02
	σ^2_E	0.69	0.69	0.69	0.69

Table 3.1: Variance components estimates for the Technow and the averaged Giraud datasets. The considered phenotypes are GY (for the Technow dataset) and DMY (for the LSM dataset).

K_F and K_D correspond to the kinship matrices between the dent (resp. flint) parental lines, and were computed according to Astle and Balding (2009). Correlation matrix Φ corresponds to the matrix of double relatedness between hybrids, of general term

$$\widehat{\Phi}_{h,h'} = (\widehat{K}_F)_{f,f'} \times (\widehat{K}_D)_{d,d'}$$

where h (resp. h') is the hybrid resulting from the crossing between the flint and dent lines f and d (resp. f' and d'). Lastly, E is the error vector.
Regarding the NAM dataset, we assumed the genetic effects to be identical in all trials (i.e. no genetic×trial interaction), and the intercept to be specific to each trial. This results in updating Model (3.12) as follows:

$$Y = 1\mu + X_T\beta_T + Z_F G_F + Z_D G_D + Z_H G_H + E$$

with the same notations and statistical assumptions as before. Here X_T is the incidence matrix associated to trials and β_T is the vector of trial effects. Alternatively, non-informative covariance matrices were also considered for the random genetic effects, replacing the kinship matrices by identity matrices. The resulting analysis is referred to as the NAM-Id dataset in the following.

Variance Component Values Table 3.1 displays the variance component estimates obtained with the different algorithms, for the Technow and the LSM datasets. One can observe that all algorithms yield the same results that are also identical to the one obtained with the *lme*4 reference. Similar conclusions were obtained when considering other phenotypic traits and/or the NAM dataset.

Computational time Table 3.2 displays the computational time associated with the different algorithms in different settings. In order to investigate the ability of the algorithms to cope with datasets of increasing sizes, we built several intermediate versions of the NAM dataset by adding observations corresponding to different trials sequentially. This corresponds to lines 3 to 6 in Table 3.2, where 2, 4, 6 and 8 trials were successively included in the analysis, leading to an increase of the number of observations from 950 to 7,725.
Several comments can be made about the results of Table 3.2. First, for some datasets,

Dataset	Nb Trials	Nb Obs	Sparse	AI-gaston	AI-ASReml	MM	Ratio
Technow	-	1,254	No	4.77	116.61	6.57	24
LSM	-	950	No	2.88	338.89	7.76	117
NAM	2	1,891	No	18.67	702.90	31.33	37
NAM	4	3,820	No	148.02	768.98	146.88	5
NAM	6	5,749	No	515.34	619.14	471.59	1
NAM	8	7,725	No	1,226.61	610.63	1,037.85	2
NAM-Id	8	7,725	Yes	1,150.56	0.32	1,519.25	4,747

Table 3.2: Computational time (in sec.) associated to the different algorithms and different analysis. Ratio corresponds to the ratio between the slowest and the fastest algorithms.

the difference in terms of computational efficiency from an algorithm to another may be huge, as quantified using the ratio between the worst and the best computational time obtained (last column). Second, one can see that there does not exist a "best" algorithm that outperforms the other ones in all cases: for instance, AI-ASreml is $100\times$ slower than its competitors on the LSM dataset, and $1,000\times$ faster on the NAM-Id dataset.

The different configurations presented in Table 3.2 give a clear picture about what can be expected from the different algorithms. Note that for the MM and AI-gaston algorithms the computational cost is directly related to the cost of inverting matrix P_γ - a non-sparse $n \times n$ matrix - whereas the computational cost of AI-ASreml roughly depends on the cost of inverting $G(\delta)$. This last cost depends on the sparsity of $G(\delta)$ and on its size $N \times N$ with $N = n_D + n_F + n_H$. When considering the Technow and LSM datasets, one has $n = n_H < N$, and the kinship matrices are not sparse. Consequently the computational time is much higher for AI-ASreml than for its competitors. On the opposite, the case "NAM-Id" correspond to the case where the individual observations of the different trials are not merged, and where each kinship matrix is assumed to be an identity matrix. In this case, one has $n \approx 8 \times n_h >> N$, where 8 is the total number of trials, and the kinship matrices are as sparse as possible. This case is clearly favorable to AI-ASreml that has been designed for such applications. The cases where trials are sequentially added show some intermediate configurations where the number of observations n grows with the number of trials. At first n is lower than N and MM and AI-gaston outperform AI-ASreml. As the number of trials increases the gap gets smaller, and finally when $N << n$ applying the MME trick becomes relevant (even with $G(\delta)$ not being sparse) and AI-ASreml becomes more efficient than MM and AI-gaston.

GWAS with two variance components

Dataset We consider the two CornFed Association panels - named CF-Flint and CF-dent hereafter - described in Rincent et al. (2014b). The CF-Flint panel consists in 259 maize lines of the Flint heterotic group crossed with a tester from the dent heterotic group. Lines were genotyped at $\approx 50K$ biallelic markers, among which 39,076 were kept after quality control for the present study. Eight phenotypic trait were quantified: Tass, Silk, ASI, DMC, DMY, PLHT, DM_Flo and DM_Y_Flo and DMYcorr, see Rincent et al. (2014b) for details. The association study was performed using the following model: at

	AI-gaston	AI-ASreml	MM	FaSTLMMe	FaSTLMMa	lme4
DM_Y_Flo	3.81	3037.78	5.23	28.64	2.01	12886.23
Tass	6.38	4514.99	5.35	28.40	2.19	34852.63
Silk	3.14	-	4.88	9.20	2.20	-
ASI	3.85	-	5.49	28.41	2.20	-
DMC	4.03	-	5.51	28.21	2.00	-
DMY	3.99	-	5.71	28.61	2.00	-
PLHT	4.50	-	5.78	28.53	2.21	-
DMC_Flo	3.63		6.27	28.53	2.02	-

Table 3.3: Computational time (in sec.) associated to the different algorithms for the complete analysis of the panel, trait by trait.

a marker ℓ and for a given trait, one has

$$Y = 1\mu + X_\ell \beta_\ell + U + E \ ,$$

with Y the vector of phenotypes, X_ℓ the vector of observed allele (0 or 1) for each line at marker ℓ, β_ℓ the effect associated to allele 1, U the random effect accounting for the genetic background, and E the error vector. One further assumes that

$$U \sim \mathcal{N}(0, \sigma_G^2 K) \ \text{ and } E \sim \mathcal{N}(0, \sigma_E^2 I)$$

where K is the matrix of kinship between lines. U and E are assumed to be independent. The unknown parameters to infer are the fixed effects μ and β_ℓ, along with the variances of the random effects σ_G^2 and σ_E^2. In order to identify markers that are strongly related to phenotypic variation the null hypothesis $H_0 : \{\beta_\ell = 0\}$ was tested using a Wald test procedure. Following Rincent et al. (2014b), the kinship matrix was computed from the markers using the formula of Astle and Balding (2009).

The analysis of panel CF-dent led to similar conclusions and is not presented here.

Computational Time The computational times corresponding to the analysis of all markers of the panel by each algorithm are displayed in Table 3.3, for each of the 8 considered phenotypes. The computational performance of AI-ASreml is reported for phenotype DM_Y_Flo only, which is sufficient to observe that it compares poorly with its competitors, being $\times 100$ slower than FastLMMe and $\times 600$ slower than AI-gaston and the MM algorithms. All other methods performed the analysis in less than 30s. As expected the fastest method is FastLMMa that performs only an approximate optimization of the restricted likelihood, improving by a factor 10 over its exact counterpart but only by a factor 2/3 over AI-gaston and MM.

Comparison of the p-value orderings Hereafter we focus two phenotypes: Tasseling and DM_yield_Flo. For a given algorithm $\mathcal{A}$ we note $L_\mathcal{A}$ the ordered list of markers, where markers are ordered on the basis of their p-values - from the smallest to the highest one. Similarly, a same ranking L_{lme4} is also obtained using $lme4$, this last ranking being considered as the reference. One can then compute a concordance score

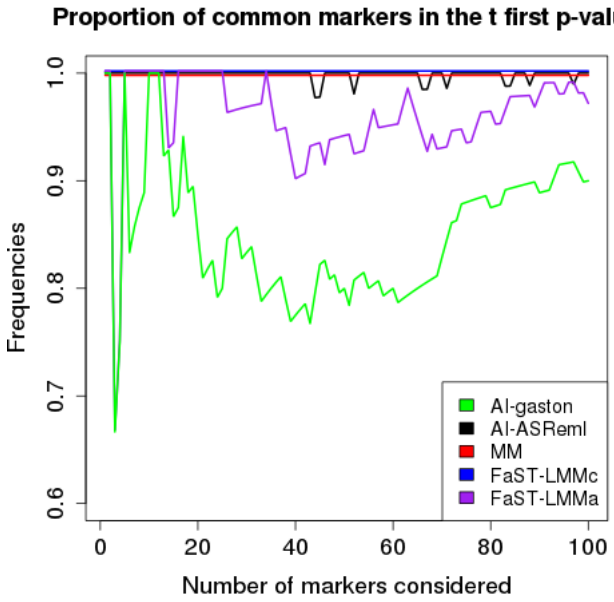


Figure 3.1: Concordance curves for the different algorithms, based on the p-values obtained from the Tasseling association analysis, and using lme4 *as a reference.*

between algorithm $\mathcal{A}$ and *lme*4 at rank r as follows:

$$C_r(\mathcal{A}, lme4) = \frac{L_{\mathcal{A}}(r) \cap L_{lme4}(r)}{r},$$

where $L_{\mathcal{A}}(r)$ contains the first r items of list $L_{\mathcal{A}}$.
Figure 3.1 displays the concordance curves of each algorithm for Tasseling. One can observe that all methods are (almost) perfectly concordant with *lme*4 except for *gaston* and to a lower extend FastLMMa. Figure 3.2 displays the same curves but for DM_yield_Flo. The discrepancy between *gaston*, FastLMMa and *lme*4 is even higher when considering DM_yield_Flo (see Fig. 3.2): focusing on the top 100 ranked markers, one can still observe a difference of ≈10% between FastLMMa/*gaston* and *lme*4.

Figure 3.2 displays the concordance curves of each algorithm. One can observe that all methods are (almost) perfectly concordant with *lme*4 except for *gaston* and FastLMMa. When considering the top 100 ranked markers, one can still observe a difference of 10% between FastLMMa and *lme*4.

QTL detection As illustrated in the previous section, the use of different algorithms for (restricted) likelihood maximization leads to different p-values lists for the markers. As a consequence the ordering of the markers with respect to their p-values varies from an algorithm to another. While this variation may seem marginal, in practice any GWAS analysis will include a multiple testing correction procedure that may be affected by the (lack of) precision of the computed p-values. To further illustrate the impact of the optimization procedure, we applied a full marker identification procedure accounting for multiple testing to each list $L_{\mathcal{A}}$. Following Rincent et al. (2014b) we performed a Bonferroni correction using M_{eff} as the number of tests, where M_{eff} is the effective

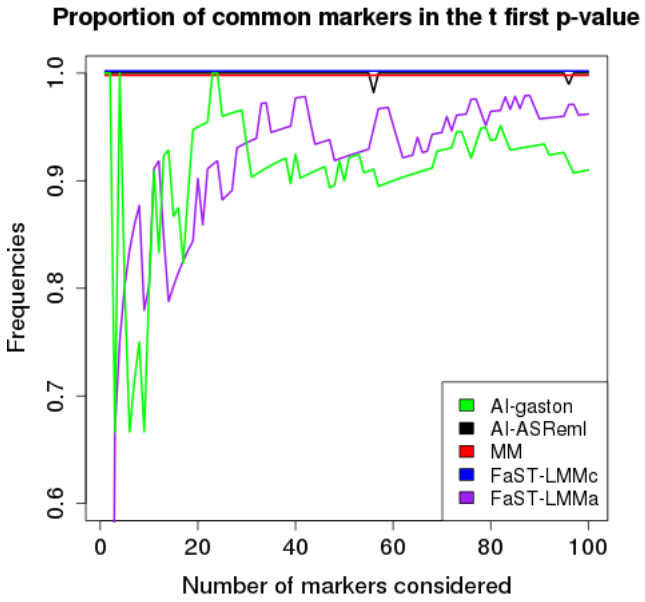


Figure 3.2: Concordance curves for the different algorithms, based on the p-values obtained from the DM_yield_Flo association analysis, and using lme4 *as a reference.*

number of tests that can be estimated using the method described in Li and Ji (2005). For the present application one obtains $M_{eff} = 3,527$. Note that this correction procedure only depends on the genotypic information (used to compute the effective number of tests) but not on the p-value distribution. Because procedures that account for the p-value distribution may be more sensitive to a wrong ordering and/or a lack of precision in p-value estimation, we also performed a Benjamini-Hochberg correction (noted BH hereafter, Benjamini and Hochberg (1995)) to obtain an alternative list of candidate QTLs. For the Li&Ji procedure the nominal level of global type I error was fixed at 0.05, and for the BH procedure 3 nominal levels were considered: 0.05, 0.1 and 0.2.

Tables 3.4 and 3.5 display the list of all markers that have been declared significant by at least one of the algorithms for Tasseling and DM_yield_Flo respectively. When considering Tasseling, only one marker is significantly detected overall. When the Li&Ji procedure is applied the marker is detected by all algorithms, and when the BH procedure is applied the marker is detected at level 0.2, except for the two FastLMM algorithms. The results are more contrasted when considering DM_yield_Flo. First, as expected from Figures 3.1 and 3.2, the AI-ASreml and MM procedures yield detection results that are identical to the ones of *lme*4. Second, the two versions of FastLMM yield more conservative results: at level 0.05 and using BH, no marker is detected. When using Li&Ji, FastLMMe fails to find one of the markers detected by AI-ASreml, MM and *lme*4, and FastLMMa only finds one among five. Lastly, one can observe that AI-gaston generally yields a larger number of detected markers, 4 of them being not found by the other methods whatever the multiple testing procedure. Overall the results are consistent with what was observed at the previous paragraph: AI-gaston and FastLMMa lead to final lists of markers that are different from the ones found with all other methods.

	AI-gaston	AI-ASReml	MM	FaSTLMMe	FaSTLMMa	lme4
PZE.101070781	5.37* °	5.37* °	5.37* °	5.17*	5.11*	5.37* °

Table 3.4: $-\log_{10}$(p-value) of markers detected by at least one algorithm for Tasseling. The star (resp. circle) indicates if the marker is detected when a Li&Ji (resp. BH FDR) multiple testing procedure is used at nominal level 5% (resp. 20%).

	AI-gaston	AI-ASReml	MM	FaSTLMMe	FaSTLMMa	lme4
SYN10537	5.92* °°°	5.61* °°°	5.61* °°°	5.4* °°	4.52 °	5.61* °°°
SYN10528	5.92* °°°	5.61* °°°	5.61* °°°	5.4* °°	4.52 °	5.61* °°°
PZE.101030022	5.37* °°	5.07* °°	5.07* °°	4.9* °	4.06	5.07* °°
PZE.101123079	4.49 °	4.87* °	4.87* °	4.71 °	4.71 °	4.87* °
SYN13856	5.23* °°	5.19* °°	5.19* °°	5.01* °	5.01* °	5.19* °°
PZE.101122758	4.86* °°	4.31 °	4.31 °	4.19	4.19	4.31 °
SYN26073	4.91* °°	4.66 °	4.66 °	4.51 °	4.51 °	4.66 °
SYN10535	4	4.46 °	4.46 °	4.33	3.57	4.46 °
PZA00240.9	4.25 °	4.25 °	4.25 °	4.13	4.08	4.25 °
PZE.101123102	4.59 °	4.59 °	4.59 °	4.45 °	4.45 °	4.59 °
PZE.105146456	4.36 °	4.66 °	4.66 °	4.52 °	4.5 °	4.66 °
PZE.109091780	4.58 °	4.4 °	4.4 °	4.27	4.2	4.4 °
SYN10536	4.7 °	4.04	4.04	3.93	3.19	4.04
SYN10531	4.29 °	4.03	4.03	3.92	3.38	4.03
PZA00583.4	4.13 °	3.81	3.81	3.72	3.63	3.81
PZE.109020361	4.13 °	3.81	3.81	3.72	3.63	3.81

Table 3.5: $-\log_{10}$(p-value) of markers detected by at least one algorithm for DMY_Flo. The star indicates if the marker is detected when a Li&Ji multiple testing procedure is used at nominal level 5%. One circle (resp. two or three circles) indicate if the marker is detected when a BH FDR multiple testing procedure is used at nominal level 20% (resp. 10% and 5%).

Moreover, the instability of the list seems greater when BH is applied, compared with Li&Ji.

GWAS with multiple variance components

We consider here the Technow dataset presented in variance component analysis. At marker ℓ there are four possible Flint $\times$ Dent allelic combinations: "00", "01", "10" and "11". The association study was performed using the following model:

$$
\begin{aligned}
&Y = 1\mu + X_\ell \beta_\ell + Z_F G_F + Z_D G_D + Z_H G_H + E \\
&G_F \sim \mathcal{N}(0, \sigma_F^2 K_F) \\
&G_D \sim \mathcal{N}(0, \sigma_D^2 K_D) \\
&G_H \sim \mathcal{N}(0, \sigma_H^2 \Phi) \\
&E \sim \mathcal{N}(0, \sigma_E^2 I) \\
&U_F \perp U_D \perp U_H \perp E
\end{aligned}
$$

where coefficients described in variance component analysis are the same and X_ℓ is the incidence matrix of the four allelic combinations and β_ℓ is a vector of effects of each

	AI-gaston	AI-ASreml	MM
GY	6.07	144.50	2.40
GM	10.98	156.68	2.44

Table 3.6: Computational time (in hour) associated to the different algorithms for the complete analysis of the data.

