Optimisation des méthodes statistiques d'analyse de la variabilité des caractères à l'aide d'informations génomiques

Laval Jacquin

Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (INP Toulouse)

Discipline ou spécialité :

Pathologie, Toxicologie, Génétique et Nutrition

Présentée et soutenue par :

M. LAVAL YANNIS JULIEN JACQUIN

le vendredi 10 octobre 2014

Titre :

Optimisation des méthodes statistiques d'analyse de la variabilité des caractères à l'aide d'informations génomiques

---

École doctorale :

Sciences Écologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)

Unité de recherche :

Génétique Physiologie et Systèmes d'Élevage (GenPhySE)

Directeurs de Thèse :

M. JEAN-MICHEL ELSEN

MME HELENE GILBERT

Rapporteurs :

M. DIDIER BOICHARD, INRA JOUY EN JOSAS

M. MIGUEL PEREZ ENCISO, UNIVERSIDAD AUTONOMA DE BARCELONE

Membres du jury :

Mme MARIA MARTINEZ, INSERM MIDI PYRENEES, Présidente

M. JEAN-MICHEL ELSEN, INRA TOULOUSE, Membre

Mme ANNE-LOUISE LEUTENEGGER, UNIVERSITE PARIS 7, Membre

Mme HELENE GILBERT, INRA TOULOUSE, Membre

Remerciements

Nombreux sont ceux qui ont contribué de près ou de loin à la réalisation de ce travail de thèse, bien qu'il n'y ait qu'un seul nom sur la couverture de ce manuscrit. Je saisis ici l'occasion de leur témoigner ma profonde gratitude.

Je tiens d'abord à remercier les financeurs de ce travail : l'Agence Nationale de la Recherche (ANR) et l'Institut National de la Recherche Agronomique (INRA), sans qui cette thèse n'aurait été possible.

Mes autres pensées de remerciements vont ensuite vers mes directeurs de thèse, Jean-Michel Elsen et Hélène Gilbert, avec qui j'ai parcouru un long chemin sur ces trois années de thèse. Vous m'avez beaucoup appris sur la génétique et vous m'avez donné de précieux conseils pendant la durée de cette thèse.

Je remercie également les autres membres du jury : Didier Boichard, Miguel Pérez-Enciso, Maria Martinez et Anne-Louise Leutenegger, de m'avoir fait l'honneur d'être les rapporteurs et examinateurs de ce travail de thèse. Merci d'avoir pris le temps de vous intéresser à mon travail.

Ensuite j'aimerais remercier tous les gens de l'unité de recherche au sein de laquelle j'ai effectué ma thèse. Merci à toi Rachel d'avoir partagé ton bureau, à deux reprises, à cause des aménagements. Merci au chef de département, Denis Milan, et aux directeurs d'unité, Xavier Fernandez et Christèle Robert, d'être attentifs aux besoins des thésards.

Merci également à tous ceux qui ne sont pas cités, car la liste est longue et j'en oublierai certainement plusieurs au passage. Quoi qu'il en soit vous avez tous ma profonde gratitude.

J'adresse bien sûr un remerciement particulier à celle qui m'a épaulé au quotidien pendant ces trois années, surtout à un moment bien précis. Merci, je n'y serais jamais arrivé sans toi.

Finalement, je tiens à dédier cette thèse à ma famille et particulièrement à la mémoire de ma mère, et de mon père, à qui je dois ma réussite et les valeurs qui font de moi ce que je suis... Vous avez été les meilleurs parents que je puisse espérer avoir, merci à vous.

A mes parents,
mon frère Pascal
et ma sœur Nadine

Résumé

L'avènement du génotypage à haut débit permet aujourd'hui de mieux exploiter le phénomène d'association, appelé déséquilibre de liaison (LD), qui existe entre les allèles de différents loci sur le génome. Dans ce contexte, l'utilité de certains modèles utilisés en cartographie de locus à effets quantitatifs (QTL) est remise en question. Les objectifs de ce travail étaient de discriminer entre des modèles utilisés en routine en cartographie et d'apporter des éclaircissements sur la meilleure façon d'exploiter le LD, par l'utilisation d'haplotypes, afin d'optimiser les modèles basés sur ce concept. On montre que les modèles uni-marqueur de liaison, développés en génétique il y a vingtaine d'années, comportent peu d'intérêts aujourd'hui avec le génotypage à haut débit. Dans ce contexte, on montre que les modèles uni-marqueur d'association comportent plus d'avantages que les modèles uni-marqueur de liaison, surtout pour des QTL ayant un effet petit ou modéré sur le phénotype, à condition de bien maîtriser la structure génétique entre individus. Les puissances et les robustesses statistiques de ces modèles ont été étudiées, à la fois sur le plan théorique et par simulations, afin de valider les résultats obtenus pour la comparaison de l'association avec la liaison. Toutefois, les modèles uni-marqueur ne sont pas aussi efficaces que les modèles utilisant des haplotypes dans la prise en compte du LD pour une cartographie fine de QTL. Des propriétés mathématiques reliées à la cartographie de QTL par l'exploitation du LD multiallélique capté par les modèles haplotypiques ont été explicitées et étudiées à l'aide d'une distance matricielle définie entre deux positions sur le génome. Cette distance a été exprimée algébriquement comme une fonction des coefficients du LD multiallélique. Les propriétés mathématiques liées à cette fonction montrent qu'il est difficile de bien exploiter le LD multiallélique, pour un génotypage à haut débit, si l'on ne tient pas compte uniquement de la similarité totale entre des haplotypes. Des études sur données réelles et simulées ont illustré ces propriétés et montrent une corrélation supérieure à 0.9 entre une statistique basée sur la distance matricielle et des résultats de cartographie. Cette forte corrélation a donné lieu à la proposition d'une méthode, basée sur la distance matricielle, qui aide à discriminer entre les modèles utilisés en cartographie.

Mot-clés *: locus à effets quantitatifs (QTL), déséquilibre de liaison (LD), puissance statistique, robustesse statistique, association, liaison, distance matricielle, haplotypes.*

Abstract

The advent of high-throughput genotyping nowadays allows better exploitation of the association phenomenon, called linkage disequilibrium (LD), between alleles of different loci on the genome. In this context, the usefulness of some models to fine map quantitative trait locus (QTL) is questioned. The aims of this work were to discriminate between models routinely used for QTL mapping and to provide enlightenment on the best way to exploit LD, when using haplotypes, in order to optimize haplotype-based models. We show that single-marker linkage models, developed twenty years ago, have little interest today with the advent of high-throughput genotyping. In this context, we show that single-marker association models are more advantageous than single-marker linkage models, especially for QTL with a small or moderate effect on the phenotype. The statistical powers and robustness of these models have been studied both theoretically and by simulations, in order to validate the comparison of single-marker association models with single-marker linkage models. However, single-marker models are less efficient than haplotype-based models for making better use of LD in fine mapping of QTL. Mathematical properties related to the multiallelic LD captured by haplotype-based models have been shown, and studied, by the use of a matrix distance defined between two loci on the genome. This distance has been expressed algebraically as a function of the multiallelic LD coefficients. The mathematical properties related to this function show that it is difficult to exploit well multiallelic LD, for a high-throughput genotyping, if one takes into account the partial and total similarity between haplotypes instead of the total similarity only. Studies on real and simulated data illustrate these properties and show a correlation above 0.9 between a statistic based on the matrix distance and mapping results. Hence a new method, based on the matrix distance, which helps to discriminate between models used for mapping is proposed.

Keywords *: quantitative trait locus (QTL), linkage disequilibrium (LD), statistical power, statistical robustness, association, linkage, matrix distance, haplotypes.*

Liste des abréviations

ADN : Acide DésoxyriboNucléique
AIP : *Allelic Identity Predictor*
BLUE : *Best Linear Unbiased Estimator*
BLUP : *Best Linear Unbiased Predictor*
DAG : *Directed Acyclic Graph*
EM : *Expectation-Maximization*
FS : *Fisher Scoring*
GWAS : *Genome Wide Association Studies*
HMM : *Hidden Markov Models*
HWE : *Hardy Weinberg Equilibrium*
IBD : *Identical By Descent*
IBS : *Identity By State*
LA : *Linkage Analysis*
LASSO : *Least Absolute Shrinkage and Selection Operator*
LD : *Linkage Disequilibrium*
LDA : *Linkage Disequilibrium Analysis*
LDLA : *Linkage Disequilibrium Linkage Analysis*
LRT : *Likelihood Ratio Test*
MCMC : *Monte Carlo Markov Chains*
ML : *Maximum Likelihood*
MVUE : *Minimum Variance Unbiased Estimator*
NR : *Newton Raphson*
OLS : *Ordinary Least Squares*
QTL : *Quantitative Trait Locus*
REML : *Restricted Maximum Likelihood*
SNP : *Single Nucleotide Polymorphism*
SNR : *Signal-to-Noise Ratio*
TP : *Trained Predictor*
VLMC : *Variable Length Markov Chains*

Table des matières

Introduction générale

Localiser les emplacements du génome impliqués dans la variation des phénotypes d'une espèce est l'un des enjeux importants de la génétique moderne. Le génome d'une espèce peut être défini par l'ensemble des gènes possédés par les individus de cette espèce et matérialisé par les molécules d'ADN constituant les chromosomes de celle-ci. Les gènes peuvent être polymorphes (i.e. variables), les différentes versions d'un gène étant qualifiées d'allèles. Le phénotype est l'expression mesurable, quantitative ou qualitative, d'un ou de plusieurs gènes. Il peut également être vu comme une variable réponse associée à un ou plusieurs gènes polymorphes. Un emplacement du génome exerçant un effet sur la variabilité d'un phénotype est connu, en génétique quantitative, sous l'acronyme anglo-saxon de QTL pour **"Quantitative Trait Locus"**. La cartographie fine de QTL, dans le cadre de la génétique quantitative, consiste à localiser un emplacement du génome ayant un effet sur le phénotype associé aux individus de l'espèce considérée. L'objectif ultime de la cartographie de QTL est l'identification des mutations causales de la variabilité des phénotypes, afin par exemple de lutter contre certaines maladies observées dans une espèce, ou de mieux sélectionner des individus en agriculture ou en élevage. Cette sélection repose sur l'évaluation du potentiel génétique de chaque individu, appelé valeur génétique.

Les approches classiques pour l'identification d'un QTL consistent à mettre en association la variabilité observée, pour un phénotype, avec la variabilité observée pour des segments chromosomiques, à divers emplacements du génome. La variabilité observée pour un segment chromosomique peut être quantifiée par **la similarité** ou la dissimilarité observée entre individus, pour ce même segment. Ces approches peuvent être classées en deux catégories de modèles linéaires, appelées **analyses d'association** et **analyses de liaison**, qui trouvent leurs fondements en génétique. Les avantages et inconvénients de ces analyses sont antagonistes. Cependant, avec le développement des possibilités de balisage très dense du génome, les analyses d'association sont généralement considérées comme plus prometteuses que les analyses de liaison. Les analyses d'association et les analyses de liaison reposent sur l'exploitation statistique du **déséquilibre de liaison (LD)**, qui se définit comme une mesure de la non indépendance probabiliste entre allèles portés par des individus en deux positions distinctes d'un chromosome. Du fait de ce LD, le poly-

morphisme observé en une position n'est pas indépendant de la variabilité existant en une autre position proche affectant le phénotype mais non observable. Les approches les plus simples explorent le génome position par position.

Cependant, le LD semble être mieux exploité, d'après la littérature, si on utilise une combinaison ordonnée de plusieurs allèles portés sur plusieurs positions d'un chromosome, également appelée **haplotype**, au lieu d'un seul allèle à une seule position. De ce fait, les modèles qui utilisent des haplotypes sont souvent décrits comme étant les plus performants pour la cartographie fine de QTL. Toutefois il n'existe pas de consensus sur la meilleure façon d'exploiter le LD par l'utilisation d'haplotypes. Une partie de ce travail de thèse vise à apporter des réponses à cette question afin d'optimiser les analyses statistiques basées sur le LD. En effet, **la précision**, **la puissance** et **la robustesse** des modèles de cartographie de QTL sont liées, entre autres, à la bonne prise en compte du LD. Ce travail de thèse, composé de trois parties, s'inscrit dans le cadre du projet "Rules & Tools" de l'ANR qui visait à améliorer les méthodes statistiques pour la cartographie de QTL.

La **Partie I** de ce manuscrit décrit le cadre général dans lequel a été réalisé ce travail de thèse. Cette partie décrit certains concepts, modèles, objets et méthodes qui sont utilisés à divers endroits du manuscrit afin de répondre aux objectifs de la thèse. La **Partie II** vise à mieux caractériser la prise en compte du LD par l'utilisation d'haplotypes. Un certain nombre de modèles d'association, qui utilisent des haplotypes en cartographie de QTL, sont comparés dans cette partie. A chacun de ces modèles correspond une manière de prédire l'identité allélique entre les allèles non observés d'un QTL à partir de la similarité entre les allèles des haplotypes entourant ce QTL. Le concept de **prédicteur d'identité allélique (AIP)** est introduit dans cette partie afin de décrire ces différentes manières de prédire. Les AIP supposent que l'on arrive à mieux exploiter le LD, entre une position testée et un QTL sur le génome, si l'on tient compte de la similarité existante entre des haplotypes. Cependant, aucun résultat algébrique n'est disponible dans la littérature quant au lien entre le LD et la prédiction d'identité allélique qu'effectuent les AIP. L'un des objectifs de la partie II de ce manuscrit est de caractériser ce lien à l'aide d'une distance matricielle définie entre une position testée et un QTL sur le génome.

Je démontre **algébriquement** dans cette partie qu'il est difficile d'exploiter le LD, entre une position testée et un QTL, si l'on ne tient pas compte uniquement de la similarité totale entre des haplotypes. Je montre également **l'efficacité** de l'AIP **le plus simple**, par rapport à d'autres AIP comparés, en présence du LD. Une étude sur **données réelles (porcs et humains)**, complétant l'étude algébrique, illustre les résultats de cette dernière quant à la prise en compte du LD par l'utilisation d'haplotypes. Enfin, je propose une méthode dans cette partie, basée sur la distance matricielle choisie, qui aide à mieux discriminer entre les AIP par rapport au LD.

La **Partie III** quant à elle vise à comparer les modèles d'association aux modèles de liaison. Les modèles de liaison sont réputés pour être robustes aux erreurs de première

espèce (faux positifs) lors des analyses pour la détection de QTL. Cependant les modèles d'association peuvent aussi être robustes à ces erreurs si l'on corrige pour la structuration génétique due aux relations entre individus. Le premier objectif de cette partie est de **quantifier la robustesse** et **la puissance** de ces modèles dans **diverses situations génétiques** générant la variabilité des phénotypes. Le second objectif est de **rechercher les situations** pour lesquelles les modèles de liaison pourraient être plus avantageux que les modèles d'association. Cette étude a été faite à la fois sur le plan **théorique** et par **simulations**. Les **facteurs de décentrage** pour des modèles de liaison et des modèles d'association sont dérivés dans cette partie afin d'étudier la robustesse de ces modèles sous l'hypothèse vraie de la présence d'un QTL (i.e. sous l'alternative).

Je montre dans cette partie que les modèles de liaison sont **antagonistes** aux modèles d'association, dans une certaine mesure, par rapport au LD mesuré sur l'ensemble d'une population et celui mesuré intra-famille. Je montre que cet antagonisme rend les **modèles de liaison plus avantageux** que les modèles d'association pour la détection de QTL, lorsque l'on a un **niveau relativement faible** de LD entre une position testée et un QTL ayant un **effet suffisamment élevé** sur le phénotype. Finalement, je montre que cet antagonisme rend aussi les **modèles de liaison peu précis** pour la cartographie fine de QTL, bien que ces modèles puissent aider, dans certaines situations, à renforcer la validité des résultats de cartographie obtenus par des modèles d'association.

Partie I :
Cadre général de l'étude

Chapitre 1

Définitions et phénomènes modélisés

L'objectif de ce chapitre est d'introduire certaines notions de génétique essentielles à la compréhension des différents modèles utilisés dans le cadre de la cartographie fine de QTL.

1.1 Définitions

1.1.1 Le génome diploïde et la recombinaison

Le support de l'information génétique est la molécule d'ADN (Acide DésoxyriboNucléique). Celle-ci constitue les *chromosomes* d'un individu appartenant à une espèce. Les individus d'une espèce dite *diploïde* sont caractérisés par n *paires* de chromosomes *homologues*, où chaque élément d'une paire est reçu d'un parent pendant la fécondation après un processus de division cellulaire appelé méiose. Pendant le processus de méiose il arrive que les chromosomes homologues d'une paire échangent leur matériel génétique à certaines positions. Cet échange se produit de façon aléatoire et se nomme la recombinaison. La figure 1.1 illustre deux recombinaisons, ayant respectivement lieu en deux positions e_1 et e_2, pour une paire de chromosomes homologues lors de la méiose. On remarquera que l'on parle parfois, de manière imprécise, de *locus* pour désigner une position sur un chromosome. Cependant, on considérera de manière équivalente la notion de position et celle de locus dans ce travail.

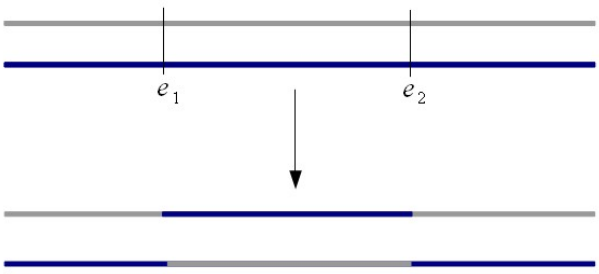


FIGURE 1.1 – Deux recombinaisons ayant respectivement lieu en e_1 et e_2 (deux loci) pour une paire de chromosomes homologues lors de la méiose.

Le phénomène de recombinaison entre deux positions est quantifié par un taux dont l'unité est le centiMorgan (cM). Ce taux définit une distance virtuelle sur un chromosome, appelée *distance génétique*, car deux positions éloignées ont une plus forte probabilité de recombiner et inversement (Strachan, 1999). Un cM correspond à environ 1% de recombinaison entre deux positions lors d'une méiose. On observe en moyenne au moins une recombinaison après une méiose pour des chromosomes ayant une taille avoisinante à 100 cM. On remarquera qu'il existe également une définition physique de la distance (en nombre de bases) sur le génome, mais celle-ci est beaucoup moins adaptée que la distance génétique pour des applications statistiques (Boitard, 2006). Une approximation couramment utilisée pour la distance physique est la suivante : 10^6 bases = 1cM.

1.1.2 Les marqueurs et cartes génétiques

Un marqueur génétique est une variation de la séquence d'ADN spécifiquement repérable à laquelle on associe sa position par rapport à la première base (le début) d'un chromosome. Les marqueurs génétiques sont utilisés pour baliser un chromosome afin de le modéliser par un intervalle réel. Il existe divers types de marqueurs génétiques mais ceux qui sont les plus couramment utilisés aujourd'hui en cartographie de QTL sont les "Single Nucleotide Polymorphism" (SNP), car ces derniers sont extrêmement nombreux sur le génome de beaucoup d'espèces.

L'utilisation des marqueurs génétiques en cartographie de QTL provient du fait que ces derniers présentent une variation observable sur l'ADN. En ce sens on dit également qu'ils sont polymorphes. Un marqueur génétique peut donc être assimilé à une variable à valeurs dans un certain alphabet $\{a_1, a_2, ..., a_s; s > 1\}$. Les éléments de cet alphabet en génétique sont appelés allèles. Par exemple, la plupart des SNP ont au plus deux allèles a_1 et a_2. De ce fait ces SNP sont dits bialléliques. Les marqueurs génétiques ayant plus de deux allèles possibles sont quant à eux dits multialléliques.

La carte génétique d'une espèce contient les marqueurs que l'on a pu identifier sur chaque chromosome associé à cette espèce. En d'autres termes elle donne une description non exhaustive du génome de l'espèce. Cette carte donne la relation d'ordre entre les marqueurs identifiés par leur distance génétique par rapport à l'origine de chaque chromosome. On notera que les cartes génétiques sont parfois des cartes physiques que l'on transforme en utilisant des règles approximatives de conversion. Par exemple, la règle approximative de conversion habituellement utilisé chez les mammifères est 10^6 bases=1cM (Strachan, 1999).

1.1.3 La notion d'haplotype et de génotype

Un haplotype est défini comme une combinaison ordonnée et finie de plusieurs allèles pour des polymorphismes (en pratique ici des SNP) portés par un chromosome. Par

exemple si on considère deux SNP M_1 et M_2, à valeurs dans $\{a_1, a_2\}$ et $\{b_1, b_2\}$ respectivement, alors on obtient l'ensemble $\{a_1b_1, a_1b_2, a_2b_1, a_2b_2\}$ de quatre haplotypes possibles définis par ces deux paires d'allèles aux marqueurs. Ainsi pour une suite de $L \geq 1$ SNP adjacents sur un chromosome on a 2^L haplotypes possibles définis par ces SNP.

Un individu diploïde possède toujours une paire d'haplotypes, pour une suite de SNP sur un chromosome, où chaque élément de la paire est reçu d'un parent. Pour les marqueurs M_1 et M_2, un individu peut par exemple recevoir la paire d'haplotypes $\{a_1b_1, a_2b_2\}$ ou $\{a_1b_2, a_2b_2\}$, et ainsi de suite, de ses parents. La notion d'haplotype est importante car les haplotypes permettent, contrairement à un seul SNP, de bien exploiter le déséquilibre de liaison (cf. 1.1.6) qui est le concept clef sur lequel repose la détection et la cartographie fine de QTL (de Bakker *et al.*, 2005 ; Boleckova *et al.*, 2012.).

Le génotype d'un individu diploïde en un marqueur est le couple d'allèles qu'il possède. Par exemple, le génotype d'un individu pour le marqueur M_1 peut être a_1/a_1 ou a_1/a_2 ou a_2/a_2, où chaque allèle du génotype de l'individu est reçu d'un parent. Un individu ayant un génotype constitué d'une même paire d'allèles (i.e. a_1/a_1 ou a_2/a_2) est dit homozygote, alors qu'un individu ayant des allèles différents (i.e. a_1/a_2) pour son génotype est dit hétérozygote.

1.1.4 L'équilibre d'Hardy-Weinberg (HWE)

Le principe d'équilibre d'Hardy-Weinberg (HWE), proposé indépendamment par le mathématicien G.H. Hardy et le médecin Wilhelm Weinberg en 1908, stipule que sous certaines conditions les fréquences des allèles et des génotypes d'un locus restent constantes d'une génération à l'autre dans une population. Les conditions les plus importantes de cet équilibre incluent l'accouplement aléatoire des individus, que ces derniers soient diploïdes et que la population soit de taille infinie. Les autres conditions sont le non-chevauchement des générations et qu'il n'y ait pas de mutation, de sélection et de migration (aucune copie d'un allèle provient de l'extérieur). Pour un locus M_1 ayant deux allèles a_1 et a_2, de fréquences f_{a_1} et f_{a_2} respectifs ($f_{a_1} + f_{a_2} = 1$), cet équilibre se traduit par :

$$\begin{cases} f_{a_1/a_1} = f_{a_1}^2, \quad f_{a_1/a_2} = 2f_{a_1}f_{a_2}, \quad f_{a_2/a_2} = f_{a_2}^2 \\ \\ f_{a_1}^2 + 2f_{a_1}f_{a_2} + f_{a_2}^2 = (f_{a_1} + f_{a_2})^2 = 1 \end{cases}$$

où f_{a_1/a_1}, f_{a_1/a_2} et f_{a_2/a_2} sont respectivement les fréquences des génotypes a_1/a_1, a_1/a_2 et a_2/a_2. En pratique, cet équilibre est très souvent vérifié car on peut montrer mathématiquement qu'une seule génération d'accouplement aléatoire, dans une population de grande taille, suffit pour atteindre cet équilibre.

1.1.5 La notion d'effet additif et de dominance à un locus

L'approche classique utilisée en génétique afin de modéliser l'effet d'un locus i sur un phénotype est décrite dans Falconer et Mackay (1996). Dans cette approche le locus i est supposé biallélique d'allèles a_1^i et a_2^i. L'effet m^i du génotype au locus i sur un individu, est ensuite paramétré $(\mathfrak{p})$ de la façon suivante :

$$(\mathfrak{p}) : m^i = \begin{cases} \mu_i + \nu_i & \text{si l'individu a le génotype } a_1^i/a_1^i \\ \mu_i + \delta_i & \text{si l'individu a le génotype } a_1^i/a_2^i \\ \mu_i - \nu_i & \text{si l'individu a le génotype } a_2^i/a_2^i \end{cases}$$

où μ_i représente l'effet moyen du locus i, ν_i son effet additif et δ_i son effet de dominance. Si $\delta_i = 0$ le modèle est purement additif, ce qui équivaut à un modèle de régression sur le nombre d'allèles a_1^i et a_2^i pour le génotype au locus i. On considère que le locus i, pour la paramétrisation $(\mathfrak{p})$, est un QTL si l'effet du génotype à ce locus est statistiquement significatif (i.e. si on a une différence de moyennes significative pour les différents génotypes). Parmi les déviations à la paramétrisation $(\mathfrak{p})$ les interactions entre les allèles de plusieurs loci sur le génome forment les phénomènes d'épistasie. L'approche de Falconer et Mackay est celle que l'on considérera dans la suite de ce travail.

1.1.6 Le déséquilibre de liaison et la recombinaison

1.1.6.1 Définition et mesures du déséquilibre de liaison (LD)

Le terme de déséquilibre de liaison entre loci, d'acronyme anglo-saxon LD pour "Linkage Disequilibrium", traduit l'association préférentielle qu'il peut y avoir entre les allèles de ces derniers. La quantification du LD est donc une grandeur mathématique qui mesure la non-indépendance probabiliste entre les allèles des loci. Soient M_1 et M_2 deux SNP respectivement à valeurs dans $\{a_1, a_2\}$ et $\{b_1, b_2\}$. Pour ces deux marqueurs on définit les fréquences $f_{a_1}, f_{a_2}, f_{b_1}$ et f_{b_2} des allèles a_1, a_2, b_1 et b_2 respectivement. Les tableaux 1.1 et 1.2 ci-dessous décrivent les relations entre les fréquences des haplotypes et celles de chacun des allèles, i.e :

Haplotypes	Fréquences
a_1b_1	$f_{a_1b_1}$
a_1b_2	$f_{a_1b_2}$
a_2b_1	$f_{a_2b_1}$
a_2b_2	$f_{a_2b_2}$

TABLEAU 1.1 – Fréquences haplotypiques.

Allèles	Fréquences
a_1	$f_{a_1} = f_{a_1b_1} + f_{a_1b_2}$
a_2	$f_{a_2} = f_{a_2b_1} + f_{a_2b_2}$
b_1	$f_{b_1} = f_{a_1b_1} + f_{a_2b_1}$
b_2	$f_{b_2} = f_{a_1b_2} + f_{a_2b_2}$

TABLEAU 1.2 – Fréquences alléliques calculées à partir des fréquences haplotypiques.

Si les allèles aux deux marqueurs M_1 et M_2 s'associent aléatoirement alors les fréquences des quatre haplotypes seront égales aux produit des fréquences des allèles portés par ces haplotypes (tableau 1.3), sinon on aura une déviation, une quantité que l'on notera Δ, dans les fréquences haplotypiques (tableau 1.4).

Fréquences haplotypiques
$f_{a_1b_1} = f_{a_1}f_{b_1}$
$f_{a_1b_2} = f_{a_1}f_{b_2}$
$f_{a_2b_2} = f_{a_2}f_{b_2}$
$f_{a_2b_1} = f_{a_2}f_{b_1}$

TABLEAU 1.3 – Fréquences haplotypiques sous l'hypothèse d'association aléatoire.

Fréquences haplotypiques
$f_{a_1b_1} = f_{a_1}f_{b_1} + \Delta$
$f_{a_1b_2} = f_{a_1}f_{b_2} - \Delta$
$f_{a_2b_2} = f_{a_2}f_{b_2} + \Delta$
$f_{a_2b_1} = f_{a_2}f_{b_1} - \Delta$

TABLEAU 1.4 – Fréquences haplotypiques dans le cas d'association non aléatoire.

La quantité Δ est le coefficient de déséquilibre de liaison. On vérifie aisément à partir des expressions du tableau 1.4 que $\Delta = f_{a_1b_1}f_{a_2b_2} - f_{a_1b_2}f_{a_2b_1}$. Δ traduit donc la différence qui existe entre les produits des fréquences des chromosomes, d'une part porteurs des haplotypes a_1b_1 et a_2b_2, et d'autre part porteurs de a_1b_2 et a_2b_1. Ce coefficient est une quantité qui varie dans l'intervalle $[-\frac{1}{4}; \frac{1}{4}]$. Le maximum et le minimum de l'intervalle sont respectivement atteints pour $f_{a_1b_1} = f_{a_2b_2} = \frac{1}{2}$ et $f_{a_1b_2} = f_{a_2b_1} = \frac{1}{2}$. En effet si $f_{a_1b_1} = f_{a_2b_2} = \frac{1}{2}$, sachant que $\sum_{i,j} f_{a_ib_j} = 1$, alors $f_{a_1b_2} = f_{a_2b_1} = 0$ et le maximum est atteint. On montre de même, par symétrie, que la valeur minimale est atteinte pour $f_{a_1b_2} = f_{a_2b_1} = \frac{1}{2}$. Comme les fréquences haplotypiques $f_{a_ib_j}$ sont toujours positives, il vient que la valeur maximale et la valeur minimale calculées de Δ, sur une population, dépendent des fréquences alléliques. En effet on a :

$$\begin{cases} f_{a_1b_2} = f_{a_1}f_{b_2} - \Delta \geq 0 \quad \text{i.e. } \Delta \leq f_{a_1}f_{b_2} \ (1) \\ f_{a_2b_1} = f_{a_2}f_{b_1} - \Delta \geq 0 \quad \text{i.e. } \Delta \leq f_{a_2}f_{b_1} \ (2) \end{cases}$$

De (1) et (2) on voit que $\Delta \leq min(f_{a_1}f_{b_2}, f_{a_2}f_{b_1})$, ce qui définit une borne supérieure pour la valeur calculée de Δ. Par symétrie on a aussi $\Delta \geq max(-f_{a_1}f_{b_1}, -f_{a_2}f_{b_2}) \Leftrightarrow \Delta \geq -min(f_{a_1}f_{b_1}, f_{a_2}f_{b_2})$, ce qui définit une borne inférieure pour la valeur calculée de Δ. On remarque que la borne inférieure et la borne supérieure prennent respectivement les valeurs -$\frac{1}{4}$ et $\frac{1}{4}$ lorsque les fréquences alléliques sont équilibrées. En considérant ces bornes on peut donc obtenir une mesure normalisée du LD, appelée D', définie par le rapport suivant (Lewontin,1964) :

$$D' = \frac{|\Delta|}{D_{max}} \text{ tel que } D' \in [0;1]$$

où :

$$D_{max} = \begin{cases} min(f_{a_1}f_{b_2}, f_{a_2}f_{b_1}) & \text{si } \Delta > 0 \\ min(f_{a_1}f_{b_1}, f_{a_2}f_{b_2}) & \text{sinon.} \end{cases}$$

Il existe d'autres mesures du LD entre marqueurs bialléliques comme, par exemple, celle du r^2 (Hill et Robertson, 1968) définie de la façon suivante :

$$r^2 = \frac{\Delta^2}{f_{a_1}f_{a_2}f_{b_1}f_{b_2}}$$

Cette mesure est également à valeurs dans [0,1]. Elle est très dépendante des fréquences alléliques et ne prendra la valeur 1 que si chaque allèle du marqueur M_1 est associé à un unique allèle au marqueur M_2.

Les mesures r^2 et D' du LD se généralisent aux cas de loci M_1 et M_2 multialléliques avec $R = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\Delta_{ij}^2}{(1-\sum_{i=1}^{I}f_{a_i}^2)(1-\sum_{j=1}^{J}f_{b_j}^2)}$ (Maruyama, 1982) et $D' = \sum_{i=1}^{I}\sum_{j=1}^{J} f_{a_i}f_{b_j}|D'_{ij}|$ (Hedrick, 1987), où I et J sont les nombres d'allèles différents aux loci M_1 et M_2 respectivement, $\Delta_{ij} = f_{a_ib_j} - f_{a_i}f_{b_j}$, $D'_{ij} = \frac{\Delta_{ij}}{D_{max_{ij}}}$ et $D_{max_{ij}}$ vaut :

$$D_{max_{ij}} = \begin{cases} min(f_{a_i}f_{b_j}, (1-f_{a_i})(1-f_{b_j})) & \text{si } \Delta_{ij} < 0 \\ min(f_{a_i}(1-f_{b_j}), (1-f_{a_i})f_{b_j}) & \text{sinon.} \end{cases}$$

Ces mesures multialléliques sont importantes pour la modélisation du déséquilibre de liaison par haplotypes tel qu'on le verra dans la partie II de ce travail. Enfin, il existe aussi d'autres mesures du LD dérivées de Δ. Des revues de ces mesures ont été réalisées par Hedrick (1987), Devlin et Risch (1995) et Cierco-Ayrolles *et al.* (2004).

1.1.6.2 Lien entre le LD et la recombinaison

Les causes du LD entre deux loci sont diverses. Parmi les causes du déséquilibre de liaison on peut citer :

▪ La mutation : la création d'un nouvel allèle à un locus crée un déséquilibre avec les allèles des loci adjacents dans une population.

▪ La sélection, qui joue un rôle très important, car on choisit des reproducteurs qui transmettront leurs chromosomes préférentiellement dans une population.

▪ Le mélange de deux populations en équilibre de liaison qui constitue un ensemble globalement en déséquilibre de liaison dès lors que leurs fréquences alléliques sont respectivement différentes.

Cependant la recombinaison casse progressivement le LD entre les loci sur un chromosome. Le paragraphe suivant illustre comment un LD initial se dégrade avec la recombinaison entre deux loci en fonction du nombre de générations.

Soit $\Delta_0 = f_{a_1b_1} - f_{a_1}f_{b_1}$ le LD initial à une génération $t = 0$, dû par exemple à une mutation, entre deux loci dans une population. Soit r le taux de recombinaison entre ces deux loci. Si on suppose qu'il n'existe pas de variation pour les fréquences alléliques entre deux générations (le principe d'équilibre d'Hardy-Weinberg) alors le LD diminue d'un facteur de $(1-r)$ à chaque génération pour des accouplements aléatoires. En effet, pour $f'_{a_1b_1}$ la fréquence haplotypique de a_1b_1 à la génération suivante, on a $f'_{a_1b_1} = (1-r)f_{a_1b_1} + rf_{a_1}f_{b_1}$ car $(1-r)$ est la probabilité de non recombinaison entre les loci et $rf_{a_1}f_{b_1}$ représente la probabilité que l'on ait un nouvel haplotype a_1b_1 à la suite d'une recombinaison entre deux chromosomes portant respectivement les allèles a_1 et b_1. Or $f'_{a_1b_1} = (1-r)f_{a_1b_1} + rf_{a_1}f_{b_1} \Leftrightarrow f'_{a_1b_1} - f_{a_1}f_{b_1} = (1-r)f_{a_1b_1} + f_{a_1}f_{b_1}(r-1)$, d'où $\Delta_1 = (1-r)(f_{a_1b_1} - f_{a_1}f_{b_1}) = (1-r)\Delta_0$. En réitérant le même raisonnement on a $\Delta_2 = (1-r)\Delta_1 = (1-r)^2\Delta_0$. Par principe de récurrence, on a donc $\Delta_t = (1-r)^t\Delta_0$ où Δ_t est le LD calculé à la t-ième génération. On peut donc aussi approximer ce résultat par $\Delta_t = e^{-rt}\Delta_0$.

Ainsi on remarque que l'amplitude du LD à la génération t (i.e. $|\Delta_t|$) est d'autant plus proche du LD initial que r est proche de zéro, donc que les loci sont proches. On remarque également que $|\Delta_t|$ diminuera quand les loci seront éloignés. Au long terme, les recombinaisons sur un chromosome mènent à l'observation d'un LD entre loci proches

uniquement, que l'on peut exploiter pour la cartographie fine de QTL.

1.2 L'IBD, IBS et le LD

1.2.1 Les concepts d'IBD et d'IBS

On dit que deux segments ou portions chromosomiques sont IBD ("Identical by Descent") s'ils ont été hérités d'un même chromosome ancestral. En absence de mutation ils possèdent donc les mêmes allèles. Deux segments chromosomiques sont dits IBS ("Identical by State") s'ils possèdent les mêmes allèles. En absence de mutation à un locus on a donc :

- $IBD \Rightarrow IBS$
- $nonIBS \Rightarrow nonIBD$

Cette propriété n'est valable que localement car deux chromosomes non IBS pour les allèles aux marqueurs peuvent partager des segments IBD. Prédire le statut IBD de plusieurs segments chromosomiques consiste à prédire si ces derniers sont porteurs d'un même allèle, ou de plusieurs mêmes allèles, en un locus que l'on n'observe pas. On en déduira donc le statut IBS pour des allèles non observés. Cette prédiction des identités alléliques repose généralement sur la connaissance des haplotypes présents sur les segments chromosomiques portant ce locus.