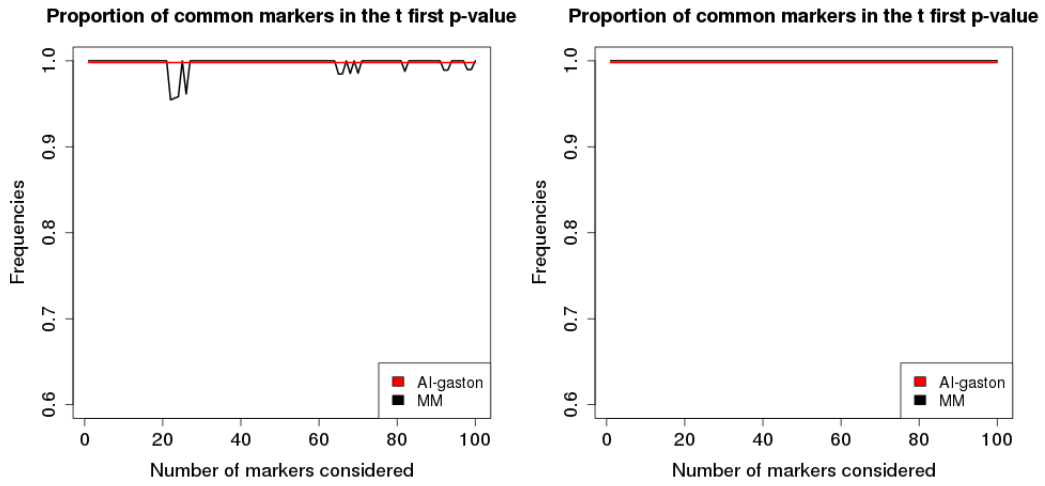


Figure 3.3: Concordance curves for AI-gaston and MM algorithms against AI-ASReml, based on the p-values obtained for the grain yield analysis (left) and grain moisture analysis (right).

combination.
As in Section 3.3.2, only 3 procedures are considered: AI-gaston, AI-ASReml and MM, AI-ASReml being considered as the reference.

Computational Time The computational times corresponding to the analysis of the 35,478 markers of hybrids by each algorithm are displayed in Table 3.6, for the grain yield (GY) and the grain moisture (GM). In this case, AI-gaston and AI-ASReml are based on the same method but the second one uses MME trick. One can see the cons about using MME trick in this analysis: computational time is 15 times longer. The MM algorithm is the fastest one for this analysis.

Comparison of the p-value orderings Figure 3.3 displays the concordance of each algorithm with AI-ASReml for GY analysis and GM analysis. For the grain moisture concordances with AI-ASReml are perfect for both algorithms. For grain yield, MM is slightly less concordant than AI-gaston with AI-ASReml.
The exact concordance of AI-gaston with AI-ASReml was expected in this study because, in this case, the two algorithms are based on the same method. Using MME trick or a direct computation of matrix P yields the same p-values.

3.4 Discussion

Computational time appeared as the most striking feature in the comparison between the different inference procedures. This high variability in performance may come as a surprise since all procedures considered here were designed for high computational efficiency. In practice, one can observe that this goal is (often) achieved when considering datasets that match the initial framework for which the procedures were developed. For instance the AI-ASReml algorithm was initially developed for the analysis of genetic data in animal breeding where relatedness coefficients are generally inferred from pedigree. These pedigree-based kinship matrices are generally sparse, which prompted the development of specific sparse matrix inversion procedures such as the one presented in Section 3.2.1. This procedure appeared orders of magnitude faster as soon as the sparsity of the kinship matrices is satisfied, as illustrated by the application to the NAM-Id dataset. Nowadays, thanks to the ever decreasing cost of the genotyping technologies, kinship can be inferred from genotypic information, leading to a more precise estimation, but yielding kinship matrices that are not sparse any longer. When applied to such data, the performance of the AI-ASReml algorithm decreases dramatically. Alternatively, FaST-LMM was developed to efficiently perform inference in mixed models without any limitation regarding the sparsity of the kinship matrix, but can only handle models with two variance components. These limitations have motivated the development of new approaches.

The MM and gaston algorithms circumvent the two limitations mentioned above, and proved to be very efficient both in the context of 2-variance component models (where the two algorithms outperform FastLMM) and in the more general case where the number of component is higher than 2. This lower computational time is explained by simpler iteration updates. This lower cost per iteration is generally counterbalanced by the requirement of a higher number of iterations to achieve convergence, but the version we implemented here benefit from the acceleration trick of Varadhan and Roland (2008) that significantly reduces the number of iterations. MM and gaston provide an attractive alternative to approximate methods, that can be quite efficient from a computational point of view but may lead to less precise p-values that may affect the QTL detection. Still, mixed model inference remains an open problem as soon as the number of observations is high and when more than one non sparse covariance matrices are considered, as illustrated by the results reported in Table 3.6.

In absence of any efficient exact procedure for the estimation in the previous configuration, one could use approximate methods to filter the candidate marker in a first step. Most of these approximate methods (see e.g. EMMAX, Kang et al. (2010)) rely on the fact that most markers have a small or even null effect. As a consequence one can use the variance estimates obtained from the null model (i.e. the model with no marker fixed effect) as relevant approximations for most markers. Still, for major QTL variance parameters may be quite different from those obtained from the null model, leading to a test statistic for the marker effect where the variance parameters are over-estimated. Therefore the filtering should be less stringent in terms of type 1 risk control in order to avoid the discarding of markers of potential interest.

The MM algorithm that we implemented here will be available soon through a R package. It will benefit from additional features, such as the possibility to test the non-genetic fixed effects, or the possibility to specify models where the marker effect can be described by more than one parameter (see Chapter 4 for an illustration of such models). For gaston package and FaSTLMM, using more than one parameter to described the marker effects needs to bypass principal functions and define incidence matrix.

Chapter 4

Detection of genes of interest in an incomplete hybrid factorial design, application to maize panel

Maize hybrids are obtained by crossing inbred lines derived from complementary heterotic groups. In this chapter, two heterotic groups were considered. Understanding the origin of phenotypic variation in such hybrids is an important issue to envisage marker based breedings strategies.

A common strategy to detect QTL for hybrid performance is to cross a population of inbred lines from one heterotic group to a same common parent derived from the other heterotic group, referred as "tester". In this case, the phenotype of a given hybrid can be associated to the genotype of its non tester parental line within the population.

This strategy enables the detection of QTL in the non tester heterotic group only, therefore detecting QTL in both groups would require two experiments (cycling on the heterotic group considered as the non tester group). In each experiment, genetic diversity is only contributed by the non tester population. As discussed in chapter 1, from a statistical point of view the data analysis can be performed using the model proposed by Yu et al. (2006) in this case. This strategy has been successfully used for productivity and related traits by Rincent et al. (2014b).

As discussed recently by Giraud et al. (2017) in the context of linkage based mapping, the use of a tester raises important issues. First, results can be dependent to a larger extent on the choice of the tester, therefore hampering a comprehensive view of the determinism of the phenotype of interest. Second, the fact that all hybrids share a common parent buffers genetic variation and restricts the number of lines that can be evaluated.

Giraud et al. (2017) have therefore advocated the study of both heterotic groups at a same time. Authors used hybrids between two sets of lines that represent the two heterotic groups. This kind of analysis allows the QTL detection in each group within a

single experiment and without selecting a tester. Moreover, a possible dominance effect between those groups could be detected in this experiment. The analysis requires to take into account two genetic diversities, each derived from a group, in the statistical modeling. Following a classical approach (Technow et al., 2014), the data can be analyzed using model (1.5) with a decomposition using the general combining ability (GCA) of each group and the specific combining ability (SCA) of the crossing between these two groups (Sprague and Tatum, 1942). These effects could be further decomposed as QTL effects and polygenic residual effects (Parisseaux and Bernardo, 2004; van Eeuwijk et al., 2010).

In this chapter, we use a panel of hybrids produced in the PIA(Programmes d'Investissements d'Avenir)/ANR project Amaizing. I will start by briefly presenting the mating design and the data available. Then I will describe models used for statistical analyses. Application of these models require the knowledge of relatedness coefficients. In practice these coefficients have to be inferred, and in this chapter inferred from bi-allelic markers, using one or an other available algorithm. Inferred values from these models will be impacted from the method chosen. One can characterize the influence of using one inference method comparing to an other one as in Rincent et al. (2014a). To this purpose, I will show results about the markers detection within each trial using relatedness inferred from Astle and Balding (2009) or from `Relatedness` (Laporte and Mary-Huard, 2017). I will discuss about the results of the detection regarding the tested hypothesis and the inference method used to estimate the relatedness parameters. Moreover, the experiment allows to take in account interaction genotype $\times$ trial (van Eeuwijk et al., 2005). so finally, I will present results about the detection by analyzing all trials jointly.

4.1 Material and Method

4.1.1 Genetic Material

Lines used for this study belong to two different populations: the flint and the dent heterotic groups. These two groups are largely used for maize production in northern Europe. Lines were crossed according to an incomplete factorial design (Comstock et al., 1952), i.e. each hybrid was derived from a cross between a flint line and a dent line. There were approximately 300 flint lines and 300 dent lines crossed to obtain 348 hybrids. Within each group, most lines were involved in a single crossing except 50 which were involved in two crosses, schematically represented in Figure 4.1.
Lines were genotyped using 600K SNP (Single Nucleotid Polymorphism) markers chip (markers are bi-allelic in this case). Missing data were imputed using Beagle (Browning and Browning, 2007) version 3.2.2. Afterwards, we used a criterion based on allele frequencies in each population: markers with a minor allele frequency lower than 4% in a given population were not considered. Note that minor allele frequency criteria does not ensure the presence of all possible allelic combinations ("00", "01", "10", "11") among hybrids. Markers where one of these combinations is not represented among hybrids were not considered furtherdown. We obtained a set of approximately 400K markers for the analyses.

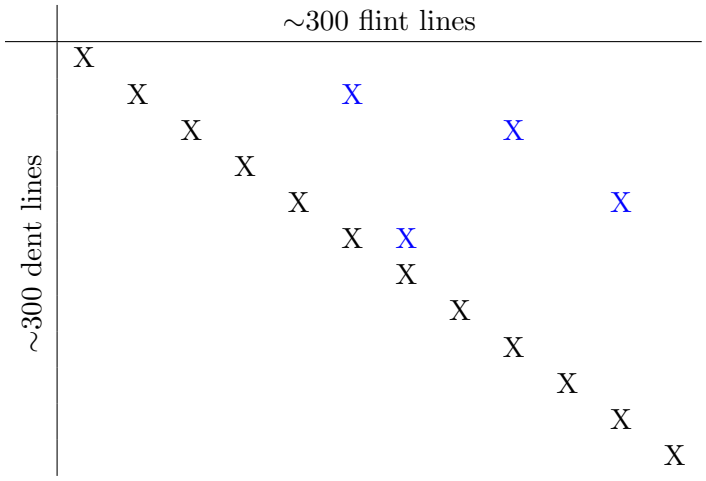


Figure 4.1: Matrix of the mating design. Blue crosses indicate hybrids with parents have been already crossed to obtain an other hybrid.

4.1.2 Phenotypic Evaluation

Phenotyping Experiment The hybrids were phenotyped two consecutive years in public and private phenotyping platforms: Saint-Martin de Hinx in 2014 (SMH14) and 2015 (SMH15), Syngenta in 2014 (SYN14), Caussade in 2014 (CAU14), Limagrain in 2015 (LMG15), Maïsadour in 2015 (MAS15), RAGT in 2015 (RAGT15) and Euralis in 2015 (EUR15), leading to a total of 8 trials. In this chapter, we will focus on two phenotypes, the anthesis (the male flowering date) and the grain yield. To avoid competition effect between hybrids with different flowering dates, hybrids were sowed within two blocks in each trial according to their flowering date evaluated in 2013. One block was composed of early and intermediate hybrids and the other one was composed of late and intermediate hybrids.

Field Effect Correction In a first step the raw data need to be corrected for field effects trial per trial. Moreover the decomposition in flowering blocks allows a correction of the raw data with respect to the flowering date of hybrids. We use the following method to correct the data. First, the following model was fitted:

$$
\begin{aligned}
&Y = \mu + X_{Flo}\beta_{Flo} + X_{Bl}\beta_{Bl} + Z_{row}R + Z_{col}C + Z_H H + E \\
&R \perp C \perp H \perp E \\
&R \sim \mathcal{N}(0, I_r\sigma_R^2) \\
&C \sim \mathcal{N}(0, I_c\sigma_C^2) \\
&H \sim \mathcal{N}(0, I_h\sigma_H^2) \\
&E \sim \mathcal{N}(0, I_n\sigma_E^2)
\end{aligned}
$$

where μ is the mean effect, X_{Flo} is the incidence matrix of the flowering group, X_{Bl} is the incidence matrix of the sawed block within the trial, Z_{row} (resp. Z_{col}) is the incidence matrix linked to the row (resp. column) in the trial, R (resp. C) is the row (resp. column) random effect , Z_H is the incidence matrix corresponding to hybrids, H

	Anthesis		Grain Yield	
	Mean	σ^2_E	Mean	σ^2_E
CAU14	192.44	0.35	85.58	88.85
SYN14	202.83	0.45	75.65	85.42
SMH14	197.50	0.40	101.49	93.46
LMG15	191.11	2.34	115.31	132.01
MAS15	197.93	3.16	88.78	95.70
RAGT15	197.22	0.97	87.84	62.64
EUR15	188.43	12.15	90.30	27.58
SMH15	195.15	1.18	86.82	151.87

Table 4.1: Table of means and residual variances among each trial of both phenotypes of interest

is the random effect associated to the polygenic background of hybrids and E is a vector of residuals. Rows and columns data correspond to spatial location of the plot where hybrids were sowed within trial. This model was inferred using ASReml (Gilmour et al., 2009) because of the sparsity of correlation matrix (see chapter 3). Once this model is adjusted, one can correct the phenotype with respect to trial effects as :

$$\tilde{Y} = Y - X_{Bl}\widehat{\beta}_{Bl} - Z_{row}\widehat{R} - Z_{col}\widehat{C}$$

Furthermore, we inferred means per trial and residual variances to filter trials. As shown in Table 4.1, the mean value of MOR14 for the anthesis is higher than the others. This trait was phenotyped lately on this trial, therefore we did not consider this trial for the anthesis. Residual variances were too high in the trial EUR15 for anthesis and too low for the grain yield. We did not consider this trial anyfurther.

From now, we rename $\tilde{Y}$ as Y. We will only use corrected phenotypes furtherdown.

4.1.3 Statistical Framework

Two types of analysis were conducted. On the one hand, we analyzed each trial separately. On the other hand, we analyzed all trials together. For each, we adjusted several models with a common random part and a variable fixed part. We defined two generic models, one for the per trial analyses and an other for the global analyses by taking all trials together. We refer to the first one by Trial model and to the second one by Global model. Firstly, we describe the two models and their random part. Secondly, we detail the inference of relatedness parameters needed in those models. Thirdly, we describe methods to infer variance components. And finally, we write tests for QTL detection in the per Trial and Global data analyses. All models here were inferred using MM algorithm presented in chapter 3 and implemented in a R package that is nearly developed.

Trial Model Because of the mating design here, model has to account for genetic variability within each population and the interaction between them. We used the

following model:

$$\begin{array}{l} Y = \{Fixed\} + Z_F U_F + Z_D U_D + Z_H U_H + E \\ U_F \perp U_D \perp U_H \perp E \\ U_F \sim \mathcal{N}(0, K_F \sigma_F^2) \\ U_D \sim \mathcal{N}(0, K_D \sigma_D^2) \\ U_H \sim \mathcal{N}(0, \Phi \sigma_H^2) \\ E \sim \mathcal{N}(0, I \sigma_E^2) \end{array} \tag{4.1}$$

where

- Y is the vector of phenotypes
- $\{Fixed\}$ represents a list of fixed effects
- Z_F (resp. Z_D and Z_H) is the incidence matrix of flint parental lines (resp. dent parental lines and hybrids)
- U_F is the polygenic effect derived from flint lines, also called General Combining Ability for flint lines (GCA_F)
- U_D is the polygenic effect derived from dent lines, also called General Combining Ability for dent lines (GCA_D)
- U_H is the polygenic effect derived from the interaction between the two populations, also called Specific Combining Ability (SCA_H)
- K_F (resp. K_D) is the simple relatedness matrix between flint (resp. dent) lines as defined in chapter 2
- Φ is the double relatedness matrix between hybrids as defined in chapter 2
- E is a vector of residuals

The full model, derived from (4.1), is composed of all possible fixed parameters. It accounts for flint (resp. dent) additive effect at a marker ℓ, quoted $\beta_{F,\ell}$ (resp. $\beta_{D,\ell}$) and the interaction effect between alleles derived from the two population at a marker ℓ, quoted $\gamma_{H,\ell}$. In this case:

$$Fixed = 1\mu + X_{F,\ell}\beta_{F,\ell} + X_{D,\ell}\beta_{D,\ell} + X_{H,\ell}\gamma_{H,\ell} \tag{4.2}$$

where

- μ is the intercept of the model
- $X_{F,\ell}$ (resp. $X_{D,\ell}$) is the genotype of flint (resp. dent) parental lines
- $X_{H,\ell}$ is a vector filled with 1 when alleles of flint and dent parental lines are different and -1 otherwise.

Note that models need to be fitted for each marker, i.e. the variances have to be estimated marker by marker.

Global Model When we study data by taking all trials, the model has to account for interaction between trial and genotype in addition to the ones already listed. The model used is quite similar as the previous one, three random terms are added, we will only detailed this three here:

$$\begin{array}{l} Y = \{Fixed\} + Z_F U_F + Z_D U_D + Z_H U_H + Z_{FT} U_{FT} + Z_{DT} U_{DT} + Z_{HT} U_{HT} + E \\ U_F \perp U_D \perp U_H \perp U_{FT} \perp U_{DT} \perp U_{HT} \perp E \\ U_{FT} \sim \mathcal{N}(0, \text{bdiag}(K_F)\sigma^2_{FT}) \\ U_{DT} \sim \mathcal{N}(0, \text{bdiag}(K_D)\sigma^2_{DT}) \\ U_{HT} \sim \mathcal{N}(0, \text{bdiag}(\Phi)\sigma^2_{HT}) \\ E \sim \mathcal{N}(0, I\sigma^2_E) \end{array} \tag{4.3}$$

where

- Z_{FT} (resp. Z_{DT} and Z_{HT}) is the incidence matrix of couples trial/flint parental lines (resp. dent parental lines and hybrids)
- U_{FT} is the random effect derived from the interaction between trials and GCA_F
- U_{DT} is the random effect derived from the interaction between trials and GCA_D
- U_{HT} is the random effect derived from the interaction between trials and SCA_H
- E is the vector of residuals

The notation bdiag(A) corresponds to:

$$\text{bdiag}(A) = \begin{pmatrix} A & 0 & 0 & ... & 0 \\ 0 & A & 0 & ... & 0 \\ 0 & 0 & A & ... & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & ... & A \end{pmatrix}$$

where the matrix A is repeated on the diagonal as times as the number of trials.

The full model is model (4.3) accounting for a global flint (resp. dent) additive effect at a marker ℓ, quoted $\beta_{F,\ell}$ (resp. $\beta_{D,\ell}$), a global interaction effect between alleles derived from the two population at a marker ℓ, quoted $\gamma_{H,\ell}$, a vector of specific flint (resp. dent) additive effects in each trial, quoted $\beta_{FT,\ell}$ (resp. $\beta_{DT,\ell}$) and a vector of specific interaction, quoted $\gamma_{HT,\ell}$. This model is model (4.3) with:

$$Fixed = 1\mu + X_{F,\ell}\beta_{F,\ell} + X_{D,\ell}\beta_{D,\ell} + X_{H,\ell}\gamma_{H,\ell} + X_{FT,\ell}\beta_{FT,\ell} + X_{DT,\ell}\beta_{DT,\ell} + X_{HT,\ell}\beta_{HT,\ell}$$

where coefficients defined in model (4.2) are the same and

- $X_{FT,\ell}$ (resp. $X_{DT,\ell}$) is the incidence matrix of couples trial/flint (resp. dent) parental line genotype
- $X_{HT,\ell}$ is the incidence matrix of couples trial/$X_{H,\ell}$

Relatedness Inference Matrix K_F, K_D and Φ involve relatedness coefficients. The mating design, detailed in subsection 4.1.1, ensures the identifiability of relatedness coefficients as explained in Section 2.2.3 because the parents of each hybrid are derived from two different populations.
We inferred these coefficients using two methods. The first one is based on Astle and Balding (2009) and will be quoted A&B method (see Chapter 2 for details). The simple relatedness between two lines i and j of the population P (flint or dent) is inferred using the following formula:

$$(\widehat{K}_P)_{i,j} = \frac{1}{L}\sum_{\ell=1}^{L} \frac{(X_{i,\ell} - p_\ell^{(P)})(X_{j,\ell} - p_\ell^{(P)})}{p_\ell^{(P)}(1 - p_\ell^{(P)})} \tag{4.4}$$

where $X_{i,\ell}$ (resp. $X_{j,\ell}$) is the genotype of the individual i (resp. j) at marker ℓ (a vector filled with 0 and 1) and $p_\ell^{(P)}$ is the allele frequency of marker ℓ in the population P inferred from individuals of the panel. Once one has inferred $\widehat{K}_F$ and $\widehat{K}_D$ using the equation (4.4), one can infer the double relatedness between hybrids $H_h = i \times j$ and $H_{h'} = i' \times j'$ to obtain the coefficient $\Phi_{h,h'}$. We used the following formula to estimate this coefficient:

$$\widehat{\Phi}_{h,h'} = (\widehat{K}_F)_{i,i'} \times (\widehat{K}_D)_{j,j'}$$

The second method to infer relatedness is using the R-package `Relatedness` as described in chapter 2.

Variance component inference One is interested here in variance component estimation, for example to study the heritability of trait. In this analysis, we use complete dataset by taking all trials. We adjust model (4.3) with:

$$Fixed = \mu + X_T\beta_T \tag{4.5}$$

According to (4.3) this model has a common residual variances for all trials. But variability of residuals with respect to trials in Table 4.1 leads to think about using Per Trial Residual Variances (PT Residual Variances). To achieve this goal, we derive model (4.5) using a vector of residuals as follow:

$$E \sim \mathcal{N}(0, \text{bdiag}((I\sigma^2_{E,t})_{1\leq t\leq N_T})) \tag{4.6}$$

where N_T is the number of trials considered and $\sigma^2_{E,t}$ is the residual variance in trial t.

Tests for QTL Detection within each Trial There are different possible effects which are interesting to explain the genotype effect on the phenotype. Each test is done using a likelihood ratio test (LRT). The ratio studied is:

$$\text{LR} = -2(\mathcal{L}_{H_0}(Y) - \mathcal{L}_{H_1}(Y))$$

where $\mathcal{L}_{H_i}(Y)$ is the complete log-likelihood under the hypothesis H_i. All models here are base on model (4.1).

Flint (resp. dent) additive effect:
Test the specific additive effect at a marker ℓ of the flint (resp. dent) population.

$$H_0 : \{Fixed = 1\mu\} \text{ vs } H_1 : \{Fixed = 1\mu + X_{F,\ell}\beta_{F,\ell}\}$$
$$H_0 : \{Fixed = 1\mu\} \text{ vs } H_1 : \{Fixed = 1\mu + X_{D,\ell}\beta_{D,\ell}\}$$

Under H_0, one has LR $\underset{+\infty}{\sim} \chi^2(1)$.

Additive effect:
Test the joint additive effect at a marker ℓ of the two populations.

$$H_0 : \{Fixed = 1\mu\} \text{ vs } H_1 : \{Fixed = 1\mu + X_{F,\ell}\beta_{F,\ell} + X_{D,\ell}\beta_{D,\ell}\}$$

Under H_0, one has LR $\underset{+\infty}{\sim} \chi^2(2)$.

Allelic interaction effect:
Test the interaction effect at a marker ℓ between alleles of the two populations.

$$H_0 : \{Fixed = 1\mu + X_{F,\ell}\beta_{F,\ell} + X_{D,\ell}\beta_{D,\ell}\} \text{ vs}$$
$$H_1 : \{Fixed = 1\mu + X_{F,\ell}\beta_{F,\ell} + X_{D,\ell}\beta_{D,\ell} + X_{H,\ell}\gamma_{H,\ell}\}$$

Under H_0, one has LR $\underset{+\infty}{\sim} \chi^2(1)$.

Full marker effect:
Test the global effect of marker ℓ.

$$H_0 : \{Fixed = 1\mu\} \text{ vs } H_1 : \{Fixed = 1\mu + X_{F,\ell}\beta_{F,\ell} + X_{D,\ell}\beta_{D,\ell} + X_{H,\ell}\gamma_{H,\ell}\}$$

Under H_0, one has LR $\underset{+\infty}{\sim} \chi^2(3)$.

Tests for QTL Detection when analyzing all trials jointly Effects studied when considering all trials together are quite similar to individual trials but, because we have access to trial effects, we study in addition the interaction between trial and genotype. All models here are based on model (4.3).

Flint (resp. dent) additive effect with trial interaction (TI):
Test jointly the global additive effect at a marker ℓ of the flint (resp. dent) population and specific additive effects within each trial.

$$H_0 : \{Fixed = 1\mu + X_T\beta_T\} \text{ vs } H_1 : \{Fixed = 1\mu + X_T\beta_T + X_{F,\ell}\beta_{F,\ell} + X_{FT,\ell}\beta_{FT,\ell}\}$$
$$H_0 : \{Fixed = 1\mu + X_T\beta_T\} \text{ vs } H_1 : \{Fixed = 1\mu + X_T\beta_T + X_{D,\ell}\beta_{D,\ell} + X_{DT,\ell}\beta_{DT,\ell}\}$$

Under H_0, one has LR $\underset{+\infty}{\sim} \chi^2(N_T)$.

Additive effect with TI:
Test jointly the global additive effect at a marker ℓ and specific additive effects within each trial.

$$H_0 : \{Fixed = 1\mu + X_T\beta_T\} \text{ vs}$$
$$H_1 : \{Fixed = 1\mu + X_T\beta_T + X_{F,\ell}\beta_{F,\ell} + X_{D,\ell}\beta_{D,\ell} + X_{FT,\ell}\beta_{FT,\ell} + X_{DT,\ell}\beta_{DT,\ell}\}$$

Under H_0, one has LR $\underset{+\infty}{\sim} \chi^2(2N_T)$.

Allelic interaction effect with TI:
Test jointly the global interaction effect at a marker ℓ between flint and dent populations and specific interaction effects within each trial.

$$\begin{aligned} H_0 : \{Fixed = \quad & 1\mu + X_T\beta_T + X_{F,\ell}\beta_{F,\ell} + X_{D,\ell}\beta_{D,\ell} + X_{FT,\ell}\beta_{FT,\ell} + X_{DT,\ell}\beta_{DT,\ell}\} \text{ vs} \\ H_1 : \{Fixed = \quad & 1\mu + X_T\beta_T + X_{F,\ell}\beta_{F,\ell} + X_{D,\ell}\beta_{D,\ell} + X_{FT,\ell}\beta_{FT,\ell} + X_{DT,\ell}\beta_{DT,\ell} \\ & + X_{H,\ell}\gamma_{H,\ell} + X_{HT,\ell}\gamma_{HT,\ell}\} \end{aligned}$$

Under H_0, one has LR $\underset{+\infty}{\sim} \chi^2(N_T)$.

Full marker effect with TI:
Test jointly the global effect of a marker ℓ and specific marker effects within each trial.

$$\begin{aligned} H_0 : \{Fixed = \quad & 1\mu + X_T\beta_T\} \text{ vs} \\ H_1 : \{Fixed = \quad & 1\mu + X_T\beta_T + X_{F,\ell}\beta_{F,\ell} + X_{D,\ell}\beta_{D,\ell} + X_{FT,\ell}\beta_{FT,\ell} + X_{DT,\ell}\beta_{DT,\ell} \\ & + X_{H,\ell}\gamma_{H,\ell} + X_{HT,\ell}\gamma_{HT,\ell}\} \end{aligned}$$

Under H_0, one has LR $\underset{+\infty}{\sim} \chi^2(3N_T)$.

4.2 Results

4.2.1 Variance Components

Inferred values of unknown variance parameters are given in Table 4.2 for anthesis and in Table 4.3 for grain yield. One can observe a preponderance of GCA effects for both traits with relatively equivalent contributions for the Flint and Dent groups. The proportion of genetic variance explained by SCA_H is quite small ($\sim 5\%$) for both traits and both relatedness inference methods. Proportions of variance explained by genetic variances (GCA_F, GCA_D and SCA_H) are slightly higher ($\sim 1\%$) using `Relatedness` than A&B method to infer relatedness coefficients.
Based on BIC values (lower is better), one can see that using A&B method is slightly better than using `Relatedness` to infer relatedness coefficients (a raise of approximately 0.5% and 0.4% for anthesis and grain yield respectively). Based on BIC values, Common Residual Variance and PT Residual Variances seem equivalent when studying grain yield. But PT Residual Variances is better than Common Residual Variance when studying anthesis (a raise of approximately 3.5% using both A&B method and `Relatedness` to infer relatedness coefficients).
Computational time of PT Residual Variance model is quite long compare to the one of Common Residual Variance model. Models for QTLs detection have to be adjusted on all available markers so computational time of one model has to be reasonable. Despite the fact that BIC values of PT Residual Variance model are better, we therefore do not consider this model furtherdown because of its associated computational time.

	Common Residual Variance		PT Residual Variances	
	A&B	**Relatedness**	A&B	**Relatedness**
GCA_F	4.49	4.94	4.56	5.11
GCA_D	3.65	3.94	3.75	4.05
SCA_H	0.51	0.48	0.69	0.60
GCA_{FT}	0.19	0.20	0.21	0.22
GCA_{DT}	0.44	0.46	0.35	0.36
SCA_{HT}	0.42	0.41	0.37	0.38
Res	1.19	1.18		
Res SMH15			0.87	0.86
Res LMG15			1.40	1.38
Res MAS15			4.39	4.37
Res RAGT15			0.44	0.43
Res SMH14			0.25	0.25
Res CAU14			0.40	0.40
BIC	8698.803	8739.158	8386.006	8430.809

Table 4.2: Variance parameters for anthesis of models (4.5) *and* (4.6) *using both A&B and* `Relatedness`*.*

	Common Residual Variance		PT Residual Variances	
	A&B	**Relatedness**	A&B	**Relatedness**
GCA_F	87.27	89.41	90.27	90.98
GCA_D	65.41	73.75	63.90	71.78
SCA_H	7.42	8.55	7.33	9.52
GCA_{FT}	26.42	28.71	22.72	24.91
GCA_{DT}	18.33	21.50	19.27	21.90
SCA_{HT}	16.03	10.78	11.35	7.06
Res	103.30	104.72		
Res SMH15			157.92	159.04
Res LMG15			173.99	175.33
Res MAS15			90.49	92.90
Res RAGT15			53.54	54.77
Res SMH14			113.14	112.11
Res CAU14			92.08	93.74
Res MOR14			77.94	78.70
BIC	20578.70	20643.57	20546.57	20611.42

Table 4.3: Variance parameters for grain yield of models (4.5) *and* (4.6) *using both A&B and* `Relatedness`*.*

4.2.2 QTL Detection per Trial

In this subsection, we study the impact of relatedness inference methods on the detection of QTL within each trial. To fulfill this purpose, we adjusted trial models where relatedness coefficients were inferred using both methods presented in Section 4.1.3.

Impact of Relatedness Inference Figure 4.2 displays QQplots of $-\log_{10}(p\text{-values})$ obtained with each method for anthesis in trial CAU14. Here we present only one trial and one phenotype because all other QQplots have the same aspect (not shown).
The $-\log_{10}(p\text{-values})$ obtained using `Relatedness` were generally higher than those obtained using A&B method and distributions of p-values obtained using `Relatedness` deviate from the "null" distribution sooner than distributions of p-values obtained using A&B. Using `Relatedness` allows to detect more markers than using A&B method. Two alternative interpretations can be made regarding these results. On the one hand, using `Relatedness` may improve the power of tests. On the other hand, using `Relatedness` to detect QTLs may increase the number of false positive discoveries.

Marker Detection within each Trial Manhattan plots are presented in Figures 4.3 and 4.4 for anthesis at CAU14 and grain yield at SMH15, respectively, using A&B method to infer relatedness. Figure 4.5 displays Manhattan plot for anthesis at CAU14 when using `Relatedness`.

The $-\log_{10}(p\text{-values})$ obtained using `Relatedness` seem globally higher than those obtained using A&B method. QTLs are not always detected for both the full marker test and a specific test. For example, in Figure 4.3, one marker is detected on chromosome 8 when one considers the flint additive effect or the global additive effect with the FDR threshold at 5%, but it is not detected when considering the full marker test. Conversely, in Figure 4.4, one marker is detected on chromosome 1 when considering the full marker test, but it is not detected with any other tests.
There are two strategies to detect QTLs. The first one is to do the full marker test and then focus on other tests to determine which effects are significant. The second one is to test flint additive, dent additive and allelic interaction effects to detect specific QTLs. As shown by Manhattan plots, in this case doing one or the other method will lead to miss some QTLs. The fact that a QTL is detected only for a specific test could be explained by the raise of degrees of freedom when one tests the full marker effect. On the opposite, a QTL detected only for the full marker test could be explained by small specific effects that become detectable once jointly accounted for. To detect all markers involved in determinism of the targeted trait, one has to study all the possible effects derived from markers.

Summary of all QTL detected in individual trials To simplify results and dimension of tables, we grouped markers into QTL with respect to their position on the genome and their correlation with other markers. Tables 4.4 and 4.5 give results for anthesis and grain yield respectively, when using A&B method to infer relatedness. Tables 4.6 and

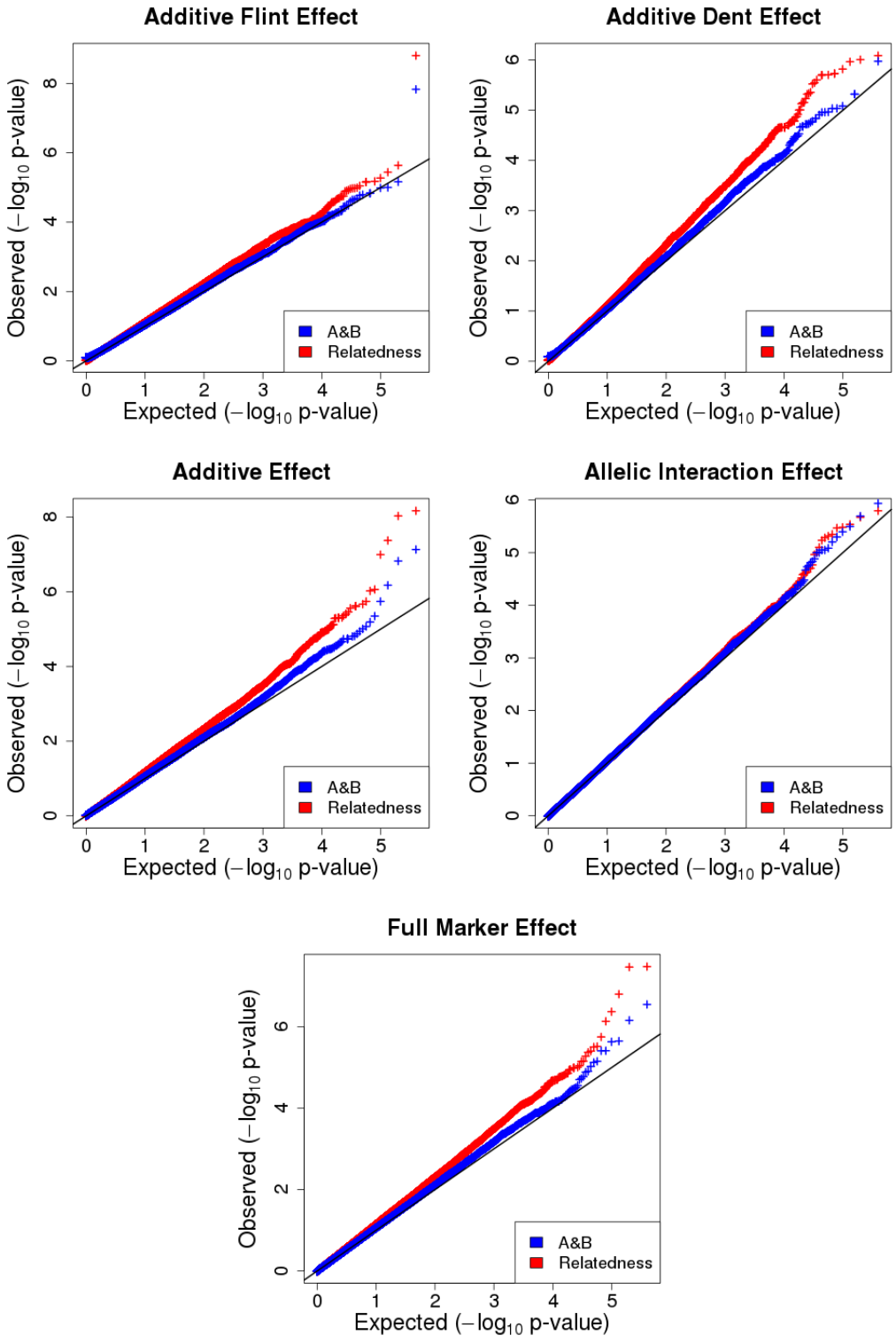


Figure 4.2: QQplots of the $-\log_{10}(p\text{-values})$ of each hypothesis tested for anthesis in the trial CAU14.