Le concept d'IBD procure un cadre conceptuel afin d'effectuer une prédiction d'identité allélique. Ce concept repose sur la spécification d'un certain horizon temporel dans le passé. Or, deux segments chromosomiques qui ne seraient pas IBD pour un certain horizon temporel dans le passé peuvent néanmoins l'être pour un horizon plus reculé. Deux segments chromosomiques IBD pour un horizon reculé peuvent également, par le biais de la méiose, partager un matériel génétique plus important que deux segments chromosomiques IBD pour un horizon moins reculé. Ainsi ce concept pose un réel problème de définition. Le concept d'IBD est cependant très utilisé en cartographie de QTL comme on le verra dans la section suivante et dans la partie II de ce travail.

1.2.2 Le concept d'IBD en cartographie de QTL

La stratégie de cartographie qui découle du concept d'IBD est la suivante. La recherche de segments chromosomiques IBD, où une mutation causale est apparue (un allèle favorable ou une délétion par exemple) chez des individus présentant un même phénotype favorable devrait permettre de localiser le QTL. Cette stratégie repose sur l'idée qu'après des recombinaisons successives, au fil des générations, le LD sera élevé dans une région IBD petite autour du QTL. Ainsi rechercher des portions chromosomiques IBD entre individus présentant des phénotypes semblables devrait, en principe, permettre de bien exploiter le LD pour une localisation fine du QTL. Cette stratégie est proposée, par exemple, dans le travail de Meuwissen et Goddard (2001). La figure 1.2 illustre l'idée de la stratégie de cartographie reposant sur le concept d'IBD.

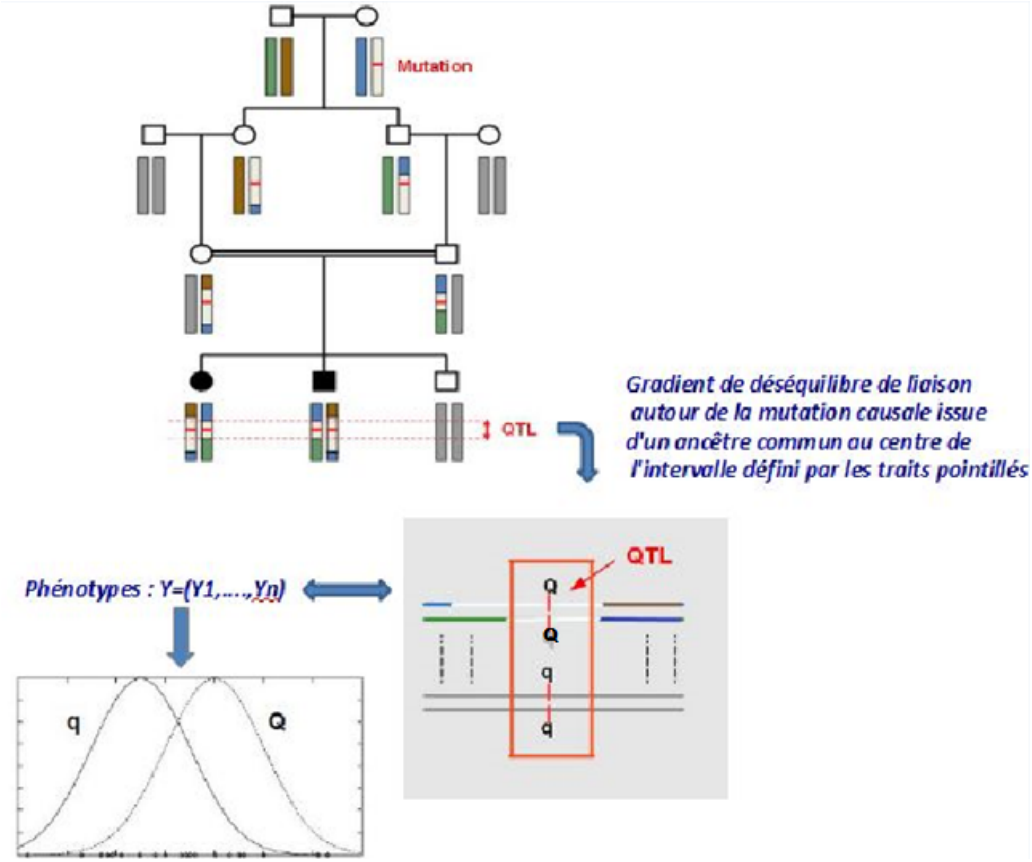


FIGURE 1.2 – Le gradient du LD dans une région IBD petite autour du QTL (d'allèle Q et q).

Chapitre 2

Modèles statistiques en cartographie de QTL

2.1 Etat de l'art sur les modèles généraux

Il existe beaucoup de modèles statistiques utilisables en cartographie de QTL. Toutefois on distingue nettement deux grandes catégories de modèles linéaires utilisés dans ce domaine. Ceux qui se placent dans le cadre où l'on a moins de paramètres que d'observations et qui ne pose pas de problème d'identifiabilité, sous certaines conditions, pour l'estimation des paramètres et ceux qui se placent dans le cadre opposé, en grande dimension, où l'on a plus de paramètres que d'observations. On notera $k \leq n$ et $k > n$ ces deux situations, où k et n représentent respectivement le nombre de variables et d'observations.

2.1.1 Le cas $k \leq n$

Considérons le modèle linéaire suivant :

$$Y = X\beta + \varepsilon$$

où la réponse Y est un vecteur de taille n, n mesures d'un phénotype par exemple, et X est le design reliant les effets de k SNP (ou haplotypes) et reliant éventuellement la moyenne aux observations. En se plaçant dans le cadre simple du modèle linéaire homoscédastique, on supposera que $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$ où σ est un paramètre positif inconnu et I_n est la matrice identité. Ce modèle admet une solution unique pour la minimisation de $\|\varepsilon\|_2^2$, où $\|.\|_2^2$ représente la norme euclidienne au carrée, lorsque les colonnes de X sont linéairement

indépendants. L'avantage d'un tel modèle, lorsque les hypothèses associées sont vérifiées, est qu'il procure des résultats d'optimalité sur le biais et la variance associés à l'estimateur de β qui minimise $\|\varepsilon\|_2^2$ (cf. 3.1.1). Cependant, si on suppose qu'il existe seulement un petit nombre de variables inférieur à k ayant réellement un effet sur Y, alors le fait de considérer les k variables conjointement ajoute du bruit dans l'estimation des paramètres de β. Ce bruit produit de la variance sur les prédictions du modèle et ce problème est connu, en statistique, sous le nom de compromis biais-variance associé au modèle. De plus, la minimisation de $\|\varepsilon\|_2^2$ n'admet plus de solution unique lorsque $k > n$ car les colonnes de X sont linéairement dépendantes dans ce cas.

2.1.2 Le cas $k > n$

Pour pallier au problème du bruit dans les estimations lorsque $k \leq n$, et afin de donner des solutions aux estimations lorsque $k > n$, diverses méthodes de régularisation ont été proposées pour construire des modèles plus parcimonieux sur les variables ayant réellement un effet. Par exemple, on peut citer la régression Ridge de Hoerl et Kennard (1970), le LASSO de Tibshirani (1996), l'Elastic Net de Zou et Hastie (2005) et certaines régressions Bayésiennes, définies en génétique, telles que Bayes A et Bayes B de Meuwissen *et al.* (2001) et Bayes Cπ et Bayes Dπ de Habier *et al.* (2011). Cependant les modèles qui se situent en grande dimension ne semblent pas encore être complètement d'actualité en cartographie de QTL. Ces modèles ont surtout été appliqués pour la prédiction des valeurs génétiques des individus en sélection. Toutes ces approches ont fait l'objet de comparaison sur données réelles ou simulées pour la détection de QTL par exemple dans le cadre de récents congrès tels que QTLMAS [1] XIII et XIV. Le LASSO, par exemple, perd en efficacité pour la sélection de variables lorsque les variables explicatives sont fortement corrélées ou lorsque le nombre de variables est nettement supérieur au nombre d'observations (Zou et Hastie, 2005), ce qui est régulièrement le cas en génétique. On remarquera au passage que la sélection de variables, dans le cadre de la génétique, revient à chercher les SNP potentiellement liés à des QTL ayant un effet significatif sur le génome.

L'utilisation pratique de certaines de ces approches posent des problèmes de paramétrisation. Les modèles Bayes A et Bayes B, par exemple, utilisent des a priori sur la

1. QTLMAS : congrès méthodologique visant à comparer des modèles et méthodes utilisés en sélection génomique et en cartographie de QTL

distribution de la variance des effets des SNP. La variance de l'effet de chaque SNP pour ces modèles est supposée suivre une Khi-deux inverse dont les paramètres doivent être spécifiés par l'utilisateur. Or ces paramètres ont une influence majeure sur la pénalisation appliquée aux effets des SNP. Le modèle Bayes B suppose également que la proportion de SNP ayant un effet est connue. Le lecteur pourra se référer à Gianola *et al.* (2009) pour une revue plus détaillée des principaux inconvénients des modèles BayesA et BayesB. Afin de palier à quelques-uns de ces inconvénients, des extensions nommées Bayes Cπ et Bayes Dπ ont été définies. Cependant, ces extensions ne permettent pas de s'affranchir totalement du problème de spécification lié aux paramètres.

2.1.3 Comparaison des différents modèles

Il n'existe pas de consensus général sur le modèle à utiliser en cartographie de QTL. En effet, les performances des différents modèles dépendent avant tout du jeu de données à analyser. Par exemple, le phénotype sera affecté par la structure génétique de la population, la distribution de l'effet du QTL et la taille de l'échantillon (Hayes, 2010). Néanmoins, les modèles mixtes tenant compte de la structuration génétique au sein d'une population semblent avoir fait leurs preuves et sont ceux les plus utilisés en cartographie de QTL. En effet, Wang *et al.* (1999) montrent que ces modèles sont suffisamment flexibles pour tenir compte de divers types d'interactions (SNP×QTL, QTL×environnement). Ces modèles permettent de modéliser des structurations complexes de la variance et de la covariance des données, et ont été appliqués dans beaucoup de situations (Smith *et al.* 2005 ; Boer *et al.*, 2007 ; Mathews *et al.*, 2008 ; van Eeuwijk *et al.*, 2010).

Bien que l'on ne puisse pas l'affirmer, les modèles utilisés dans le cadre de la grande dimension semblent ne pas apporter plus de puissance et de précision lorsque les cartes génétiques sont denses. Des analyses sur plusieurs jeux de données réelles ont montré, par exemple, que les tests en chaque point du génome associés à un modèle mixte uni-SNP et Bayes Cπ, étaient fortement corrélés, de l'ordre de 0.9 (Teyssèdre, 2011). Notons finalement que bon nombre des approches utilisées dans le cadre de la grande dimension peuvent être interprétées dans le cadre du modèle mixte. Le modèle correspondant au LASSO, par exemple, peut être formulé comme un modèle à effets aléatoires où les paramètres estimés sont distribués dans une double exponentielle (Tibshirani, 1996 ; Foster, 2006). De

même, la régression Ridge peut être interprétée, d'un point de vue Bayésien, comme un modèle à effets aléatoires où les paramètres estimés, et les résidus, sont distribués dans des gaussiennes (Lindley et Smith, 1972 ; Chen, 2009).

2.1.4 Conclusions

Quelque soit le modèle utilisé, le choix des variables explicatives pour la cartographie fine de QTL, un SNP à la fois, les haplotypes autour d'un locus, l'ensemble des SNP ou l'ensemble des haplotypes sur le génome, reste un problème ouvert qui suscite encore des recherches. Le choix des variables explicatives pour la cartographie de QTL devrait être fait en fonction du niveau de LD entre ces variables et les QTL sur le génome (Calus *et al.*, 2009).

Par exemple un LD entre une position testée et un QTL peut être mieux décrit, localement sur le génome, par l'utilisation d'haplotypes au lieu d'un seul SNP. En revanche l'utilisation d'un trop grand nombre de SNP peut diminuer le potentiel du LD en cartographie fine de QTL, car elle ne tient pas compte que du LD local, nécessaire à une cartographie fine, mais également du LD entre un QTL et des SNP éloignés. Ce LD à longue distance induit une sous-estimation de l'effet d'un QTL s'il est présent.

A ce jour, l'utilisation d'un modèle linéaire mixte, à travers une procédure de génome scan où l'on analyse séquentiellement des portions du génome afin de déterminer l'existence, ou non, de QTL sur ces portions est toujours une méthode de choix pour la cartographie fine de QTL. La section suivante décrit ce modèle dans le cadre général et ses particularisations en cartographie de QTL.

2.2 Le modèle mixte à effets fixes et aléatoires

2.2.1 Le modèle général

Le modèle mixte, à $L+1$ facteurs aléatoires indépendants, est couramment utilisé en génétique quantitative afin de traiter les effets dus à la structuration génétique, qu'on appelle effets polygéniques en génétique, en aléatoires (Fisher, 1918). Il est couramment utilisé, par exemple, pour les analyses d'association en cartographie fine de QTL. Avec

les notations de Rao et de Kleffe (1988) ce modèle (2.1) s'écrit de la façon suivante :

$$Y = X\beta + \epsilon\,;\ \epsilon \sim \mathcal{N}_n(0, V) \tag{2.1}$$

où ϵ se décompose comme une combinaison linéaire de vecteurs aléatoires structuraux $(u_l)_{0\leq l\leq L}$ non observables, i.e. $\epsilon = \sum_{l=0}^{L} Z_l u_l$ où $(Z_l)_{0\leq l\leq L}$ correspond aux matrices d'incidence respectives de chacun de ces vecteurs aléatoires. Etant donné que le modèle polygénique suppose que le phénotype d'un individu est la somme des effets d'un très grand nombre de gènes (modèle infinitésimal), on peut donner une approximation de la matrice de variance-covariance de ces effets dans un modèle mixte. Le modèle (2.2) avec une composante aléatoire pour les effets polygéniques s'écrit habituellement avec les notations suivantes :

$$Y = X\beta + \sum_{l=0}^{L=1} Z_l u_l\,;\quad u_1 \sim \mathcal{N}_n(0, A\sigma^2_{u_1}),\ \ u_0(=\varepsilon) \sim \mathcal{N}_n(0, I_n\sigma^2_{u_0}) \tag{2.2}$$

ou encore :

$$Y = X\beta + Zu + \varepsilon\ \ ;\ \ u \sim \mathcal{N}_n(0, A\sigma^2_u),\ \ \varepsilon \sim \mathcal{N}_n(0, I_n\sigma^2_\varepsilon)$$

où :

- Y (n x 1) est le vecteur des phénotypes
- β (k x 1) est le vecteur des effets fixes
- X (n x k) est la matrice d'incidence reliant les effets fixes aux individus
- $Z = I_n$ (n x n) est la matrice d'incidence reliant les effets polygéniques aux individus
- u (n x 1) est le vecteur aléatoire des effets polygéniques
- A (n x n) est la matrice de variance-covariance de u (la matrice d'apparentement)
- ε (n x 1) est le vecteur aléatoire des résidus
- σ^2_u et σ^2_ε sont respectivement la variance polygénique et résiduelle

L'idée générale du modèle (2.2) est de tenir compte des corrélations entre les effets polygéniques d'individus apparentés via la matrice A. Cette matrice peut être construite soit à partir de l'information marqueur ou pedigree ou à partir d'une combinaison de ces deux sources d'information.

2.2.1.1 Le modèle à effets fixes au QTL

Ce modèle est le même que celui défini par l'équation (2.2), sauf que le vecteur β inclue les effets des allèles au locus testé parmi d'autres effets fixes. Les autres effets fixes peuvent par exemple correspondre à des moyennes familiales ou une moyenne générale et ainsi de suite. Ce modèle est utilisé dans la partie III de ce travail pour la discrimination entre des analyses d'association et des analyses de liaison. Remarquons que les "Genome Wide Association Studies" (GWAS) dans la littérature sont basés sur ce type de modèle.

2.2.1.2 Le modèle à effets aléatoires au QTL

Ce modèle est une extension de celui défini par l'équation (2.2) en rajoutant une composante aléatoire additionnelle. Ce modèle est utilisé dans la partie II de ce travail pour la discrimination entre des analyses d'association utilisant des haplotypes. Remarquons au passage que les allèles en un seul marqueur, un SNP par exemple, définissent des haplotypes en un seul locus. Ce modèle se décompose de la manière suivante :

$$Y = X\beta + Z_1 u_1 + Z_2 u_2 + \varepsilon \ ; \quad u_1 \sim \mathcal{N}_n(0, A\sigma^2_{u_1}), u_2 \sim \mathcal{N}_h(0, H\sigma^2_{u_2}), \ \varepsilon \sim \mathcal{N}_n(0, I_n\sigma^2_\varepsilon) \tag{2.3}$$

où :

- Z_2 (n x h) est la matrice d'incidence reliant les effets des h haplotypes au QTL aux individus
- u_2 (h x 1) est le vecteur aléatoire des effets des h haplotypes au QTL
- H (h x h) est la matrice de variance-covariance de u_2
- $\sigma^2_{u_2}$ est la variance de l'effet du QTL

Diverses méthodes ont été proposées pour la construction de la matrice H par l'utilisation des haplotypes. L'un des objectifs des chapitres 4 et 5 est de comparer ces méthodes que l'on nommera prédicteurs d'identité allélique, on y reviendra. Bien que l'on ne puisse pas l'affirmer, il semblerait que les modèles mixtes avec des effets aléatoires au QTL sont plus robustes et précis pour la cartographie que ceux où ces effets sont fixés (Kolbehdari *et al.*, 2005 ; Boleckova *et al.*, 2012). Cela est probablement dû à la propriété des modèles à effets aléatoires à rétrécir les effets estimés pour des petits effectifs (i.e. allèles ou haplotypes peu fréquents au QTL).

En supposant que les observations $Y \in \mathbb{R}^n$ suivent bien une loi normale multidimensionnelle, i.e. $Y \sim \mathcal{N}_n(X\beta, V(\theta))$, la vraisemblance pour le modèle (2.3) est donnée par :

$$L(\beta, \theta; Y) = |V|^{-\frac{1}{2}}(2\pi)^{-\frac{n}{2}} exp\Big(-\frac{1}{2}(Y - X\beta)' V^{-1}(Y - X\beta)\Big) \tag{2.4}$$

où :

- $V = V(\theta) = Z_1 A \sigma^2_{u_1} Z_1' + Z_2 H \sigma^2_{u_2} Z_2' + I_n \sigma^2_\varepsilon$
- $\theta = (\sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_\varepsilon)$

On rappelle que la vraisemblance d'un modèle probabiliste-statistique est la probabilité que les observations soient la réalisation d'un échantillon théorique de la loi de probabilité associée au modèle. Des estimations pour les composantes du vecteur θ, nécessaires au calcul d'une statistique de test pour la détection de QTL, peuvent être obtenues en maximisant directement $L(\beta, \theta; Y)$. Cependant on verra en 3.1.1 pourquoi il est préférable d'utiliser une variante de cette fonction qui est la vraisemblance restreinte (Searle, 1987, 1989 ; Searle, Casella et McCulloch 1992 ; Foulley, 1993).

2.3 Les modèles de liaison et les modèles d'association

Il existe deux grandes classes de modèles linéaires qui trouvent leurs fondements en génétique et qui sont utilisés en génome scan. Ces modèles exploitent différemment le LD entre les allèles observés d'un marqueur et les allèles non observés d'un QTL. La première classe est appelée analyse de liaison, d'acronyme anglo-saxon LA pour "Linkage Analysis", et la deuxième est appelée analyse d'association d'acronyme anglo-saxon LDA pour "Linkage Disequilibrium Analysis". Les sous-sections suivantes décrivent ces modèles dans le cadre d'un seul marqueur testé sur le génome (approche uni-SNP), bien qu'ils peuvent être généralisés à l'utilisation d'haplotypes. Ces modèles sont des cas particuliers du modèle mixte général défini en 2.2.1.

2.3.1 Les modèles de liaison (LA)

Les modèles LA, très largement utilisés ces 20 dernières années pour la cartographie, sont basés sur l'observation de la transmission d'allèles au sein de familles d'individus au cours de quelques générations. Ces modèles font l'hypothèse de la co-transmission d'allèles

à un marqueur testé avec certains allèles au QTL, dans une famille, générant ainsi une variation du phénotype étudié au sein de celle-ci. Ces approches reposent sur le LD intra famille entre les allèles au marqueur testé et les allèles au QTL. Elles testent le contraste statistique, au marqueur testé, entre deux groupes d'individus ayant respectivement reçu un allèle différent d'un parent hétérozygote. Par exemple, un modèle de liaison pour le phénotype Y_{ils}, d'un individu s ayant reçu l'allèle $l \in \{a_1, a_2\}$ d'un père i, peut s'écrire de la façon suivante :

$$Y_{ils} = \mu + \nu_i + X_{ils}\eta_i + \varepsilon_{ils}$$

où μ est la moyenne générale de la population, ν_i correspond à un effet moyen du père i, η_i est l'effet du marqueur testé dans la famille i et ε_{ils} est un bruit que l'on suppose gaussien. X_{ils} est la variable du design décrivant l'allèle l transmis à l'individu s par le père i, X_{ils} peut par exemple être à valeur dans $\{-1, 1\}$. L'estimateur de η_i pour un tel modèle s'écrit fréquemment comme la différence de moyennes (i.e. le contraste) entre les individus porteurs de a_1 et ceux porteurs de a_2 dans la famille i.

Les modèles LA appliqués à une ou deux générations sont peu précis pour la localisation d'un QTL puisqu'il ne se produit que peu d'évènements de recombinaison, intra famille, entre ce QTL et des marqueurs éloignés (Bodmer, 1986 ; Boehnke, 1994 ; Fan et Xiong, 2002). Par définition, l'application des modèles LA nécessite d'avoir plusieurs familles de grande taille. En effet, ces approches nécessitent d'une part d'avoir des pères (ou des mères) hétérozygotes, et d'autre part, d'être en mesure de repérer la transmission des allèles à la descendance ce qui n'est pas toujours possible (descendants hétérozygotes de mères non génotypées ou trio hétérozygotes).

Ces modèles sont donc parfois sujets, par manque d'effectifs adaptés (i.e. familles non adaptées), à un manque de puissance pour la détection de QTL (Sham *et al.*, 2000). Par ailleurs ces modèles font intervenir un grand nombre de paramètres à estimer, ce qui contribue également parfois à un manque de puissance. A l'inverse, ces approches sont généralement robustes aux faux positifs contrairement aux modèles LDA (Jung *et al.*, 2005). En effet, en comparant les effets des allèles intra famille, on s'affranchit de confusion entre effets alléliques et effets familiaux. Les approches LA peuvent aussi donner de la puissance, contrairement aux approches LDA, lorsqu'il y a très peu de marqueurs.

L'un des modèles LA des plus connus est l'"interval mapping" (IM) de Lander et Botstein (1989). Ce modèle exploite le fait que les QTL sont localisés entre deux marqueurs et prend en compte les génotypes et les taux de recombinaison, à gauche et à droite, de la position testée afin de déterminer les probabilités des allèles portés par les individus à cette position. Une fois ces probabilités déterminées, les contrastes statistiques entre les allèles transmis intra familles sont testés pour valider la présence ou non d'un QTL. Le principal inconvénient de ce modèle de liaison est qu'il est généralement peu précis pour la cartographie (Haley et Knott, 1992 ; Martinez et Curnow, 1992).

Ce modèle et des variantes de celui-ci ont été développés dans le cadre du maximum de vraisemblance et de la régression (Kao, 2000). Par exemple, on peut citer Elsen *et al.* (1999), pour des dispositifs de demi frères et/ou plein frères, Le Roy *et al.* (1998) pour le cas d'un mélange de plein frère et demi frères, et Knott *et al.* (1996) pour un dispositif de demi frères uniquement. Remarquons que les modèles LA furent développés il y a quelques décennies, quand on ne disposait pas encore de cartes génétiques très denses. Les nouvelles cartes génétiques permettent aujourd'hui d'appréhender le LD d'une autre façon, non plus forcément au niveau d'une famille mais au niveau de toute une population. Et c'est de cette façon d'appréhender le LD qu'ont émergé la classe des modèles dits LDA.

2.3.2 Les modèles d'association (LDA)

Les modèles LDA sont connus pour leur haute précision en cartographie par l'exploitation local du LD populationnel (Fan et Xiong, 2002 ; Aranzana *et al.*, 2005 ; Fan *et al.*, 2006). En contraste du LD intra famille, le LD populationnel est celui mesuré sur l'ensemble des individus d'une population. Par exemple, un modèle d'association couramment utilisé en génétique s'écrit de la façon suivante :

$$Y_i = \mu + \underline{X}_i\alpha + \varepsilon_i$$

où Y_i est le phénotype d'un individu i, μ est la moyenne générale, α est l'effet du marqueur testé et ε_i est un bruit que l'on suppose gaussien. $\underline{X}_i$ est la variable du design décrivant le génotype de l'individu i au marqueur testé, $\underline{X}_i$ peut par exemple être à valeur dans $\{-2, 0, 2\}$. L'estimateur de α pour un tel modèle s'écrit fréquemment comme la différence de moyennes entre les individus qui ont le génotype a_1/a_1 et ceux qui ont le génotype a_2/a_2. Les modèles LDA exploitent le LD entre les allèles au marqueur testé et les allèles au QTL,

sur l'ensemble de la population, contrairement aux modèles LA. Ces modèles sont réputés pour être puissants (Long et Langley, 1999 ; Sham *et al.*, 2000 ; Flint-Garcia *et al.*, 2005). Les GWAS sont basés sur des modèles LDA (Stranger *et al.*, 2011 ; Bush et Moore, 2012 ; Korte et Farlow, 2013). La bonne puissance associée aux modèles d'association provient, d'une part, du fait qu'ils ne demandent pas la construction de familles adaptées et qu'ils utilisent en conséquence tous les individus de la population, et d'autre part, ces modèles reposent fréquemment sur un nombre restreint de paramètres à estimer. Remarquons que les modèles LDA peuvent s'écrire en fonction des familles dans une population, tel qu'on le verra dans le chapitre 6, bien que ces modèles ne trouvent pas leur fondement dans des approches familiales comme les modèles LA.

Les modèles d'association ont cependant une grande faiblesse qui est le manque de maîtrise des faux positifs, si on ne corrige pas pour une structuration génétique existante, due aux relations entre individus dans une population (Jung *et al.*, 2005 ; Manenti *et al.*, 2009 ; Platt *et al.*, 2010). De ce fait, des modèles d'association corrigés pour la structure de population ont été proposés afin de tenir compte d'effets non attribuables au QTL (i.e. les effets polygéniques). Ces modèles d'association sont actuellement décrits dans la littérature comme étant les plus performants pour la cartographie fine, car ils peuvent aider à disséquer les effets polygéniques de l'effet d'un QTL s'il existe (Newman *et al.*, 2001 ; Yu *et al.*, 2005 ; Zhang *et al.*, 2009).

Chapitre 3

Méthodes d'estimation et statistiques de test en cartographie de QTL

L'objectif de ce chapitre est de construire une démarche sur le choix des méthodes d'estimation et des statistiques de test utilisées dans la partie II et III de ce travail de thèse. Des éléments de dérivation sont abordés afin de justifier la démarche à la fois d'un point de vue théorique et pratique. Certaines notions sont également rappelées, telles que la projection et la convexité, car elles s'avéreront utilisées dans la partie II et III de ce travail. Le lecteur pourra simplement, s'il le souhaite, retenir les éléments clefs en début ou en fin de chaque section (ou sous-section).

3.1 Méthodes d'estimation

On verra dans la sous-section suivante pourquoi il est préférable de maximiser une autre fonction objective, que celle définie en (2.4) dans le chapitre 2, afin d'obtenir des estimations pour les composantes de la variance associées au modèle mixte utilisé en cartographie de QTL.

3.1.1 Les méthodes ML et REML

Avant de redéfinir la méthode REML (Patterson et Thompson, 1971) il est utile de rappeler brièvement la méthode du maximum de vraisemblance (ML), dans le cadre simple du modèle linéaire classique (homoscédastique), afin de comprendre pourquoi la méthode REML est la plus adaptée en détection de QTL.

3.1.1.1 La méthode ML ("Maximum likelihood")

Soit le modèle linéaire classique (homoscédastique) suivant :

$$Y = X\beta + \varepsilon \,;\, \varepsilon \sim \mathcal{N}_n(0, I_n\sigma^2) \tag{3.1}$$

Ce modèle est emboîté au modèle (2.1) et on suppose ici, par définition, que les variables $(Y_i)_{i\in\{1,..,n\}}$ sont indépendantes et identiquement distribuées (i.i.d) contrairement à (2.1). La log-vraisemblance pour ce sous-modèle est donnée par : $lnL(\beta, \sigma^2; Y) = -\frac{n}{2}ln\sigma^2 - \frac{n}{2}ln2\pi - \frac{1}{2\sigma^2}\|Y - X\beta\|_2^2$. Maximiser $lnL(\beta, \sigma^2; Y)$ équivaut aussi à minimiser la fonction $H(\beta, \sigma^2) = -lnL(\beta, \sigma^2; Y)$.

Minimisation de H en β :

On a $H(\beta, \sigma^2) = \frac{n}{2}ln\sigma^2 + \frac{n}{2}ln2\pi + \frac{1}{2\sigma^2}\|Y - X\beta\|_2^2$. Or comme tous les termes définis par H sont positifs et ne dépendent pas de β, sauf pour $\|Y - X\beta\|_2^2$, il vient que la minimisation de H en β revient à la minimisation de ce seul terme. Soit E l'espace vectoriel engendré par les colonnes de X, i.e. $E = \{X\beta; \beta \in \mathbb{R}^k\} = vect(X)$. E est un sous-espace vectoriel de $\mathbb{R}^n$ et on suppose que X est de plein rang, i.e. $rang(X) = dim(vect(X)) = k = dim(E)$. Donc si on pose $Z = X\beta$ alors $\hat{Z} = \underset{Z\in E}{min}\|Y - Z\|_2^2 = P_E Y$ (projection orthogonale de Y sur E). Comme la projection orthogonale est unique on a une unique solution pour Z par injectivité. Donc on a $\hat{Z} = X\hat{\beta} = P_E Y$ par injectivité (i.e. $\mathbb{E}[Y|X] = P_E Y = X\hat{\beta}$ dans ce cadre gaussien particulier).

Proposition (Projecteur orthogonal P_E)**:** $P_E = X(X'X)^{-1}X'$

Un projecteur orthogonal P_E est caractérisé par *i)* $\forall x \in \mathbb{R}^n, P_E x \in E$ et *ii)* $\forall x \in \mathbb{R}^n, \forall y \in E, x - P_E x \perp y$. Ces deux propriétés sont relativement simples à démontrer et aident à la compréhension de la notion d'orthogonalité associée au projecteur P_E.

Preuve de *i)* et *ii)* :

i) Soit $x \in \mathbb{R}^n$, on a $X(X'X)^{-1}X'x = \underset{n\times k}{X}\Big[\underset{k\times 1}{(X'X)^{-1}X'x}\Big] \in E$

ii) Soit $y \in E$ donc $y = X\beta$ et on dénote par $< .,. >$ le produit scalaire usuel sur $\mathbb{R}^n$, on a :

$$\begin{aligned} < y, x - X(X'X)^{-1}X'x > &=< X\beta, x - X(X'X)^{-1}X'x > \\ &=< \beta, X'\left(x - X(X'X)^{-1}X'x\right) > \text{ (propriété de } < .,. >) \end{aligned}$$

$$=< \beta, X^{'}x - X^{'}X\underbrace{(X^{'}X)^{-1}X^{'}x}_{Id} >$$
$$=< \beta, 0 >= 0 \qquad \square$$

Donc on a d'une part que P_E est une matrice (symétrique) de projection orthogonale, sur l'espace vectoriel engendré par les colonnes de X, et d'autre part que $X\hat{\beta} = P_E Y$. On a donc $(X^{'}X)^{-1}X^{'}.X\hat{\beta} = (X^{'}X)^{-1}X^{'}.P_E Y \iff \hat{\beta}_{ML} = (X^{'}X)^{-1}X^{'}Y$. On remarquera que $\hat{\beta}_{ML}$, dans le cadre gaussien, est le même que celui obtenu par moindre carrés ordinaires (OLS : "Ordinary Least Squares"). De plus $\hat{\beta}_{ML}$ dans ce cadre est un estimateur optimal parmi les estimateurs sans biais (MVUE : "Minimum Variance Unbiased Estimator"), car il atteint la borne de Cramer-Rao. Si les observations ne sont pas gaussiennes $\hat{\beta}_{ML}$ devient alors, d'après le théorème de Gauss-Markov, un estimateur optimal seulement parmi les estimateurs linéaires et sans biais (BLUE : "Best Linear Unbiased Estimator").

Remarque :

L'objet P_E, défini par la proposition précédente, s'avère être fondamentale et utile pour la construction des statistiques de test en général (test du rapport de vraisemblance, test de Fisher...). Cet objet intervient également dans le calcul des paramètres de décentrage, tel qu'on le verra dans la partie III de ce travail, où l'on compare la puissance et la robustesse de certains modèles uni-SNP de liaison avec des modèles uni-SNP d'association.

Minimisation de H en σ^2 :

On a $\dfrac{\partial H(\hat{\beta}_{ML}, \sigma^2)}{\partial \sigma^2} = \dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\|Y - X\hat{\beta}_{ML}\|_2^2 = 0 \iff \hat{\sigma}^2_{ML} = \dfrac{1}{n}\|Y - X\hat{\beta}_{ML}\|_2^2$. Cet estimateur $\hat{\sigma}^2_{ML}$ de la variance est biaisé. En effet on peut montrer que $\mathbb{E}[\hat{\sigma}^2_{ML}] = \dfrac{n-k}{n}\sigma^2$ et l'estimateur non biaisé de σ^2 est donc donné par $\dfrac{1}{n-k}\|Y - X\hat{\beta}_{ML}\|_2^2$. On remarque cependant que si le vecteur Y est d'espérance le vecteur nul, i.e. $\mathbb{E}(Y) = X\beta = 0_n$, alors le nouvel estimateur $\hat{\sigma}^2_{ML} = \dfrac{1}{n}\|Y\|^2$ est sans biais. En effet comme on suppose que les $(Y_i)_{i\in\{1,..,n\}}$ sont i.i.d on a $\mathbb{E}(\hat{\sigma}^2_{ML}) = \dfrac{1}{n}\mathbb{E}(< Y, Y >) = \dfrac{1}{n}\sum_i^n \mathbb{E}(Y_i^2) = \mathbb{E}(Y_i^2) = \sigma^2$.

Il vient donc que l'on cherchera préférentiellement à construire un vecteur U, qui sera fonction de Y et dont l'espérance sera nulle, tel que lorsque l'on maximisera sa vraisemblance alors l'estimation de σ^2 sera sans biais. La méthode REML est basée sur cette transformation de la réponse Y et elle est très utilisée en cartographie de QTL.

3.1.1.2 La méthode REML ("Restricted Maximum likelihood")

La méthode REML consiste à maximiser la vraisemblance du vecteur $U^{=}A'Y$, où A (n x $n-k$) est une base orthonormée de $E^{\perp}$ qui est l'orthogonal de E. En effet on peut décomposer l'espace initial $\mathbb{R}^n$ en deux sous-espaces supplémentaires de la façon suivante $\mathbb{R}^n = E \oplus E^{\perp}$, et il existera ainsi toujours au moins une base orthonormée de $E^{\perp}$. $E^{\perp}$, aussi appelé l'espace des contrastes d'erreur (Harville, 1977), est de dimension $n-k$ et les coordonnées de U dans A s'écrivent donc de la manière suivante $U = (u_1, .., u_{n-k})$.

L'espérance du vecteur U est nul par construction car les coordonnées des vecteurs colonnes de $X\beta$ dans la base A valent toutes 0 par orthogonalité. En effet, en choisissant la base canonique pour A, on a $\mathbb{E}(U) = \mathbb{E}(A'Y) = A'\mathbb{E}(Y) = A'.X\beta = 0$ et $Var(U) = A'Var(Y)A = \sigma^2 A'A = \sigma^2 I_{n-k}$. Il vient donc que la log-vraisemblance de U est donnée par $lnL(\sigma^2) = -\frac{n-k}{2} ln\sigma^2 - \frac{n-k}{2} ln2\pi - \frac{1}{2\sigma^2}\|U\|^2$. Et l'estimateur de la variance de son maximum de vraisemblance est donnée par $\hat{\sigma}^2_{ML(U)} = \frac{1}{n-k}\|U\|^2 = \frac{1}{n-k} < U, U >= \frac{1}{n-k} Y'AA'Y = \frac{1}{n-k} Y'A(A'A)^{-1}A'Y$ (puisque $A'A = I_{n-k}$). On note $P_{E^{\perp}} = A(A'A)^{-1}A'$ la matrice de projection sur $E^{\perp}$ pour la métrique A. Comme $P_{E^{\perp}}$ est symétrique ($P_{E^{\perp}} = P'_{E^{\perp}}$) et que $P_{E^{\perp}} = P_{E^{\perp}} P_{E^{\perp}}$ on a $\hat{\sigma}^2_{ML(U)} = \frac{1}{n-k} Y' P_{E^{\perp}} Y = \frac{1}{n-k} Y' P'_{E^{\perp}} P_{E^{\perp}} Y = \frac{1}{n-k}\|P_{E^{\perp}} Y\|^2$. Or $P_{E^{\perp}} Y = (I_n - P_E)Y = Y - P_E Y = Y - X\hat{\beta}_{ML} \Longrightarrow \hat{\sigma}^2_{ML(U)} = \frac{1}{n-k}\|Y - X\hat{\beta}_{ML}\|^2$ (l' estimateur non biaisé de σ^2).

En conclusion, maximiser la vraisemblance de U (ou la vraisemblance restreinte de Y) permet d'obtenir un estimateur de la variance $\hat{\sigma}^2_{ML(U)}$ (ou $\hat{\sigma}^2_{REML}$) non biaisé et indépendamment du choix de A.

Cependant si ε a une structure quelconque pour sa matrice V de variance covariance, tel que dans le cadre du modèle (2.1) défini dans le chapitre 2 où $\varepsilon = \epsilon$, alors le raisonnement précédent reste valable. V est par définition semi-définie positive, car c'est une matrice de variance-covariance, et on peut donc lui appliquer une décomposition de Cholesky. On peut donc écrire $V = \sigma^2 LL'$, à une constante σ^2 près, où L est une matrice triangulaire inférieure de plein rang. Or dans le cadre du modèle (2.1) on a :

$$Y = X\beta + \epsilon \Longleftrightarrow L^{-1}Y = L^{-1}X\beta + L^{-1}\epsilon \Longleftrightarrow Y^* = X^*\beta + \epsilon^*$$

où $Y^* = L^{-1}Y$, $X^* = L^{-1}X$ et $\epsilon^* = L^{-1}\epsilon$. Le modèle correspondant à Y^* est équivalent au modèle (2.1) et on a $\mathbb{E}(Y^*) = X^*\beta$ et $Var(Y^*) = L^{-1}Var(\epsilon)(L^{-1})' = L^{-1}\sigma^2 LL'(L^{-1})'$ $= \sigma^2 L'(L^{-1})' = \sigma^2(L^{-1}L)' = \sigma^2 I_n$. Ainsi, par transformation linéaire, on se ramène au cas précédent du modèle linéaire classique (3.1) avec des résidus indépendants et homoscédastiques et on montre la validité du raisonnement pour (2.1). On peut vérifier aisément que l'expression de la log-vraisemblance restreinte associée au modèle (2.1) est donnée par :

$$lnL(\theta; A'Y) = -\frac{1}{2}((n-k)ln(2\pi) + ln|A'VA| + Y'A(A'VA)^{-1}A'Y)$$

Lorsque A est la base canonique de $E^{\perp}$ on a $A'A = I_{n-k}$ et l'expression précédente peut se mettre sous la forme suivante (Harville, 1977 ; LaMotte, 2007) :

$$\begin{aligned} lnL(\theta; A'Y) = -\frac{1}{2}\Big((n-k)ln(2\pi) - ln|X'X| + ln|V| + ln|X'V^{-1}X| + \\ (Y - X\hat{\beta})'V^{-1}(Y - X\hat{\beta})\Big) \end{aligned} \tag{3.2}$$

Toutefois il existe une autre dérivation de la vraisemblance restreinte dans le cadre bayésien, qui est plus rapide d'accès, tel qu'on le verra en 3.1.2. Cette dérivation est importante car elle permet de faciliter la mise en oeuvre d'algorithmes, tel que l'EM par exemple, pour l'estimation des composantes de la variance dans le cadre du modèle mixte. Enfin, retenons de cette sous-section que la méthode REML permet d'avoir des estimations non-biaisés des composantes de la variance, ce qui est un avantage du REML par rapport au ML en cartographie de QTL.

3.1.2 Interprétation bayésienne de la vraisemblance restreinte

La vraisemblance d'un modèle statistique est une fonction de probabilité des observations et des paramètres associés au modèle. Elle est la probabilité que les observations soient la réalisation d'un échantillon théorique de la loi de probabilité associée au modèle sachant les paramètres de cette dernière. On peut donc aussi écrire la vraisemblance associée à un modèle de la façon suivante : $L(\beta, \theta; Y) = P(y|\beta, \theta)$ où $\beta = (\beta_1, .., \beta_k)$ et $\theta = (\sigma^2_{u_0}, \sigma^2_{u_1}, .., \sigma^2_{u_L})$ sont les paramètres définis dans le cadre du modèle (2.1). Cette façon de ré-écrire la vraisemblance a l'avantage de mieux expliciter les densités conditionnelles utilisées dans le cadre bayésien. La vraisemblance restreinte dans le cadre bayésien se définit comme étant la vraisemblance marginale $P(y|\theta)$ obtenue en intégrant β selon une

loi *a priori* uniforme et impropre $\pi(\beta|\theta) = \mathbb{1}_{]-\infty;+\infty[^k}(\beta)$ (Harville, 1974), i.e :

$$P(y|\theta) = \int_{\beta\in\mathbb{R}^k} P(y,\beta|\theta)\mathbf{d}\beta = \int_{\beta\in\mathbb{R}^k} P(y|\beta,\theta)\pi(\beta|\theta)\mathbf{d}\beta = \int_{\beta\in\mathbb{R}^k} P(y|\beta,\theta)\mathbf{d}\beta \qquad (3.3)$$

où $\mathbf{d}\beta = \mathbf{d}\beta_1\mathbf{d}\beta_2...\mathbf{d}\beta_k$. On notera que β est considéré ici comme un vecteur aléatoire parasite, que l'on élimine par intégration, et la deuxième égalité de (3.3) s'obtient aisément en reportant l'égalité $P(y,\beta,\theta) = P(y|\beta,\theta)P(\beta,\theta)$ dans l'égalité $P(y,\beta|\theta) = \dfrac{P(y,\beta,\theta)}{P(\theta)} = P(y|\beta,\theta)P(\beta|\theta)$, bien que l'on choisisse ici une loi *a priori* uniforme et impropre $\pi(\beta|\theta)$ pour la loi de $\beta|\theta$. On rappelle aussi que dans le cadre du modèle (2.1) la vraisemblance $P(y|\beta,\theta)$ est donnée par :

$$P(y|\beta,\theta) = (2\pi)^{-\frac{n}{2}}|V|^{-\frac{1}{2}}exp(-\frac{1}{2}(Y-X\beta)'V^{-1}(Y-X\beta))$$

Or l'expression $(Y-X\beta)'V^{-1}(Y-X\beta)$ peut aussi se décomposer en $(Y-X\hat{\beta})'V^{-1}(Y-X\hat{\beta})+(\beta-\hat{\beta})'X'V^{-1}X(\beta-\hat{\beta})$ ce qui est nécessaire au calcul de la vraisemblance marginale (Gianola *et al.*, 1986). En effet, on a :

$$\begin{aligned}
(Y-X\beta)'V^{-1}(Y-X\beta) &= (Y-X(\beta-\hat{\beta}+\hat{\beta}))'V^{-1}(Y-X(\beta-\hat{\beta}+\hat{\beta})) \\
&= ((Y-X\hat{\beta})-X(\beta-\hat{\beta}))'V^{-1}((Y-X\hat{\beta})-X(\beta-\hat{\beta})) \\
&= ((Y-X\hat{\beta})'-(\beta-\hat{\beta})'X')V^{-1}((Y-X\hat{\beta})-X(\beta-\hat{\beta})) \\
&= \Big[(Y-X\hat{\beta})'V^{-1}(Y-X\hat{\beta})-(Y-X\hat{\beta})'V^{-1}X(\beta-\hat{\beta}) \\
&\quad -(\beta-\hat{\beta})'X'V^{-1}(Y-X\hat{\beta})+(\beta-\hat{\beta})'X'V^{-1}X(\beta-\hat{\beta})\Big]
\end{aligned}$$

En remplaçant $\hat{\beta}$ par $(X'V^{-1}X)^{-1}X'V^{-1}Y$ (l'estimateur des moindres carrés généralisés (MCG)) dans les termes $(Y-X\hat{\beta})'V^{-1}X(\beta-\hat{\beta})$ et $(\beta-\hat{\beta})'X'V^{-1}(Y-X\hat{\beta})$ de la dernière égalité, et en continuant les calculs, on vérifie aisément que $(Y-X\hat{\beta})'V^{-1}X(\beta-\hat{\beta}) = 0 = (\beta-\hat{\beta})'X'V^{-1}(Y-X\hat{\beta})$ d'où on obtient la décomposition. Le premier terme du membre droit de l'égalité $(Y-X\beta)'V^{-1}(Y-X\beta) = (Y-X\hat{\beta})'V^{-1}(Y-X\hat{\beta})+(\beta-\hat{\beta})'X'V^{-1}X(\beta-\hat{\beta})$ ne dépend pas de la variable d'intégration β dans le calcul de la vraisemblance marginale. On peut donc le factoriser comme une constante et on obtient l'égalité suivante pour la vraisemblance marginale :

$$P(y|\theta) = (2\pi)^{-\frac{n}{2}}|V|^{-\frac{1}{2}}exp(-\frac{1}{2}(Y-X\hat{\beta})'V^{-1}(Y-X\hat{\beta}))\int_{\beta\in\mathbb{R}^k} exp(-\frac{1}{2}(\beta-\hat{\beta})'X'V^{-1}X(\beta-\hat{\beta}))\mathbf{d}\beta$$

Or $exp(-\frac{1}{2}(\beta-\hat{\beta})'X'V^{-1}X(\beta-\hat{\beta}))$ est la densité de $\beta|y,\theta \sim \mathcal{N}_k(\hat{\beta},(X'V^{-1}X)^{-1})$ et en divisant par le facteur de normalisation $(2\pi)^{\frac{k}{2}}|X'V^{-1}X|^{-\frac{1}{2}}$ on a que :

$$(2\pi)^{-\frac{k}{2}}|X'V^{-1}X|^{\frac{1}{2}}\int_{\beta\in\mathbb{R}^k} exp(-\frac{1}{2}(\beta-\hat{\beta})'X'V^{-1}X(\beta-\hat{\beta}))\mathbf{d}\beta = 1$$

$$\Longrightarrow \int_{\beta\in\mathbb{R}^k} exp(-\frac{1}{2}(\beta-\hat{\beta})'X'V^{-1}X(\beta-\hat{\beta}))\mathbf{d}\beta = (2\pi)^{\frac{k}{2}}|X'V^{-1}X|^{-\frac{1}{2}}$$

Il vient donc finalement que l'expression de la vraisemblance marginale est donnée par :

$$P(y|\theta) = L_{REML}(\theta;Y) = (2\pi)^{-\frac{(n-k)}{2}}|V|^{-\frac{1}{2}}|X'V^{-1}X|^{-\frac{1}{2}}exp(-\frac{1}{2}(Y-X\hat{\beta})'V^{-1}(Y-X\hat{\beta}))$$

, ou encore sous sa forme la plus connue avec le logarithme :

$$\begin{aligned} ln(L_{REML}(\theta;Y)) = -\frac{1}{2}\Big((n-k)ln(2\pi) + ln(|V|) + ln(|X'V^{-1}X|) \\ + (Y-X\hat{\beta})'V^{-1}(Y-X\hat{\beta})\Big) \end{aligned} \tag{3.4}$$

Maximiser l'expression (3.2) par rapport à θ revient à maximiser l'expression (3.4), car ces deux expressions sont les mêmes à une constante près qui est $ln|X'X|$ et cette constante ne dépend pas de θ.

On pourra retenir de cette sous-section que l'on se place dans le cadre de la vraisemblance restreinte, associée au modèle (2.1), lorsque l'on choisit une loi *a priori* uniforme et impropre $\pi(\beta|\theta)$ pour la loi de $\beta|\theta$. Cette propriété est très utile, comme on le verra dans la sous-section suivante, car elle simplifie beaucoup l'implémentation de l'EM dans le cadre du modèle mixte.

3.1.3 Estimation des composantes du modèle mixte par EM ("Expectation-Maximization")

Il existe plusieurs algorithmes d'optimisation possibles, tels que Newton-Raphson (NR), Fisher Scoring (FS ; variante de Newton-Raphson), ou des méthodes de Monte-Carlo par chaîne de Markov (MCMC), afin d'estimer les composantes associées au modèle mixte pour la détection de QTL. L'algorithme EM (Dempster *et al.*, 1977) possède

cependant plusieurs avantages, par rapport aux autres algorithmes, qui lui confère une célébrité reconnue (Couvreur, 1996). En effet l'EM, dans le cadre général, garantit des estimations qui font croître la fonction de vraisemblance à chaque itération. Cette garantie peut aussi être obtenue pour d'autres algorithmes, tels que NR ou FS, conditionnellement à une recherche linéaire, c'est à dire de chercher une direction de descente pour la fonction objective à minimiser. Cependant cette recherche linéaire peut être plus ou moins longue et numériquement instable. Les méthodes MCMC dans un cadre bayésien nécessitent quant à elles la spécification, plus ou moins arbitraire, de lois a priori utilisées pour le calcul de lois a posteriori associées aux composantes estimées par le schéma de Gibbs (cas particulier de l'algorithme de Metropolis-Hasting). De plus il existe peu de résultats théoriques sur le nombre d'itérations requis pour la convergence de l'algorithme de Gibbs, et ce nombre d'itérations s'avère en pratique assez élevé (Rosenthal, 1995).

L'estimation par EM, pouvant aussi être longue, garantit cependant que les estimations appartiennent à l'espace paramétrique et que l'algorithme converge au moins à un maximum local pour la vraisemblance (Foulley, 2002). De plus les expressions de l'EM pour les estimations à chaque itération sont souvent simples et facile à implémenter. Notons finalement que cet algorithme permet d'obtenir, à chaque itération, des estimations pour des variables non-observées (manquantes ou latentes) nécessaires à la maximisation de la vraisemblance et à l'estimation des paramètres à l'itération suivante. On donne dans la sous-section suivante une introduction simple de cet algorithme afin d'expliciter les expressions générales utilisées dans le cadre du modèle mixte. Ces expressions seront utilisées dans le cadre du chapitre 5 sur la discrimination des modèles d'association utilisant des haplotypes.

3.1.3.1 L'algorithme EM

On donne ici une description simple de l'algorithme, inspirée de Borman (2004), dans le cas discret qui se généralise sans trop de complexité au cas continu. On rappelle d'abord quelques éléments d'analyse convexe réelle nécessaires à la dérivation de l'algorithme.

Définition (Fonction convexe)**:** Une application $f : [a, b] \to \mathbb{R}$ est dite *convexe* sur $[a, b]$ si $\forall x_1, x_2 \in [a, b]$ et $\forall \lambda \in [0, 1]$ on a,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2)$$

Théorème (Inégalité de Jensen)**:** Soit f une fonction convexe définie sur un intervalle fermé I. *Si* $x_1, ..., x_n \in I$ et $\lambda_1, ..., \lambda_n \geq 0$ tels que $\sum_{i=1}^{n} \lambda_i = 1$ alors :

$$f(\sum_{i=1}^{n} \lambda_i x_i) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$$

Dérivation de l'algorithme :

Soit Z un vecteur aléatoire de données manquantes, ayant pour réalisation z, dont la connaissance rendrait possible la maximisation de la vraisemblance des données dite "complètes" $P(Y, Z|\theta)$. On remarque que la vraisemblance des données peut se ré-écrire de la façon suivante :

$$P(Y|\theta) = \sum_z P(Y, z|\theta) = \sum_z P(Y|z, \theta)P(z|\theta)$$

Soit $L(\theta^{[t]})$ la valeur de la log-vraisemblance des données à l'itération t. On a :

$$\begin{aligned}
L(\theta) - L(\theta^{[t]}) &= ln\Big(\sum_z P(Y|z,\theta)P(z|\theta)\Big) - ln\Big(P(Y|\theta^{[t]})\Big) \\
&= ln\Big(\sum_z P(Y|z,\theta)P(z|\theta).\frac{P(z|Y,\theta^{[t]})}{P(z|Y,\theta^{[t]})}\Big) - ln\Big(P(Y|\theta^{[t]})\Big) \\
&= ln\Big(\sum_z P(z|Y,\theta^{[t]}).\frac{P(Y|z,\theta)P(z|\theta)}{P(z|Y,\theta^{[t]})}\Big) - ln\Big(P(Y|\theta^{[t]})\Big) \\
&\geq \sum_z P(z|Y,\theta^{[t]}).ln\Big(\frac{P(Y|z,\theta)P(z|\theta)}{P(z|Y,\theta^{[t]})}\Big) - ln\Big(P(Y|\theta^{[t]})\Big)
\end{aligned}$$

Le passage à l'inégalité est obtenu par l'inégalité de Jensen en remarquant que l'on a $\sum_z P(z|Y, \theta^{[t]}) = 1$. De plus comme on a $ln\Big(P(Y|\theta^{[t]})\Big) = \sum_z P(z|Y, \theta^{[t]}).ln\Big(P(Y|\theta^{[t]})\Big)$ il vient que :

$$\begin{aligned}
L(\theta) - L(\theta^{[t]}) &\geq \sum_z P(z|Y,\theta^{[t]}).ln\Big(\frac{P(Y|z,\theta)P(z|\theta)}{P(z|Y,\theta^{[t]})P(Y|\theta^{[t]})}\Big) \\
L(\theta) &\geq L(\theta^{[t]}) + \sum_z P(z|Y,\theta^{[t]}).ln\Big(\frac{P(Y|z,\theta)P(z|\theta)}{P(z|Y,\theta^{[t]})P(Y|\theta^{[t]})}\Big) = \varphi(\theta|\theta^{[t]})
\end{aligned}$$

On a donc que la fonction $\varphi(\theta|\theta^{[t]})$ est bornée supérieurement par $L(\theta)$. Cependant, cette condition n'est pas totalement suffisante en soi pour faire croître $L(\theta)$ par la maximisation de $\varphi(\theta|\theta^{[t]})$. La condition complémentaire est donnée par le fait que $L(\theta)$ soit égale à $\varphi(\theta|\theta^{[t]})$ à l'itération t. En effet on a :

$$\begin{aligned}\varphi(\theta^{[t]}|\theta^{[t]}) &= L(\theta^{[t]}) + \sum_z P(z|Y,\theta^{[t]}).ln\Big(\frac{P(Y|z,\theta^{[t]})P(z|\theta^{[t]})}{P(z|Y,\theta^{[t]})P(Y|\theta^{[t]})}\Big) \\ &= L(\theta^{[t]}) + \sum_z P(z|Y,\theta^{[t]}).ln\Big(\frac{P(Y,z|\theta^{[t]})}{P(Y,z|\theta^{[t]})}\Big) \\ &= L(\theta^{[t]}) + \sum_z P(z|Y,\theta^{[t]}).ln1 = L(\theta^{[t]})\end{aligned}$$

On a donc $L(\theta^*) \geq \varphi(\theta^*|\theta^{[t]}) \geq \varphi(\theta^{[t]}|\theta^{[t]}) = L(\theta^{[t]})$, où $\theta^* = \underset{\theta}{argmax}\{\varphi(\theta|\theta^{[t]})\}$. Ainsi il suffit de choisir θ à l'itération suivante de manière à faire croître $\varphi(\theta|\theta^{[t]})$ afin de faire croître $L(\theta)$. L'itération $t+1$ de l'algorithme est donc donné par :

$$\begin{aligned}\theta^{[t+1]} &= \underset{\theta}{argmax}\{\varphi(\theta|\theta^{[t]})\} \\ &= \underset{\theta}{argmax}\Big\{\sum_z P(z|Y,\theta^{[t]}).ln\Big(P(Y|z,\theta)P(z|\theta)\Big)\Big\} \\ &= \underset{\theta}{argmax}\Big\{\sum_z P(z|Y,\theta^{[t]}).ln\Big(P(Y,z|\theta)\Big)\Big\} \\ &= \underset{\theta}{argmax}\Big\{\mathbb{E}_{Z|Y,\theta^{[t]}}\Big[ln(P(Y,Z|\theta))\Big]\Big\}\end{aligned}$$

Finalement on voit apparaître les deux étapes clés de l'algorithme qui sont :

1. Etape-E : le calcul de l'espérance conditionnelle $\mathbb{E}_{Z|Y,\theta^{[t]}}\Big[ln(P(Y,Z|\theta))\Big]$
2. Etape-M : la maximisation de cette espérance par rapport à θ

3.1.3.2 Estimation des composantes par EM(-REML)

On se place dans le cadre simple du modèle mixte à un facteur aléatoire afin d'expliciter les expressions de l'algorithme, bien que celles-ci se généralisent sans trop de difficulté au cas de plusieurs facteurs aléatoires. Soit le modèle mixte suivant :

$$Y = X\beta + Zu + \varepsilon \;\;;\;\; u \sim \mathcal{N}_q(0, K\sigma_u^2),\;\; \varepsilon \sim \mathcal{N}_n(0, I_n\sigma_\varepsilon^2)$$

où u est de dimension q. On pose $V = var(Y) = ZK\sigma_u^2Z' + I_n\sigma_\varepsilon^2$. Soit $\tilde{Z} = (\beta, u)$ le vecteur des données manquantes et $\theta = (\sigma_u^2, \sigma_\varepsilon^2)$ le vecteur des paramètres à estimer. On a :

$$P(Y,\tilde{Z}|\theta) = P(Y,\beta,u|\theta) = P(Y|\beta,u,\theta).P(u|\beta,\theta).P(\beta|\theta)$$

On a vu précédemment (cf. 3.1.2) que l'on se plaçait dans le cadre du REML pour la

fonction à maximiser lorsque l'on choisit une loi *a priori* uniforme et impropre $\pi(\beta|\theta)$ pour la loi de $\beta|\theta$. Dans ce cas on a :

$$P(Y, \tilde{Z}|\theta) \propto P(Y|\beta, u, \theta).P(u|\beta, \theta)$$

où $Y|\beta, u, \theta \sim \mathcal{N}_n(X\beta + Zu, I_n\sigma_\varepsilon^2)$ et $u|\beta, \theta \sim \mathcal{N}_q(0, K\sigma_u^2)$. Il vient également que :

$$lnP(Y, \tilde{Z}|\theta) = lnP(Y|\beta, u, \theta) + lnP(u|\beta, \theta) + cste$$

$$\Rightarrow \mathbb{E}_{\tilde{Z}|Y,\theta^{[t]}}\Big[lnP(Y, \tilde{Z}|\theta)\Big] = \mathbb{E}_{\tilde{Z}|Y,\theta^{[t]}}\Big[lnP(Y|\beta, u, \theta)\Big] + \mathbb{E}_{\tilde{Z}|Y,\theta^{[t]}}\Big[lnP(u|\beta, \theta)\Big] + cste$$

où

$$\begin{cases} \mathbb{E}_{\tilde{Z}|Y,\theta^{[t]}}\Big[lnP(Y|\beta, u, \theta)\Big] = -\dfrac{1}{2}\left[nln2\pi + nln\sigma_\varepsilon^2 + \dfrac{\mathbb{E}_{\tilde{Z}|Y,\theta^{[t]}}\Big[\varepsilon'\varepsilon\Big]}{\sigma_\varepsilon^2}\right] & (1) \\ \mathbb{E}_{\tilde{Z}|Y,\theta^{[t]}}\Big[lnP(u|\beta, \theta)\Big] = -\dfrac{1}{2}\left[qln2\pi + qln\sigma_u^2 + \dfrac{\mathbb{E}_{\tilde{Z}|Y,\theta^{[t]}}\Big[u'K^{-1}u\Big]}{\sigma_u^2}\right] & (2) \end{cases}$$

En dérivant (1) et (2) par rapport à σ_ε^2 et σ_u^2 respectivement, et en égalant à zéro, on a :

$$\begin{cases} {\sigma_\varepsilon^2}^{[t+1]} = \dfrac{\mathbb{E}_{\tilde{Z}|Y,\theta^{[t]}}\Big[\varepsilon'\varepsilon\Big]}{n} = \dfrac{\mathbb{E}(\varepsilon|Y, \theta^{[t]})'\mathbb{E}(\varepsilon|Y, \theta^{[t]}) + tr(var(\varepsilon|Y, \theta^{[t]}))}{n} & (3) \\ {\sigma_u^2}^{[t+1]} = \dfrac{\mathbb{E}_{\tilde{Z}|Y,\theta^{[t]}}\Big[u'K^{-1}u\Big]}{q} = \dfrac{\mathbb{E}(u|Y, \theta^{[t]})'K^{-1}\mathbb{E}(u|Y, \theta^{[t]}) + tr(K^{-1}var(u|Y, \theta^{[t]}))}{q} & (4) \end{cases}$$

En identifiant (3) et (4) aux composantes estimées par résolution du système d'équations du modèle mixte (Henderson, 1973, 1984), et après quelques manipulations matricielles, on montre que :

$$\begin{cases} {\sigma_\varepsilon^2}^{[t+1]} = \dfrac{Y'Y - [\hat{\beta}^{[t]'}\hat{u}^{[t]'}].[X'Z']'.Y - \lambda^{[t]}\hat{u}^{[t]'}K^{-1}\hat{u}^{[t]} + \Big[rang(X) + q - \lambda^{[t]}tr(K^{-1}C_{uu}^{[t]})\Big].{\sigma_\varepsilon^2}^{[t]}}{n} \\ {\sigma_u^2}^{[t+1]} = \dfrac{\hat{u}^{[t]'}K^{-1}\hat{u}^{[t]} + tr(K^{-1}C_{uu}^{[t]}){\sigma_\varepsilon^2}^{[t]}}{q} \end{cases}$$

où $\lambda^{[t]} = {\sigma_\varepsilon^2}^{[t]}/{\sigma_u^2}^{[t]}$, $\hat{\beta}^{[t]}$ est l'estimateur des moindres carrés généralisés, $\hat{u}^{[t]} =$

$\mathbb{E}(u|Y,\theta^{[t]}) = \sigma_u^{2\ [t]} KZ'(V^{[t]})^{-1} \times(Y - X\hat{\beta}^{[t]})$ est le "Best Linear Unbiased Predictor" (BLUP) et $C_{uu}^{[t]}\sigma_\varepsilon^{2\ [t]} = var(u|Y,\theta^{[t]})$, avec $C_{uu}^{[t]}$ étant le bloc relatif aux effets aléatoires dans l'inverse de la matrice des coefficients des équations d'Henderson.

Ces formules se généralisent aux expressions suivantes dans le cas de L facteurs aléatoires ($k \in \{1,..,L\}$) :

$$\begin{cases} \sigma_\varepsilon^{2\ [t+1]} = \Big[Y'Y - \tilde{Z}^{[t]'}.W'.Y - \sum_{k=1}^{L} \lambda_k^{[t]} \hat{u}_k^{[t]'} K_k^{-1} \hat{u}_k^{[t]} \\ \qquad\qquad + \Big[rang(X) + \sum_{k=1}^{L} q_k - \sum_{k=1}^{L} \lambda_k^{[t]} tr(K_k^{-1} C_{kk}^{[t]}) \Big] . \sigma_\varepsilon^{2\ [t]} \Big] \Big/ n \\ \\ \sigma_{u_k}^{2\ [t+1]} = \dfrac{\hat{u}_k^{[t]'} K_k^{-1} \hat{u}_k^{[t]} + tr(K_k^{-1} C_{kk}^{[t]}) \sigma_\varepsilon^{2\ [t]}}{q_k} \end{cases}$$

où $\lambda_k^{[t]} = \sigma_\varepsilon^{2\ [t]} / \sigma_{u_k}^{2\ [t]}$, $\tilde{Z}^{[t]} = [\hat{\beta}^{[t]} \hat{u_1}^{[t]} ... \hat{u_L}^{[t]}]$ et $W = [XZ_1...Z_L]$.

Cet EM a été implémenté (en Fortran 90) dans le cadre de ce travail de thèse, pour les fonctions de vraisemblance restreinte sous l'hypothèse nulle et alternative, afin de comparer les différents modèles d'association utilisant des haplotypes décrits dans le chapitre 5.

3.2 Statistiques de test

3.2.1 Tests généraux

Il existe plusieurs statistiques de test possibles pour la détection de QTL dans le cadre du modèle mixte. Parmi celles-ci il y a le test du rapport de vraisemblances (LRT : "Likelihood ratio test"), le test de Wald et le test du Score. Ces trois tests sont tous équivalents pour le développement de Taylor à l'ordre 1 lorsque le nombre d'observations croît à l'infini (Rao, 1973 ; Gourieroux et Monfort, 1989 ; Yi et Wang, 2011). Cependant ils peuvent néanmoins différer dans leurs comportements lorsque l'on a un nombre fini d'observations. Le lemme fondamental de Neyman-Pearson affirme également que le test du rapport de vraisemblances (LRT) est le plus puissant de ces tests si les hypothèses

testées sont dites "simples", c'est à dire si elles sont de la forme suivante :

$$H_0 : \theta = \theta_0 \text{ et } H_1 : \theta = \theta_1 \text{ où } \theta_1 \neq \theta_0$$

On rappelle que la puissance, $1-\beta$, d'un test est définie comme étant la probabilité de rejeter l'hypothèse nulle, pour un seuil s de rejet fixé, sachant que l'hypothèse alternative est vraie, i.e. $1-\beta = P_s(\{\text{rejeter } H_0 | H_1 \text{ est vraie}\})$ où le complément β est le risque d'erreur de deuxième espèce (faux négatif). On rappelle également que la probabilité d'erreur de première espèce (faux positif), α, d'un test est définie comme étant la probabilité de rejeter à tort H_0, i.e. $\alpha = P_s(\{\text{rejeter } H_0 | H_0 \text{ est vraie}\})$. Les hypothèses testées pour la détection de QTL dans le cadre du modèle (2.1), lorsqu'il existe un effet polygénique et un effet QTL, sont les suivantes :

$$\begin{cases} H_1 : V = Z_1 \sigma^2_{u_1} A Z_1{}' + Z_2 \sigma^2_{u_2} H Z_2' + \sigma^2_{u_0} I_n & \iff H_1 : \sigma^2_{u_2} > 0 \\ H_0 : V = Z_1 \sigma^2_{u_1} A Z_1{}' + \sigma^2_{u_0} I_n & \iff H_0 : \sigma^2_{u_2} = 0 \end{cases}$$

où l'on rappelle que A est la matrice de variance-covariance des effets polygéniques et H est la matrice de variance-covariance des effets des allèles (ou des haplotypes) au QTL. H_1 est donc l'hypothèse de la présence d'effet d'un QTL que l'on teste contre H_0 qui est l'hypothèse de l'absence d'effet de celui-ci. On voit que ces hypothèses ne sont pas simples mais on verra, dans la sous-section suivante, comment justifier l'utilisation du LRT pour tester ces hypothèses.

3.2.2 Le test du rapport de vraisemblances (LRT)

Le test du rapport de vraisemblances est un test dit uniformément de plus grande puissance ($U.P.P$) si les hypothèses testées sont simples. Cela signifie que lorsque la probabilité d'erreur de première espèce, α, est fixée le LRT est le test qui minimise la probabilité de l'erreur de deuxième espèce β, i.e. $\beta = P_s(\{\text{non rejet } H_0 | H_1 \text{ est vraie}\})$. Néanmoins on ne peut pas effectuer ce test si on ne dispose pas d'une estimation de $\sigma^2_{u_2}$, sachant que $\sigma^2_{u_2} \in]0; +\infty[$ sous H_1 alors que $\sigma^2_{u_2} = 0$ sous H_0. On cherchera donc toujours à avoir des estimations ponctuelles et préférentiellement non biaisées, par le REML, des composantes $(\sigma^2_i)_{0 \leq i \leq 2}$ qui maximisent les vraisemblances (restreintes) et les hypothèses testées seront en conséquence simples.

Le rapport des vraisemblances restreintes maximales (RLRT), à une position testée et nommée i, que l'on établit pour la détection de QTL est alors donné par :

$$\lambda_i = -2ln\left(\frac{sup_{\sigma^2_{u_0},\sigma^2_{u_1}} L_{REML}(\sigma^2_{u_0},\sigma^2_{u_1};Y)}{sup_{\sigma^2_{u_0},\sigma^2_{u_1},\sigma^2_{u_2}} L_{REML}(\sigma^2_{u_0},\sigma^2_{u_1},\sigma^2_{u_2};Y)}\right) = -2ln\left(\frac{sup_{\sigma^2_{u_0},\sigma^2_{u_1}} L_{REML}(H0)}{sup_{\sigma^2_{u_0},\sigma^2_{u_1},\sigma^2_{u_2}} L_{REML}(H1)}\right)$$

La distribution de cette statistique de test n'est pas connue mais Self et Liang (1987) montrent qu'elle est proche de $\frac{1}{2}\chi^2(1) + \frac{1}{2}\chi^2(2)$. Finalement, pour un ensemble $\mathcal{I} = \{i_1, .., i_r\}$ de positions testées sur le génome, la position d'un QTL est estimée par :

$$\hat{\lambda}_{RLRT_{max}} = \underset{i\in\mathcal{I}}{argmax}\{\hat{\lambda}_i\}$$

On retiendra donc de cette sous-section que le test du rapport de vraisemblances est un test uniformément plus puissant pour la cartographie de QTL. Ce test a aussi l'avantage de procurer un cadre naturel d'application de l'algorithme EM, qui augmente naturellement les vraisemblances de ce rapport en estimant les paramètres associés (cf. 3.1.3.1).

Partie II :
Discrimination entre modèles d'association utilisant des haplotypes

Chapitre 4

Modèles d'association utilisant des haplotypes

4.1 Contexte et importance de l'étude

Un prédicteur d'identité allélique ou AIP pour "Allelic Identity Predictor" repose sur l'idée générale que, quand la densité des SNP est assez grande un déséquilibre de liaison entre marqueurs peut être observé, et que deux segments chromosomiques ont d'autant plus de chance de porter le même allèle en un locus caché (non observable) qu'ils se ressemblent aux SNP voisins. Ainsi un AIP associe une valeur positive à un couple de segments chromosomiques, que l'on peut normaliser à l'intervalle [0,1], si ces derniers portent des allèles IBS aux SNP constituant les haplotypes dont ils sont porteurs. On note dès à présent que chaque position testée dans un génome scan est choisie par convention comme le centre des segments chromosomiques considérés.

Chaque AIP produit une matrice H de similitude entre haplotypes qui lui est propre à chaque position testée, que l'on interprète dans la littérature comme une matrice de variance-covariance des effets des allèles au QTL (Meuwissen et Goddard, 2000). Ainsi chaque AIP définit un modèle d'association, dans le cadre du modèle mixte, qui lui est propre. Aucun consensus dans la littérature ne semble se dégager autour d'un AIP particulier. Parmi les raisons expliquant cette absence de consensus figure certainement le manque de connaissances sur le lien entre le déséquilibre de liaison et la prédiction d'identité allélique qu'effectuent les AIP. L'un des objectifs du chapitre 5 est de caractériser ce lien à l'aide d'une distance matricielle, définie entre une position testée et un QTL sur le génome, et de proposer des règles à partir de celle-ci pour la discrimination des AIP en cartographie.

4.2 Les prédicteurs d'identité allélique (AIP)

Les AIP proposés dans la littérature sont basés sur des concepts très différents. Parmi ces prédicteurs, on trouve ceux qui sont basés sur l'observation simple de la ressemblance entre haplotypes (e.g. IBS_{hap} ; *Score* de Li et Jiang, 2005) et d'autres basés sur des principes d'évolution des populations tels que le processus de coalescence (e.g. $P(IBD)$ de Meuwissen et Goddard, 2001), les arbres phylogénétique (e.g. *Blossoc* de Mailund *et al.*, 2006), l'analyse cladistique (e.g. Durrant *et al.*, 2004) et les graphes de recombinaisons ancestrales (e.g. Minichiello et Durbin, 2006). Il existe également des approches basées sur des théories Markoviennes telles que *Beagle* de Browning et Browning (2006), *Phase* de Stephens *et al.* (2001) et $FastPhase$ de Scheet et Stephens (2006). Cependant, par souci de simplification, on ne détaillera ici qu'un nombre limité de ces AIP que l'on comparera dans le chapitre 5.

4.2.1 IBS_{hap}

IBS_{hap} est une fonction booléenne qui associe 1, à un couple de segments chromosomiques, si les allèles des haplotypes dont ils sont porteurs sont IBS à tous les SNP et 0 sinon. Par exemple pour deux haplotypes h_i et h_j tel que $h_i = (122121)$ et $h_j = (122122)$, portés par un ensemble de segments chromosomiques distincts, on a $IBS_{hap}(h_i,h_i)$=1 et $IBS_{hap}(h_i,h_j)$=0. C'est le plus simple des prédicteurs d'identité allélique.

4.2.2 *Score* : le score de similarité de Li et Jiang (2005)

Le score de Li et Jiang (2005) est une fonction somme qui se décompose en deux sous-fonctions sommes. La première sous-fonction, identifiable à une mesure de similarité de Hamming, compte le nombre d'allèles IBS entre les haplotypes portés par un couple de segments chromosomiques et centré sur la position dont on teste si elle porte un QTL. La deuxième sous-fonction calcule la longueur du segment continu, et passant par le centre, partagé en commun par les haplotypes. Le score entre deux segments chromosomiques représentés par leurs haplotypes respectifs h_i et h_j est donné par :

$$s_{i,j} = \sum_{k=-l}^{r} w_1(x_k)\mathbb{1}(h_i(k), h_j(k)) + \sum_{k=-l'}^{r'} w_2(x_k)$$

où x_k est la distance réelle des allèles $h_i(k)$ et $h_j(k)$, au marqueur k ou à la position k ($-l \leq k \leq r$, $-l' \geq -l$ et $r' \leq r$), par rapport au centre des haplotypes qui se situe à la distance x_t. w_1 et w_2 sont des fonctions poids qui ont pour objet de décroître l'importance que l'on attribue aux SNP quand ceux-ci s'éloignent du locus testé x_t.

Ces poids peuvent être des fonctions linéaires, quadratiques ou exponentielles décroissantes. Les auteurs n'imposent pas le choix de ces fonctions à l'utilisateur et ont illustré l'application de leur méthode avec $w_1(x_k) = w_2(x_k) = 1 - \sum_{i=t}^{k-1} |x_i - x_{i+1}|$, où $|x_i - x_{i+1}|$ est la longueur réelle en centiMorgan entre les SNP situés aux distances x_i et x_{i+1}. La deuxième sous-fonction est justifiée par l'hypothèse que deux segments chromosomiques, partageant un segment continu, ont plus de chance d'être hérités d'un ancêtre en commun. La première sous-fonction quant à elle évite que l'on sous-estime le degré de similitude qu'engendrerait une mutation ou une erreur de génotypage par rapport à l'un des marqueurs sur les haplotypes.