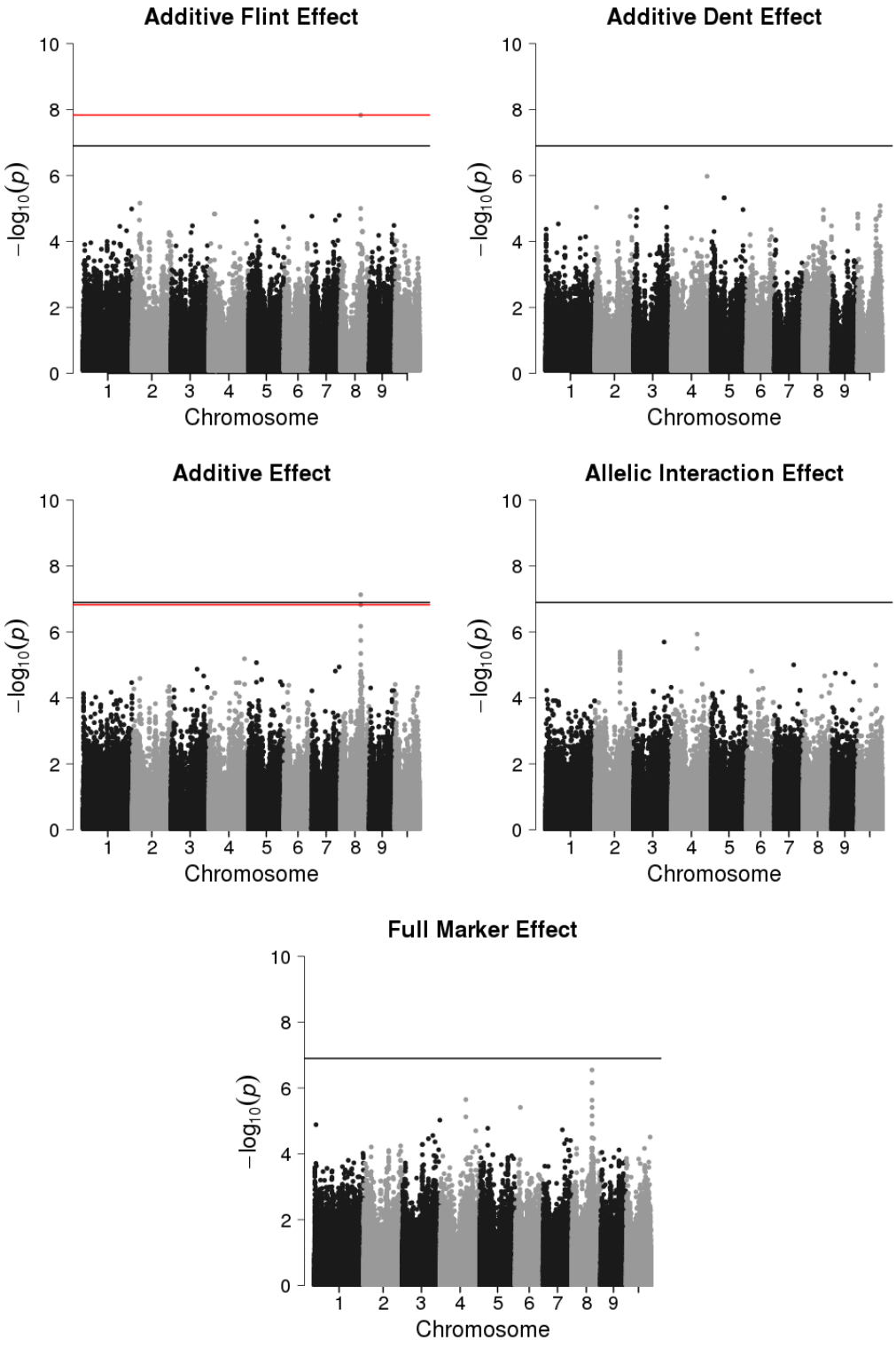


Figure 4.3: Manhattan plots of $-\log_{10}(p\text{-values})$ for the anthesis in trial CAU14, using A&B method to infer relatedness coefficients. Black line corresponds to the bonferroni threshold. Red line corresponds to the minimum $-\log_{10}(p\text{-value})$ higher than $-\log_{10}(T)$ with T the FDR threshold at 5%.

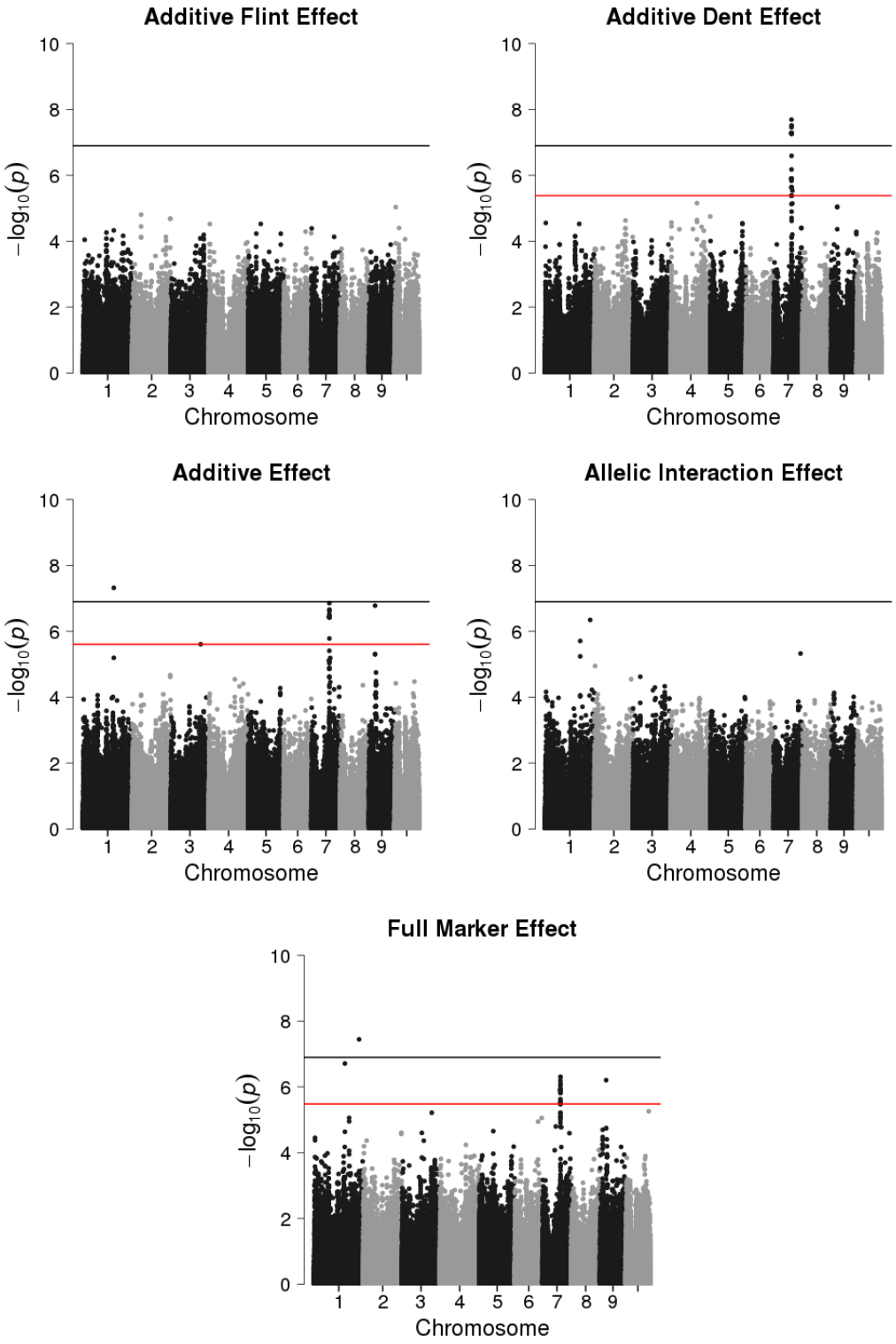


Figure 4.4: Manhattan plots of $-\log_{10}(p\text{-values})$ *for the grain yield in trial SMH15, using A&B method to infer relatedness coefficients. Black line corresponds to the bonferroni threshold. Red line corresponds to the minimum* $-\log_{10}(p\text{-value})$ *higher than* $-\log_{10}(T)$ *with* T *the FDR threshold at 5%.*

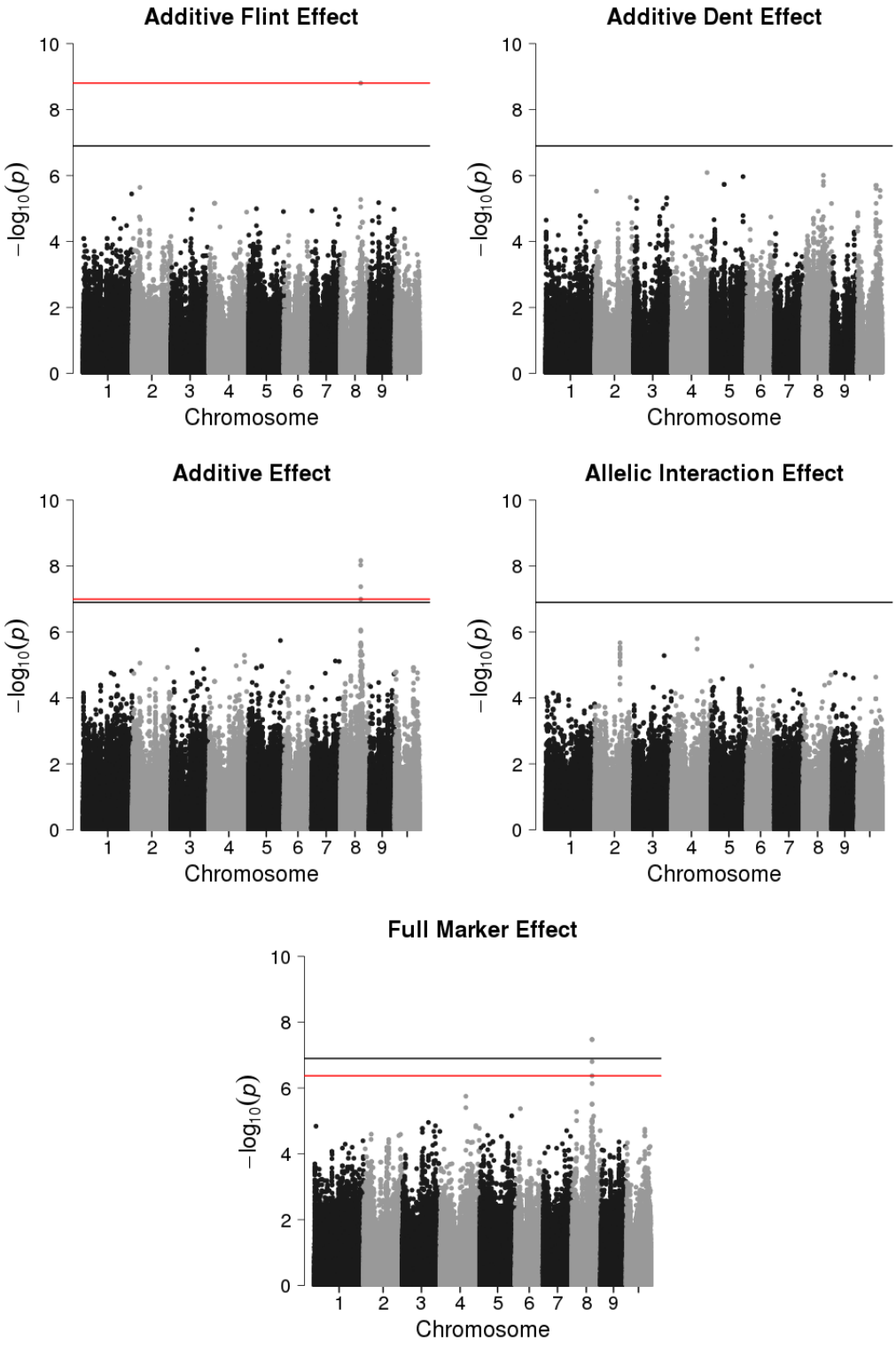


Figure 4.5: Manhattan plots of $-\log_{10}(p\text{-values})$ for the anthesis in trial CAU14, using `Relatedness` *to infer relatedness coefficients. Black line corresponds to the bonferroni threshold. Red line corresponds to the minimum $-\log_{10}(p\text{-value})$ higher than $-\log_{10}(T)$ with T the FDR threshold at 5%.*

	Chr	MinPos	MaxPos	SpikePos	Pval	Type	Trial
QTL1	3	207227085	207273133	207250092	6.60e-07	Int	MAS15
QTL2	4	33073973	34691176	33882328	5.55e-08	AddF;Add	SMH14;LMG15;MAS15
QTL3	4	35498565	37115386	36306920	7.40e-07	Add	MAS15
QTL4	5	22977027	23058617	23017884	6.77e-07	Int	MAS15
QTL5	5	137239697	138591297	137919829	6.99e-07	Add	MAS15
QTL6	8	122931493	122967814	122949646	1.47e-08	AddF;Add	CAU14
QTL7	8	123486667	123523088	123504889	1.51e-07	Add	CAU14
QTL8	9	118682726	118752544	118717549	1.29e-07	Int	MAS15

Table 4.4: Summary of all QTLs detected in individual trial analyses for the anthesis, using A&B method to infer relatedness coefficients. Chr is the chromosome number, MinPos (resp. MaxPos) is the minimum (resp. maximum) physical position of the QTL, SpikePos is the physical position of the marker associated to the lowest p-value, Pval is the p-value of the marker located in SpikePos, Type is the tested hypotheses which led to the detection, Trial is the list of trials where the QTL is detected.

	Chr	MinPos	MaxPos	SpikePos	Pval	Type	Trial
QTL1	1	187925319	188103693	188014659	4.78e-08	Add;Marker	SMH15
QTL2	1	275024696	275155215	275090036	3.58e-08	Marker	SMH15
QTL3	3	181856980	181899492	181878262	2.47e-06	Add	SMH15
QTL4	5	188687491	188736959	188712217	9.48e-08	AddD	SMH14
QTL5	7	108999484	110601741	109740799	2.03e-08	AddD;Add;Marker	SMH15
QTL6	7	114737557	115831136	115284347	2.91e-06	AddD	SMH15
QTL7	8	129936264	130002097	129978850	1.72e-07	Int	LMG15
QTL8	8	136032803	137312565	136713760	1.89e-07	AddD;Marker	RAGT15
QTL9	9	36510612	38307857	37409235	1.66e-07	Add;Marker	SMH15
QTL10	10	24440373	26069606	25254990	1.24e-07	Marker	MOR14

Table 4.5: Summary of all QTLs detected in individual trial analyses for the grain yield, using A&B method to infer relatedness coefficients. Chr is the chromosome number, MinPos (resp. MaxPos) is the minimum (resp. maximum) physical position of the QTL, SpikePos is the physical position of the marker associated to the lowest p-value, Pval is the p-value of the marker located in SpikePos, Type is the tested hypotheses which led to the detection, Trial is the list of trials where the QTL is detected.

4.7 give results for anthesis and grain yield respectively, when using `Relatedness`.

All QTLs detected when using A&B are also detected when using `Relatedness`. Except one QTL, all are detected with `Relatedness` for same hypotheses or supplementary hypotheses related to those detected with A&B method. For example, QTL3 in Table 4.4 is detected for a additive effect whereas the same QTL in Table 4.6 (QTL21) is detected for flint additive effect, additive effect and full marker effect. There is only one exception for grain yield: QTL8 in Table 4.5 is detected for dent additive effect and full marker effect when using A&B whereas QTL14 in Table 4.7 is detected for additive effect and full marker effect but not for dent additive effect when using `Relatedness`. The number of QTLs detected when using A&B method is 8 for the anthesis and 10 for grain yield, whereas the number of QTLs detected when using `Relatedness` is 40 and 17 respectively. In both cases, we detect a QTL already reported in the literature and that corresponds to a major QTL for flowering time "vgt2" (QTL7 in Table 4.4 and QTL33 in Table 4.6). Both inference methods lead to detect principally flint additive effect,

	Chr	MinPos	MaxPos	SpikePos	Pval	Type	Trial
QTL1	1	29759110	29889994	29824504	1.66e-06	Int	MAS15
QTL2	1	51846200	52458709	52152455	4.00e-06	Int	MAS15
QTL3	1	72139068	72330839	72235348	1.42e-06	AddF	MAS15
QTL4	1	187925319	188103693	188014659	5.20e-07	Add	SMH15
QTL5	1	249481328	250188927	249860569	1.14e-06	Add	SMH15
QTL6	1	296537860	296604671	296571258	9.21e-07	AddF;Add	LMG15
QTL7	2	29962136	30039666	30000856	3.17e-06	Int	MAS15
QTL8	2	34394474	34474721	34434499	1.62e-06	Marker	MAS15
QTL9	2	42755878	42847764	42801764	6.60e-07	AddF;Add	LMG15
QTL10	2	183397532	183418800	183408153	2.05e-06	Add	LMG15
QTL11	2	185645697	185666158	185655920	2.57e-06	Add;Marker	MAS15
QTL12	2	214990030	215033365	215011710	3.90e-07	Add	SMH14;SMH15
QTL13	3	1530506	1550179	1541536	3.67e-07	Marker	MAS15
QTL14	3	32711250	34404141	33557696	1.35e-06	AddF	MAS15
QTL15	3	121214088	122906979	122060534	3.30e-07	AddF	SMH14
QTL16	3	169062211	169099373	169080793	7.22e-07	AddF	MAS15
QTL17	3	199279355	199333417	199306388	1.92e-06	Add	SMH15
QTL18	3	207227085	207273133	207250148	4.29e-07	Int;Marker	MAS15
QTL19	4	11565938	11587401	11576663	3.79e-06	Int	MAS15
QTL20	4	32886547	34691176	33882328	6.00e-09	AddF;Add	SMH14;RAGT15;LMG15;MAS15
QTL21	4	35498565	37115386	36306920	2.52e-07	AddF;Add;Marker	MAS15
QTL22	4	216614568	218231277	217422923	3.53e-06	Add	LMG15
QTL23	5	22977027	23058617	23017932	1.21e-06	Int	MAS15
QTL24	5	119491999	121109100	120178761	2.27e-06	Int;Marker	MAS15
QTL25	5	137239697	138594982	137898176	1.28e-06	AddF;Add;Marker	MAS15
QTL26	5	186933193	186986427	186959785	3.11e-06	Int;Marker	MAS15
QTL27	5	195282149	195325351	195303760	4.69e-07	Add	LMG15
QTL28	7	168081333	168087858	168084596	6.37e-07	AddF;Add	LMG15
QTL29	8	28249381	29243930	28746656	5.88e-06	Add	SMH15
QTL30	8	122931493	122967814	122949646	1.57e-09	AddF;Add;Marker	CAU14;SMH14;LMG15;SMH15
QTL31	8	123190729	123210729	123200729	6.33e-07	Add;Marker	LMG15;MAS15
QTL32	8	123254290	123290662	123272484	4.24e-08	Add;Marker	CAU14;SMH15;LMG15;MAS15
QTL33	8	123486131	123530355	123504889	6.81e-09	Add;Marker	CAU14;SMH14;SMH15;LMG15;MAS15
QTL34	8	133024400	133044400	133034400	3.00e-06	Add	LMG15
QTL35	9	57736343	59751203	58852581	2.58e-07	AddF;Add	LMG15
QTL36	9	118682726	118752544	118717549	4.08e-08	Int;Marker	MAS15
QTL37	9	136536072	136572665	136554345	1.24e-07	AddD;Add;Marker	MAS15
QTL38	9	142733128	142760181	142746618	2.19e-06	Add	SMH15
QTL39	9	150819939	150831476	150825715	1.37e-06	AddF	LMG15
QTL40	10	138823997	138834442	138829251	7.92e-08	Int	SMH14