Par exemple si $\forall i \in \{-l, .., r\}$ $|x_i - x_{i+1}| = 0,1$ (distance constante entre les marqueurs) alors pour deux segments chromosomiques de 6 marqueurs, $h_i = (\mathbf{11212}2)$ et $h_j = (1\mathbf{2212}1)$ centrés à mi-chemin entre l'allèle **2** et **1**, on a :

$$\begin{aligned}
\sum_{k=-3}^{3} w_1(x_k)\mathbb{1}(h_i(k), h_j(k)) &= 2\left[1 - \frac{0,1}{2}\right] + \left[1 - (0,1 + \frac{0,1}{2})\right] + \left[1 - (0,1 + 0,1 + \frac{0,1}{2})\right] \\
&= 3,5 \\
\text{et } \sum_{k=-3}^{3} w_2(x_k) &= 2\left[1 - \frac{0,1}{2}\right] + \left[1 - (0,1 + \frac{0,1}{2})\right] = 2,75 \\
&\Rightarrow s_{i,j} = 3,5 + 2,75 = 6.25
\end{aligned}$$

De la même façon on montre que le score maximum entre deux haplotypes de 6 marqueurs est donné par $s_{max} = s_{i,i} = s_{j,j} = 5,1 + 5,1 = 10,2$. On a donc que le score normalisé entre h_i et h_j vaut :

$$s_{i,j}^{norm} = \frac{s_{i,j}}{s_{max}} = 0.61$$

L'exemple suivant illustre le concept de ce score de similarité (sans considération des

fonctions poids cette fois-ci) :

Si l'on considère les quatre haplotypes suivants $h_1 = (\mathbf{11}2\mathbf{1}2\mathbf{1})$, $h_2 = (\mathbf{1}22\mathbf{2}2\mathbf{2})$, $h_3 = (1\mathbf{122}12)$ et $h_4 = (2\mathbf{122}21)$, et que l'on considère seulement la longueur du segment continu comme mesure de similitude, alors $s_{3,4} = 3$ tandis que $s_{1,2} = 0$ sachant que h_1 et h_2 partagent 3 allèles en commun. Donc si on a une erreur de génotypage, ou une mutation, au 4-ième marqueur de h_1 et h_2 alors la deuxième sous-fonction, à elle seule, sous-estimerait réellement le niveau de similitude entre ces haplotypes. Les deux mesures combinées définissent ainsi un score supposé plus robuste que chacune des mesures prise séparément.

4.2.3 $P(IBD)$: la probabilité d'IBD de Meuwissen et Goddard (2001)

La méthode de Meuwissen et Goddard (2001) intègre les hypothèses d'un modèle de coalescence sous-jacent, analogue à celui de Wright-Fisher (Fisher 1922, 1930 ; Wright, 1931 ; Charlesworth, 2009) pour calculer la probabilité qu'un allèle à un locus testé A soit IBD, dans une paire de chromosomes homologues, conditionnellement à l'état IBS ($S = 1$ ou 0) des allèles aux marqueurs adjacents. Par exemple la probabilité que deux segments chromosomiques, ayant pour haplotypes $h_i = (122122)$ et $h_j = (122121)$, aient des allèles IBD au centre (le locus A) est donnée par :

$$P(A = IBD|S) = \frac{P(A = IBD\&S)}{P(S\&A = IBD) + P(S\&A = nonIBD)}$$

où S décrit les statuts IBS des allèles aux marqueurs entre les deux haplotypes. Afin d'effectuer ce calcul les auteurs définissent un vecteur ϕ qui contient les différents états d'IBD (0 ou 1) des allèles des deux segments chromosomiques et aussi l'information de recombinaison historique, éventuel, entre chacun de ces allèles.

Par exemple pour des haplotypes de 3 marqueurs, donc de 3 allèles, le vecteur : $\phi = [\phi(-2)\ \phi(-1)\ \phi(0)\ \phi(1)\ \phi(2)] = [1\_1 \times 0]$ décrit la situation où les allèles aux positions -2 et 0 sont IBD ($\phi(-2) = 1$ et $\phi(0) = 1$) sans aucun évènement de recombinaison au milieu ($\phi(-1) =$ "_") et l'allèle à droite du locus 0 est nonIBD ($\phi(2) = 0$) précédé d'une recombinaison ($\phi(1) =$ " $\times$ "). Cette description se généralise à tout haplotype de

plusieurs marqueurs. Si on suppose que le locus A est à la position 0, qui est le centre des haplotypes, alors on a $P(S\&A = IBD) = \sum_{\phi|\phi(0)=1} P(S|\phi)P(\phi)$ et $P(S\&A = nonIBD) = \sum_{\phi|\phi(0)=0} P(S|\phi)P(\phi)$. Pour $\phi = [\phi(0)\ \phi(1)] = [1\_\ 1]$, $P(\phi)$ est calculée selon les hypothèses d'un modèle de coalescence. C'est à dire que la probabilité qu'il n'y ait pas d'ancêtre commun à deux gamètes pendant $t-1$ générations est pour des organismes diploïdes de $\left(1-\frac{1}{2Ne}\right)^{t-1}$, où Ne est la taille efficace de la population et $t \in \{1,..,T\}$ tel que $T =$ génération actuelle.

De plus, les auteurs imposent que la probabilité qu'il n'y ait pas de recombinaison entre deux loci est modélisée par une loi de Poisson de $k = 1$ occurrence dans un intervalle de longueur c Morgan (i.e. $exp(-c)$). Donc la probabilité que soient IBD deux segments chromosomiques issus d'un ancêtre en commun t générations dans le passé et n'ayant pas recombinés vaut $\frac{1}{2Ne}\left(1-\frac{1}{2Ne}\right)^{t-1} exp(-c)^{2t}$ (*). Les auteurs approximent ce résultat par $\frac{1}{2Ne} exp\left(-\frac{t-1}{2Ne} - 2ct\right)$ (**) en utilisant la limite suivante ; $\lim_{n\to+\infty}\left(1+\frac{a}{n}\right) = exp(a)$. En effet $\left(1-\frac{1}{2Ne}\right)^{t-1} = \left(\left(1+\frac{-1}{2Ne}\right)^{2Ne}\right)^{\frac{t-1}{2Ne}} \approx exp(-1)^{\frac{t-1}{2Ne}}$, pour Ne suffisamment grand, et en reportant cette approximation dans l'expression (*) on obtient l'expression (**). Finalement, comme l'évènement de coalescence peut arriver à un temps t quelconque dans le passé jusqu'à la génération actuelle, la probabilité que deux segment chromosomiques de longueur c soient IBD est donnée par :

$$f(c) = P(\phi = [1\_\ 1]) = \frac{1}{2Ne} exp(-2c) \sum_{t=1}^{T} exp\left[-(t-1)\left(\frac{1}{2Ne} + 2c\right)\right]$$

De cette dernière expression, les auteurs dérivent des expressions pour le calcul de la probabilité de ϕ selon qu'il ait une recombinaison entre le locus testé à la position 0 et le deuxième à la position 1. Ces calculs sont ensuite généralisés à tout haplotype d'un nombre quelconque de marqueurs.

4.2.4 *Beagle* : le modèle de regroupement local d'haplotypes de Browning et Browning (2006)

Le modèle de Browning et Browning (2006), implémenté selon une heuristique dans le logiciel Beagle, est un modèle de regroupement local d'haplotypes, ou de segments chro-

mosomiques plus précisément, basé sur des VLMCs ("Variable Length Markov Chains"). On rappelle qu'une chaîne de Markov (d'ordre 1) est une suite de variables aléatoires, ayant la propriété de Markov, à valeurs dans un espace d'états. La propriété de Markov suppose que la probabilité de l'état futur d'une variable aléatoire dépend uniquement de son état présent et non de ses états passés. La propriété de Markov n'est généralement pas vérifiée pour les SNP à cause du déséquilibre de liaison.

Cependant les chaînes de Markov d'ordre supérieur (d'ordre m ; $m > 1$) peuvent, éventuellement, tenir compte du déséquilibre de liaison. L'utilisation de ces chaînes dans le cadre des HMMs, pour "Hidden Markov Models", nécessite une spécification de ce type de modèle en fonction du lien entre la variable représentant l'émission (la valeur d'un SNP par exemple) et celle représentant l'état caché (l'état d'IBD par exemple). Par exemple, un SNP (i.e. la variable d'émission), dans une séquence sur un chromosome, peut être modélisé conditionnellement à l'émission de plusieurs autres SNP sur le chromosome et une variable représentant l'état IBD qui lui est sous-jacent. Le modèle devient complexe lorsque le nombre de SNP intervenant dans ce conditionnement augmente.

L'objectif du modèle de Browning et Browning, par l'utilisation des VLMCs, est d'apporter une modélisation plus flexible que les HMMs en ne faisant aucune hypothèse quant à la nature de ce lien. Ce modèle serait, selon les auteurs, suffisamment flexible afin d'approcher un HMM. Les auteurs affirment également que ce modèle permettrait de bien tenir compte du déséquilibre de liaison dans les analyses d'association. Un modèle basé sur des VLMCs peut être représenté par un graphe acyclique orienté ou DAG pour "Directed acyclic graph". Par exemple pour la liste d'haplotypes, et de leurs comptes associés, représentés dans le tableau 4.1 on a le DAG suivant (avant clusterisation) :

Haplotypes	Total
1111	21
1112	79
1122	95
1221	116
2111	25
2112	112
2122	152

TABLEAU 4.1 – Les haplotypes associés au DAG de la figure 4.1

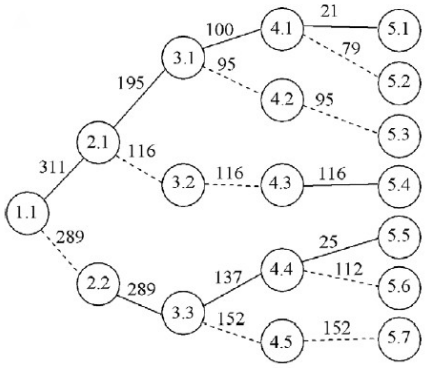


FIGURE 4.1 – Le DAG associé au tableau 4.1

Chaque arrête entre les nœuds (les cercles) du DAG, dans la figure 4.1, représente un SNP. Les arrêtes continues et en pointillés représentent l'allèle 1 et 2 respectivement. Sur chaque arrête est représenté le nombre d'haplotypes ayant une histoire particulière. Par exemple, à la troisième arrête 137 haplotypes ont l'allèle 2 au premier SNP et l'allèle 1 au deuxième et troisième SNP. Finalement les niveaux du DAG, qui se déclinent en sous-niveaux, sont représentés à l'intérieur des nœuds. Par exemple, au niveau 3 on a les sous-niveaux 3.1, 3.2 et 3.3.

L'heuristique de clusterisation des nœuds, basée sur le travail de Ron *et al.* (1998), aux différents niveaux est définie de la façon suivante. Au premier niveau du DAG correspond un seul nœud et il n'y a donc pas de clusterisation. Au niveau i ($2 \leq i \leq L$) on définit les quantités suivantes. Soient n_x le nombre d'haplotypes sortant d'un nœud x, et $n_x(a_i a_{i+1}...a_k)$ le nombre d'haplotypes qui ont la séquence $a_i a_{i+1}...a_k$ aux SNP i,$i+1$,..,k en sortant du nœud x. On notera que a_k peut être égale à 1 ou 2 selon la séquence considérée. Pour deux nœuds x et y, au niveau i, la différence des probabilités conditionnelles observées pour la séquence $a_i a_{i+1}...a_k$ est donnée par :

$$diff_{xy}(a_i a_{i+1}...a_k) = \Big|\frac{n_x(a_i a_{i+1}...a_k)}{n_x} - \frac{n_y(a_i a_{i+1}...a_k)}{n_y}\Big|$$

Les auteurs définissent ensuite un score de similarité $ss(x, y)$, sur l'ensemble des séquences possibles et de longueur variable, à partir de $diff_{xy}(a_i a_{i+1}...a_k)$, i.e.

$$\begin{aligned} ss(x,y) &= \max_{k=i,..,L-i}\Big(\max_{a_i a_{i+1}...a_k}\Big\{\Big|\frac{n_x(a_i a_{i+1}...a_k)}{n_x} - \frac{n_y(a_i a_{i+1}...a_k)}{n_y}\Big|\Big\}\Big) \\ &= \max_{k=i,..,L-i}\Big(\max_{a_i a_{i+1}...a_k}\Big\{diff_{xy}(a_i a_{i+1}...a_k)\Big\}\Big) \end{aligned}$$

Ce score de similarité, qui est plutôt un score de dissimilarité entre les nœuds x et y, sera petit si les différentes paires de probabilités conditionnelles observées, sur les différentes séquences de longueur variable, sont similaires. Le critère de regroupement de x et y est alors le suivant : si $ss(x, y)$ est plus petit qu'une certaine quantité α alors les nœuds x et y fusionneront. La fusion de x et y traduit simplement le regroupement de chromosomes, au SNP i, dont les probabilités d'avoir des suites d'allèles spécifiques après ce SNP sont suffisamment similaires. La quantité α, choisit par les auteurs, est basée sur

le travail de Ron *et al.* (1998) et est donnée par :

$$\alpha = m\sqrt{\frac{1}{n_x} + \frac{1}{n_y}} + b$$

où m et b sont des paramètres de "scale" et "shift" à définir par l'utilisateur ou à utiliser avec les valeurs fournies par défaut.

4.2.5 TP : le "Trained Predictor"

Cet AIP effectue un apprentissage par moindres carrés sur l'ensemble des SNP disponibles dans le jeu de données considéré. Il est basé sur une distance matricielle similaire, dans une certaine mesure, à celle utilisée dans l'article 1 associé au chapitre 5. Il a été considéré pour cette étude afin de comparer un nombre plus important de AIP basés sur des concepts différents. Ce prédicteur est pleinement détaillé dans l'article 1 et il ne sera donc pas décrit ici.

Chapitre 5

Méthodes de discrimination et comparaison des modèles d'association

5.1 Cadre de l'étude

L'indicateur classique utilisé pour comparer les performances des AIP en cartographie de QTL est leur précision de cartographie. Cependant, cette précision ne renseigne pas sur les mécanismes sous-jacents à une bonne ou mauvaise performance de cartographie. D'après la littérature, les données utilisées ; phénotypes, allèles aux marqueurs et QTL, sont généralement simulées afin d'évaluer la précision des AIP. On propose pour l'étude faite dans le cadre de l'Article 1, de comparer les AIP à la fois sur leur précision en cartographie et sur une efficacité relative basée sur une distance matricielle. Cette distance est définie entre les matrices de similitude produites par les AIP à une position testée et la matrice contenant les identités alléliques connues au QTL.

Cette efficacité relative s'avère fortement corrélée, d'au moins 0.9, à la précision des AIP en cartographie (cf. Article 1), montrant son utilité dans la comparaison des performances des AIP en cartographie. Les propriétés algébriques de la distance matricielle associée à cette efficacité relative ont été étudiées, afin de comprendre les mécanismes sous-jacents à la précision des AIP par l'exploitation du LD multiallélique. Des simulations basées sur des données réelles (pedigree et allèles aux marqueurs) ont permis de quantifier ces comparaison.

On suppose pour cette étude qu'il existe un QTL biallélique dans une zone sur le génome balisée par des SNP. Les haplotypes définis par ces SNP définissent un ensemble $\mathcal{I} = \{i_1, i_2, ..., i_{r-1}, i_r\}$ de positions que l'on teste pour la présence d'un QTL. Chacune de

ces positions est le centre de segments chromosomiques définis par une fenêtre glissante dans un génome scan. La figure 5.1 illustre un ensemble $\mathcal{I}$ de positions testées, définit par une fenêtre glissante de 6 SNP dans un génome scan, pour des organismes diploïdes.

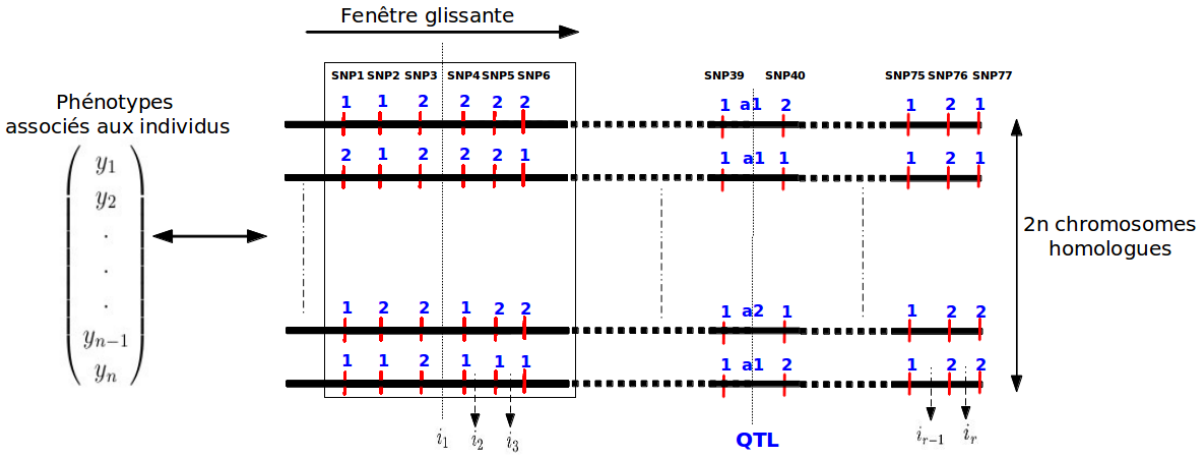


FIGURE 5.1 – Genome scan pour un ensemble $\mathcal{I}$ de positions testées définit par une fenêtre glissante de 6 SNPpour des organismes diploïdes

On remarquera que deux segments chromosomiques distincts peuvent porter les mêmes haplotypes et des allèles différents au QTL. La notion de segment chromosomique fait donc référence à l'échantillonnage existant pour un haplotype particulier. A une position testée $i \in \mathcal{I}$ on peut construire une matrice de similitude $M^{\mathcal{P},i} = (s^{\mathcal{P}}_{i,c_1,c_2})_{1\leq c_1,c_2\leq 2n}$, où $s^{\mathcal{P}}_{i,c_1,c_2} \in [0,1]$ est la prédiction d'identité allélique, à la position i pour un couple (c_1,c_2) de segments chromosomiques, associée à un prédicteur $\mathcal{P}$. La matrice $M^{\mathcal{P},i}$ contient toute l'information de similitude entre haplotypes utilisée en cartographie de QTL. De la même façon on peut construire à un QTL une matrice $M^{QTL} = (u^{QTL}_{c_1,c_2})_{1\leq c_1,c_2\leq 2n}$, où $u^{QTL}_{c_1,c_2}$ est l'identité allélique réelle au QTL pour un couple (c_1,c_2) de segments chromosomiques. On peut définir une distance normalisée entre $M^{\mathcal{P},i}$ et M^{QTL} de la façon suivante :

$$d_1(M^{\mathcal{P},i}, M^{QTL}) = \frac{1}{4n^2}\|M^{\mathcal{P},i} - M^{QTL}\|_1 = \frac{1}{4n^2}\sum_{c_1=1}^{2n}\sum_{c_2=1}^{2n}|s^{\mathcal{P}}_{i,c_1,c_2} - u^{QTL}_{c_1,c_2}| \qquad (5.1)$$

Un prédicteur $\mathcal{P}$ peut être considéré comme efficace s'il induit une distance minimale, pour la mesure d_1, à la position testée la plus proche du QTL. Autrement dit $\mathcal{P}$ sera dit efficace si $\underset{i\in\mathcal{I}}{argmin}\{\ d_1(M^{\mathcal{P},i}, M^{QTL})\ \}$ est proche du QTL. On verra dans la section suivante comment exploiter au mieux cette définition pour effectuer cette comparaison.

5.2 Principaux résultats de l'étude

5.2.1 Caractérisation de la prise en compte du LD par l'utilisation d'haplotypes

Afin de caractériser le lien existant entre le déséquilibre de liaison et la prédiction d'identité allélique par l'utilisation d'haplotypes, on propose la notion d'efficacité relative et la distance matricielle, définie en fonction des coefficients du LD multiallélique, décrites dans les sous-sections suivantes.

5.2.1.1 Outil numérique : l'efficacité relative des prédicteurs

Afin de comparer la capacité de prédiction des différents AIP par rapport au LD, on définit $\theta^{\mathcal{P}}_{r.e.}$ comme la position testée pour un ensemble $\mathcal{I}$ où $d_1(M^{\mathcal{P},i}, M^{QTL})$ est minimale. Autrement dit on a :

$$\theta^{\mathcal{P}}_{r.e.} = \underset{i\in\mathcal{I}}{argmin}\{\ d_1(M^{\mathcal{P},i}, M^{QTL})\ \}$$

On définit en conséquence l'efficacité relative d'un prédicteur $\mathcal{P}$ de la manière suivante. Soit θ_{QTL} la position du QTL sur le génome. Un prédicteur $\mathcal{P}$ est dit plus efficace qu'un prédicteur $\mathcal{P}'$ si on a :

$$\begin{cases} |\theta^{\mathcal{P}}_{r.e.} - \theta_{QTL}| < |\theta^{\mathcal{P}'}_{r.e.} - \theta_{QTL}| & (a) \\ \\ d_1(M^{\mathcal{P},\theta^{\mathcal{P}}_{r.e.}}, M^{QTL}) < d_1(M^{\mathcal{P}',\theta^{\mathcal{P}'}_{r.e.}}, M^{QTL}) & (b) \end{cases}$$

où $|.|$ est la valeur absolue. (a) stipule que la position associée à la meilleure prédiction d'identité allélique au QTL, pour le prédicteur $\mathcal{P}$, est plus proche du QTL que celle associée à $\mathcal{P}'$. (b) stipule que la prédiction d'identité allélique pour $\mathcal{P}$ à $\theta^{\mathcal{P}}_{r.e.}$ est meilleure que celle de $\mathcal{P}'$ à $\theta^{\mathcal{P}'}_{r.e.}$.

5.2.1.2 Distance matricielle en fonction de coefficients du LD

La distance définie en (5.1) peut se ré-écrire, pour un QTL biallélique, de la façon suivante :

$$d_1(M^{\mathcal{P},i}, M^{QTL}) = \sum_{p=1}^{K}\left[\ 4\Big(\sum_{q\neq p}^{K} s^{\mathcal{P}}_{i,h_p,h_q} - s^{\mathcal{P}}_{i,h_p,h_p}\Big)\mathbf{\Delta^2_p} + \Psi^{\mathcal{P}}_{pq}(\Delta_{l\neq p,q})\mathbf{\Delta_p} + \Phi^{\mathcal{P}}_{pq}(\Delta_{l\neq p,q})\ \right]$$

$$= \xi^{\mathcal{P}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K}) \tag{5.2}$$

où $(\boldsymbol{\Delta_p})_{1 \leq p \leq K}$ sont les coefficients du LD multiallélique entre les haplotypes à la position i et les allèles au QTL. $\Psi^{\mathcal{P}}_{pq}(\Delta_{l \neq p,q})$ et $\Phi^{\mathcal{P}}_{pq}(\Delta_{l \neq p,q})$ sont des termes comprenant des sommes et des produits de fréquences marginales, de prédictions d'identité allélique et de coefficients de LD. De la formulation définie en (5.2), on voit qu'il est difficile de caractériser le comportement de la distance d_1 par rapport aux coefficients du LD multiallélique pour un prédicteur continu à valeurs dans [0,1]. Cependant pour le cas particulier où l'on observe seulement deux haplotypes (5.2) se réduit à une fonction réelle d'un seul coefficient de LD entre les haplotypes et les allèles au QTL :

$$\xi^{\mathcal{P}}(\boldsymbol{\Delta_1}) = \Big[-4s^{\mathcal{P}}_{i,h_1,h_1} - 4s^{\mathcal{P}}_{i,h_2,h_2} + 8s^{\mathcal{P}}_{i,h_1,h_2} \Big] \boldsymbol{\Delta_1^2} + \Psi^{\mathcal{P}} \boldsymbol{\Delta_1} + \Phi^{\mathcal{P}} \tag{5.3}$$

où $\Psi^{\mathcal{P}}$ et $\Phi^{\mathcal{P}}$ sont des termes qui ne dépendent pas du LD. De (5.3), on voit que la distance aura une vitesse de décroissance, par rapport au LD, d'autant plus faible lorsque la prédiction $s^{\mathcal{P}}_{i,h_1,h_2} \in [0,1]$ sera d'autant plus grande. Or, les AIP par construction affecteront une valeur d'autant plus grande à $s^{\mathcal{P}}_{i,h_1,h_2}$ lorsque les haplotypes h_1 et h_2 seront d'autant plus similaires. Ainsi l'extension de la prédiction d'identité allélique de l'ensemble discret $\{0,1\}$ à l'intervalle [0,1] mène à une détérioration de la prise en compte du LD par la prédiction.

De (5.3), on remarque également que la meilleure vitesse de décroissance $-8\boldsymbol{\Delta_1^2}$ est obtenue lorsque $s^{\mathcal{P}}_{i,h_1,h_1} = s^{\mathcal{P}}_{i,h_2,h_2} = 1$ et $s^{\mathcal{P}}_{i,h_1,h_2} = 0$, ce qui correspond aux prédictions associées au prédicteur IBS_{hap}. On peut montrer que la distance définie en (5.3) ne peut atteindre zéro, pour un LD maximal, que si les prédictions d'identité allélique entre haplotypes correspondent à celles de IBS_{hap}. Autrement dit, il existe une borne inférieure strictement positive pour la distance définie en (5.3) dès lors que les prédictions associées à un AIP ne correspondent pas à celles de IBS_{hap}. Cette borne est donnée par $\frac{1}{2} s^{\mathcal{P}}_{i,h_1,h_2}$ comme on le verra dans l'article 1 suivant.

Finalement, on peut aussi montrer que la distance définie dans le cas général en (5.2) se réduit à la forme suivante pour $\mathcal{P} = IBS_{hap}$:

$$\xi^{IBS_{hap}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K}) = \sum_{p=1}^{K} \left[-4\boldsymbol{\Delta_p^2} + \Psi^{IBS_{hap}}_{pq} \boldsymbol{\Delta_p} + \Phi^{IBS_{hap}}_{pq} \right]$$

$$= \sum_{p=1}^{K} \left[Q_p(\mathbf{\Delta_p}) \right] \tag{5.4}$$

La distance définie en (5.4) est une somme de fonctions concaves, notées $(Q_p)_{1\leq p\leq K}$, en chacun des coefficients du LD multiallélique. On montre dans l'article 1, dans la section suivante, que les coefficients du LD multiallélique tendent à s'éloigner des valeurs critiques associées à ces fonctions lorsque la position testée se rapproche du QTL. Autrement dit la somme des fonctions $(Q_p)_{1\leq p\leq K}$, ou la distance induite par IBS_{hap} dans le cas général, tendra à diminuer lorsque la position testée se rapprochera du QTL. Ces résultats algébriques sont également valides pour le cas d'un QTL multiallélique (cf. article 1).

5.3 Article 1

L'article 1 suivant caractérise le lien entre le déséquilibre de liaison et la prédiction d'identité allélique en utilisant la distance définie en (5.1). Cet article compare également plusieurs AIP sur des jeux de chromosomes réels chez le porc et l'humain. Les prédicteurs comparés sont IBS_{hap}, $Score$, $P(IBD)$, $Beagle$ et un prédicteur nommé TP basé sur les moindres carrés. Le prédicteur d'identité allélique entre les allèles d'un marqueur, nommé IBS_m, a également été ajouté à la liste des AIP comparés afin de rendre l'étude plus complète dans le cadre d'un QTL biallélique. La comparaison des AIP a été effectuée sur la base de la distance matricielle définie en (5.1) et la performance de ces derniers en cartographie d'un QTL simulé. Le QTL a été simulé sur des chromosomes porcins, de la race Large White, dans des zones exhibant des niveaux de LD différents. La distance matricielle a été évaluée à la fois sur ces chromosomes porcins et sur des chromosomes humains provenant d'individus non apparentés de Beijing en Chine.

L'étude a montré que la quantité $\theta^{\mathcal{P}}_{r.e.}$ est corrélée d'au moins 0.9 à la position estimée du QTL, sur les chromosomes porcins, lorsque ce dernier explique au moins 8% de la variance totale du phénotype. L'étude a aussi montré qu'il y a peu de différence dans le comportement des AIP, par rapport à la distance matricielle, selon le jeu de données étudié (porc ou humain). Enfin, l'étude a montré la validité des expressions algébriques dans le cadre de données réelles, dans l'appréhension du comportement des différents AIP. Toutes les dérivations, associées aux principaux résultats algébriques, sont données dans les fichiers additionnels 1 et 2 en fin de l'article 1.

Jacquin *et al. Genetics Selection Evolution* 2014, **46**:45
http://www.gsejournal.org/content/46/1/45

GSE Genetics Selection Evolution

RESEARCH Open Access

Using haplotypes for the prediction of allelic identity to fine-map QTL: characterization and properties

Laval Jacquin[1,2,3*], Jean-Michel Elsen[1,2,3] and Hélène Gilbert[1,2,3]

Abstract

Background: Numerous methods have been developed over the last decade to predict allelic identity at unobserved loci between pairs of chromosome segments along the genome. These loci are often unobserved positions tested for the presence of quantitative trait loci (QTL). The main objective of this study was to understand from a theoretical standpoint the relation between linkage disequilibrium (LD) and allelic identity prediction when using haplotypes for fine mapping of QTL. In addition, six allelic identity predictors (AIP) were also compared in this study to determine which one performed best in theory and application.

Results: A criterion based on a simple measure of matrix distance was used to study the relation between LD and allelic identity prediction when using haplotypes. The consistency of this criterion with the accuracy of QTL localization, another criterion commonly used to compare AIP, was evaluated on a set of real chromosomes. For this set of chromosomes, the criterion was consistent with the mapping accuracy of a simulated QTL with either low or high effect. As measured by the matrix distance, the best AIP for QTL mapping were those that best captured LD between a tested position and a QTL. Moreover the matrix distance between a tested position and a QTL was shown to decrease for some AIP when LD increased. However, the matrix distance for AIP with continuous predictions in the [0,1] interval was algebraically proven to decrease less rapidly up to a lower bound with increasing LD in the simplest situations, than the discrete predictor based on identity by state between haplotypes (IBS_{hap}), for which there was no lower bound. The expected LD between haplotypes at a tested position and alleles at a QTL is a quantity that increases naturally when the tested position gets closer to the QTL. This behavior was demonstrated with pig and unrelated human chromosomes.

Conclusions: When the density of markers is high, and therefore LD between adjacent loci can be assumed to be high, the discrete predictor IBS_{hap} is recommended since it predicts allele identity correctly when taking LD into account.

Background

Numerous methods have been developed to predict allelic identity at an unobserved locus between pairs of chromosome segments. Such predictions are generally carried out by observing allelic similarities between the pairs of chromosome segments that surround this locus [1-3]. It is assumed that chromosome segments that exhibit more similarities have a higher chance of harboring the same allele(s) at this locus. Many of these methods [1-5] use either directly or implicitly the concept of identity-by-descent (IBD), and therefore predict allelic identity based on allelic likeness. Such predictions of allelic identity can be either continuous or discrete in the [0,1] interval. The matrices that contain these predictions for pairs of chromosome segments, at an unobserved locus, can be used in a statistical procedure to detect association between the locus and some phenotypes of interest. For example, these matrices can be interpreted as being proportional to the

*Correspondence: Julien.Jacquin@toulouse.inra.fr
[1] INRA, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), F-31326, Castanet-Tolosan, France
[2] Université de Toulouse, INP, ENSAT, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), F-31326, Castanet-Tolosan, France
Full list of author information is available at the end of the article

covariance matrices of the effect of the locus on phenotypes of interest [1,4,6] and therefore play a central role in the statistical analysis of the variability. The similarity between chromosome segments can be measured based on the haplotypes of markers carried by the segments. Indeed, it has been shown that haplotype-based methods have a higher potential to detect trait-marker associations than single-marker methods in some cases [7-16]. Different methods for predicting allelic identity, hereafter named Allelic Identity Predictors (AIP), have been proposed and in this study, we have compared some of these methods i.e.: (1) the probability measure described by Meuwissen and Goddard [1] is the conditional probability of being IBD at an unobserved locus for pairs of haplotypes, given the identical-by-state (IBS) status of alleles spanning that position; (2) the similarity score of Li and Jiang [2] calculates the sum of the number of shared alleles and the length of the longest shared substring that spans an unobserved locus for pairs of haplotypes; (3) the probability model of Browning [3] is based on Variable Length Markov Chains (VLMC) and performs chromosome clustering at a given marker, and in this model, chromosomes that belong to a given cluster are considered as potentially harboring the same unobserved allele(s) locally; and (4) the IBS status of all marker alleles between pairs of haplotypes and (5) the IBS status of single markers alleles, which are the simplest AIP.

In some association studies, such as those that use random effect models for example, the only input that differs from one AIP to another is the similarity (covariance) matrix built for the tested location. Thus, investigating the properties of similarity matrices is another strategy when comparing AIP, since this comparison is generally based on the accuracy of quantitative trait locus (QTL) localization (e.g. root mean square error). The main objective of the present study was to understand the relation between linkage disequilibrium (LD) and allelic identity prediction when using haplotypes, by identifying the properties of similarity matrices in the neighborhood of a QTL and at the QTL. This was performed using a simple distance measure between these matrices and the similarity matrix at the QTL based on the observed allelic identity (IBS). This distance measure was expressed analytically in terms of LD coefficients. There has been an increasing interest in taking advantage of LD for fine-mapping of complex disease genes [17-20] and QTL [21-24]. Nevertheless, to the best of our knowledge, no study has yet used analytical methods to compare AIP in relation to LD. Here, we define a new criterion based on the chosen matrix distance measure, which allows discrimination between the six AIP. We evaluated the consistency of this criterion with the mapping accuracy of the six AIP for a QTL simulated according to different LD patterns and populations.

The simulations were based on two population types, a set of human chromosomes and a set of porcine chromosomes, with different LD and density patterns. In each case, the QTL was a hidden SNP that simulated a biallelic QTL, as previously proposed [4,8,23,25]. Hence, the present study was framed around the common idea that there is a favorable allele at the QTL, which affects an observed trait. In this context, the aim of AIP is to predict, at the QTL, whether both chromosomal segments of any pair harbor the same unobserved favorable allele or not, which is the same as predicting the IBS or non-IBS state of the alleles. A new (6) unreferenced AIP, named trained predictor and abbreviated as TP, is also compared in this paper. This new predictor, based on a matrix distance concept similar to the one used to discriminate between the AIP, performs least squares prediction in a global fashion over chromosomes. The purpose of this predictor was to investigate the behavior of an AIP which performs global training over the chromosomes in relation to local patterns of LD.

Methods

Matrix distance comparison

Let $\mathcal{I} = \{i_1, \ldots, i_r\}$ be a set of positions that are tested for the presence of a QTL on $2n$ phased homologous chromosomes for n diploid individuals. Only one QTL is considered to be in the screened region. In a sliding window approach, each position tested is considered to be the unobserved center of the haplotypes carried by different chromosome segments. Figure 1 shows an example of tested positions for a sliding window of six markers and a QTL located between SNPs 39 and 40.

Let $s^{\mathcal{P}}_{i,c_1,c_2} \in [0,1]$ be the IBS or IBD prediction of allelic identity, depending on an AIP $\mathcal{P}$ at a tested position $i \in \mathcal{I}$, for a couple (c_1, c_2) of chromosome segments. Note that $s^{\mathcal{P}}_{i,c_1,c_2}$ is calculated according to the observed similarity between the haplotypes carried by c_1 and c_2. Hence, c_1 and c_2 can harbor different unobserved alleles at i even if these segments carry the same haplotype. We define $\mathbf{M}^{\mathcal{P},i} = \left(s^{\mathcal{P}}_{i,c_1,c_2}\right)_{1 \leq c_1,c_2 \leq 2n}$ as the similarity matrix built from the predictions of allelic identity at locus i for $\mathcal{P}$. Matrix $\mathbf{M}^{\mathcal{P},i}$ can be used in a statistical procedure to detect association between i and some phenotype of interest.

Let $u^{QTL}_{c_1,c_2} \in \{0,1\}$ be the true allelic identity observed at the QTL (IBS) for a couple (c_1, c_2) of chromosome segments. On the basis of known alleles at the QTL, the similarity $\mathbf{M}^{QTL} = \left(u^{QTL}_{c_1,c_2}\right)_{1 \leq c_1,c_2 \leq 2n}$ can be built with the real allelic identities. Note that $\mathbf{M}^{QTL}$ is simply a similarity matrix that describes the IBS or non-IBS state of alleles at the QTL.