Table 4.6: Summary of all QTLs detected in individual trial analyses for the anthesis, using `Relatedness` *to infer relatedness coefficients. Chr is the chromosome number, MinPos (resp. MaxPos) is the minimum (resp. maximum) physical position of the QTL, SpikePos is the physical position of the marker associated to the lowest p-value, Pval is the p-value of the marker located in SpikePos, Type is the tested hypotheses which led to the detection, Trial is the list of trials where the QTL is detected.*

	Chr	MinPos	MaxPos	SpikePos	Pval	Type	Trial
QTL1	1	9410068	9460154	9435138	2.19e-07	Add;Marker	RAGT15
QTL2	1	179173582	179302324	179238996	2.86e-06	Add	RAGT15
QTL3	1	187925319	188179616	188014659	1.08e-08	Add;Marker	SMH15
QTL4	1	214280372	214383338	214331897	4.86e-06	Marker	SMH15
QTL5	1	215146259	215246254	215196304	6.58e-07	Add;Marker	RAGT15
QTL6	1	251315626	251543607	251429693	2.26e-07	Marker	MOR14
QTL7	1	275024696	275155215	275090036	1.45e-08	Marker	SMH15
QTL8	3	181856980	181899492	181878262	1.78e-06	Add;Marker	SMH15
QTL9	5	188687491	188736959	188712217	7.33e-09	AddD;Add	SMH14
QTL10	7	108999484	110601741	109740799	1.42e-08	AddD;Add;Marker	SMH15
QTL11	7	114715042	115831136	115284347	1.11e-06	AddD;Add;Marker	SMH15
QTL12	8	2412401	2438579	2425454	3.78e-07	Int	LMG15
QTL13	8	129936264	130002097	129978850	1.22e-07	Int	LMG15
QTL14	8	136032803	137312565	136716294	9.07e-07	Add;Marker	RAGT15
QTL15	9	36510612	38410771	37409235	4.73e-08	Add;Marker	SMH15
QTL16	10	24440373	26069606	25254990	2.71e-08	Marker	MOR14
QTL17	10	141956077	141976077	141966077	5.13e-06	Marker	SMH15

Table 4.7: Summary of all QTLs detected in individual trial analyses for the grain yield, using `Relatedness` *to infer relatedness coefficients. Chr is the chromosome number, MinPos (resp. MaxPos) is the minimum (resp. maximum) physical position of the QTL, SpikePos is the physical position of the marker associated to the lowest p-value, Pval is the p-value of the marker located in SpikePos, Type is the tested hypotheses which led to the detection, Trial is the list of trials where the QTL is detected.*

additive effect or allelic interaction effect for anthesis and dent additive effect, additive effect, full marker effect or allelic interaction effect for grain yield. In this experiment, hybrids are derived from crossing flint lines, to bring precocity, and dent lines, to bring productivity. This is consistent with the type of effects detected for each trait.

Common QTLs detected using A&B method and `Relatedness` could indicate a raise of power when using the second method to infer relatedness, because new effects detected when using `Relatedness` are related with the ones detected when using A&B method. The proportion of allelic interaction QTLs is quite small and consistent with the SCA_H variances inferred in subsection 4.2.1, which were small too. QTLs are mostly detected in only one trial. QTLs can be further decomposed in two groups. The first one is composed of QTLs detected in one trial and with sub-significant effect in other trials. The second one is composed of QTLs detected in one trial and with no significant effects in other trials. A joint analysis of all trials could improve the power of detection for the first group and accounting for interaction between trials and genotype could improve the power of detection for the second group.

4.2.3 QTL Detection by joint analysis of all trials

The goal of this subsection is to study the impact of using all trials on QTL detection. To test our method and because of computational time, we only analyzed anthesis in this part. When using all trials, for some markers, the matrix X combining all incidence matrices for fixed effects did not have the expected rank. The corresponding markers were not considered for the analyses.

	Common Residual Variance	PT Residual Variances
$M_{F,5}$	6.69e-11	3.51e-09
$M_{D,10}$	2.13e-11	5.86e-14
$M_{A,6}$	1.44e-10	4.40e-07

Table 4.8: p-values associated to markers $M_{F,5}$, $M_{D,10}$ and $M_{A,6}$ with Common Residual Variance and PT Residual Variances

Impact of relatedness inference method Figure 4.6 displays QQplots of $-\log_{10}(p$-values) of each tested hypothesis using A&B method and `Relatedness`. Contrary to results in QTL detection within each trial the two curves are similar and both of them diverge from the first bisector quite soon. One can see in Figures 4.7 and 4.8 the distribution of p-values with respect to tested hypotheses using A&B and `Relatedness`. The distribution of p-values for both methods seems sub-uniform. Note that, the correlation between markers is not accounted for in this analysis. The non-independence of tests may contribute to the shape of histograms.

Impact of analyzing jointly all trials Figure 4.9 displays the Manhattan plots of $-\log_{10}(p$-values) for each hypothesis tested when using A&B method to infer relatedness coefficients. One can see that $-\log_{10}(p$-values) values are higher than in Manhattan plots 4.3. New regions are detected through these analyses, like on chromosomes 5, 10 and 6 for the additive flint TI effect, dent flint TI effect and additive TI effect. This could be due to a strong interaction effect between trials and genotype or to the fact that effects within each trial were almost significant and their combination increases the power of detection.

To simplify notations, most significant markers of regions on chromosomes 5, 10 and 6 are quoted $M_{F,5}$, $M_{D,10}$ and $M_{A,6}$, respectively. Figure 4.10 displays boxplots of inferred fixed values for the three markers. In these three graphics, one can see that changing allele values from 0 to 1 leads to increase flowering dates in each trial except in trial MAS15. The inversion or absence of effect in this particular trial increases the effect of interaction between trials and genotype and therefore increases the significativity of tests.

Analysis of MAS15 impact These results prompted us to further investigate the specificity of trial MAS15. As displayed in Table 4.2, residual variances vary with respect to trials, and the residual variance of MAS15 is the highest. We considered two methods to correct the effect of MAS15. The first method was to use per trial residual variances in each trial as in Section 4.1.3. The second method was simply to dismiss the trial MAS15.

When using PT Residual Variances, p-values associated to $M_{F,5}$ and $M_{A,6}$ increased whereas p-value associated to $M_{D,10}$ decreased as shown in Table 4.8. The decrease of p-value associated to $M_{D,10}$ could be due to a stronger dent additive effect within trials with the lowest error variances.

Table 4.9 displays the impact of not considering trial MAS15 on p-values associated

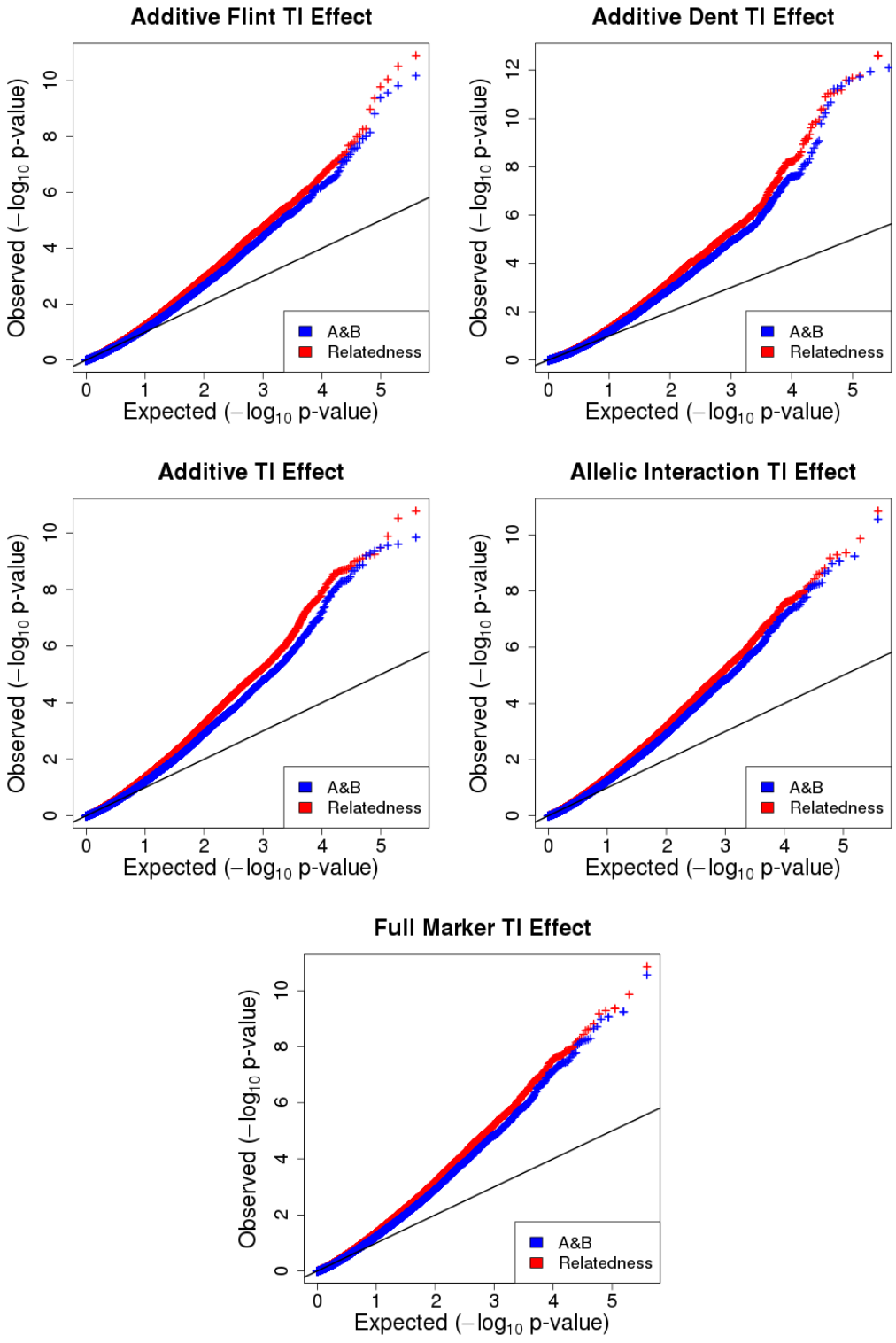


Figure 4.6: QQplots of the $-\log_{10}$(p-values) of each hypothesis tested for anthesis by analyzing all trials jointly, with both methods to infer relatedness coefficients.

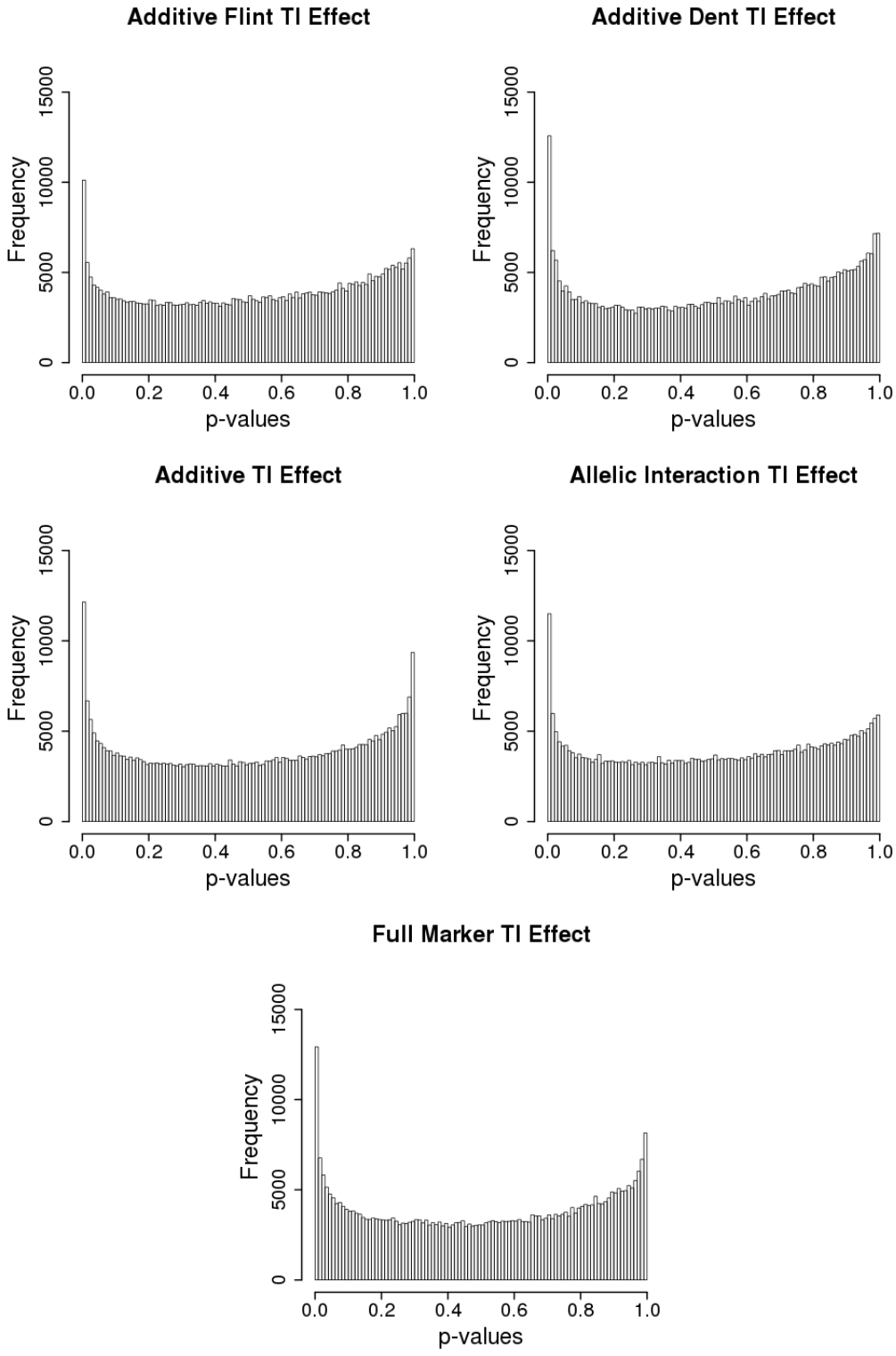


Figure 4.7: Histograms of the p-values of each hypothesis tested for anthesis by analyzing all trials jointly, using A&B method to infer relatedness coefficients.

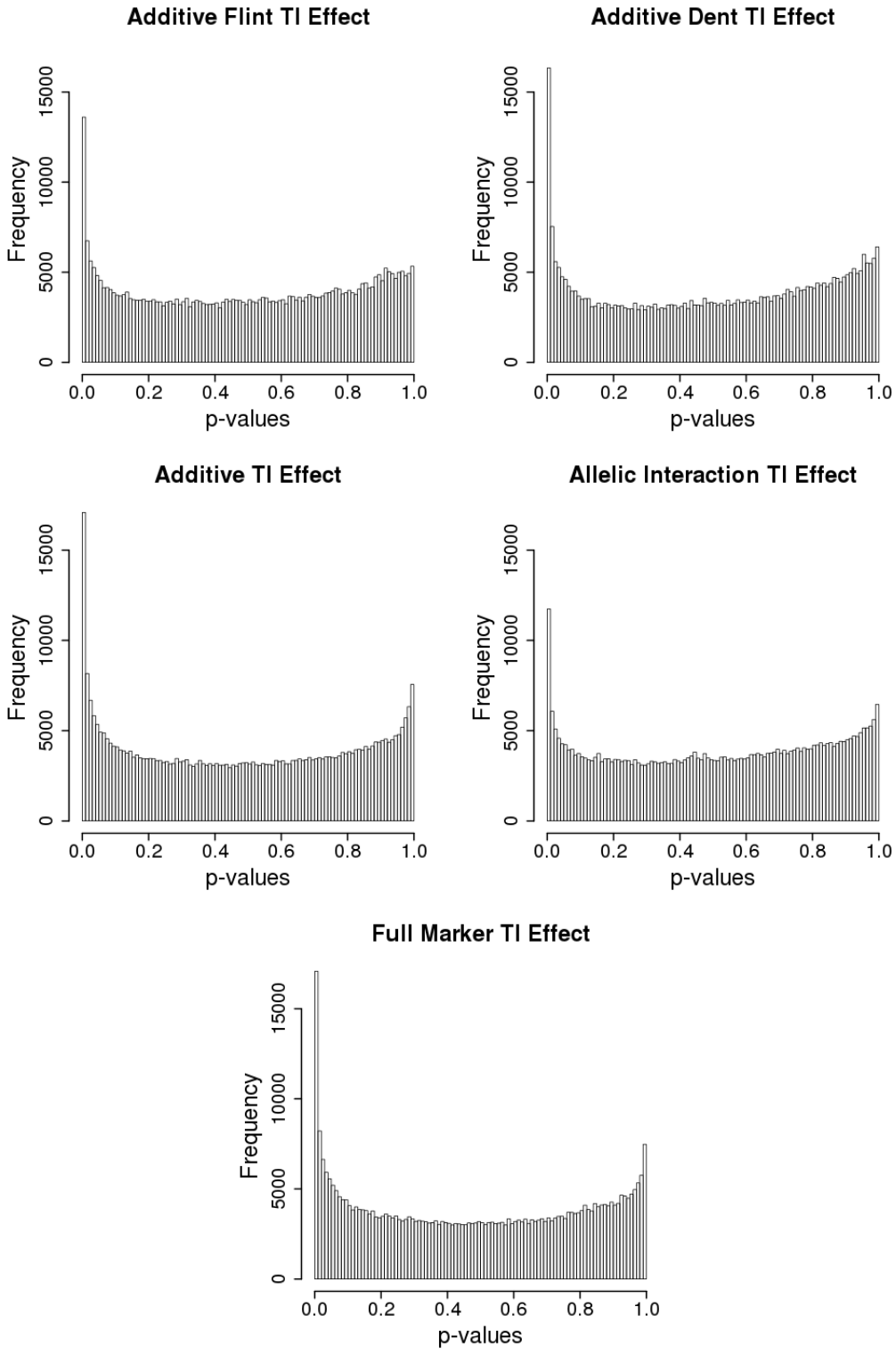


Figure 4.8: Histograms of the p-values of each hypothesis tested for anthesis by analyzing all trials jointly, using `Relatedness` *to infer relatedness coefficients.*