Let d_1 be a normalized distance measure between $\mathbf{M}^{\mathcal{P},i}$ and $\mathbf{M}^{QTL}$ induced by the entrywise 1-norm, which is the

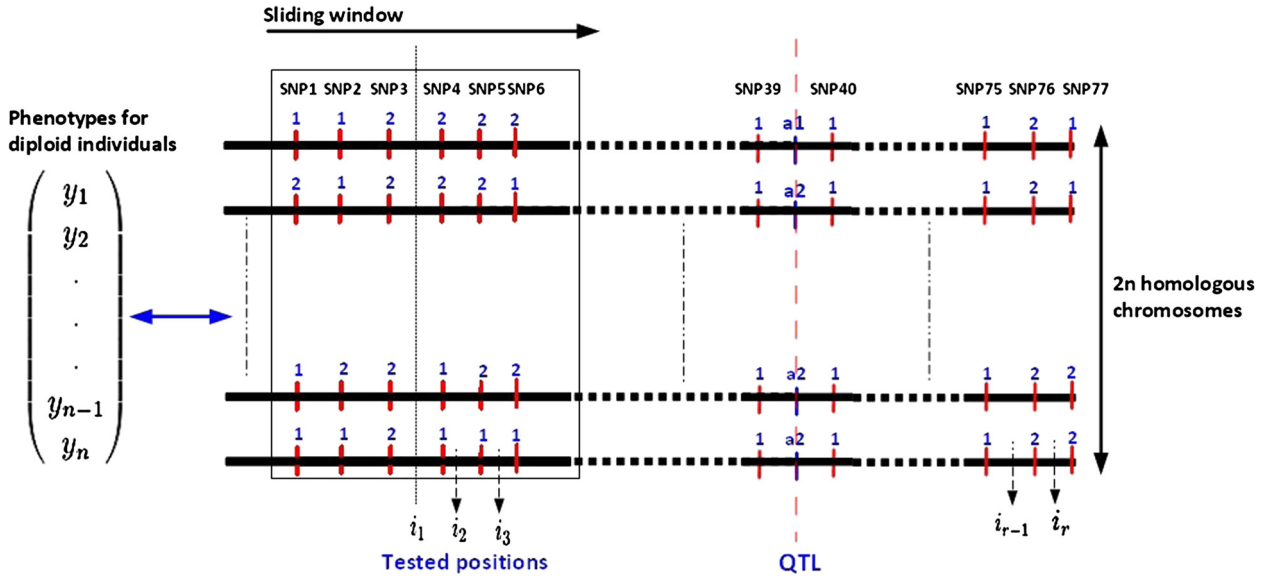


Figure 1 Principle of the sliding window approach when screening a chromosome for the set $\{i_1, \ldots, ir\}$ of tested positions for associations between phenotypes and 6-marker haplotypes.

sum of the absolute differences between the elements of two matrices or vectors, i.e.

$$d_1\left(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}\right) = \frac{1}{4n^2}\|\mathbf{M}^{\mathcal{P},i} - \mathbf{M}^{QTL}\|_1 = \frac{1}{4n^2}\sum_{c_1=1}^{2n}\sum_{c_2=1}^{2n}|s^{\mathcal{P}}_{i,c_1,c_2} - u^{QTL}_{c_1,c_2}|$$

Note that some AIP have continuous prediction errors $|s^{\mathcal{P}}_{i,c_1,c_2} - u^{QTL}_{c_1,c_2}|$ in $[0, 1]$, while for others, prediction errors are limited to the discrete set $\{0, 1\}$. Measure d_1 is therefore more appropriate than the euclidean metric d_2, for example, because it does not shrink continuous prediction errors in $[0, 1]$. Let θ_{QTL} be the position of the QTL. When a predictor $\mathcal{P}$ performs well, $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ should be minimum at the tested position closest to θ_{QTL}. Hence $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ can be used to compare different AIP for a set of tested positions. Note that $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ can also be expressed as [see Additional file 1]:

$$d_1\left(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}\right) = \sum_{p=1}^{K} f_{i,h_p} \sum_{q=1}^{K} f_{i,h_q}\left[p^{QTL}_{i,h_p,h_q} \times \left(1 - s^{\mathcal{P}}_{i,h_p,h_q}\right) + \left(1 - p^{QTL}_{i,h_p,h_q}\right) s^{\mathcal{P}}_{i,h_p,h_q}\right] \quad (1)$$

where $K = 2^t$ is the number of possible observed haplotypes at position i, for a sliding window of t markers. f_{i,h_p} and f_{i,h_q} are the frequencies of haplotypes h_p and h_q at position i, respectively. Note that some haplotypes among the K possible haplotypes may not be observed in practice. Hence, the corresponding frequencies for these haplotypes will naturally be equal to 0 in expression (1). p^{QTL}_{i,h_p,h_q} is the proportion of identical alleles shared at the QTL by the pairs of chromosomes that carry h_p and h_q, at position i, and $s^{\mathcal{P}}_{i,h_p,h_q}$ is the prediction of allelic identity at locus i for the predictor $\mathcal{P}$ and a pair (h_p, h_q) of haplotypes. Expression (1) will be used subsequently to express d_1 as a function of LD coefficients, and to understand the trained predictor defined in this paper.

Measures of AIP evaluated

The AIP evaluated in this study were IBS_m (IBS status of alleles at single markers), IBS_{hap} (IBS status of all marker alleles between pairs of haplotypes), P(IBD) (IBD probability of Meuwissen and Goddard [1]), Score (similarity score of Li and Jiang [2]), Beagle (cluster-based probability model of Browning [3]) and TP (the trained predictor). Note that the tested positions coincide with marker positions for IBS_m and Beagle. These positions are therefore different from those in Figure 1. The tested positions for IBS_{hap}, P(IBD), Score and TP are defined as presented in Figure 1.

IBS_m gives an allelic identity prediction of 1 if a pair of chromosome segments carries the same allele at a tested marker and 0 otherwise. With IBS_{hap} the prediction of allelic identity is equal to 1 if both chromosome

segments of a pair carry the same marker alleles for haplotypes that span the tested position i, and 0 otherwise. P(IBD) is an estimation of the conditional probability of being IBD at i for a pair of chromosome segments, given the IBS status of marker alleles of the haplotypes spanning i. This measure of probability is based on a coalescence process and models recombination between markers. The P(IBD) function was applied here with an ancestral effective population size of 100 and 100 generations from the base population, as in Meuwissen and Goddard [1]. Meuwissen and Goddard [23] showed that violations of these assumptions, i.e. that alter the effective population size and the number of generations since the base population, had no effect on the mapping accuracy of their methods [23,26]. For a pair of haplotypes carried by two chromosome segments, Score is the summation of the number of IBS alleles and the length of the longest common substring of IBS alleles that span i. Score integrates weight functions that decrease the significance of markers based on their genetic distance from i. As proposed in Li and Jiang [2], these functions were chosen to be one minus the distance, in centiMorgan (cM), of each marker from i on the haplotypes within the sliding window (as presented in Figure 1). Beagle clusters chromosomes or haplotypes locally at a tested marker if they have similar probabilities of carrying the same alleles at following adjacent markers. The Beagle probability model was built at each marker by running the Beagle software (Beagle 3.3.2; http://faculty.washington.edu/browning/beagle/beagle.html, Browning [3], Browning and Browning [12]) and fitting all the chromosome markers at one time. The Beagle probability model needs two parameters (scale and shift) to be built. These parameters were first estimated from the data using a cross-validation procedure. However, the mapping results were less accurate than those obtained with the default values for these parameters that were proposed by the authors. According to the authors, the default values have performed well in simulation studies and real data analyses [12,27]. Hence the default values scale = 4.0 and shift = 0.2 [12] were used.

The trained predictor (TP), built by least squares prediction, is based on the idea that pairs of haplotypes that exhibit the same amount of allelic similarity should have the same probability of harboring identical alleles, regardless of the tested positions they span. Estimates for $\left(s^{TP}_{i,h_p,h_q}\right)_{(p,q)\in\{1,..,K\}^2}$ can be obtained as follows. Let $\mathcal{J} = \{j_1, \ldots, j_T\}$ be a set of observed SNPs on chromosomes, which are called target SNPs. Each target SNP j is defined as the middle marker of a sliding window of $t+1$ loci, where t is the number of observed flanking markers used to predict allelic identity at the target SNP. Let $u_{j,c_1,c_2} \in \{0,1\}$ be the real allele identity at j for (c_1, c_2) and let $\mathcal{E}^{TP}$ be the mean squared prediction errors over $\mathcal{J}$ for TP, i.e.

$$\begin{aligned}
\mathcal{E}^{TP} &= \frac{1}{T}\sum_{j=j_1}^{j_T}\left[\frac{1}{4n^2}\sum_{c_1=1}^{2n}\sum_{c_2=1}^{2n}\left(s^{TP}_{c_1,c_2} - u_{j,c_1,c_2}\right)^2\right] \\
&= \frac{1}{T}\sum_{j=j_1}^{j_T}\left[\, d_2\left(\mathbf{M}^{TP,j}, \mathbf{M}^{j}\right)\,\right] \\
&= \frac{1}{T}\sum_{j=j_1}^{j_T}\left[\sum_{p=1}^{K} f_{j,h_p}\sum_{q=1}^{K} f_{j,h_q}\left[\, p_{j,h_p,h_q}\left(s^{TP}_{h_p,h_q} - 1\right)^2\right.\right. \\
&\quad \left.\left. + \left(1 - p_{j,h_p,h_q}\right)\left(s^{TP}_{h_p,h_q} - 0\right)^2\right]\right]
\end{aligned}$$

Note that the expression of the normalized squared euclidean distance, d_2, in terms of frequencies and proportions is analogous to that of d_1 in (1).

Indeed f_{j,h_p}, f_{j,h_q} and p_{j,h_p,h_q} at locus j are defined as in (1). Estimates for $\left(s^{TP}_{i,h_p,h_q}\right)_{(p,q)\in\{1,..,K\}^2}$ are obtained by differentiating $\mathcal{E}^{TP}$ with respect to $s^{TP}_{h_p,h_q}$, i.e.

$$\frac{\partial \mathcal{E}^{TP}}{\partial s^{TP}_{h_p,h_q}} = 0 \Longleftrightarrow \hat{s}^{TP}_{h_p,h_q} = \frac{\sum_{j=j_1}^{j_T} f_{j,h_p} f_{j,h_q} p_{j,h_p,h_q}}{\sum_{j=j_1}^{j_T} f_{j,h_p} f_{j,h_q}}$$

Note that the second derivative of $\mathcal{E}^{TP}$ with respect to $s^{TP}_{h_p,h_q}$ is positive since it is a sum of frequencies. This implies that $\mathcal{E}^{TP}$ reaches a minimum for the set of estimates $\left(\hat{s}^{TP}_{h_p,h_q}\right)_{(p,q)\in\{1,..,K\}^2}$, since $\mathcal{E}^{TP}$ is a sum of convex functions of each $s^{TP}_{h_p,h_q}$. Hence, TP associates $\hat{s}^{TP}_{h_p,h_q}$ to any observed couple $\left(h_p, h_q\right)$ at any tested position $i \in \mathcal{I}$. The observed target SNPs ($j \in \mathcal{J}$) are used to estimate the predictions of allelic identity for TP and should not be confused with the unobserved tested positions ($i \in \mathcal{I}$).

Statistical models, test statistic and relative efficiency

Mixed models

The following mixed models were used to test for the presence of a QTL at a given position $i \in \mathcal{I}$ for all AIP:

$$\begin{cases} \mathbf{Y} = \mathbf{X}\beta + \mathbf{Z_h}\mathbf{h} + \mathbf{Z_u}\mathbf{u} + \boldsymbol{\varepsilon} & (\mathrm{H}_1) \\ \mathbf{Y} = \mathbf{X}\beta + \mathbf{Z_u}\mathbf{u} + \boldsymbol{\varepsilon} & (\mathrm{H}_0) \end{cases}$$

where β is a fixed effect, which is the overall mean, and $\mathbf{X} = \mathbb{1}_n$ is a vector of n ones. Vector $\mathbf{u}$ represents the random polygenic effects due to relationships among individuals, i.e. $\mathbf{u} \sim \mathcal{N}_n\left(0, \mathbf{A}\sigma^2_{\mathrm{u}}\right)$ where $\mathbf{A}$ is the additive relationship matrix built from the pedigree [28,29]. $\mathbf{Z_h}$ and $\mathbf{Z_u}$ are design matrices that link random effects to individuals and $\boldsymbol{\varepsilon}$ is the vector of homoscedastic error terms, i.e. $\boldsymbol{\varepsilon} \sim \mathcal{N}_n\left(0, \mathbf{I}_n\sigma^2_{\varepsilon}\right)$.

In the model corresponding to (H_1), $\mathbf{h}$ is a vector of random effects of haplotypes at position i, i.e. $\mathbf{h} \sim \mathcal{N}_\kappa(0, \mathbf{H}^{\mathcal{P},i}\sigma^2_{\mathrm{h}})$, where κ $(\kappa \leq K)$ corresponds to the number of observed haplotypes, or alleles, at position i. Note that $\mathbf{h}$ has the same dimension κ for all AIP except for $\mathrm{IBS_m}$ and Beagle. The tested positions coincide with marker positions for these two predictors. At a tested marker i, $\kappa = 2$ for $\mathrm{IBS_m}$ and κ is equal to the number of local clusters for Beagle. Therefore, depending on the predictor $\mathcal{P}$, $\mathbf{H}^{\mathcal{P},i}$ is a similarity matrix based on either distinct observed haplotypes (e.g. $\mathcal{P}$ = Score) or distinct clusters (e.g. $\mathcal{P}$ = Beagle). Note that $\mathbf{H}^{\mathcal{P},i}$ and $\mathbf{M}^{\mathcal{P},i}$ are equivalent sources of data contingent upon the list of haplotypes, or distinct local clusters, for the chromosome segments at any tested position. Indeed, depending on $\mathcal{P}$, one can build $\mathbf{M}^{\mathcal{P},i}$ from $\mathbf{H}^{\mathcal{P},i}$ in one of the two following ways. (1) $\mathbf{M}^{\mathcal{P},i}_{c_1,c_2} = \mathbf{H}^{\mathcal{P},i}_{h(c_1),h(c_2)}$, where $h(c_1)$ and $h(c_2)$ are the haplotype numbers carried by chromosomes c_1 and c_2, respectively or (2) $\mathbf{M}^{\mathcal{P},i}_{c_1,c_2} = \mathbf{H}^{\mathcal{P},i}_{\mathcal{C}(c_1),\mathcal{C}(c_2)}$, where $\mathcal{C}(c_1)$ and $\mathcal{C}(c_2)$ are the cluster numbers to which chromosomes c_1 and c_2 belong, respectively.

RLRT statistic

The Expectation-Maximization algorithm was used for the restricted maximum likelihoods of the mixed models [30-33], to estimate the components β, $\mathbf{h}$, $\mathbf{u}$, ε and the variance terms $\sigma^2_{\mathrm{h}}, \sigma^2_{\mathrm{u}}, \sigma^2_{\varepsilon}$. Let $\lambda^{\mathcal{P}}_i$ be the restricted maximum likelihood ratio test (RLRT) of (H_1) versus (H_0) for position i, i.e.

$$\lambda^{\mathcal{P}}_i = -2ln\left(\frac{L_{REML}(\mathrm{H}_0)}{L^{\mathcal{P}}_{REML}(\mathrm{H}_1)}\right)$$

We defined $\theta^{\mathcal{P}}_{\mathrm{m.a.}}$ as the estimated position of a QTL for a predictor $\mathcal{P}$, i.e.

$$\theta^{\mathcal{P}}_{\mathrm{m.a.}} = \underset{i\in\mathcal{I}}{argmax}\left\{\hat{\lambda}^{\mathcal{P}}_i\right\}$$

Relative efficiency

To compare the predictive ability of the different predictors in relation to LD, we defined $\theta^{\mathcal{P}}_{\mathrm{r.e.}}$ as the tested position where $d_1\left(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}\right)$ is minimized for a predictor $\mathcal{P}$, i.e.

$$\theta^{\mathcal{P}}_{\mathrm{r.e.}} = \underset{i\in\mathcal{I}}{argmin}\left\{d_1\left(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}\right)\right\}$$

Consequently, we defined the relative efficiency of a predictor $\mathcal{P}$ as follows. Predictor $\mathcal{P}$ is considered to be more efficient than a predictor $\mathcal{P}'$ if

$$\begin{cases} |\theta^{\mathcal{P}}_{\mathrm{r.e.}} - \theta_{QTL}| < |\theta^{\mathcal{P}'}_{\mathrm{r.e.}} - \theta_{QTL}| & (a) \\ d_1\left(\mathbf{M}^{\mathcal{P},\theta^{\mathcal{P}}_{\mathrm{r.e.}}}, \mathbf{M}^{QTL}\right) < d_1\left(\mathbf{M}^{\mathcal{P}',\theta^{\mathcal{P}'}_{\mathrm{r.e.}}}, \mathbf{M}^{QTL}\right) & (b) \end{cases}$$

where $|.|$ is the absolute value. When $\theta^{\mathcal{P}}_{\mathrm{r.e.}}$ was not unique, the mean of the different argmins was retained as $\theta^{\mathcal{P}}_{\mathrm{r.e.}}$. Inequality (a) states that the tested position associated with the best prediction, of the allele identity at the QTL, is closer to the QTL for $\mathcal{P}$ than that for $\mathcal{P}'$. Inequality (b) states that the true allelic identity at the QTL is better predicted by $\mathcal{P}$ at $\theta^{\mathcal{P}}_{\mathrm{r.e.}}$ than by $\mathcal{P}'$ at $\theta^{\mathcal{P}'}_{\mathrm{r.e.}}$.

Comparison criteria

N simulations $(w = 1, .., N)$ were performed to evaluate the mapping accuracy and the relative efficiency of the different AIP in different situations.

Mapping accuracy

The mapping accuracy of the simulated QTL was evaluated for each AIP with the root mean square error (RMSE):

$$\mathrm{RMSE}^{\mathrm{m.a.}} = \sqrt{\frac{1}{N}\sum_{w=1}^{N}\left(\theta^{\mathcal{P},w}_{\mathrm{m.a.}} - \theta_{QTL}\right)^2}$$

Relative efficiency

The relative efficiency of each AIP was evaluated by considering the three following quantities:

$$\begin{cases} \mathrm{RMSE}^{\mathrm{r.e.}} = \sqrt{\frac{1}{N}\sum_{w=1}^{N}\left(\theta^{\mathcal{P},w}_{\mathrm{r.e.}} - \theta_{QTL}\right)^2} \\ \hat{\mathbb{E}}^{\mathrm{r.e.}} = \frac{1}{N}\sum_{w=1}^{N} d_1\left(\mathbf{M}^{\mathcal{P},\theta^{\mathcal{P},w}_{\mathrm{r.e.}}}, \mathbf{M}^{QTL,w}\right) \\ \hat{\sigma}^{\mathrm{r.e.}} = \sqrt{\frac{1}{N}\sum_{w=1}^{N}\left(d_1\left(\mathbf{M}^{\mathcal{P},\theta^{\mathcal{P},w}_{\mathrm{r.e.}}}, \mathbf{M}^{QTL,w}\right) - \hat{\mathbb{E}}^{\mathrm{r.e.}}\right)^2} \end{cases}$$

where $\mathrm{RMSE}^{\mathrm{r.e.}}$ and $\hat{\mathbb{E}}^{\mathrm{r.e.}}$ measure conditions (a) and (b), defined in the paragraph on relative efficiency, and $\hat{\sigma}^{\mathrm{r.e.}}$ measures the standard deviation of the matrix distance at $\theta^{\mathcal{P}}_{\mathrm{r.e.}}$.

Data for simulation

A sliding window of $t = 6$ markers was chosen for all analyses, except for $\mathrm{IBS_m}$ and Beagle. Windows of six and 12 markers were previously shown to be optimal for QTL mapping accuracy [34,35] with 60K type

SNP chips. Hence, all analyses were done using a sliding window of $t = 6$ markers, except for IBS_m and Beagle, to make comparison between the series of results easier. A set of 90 human chromosomes 21 from unrelated Han Chinese individuals from Beijing (HCB), and a set of 235 swine chromosomes 18 from French Large White (FLW) pigs, were used for LD and matrix distance computations. The 90 HCB chromosomes were genotyped for 16 881 SNPs and are available from the HapMap project website (http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/2005-03_phaseI/full/). The FLW chromosomes were genotyped for 1252 SNPs using the Illumina Porcine 60K+SNP iSelect Beadchip [36]. Only 14 976 SNPs on the HCB chromosomes and 969 SNPs on the FLW chromosomes for which the minor allele frequency was greater than 5% were retained for analysis. The LD and matrix distance computations were conducted for the HCB and the FLW chromosomes. The QTL simulations were only conducted for the FLW chromosomes for which a pedigree was available. The marker density varied across the FLW chromosomes based on physical distance in kilobase. One megabase was considered equivalent to 1 cM for conversion in this study.

Variation of LD between tested positions and a QTL

LD between a tested position i and a QTL was measured using the multiallelic measure R of LD as suggested by [37-39]. Let $\mathbf{\Delta_p} = f^{QTL}_{i,h_p a_1} - f_{i,h_p} f_{a_1}$ be the LD coefficient between haplotype h_p at position i and allele a_1 at the QTL. $f^{QTL}_{i,h_p a_1}$ is the frequency of haplotype $h_p a_1$ defined by the marker haplotype h_p that spans position i and allele a_1 at the QTL. f_{a_1} is the frequency of allele a_1 at the QTL and f_{i,h_p} is haplotype h_p frequency at i. Note that $-\mathbf{\Delta_p} = f^{QTL}_{i,h_p a_2} - f_{i,h_p} f_{a_2}$. Hence, for a biallelic QTL, R can be expressed as:

$$R_{i,QTL} = \frac{\sum_{p=1}^{K}\sum_{l=1}^{2}\left(f^{QTL}_{i,h_p a_l} - f_{i,h_p} f_{a_l}\right)^2}{\left(1-\sum_{p=1}^{K} {f^{QTL}_{i,h_p}}^2\right)\left(1-\sum_{l=1}^{2} {f_{a_l}}^2\right)}$$

$$= \frac{\sum_{p=1}^{K}\left[\left(\mathbf{\Delta_p}\right)^2 + \left(-\mathbf{\Delta_p}\right)^2\right]}{\left(1-\sum_{p=1}^{K} {f^{QTL}_{i,h_p}}^2\right)\left(1-\sum_{l=1}^{2} {f_{a_l}}^2\right)}$$

$$= \frac{2\sum_{p=1}^{K}\mathbf{\Delta_p^2}}{\left(1-\sum_{p=1}^{K} {f^{QTL}_{i,h_p}}^2\right)\left(1-\sum_{l=1}^{2} {f_{a_l}}^2\right)} = \frac{D^2_{i,QTL}}{H_i.H_{QTL}}$$

where $H_i = 1 - \sum_{p=1}^{K} {f^{QTL}_{i,h_p}}^2$ and $H_{QTL} = 1 - \sum_{l=1}^{2} {f_{a_l}}^2$ are the Hardy-Weinberg heterozygosities at i and the QTL respectively and $D^2_{i,QTL} = 2\sum_{p=1}^{K}\mathbf{\Delta_p^2}$. $R_{i,QTL}$ and $D^2_{i,QTL}$ are expected to increase as the tested position i gets closer to a QTL. The general behaviors of the normalized measure $R_{i,QTL}$ and the non-normalized measure $D^2_{i,QTL}$ were described by computing the LD between the haplotypes at successive distinct positions, using a sliding window, and the alleles of a fixed SNP centered over a region of 81 markers on the chromosomes. The fixed SNP was centered over a region of 76 distinct overlapping sliding windows available within the region of 81 markers. The 76 distinct positions associated to the windows played the role of the tested positions of an association study. The fixed SNP played the role of a biallelic QTL. The computation was repeated for all possible regions of 81 successive markers. Since 969 SNPs were retained on the 235 porcine chromosomes, computation was performed for 889 ($969 - 81 + 1 = 889$) regions of 81 markers. The same procedure was performed on the HCB chromosomes, thus leading to 14 896 possible regions for this set of chromosomes. The empirical means of the 889 FLW and the 14 896 HCB LD profiles were then computed to describe the expected behaviors of $R_{i,QTL}$ and $D^2_{i,QTL}$. Another major purpose of these computations was to help the analytical comparison of the AIP and the associated matrix distances, which can be expressed as elements of multiallelic LD (see Results section).

Distributions of matrix distance as a function of multiallelic LD

The distributions of the matrix distance for the six compared AIP, as function of local multiallelic LD, were also evaluated on the FLW and HCB chromosomes. The matrix distances for the six AIP were calculated at 966 and 14 973 possible target SNPs for the FLW and HCB chromosomes, respectively. The target SNPs were defined in exactly the same way as used for the trained predictor (TP). The matrix distances calculated at each window that harbors a target SNP for the six AIP were then plotted against the multiallelic measure R of LD between the haplotypes and the target alleles within the window.

QTL simulation on FLW chromosomes

The 235 FLW chromosomes were included in $N = 200$ gene-drop simulations, in a 25-generation pedigree for the FLW breed, using the LDSO software [40]. The pedigree was composed of 1594 founders, 3373 sires and 7100 dams. The gene-drop procedure was used to generate different realistic genealogy structures between the chromosomes. For each gene-drop the 235 FLW chromosomes were uniformly distributed, with replacement, among the 1594 founders of the pedigree. Hence, the measured LD structure for mapping among descendant individuals at the end of each gene-drop was almost the same as on the 235 FLW chromosomes. It must be emphasized that the use of replicates of only 235 chromosomes to populate 1594 diploid founders, followed by 25 generations

of recombinations events, means that the number of different haplotypes at a position is much lower than 3188 (2×1594). Thus, the results correspond to medium range population sizes. After each gene-drop, only the chromosomes and phenotypes of the $n = 485$ individuals of generation 25 were retained for subsequent analyzes.

Three distant SNPs were chosen as putative QTL, in order to have different LD levels with the six-marker haplotype that surrounds them on the 235 initial FLW chromosomes. Two different QTL effects were simulated for each of these SNPs, thus leading to six different scenarios. The LD between these SNPs and the observed haplotypes that harbored them was measured using the multiallelic measure R of LD. The LD levels around the three SNPs were equal to 0.52, 0.18 and 0.08, and the lengths of the haplotypes harboring them were equal to 0.09 cM, 0.37 cM and 0.75 cM, respectively. Note that these differences in length were due to the different marker densities in the distinct regions that harbor each putative QTL. The length of the region scanned for QTL mapping around each simulated QTL was approximately 3 cM.

The phenotypes in the pedigree were computed as $y_i = \frac{1}{2}\left(p_i^f + p_i^m\right) + \phi_i + g_i^{QTL} + \delta$, where p_i^f, p_i^m are normal random polygenic effects of the parents with variance 0.5, ϕ_i is a normal random mendelian sampling effect with variance 0.25 and δ is a normal random environmental effect with variance 1. g_i^{QTL} is the QTL genotype effect of individual i. QTL genotype effect was first computed as $g_i^{QTL} = 2$ or 0 or -2, if the QTL genotype of individual i was a_1a_1 or a_1a_2 or a_2a_2 respectively. In the same way a second set of simulations was carried out with the QTL genotype effect computed as $g_i^{QTL} = 0.5$ or 0 or -0.5. Only the gene-drop simulations for which the minor allele frequency at the QTL was greater or equal to 0.1 were retained. Each simulated QTL was verified for Hardy-Weinberg equilibrium during simulations. Hence, under the standard model, where the dominance effect is equal to 0 as in this study, the first simulated QTL effect explained at most 57% of the phenotypic variance for equal frequencies at the QTL. In the same way, the second simulated QTL effect explained at most 8% of the phenotypic variance.

Results

This section gives theoretical and empirical results that show that, compared to others, some AIP exhibit a better behavior for the decrease of their matrix distance, as defined by expression (1), when the multiallelic LD between a tested position and a QTL increases. In summary, the theoretical results show that expression (1) can be written as a function of the multiallelic LD coefficients of R, and that the decreasing behavior of this function depends on the nature of the AIP (see equations (2), (3), (4), (5) and (6) of this section). The empirical results show that R is expected to be highest when the tested position is closest to the QTL (see Figure 2 of this section). The expectation taken for the multiallelic LD was the empirical mean, which was found to converge for distant regions on the chromosomes. These regions can be assumed to be independent, thus showing an expected behavior for the multiallelic LD. The empirical results also show that the tested position that minimizes the matrix distance is highly correlated with the mapping accuracy of the AIP (see sub-section on mapping accuracy and relative efficiency of this section).

Variation of LD between tested positions and a QTL

Figure 2 shows the empirical means of the 889 FLW and the 14 896 HCB LD profiles for $R_{i,QTL}$ and $D^2_{i,QTL}$.

In Figure 2 the values of $R_{i,QTL}$ and $D^2_{i,QTL}$ increase, as expected, as the tested position i moves closer to the QTL. This implies that the sum of the $\mathbf{\Delta_p^2}$ terms increases on average as position i moves toward the QTL. The highest expected values for $R_{i,QTL}$ and $D^2_{i,QTL}$, in Figure 2 are reached for the tested position closest to the QTL. Note that the range of values for $D^2_{i,QTL}$, in Figure 2 is smaller than that of $R_{i,QTL}$. This is due to the lack of a normalization factor for $D^2_{i,QTL}$.

Matrix distance as function of multiallelic LD coefficients

Based on expression (1), $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ can be re-written as [see Additional file 1]:

$$\begin{aligned} d_1\left(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}\right) = \sum_{p=1}^{K}\sum_{q=1}^{K}\Big[\Big[f^{QTL}_{i,h_pa_1}f^{QTL}_{i,h_qa_1} \\ + f^{QTL}_{i,h_pa_2}f^{QTL}_{i,h_qa_2}\Big]\left(1 - s^{\mathcal{P}}_{i,h_p,h_q}\right) \\ + \left[f^{QTL}_{i,h_pa_2}f^{QTL}_{i,h_qa_1} + f^{QTL}_{i,h_pa_1}f^{QTL}_{i,h_qa_2}\right]s^{\mathcal{P}}_{i,h_p,h_q}\Big] \end{aligned} \quad (2)$$

Replacing the $2K$ frequencies in expression (2) by the $(\mathbf{\Delta_p})_{1\le p\le K}$ LD coefficient terms and the product of marginal frequencies gives [see Additional file 1]:

$$\begin{aligned} d_1\left(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}\right) = \sum_{p=1}^{K}\Bigg[4\left(\sum_{q\neq p}^{K} s^{\mathcal{P}}_{i,h_p,h_q} - s^{\mathcal{P}}_{i,h_p,h_p}\right)\mathbf{\Delta_p^2} \\ + \Psi^{\mathcal{P}}_{pq}\left(\Delta_{l\neq p,q}\right)\mathbf{\Delta_p} + \Phi^{\mathcal{P}}_{pq}\left(\Delta_{l\neq p,q}\right)\Bigg] \\ = \xi^{\mathcal{P}}\left(\mathbf{\Delta_1}, .., \mathbf{\Delta_K}\right), \end{aligned} \quad (3)$$

where $\Psi^{\mathcal{P}}_{pq}\left(\Delta_{l\neq p,q}\right)$ and $\Phi^{\mathcal{P}}_{pq}\left(\Delta_{l\neq p,q}\right)$ are sums and products of marginal frequencies, allelic identity predictions and LD coefficient terms. The general behavior

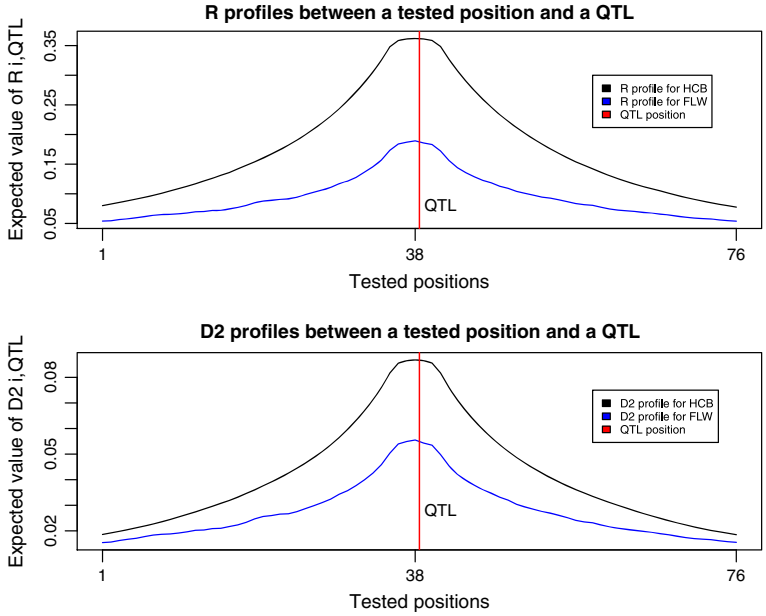


Figure 2 Empirical means of the 889 FLW and 14 896 HCB LD profiles, obtained for the normalized and the non-normalized multiallelic LD between tested positions (tested position *i* = center of six marker haplotypes) and a biallelic QTL (red vertical line), for regions of 81 markers on chromosomes.

of $\xi^{\mathcal{P}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$, with respect to $R_{i,QTL}$, is unspecifiable due to its complexity. For instance, the behavior of $\xi^{\mathcal{P}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$ cannot be specified for continuous AIP in [0,1]. However for $\mathcal{P} = \text{IBS}_{\text{hap}}$, $\xi^{\mathcal{P}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$ reduces to a sum of strictly concave functions of each LD coefficient [see Additional file 1], i.e.

$$\begin{aligned}\xi^{\text{IBS}_{\text{hap}}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K}) &= \sum_{p=1}^{K}\left[-4\Delta_{\mathbf{p}}^{\mathbf{2}} + \Psi_{pq}^{\text{IBS}_{\text{hap}}}\boldsymbol{\Delta_p} + \Phi_{pq}^{\text{IBS}_{\text{hap}}}\right] \\ &= \sum_{p=1}^{K}\left[Q_p(\boldsymbol{\Delta_p})\right], \end{aligned} \quad (4)$$

where $\Psi_{pq}^{\text{IBS}_{\text{hap}}}$ and Φ_{pq} are sums and products of marginal frequencies that do not depend on LD coefficients. Let $\boldsymbol{\Delta_p^*} = \frac{\Psi_{pq}^{\text{IBS}_{\text{hap}}}}{8}$ be the critical value for each Q_p function. $\xi^{\text{IBS}_{\text{hap}}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$ will decrease if the squared or absolute deviation of each $\boldsymbol{\Delta_p}$ term from its corresponding $\boldsymbol{\Delta_p^*}$ increases [see Additional file 2]. However note that the squared deviations of all $\boldsymbol{\Delta_p}$ terms from their corresponding critical values do not need to increase simultaneously for $\xi^{\text{IBS}_{\text{hap}}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$ to decrease. For example, some Q_p functions corresponding to haplotypes with low frequencies can be negligible in expression (4). Hence, if $\sum_{p=1}^{K}\left(\boldsymbol{\Delta_p} - \boldsymbol{\Delta_p^*}\right)^2$ increases sufficiently, $\xi^{\text{IBS}_{\text{hap}}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$ will decrease. It can be shown that $\sum_{p=1}^{K}\left(\boldsymbol{\Delta_p} - \boldsymbol{\Delta_p^*}\right)^2$ will increase if $\sum_{p=1}^{K}\boldsymbol{\Delta_p}^2$ increases and that these two quantities share almost the same pattern for their expected values [see Additional file 2]. Thus, according to the $D_{i,QTL}^2$ profiles in Figure 2, $\xi^{\text{IBS}_{\text{hap}}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$ is expected to decrease as position i moves toward the QTL position.

An important result for $\xi^{\mathcal{P}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$ is obtained when only two haplotypes are observed among the K possible haplotypes. In this case, $\xi^{\mathcal{P}}(\boldsymbol{\Delta_1}, .., \boldsymbol{\Delta_K})$ reduces to a real function of one LD coefficient [see Additional file 1], i.e.:

$$\begin{aligned}\xi^{\mathcal{P}}(\boldsymbol{\Delta_1}) = &\left[-4s_{i,h_1,h_1}^{\mathcal{P}} - 4s_{i,h_2,h_2}^{\mathcal{P}} + 8s_{i,h_1,h_2}^{\mathcal{P}}\right]\boldsymbol{\Delta_1^2} \\ &+ \Psi^{\mathcal{P}}\boldsymbol{\Delta_1} + \Phi^{\mathcal{P}}, \end{aligned} \quad (5)$$

where $\Psi^{\mathcal{P}}$ and $\Phi^{\mathcal{P}}$ are terms independent of LD, and the minimum and maximum possible values for $\boldsymbol{\Delta_1}$ are given by $-\frac{1}{4}$ and $\frac{1}{4}$, respectively. If $\mathcal{P} = \text{IBS}_{\text{hap}}$ we have:

$$\xi^{\text{IBS}_{\text{hap}}}(\boldsymbol{\Delta_1}) = -8\boldsymbol{\Delta_1^2} + \Psi^{\text{IBS}_{\text{hap}}}\boldsymbol{\Delta_1} + \Phi^{\text{IBS}_{\text{hap}}} \quad (6)$$

The minimum and maximum possible values for the critical value, $\boldsymbol{\Delta_1^*}$, of $\xi^{\text{IBS}_{\text{hap}}}$ are given by $-\frac{1}{4}$ and $\frac{1}{4}$, respectively, if the tested locus and the QTL are monomorphic [see Additional file 1]. In other words, the

critical value of this function will always lie within the range of the LD coefficient when LD exist. In expression (5), the coefficient $\left[-4s^{\mathcal{P}}_{i,h_1,h_1} - 4s^{\mathcal{P}}_{i,h_2,h_2} + 8s^{\mathcal{P}}_{i,h_1,h_2}\right]$ is always greater or equal to -8 for any other predictor $\mathcal{P}$ than IBS_{hap}, since $s^{\mathcal{P}}_{i,h_1,h_2} \in [0,1]$. For instance, AIP by construction assign positive values to $s^{\mathcal{P}}_{i,h_1,h_2}$ when haplotypes h_1 and h_2 share allele similarity. This property is even truer if h_1 and h_2 are very similar. In such cases, the highest rate of decrease for $\xi^{\mathcal{P}}$, with respect to the absolute deviation of $\mathbf{\Delta_1}$ from $\mathbf{\Delta_1^*}$, is thus induced by $\mathcal{P} = \text{IBS}_{\text{hap}}$. Moreover, for such cases, we also have $\xi^{\mathcal{P}}\left(-\frac{1}{4}\right) = \xi^{\mathcal{P}}\left(\frac{1}{4}\right) \in \left[\frac{1}{2}s^{\mathcal{P}}_{i,h_1,h_2}, 1\right]$, which expresses a lower bound for $\xi^{\mathcal{P}}$ (i.e. $\frac{1}{2}s^{\mathcal{P}}_{i,h_1,h_2}$, [see Additional file 1]). Finally, $\xi^{\mathcal{P}}\left(-\frac{1}{4}\right) = \xi^{\mathcal{P}}\left(\frac{1}{4}\right) = 0$ if and only if $\mathcal{P} = \text{IBS}_{\text{hap}}$. In other words, when LD between the haplotypes and the QTL alleles is complete, the matrix distance is equal to 0 if and only if $\mathcal{P} = \text{IBS}_{\text{hap}}$ [see Additional file 1]. The decreasing behavior of $\xi^{\mathcal{P}}$ between a tested position and a QTL for a substantial increase of LD is therefore deteriorated for AIP with continuous predictions in $[0, 1]$. Hence, this result questions the behavior of AIP with continuous predictions in $[0, 1]$ in relation to LD, in the general case where K is greater than 2.