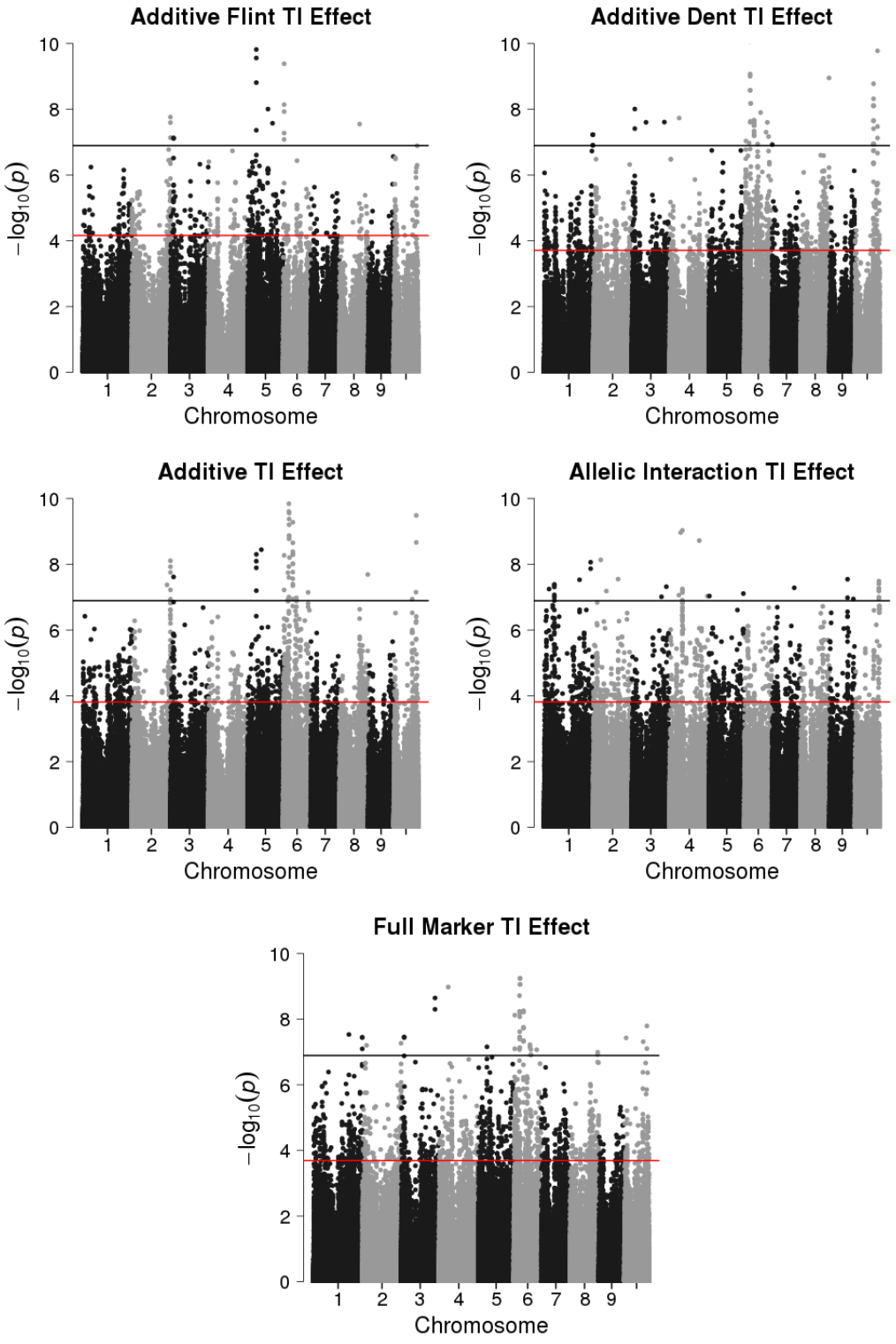


Figure 4.9: Manhattan plots of $-\log_{10}(p\text{-values})$ *for the anthesis by analyzing all trials jointly, using A&B method to infer relatedness coefficients. Black line corresponds to the bonferroni threshold. Red line corresponds to the minimum* $-\log_{10}(p\text{-value})$ *higher than* $-\log_{10}(T)$ *with* T *the FDR threshold at 5%.*

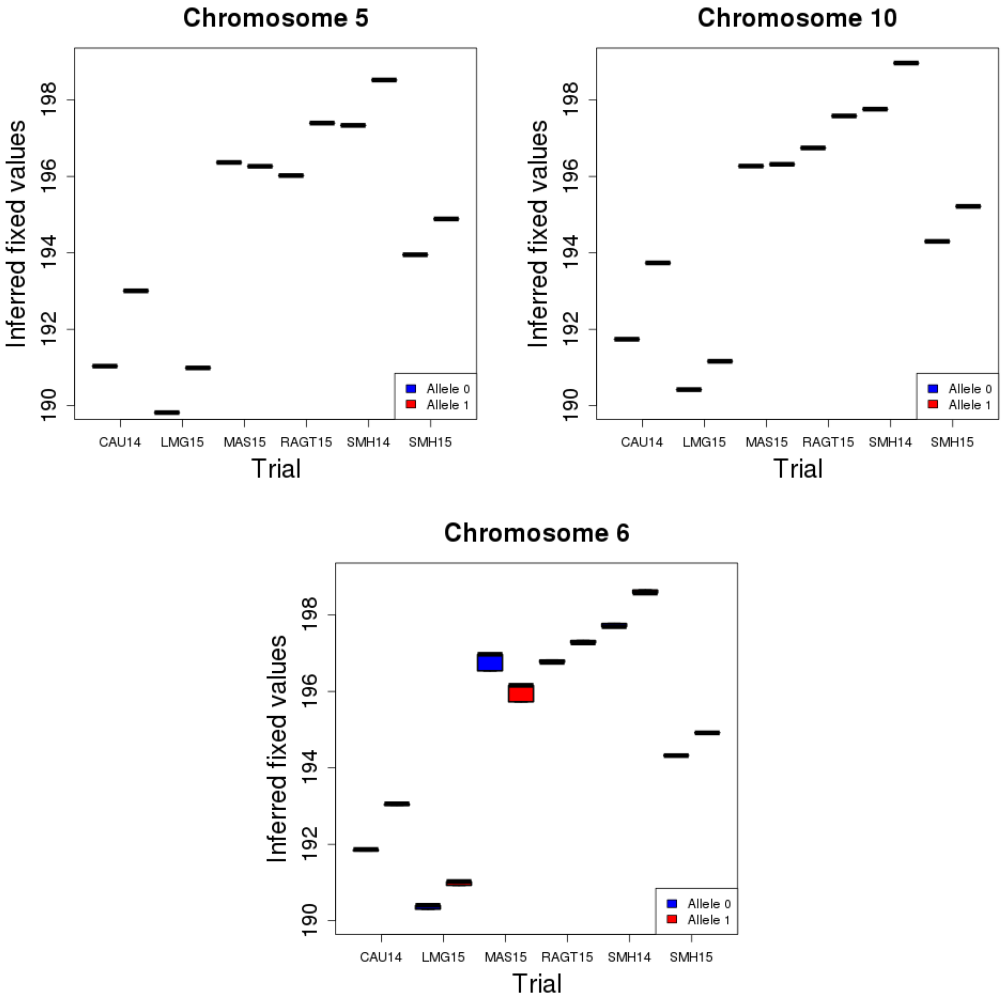


Figure 4.10: Boxplots of inferred fixed values for $M_{F,5}$ (topleft), $M_{D,10}$ (topright) and $M_{A,6}$ (bottom) with respect to trials and genotype of flint, dent and dent lines, respectively.

	With MAS15	Without MAS15
$M_{F,5}$	6.69e-11	3.57e-05
$M_{D,10}$	2.13e-11	1.24e-10
$M_{A,6}$	1.44e-10	5.25e-2

Table 4.9: p-values associated to markers $M_{F,5}$, $M_{D,10}$ and $M_{A,6}$ with and without considering trial MAS15

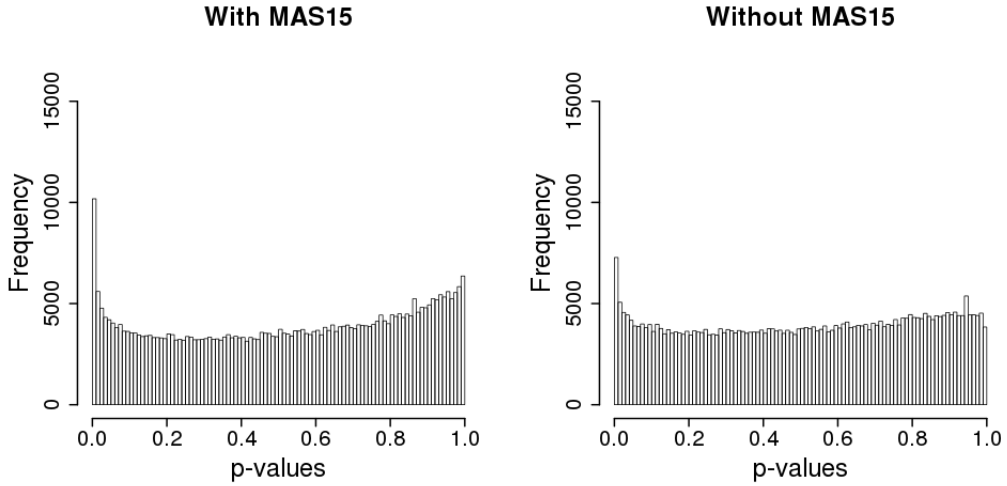


Figure 4.11: Histograms of p-values associated to flint additive TI test with trial MAS15 (left) and without trial MAS15 (right) and with A&B method to infer relatedness coefficients

to markers $M_{F,5}$, $M_{D,10}$ and $M_{A,6}$. In this case, the three p-values increased and two of them became non-significant.

QTL detection for flint additive TI effect without MAS15 The strong impact associated with the inclusion/exclusion of trial MAS15 on the p-values for these three loci prompted us to further investigate the impact of excluding trial MAS15. We tested only the flint additive TI effect on all genome without MAS15 to compare results.
One can see in Figures 4.11 and 4.12 that excluding trial MAS15 leads to a distribution of p-values closer to that expected under the null hypothesis. We considered that this could indicate an error in data from trial MAS15. We looked more carefully at the distribution of flowering date for hybrids in this trial. In Figure 4.13, the distribution seems bimodal whereas unimodal distributions were observed in other trials. Going back to the data, we identified for this trial an error in the preprocessing of the data explained in Section 4.1.2. Hybrids were wrongly assigned to flowering groups, so that the field correction was not correctly performed. The error impacts particularly results about anthesis. Considering this error, we display results about variance component inference and QTL detection within each trial without considering trial MAS15 in Appendix A.

Figure 4.14 displays Manhattan plot of $-\log_{10}(p\text{-values})$ of the test. One can see that values are smaller than those in Figure 4.3 when considering MAS15. Fewer

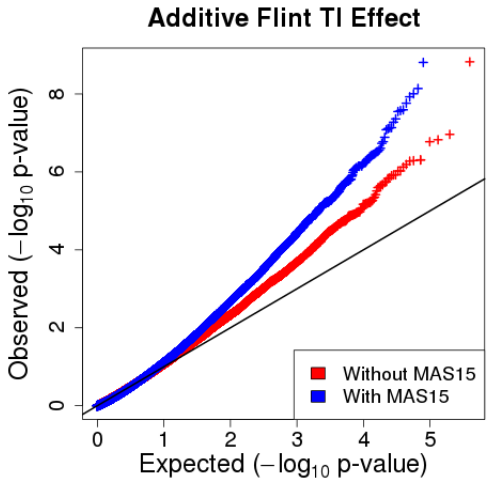


Figure 4.12: QQplot of the $-\log_{10}(p\text{-value})$ *for flint additive TI effect for anthesis by analyzing all trials jointly, using A&B method to infer relatedness coefficients (with and without trial MAS15).*

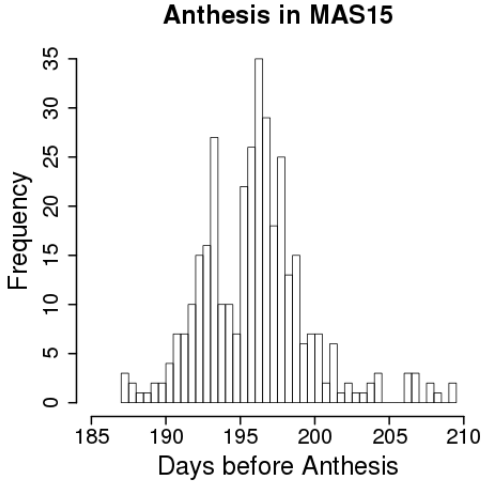


Figure 4.13: Histogram of flowering date of hybrids in trial MAS15.

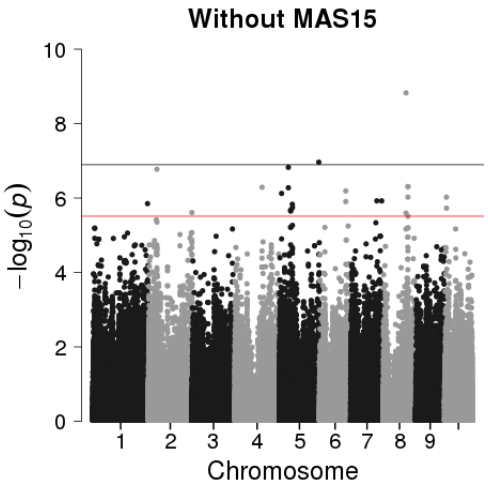


Figure 4.14: Manhattan plot of $-\log_{10}(p\text{-values})$ *associated to the additive flint TI effect, without trial MAS15, for the anthesis, using A&B method to infer relatedness coefficients. Black line corresponds to the bonferroni threshold. Red line corresponds to the minimum* $-\log_{10}(p\text{-value})$ *higher than* $-\log_{10}(T)$ *with T the FDR threshold at 5%.*

regions are detected. Regions are grouped as in subsection 4.2.2 and results are displayed in Table 4.10. Table 4.4 includes QTL detected in trial MAS15. To facilitate the comparison, we rebuilt this table without the trial MAS15 in Table 4.14 and only with the flint additive effect.
The number of QTLs detected with the joint analysis of all trials is higher than the number of QTLs detected with analyses within each trial. Note however, that QTL1 in 4.10 was detected when analyzing separately each trial with `Relatedness` to infer relatedness coefficients.
More surprisingly, only one of the two QTLs (in vgt2 region on Chromosome 8) detected with individual trial analyses was detected by analyzing all trials together. Despite the detection of QTL1 in Table 4.11 in two trials, this QTL was not detected when taking all trials together. But the QTL2 is detected when taking all trials.
If one looks at the boxplots of inferred fixed values with respect to the flint allele in Figure 4.15, the effects of the two markers in the different trials appear quite similar but one can observe a slight tendency towards more stable effects for QTL1, suggesting a lower interaction for QTL1 than for QTL2. Considering the number of degrees of freedom added to integrate, this may explain that QTL2 is detected whereas QTL1 is not. Such differences in the behaviour of QTL in single and multiple environment analyses have been reported previously by Moreau et al. (2004) in the context of linkage QTL mapping.

4.3 Conclusions

Inferring relatedness coefficients using `Relatedness` leads to detect more regions but it is not possible to tell on the basis of present results if these regions are real QTLs or false positives. To investigate this point, one could use simulated data, as in Rincent

	Chr	MinPos	MaxPos	SpikePos	Pval	Type
QTL1	1	296537860	296604671	296571258	1.41e-06	AddF TI
QTL2	2	45936108	46042320	45989148	1.68e-07	AddF TI
QTL3	2	235839119	235865570	235852341	2.45e-06	AddF TI
QTL4	4	146320528	147937237	147128883	5.13e-07	AddF TI
QTL5	5	11773790	11801440	11787642	7.51e-07	AddF TI
QTL6	5	47971637	49288925	48630447	1.50e-07	AddF TI
QTL7	5	61166367	61271414	61215284	2.13e-06	AddF TI
QTL8	5	69494886	71542501	70153365	1.47e-06	AddF TI
QTL9	5	213725489	213738702	213732099	1.09e-07	AddF TI
QTL10	6	141925997	142763094	142344596	6.44e-07	AddF TI
QTL11	7	141637892	141675462	141656663	1.19e-06	AddF TI
QTL12	7	168081333	168087858	168084596	1.19e-06	AddF TI
QTL13	8	122931493	122969131	122949646	1.49e-09	AddF TI
QTL14	8	133512273	133610933	133560305	4.92e-07	AddF TI
QTL15	10	9888472	9906991	9897742	9.46e-07	AddF TI
QTL16	10	10572689	10592611	10582540	1.87e-06	AddF TI

Table 4.10: Summary of all QTLs detected in joint analysis of all trials, except trial MAS15, for the anthesis and only for flint additive TI effect, using A&B method to infer relatedness coefficients. Chr is the chromosome number, MinPos (resp. MaxPos) is the minimum (resp. maximum) physical position of the QTL, SpikePos is the physical position of the marker associated to the lowest p-value, Pval is the p-value of the marker located in SpikePos, Type is the tested hypotheses which led to the detection.

	Chromosome	MinPos	MaxPos	SpikePos	Pval	Type	Trial
QTL1	4	33073973	34691176	33882328	5.55e-08	AddF;Add	SMH14;LMG15
QTL2	8	122931493	122967814	122949646	1.47e-08	AddF;Add	CAU14

Table 4.11: Summary of all QTLs detected in individual trial analyses, except trial MAS15, for the anthesis and only for additive flint effect, using A&B method to infer relatedness coefficients. Chr is the chromosome number, MinPos (resp. MaxPos) is the minimum (resp. maximum) physical position of the QTL, SpikePos is the physical position of the marker associated to the lowest p-value, Pval is the p-value of the marker located in SpikePos, Type is the tested hypotheses which led to the detection, Trial is the list of trials where the QTL is detected.

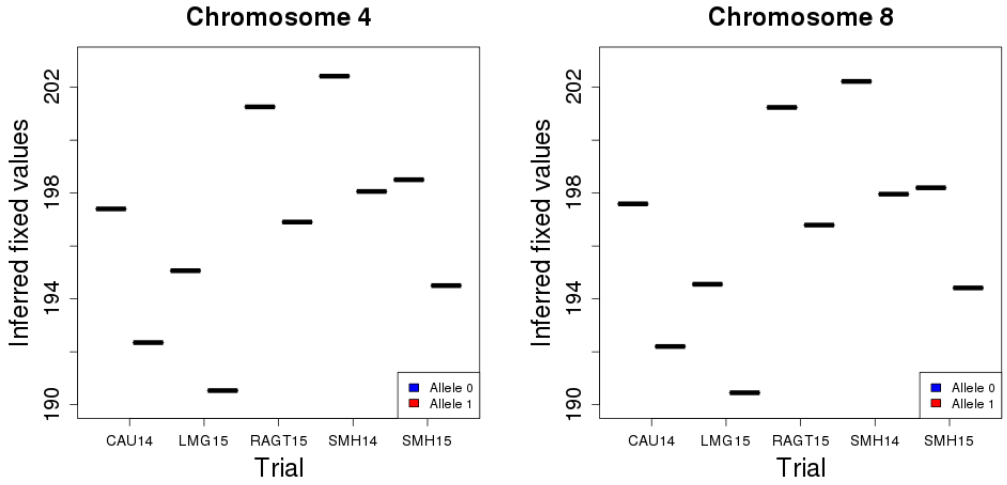


Figure 4.15: Boxplots of inferred fixed values with respect to trials and genotype of flint lines.

et al. (2014a), to evaluate the balance between power and type I risk control when using `Relatedness` to infer relatedness coefficients. Furthermore, one could compare QTL detection using `Relatedness` to the method proposed by Rincent et al. (2014a) which infers kinship matrices chromosome by chromosome. For analysis of a marker ℓ, the kinship matrix used in mixed model is inferred using all markers except those on the same chromosome than marker ℓ. This aims at avoiding "proximal contamination" that leads to a loss of power due to the colinearity between the marker tested as fixed effects and those considered to estimate the kinship. One could also compare `Relatedness` to the method proposed by Vitezica et al. (2013) for double relatedness coefficients.

At the biological level, our results bring information regarding genotype x environment interaction, additive vs. interaction allelic effects and the combination of both. Analyses in individual trials suggest that QTL are mostly specific to one trial. In addition, considering all trials jointly leads to detect more QTLs than within each location. This supports the hypothesis that the expression of genes depends to a large extent on the environment and that although variable across locations, there are trends in effects that contribute to a higher power of the global analysis. It would be interesting to study environment parameters to understand the impact of environmental effects on the expression of alleles. Note however that the problem encountered with trial MAS15 suggests that tests involving a QTL x environment interaction can be highly sensitive to errors in the initial data, which therefore require a specific attention.
QTLs in individual trial analyses are mostly detected for additive effects (flint, dent or global). There are few QTLs detected for allelic interaction effect. This results is consistent the one of Giraud et al. (2017) who conducted a linkage based QTL detection in hybrids between the two same genetic groups (flint and dent) and also mostly found additive QTL effects. It supports the idea that this organization of diversity into heterotic groups is an efficient way to limit allelic interaction effects, which is useful for the practical management of breeding programs.

Altogether, results from this experiment (see below) suggest that, to detect QTLs involved in the determinism of a targeted trait, one has to test different hypotheses, accounting for different biological effects. For instance, it would be interesting to test only additive effects with no interaction with the environment in the joint analysis. This may lead to discover additional QTLs like QTL4, because the number of degrees of freedom will decrease. The multiplication of tests prompts us to think about hierarchical testing procedure (Buzdugan et al., 2016) and the use of adequate multiple testing correction, like the higher criticism.

	Common Residual Variance		PT Residual Variances	
	AB General	Rel General	AB Specific	Rel Specific
GCA_F	4.64	5.09	4.65	5.20
GCA_D	3.91	4.17	3.90	4.19
SCA_H	0.52	0.53	0.67	0.58
GCA_{FT}	0.21	0.20	0.23	0.22
GCA_{DT}	0.26	0.26	0.31	0.30
SCA_{HT}	0.24	0.26	0.35	0.37
Res	0.81	0.81		
Res SMH15			0.89	0.88
Res LMG15			1.43	1.41
Res RAGT15			0.45	0.45
Res SMH14			0.24	0.25
Res CAU14			0.38	0.38
BIC	6811.01	6849.97	6735.18	6777.29

Suppl. Table 4.12: Variance parameters for anthesis of models (4.5) *and* (4.6) *using both A&B and* `Relatedness` *and without considering trial MAS15.*

Appendix

Appendix A: Results without considering trial MAS15

Results which can be affected by the error on data derived from MAS15 are the inference of variance components and tables of QTLs detected within each trial.

Variance Components Suppl. Tables 4.12 and 4.13 display estimation of variance components when studying anthesis and grain yield respectively without considering data derived from trial MAS15.

All BIC values have decreased comparing to the analysis with trial MAS15, which can be explained to a large extent by the decrease in the number of observations. The advantage of a per trial residual variance model compared with a common residual variance model regarding BIC values is not as pronounced as before when studying anthesis (a raise of approximately 1%). Using A&B method to infer relatedness still seems better than using `Relatedness`.

Summary of QTLs detection in individual trials Suppl. Tables 4.14 and 4.15 display QTLs detected for anthesis and grain yield when using A&B method to infer relatedness. Suppl. Tables 4.16 and 4.17 display QTLs detected for anthesis and grain yield when using `Relatedness`.
The number of regions detected for anthesis is lower without trial MAS15. This result is expected regarding the problem about the correction. There is no more QTL detected for allelic interaction between alleles among all trials when considering A&B method to infer relatedness. When using `Relatedness`, one marker is detected for allelic interaction effect. This could be due to a better modelization of the double relatedness coefficients between hybrids.

	Common Residual Variance		PT Residual Variances	
	AB General	Rel General	AB Specific	Rel Specific
GCA_F	85.30	86.90	87.79	88.30
GCA_D	66.51	73.67	64.71	71.41
SCA_H	6.02	7.60	6.19	8.59
GCA_{FT}	27.63	27.99	23.98	24.35
GCA_{DT}	18.95	21.15	19.46	21.05
SCA_{HT}	13.29	9.92	8.86	6.35
Res	106.77	108.51		
Res SMH15			154.45	156.17
Res LMG15			179.58	182.50
Res RAGT15			57.58	59.09
Res SMH14			110.85	110.27
Res CAU14			92.29	94.79
Res MOR14			77.72	78.66
BIC	17783.72	17818.01	17749.80	17784.35

Suppl. Table 4.13: Variance parameters for grain yield of models (4.5) *and* (4.6) *using both A&B and* `Relatedness` *and without considering trial MAS15.*

	Chr	MinPos	MaxPos	SpikePos	Pval	Type	Trial
QTL1	4	33073973	34691176	33882328	5.55e-08	AddF;Add	SMH14;LMG15
QTL2	8	122931493	122967814	122949646	1.47e-08	AddF;Add	CAU14
QTL3	8	123486667	123523088	123504889	1.51e-07	Add	CAU14

Suppl. Table 4.14: Summary of all QTLs detected in individual trial analyses, except trial MAS15, for the anthesis, using A&B method to infer relatedness coefficients. Chr is the chromosome number, MinPos (resp. MaxPos) is the minimum (resp. maximum) physical position of the QTL, SpikePos is the physical position of the marker associated to the lowest p-value, Pval is the p-value of the marker located in SpikePos, Type is the tested hypotheses which led to the detection, Trial is the list of trials where the QTL is detected.

Results about grain yield are not impacted by the error on the correction. Indeed, no QTL was found in trial MAS15 during the study. Correction of data at MAS15 may increase the number of detected QTLs.

	Chr	MinPos	MaxPos	SpikePos	Pval	Type	Trial
QTL1	1	187925319	188103693	188014659	4.78e-08	Add;Marker	SMH15
QTL2	1	275024696	275155215	275090036	3.58e-08	Marker	SMH15
QTL3	3	181856980	181899492	181878262	2.47e-06	Add	SMH15
QTL4	5	188687491	188736959	188712217	9.48e-08	AddD	SMH14
QTL5	7	108999484	110601741	109740799	2.03e-08	AddD;Add;Marker	SMH15
QTL6	7	114737557	115831136	115284347	2.91e-06	AddD	SMH15
QTL7	8	129936264	130002097	129978850	1.72e-07	Int	LMG15
QTL8	8	136032803	137312565	136713760	1.89e-07	AddD;Marker	RAGT15
QTL9	9	36510612	38307857	37409235	1.66e-07	Add;Marker	SMH15
QTL10	10	24440373	26069606	25254990	1.24e-07	Marker	MOR14

Suppl. Table 4.15: Summary of all QTLs detected in individual trial analyses, except trial MAS15, for the grain yield, using A&B method to infer relatedness coefficients. Chr is the chromosome number, MinPos (resp. MaxPos) is the minimum (resp. maximum) physical position of the QTL, SpikePos is the physical position of the marker associated to the lowest p-value, Pval is the p-value of the marker located in SpikePos, Type is the tested hypotheses which led to the detection, Trial is the list of trials where the QTL is detected.

	Chr	MinPos	MaxPos	SpikePos	Pval	Type	Trial
QTL1	1	187925319	188103693	188014659	5.20e-07	Add	SMH15
QTL2	1	249481328	250188927	249860569	1.14e-06	Add	SMH15
QTL3	1	296537860	296604671	296571258	9.21e-07	AddF;Add	LMG15
QTL4	2	42755878	42847764	42801764	6.60e-07	AddF;Add	LMG15
QTL5	2	183397532	183418800	183408153	2.05e-06	Add	LMG15
QTL6	2	214990030	215033365	215011710	3.90e-07	Add	SMH14;SMH15
QTL7	3	121214088	122906979	122060534	3.30e-07	AddF	SMH14
QTL8	3	199279355	199333417	199306388	1.92e-06	Add	SMH15
QTL9	4	33073973	34691176	33882328	6.00e-09	AddF;Add	SMH14;RAGT15;LMG15
QTL10	4	216614568	218231277	217422923	3.53e-06	Add	LMG15
QTL11	5	195282149	195325351	195303760	4.69e-07	Add	LMG15
QTL12	7	168081333	168087858	168084596	6.37e-07	AddF;Add	LMG15
QTL13	8	28249381	29243930	28746656	5.88e-06	Add	SMH15
QTL14	8	122931493	122967814	122949646	1.57e-09	AddF;Add;Marker	CAU14;SMH14;LMG15;SMH15
QTL15	8	123190729	123210729	123200729	6.33e-07	Add	LMG15
QTL16	8	123254290	123290662	123272484	4.24e-08	Add;Marker	CAU14;SMH15;LMG15
QTL17	8	123486131	123530355	123504889	6.81e-09	Add;Marker	CAU14;SMH14;SMH15;LMG15
QTL18	8	133024400	133044400	133034400	3.00e-06	Add	LMG15
QTL19	9	57736343	59751203	58852581	2.58e-07	AddF;Add	LMG15
QTL20	9	142733128	142760181	142746618	2.19e-06	Add	SMH15
QTL21	9	150819939	150831476	150825715	1.37e-06	AddF	LMG15
QTL22	10	138823997	138834442	138829251	7.92e-08	Int	SMH14

Suppl. Table 4.16: Summary of all QTLs detected in individual trial analyses, except trial MAS15, for the anthesis, using `Relatedness` to infer relatedness coefficients. Chr is the chromosome number, MinPos (resp. MaxPos) is the minimum (resp. maximum) physical position of the QTL, SpikePos is the physical position of the marker associated to the lowest p-value, Pval is the p-value of the marker located in SpikePos, Type is the tested hypotheses which led to the detection, Trial is the list of trials where the QTL is detected.

	Chr	MinPos	MaxPos	SpikePos	Pval	Type	Trial
QTL1	1	9410068	9460154	9435138	2.19e-07	Add;Marker	RAGT15
QTL2	1	179173582	179302324	179238996	2.86e-06	Add	RAGT15
QTL3	1	187925319	188179616	188014659	1.08e-08	Add;Marker	SMH15
QTL4	1	214280372	214383338	214331897	4.86e-06	Marker	SMH15
QTL5	1	215146259	215246254	215196304	6.58e-07	Add;Marker	RAGT15
QTL6	1	251315626	251543607	251429693	2.26e-07	Marker	MOR14
QTL7	1	275024696	275155215	275090036	1.45e-08	Marker	SMH15
QTL8	3	181856980	181899492	181878262	1.78e-06	Add;Marker	SMH15
QTL9	5	188687491	188736959	188712217	7.33e-09	AddD;Add	SMH14
QTL10	7	108999484	110601741	109740799	1.42e-08	AddD;Add;Marker	SMH15
QTL11	7	114715042	115831136	115284347	1.11e-06	AddD;Add;Marker	SMH15
QTL12	8	2412401	2438579	2425454	3.78e-07	Int	LMG15
QTL13	8	129936264	130002097	129978850	1.22e-07	Int	LMG15
QTL14	8	136032803	137312565	136716294	9.07e-07	Add;Marker	RAGT15
QTL15	9	36510612	38410771	37409235	4.73e-08	Add;Marker	SMH15
QTL16	10	24440373	26069606	25254990	2.71e-08	Marker	MOR14
QTL17	10	141956077	141976077	141966077	5.13e-06	Marker	SMH15

Suppl. Table 4.17: Summary of all QTL detected in individual trial analyses, except trial MAS15, for the grain yield, using `Relatedness` *to infer relatedness coefficients. Chr is the chromosome number, MinPos (resp. MaxPos) is the minimum (resp. maximum) physical position of the QTL, SpikePos is the physical position of the marker associated to the lowest p-value, Pval is the p-value of the marker located in SpikePos, Type is the tested hypotheses which led to the detection, Trial is the list of trials where the QTL is detected.*

Chapter 5

Conlusions

5.1 Conclusion Générale

Au cours de mon doctorat, je me suis intéressé à plusieurs méthodes statistiques nécessaires à la détection de QTLs par GWAS en génétique quantitative. Ma première contribution a été de proposer un cadre d'étude rigoureux pour l'identifiabilité et l'estimation des coefficients d'apparentement à partir de marqueurs bi-alléliques. Il s'agit à ce jour de la première étude de l'apparentement qui prend en compte la nature des données (individus phasés ou non) et le plan d'expérience (possible(s) parent(s) commun(s), structure en populations). Elle a permis de conclure que l'identifiabilité d'une partie des paramètres d'apparentement dépend de la structure du plan de croisement, c'est-à-dire de la manière dont des individus de la même population ou au contraire de populations différentes sont croisés pour obtenir les hybrides. Ainsi certains plans tels que les plans factoriels (Comstock et al., 1952) permettent l'estimation de l'ensemble des coefficients d'apparentement. A l'inverse, les plans diallèles (Griffing, 1956) ne permettent pas d'estimer le coefficient de double parenté entre les hybrides à partir de marqueurs bi-alléliques. Cette non-identifiabilité peut avoir des conséquences sur l'analyse statistique : le coefficient de double parenté est nécéssaire pour la modélisation de l'interaction entre allèles en effet polygénique, et des erreurs sur son estimation peuvent dégrader les performances de l'analyse d'association ou de sélection génomique à suivre. Le package R `Relatedness` permet de prendre en compte la structure des plans de croisement et de vérifier l'estimabilité des coefficients d'apparentement. Ce package est disponible, en version 2.0, sur le CRAN (Comprehensive R Archive Network) (R Core Team, 2015).

J'ai aussi étudié et implémenté un algorithme pour l'estimation des paramètres des modèles mixtes à composantes de la variance. Cet algorithme, bientôt disponible sous la forme d'un package R, sera particulièrement utile dans le cadre de la génétique quantitative pour la détection de QTLs. En effet, l'algorithme MM est plus rapide que ses concurrents directs lorsque les matrices de corrélation ne sont pas creuses. De nos jours en génétique des plantes, ces matrices sont largement estimées sur la base des marqueurs bi-alléliques, ce qui les empêche d'être creuses. De plus, cet algorithme s'est montré plus rapide que l'algorithme FaST-LMM dans sa version exacte (Lippert et al., 2011) qui

est l'algorithme de référence en génétique des plantes lorsque le modèle génétique ne contient qu'un seul effet aléatoire génétique.
Utiliser l'algorithme le mieux adapté aux données analysées est crucial en terme de temps de calcul : il est dans certains cas possible d'économiser plusieurs heures voire plusieurs jours de calcul lors de l'étape d'inférence. Notre étude a permis de caractériser avec précision les différents cas d'applications de prédilection des algorithmes étudiés, particulièrement en génétique végétale. Ainsi, si on considère des matrices de corrélations creuses, il est préférable d'utiliser un algorithme utilisant l'astuce de l'équation d'Henderson tel que ASReml (Johnson and Thompson, 1995; Gilmour et al., 2009).
Il n'existe pas à ce jour un algorithme permettant l'estimation la plus rapide et la plus précise des paramètres sur l'ensemble des cadres d'étude.

Ces développements ont été appliqués à un panel d'hybrides de maïs. La première remarque sur cette étude est le manque de puissance lors de la détection de QTLs. Nous avons donc suivi deux pistes pour essayer d'augmenter le nombre de QTLs détectés : utiliser `Relatedness` pour estimer les coefficients de l'apparentement et faire une analyse jointe de tous les essais. Lors de l'analyse essai par essai, très peu de QTLs ont été trouvés lorsque nous avons utilisé la méthode d'estimation de l'apparentement décrite par Astle and Balding (2009). Par comparaison, l'utilisation de l'algorithme `Relatedness` s'est traduit par une augmentation du nombre de régions détectées par les analyses. De plus, cela a permis de détecter une région "proche" d'un autre QTL majeur de floraison "vgt1" (chromosome 8, 131.000.000), ce qui tend à favoriser l'utilisation de notre package. Dans les deux cas, et malgré un manque de puissance avec la méthode A&B, nous avons été capable de détecter un QTL correspondant exactement à "vgt2". Bien entendu, il reste à étudier de manière propre l'impact de l'utilisation de `Relatedness` sur la puissance de détection. Il existe peu de stratégies disponibles pour cette étude au vu de la complexité des données, nous pensons à utiliser des données simulées pour analyser empiriquement cet impact.