Distributions of matrix distance as function of multiallelic LD

Figures 3 and 4 show the distributions of the matrix distance for the six AIP against the local multiallelic LD. Figures 3 and 4 convey only local information for the case where the tested position is closest to the QTL, as opposed to Figure 2. Darker and lighter blue regions in Figures 3 and 4 correspond to higher and lower density of points. The red lines in Figures 3 and 4 correspond to non-parametric LOESS regressions of the matrix distance on the multiallelic LD.

Figures 3 and 4 show a better behavior of IBS_{hap} and P(IBD) for the decrease of their matrix distance, with lower variability around the LOESS curves compared to the other predictors, when the LD between the haplotypes and the target alleles increases. The distributions of the matrix distance for IBS_{hap} and P(IBD) in these figures show similar trends on the FLW and HCB chromosomes. This is due to the fact that these two predictors perform similarly in some conditions (see sub-section on mapping accuracy and relative efficiency). However IBS_{hap} shows a better behavior compared to all other predictors in Figures 3 and 4, for the decrease of its matrix distance with increasing multiallelic LD. The good behavior of IBS_{hap} for the decrease of its matrix distance in Figures 3 and 4 is totally explained by equation (4), where the sum of the concave polynomials decreases as the multiallelic LD increases. The better behavior of IBS_{hap}, compared to the other predictors in Figures 3 and 4, is explained by equations (3) and (5), which show that continuous predictions in [0,1] will deteriorate the decrease of the matrix distance with respect to LD.

The matrix distances for Beagle and IBS_{m} were also plotted against the local multiallelic LD between the haplotypes and the target alleles in Figures 3 and 4, although these two predictors are defined for marker positions only. Indeed, one of the aims of this study was to compare the AIP based on local LD between haplotypes and alleles at a hidden locus. TP, Score, Beagle and IBS_{m} showed poor relationships for the decrease of their matrix distance with the increasing multiallelic LD. The matrix distance distributions showed high variability for these predictors with respect to R on the FLW and HCB chromosomes. Note that the length of the six marker haplotypes on the HCB chromosomes were equal to 0.01 cM, on average, compared to 0.31 cM on average for those on the FLW chromosomes.

Mapping accuracy and relative efficiency

Table 1 relates the relative efficiency of the six AIP that were compared, and their mapping accuracies, for a QTL simulated under six scenarios on the FLW chromosomes for $N = 200$ simulations. $R_{i^*,QTL}$ in Table 1 corresponds to the multiallelic LD at position i^*, measured between the marker-haplotypes that harbor the simulated QTL and the QTL alleles. Note that the tested position i^* does not necessarily coincide with the QTL position. Thus, i^* can be defined as the tested position closest to the simulated QTL.

In Table 1, $\text{IBS}_{\text{m}}^{\text{QTL}}$ refers to the IBS_{m} predictor applied to the data set containing the causal variants. This situation was examined as a gold standard. As shown in Table 1 and as expected, $\text{IBS}_{\text{m}}^{\text{QTL}}$ provided the best mapping accuracy since the data set used contained the causal variants and both the simulated QTL and the analyzed markers were biallelic. However, it should be noted that the $\text{RMSE}^{\text{m.a.}}$ for $\text{IBS}_{\text{m}}^{\text{QTL}}$ was never equal to 0. This is principally due to the error term in the probabilistic models for hypothesis testing. $\text{RMSE}^{\text{r.e.}}$ for $\text{IBS}_{\text{m}}^{\text{QTL}}$ was also not equal to 0 when LD was highest ($R_{i^*,QTL} = 0.52$). This was due to a nearby marker which was in complete LD with the SNP that simulated the QTL (i.e. the biallelic LD was complete). Consequently the argument of the minimum (argmin) for the set of matrix distances was not unique.

In Table 1 both $\text{RMSE}^{\text{r.e.}}$ and $\text{RMSE}^{\text{m.a.}}$ increased globally for all predictors when LD decreased in the vicinity of the QTL. $\text{RMSE}^{\text{r.e.}}$ and $\text{RMSE}^{\text{m.a.}}$ were highly correlated, regardless of the QTL effect. Across all LD levels, the Spearman correlation coefficient between these two quantities was equal to 0.89 (or 0.91) when the QTL

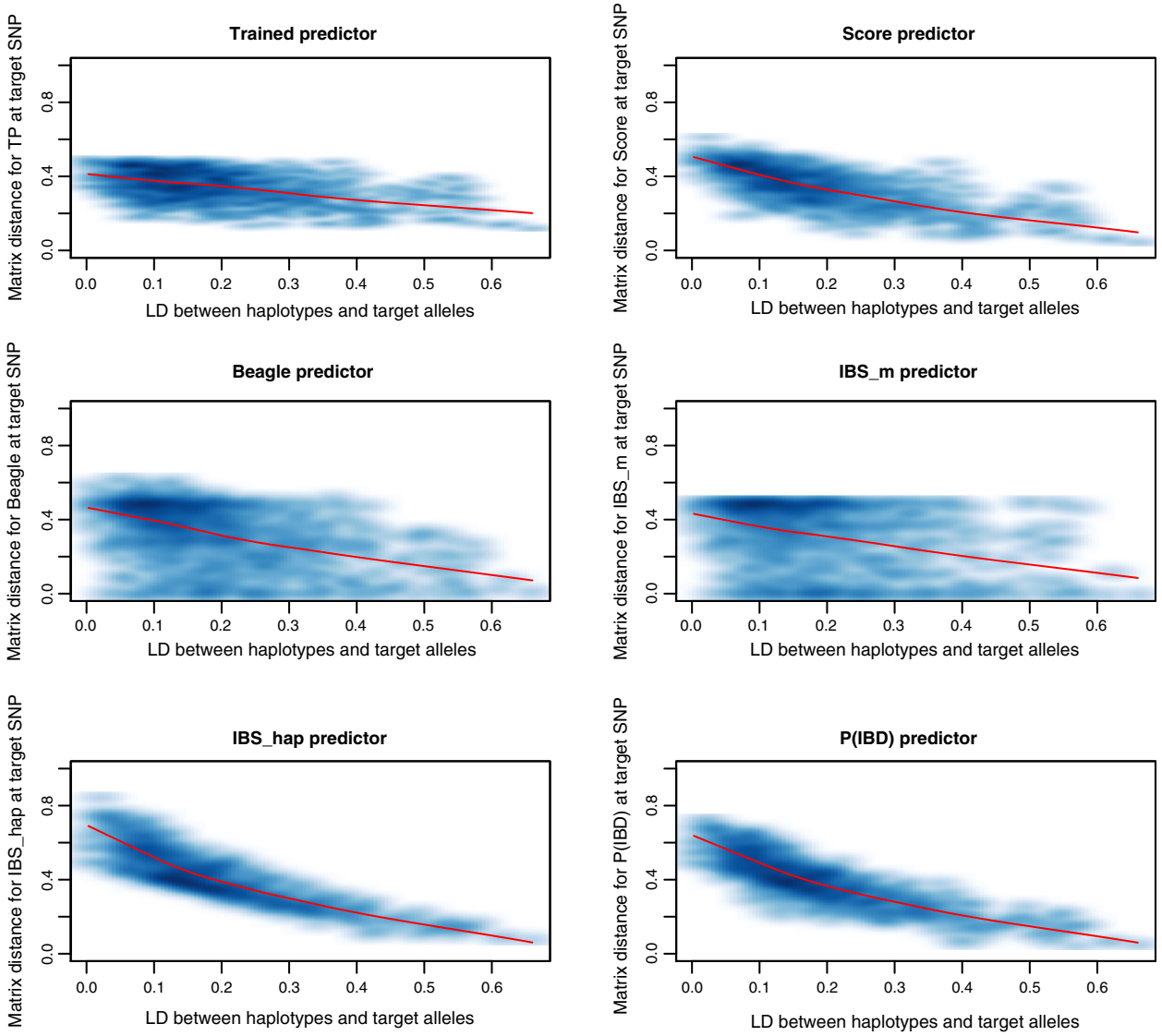


Figure 3 The matrix distance at 966 target SNPs on the 235 FLW chromosomes for the six compared AIP. Darker and lighter blue regions correspond to higher and lower density of points. The red lines correspond to LOESS regressions of the matrix distance on the multiallelic LD.

effect explained at most 57% (or 8%) of the total variance, respectively (Figure 5).

Each dot in Figure 5 represents $\text{RMSE}^{\text{m.a.}}$ against $\text{RMSE}^{\text{r.e.}}$ for one of the AIP at a particular LD level. In Table 1, the IBS_{hap} predictor was often the most accurate and efficient AIP when the data was analyzed without the QTL. However, the P(IBD) predictor showed similar mapping and efficiency results to IBS_{hap}. As defined by [1], the P(IBD) predictor relies on the IBS state of alleles between haplotype markers which suggests that IBS_{hap} and P(IBD) may perform similarly in some conditions [41]. Indeed, the distribution of IBD probabilities in the vicinity of a simulated QTL was almost bimodal (0 or 1) among the different pairs of chromosome segments for the different sets of simulations, and thus similar to the distribution of the values for IBS_{hap} between the segments. To illustrate this phenomenon, Figure 6 provides an example of distributions for the values of P(IBD) and IBS_{hap}, for one gene-drop simulation, between pairs of chromosome segments around the simulated QTL for the moderate LD situation ($R_{i^*,QTL} = 0.18$).

IBS_{hap} and P(IBD) also showed similar patterns at a set of tested positions for the matrix distances $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$. Figure 7 shows an example for the mean and the sample quantiles at 2.5 and 97.5% for $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ at each tested position for the six AIP, from 200 gene-drop simulations with a QTL simulated for the moderate LD situation ($R_{i^*,QTL} = 0.18$).

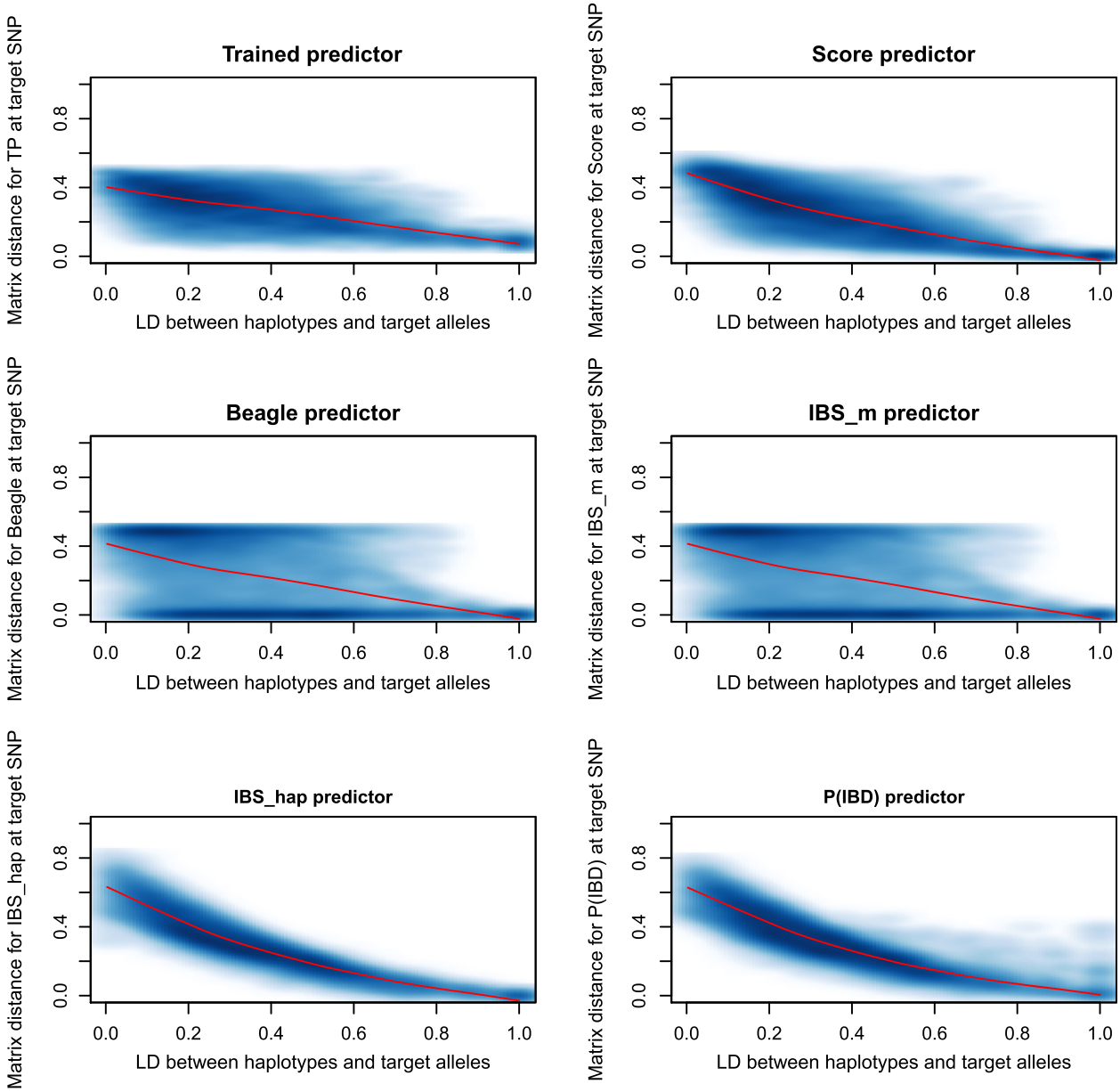


Figure 4 The matrix distance at 14973 target SNPs on the 90 HCB chromosomes for the six compared AIP. Darker and lighter blue regions correspond to higher and lower density of points. The red lines correspond to LOESS regressions of the matrix distance on the multiallelic LD.

As observed in Figure 7, the minima of the curves for the mean and the sample quantiles at 2.5 and 97.5% of the matrix distance distributions almost coincide with the QTL position for IBS_{hap} and P(IBD). For these two predictors, the three curves also show a smooth decreasing behavior as the tested position gets closer to the simulated QTL. This behavior shows the ability of IBS_{hap} and P(IBD) to capture LD structure along the chromosomes with respect to the simulated QTL, for different gene-drop simulations. It is interesting to note that IBS_{hap} and P(IBD) show similar patterns for the mean and the sample quantiles curves. However, the minimum of each of the three curves in Figure 7 is lower for IBS_{hap} than for P(IBD). Note that the patterns of the matrix distances for IBS_{hap} in Figure 7 are explained by equation (4) and Figure 2. That is, the matrix distance will decrease for IBS_{hap} due to the expected increase of the multiallelic LD, as the tested position moves toward the QTL position. In the same way, the patterns of the matrix distances for P(IBD) in Figure 7 are explained according to Figures 2 and 6. That is, P(IBD)

Table 1 Relative efficiencies and mapping accuracies for different QTL effects

	AIP			IBS_m^{QTL}	IBS_m	TP	Score	IBS_{hap}	P(IBD)	Beagle
$R_{i^*,QTL} = 0.52$	Relative efficiency		$RMSE^{r.e.}$	0.02	0.16	0.17	0.15	0.06	0.10	0.54
			$\hat{\mathbb{E}}^{r.e.}$	0.03	0.01	0.23	0.14	0.12	0.14	0.28
			$\hat{\sigma}^{r.e.}$	0.09	0.02	0.02	0.03	0.04	0.03	0.04
	Mapping	$\leq 57\%$	$RMSE^{m.a.}$	0.11	0.17	0.23	0.17	0.10	0.11	0.16
	accuracy	$\leq 8\%$	$RMSE^{m.a.}$	0.17	0.22	0.32	0.45	0.28	0.26	0.40
$R_{i^*,QTL} = 0.18$	Relative efficiency		$RMSE^{r.e.}$	0.00	0.18	0.46	0.21	0.14	0.14	0.40
			$\hat{\mathbb{E}}^{r.e.}$	0.00	0.18	0.39	0.35	0.31	0.34	0.31
			$\hat{\sigma}^{r.e.}$	0.00	0.06	0.02	0.03	0.05	0.04	0.06
	Mapping	$\leq 57\%$	$RMSE^{m.a.}$	0.06	0.29	0.27	0.33	0.16	0.21	0.28
	accuracy	$\leq 8\%$	$RMSE^{m.a.}$	0.10	0.34	0.36	0.46	0.29	0.30	0.31
$R_{i^*,QTL} = 0.08$	Relative efficiency		$RMSE^{r.e.}$	0.00	0.76	1.00	1.00	1.04	1.04	0.72
			$\hat{\mathbb{E}}^{r.e.}$	0.00	0.24	0.35	0.33	0.31	0.37	0.34
			$\hat{\sigma}^{r.e.}$	0.00	0.06	0.04	0.05	0.06	0.05	0.08
	Mapping	$\leq 57\%$	$RMSE^{m.a.}$	0.13	0.66	0.58	0.54	0.51	0.58	0.55
	accuracy	$\leq 8\%$	$RMSE^{m.a.}$	0.18	0.71	0.66	0.71	0.62	0.69	0.59

- $R_{i^*,QTL}$: Multiallelic measure of LD between the simulated QTL and the haplotypes harboring it.
- $RMSE^{r.e.}$: Root mean square error of $\theta^{\mathcal{P}}_{r.e.}$ with respect to θ_{QTL} (cM).
- $\hat{\mathbb{E}}^{r.e.}$: Expected value of the matrix distance at $\theta^{\mathcal{P}}_{r.e.}$.
- $\hat{\sigma}^{r.e.}$: Standard error of the matrix distance at $\theta^{\mathcal{P}}_{r.e.}$.
- $RMSE^{m.a.}$: Root mean square error of $\theta^{\mathcal{P}}_{m.a.}$ with respect to θ_{QTL} (cM).

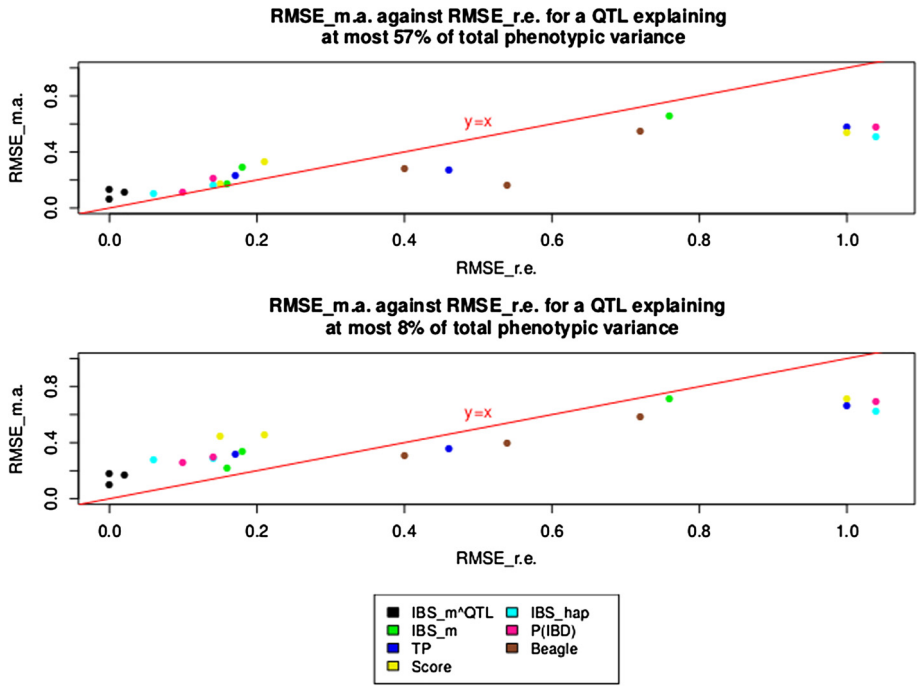


Figure 5 $RMSE^{m.a.}$ **against** $RMSE^{r.e.}$ **for the compared AIP according to different QTL effects across all LD levels.**

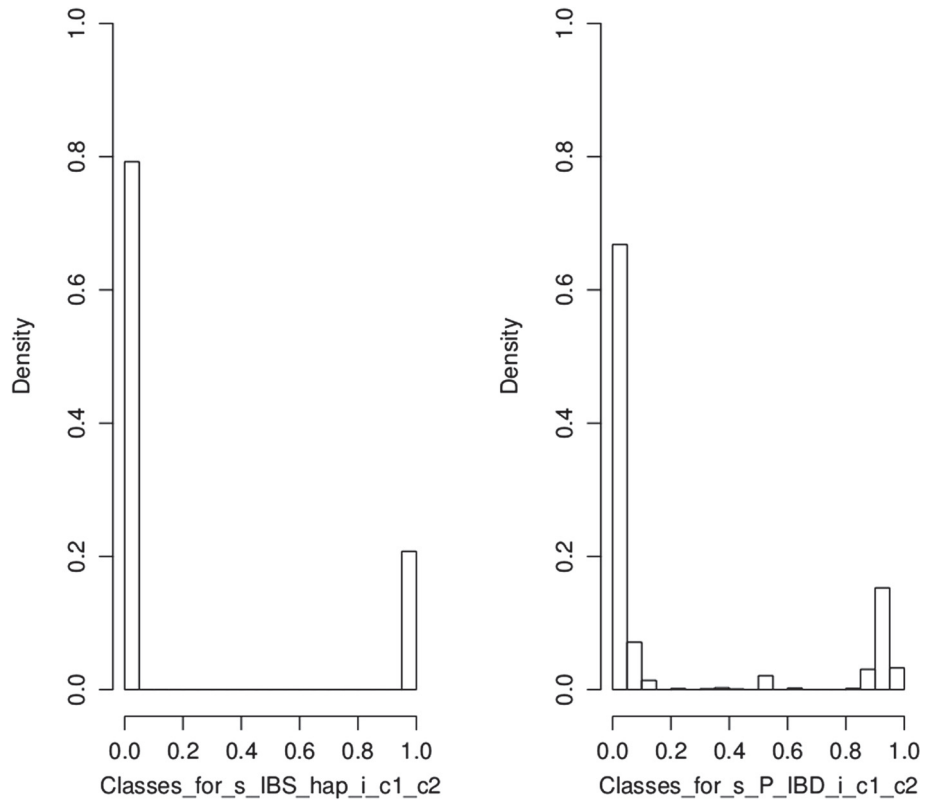


Figure 6 Distribution of values for IBS_{hap} and P(IBD) between chromosome segments around the simulated QTL for the moderate LD situation ($R_{i^*,QTL} = 0.18$), example for one simulation. The class width for the IBD probabilities is equal to 0.05.

will behave slightly differently from IBS_{hap}, according to Figures 2 and 6, when taking equations (3) and (5) into account.

As shown in Figure 7, the other predictors cannot capture the LD structure along the chromosomes with respect to the simulated QTL as well as IBS_{hap} and P(IBD); this is particularly the case for Score and even more for TP. For the latter two predictors, $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ shows little variability and is low on average across the tested positions. This could explain the lack of a clear ranking between the mapping accuracies of TP and Score in Table 1. For Beagle, a good relative efficiency and mapping accuracy was observed for the lowest LD situation ($R_{i^*,QTL} = 0.08$) in Table 1, compared to all the other predictors, when the QTL effect was low. Note that AIP that are based on haplotypes and that do not perform haplotype clustering like Beagle, may not be at an advantage for a low LD situation. For example, the matrix distance for IBS_{hap}, as defined by equation (4), will not decrease if there is little LD between local haplotypes and QTL alleles. Therefore, haplotype clustering is necessary for such situations. Moreover, these AIP will intrinsically provide an excess of degrees of freedom for testing association if the QTL is biallelic, while not compensating for the low LD captured in the matrix distance. Hence, AIP based on haplotype clustering can provide higher mapping accuracy for low LD situations.

Discussion

Matrix distance properties

The present study showed that the QTL mapping accuracy of AIP is highly correlated to the tested position that minimizes the matrix distance defined for comparison. The use of the matrix distance to compare various AIP has many advantages for methodology development and validation. First, it is independent of phenotype simulation processes and statistical tests that are commonly used to compare QTL mapping accuracy of different AIP [4,8,23,25]. Indeed the phenotype simulation process, when based on certain specific assumptions, may favor some AIP over others: for example, IBD-based AIP might be at an advantage if the phenotypes are simulated only according to population history. The statistical test used may also favor some AIP, such as IBS_{hap}, IBS_m and Beagle, over others due to numerical stability when estimating variance components. As such, solving mixed model equations when covariance matrices are close to singularity due to AIP computation has been reported as an

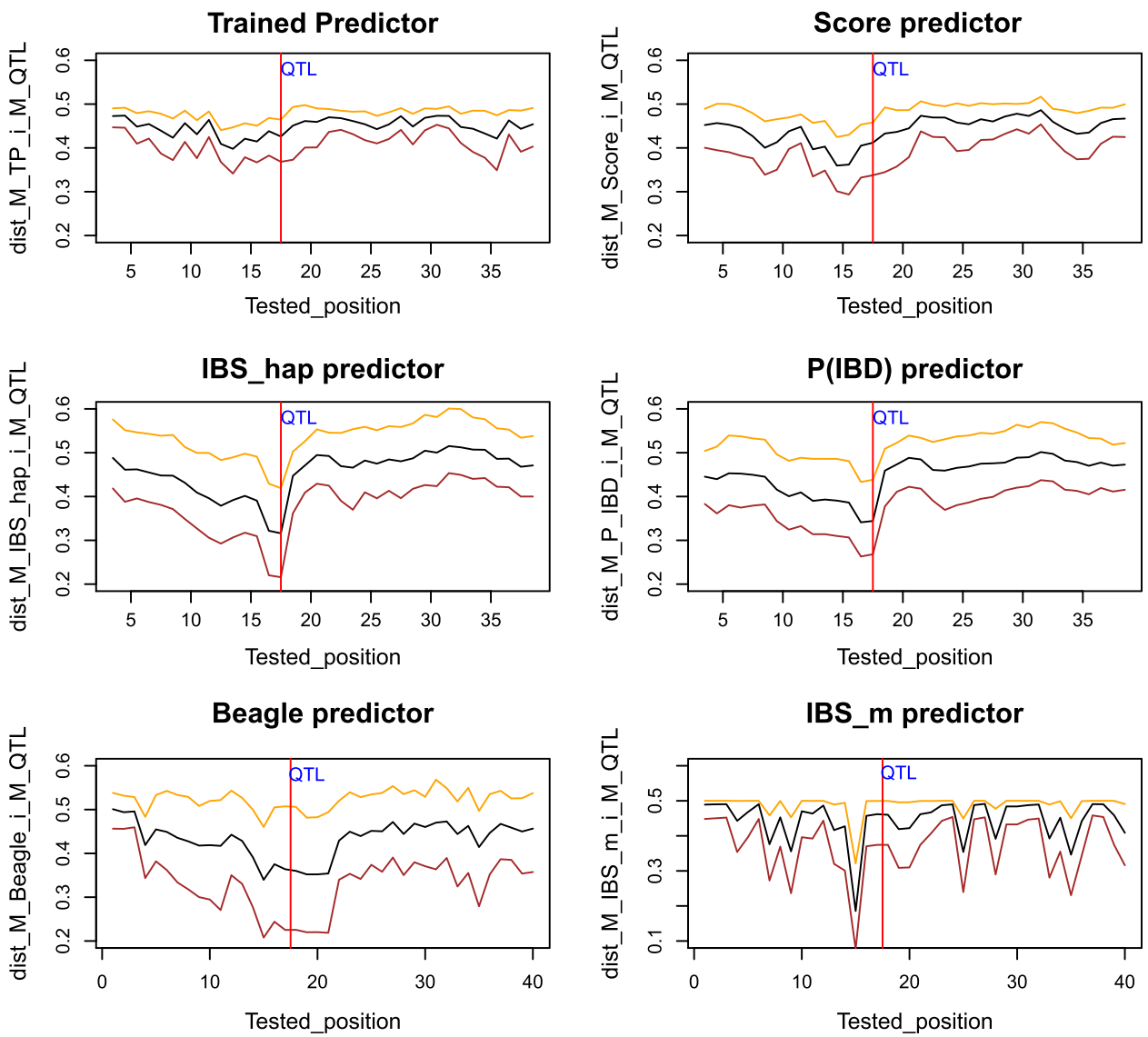


Figure 7 The mean and sample quantiles at 2.5% and 97.5% of $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ (black, brown and orange respectively) for TP, Score, IBS_{hap}, P(IBD), Beagle and IBS_m, with respect to the tested position and the simulated QTL position (red line) for the moderate LD situation ($R_{i^*,QTL} = 0.18$).

issue, and clustering strategies for haplotypes, which actually modify the properties of the AIP matrices, have been proposed to facilitate computation [42,43]. The major drawback of the matrix distance approach is related to this advantage: a particularly efficient AIP or a particularly efficient haplotype size, identified from the matrix distance, that can not be used in association studies would be of no value. Another advantage of the matrix distance approach is that computation time is highly reduced compared to association studies, so numerous comparisons can be done. In the present study, the relative efficiency of the AIP was consistent with the results for QTL mapping accuracy, regardless of the QTL effects and LD patterns. Therefore, the concept of relative efficiency was proven useful to compare AIP and avoid time-consuming association studies on simulated data. Combining the relative efficiency with the mapping accuracy of predictors could also be helpful to gain a better understanding of the underlying mechanisms in an association study.

Comparing AIP

The results showed that the most accurate AIP for mapping were those that best captured LD between a tested position and a QTL. This was proposed from algebraic developments in the simplest situations and validated

using real data and simulations. The matrix distance can be written for any AIP as a sum of functions of LD coefficients, and more precisely for the IBS_{hap} predictor as a sum of concave polynomials of LD coefficients. When LD was moderate to high around the QTL, the IBS_{hap} predictor was the most efficient and accurate matrix for mapping. For a biallelic QTL, the domains of values for which some of these concave polynomials can either decrease or increase with increasing LD was shown in our developments as limited to extreme allele frequencies for the haplotypes and QTL. Additionally, continuous AIP in [0,1] were shown to deteriorate the matrix distance generally when LD between a tested position and the QTL increased. This was observed on two unrelated data sets, which showed that this behavior is not related to the marker density or population history. All LD measures are based on counting occurrences for discrete events at distinct loci to quantify non-random association [37,38], which thus explains the algebraic and simulation results for discrete and continuous AIP when a relatively high LD is available for detection. The pig example was built using 235 haplotypes and 25 gene-drop generations, a realistic situation with regard to the effective population size. However, the impact of the resulting long-range haplotypic identity, which depends strongly on the population size and mating strategies, on the relative values of the considered AIP should be investigated.

Despite using two contrasting data sets in terms of marker density and population history, P(IBD) always behaved very similarly to IBS_{hap}. When extending the calculations to longer haplotypes (results not shown), a similar behavior was observed. Yet advantages have been reported for P(IBD) compared to IBS_{hap} in some situations. For example, Roldan *et al.* [43] showed better accuracy for P(IBD) compared to IBS_{hap}, after a clustering step for haplotypes when marker intervals were equal to 0.05 cM between SNPs, but not when they reached 0.25 cM. However in Roldan *et al.* [43], different statistical models were applied to P(IBD) versus IBS_{hap} (mixed model *versus* fixed effects model respectively). Hence, these two AIP were not compared on the same basis. For instance, Boleckova *et al.* [44] showed that statistical models in which haplotypes were fitted as random effects performed better than those in which they were fitted as fixed effects. When both LD and the QTL effect were low, Beagle showed a relatively good efficiency and mapping accuracy. It was not possible to derive algebraical comparisons between AIP when LD was low, but this, together with earlier studies that point out that continuous advanced methods are more efficient than simple IBS_{hap}, suggests that some continuous AIP in [0,1] may provide efficiency when LD between markers and a QTL is reduced.

Extending the results to multiallelic QTL

In the present study, we considered a biallelic QTL for algebra and simulations. Yet the algebraic derivation of the matrix distance can be generalized to a multiallelic QTL without difficulty [see Additional file 1]. As suggested by these developments, for a multiallelic QTL, the relationship between continuous predictions of allelic identity at a tested position and the corresponding LD coefficients will tend to be looser than for discrete predictions. In addition, the matrix distance for the IBS_{hap} predictor can always be written as a sum of concave polynomials of LD coefficients for any degree of allelism at the QTL.

Conclusion

The IBS_{hap} predictor can always capture multiallelic LD between a tested position and a QTL, regardless of the degree of allelism at the QTL. The IBS_{hap} predictor also has the advantage of being simple, fast and numerically stable when used in association studies. Therefore, it is suggested that, for studies with a high density of markers and for which LD between markers and the causal variants is likely to be high, the use of the IBS_{hap} predictor is recommended.

Additional files

Additional file 1: Algebraic derivations of formulas in the main text. This file contains all the algebraic derivations for expressions (1) to (6) and the generalization of the matrix distance, as a sum of concave functions of LD coefficients when $\boldsymbol{\mathcal{P}} = IBS_{hap}$, for the case of a multiallelic QTL.

Additional file 2: Domains of LD coefficients and boundary conditions for the critical values of each Q_p function. This file contains the domain of values for the multiallelic LD coefficients, the boundary conditions for the critical value of each Q_p function in expression (4) and the relation between the sum of the squared deviations and $D^2_{i,QTL}$.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
LJ derived the analytical results, performed the simulations and wrote the manuscript. LJ, JME and HG were involved in the conception of the study. All authors read and approved the final manuscript.

Acknowledgements
The study was supported by the French National Research Agency (ANR-09-GENM-002 Rules & Tools Project).

Author details
[1]INRA, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), F-31326, Castanet-Tolosan, France. [2]Université de Toulouse, INP, ENSAT, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), F-31326, Castanet-Tolosan, France. [3]Université de Toulouse, INP, ENVT, GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), F-31076, Toulouse, France.

Received: 20 November 2013 Accepted: 20 May 2014
Published: 14 July 2014

References

1. Meuwissen THE, Goddard ME: **Prediction of identity by descent probabilities from marker haplotypes.** *Genet Sel Evol* 2001, **33:**605–634.