Dans un second temps nous avons réalisé l'analyse de tous les lieux conjointement. Cette dernière nous a tout d'abord permis d'identifier une erreur lors de la préparation des données sur un lieu. En effet, une forte incohérence au sein d'un lieu affecte les effets d'interaction entre le génotype et les essais et augmente ainsi le nombre de régions détectées. Bien que l'étude multi-essais ne soit pas achevée, le nombre de régions détectées lors de l'analyse sans l'essai MAS15 est supérieur au nombre de régions détectées dans l'analyse essai par essai. Ce potentiel gain de puissance est un message encourageant pour la suite des études de GWAS. Les résultats de l'analyse jointe, une fois complétée, pourraient permettre de revoir la manière de détecter les QTLs en génétique des plantes et inciter les chercheurs à travailler dans cette voie.

5.2 Perspectives

Il y a beaucoup d'ouvertures sur les modèles mixtes suite à notre étude. La première serait d'étudier la possibilité d'utiliser des astuces de sparsité des matrices de corrélation

pour la méthode MM (Misztal and Perez-Enciso, 1993; Masuda et al., 2015) mais aussi d'utiliser des méthodes approchées des calculs de trace (Brezinski et al., 2012; Fika and Koukouvinos, 2017). On pourrait aussi s'intéresser à diminuer la taille des matrices à inverser. Pour cela, il serait possible d'orthogonaliser les modèles selon les vecteurs propres des matrices de corrélation en gardant seulement les valeurs propres les plus fortes. Il faudrait bien sûr, par la suite, étudier l'impact de cette projection, à la fois sur le temps de calcul (normalement plus court) et sur la précision des estimations.

La multiplication des effets aléatoires et le nombres grandissant de marqueurs disponibles entraînent des temps de calcul très longs pour la détection de QTL. Comme présentées dans le Chapitre 3, des méthodes d'inférence approchées (Kang et al., 2010; Lippert et al., 2011) permettent d'accélérer grandement le processus de détection mais ne permettent pas forcément de détecter tous les marqueurs. Il pourrait être intéressant de dériver une méthode de détection couplant méthodes approchées et méthodes exactes d'inférence. Cette méthode se composerait de deux parties. La première étant d'effectuer les tests sur l'ensemble des marqueurs en utilisant une méthode d'estimation approchée et de sélectionner les marqueurs passant un certain seuil de sélection peu strict (à déterminer). La seconde étape serait d'utiliser une méthode d'inférence exacte sur les marqueurs précédemment sélectionnés et de les déclarer comme ayant un effet significatif suivant un seuil de détection plus strict (à déterminer). Il faut donc étudier la valeur des différents seuils pour contrôler le nombre de faux positifs et conserver une puissance suffisante lors de ce processus de détection.

Certaines méthodes n'ont pas du tout été envisagées dans ce manuscrit, en particulier la méthode MCMC (Gilks et al., 1996) ou les méthodes de détection de QTLs bayesiennes. Il serait éventuellement intéressant d'étudier empiriquement ces méthodes comme dans le Chapitre 3, afin de considérer leur apport contre la méthode de détection présentée ici. On pourrait ainsi coupler ou dériver ces méthodes à la détection de QTLs par maximisation de la vraisemblance des modèles mixtes.

Nous avons pu voir l'effet d'insérer des données erronées sur un lieu dans une étude multi-environnements. Je pense qu'il serait intéressant d'effectuer une étude sur la puissance des tests (Saïdou et al., 2014) et la variation de cette dernière suivant la qualité des données insérées dans l'étude. De plus, les marqueurs observés sont corrélés entre eux et pour l'instant cette corrélation n'est pas prise en compte dans les modèles ni pour les corrections de tests multiples. Il pourrait être intéressant de quantifier la corrélation entre marqueurs à l'aide du déséquilibre de liaison (Hill and Weir, 1994; Doligez et al., 2011; Nicolas et al., 2016). Ce dernier commence à être pris en compte lors des corrections de test multiples (Li and Ji, 2005; Gao et al., 2010) mais les propriétés statistiques de ces seuils restent encore à préciser. L'application de ces seuils à la détection de QTLs détaillée dans le Chapitre 4 permettra éventuellement de détecter de nouvelles régions impactant le déterminisme des traits étudiés.

Il pourrait être aussi très intéressant d'effectuer une détection de QTLs sur le panel présenté dans la chapitre 4 en étudiant plusieurs phénotypes en même temps (Henderson and Quaas, 1976; Scutari et al., 2014), ce qui, selon les auteurs, augmente la

puissance de détection des QTLs pléiotropiques dans la cas d'étude d'effets additifs. Ces analyses permettraient de détecter des QTLs impactant simultanément plusieurs phénotypes d'intérêt en même temps, tel que la précocité de floraison et le rendement dans notre cas où une (resp. deux) région "commune" est détectée en utilisant A&B (resp. `Relatedness`) pour ces deux caractères. Bien évidemment, ce type d'étude entraîne une complexification des modèles à travers la prise en compte de la corrélation entre les phénotypes.
Cette étude multi-caractères pourrait se coupler à l'étude multi-environnements effectuée dans le Chapitre 4 (Alimi et al., 2013).

Outre la détection de QTL, les modèles mixtes présentés ont aussi des applications en sélection génomique (Heffner et al., 2009). Les analyses permettent de prédire sur la base de l'observation d'un pool d'individus les valeurs phénotypiques d'individus non observés. Il serait intéressant d'étudier l'impact de l'estimation de l'apparentement entre les individus sur la prédiction, une "meilleure" modélisation du double apparentement pourrait augmenter la qualité des prédictions.

Bibliography

Alimi, N., Bink, M., Dieleman, J., Magán, J., Wubs, A., Palloix, A., and Van Eeuwijk, F. (2013). Multi-trait and multi-environment QTL analyses of yield and a set of physiological traits in pepper. *Theoretical and applied genetics*, 126(10):2597–2625.

Astle, W. and Balding, D. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, pages 451–471.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bernardo, R. (1994). Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Sci.*, 34:20–25.

Bink, M., Anderson, A., van de Weg, W., and Thompson, E. (2008). Comparison of marker-based pairwise relatedness estimators on a pedigreed plant population. *Theoretical and Applied Genetics*, 117(6):843–855.

Brezinski, C., Fika, P., and Mitrouli, M. (2012). Moments of a linear operator, with applications to the trace of the inverse of matrices and the solution of equations. *Numerical Linear Algebra with Applications*, 19(6):937–953.

Browning, S. and Browning, B. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097.

Buzdugan, L., Kalisch, M., Navarro, A., Schunk, D., Fehr, E., and Bühlmann, P. (2016). Assessing statistical significance in multivariable genome wide association analysis. *Bioinformatics*, 32(13):1990–2000.

Comstock, R., Robinson, H., et al. (1952). Estimation of average dominance of genes. *Estimation of average dominance of genes.*

Crow, J. and Kimura, M. (1970). *An Introduction to Population Genetics Theory*. Harper and Row, Publishers, Inc.

Csuros, M. (2014). Non-identifiability of identity coefficients at biallelic loci. *Theoretical Population Biology*, 92:22–29.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, pages 1–38.

Doligez, A., Siberchicot, A., Mangin, B., Cierco-Ayrolles, C., This, P., and Nicolas, S. (2011). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity*, 108(3):285.

Fika, P. and Koukouvinos, C. (2017). Stochastic estimates for the trace of functions of matrices via hadamard matrices. *Communications in Statistics-Simulation and Computation*, 46(5):3491–3503.

Gallais, A. (1990). *Théorie de la Sélection en Amélioration des Plantes*. Masson.

Gao, X., Becker, L., Becker, D., Starmer, J., and Province, M. (2010). Avoiding the high bonferroni penalty in genome-wide association studies. *Genetic epidemiology*, 34(1):100–105.

Gibbons, R. and Hedeker, D. (2000). Applications of mixed-effects models in biostatistics. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 70–103.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). Introducing markov chain monte carlo. *Markov chain Monte Carlo in practice*, 1:19.

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

Gilmour, A. R., Gogel, B., Cullis, B., Thompson, R., Butler, D., et al. (2009). ASReml user guide release 3.0. *VSN International Ltd, Hemel Hempstead, UK*.

Giraud, H., Bauland, C., Falque, M., Madur, D., Combes, V., Jamin, P., Monteil, C., Laborde, J., Palaffre, C., Gaillard, A., et al. (2017). Reciprocal genetics: Identifying QTL for general and specific combining abilities in hybrids between multiparental populations from two maize (zea mays l.) heterotic groups. *Genetics*, 207(3):1167–1180.

Griffing, B. (1956). Concept of general and specific combining ability in relation to diallel crossing systems. *Australian journal of biological sciences*, 9(4):463–493.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338.

Hayman, B. (1954). The theory and analysis of diallel crosses. *Genetics*, 39:789–809.

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

Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics*, 9(2):226–252.

Henderson, C. (1973). Sire evaluation and genetic trends. *Journal of Animal Science*, 1973(Symposium):10–41.

Henderson, C. and Quaas, R. (1976). Multiple trait evaluation using relatives' records. *Journal of Animal Science*, 43(6):1188–1197.

Hepler, A. (2005). Improving forensic identification using bayesian networks and relatedness estimation. *North Carolina State University, Raleigh*.

Hill, W. and Weir, B. (1994). Maximum-likelihood estimation of gene location by linkage disequilibrium. *American journal of human genetics*, 54(4):705.

Hunter, D. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37.

Johnson, D. and Thompson, R. (1995). Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *Journal of dairy science*, 78(2):449–456.

Kalinowski, S., Wagner, A., and Taper, M. (2006). ML-Relate: a computer program for maximum likelihood estimation of relatedness and relationship. *Mol. Ecol. Notes*, 6.

Kang, H., Sul, J., Zaitlen, N., Kong, S., Freimer, N., Sabatti, C., Eskin, E., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354.

Kang, H., Zaitlen, N., Wade, C., Kirby, A., Heckerman, D., Daly, M., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723.

Laporte, F., Charcosset, A., and Mary-Huard, T. (2017). Estimation of the relatedness coefficients from biallelic markers, application in plant mating designs. *Biometrics*.

Laporte, F. and Mary-Huard, T. (2017). *Relatedness: Maximum Likelihood Estimation of Relatedness using EM Algorithm*. R package version 2.0.

Li, J. and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95(3):221–227.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C., Davidson, R., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833–835.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc.

Masuda, Y., Aguilar, I., Tsuruta, S., and Misztal, I. (2015). Acceleration of sparse operations for average-information reml analyses with supernodal methods and sparse-storage refinements. *Journal of animal science*, 93(10):4670–4674.

McPeek, M. and Sun, L. (2000). Statistical tests for detection of misspecified relationships by use of genome-screen data. *The American Journal of Human Genetics*, 66(3):1076–1094.

Milligan, B. (2003). Maximum-likehood estimation of relatedness. *Genetics*, 163:1153–1167.

Misztal, I. and Perez-Enciso, M. (1993). Sparse matrix inversion for restricted maximum likelihood estimation of variance components by expectation-maximization. *Journal of dairy science*, 76(5):1479–1483.

Moreau, L., Charcosset, A., and Gallais, A. (2004). Use of trial clustering to study qtl× environment effects for grain yield and related traits in maize. *Theoretical and applied genetics*, 110(1):92–105.

Nicolas, S., Péros, J.-P., Lacombe, T., Launay, A., Le Paslier, M.-C., Bérard, A., Mangin, B., Valière, S., Martins, F., Le Cunff, L., et al. (2016). Genetic diversity, linkage disequilibrium and power of a large grapevine (vitis vinifera L) diversity panel newly designed for association studies. *BMC plant biology*, 16(1):74.

Parisseaux, B. and Bernardo, R. (2004). In silico mapping of quantitative trait loci in maize. *Theoretical and Applied Genetics*, 109(3):508–514.

Perdry, H. and Dandine-Roulland, C. (2017). *gaston: Genetic Data Handling (QC, GRM, LD, PCA) & Linear Mixed Models*. R package version 1.5.

Petersen, K., Pedersen, M., et al. (2008). The matrix cookbook. *Technical University of Denmark*, 7:15.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rincent, R., Moreau, L., Monod, H., Kuhn, E., Melchinger, A., Malvar, R., Moreno-Gonzalez, J., Nicolas, S., Madur, D., Combes, V., et al. (2014a). Recovering power in association mapping panels with variable levels of linkage disequilibrium. *Genetics*, 197(1):375–387.

Rincent, R., Nicolas, S., Bouchet, S., Altmann, T., Brunel, D., Revilla, P., Malvar, R., Moreno-Gonzalez, J., Campo, L., Melchinger, A., et al. (2014b). Dent and flint maize diversity panels reveal important genetic potential for increasing biomass production. *Theoretical and applied genetics*, 127(11):2313–2331.

Saïdou, A.-A., Thuillet, A.-C., Couderc, M., Mariac, C., and Vigouroux, Y. (2014). Association studies including genotype by environment interactions: prospects and limits. *BMC genetics*, 15(1):3.

Scheffe, H. (1956). A "mixed model" for the analysis of variance. *The Annals of Mathematical Statistics*, pages 23–36.

Scutari, M., Howell, P., Balding, D., and Mackay, I. (2014). Multiple quantitative trait analysis using bayesian networks. *Genetics*, 198(1):129–137.

Searle, S., Casella, G., and McCulloch, C. (1992). *Variance components*, volume 391. John Wiley & Sons.

Shull, G. (1908). The composition of a field of maize. *Am. Breeders' Assoc. Rep.*, 5:51–59.

Shull, G. (1914). Duplicate genes for capsule-form inbursa bursa-pastoris. *Molecular and General Genetics MGG*, 12(1):97–149.

Sprague, G. and Tatum, L. (1942). General vs. specific combining ability in single crosses of corn. *Agronomy journal*, 34(10):923–932.

Technow, F., Riedelsheimer, C., Schrag, T., and Melchinger, A. (2012). Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *Theor Appl Genet*, 125:1181–1194.

Technow, F., Schrag, T., Schipprack, W., Bauer, E., Simianer, H., and Melchinger, A. (2014). Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. *Genetics*, 197(4):1343–1355.

Thompson, E. (1975). Estimation of pairwise relationships. *Ann. hum. genet.*, 39(2):173–188.

van Eeuwijk, F., Boer, M., Totir, L., Bink, M., Wright, D., Winkler, C., Podlich, D., Boldman, K., Baumgarten, A., Smalley, M., et al. (2010). Mixed model approaches for the identification of qtls within a maize hybrid breeding program. *Theoretical and Applied Genetics*, 120(2):429–440.

van Eeuwijk, F., Malosetti, M., Yin, X., Struik, P., and Stam, P. (2005). Statistical models for genotype by environment data: from conventional anova models to eco-physiological qtl models. *Australian Journal of Agricultural Research*, 56(9):883–894.

Varadhan, R. and Roland, C. (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scandinavian Journal of Statistics*, 35(2):335–353.

Vitezica, Z., Varona, L., and Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics*, 195(4):1223–1230.

Wang, J. (2011). COANCESTRY: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Molecular Ecology Resources*, 11(1):141–145.

Wright, S. (1922). Coefficients of inbreeding and relationship. *Am. Nat.*, pages 330–338.

Yu, J., Pressoir, G., Briggs, W., Vroh Bi, I., Yamasaki, M., Doebley, J., McMullen, M., Gaut, B., Nielsen, D., Holland, J., Kresovich, S., and Buckler, E. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208.

Zhou, H., Hu, L., Zhou, J., and Lange, K. (2015). MM algorithms for variance components models. *arXiv preprint arXiv:1509.07426*.

Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–824.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Développement de méthodes statistiques pour l'identification de gènes d'intérêt en présence d'apparentement et de dominance, application à la génétique du maïs.

Mots Clefs : Modèle de mélange, Modèle mixte, Identifiabilité, Apparentement, Génétique d'association.

Résumé : La détection de gènes est une étape importante dans la compréhension des effets de l'information génétique d'un individu sur ses caractères phénotypiques. Durant mon doctorat, j'ai étudié les méthodes statistiques pour conduire les analyses de génétique d'association, avec les hybrides de maïs comme modèle d'application. Je me suis tout d'abord intéressé à l'estimation des paramètres d'apparentement entre individus à partir de données de marqueurs bialléliques. Cette estimation est réalisée dans le cadre d'un modèle de mélange paramétrique. J'ai étudié l'identifiabilité de ce modèle dans un cadre général mais aussi dans un cadre plus spécifique où les individus étudiés étaient issus de croisements entre lignées, cadre représentatif des plans de croisement classiquement utilisés en génétique végétale. Je me suis ensuite intéressé à l'estimation des paramètres des modèles mixtes à plusieurs composantes de variance et plus particulièrement à la performance des algorithmes pour tester l'effet de très nombreux marqueurs. J'ai comparé pour cela des logiciels existants et optimisé un algorithme Min-Max. La pertinence des différentes méthodes développées a finalement été illustrée dans le cadre de la détection de QTL à travers une analyse d'association réalisée sur un panel d'hybrides de maïs.

Title : Development of statistical methods to identify genes of interest in presence of relatedness and dominance, application to maize genetics.

Keys words : Mixture model, Mixed model, Identifiability, Relatedness, Genome-wide association studies.

Abstract : The detection of genes is a first step to understand the impact of the genetic information of individuals on their phenotypes. During my PhD, I studied statistical methods to perform genome-wide association studies, with maize hybrids as an application case. Firstly, I studied the inference of relatedness coefficients between individuals from biallelic marker data. This estimation is based on a parametric mixture model. I studied the identifiability of this model in the generic case but also in the specific case of mating design where observed individuals are obtained by crossing lines, a representative case of classical mating design in plant genetics. Then I studied inference of variance component mixed model parameters and particularly the performance of algorithms to test effects of numerous markers. I compared existing programs and I optimized a Min-Max algorithm. Relevance of developed methods had been illustrated for the detection of QTLs through a genome-wide association analysis in a maize hybrids panel.

Université Paris-Saclay, Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128, 91190 Saint-Aubin, France