2. Li J, Jiang T: **Haplotype-based linkage disequilibrium mapping via direct data mining.** *Bioinformatics* 2005, **21:**4384–4393.
3. Browning SR: **Multilocus association mapping using variable-length Markov chains.** *Am J Hum Genet* 2006, **78:**903–913.
4. Pong-Wong R, George AW, Woolliams JA, Haley CS: **A simple and rapid method for calculating identity-by-descent matrices using multiple markers.** *Genet Sel Evol* 2001, **33:**453–471.
5. Bercovici S, Meek C, Wexler Y, Geiger D: **Estimating genome-wide IBD sharing from SNP data via an efficient hidden Markov model of LD with application to gene mapping.** *Bioinformatics* 2010, **26:**i175–i182.
6. Druet T, Georges M: **A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping.** *Genetics* 2010, **184:**789–798.
7. Akey J, Jin L: **Xiong M: Haplotypes vs single marker linkage disequilibrium tests: what do we gain?** *Eur J Hum Genet* 2001, **9:**291–300.
8. Abdallah J, Goffinet B, Cierco-Ayrolles C, Pérez-Enciso M: **Linkage disequilibrium fine mapping of quantitative trait loci: a simulation study.** *Genet Sel Evol* 2003, **35:**513–532.
9. Cardon LR, Abecasis GR: **Using haplotype blocks to map human complex trait loci.** *Trends Genet* 2003, **19:**136–140.
10. Clark AG: **The role of haplotypes in candidate gene studies.** *Genet Epidemiol* 2004, **27:**321–333.
11. Schaid DJ: **Evaluating associations of haplotypes with traits.** *Genet Epidemiol* 2004, **27:**348–364.
12. Browning BL, Browning SR: **Efficient multilocus association testing for whole genome association studies using localized haplotype clustering.** *Genet Epidemiol* 2007, **31:**365–375.
13. Chen Y, Li X, Li J: **A novel approach for haplotype-based association analyzis using family data.** *BMC Bioinformatics* 2010, **11:**S45.
14. Lin WY, Yi N, Zhi D, Zhang K, Gao G, Tiwari HK, Liu N: **Haplotype-based methods for detecting uncommon causal variants with common SNPs.** *Genet Epidemiol* 2012, **36:**572–582.
15. Knüppel S, Esparza-Gordillo J, Marenholz I, Holzhütter HG, Bauerfeind A, Ruether A, Weidinger S, Lee YA, Rohde K: **Multi-locus stepwise regression: a haplotype based algorithm for finding genetic associations applied to atopic dermatitis.** *BMC Med Genet* 2012, **13:**8.
16. Li M, Wing HW, Art BO: **A sparse transmission disequilibrium test for haplotypes based on Bradley-Terry graphs.** *Hum Hered* 2012, **73:**52–61.
17. Risch N, Merikangas K: **The future of genetic studies of complex human diseases.** *Science* 1996, **273:**1516–1517.
18. Terwilliger JD, Weiss KM: **Linkage disequilibrium mapping of complex disease: fantasy or reality?** *Curr Opin Biotechnol* 1998, **9:**578–594.
19. Jorde LB: **Linkage disequilibrium and the search for complex disease genes.** *Genome Res* 2000, **10:**1435–1444.
20. Weiss KM, Clark AG: **Linkage disequilibrium and the mapping of complex human traits.** *Trends in Genet* 2002, **18:**19–24.
21. Slatkin M: **Disequilibrium mapping of a quantitative-trait locus in an expanding population.** *Am J Hum Genet* 1999, **64:**1764–1772.
22. Farnir F, Coppieters W, Arranz JJ, Berzi P, Cambisano N, Grisart B, Karim L, Marcq F, Moreau L, Mni M, Nezer C, Simon P, Vanmanshoven P, Wagenaar D, George M: **Extensive genome-wide linkage disequilibrium in cattle.** *Genome Res* 2000, **10:**220–227.
23. Meuwissen THE, Goddard ME: **Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci.** *Genetics* 2000, **155:**421–430.
24. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES: **Structure of linkage disequilibrium and phenotypic associations in the maize genome.** *Proc Natl Acad Sci USA* 2001, **98:**11479–11484.
25. He W, Fernando RL, Dekkers JCM, Gilbert H: **A gene frequency model for QTL mapping using Bayesian inference.** *Genet Sel Evol* 2010, **42:**21.
26. Grapes L, Dekkers JCM, Rothschild MF, Fernando RL: **Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci.** *Genetics* 2004, **166:**1561–1570.
27. Browning BL, Browning SR: **Haplotypic analysis of Wellcome Trust Case Control Consortium data.** *Human Genet* 2008, **123:**273–280.
28. Henderson CR: **Best linear unbiased estimation and prediction under a selection model.** *Biometrics* 1975, **31:**423–447.
29. Henderson CR: **A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values.** *Biometrics* 1976, **32:**69–83.
30. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from in-complete data via the EM algorithm.** *Roy Statist Soc Ser B* 1977, **39:**1–38.
31. Patterson HD, Thompson R: **Recovery of inter-block information when block sizes are unequal.** *Biometrika* 1971, **58:**545–554.
32. Harville DA: **Bayesian inference for variance components using only error contrasts.** *Biometrika* 1974, **61:**383–385.
33. Foulley JL: **EM algorithm: theory and application to the mixed model.** *J Soc Fr Stat* 2002, **143:**57–109.
34. Calus MPL, Meuwissen THE, Windig JJ, Knol EF, Schrooten C, Vereijken ALJ, Veerkamp RF: **Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values.** *Genet Sel Evol* 2009, **41:**11.
35. Grapes L, Firat MZ, Dekkers JCM, Rothschild MF, Fernando RL: **Optimal haplotype structure for linkage disequilibrium-based fine mapping of quantitative trait loci using identity by descent.** *Genetics* 2005, **172:**1955–1965.
36. Ramos AM, Crooijmans RP, Affara NA, Amaral AJ, Archibald AL, Beever JE, Bendixen C, Churcher C, Clark R, Dehais P, Hansen MS, Hedegaard J, Hu ZL, Kerstens HH, Law AS, Megens HJ, Milan D, Nonneman DJ, Rohrer GA, Rothschild MF, Smith TP, Schnabel RD, Van Tassell CP, Taylor JF, Wiedmann RT, Schook LB, Groenen MA: **Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology.** *PLoS ONE* 2009, **4:**e6524.
37. Hedrick PW, Thomson G: **A two-locus neutrality test: applications to humans, E. coli and Lodgepole pine.** *Genetics* 1985, **112:**135–156.
38. Hedrick PW: **Gametic disequilibrium measures: proceed with caution.** *Genetics* 1987, **117:**331–341.
39. Maurer HP, Knaak C, Melchinger AE, Ouzunova M, Frisch M: **Linkage disequilibrium between SSR markers in six pools of elite lines of an european breeding program for hybrid maize.** *Maydica* 2006, **51:**269–279.
40. Ytournel F, Teyssèdre S, Roldan D, Erbe M, Simianer H, Boichard D, Gilbert H, Druet T, Legarra A: **LDSO: A program to simulate pedigrees and molecular information under various evolutionary forces.** *J Anim Breed Genet* 2012, **129:**417–421.
41. Ytournel F, Gilbert H, Boichard D: **Concordance between IBD probabilities and linkage disequilibrium.** In *Proceedings of European Federation of Animal Science Annual Meeting; 26 August 2007; Dublin*; 2007:1248. [http://www.eaap.org/Previous_Annual_Meetings/2007Dublin/Papers/S38_1248_Ytournel.pdf]
42. Druet T, Fritz S, Boussaha M, Ben-Jemaa S, Guillaume F, Derbala D, Zelenika D, Lechner D, Charon C, Boichard D, Gut I, Eggen A, Gautier M: **Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map.** *Genetics* 2008, **178:**2227–2235.
43. Roldan DL, Gilbert H, Henshall JM, Legarra A, Elsen JM: **Fine-mapping quantitative trait loci with a medium density marker panel: efficiency of population structures and comparison of linkage disequilibrium linkage analysis models.** *Genet Res Camb* 2012, **94:**223–234.
44. Boleckova J, Christensen OF, Sørensen P, Sahana G: **Strategies for haplotype-based association mapping in a complex pedigreed population.** *Czech J Anim Sci* 2012, **1:**1–9.

doi:10.1186/1297-9686-46-45
Cite this article as: Jacquin *et al.*: Using haplotypes for the prediction of allelic identity to fine-map QTL: characterization and properties. *Genetics Selection Evolution* 2014 **46**:45.

Additional file 1

Details of the algebraic derivations of the formulas in the main text

Algebraic derivation of expression (1)

Let $K = 2^t$ be the number of possible haplotypes, at locus i, for a sliding window of t markers. Let $\mathbb{1}_{\{u_{QTL,c_1,c_2}=1\}}$ and $\mathbb{1}_{\{u_{QTL,c_1,c_2}=0\}}$ be the following indicator functions:

$$\mathbb{1}_{\{u_{QTL,c_1,c_2}=1\}} = \begin{cases} 1 \text{ if } (c_1, c_2) \text{ have identical alleles at the QTL} \\ 0 \text{ else} \end{cases}$$

$$\mathbb{1}_{\{u_{QTL,c_1,c_2}=0\}} = \begin{cases} 1 \text{ if } (c_1, c_2) \text{ have non-identical alleles at the QTL} \\ 0 \text{ else} \end{cases}$$

We have:

$$\begin{aligned} d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}) &= \frac{1}{4n^2} \sum_{c_1=1}^{2n} \sum_{c_2=1}^{2n} |s^{\mathcal{P}}_{i,c_1,c_2} - u_{QTL,c_1,c_2}| \\ &= \frac{1}{4n^2} \sum_{c_1=1}^{2n} \sum_{c_2=1}^{2n} \mathbb{1}_{\{u_{QTL,c_1,c_2}=1\}} |s^{\mathcal{P}}_{i,c_1,c_2} - 1| + \frac{1}{4n^2} \sum_{c_1=1}^{2n} \sum_{c_2=1}^{2n} \mathbb{1}_{\{u_{QTL,c_1,c_2}=0\}} |s^{\mathcal{P}}_{i,c_1,c_2} - 0| \end{aligned}$$

Let E_{h_p} be the set of chromosome segments carrying haplotype h_p ($p \in \{1, .., K\}$) at locus i. We have:

$$\begin{aligned} d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}) =& \frac{1}{4n^2} \sum_{c_1=1}^{2n} \sum_{c_2=1}^{2n} \sum_{p=1}^{K} \mathbb{1}_{\{c_1 \in E_{h_p}\}} \sum_{q=1}^{K} \mathbb{1}_{\{c_2 \in E_{h_q}\}} \mathbb{1}_{\{u_{QTL,c_1,c_2}=1\}} |s^{\mathcal{P}}_{i,c_1,c_2} - 1| \\ &+ \frac{1}{4n^2} \sum_{c_1=1}^{2n} \sum_{c_2=1}^{2n} \sum_{p=1}^{K} \mathbb{1}_{\{c_1 \in E_{h_p}\}} \sum_{q=1}^{K} \mathbb{1}_{\{c_2 \in E_{h_q}\}} \mathbb{1}_{\{u_{QTL,c_1,c_2}=0\}} |s^{\mathcal{P}}_{i,c_1,c_2} - 0| \end{aligned}$$

where $\mathbb{1}_{\{c_1 \in E_{h_p}\}}$ and $\mathbb{1}_{\{c_2 \in E_{h_q}\}}$ are the indicator functions of the events $\{c_1 \in E_{h_p}\}$ and $\{c_2 \in E_{h_q}\}$ respectively. Indeed $\sum_{p=1}^{K} \mathbb{1}_{\{c_1 \in E_{h_p}\}}$ and $\sum_{q=1}^{K} \mathbb{1}_{\{c_2 \in E_{h_q}\}}$ are always equal to one. $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ can thus be expressed as:

$$\begin{aligned} d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}) = & \sum_{p=1}^{K}\sum_{q=1}^{K} \frac{1}{4n^2} \sum_{c_1=1}^{2n}\sum_{c_2=1}^{2n} \mathbb{1}_{\left\{u_{QTL,c_1 \in E_{h_p}, c_2 \in E_{h_q}}=1\right\}} |s^{\mathcal{P}}_{i,h_p,h_q} - 1| \\ & + \sum_{p=1}^{K}\sum_{q=1}^{K} \frac{1}{4n^2} \sum_{c_1=1}^{2n}\sum_{c_2=1}^{2n} \mathbb{1}_{\left\{u_{QTL,c_1 \in E_{h_p}, c_2 \in E_{h_q}}=0\right\}} |s^{\mathcal{P}}_{i,h_p,h_q} - 0| \end{aligned}$$

where $\frac{1}{4n^2} \sum_{c_1=1}^{2n}\sum_{c_2=1}^{2n} \mathbb{1}_{\left\{u_{QTL,c_1 \in E_{h_p}, c_2 \in E_{h_q}}=1\right\}} = f(c_1 \in E_{h_p}, c_2 \in E_{h_q}, \text{identical alleles at the QTL})$ is the frequency of chromosome segments, at locus i, carrying h_p and h_q and having identical alleles at the QTL. Since $\{c_1 \in E_{h_p}\}$ and $\{c_2 \in E_{h_q}\}$ are independent events we have:

$$f(c_1 \in E_{h_p}, c_2 \in E_{h_q}, \text{identical alleles at the QTL}) = p^{QTL}_{i,h_p,h_q} f_{i,h_p} f_{i,h_q}$$

where p^{QTL}_{i,h_p,h_q} is the proportion of identical alleles shared at the QTL by the couples of chromosomes carrying h_p and h_q at position i. f_{i,h_p} and f_{i,h_q} are the frequencies of haplotypes h_p and h_q at position i respectively.

Similarly we have:

$$\begin{aligned} \frac{1}{4n^2} \sum_{c_1=1}^{2n}\sum_{c_2=1}^{2n} \mathbb{1}_{\left\{u_{QTL,c_1 \in E_{h_p}, c_2 \in E_{h_q}}=0\right\}} &= f(c_1 \in E_{h_p}, c_2 \in E_{h_q}, \text{non-identical alleles at the QTL}) \\ &= (1 - p^{QTL}_{i,h_p,h_q}) f_{i,h_p} f_{i,h_q} \end{aligned}$$

Consequently $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ can be written as (1), i.e.

$$d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}) = \sum_{p=1}^{K}\sum_{q=1}^{K} f_{i,h_p} f_{i,h_q} \left[p^{QTL}_{i,h_p,h_q}(1 - s^{\mathcal{P}}_{i,h_p,h_q}) + (1 - p^{QTL}_{i,h_p,h_q}) s^{\mathcal{P}}_{i,h_p,h_q} \right] \quad (1)$$

Algebraic derivation of expression (2)

Let $(n_{h_p})_{1\leq p\leq K}$ be the counts of the possible haplotypes at a tested position i. And let $(n_{h_p a_l})_{\substack{1\leq p\leq K \\ 1\leq l\leq 2}}$ be the counts of the $2K$ possible haplotypes defined between i and a QTL. Expression (1) can be re-written as:

$$\begin{aligned} d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}) &= \sum_{p=1}^{K}\sum_{q=1}^{K} \frac{n_{h_p}}{2n}\frac{n_{h_q}}{2n} \left[\frac{[n_{h_p a_1} n_{h_q a_1} + n_{h_p a_2} n_{h_q a_2}]}{n_{h_p} n_{h_q}} (1 - s^{\mathcal{P}}_{i,h_p,h_q}) \right. \\ &\qquad \left. + \left(\frac{n_{h_p} n_{h_q} - [n_{h_p a_1} n_{h_q a_1} + n_{h_p a_2} n_{h_q a_2}]}{n_{h_p} n_{h_q}} \right) s^{\mathcal{P}}_{i,h_p,h_q} \right] \\ &= \sum_{p=1}^{K}\sum_{q=1}^{K} \left[\left[f^{QTL}_{i,h_p a_1} f^{QTL}_{i,h_q a_1} + f^{QTL}_{i,h_p a_2} f^{QTL}_{i,h_q a_2} \right] (1 - s^{\mathcal{P}}_{i,h_p,h_q}) \right. \\ &\qquad \left. + \left(\frac{(n_{h_p a_1} + n_{h_p a_2})(n_{h_q a_1} + n_{h_q a_2})}{4n^2} - f^{QTL}_{i,h_p a_1} f^{QTL}_{i,h_q a_1} - f^{QTL}_{i,h_p a_2} f^{QTL}_{i,h_q a_2} \right) s^{\mathcal{P}}_{i,h_p,h_q} \right] \end{aligned}$$

which simplifies to (2):

$$\begin{aligned} d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}) = \sum_{p=1}^{K}\sum_{q=1}^{K} \Bigg[& \left[f^{QTL}_{i,h_p a_1} f^{QTL}_{i,h_q a_1} + f^{QTL}_{i,h_p a_2} f^{QTL}_{i,h_q a_2} \right] (1 - s^{\mathcal{P}}_{i,h_p,h_q}) \\ & + \left[f^{QTL}_{i,h_p a_2} f^{QTL}_{i,h_q a_1} + f^{QTL}_{i,h_p a_1} f^{QTL}_{i,h_q a_2} \right] s^{\mathcal{P}}_{i,h_p,h_q} \Bigg] \quad (2) \end{aligned}$$

Algebraic derivation of expression (3)

Let $f^{QTL}_{i,h_p a_1} = f_{i,h_p} f_{a_1} + \mathbf{\Delta_p} = \alpha_p + \mathbf{\Delta_p}$ and $f^{QTL}_{i,h_p a_2} = f_{i,h_p} f_{a_2} - \mathbf{\Delta_p} = \tilde{\alpha}_p - \mathbf{\Delta_p}$. Note that we have $\sum_{p=1}^{K} \mathbf{\Delta_p} = \sum_{p=1}^{K} (f^{QTL}_{i,h_p a_1} - f_{i,h_p} f_{a_1}) = f_{a_1} - f_{a_1} \sum_{p=1}^{K} f_{i,h_p} = 0$. Replacing the haplotype frequencies in

expression (2) with the expressions including the α_p, $\tilde{\alpha}_p$ and $\mathbf{\Delta_p}$ terms (same for the frequencies depending on the $\alpha_q, \tilde{\alpha}_q$ and $\mathbf{\Delta_q}$ terms) gives:

$$
\begin{aligned}
d(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}) = \sum_{p=1}^{k} \Bigg[& -4 s^{\mathcal{P}}_{i,h_p,h_p} \mathbf{\Delta_p}^2 + \Big[\alpha_p - \tilde{\alpha}_p + 4 s^{\mathcal{P}}_{i,h_p,h_p} (\tilde{\alpha}_p - \alpha_p) + \sum_{q \neq p}^{k} (\alpha_q - \tilde{\alpha}_q) \Big] \mathbf{\Delta_p} \\
& + \alpha_p^2 + \tilde{\alpha}_p^2 + s^{\mathcal{P}}_{i,h_p,h_p} (-\alpha_p^2 - \tilde{\alpha}_p^2 + 2\alpha_p \tilde{\alpha}_p) + \sum_{q \neq p}^{k} \alpha_p \alpha_q + \tilde{\alpha}_p \tilde{\alpha}_q \\
& + \sum_{q \neq p}^{k} s^{\mathcal{P}}_{i,h_p,h_q} \Big(-4 \mathbf{\Delta_p} \mathbf{\Delta_q} + 2(\tilde{\alpha}_p - \alpha_p) \mathbf{\Delta_q} + 2(\tilde{\alpha}_q - \alpha_q) \mathbf{\Delta_p} \\
& - \alpha_p \alpha_q - \tilde{\alpha}_p \tilde{\alpha}_q + \tilde{\alpha}_p \alpha_q + \alpha_p \tilde{\alpha}_q \Big) \Bigg] \quad (*)
\end{aligned}
$$

Replacing $\mathbf{\Delta_q}$ with $-\mathbf{\Delta_p} - \sum_{l \neq p,q}^{K} \mathbf{\Delta_l}$ $\Big($ since $\sum_{p=1}^{k} \mathbf{\Delta_p} = 0 \Big)$ in (*) finally gives:

$$
\begin{aligned}
d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}) = \sum_{p=1}^{K} \Bigg[& 4 \Big(\sum_{q \neq p}^{K} s^{\mathcal{P}}_{i,h_p,h_q} - s^{\mathcal{P}}_{i,h_p,h_p} \Big) \mathbf{\Delta_p^2} + \Big[\alpha_p - \tilde{\alpha}_p + 4 s^{\mathcal{P}}_{i,h_p,h_p} (\tilde{\alpha}_p - \alpha_p) \\
& + \sum_{q \neq p}^{K} \Big(\alpha_q - \tilde{\alpha}_q + s^{\mathcal{P}}_{i,h_p,h_q} (4 \sum_{l \neq p,q}^{K} \mathbf{\Delta_l} + 2(\alpha_p - \tilde{\alpha}_p) + 2(\tilde{\alpha}_q - \alpha_q) \ ) \Big) \Big] \mathbf{\Delta_p} \\
& + \alpha_p^2 + \tilde{\alpha}_p^2 + s^{\mathcal{P}}_{i,h_p,h_p} (-\alpha_p^2 - \tilde{\alpha}_p^2 + 2\alpha_p \tilde{\alpha}_p) + \sum_{q \neq p}^{K} \alpha_p \alpha_q + \tilde{\alpha}_p \tilde{\alpha}_q \\
& + \sum_{q \neq p}^{K} s^{\mathcal{P}}_{i,h_p,h_q} \Big(2(\alpha_p - \tilde{\alpha}_p) \sum_{l \neq p,q}^{K} \mathbf{\Delta_l} - \alpha_p \alpha_q - \tilde{\alpha}_p \tilde{\alpha}_q + \tilde{\alpha}_p \alpha_q + \alpha_p \tilde{\alpha}_q \Big) \Bigg]
\end{aligned}
$$

Hence $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ can be expressed as (3):

$$
\begin{aligned}
d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}) = & \sum_{p=1}^{K} \Big[\ 4 \Big(\sum_{q \neq p}^{K} s^{\mathcal{P}}_{i,h_p,h_q} - s^{\mathcal{P}}_{i,h_p,h_p} \Big) \mathbf{\Delta_p^2} + \Psi^{\mathcal{P}}_{pq}(\Delta_{l \neq p,q}) \mathbf{\Delta_p} + \Phi^{\mathcal{P}}_{pq}(\Delta_{l \neq p,q}) \ \Big] \\
= & \xi^{\mathcal{P}}(\mathbf{\Delta_1}, .., \mathbf{\Delta_K}) \qquad (3)
\end{aligned}
$$

Algebraic derivation of expression (4)

Expression (4) is obtained directly from expression (3) for $s^{\mathcal{P}}_{i,h_p,h_p} = 1$ and $s^{\mathcal{P}}_{i,h_p,h_q} = 0$ when $\mathcal{P} = \text{IBS}_{\text{hap}}$.

Algebraic derivation of expressions (5) and (6)

Considering that only two haplotypes exist among the K possible ones is the same as setting $K = 2$. Expression (5) can be obtained directly by reducing expression (4) for the case where $K = 2$. However another derivation is given here so as to exhibit other properties, such as a lower bound, for the matrix distance. For $K = 2$ expression (2) becomes:

$$\begin{aligned} d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}) =& (f^{QTL}_{i,h_1a_1}{}^2 + f^{QTL}_{i,h_1a_2}{}^2) - s^{\mathcal{P}}_{i,h_1,h_1}(f^{QTL}_{i,h_1a_1} - f^{QTL}_{i,h_1a_2})^2 \\ &+ 2(f^{QTL}_{i,h_1a_1} f^{QTL}_{i,h_2a_1} + f^{QTL}_{i,h_1a_2} f^{QTL}_{i,h_2a_2})(1 - s^{\mathcal{P}}_{i,h_1,h_2}) \\ &+ 2(f^{QTL}_{i,h_1a_2} f^{QTL}_{i,h_2a_1} + f^{QTL}_{i,h_1a_1} f^{QTL}_{i,h_2a_2}) s^{\mathcal{P}}_{i,h_1,h_2} \\ &- s^{\mathcal{P}}_{i,h_2,h_2}(f^{QTL}_{i,h_2a_1} - f^{QTL}_{i,h_2a_2})^2 + (f^{QTL}_{i,h_2a_1}{}^2 + f^{QTL}_{i,h_2a_2}{}^2) \end{aligned}$$

and the frequencies in expression (2) can be written as:

$$\begin{aligned} f^{QTL}_{i,h_1a_1} &= f_{i,h_1} f_{a_1} + \mathbf{\Delta_1} = \alpha_1 + \mathbf{\Delta_1} \\ f^{QTL}_{i,h_1a_2} &= f_{i,h_1} f_{a_2} - \mathbf{\Delta_1} = \tilde{\alpha}_1 - \mathbf{\Delta_1} \\ f^{QTL}_{i,h_2a_1} &= f_{i,h_2} f_{a_1} - \mathbf{\Delta_1} = \alpha_2 - \mathbf{\Delta_1} \\ f^{QTL}_{i,h_2a_2} &= f_{i,h_2} f_{a_2} + \mathbf{\Delta_1} = \tilde{\alpha}_2 + \mathbf{\Delta_1} \end{aligned}$$

Note that $\mathbf{\Delta_1}$, in this case, can be expressed as $\mathbf{\Delta_1} = f^{QTL}_{i,h_1a_1} f^{QTL}_{i,h_2a_2} - f^{QTL}_{i,h_1a_2} f^{QTL}_{i,h_2a_1}$ with its maximum and minimum value given by $\frac{1}{4}$ and $-\frac{1}{4}$ respectively. The maximum value of $\mathbf{\Delta_1}$ is given by $f^{QTL}_{i,h_1a_1} =$

$f^{QTL}_{i,h_2a_2} = \frac{1}{2}$ and $f^{QTL}_{i,h_1a_2} = f^{QTL}_{i,h_2a_1} = 0$, and its minimum value is given by $f^{QTL}_{i,h_1a_1} = f^{QTL}_{i,h_2a_2} = 0$ and $f^{QTL}_{i,h_1a_2} = f^{QTL}_{i,h_2a_1} = \frac{1}{2}$. Replacing the haplotype frequencies in $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ with the expressions including the $\alpha_1, \tilde{\alpha}_1, \tilde{\alpha}_2, \alpha_2$ and $\mathbf{\Delta_1}$ terms gives:

$$\begin{aligned} d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}) = & \left[-4s^{\mathcal{P}}_{i,h_1,h_1} - 4s^{\mathcal{P}}_{i,h_2,h_2} + 8s^{\mathcal{P}}_{i,h_1,h_2}\right] \mathbf{\Delta_1}^2 \\ & + \left[4s^{\mathcal{P}}_{i,h_1,h_1}(\tilde{\alpha}_1 - \alpha_1) + 4s^{\mathcal{P}}_{i,h_2,h_2}(\alpha_2 - \tilde{\alpha}_2) - 4s^{\mathcal{P}}_{i,h_1,h_2}(\tilde{\alpha}_1 + \alpha_2 - (\alpha_1 + \tilde{\alpha}_2))\right] \mathbf{\Delta_1} \\ & - s^{\mathcal{P}}_{i,h_1,h_1}(\tilde{\alpha}_1 - \alpha_1)^2 - s^{\mathcal{P}}_{i,h_2,h_2}(\alpha_2 - \tilde{\alpha}_2)^2 + 2s^{\mathcal{P}}_{i,h_1,h_2}(\tilde{\alpha}_1 - \alpha_1)(\alpha_2 - \tilde{\alpha}_2) \\ & + (\alpha_1 + \alpha_2)^2 + (\tilde{\alpha}_2 + \tilde{\alpha}_1)^2 \\ & = \xi^{\mathcal{P}}(\mathbf{\Delta_1}) \end{aligned}$$

Hence $d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL})$ can be expressed as:

$$\xi^{\mathcal{P}}(\mathbf{\Delta_1}) = \left[-4s^{\mathcal{P}}_{i,h_1,h_1} - 4s^{\mathcal{P}}_{i,h_2,h_2} + 8s^{\mathcal{P}}_{i,h_1,h_2}\right] \mathbf{\Delta_1^2} + \Psi^{\mathcal{P}}\mathbf{\Delta_1} + \Phi^{\mathcal{P}} \qquad (5)$$

For the extreme values of $\mathbf{\Delta_1}$ we have:

$$\xi^{\mathcal{P}}\left(\frac{1}{4}\right) = \xi^{\mathcal{P}}\left(-\frac{1}{4}\right) = \frac{1}{2} + \frac{1}{2}s^{\mathcal{P}}_{i,h_1,h_2} - \frac{1}{4}(s^{\mathcal{P}}_{i,h_1,h_1} + s^{\mathcal{P}}_{i,h_2,h_2})$$

This quantity can also be obtained simply from expression (2), when $K = 2$, by replacing f^{QTL}_{i,h_1a_1}, f^{QTL}_{i,h_2a_2}, f^{QTL}_{i,h_1a_2} and f^{QTL}_{i,h_2a_1} by their corresponding values for the maximum and minimum value of $\mathbf{\Delta_1}$. For $\mathbf{\Delta_1} = \frac{1}{4}$ we have:

$$\begin{aligned} d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}) = & \left(\left(\frac{1}{2}\right)^2 + 0^2\right) - s^{\mathcal{P}}_{i,h_1,h_1}\left(\frac{1}{2} - 0\right)^2 + 2\left(\ \frac{1}{2}.0\ +\ 0.\frac{1}{2}\ \right)(1 - s^{\mathcal{P}}_{i,h_1,h_2}) \\ & + 2\left(\ 0.0\ +\ \frac{1}{2}.\frac{1}{2}\ \right)s^{\mathcal{P}}_{i,h_1,h_2} - s^{\mathcal{P}}_{i,h_2,h_2}\left(0 - \frac{1}{2}\right)^2 + \left(0^2 + \left(\frac{1}{2}\right)^2\right) \\ = & \frac{1}{2} + \frac{1}{2}s^{\mathcal{P}}_{i,h_1,h_2} - \frac{1}{4}(s^{\mathcal{P}}_{i,h_1,h_1} + s^{\mathcal{P}}_{i,h_2,h_2}) \\ = & \xi^{\mathcal{P}}\left(\frac{1}{4}\right) \end{aligned}$$

In the same manner we can show that $\xi^{\mathcal{P}}\left(-\frac{1}{4}\right) = \xi^{\mathcal{P}}\left(\frac{1}{4}\right)$ since the squares and the products of the frequencies in expression (2) when $K = 2$ are symmetric. $\frac{1}{2} + \frac{1}{2}s^{\mathcal{P}}_{i,h_1,h_2} - \frac{1}{4}(s^{\mathcal{P}}_{i,h_1,h_1} + s^{\mathcal{P}}_{i,h_2,h_2})$ is greater or equal to $\frac{1}{2}s^{\mathcal{P}}_{i,h_1,h_2}$ since the maximum possible value of $s^{\mathcal{P}}_{i,h_1,h_1}$ and $s^{\mathcal{P}}_{i,h_2,h_2}$ is equal to one. Hence if haplotypes h_1 and h_2 share allele similarity $s^{\mathcal{P}}_{i,h_1,h_2}$ will be positive and $\xi^{\mathcal{P}}\left(-\frac{1}{4}\right) = \xi^{\mathcal{P}}\left(\frac{1}{4}\right) \in [\frac{1}{2}s^{\mathcal{P}}_{i,h_1,h_2}, 1]$. For $\mathcal{P} = \text{IBS}_{\text{hap}}$, i.e. $s^{\mathcal{P}}_{i,h_1,h_1} = s^{\mathcal{P}}_{i,h_2,h_2} = 1$ and $s^{\mathcal{P}}_{i,h_1,h_2} = 0$, we have $\xi^{\mathcal{P}}\left(-\frac{1}{4}\right) = \xi^{\mathcal{P}}\left(\frac{1}{4}\right) = 0$. Note that for $\mathcal{P} = \text{IBS}_{\text{hap}}$ $\xi^{\mathcal{P}}(\mathbf{\Delta_1})$ becomes:

$$\xi^{\text{IBS}_{\text{hap}}}(\mathbf{\Delta_1}) = -8\mathbf{\Delta_1}^2 + 4\left[(\tilde{\alpha}_1 - \alpha_1) + (\alpha_2 - \tilde{\alpha}_2)\right]\mathbf{\Delta_1} - (\tilde{\alpha}_1 - \alpha_1)^2 - (\alpha_2 - \tilde{\alpha}_2)^2 + (\alpha_1 + \alpha_2)^2 + (\tilde{\alpha}_2 + \tilde{\alpha}_1)^2 \qquad (6)$$

Thus differentiating $\xi^{\text{IBS}_{\text{hap}}}$ with respect to $\mathbf{\Delta_1}$ gives:

$$\mathbf{\Delta_1^*} = \frac{\tilde{\alpha}_1 - \alpha_1 + \alpha_2 - \tilde{\alpha}_2}{4} = \frac{(2f_{a_1} - 1)(1 - 2f_{i,h_1})}{4}$$

The minimum of $\mathbf{\Delta_1^*}$, which is equal to $-\frac{1}{4}$, is given by either $f_{a_1} = f_{i,h_1} = 1$ or $f_{a_1} = f_{i,h_1} = 0$. Its maximum value, which is equal to $\frac{1}{4}$, is given by either $f_{a_1} = 0$ and $f_{i,h_1} = 1$ or $f_{a_1} = 1$ and $f_{i,h_1} = 0$. Hence $\mathbf{\Delta_1^*}$ takes its minimum and maximum values when both the tested locus and the QTL are monomorphic.

Algebraic derivation of the matrix distance for a multiallelic QTL

For S distinct alleles at the QTL expression (2) generalizes to:

$$d_1(\mathbf{M}^{\mathcal{P},i}, \mathbf{M}^{QTL}) = \sum_{p=1}^{K}\sum_{q=1}^{K}\left[(1 - s^{\mathcal{P}}_{i,h_p,h_q})\sum_{l=1}^{S} f^{QTL}_{i,h_p a_l} f^{QTL}_{i,h_q a_l} + s^{\mathcal{P}}_{i,h_p,h_q}\sum_{l=1}^{S}\sum_{m\neq l}^{S} f^{QTL}_{i,h_p a_l} f^{QTL}_{i,h_q a_m}\right] \qquad (7)$$

For $\mathcal{P} = \text{IBS}_{\text{hap}}$ expression (7) becomes:

$$d_1(\mathbf{M}^{\mathrm{IBS_{hap}},i},\mathbf{M}^{QTL}) = \sum_{p=1}^{K}\sum_{l=1}^{S}\sum_{m\neq l}^{S} f_{i,h_pa_l}^{QTL} f_{i,h_pa_m}^{QTL} + \sum_{p=1}^{K}\sum_{l=1}^{S}\sum_{q\neq p}^{K} f_{i,h_pa_l}^{QTL} f_{i,h_qa_l}^{QTL} \quad (7)$$

Let $\mathbf{\Delta_{pl}} = f_{i,h_pa_l}^{QTL} - f_{i,h_p}f_{a_l} = f_{i,h_pa_l}^{QTL} - \alpha_{pl}$, which is equivalent to $f_{i,h_pa_l}^{QTL} = \alpha_{pl} + \mathbf{\Delta_{pl}}$. Note that:

$$D_{i,QTL}^2 = \sum_{p=1}^{K}\sum_{l=1}^{S}(f_{i,h_pa_l}^{QTL} - f_{i,h_p}.f_{a_l})^2 = \sum_{p=1}^{K}\sum_{l=1}^{S}\mathbf{\Delta_{pl}}^2$$

Since $\sum_{p=1}^{K}\mathbf{\Delta_{pl}} = 0$ we have $\sum_{q\neq p}^{K}\mathbf{\Delta_{ql}} = -\mathbf{\Delta_{pl}}$ and $\sum_{m\neq l}^{S}\mathbf{\Delta_{pm}} = -\mathbf{\Delta_{pl}}$. Replacing the haplotype frequencies in expression (7) with the LD coefficients and the product of frequencies terms, and subsequently replacing $\sum_{q\neq p}^{K}\mathbf{\Delta_{ql}}$ and $\sum_{m\neq l}^{S}\mathbf{\Delta_{pm}}$ with $-\mathbf{\Delta_{pl}}$, gives:

$$\begin{aligned} d_1(\mathbf{M}^{\mathrm{IBS_{hap}},i},\mathbf{M}^{QTL}) = \sum_{p=1}^{K}\sum_{l=1}^{S}\Bigg[&-2\mathbf{\Delta_{pl}^2} + \left(\Psi_{pl}^{\mathrm{IBS_{hap}},(1)} + \Psi_{pl}^{\mathrm{IBS_{hap}},(2)}\right)\mathbf{\Delta_{pl}} \\ &+ \left(\Phi_{pl}^{\mathrm{IBS_{hap}},(1)} + \Phi_{pl}^{\mathrm{IBS_{hap}},(2)}\right)\Bigg] \\ = \xi^{\mathrm{IBS_{hap}}}&(\mathbf{\Delta_{11}},\mathbf{\Delta_{12}},...,\mathbf{\Delta_{KS}}) \end{aligned}$$

As for expression (3) the general behavior of the matrix distance for continuous predictors in $[0,1]$, as function of LD coefficients, is unspecifiable for the multiallelic QTL case. Hence we did not express the matrix distance, here in the multiallelic QTL case, for continuous predictors in $[0,1]$.

Additional file 2

Domains of LD coefficients and boundary conditions for the critical values of each Qp function

Domains of values for the multiallelic LD coefficients

Let $\mathbf{\Delta_p}$ be the LD coefficient between haplotype h_p and allele a_1 at the QTL, i.e.

$$\begin{cases} \mathbf{\Delta_p} = f^{QTL}_{i,h_pa_1} - f_{i,h_p}f_{a_1} \ (a) \\ -\mathbf{\Delta_p} = f^{QTL}_{i,h_pa_2} - f_{i,h_p}f_{a_2} \Leftrightarrow \mathbf{\Delta_p} = f_{i,h_p}f_{a_2} - f^{QTL}_{i,h_pa_2} \ (b) \end{cases} \qquad \begin{cases} f^{QTL}_{i,h_pa_1} = f_{i,h_p}f_{a_1} + \mathbf{\Delta_p} \ (c) \\ f^{QTL}_{i,h_pa_2} = f_{i,h_p}f_{a_2} - \mathbf{\Delta_p} \ (d) \end{cases}$$

$\mathbf{\Delta_p}$ is maximum in (a) and (b) when $f^{QTL}_{i,h_pa_1} = f_{a_1}$ and $f^{QTL}_{i,h_pa_2} = 0$ respectively. Under these conditions we also have $f_{i,h_p} = f^{QTL}_{i,h_pa_1} = f_{a_1}$ since $f_{i,h_p} = f^{QTL}_{i,h_pa_1} + f^{QTL}_{i,h_pa_2} = f_{a_1} + 0$. Hence $\mathbf{\Delta_p}$ can be written as $\mathbf{\Delta_p} = f_{i,h_p} - f^2_{i,h_p} = f_{i,h_p}(1 - f_{i,h_p})$ under these conditions. $f_{i,h_p}(1 - f_{i,h_p})$ is identifiable to the function $x \mapsto x(1-x)$ which takes a maximum value of $\frac{1}{4}$ for $x = \frac{1}{2}$. One of the maximum possible value for $\mathbf{\Delta_p}$ is thus given by $\frac{1}{4}$. In the same manner we can show that one of the minimum possible value for $\mathbf{\Delta_p}$ is given by $-\frac{1}{4}$. Since $f^{QTL}_{i,h_pa_1} \geq 0$ and $f^{QTL}_{i,h_pa_2} \geq 0$ we also have $\mathbf{\Delta_p} \geq -f_{i,h_p}.f_{a_1}$ and $\mathbf{\Delta_p} \leq f_{i,h_p}.f_{a_2}$ from (c) and (d) respectively. Hence the complete domain of values for each $\mathbf{\Delta_p}$ term is given by: $\mathbf{\Delta_p} \in \left[max(-\frac{1}{4}, -f_{i,h_p}.f_{a_1}), min(\frac{1}{4}, f_{i,h_p}.f_{a_2})\right]$.

Boundary conditions for the critical value of each Q_p function

Each Q_p function is given by:

$$Q_p(\mathbf{\Delta_p}) = -4\mathbf{\Delta_p^2} + \Psi^{\text{IBS}_{\text{hap}}}_{pq}\mathbf{\Delta_p} + \Phi^{\text{IBS}_{\text{hap}}}_{pq}$$

Differentiating Q_p with respect to $\mathbf{\Delta_p}$ gives $\mathbf{\Delta^*_p} = \frac{\Psi^{\text{IBS}_{\text{hap}}}_{pq}}{8}$ where $\Psi^{\text{IBS}_{\text{hap}}}_{pq}$ is given by:

$$\Psi_{pq}^{\mathrm{IBS_{hap}}} = 3(\tilde{\alpha}_p - \alpha_p) + \sum_{q \neq p}^{K} (\alpha_q - \tilde{\alpha}_q)$$

See expression (3), with $s^{\mathcal{P}}_{i,h_p,h_p} = 1$ and $s^{\mathcal{P}}_{i,h_p,h_q} = 0$, in Additional file 1 for $\Psi_{pq}^{\mathrm{IBS_{hap}}}$ and the corresponding products of frequencies for α_p and $\tilde{\alpha}_p$.

$$\begin{aligned}
\Psi_{pq}^{\mathrm{IBS_{hap}}} &= 3(\tilde{\alpha}_p - \alpha_p) + \sum_{q \neq p}^{K} (\alpha_q - \tilde{\alpha}_q) \\
&= 3(\tilde{\alpha}_p - \alpha_p) + \sum_{q \neq p}^{K} (\alpha_q - \tilde{\alpha}_q) + \alpha_p - \tilde{\alpha}_p - (\alpha_p - \tilde{\alpha}_p) \\
&= 4(\tilde{\alpha}_p - \alpha_p) + \sum_{q=1}^{K} (\alpha_q - \tilde{\alpha}_q) \\
&= 4f_{i,h_p}(f_{a_2} - f_{a_1}) + (f_{a_1} - f_{a_2}) \sum_{q=1}^{K} f_{i,h_q} = (f_{a_1} - f_{a_2})[1 - 4f_{i,h_p}]
\end{aligned}$$

Hence we have $\mathbf{\Delta_p^*} = \dfrac{(f_{a_1} - f_{a_2})[1 - 4f_{i,h_p}]}{8} = \dfrac{(2f_{a_1} - 1)[1 - 4f_{i,h_p}]}{8}$. Let $\mathbf{\Delta_{p_{min}}} = max(-\dfrac{1}{4}, -f_{i,h_p}.f_{a_1})$ and $\mathbf{\Delta_{p_{max}}} = min(\dfrac{1}{4}, f_{i,h_p}.f_{a_2})$. Note that $0 \in \left]\mathbf{\Delta_{p_{min}}}, \mathbf{\Delta_{p_{max}}}\right[$. Hence if $\mathbf{\Delta_p^*} \in \left]\mathbf{\Delta_{p_{min}}}, \mathbf{\Delta_{p_{max}}}\right[$ and the magnitude (absolute value) of $\mathbf{\Delta_p}$ increases sufficiently Q_p will decrease. Figure 8 gives an example of a Q_p function with $\mathbf{\Delta_p^*} \in \left]\mathbf{\Delta_{p_{min}}}, \mathbf{\Delta_{p_{max}}}\right[$.

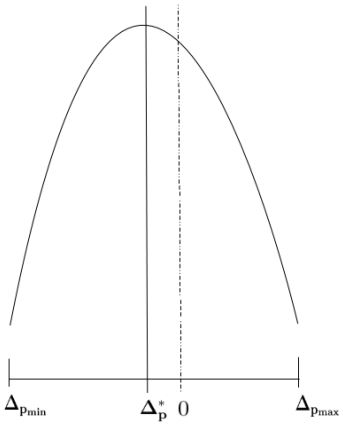


FIGURE 8: An example of a Q_p function with $\mathbf{\Delta_p^*} \in \left]\mathbf{\Delta_{p_{min}}}, \mathbf{\Delta_{p_{max}}}\right[$.

The only situations for which it is not possible to tell if Q_p will decrease, as the magnitude of $\mathbf{\Delta_p}$ increases sufficiently, are given by the following conditions; $\mathbf{\Delta_p^* \leq \Delta_{p_{min}}}$ (e) or $\mathbf{\Delta_p^* \geq \Delta_{p_{max}}}$ (f). In these situations Q_p can either decrease or increase if the magnitude of $\mathbf{\Delta_p}$ increases sufficiently. Figure 9 gives an example of a Q_p function with $\mathbf{\Delta_p^* \leq \Delta_{p_{min}}}$.

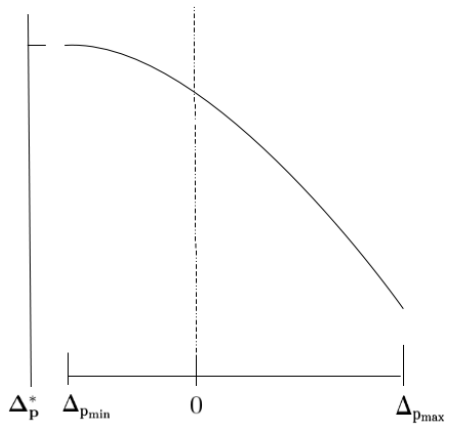


FIGURE 9: An example of a Q_p function with $\mathbf{\Delta_p^* \leq \Delta_{p_{min}}}$.

Conditions (e) and (f) can be written as follows:

$$(e)\begin{cases} \text{If } \mathbf{\Delta_{p_{min}}} = -\dfrac{1}{4} \text{ then } \mathbf{\Delta_p^* \leq \Delta_{p_{min}}} \Leftrightarrow (2f_{a_1}-1)[1-4f_{i,h_p}] \leq -2 \\ \text{or} \\ \text{If } \mathbf{\Delta_{p_{min}}} = -f_{i,h_p}.f_{a_1} \text{ then } \mathbf{\Delta_p^* \leq \Delta_{p_{min}}} \Leftrightarrow \dfrac{(2f_{a_1}-1)[1-4f_{i,h_p}]}{f_{i,h_p}.f_{a_1}} \leq -8 \end{cases}$$

$$(f)\begin{cases} \text{If } \mathbf{\Delta_{p_{max}}} = \dfrac{1}{4} \text{ then } \mathbf{\Delta_p^* \geq \Delta_{p_{max}}} \Leftrightarrow (2f_{a_1}-1)[1-4f_{i,h_p}] \geq 2 \\ \text{or} \\ \text{If } \mathbf{\Delta_{p_{max}}} = f_{i,h_p}(1-f_{a_1}) \text{ then } \mathbf{\Delta_p^* \geq \Delta_{p_{max}}} \Leftrightarrow \dfrac{(2f_{a_1}-1)[1-4f_{i,h_p}]}{f_{i,h_p}(1-f_{a_1})} \geq 8 \end{cases}$$

Let w, t, s and u be the following functions of f_{a_1} and f_{i,h_p}; $w(f_{a_1}, f_{i,h_p}) = (2f_{a_1} - 1)[1 - 4f_{i,h_p}] + 2$, $t(f_{a_1}, f_{i,h_p}) = \dfrac{(2f_{a_1} - 1)[1 - 4f_{i,h_p}]}{f_{i,h_p} \cdot f_{a_1}} + 8$, $s(f_{a_1}, f_{i,h_p}) = (2f_{a_1} - 1)[1 - 4f_{i,h_p}] - 2$ and $u(f_{a_1}, f_{i,h_p}) = \dfrac{(2f_{a_1} - 1)[1 - 4f_{i,h_p}]}{f_{i,h_p}(1 - f_{a_1})} - 8$. Conditions (e) and (f) are the same as searching values of f_{a_1} and f_{i,h_p} for which we have:

$$(e) \begin{cases} w \leq 0 \\ \text{or} \\ t \leq 0 \end{cases} \qquad (f) \begin{cases} s \geq 0 \\ \text{or} \\ u \geq 0 \end{cases}$$

Figure 10 shows the regions (red colored), for different a_1 and h_p frequencies, where condition (e) or (f) is realized. As can be seen in figure 10, the conditions for w, t, s and u are verified when f_{a_1} and f_{i,h_p} are both high or both low, or one of these two frequencies is high and the other one is low. Note that these frequencies correspond to situations where Q_p can still decrease as suggested by figure 11 (see relation between the sum of the squared deviations and $D^2_{i,QTL}$).

Moreover these frequencies correspond to situations which are unfavorable for QTL analysis as low frequencies do not allow for reliable estimation and comparison of contrasts between groups of individuals. Finally note that LD requires variation of alleles between loci to exist. Hence these high or low frequencies correspond to situations which are unfavorable for LD mapping of QTL.

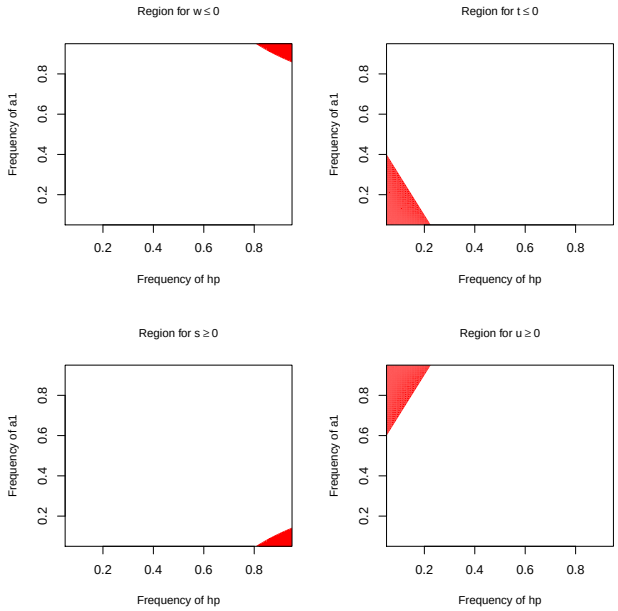


FIGURE 10: The different regions (red colored) where condition (e) or (f) is realized.

Relation between the sum of the squared deviations and $D^2_{i,QTL}$

Let $D^2_{i,QTL} = 2\sum_{p=1}^{K} \mathbf{\Delta_p^2}$ be the non-normalized multiallelic measure of LD and $SD_{i,QTL} = \sum_{p=1}^{K}(\mathbf{\Delta_p} - \mathbf{\Delta_p^*})^2$ be the sum of the squared deviations of the multiallelic LD coefficients from their corresponding $\mathbf{\Delta_p^*}$ critical values. $SD_{i,QTL}$ can be written as a sum of convex U_p functions of each LD coefficient, i.e.

$$SD_{i,QTL} = \sum_{p=1}^{K}(\mathbf{\Delta_p} - \mathbf{\Delta_p^*})^2 = \sum_{p=1}^{K} \mathbf{\Delta_p^2} - \omega\mathbf{\Delta_p} + \upsilon = \sum_{p=1}^{K} U_p(\mathbf{\Delta_p})$$

where $\omega = 2\mathbf{\Delta_p^*}$ and $\upsilon = \mathbf{\Delta_p^*}^2$, and the critical value of each U_p function is $\mathbf{\Delta_p^*}$. Hence there is an implicit relationship between $D^2_{i,QTL}$ and $SD_{i,QTL}$. If the sum of the squared $\mathbf{\Delta_p}$ terms increases sufficiently (i.e. $\frac{D^2_{i,QTL}}{2}$ increases sufficiently) $SD_{i,QTL}$ will increase. The same procedure as the one used to describe the expected behavior of $D^2_{i,QTL}$ was repeated for $SD_{i,QTL}$ and $\frac{D^2_{i,QTL}}{2}$ on the 889 FLW regions (see variation of LD subsection in methods). That is both $\mathbf{\Delta_p}$ and $\mathbf{\Delta_p^*}$ were computed at

each tested position, in order to compute $SD_{i,QTL}$ and $\frac{D^2_{i,QTL}}{2}$, while screening the 889 regions. Figure 11 shows the profiles of the empirical means of the 889 FLW curves for $SD_{i,QTL}$ and $\frac{D^2_{i,QTL}}{2}$, and the deviation $\left(\mathbb{E}[SD_{i,QTL} - \frac{D^2_{i,QTL}}{2}] = \mathbb{E}[-\omega\boldsymbol{\Delta}_{\mathbf{p}} + v]\right)$ between these two profiles. As observed in figure 11, the profiles for the expected values of $SD_{i,QTL}$ and $\frac{D^2_{i,QTL}}{2}$ exhibit similar patterns with a relatively small increasing deviation, between these two profiles, as the tested position moves toward the QTL. Note that the profile for the deviation exhibit a similar trend to the profiles for the expected values of $SD_{i,QTL}$ and $\frac{D^2_{i,QTL}}{2}$.

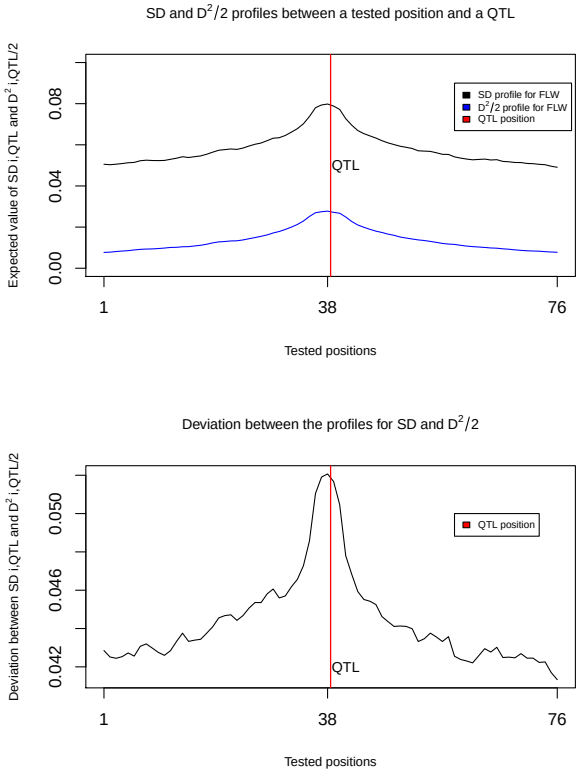


FIGURE 11: Empirical means of the 889 FLW curves, obtained for $D^2_{i,QTL}$ and $SD_{i,QTL}$ between tested positions (tested position i = center of 6 marker haplotypes) and a biallelic QTL (red vertical line) for regions of 81 markers on chromosomes, and the deviation between the mean curves.

5.4 Discussion et perspectives de l'étude

J'ai proposé pour cette étude une méthode basée sur une distance matricielle, qui aide à mieux discriminer entre les méthodes haplotypiques effectuant une prédiction d'identité allélique, à une position testée, entre des allèles non observés. Cette méthode, appelée efficacité relative, a permis de valider des résultats de cartographie et a mis en évidence que IBS_{hap} est un prédicteur de choix en cartographie de QTL pour un marquage à haute densité.

Néanmoins, cette méthode ne donne pas d'indication sur la faisabilité d'application, en cartographie de QTL, d'un AIP que l'on pourrait identifier comme étant efficace. Un AIP ayant une meilleure efficacité relative que d'autres prédicteurs pourrait, par exemple, produire une matrice de corrélations entre haplotypes qui serait mal conditionnée ou voir même singulière. La notion d'efficacité relative a donc des limites dans la discrimination entre des prédicteurs d'identité allélique, puisqu'elle ne caractérise pas la faisabilité de la détection.

Toutefois, l'efficacité relative reste un moyen rapide et pratique pour tester la capacité de prédiction d'identité allélique d'un AIP par rapport au LD. En effet, cette méthode ne nécessite pas que l'on effectue des simulations pour comparer des AIP. L'efficacité relative peut être vue comme une notion complémentaire à celle de l'étude de la capacité de cartographie par simulations et réciproquement. Remarquons que la distance matricielle choisie pour la définition de l'efficacité relative dans cette étude procure un cadre théorique pour la compréhension des mécanismes sous-jacents à la cartographie de QTL par l'exploitation du LD multiallélique.

En effet, la statistique $\text{RMSE}^{r.e.}$ basée sur cette distance s'est avérée corrélée d'au moins 0.9 à la statistique $\text{RMSE}^{m.a.}$ dans le cadre de l'article 1. De plus, les résultats algébriques basés sur cette distance montrent que IBS_{hap} prédit correctement l'identité allélique à une position testée, par rapport au LD existant entre ce dernier et un QTL, sachant que ce prédicteur a été majoritairement plus précis que les autres. Les résultats algébriques montrent également que la bonne relation entre la prédiction d'identité allélique et le LD pour IBS_{hap} est toujours vraie quelque soit le degré d'allélisme au QTL. Ces résultats montrent aussi, pour le cas particulier de deux haplotypes, que l'extension de la

prédiction d'identité allélique de l'ensemble discret $\{0, 1\}$ à l'intervalle [0,1] mène à une détérioration de la prise en compte du LD par la prédiction. Ainsi, ce cas particulier met en défaut l'hypothèse généralement avancée dans la littérature (Meuwissen et Goddard, 2001 ; Meuwissen *et al.*, 2002 ; Li et Jiang, 2005), qui est que l'on exploite mieux le LD entre une position testée et un QTL si l'on tient compte de la similarité partielle et totale entre des haplotypes (i.e. l'extension de $\{0, 1\}$ à [0,1] pour la prédiction). Les résultats numériques de l'article 1 ont validé et illustré ces résultats algébriques.

Au vu de l'ensemble des résultats de l'article 1, on peut conclure que IBS_{hap} est un prédicteur qui a plusieurs avantages par rapport aux autres AIP comparés. Ce prédicteur possède de bonnes propriétés théoriques par rapport au LD multiallélique, il a l'avantage d'être rapide en temps de calcul, simple pour l'implémentation et stable numériquement pour les analyses d'association. Néanmoins, on rappelle qu'un marquage à haute densité, de 50K (cf. article 1) ou voir bien plus, est nécessaire pour l'utilisation de cet AIP. Un tel niveau de marquage n'est pas toujours accessible en pratique. Une telle densité devrait cependant, sous réserve d'existence de variants causaux, permettre de capter le LD multiallélique entre des haplotypes et les variants cachés (i.e. non observés).

En plus de leur capacité à bien décrire un LD local, les approches haplotypiques sont aussi cohérentes d'un point de vue biologique car elles ne font pas d'hypothèse sur le degré d'allélisme au QTL. Par exemple, IBS_{hap} a été majoritairement plus précis que IBS_m sur l'ensemble des situations étudiées dans l'article 1, alors que les QTL simulés étaient bialléliques. De cette manière les approches haplotypiques peuvent potentiellement rendre compte d'un multiallélisme quelconque au QTL. Cependant, le phasage correct des chromosomes est un élément important dont il faut tenir compte si on veut exploiter au mieux le LD multiallélique. La reconstruction des phases en pratique peut parfois être difficile et constitue donc un facteur limitant dans l'application des approches haplotypiques. Ainsi, d'autres AIP que IBS_{hap} peuvent être envisagés dans les situations où le LD multiallélique serait peu exploitable à cause d'un phasage erroné et/ou d'un faible niveau de marquage.

Finalement, le choix de la taille des haplotypes en nombre de marqueurs est également un autre facteur limitant dans l'application des approches haplotypiques. Le choix de ce paramètre peut être variable selon la densité de marquage, ou plus précisément selon la distance physique entre les marqueurs, afin de mieux capter le LD multiallélique. Or, le

nombre de paramètres à estimer dans un modèle haplotypique sera d'autant plus important que la taille des haplotypes sera grande. Un nombre trop important de paramètres mène généralement à une perte de puissance et une augmentation du temps de calcul dans les analyses. Des stratégies de regroupement d'haplotypes ont donc été proposées afin de palier ces problèmes (Browning et Browning, 2006 ; Druet *et al.*, 2008 ; Roldan *et al.*, 2012). Cependant, les résultats algébriques (et numériques) obtenus dans le cadre de l'article 1, pour le cas de deux haplotypes en particulier, montrent que la stratégie proposée par Browning et Browning (2006) ne permet pas de bien exploiter le LD. Une taille de 4 à 6 marqueurs telle que proposée par Calus *et al.* (2009) constituerait peut-être le choix approprié à utiliser pour un marquage à haute densité. Toutefois, plus de recherches devraient être menées afin de justifier le choix de la taille en fonction de la densité de marquage.

Partie III :
Discrimination entre modèles uni-SNP d'association et modèles uni-SNP de liaison

Chapitre 6

Les modèles uni-SNP d'association et uni-SNP de liaison

6.1 Contexte et importance de l'étude

Les analyses d'association sont des analyses précises et puissantes, au sens statistique, pour la cartographie de QTL. Cependant, ces approches sont aussi connues pour ne pas être robustes aux erreurs de première espèce (cf. 2.3.2). A contrario, les analyses de liaison sont connues pour leur robustesse, leur puissance limitée par le nombre de parents hétérozygotes et leur faible précision en cartographie de QTL (cf. 2.3.1). La robustesse des modèles de liaison vient de la prise en compte d'effets familiaux génétiques ou environnementaux. De ce fait, des modèles d'association corrigés pour la structure familiale ont été proposés afin de tenir compte de cette structure génétique. Ces modèles d'association corrigés sont actuellement décrits, dans la littérature, comme étant suffisamment robustes, puissants et précis pour potentiellement permettre de mener à bien une analyse en cartographie de QTL (cf. 2.3.2).

L'enjeu de cette partie consiste à caractériser les situations dans lesquelles les analyses d'association et les analyses de liaison sont robustes, ou non, aux erreurs de première espèce, et de quantifier leurs puissances relatives dans ces situations. L'objectif final de cette partie est de savoir s'il existe encore des situations où les analyses de liaison pourraient être plus avantageuses que les analyses d'association. Dans le chapitre 5 on a considéré des modèles d'association avec des effets haplotypiques, en aléatoires, et corrigés pour des structures familiales quelconques. Dans cette nouvelle partie, l'exploration se limitera à des structures de familles de demi-frères, et les modèles d'association comme de liaison

portent sur les effets fixés des allèles à un SNP.

Ces simplifications ont pour but de faciliter, comme on le verra dans le chapitre 7, la dérivation des facteurs de décentrage associés aux statistiques de test. Elles se justifient aussi par l'observation que les familles de demi-frères sont la norme chez les ruminants d'élevage et particulièrement chez les bovins. On se focalisera sur le modèle de liaison de Knott *et al.* (1996), pour la comparaison des analyses de liaison avec les analyses d'association, car ce modèle est une approche classique utilisée en cartographie de QTL. Ce chapitre présente les modèles statistiques d'analyse que l'on discriminera dans le chapitre 7, et la construction des tests associés afin d'effectuer cette discrimination.

6.2 Les modèles statistiques discriminés

On considère un dispositif de p familles, de p pères non apparentés, avec m descendants par famille. Les descendants considérés ont des mères différentes intra et inter famille de père. On a donc p familles différentes de m demi-frères et un nombre total d'individus donné par $n = mp$. On indice respectivement par A ($= Association$) et T ($= Transmission$) les éléments des modèles d'association et ceux des modèles de liaison. On suppose que le marqueur testé vérifie HWE et qu'il est identifiable au QTL, sauf si précisé autrement, et on note α l'effet de l'allèle a_1 au marqueur.

6.2.1 Les modèles d'association

Soient les génotypes a_2a_2, a_1a_2 et a_1a_1 au marqueur/QTL, que l'on codifie en -2, 0 et 2 respectivement, et soient n^i_{-2}, n^i_0 et n^i_2 les nombres d'individus ayant ces génotypes dans la famille i. $\forall i \in \{1, .., p\}$ on a $m = n^i_{-2} + n^i_0 + n^i_2$. Le modèle d'association corrigé pour la structure familiale, considéré ici, porte sur des effets fixés pour les allèles au QTL. Ce modèle s'écrit de la façon suivante :

$$Y_A = X_A\beta_A + \epsilon_A \,;\, \epsilon_A \sim \mathcal{N}_n(0, V) \tag{6.1}$$

$$\Longleftrightarrow \begin{pmatrix} Y_A^{(1)} \\ \vdots \\ \vdots \\ Y_A^{(i)} \\ \vdots \\ \vdots \\ Y_A^{(p)} \end{pmatrix} = \begin{pmatrix} X_A^{(1)} \\ \vdots \\ \vdots \\ X_A^{(i)} \\ \vdots \\ \vdots \\ X_A^{(p)} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \end{pmatrix} + \begin{pmatrix} \epsilon_A^{(1)} \\ \vdots \\ \vdots \\ \epsilon_A^{(i)} \\ \vdots \\ \vdots \\ \epsilon_A^{(p)} \end{pmatrix} \text{ avec } Y_A^{(i)} = \begin{pmatrix} Y_{i21} \\ \vdots \\ Y_{i2n_2^i} \\ Y_{i01} \\ \vdots \\ Y_{i0n_0^i} \\ Y_{i^-21} \\ \vdots \\ Y_{i^-2n_{-2}^i} \end{pmatrix} \text{ et } X_A^{(i)} = \begin{pmatrix} 1 & 2 \\ \vdots & \vdots \\ 1 & 2 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & -2 \\ \vdots & \vdots \\ 1 & -2 \end{pmatrix} \begin{matrix} \Big\} mf_{a_1}^2 \\ \Big\} 2mf_{a_1}f_{a_2} \\ \Big\} mf_{a_2}^2 \end{matrix}$$

Dans ce modèle Y_{ijk} correspond au phénotype de l'individu k ($k \in \{1, .., n_j^i\}$) ayant le génotype j ($j \in \{^-2, 0, 2\}$) dans la famille i. Les valeurs espérées sous HWE pour n_{-2}^i, n_0^i et n_2^i sont données par $n_{-2}^i = mf_{a_2}^2$, $n_0^i = 2mf_{a_1}f_{a_2}$ et $n_2^i = mf_{a_1}^2$. μ correspond ici à la moyenne générale, α à l'effet de l'allèle a_1 au marqueur et $X_A^{(i)}$ est le bloc du design reliant ces effets aux individus de la famille i. Le vecteur ϵ_A dans ce modèle se décompose également de la façon suivante $\epsilon_A = u + \varepsilon$, où $u \sim \mathcal{N}_n(0, A\sigma_u^2)$ et $\varepsilon \sim \mathcal{N}_n(0, I_n\sigma_\varepsilon^2)$ sont respectivement les vecteurs aléatoires d'effets polygéniques et résiduels que l'on suppose gaussiens. On a donc $V = A\sigma_u^2 + I_n\sigma_\varepsilon^2$, où $A = \bigoplus_{i=1}^{p} S_i$ tel que $\forall i \in \{1, .., p\}$ $S_i = \frac{3}{4}I_m + \frac{1}{4}J_m$ (J_m est la matrice carrée de dimension m remplie de 1). Cette structuration de V, en fonction de la matrice pedigree A, permet de se prémunir des effets de structures familiales que l'on pourrait attribuer aux effets dus un QTL en cas de confusion entre effets. On remarque que l'écriture de A, comme la somme directe des S_i, correspond simplement à une matrice diagonale par blocs dont ces derniers sont les S_i. On dira que le modèle (6.1) correspond à un modèle d'association homoscédastique lorsque $V = I_n\sigma_\varepsilon^2$ (i.e. $\epsilon_A = \varepsilon$).

6.2.2 Les modèles de liaison (Knott *et al.*, 1996)

Soient $\tilde{n}_{-1}^i$ et $\tilde{n}_1^i$ les nombres d'individus, que l'on identifie dans la famille i, comme ayant respectivement reçu les allèles a_2 et a_1 du père i. On associe donc la codification -1 aux individus ayant reçu l'allèle a_2 et 1 à ceux ayant reçu l'allèle a_1, lorsque que l'on peut identifier ces individus dans une famille. Le nombre espéré de pères hétérozygotes sous HWE est donné par $pe = 2f_{a_1}f_{a_2}p$. Le modèle hétéroscédastique de liaison de Knott *et*

al. (1996) dans ce cadre, tenant éventuellement compte d'une hétérogénéité de variances résiduelles inter familles, s'écrit de la façon suivante :

$$Y_T = X_T\beta_T + \epsilon_T \,;\, \epsilon_T \sim \mathcal{N}_n(0, \tilde{V}) \tag{6.2}$$

$$\Longleftrightarrow \begin{pmatrix} Y_T^{(1)} \\ \vdots \\ \vdots \\ Y_T^{(i)} \\ \vdots \\ \vdots \\ Y_T^{(pe)} \end{pmatrix} = \begin{pmatrix} X_T^{(1)} \\ \vdots \\ \vdots \\ X_T^{(i)} \\ \vdots \\ \vdots \\ X_T^{(pe)} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_{pe} \\ \delta_1 \\ \vdots \\ \delta_{pe} \end{pmatrix} + \begin{pmatrix} \epsilon_T^{(1)} \\ \vdots \\ \vdots \\ \epsilon_T^{(i)} \\ \vdots \\ \vdots \\ \epsilon_T^{(pe)} \end{pmatrix} \text{ avec } Y_T^{(i)} = \begin{pmatrix} Y_{i11} \\ \vdots \\ Y_{i1\tilde{n}_1^i} \\ Y_{i^-11} \\ \vdots \\ Y_{i^-1\tilde{n}_{-1}^i} \end{pmatrix} \text{ et } X_T^{(i)} = \begin{bmatrix} X_{T\mu}^{(i)} & X_{T\delta}^{(i)} \end{bmatrix}$$

$$\text{où } X_{T\mu}^{(i)} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & 1 & \vdots & \vdots \\ \vdots & \vdots & 1 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} \text{ et } X_{T\delta}^{(i)} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & 1 & \vdots & \vdots \\ \vdots & \vdots & -1 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & -1 & 0 & 0 \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \} \frac{mf_{a_1}}{2} \\ \\ \\ \} \frac{mf_{a_2}}{2} \\ \\ \end{matrix}$$

$$\begin{matrix} & & i\text{-}ème & & \\ & & & & \end{matrix}$$

Dans ce modèle Y_{ils} correspond au phénotype de l'individu s ($s \in \{1, .., \tilde{n}_l^i\}$) ayant reçu l'allèle codifié l ($l \in \{-1, 1\}$) du père i ($i \in \{1, .., pe\}$). Les valeurs espérées sous HWE pour $\tilde{n}_{-1}^i$ et $\tilde{n}_1^i$ sont données par $\tilde{n}_{-1}^i = \frac{mf_{a_2}}{2}$ et $\tilde{n}_1^i = \frac{mf_{a_1}}{2}$. On remarque que seuls les descendants homozygotes sont informatifs ici, car on suppose que les génotypes des mères ne sont pas connus, d'où l'on a $\frac{mf_{a_2}}{2}$ et $\frac{mf_{a_1}}{2}$ descendants homozygotes, qui sont respectivement a_2a_2 et a_1a_1, pour chaque père hétérozygote. On a donc $\frac{m}{2} = \tilde{n}_{-1}^i + \tilde{n}_1^i$. μ_i et δ_i pour ce modèle correspondent respectivement à la moyenne et à l'effet additif des allèles au QTL dans la famille i, dont les blocs correspondants du design sont $X_{T\mu}^{(i)}$ et $X_{T\delta}^{(i)}$. On notera que la matrice $X_{T\delta}^{(i)}$, qui décrit la transmission des allèles du père i, n'est pas toujours déterminée et que ses éléments peuvent être des probabilités de transmission

calculées (Knott *et al.*, 1996).

On remarquera aussi que $X_{T\delta}^{(i)}$ décrit également, ici, la transmission des génotypes a_1a_1 et a_2a_2 aux individus, car les génotypes des mères ne sont pas connus. De plus, si l'effet α de l'allèle a_1 transmis par le père est le même que celui de la mère alors on aura un modèle de régression sur les génotypes intra famille, car le modèle est additif par construction. Sous cette hypothèse, on peut montrer que l'estimateur de δ_i est égal à deux fois l'estimateur de α au sein d'une famille (cf.7.4.1 et 7.4.2). Si on suppose, en outre, que l'effet α ne change pas entre les familles alors on aura $\forall i \in \{1, ..., pe\}\ \delta_i = 2\alpha$. Finalement on remarque que la matrice $\tilde{V}$, contrairement à V pour l'analyse d'association, n'est pas structurée en utilisant l'information pedigree car ce modèle de liaison tient déjà compte d'un effet familial. La matrice $\tilde{V}$ permet de tenir compte d'une hétérogénéité de variances entre familles s'il en existe une. Dans ce cas cela revient à poser $\tilde{V} = \bigoplus_{i=1}^{pe} \sigma^2_{\varepsilon_i} I_{\frac{m}{2}}$ avec $\forall (i,j) \in \{1, .., pe\}^2$, $\sigma^2_{\varepsilon_i} \neq \sigma^2_{\varepsilon_j}$. On dira que le modèle (6.2) correspond à un modèle de liaison homoscédastique lorsque $\tilde{V} = I_{ne}\sigma^2_{\varepsilon}$, où $ne = \frac{mpe}{2}$.

6.3 Le test de Fisher, seuil de rejet, puissance et risque

Le test de Fisher est un test mathématiquement équivalent au LRT, vu en 3.2, dans le cadre gaussien. En effet, l'estimateur du maximum de vraisemblance (ML) et celui des moindres carrés ordinaires (OLS), dans ce cas, sont les mêmes et on peut montrer que la statistique de Fisher peut se ré-écrire comme une fonction monotone croissante du LRT (Guyader, 2012). Néanmoins un avantage de la statistique de Fisher, par rapport au LRT, est qu'elle peut être parfois plus simple à calculer et à manipuler analytiquement.

On explicite dans cette section les éléments clefs, de manière relativement succincte, nécessaires à la construction des tests F pour les modèles d'association et les modèles de liaison considérés. Le lecteur pourra simplement, s'il le souhaite, éluder les étapes de la construction des tests pour l'association et la liaison. Cependant, une bonne compréhension de ces étapes permet de comprendre certaines propriétés et la provenance du facteur de décentrage associés à ce test.

6.3.1 Le test de Fisher (test F)

Le test F est un test géométrique et général dans le cadre de la régression. Ce test est basé sur le théorème de Cochran, qui permet de connaître les lois du numérateur et du dénominateur de la statistique de test, et sur le théorème des trois perpendiculaires (Azaïs et Bardet, 2006). Le théorème de Cochran, énoncé ci-après, et la notion de projecteur vu en 3.1.1, s'avèrent fondamentaux pour la construction des statistiques de test pour les modèles d'association et les modèles de liaison considérés.

Théorème (Cochran)**:** Soit E un sous-espace vectoriel de $\mathbb{R}^n$, de dimension k ($k \leq n$), et P_E un projecteur orthogonal sur E. Si $\varepsilon \sim \mathcal{N}_n(0, I_n\sigma^2)$ alors on a :

$$\|P_E\varepsilon\|_2^2 \sim \sigma^2\chi^2(tr(P_E)) \Longleftrightarrow \varepsilon' P_E\varepsilon \sim \sigma^2\chi^2(k)$$

Le test F :

Soient les hypothèses testées suivantes :

$$\begin{cases} H_1 : Y = X\beta + \varepsilon\,;\, \varepsilon \sim \mathcal{N}_n(0, I_n\sigma^2) \\ H_0 : Y = X_0\beta_0 + \varepsilon\,;\, \varepsilon \sim \mathcal{N}_n(0, I_n\sigma^2) \end{cases}$$

avec $vect(X_0) \subset vect(X)$, $rang(X_0) = k_0$ et $rang(X) = k$. $q = k - k_0$ représente le nombre de paramètres, dans le modèle associé à H_1, dont on veut tester la nullité simultanée. La statistique de Fisher, qui suit la loi $F(q, n-k)$ sous H_0, est donnée par :

$$\hat{F} = \frac{\|\hat{Y} - \hat{Y}_0\|_2^2/q}{\|Y - \hat{Y}\|_2^2/n-k} = \frac{\|\hat{Y} - \hat{Y}_0\|_2^2/q}{\hat{\sigma}^2} \underset{H_0}{\sim} F(q, n-k)$$

où $\hat{Y} = X\hat{\beta} = P_E Y$ et $\hat{Y}_0 = X_0\hat{\beta}_0 = P_{E_0}Y$ sont respectivement les projections de Y sur les espaces E et E_0 ($E_0 \subset E$) engendrés par les colonnes de X et X_0 (i.e. $E = vect(X)$ et $E_0 = vect(X_0)$). Le principe de ce test est naturel : si la distance entre $\hat{Y}$ et $\hat{Y}_0$ relativement à $\hat{\sigma}^2$ est grande, alors on rejettera H_0 pour un seuil prédéfini par rapport à la distribution de Fisher à q et $n-k$ degrés de libertés ($F(q, n-k)$). Autrement dit si cette distance est petite, c'est à dire que si le modèle le plus simple sous H_0 apporte la même information que celui sous H_1, alors on ne rejettera pas H_0. C'est une application du principe de parcimonie. On remarque que plus $\hat{\sigma}^2$ sera petit et plus on rejettera H_0 aisément et inversement. Ce test est un rapport signal sur bruit (SNR : "Signal-to-Noise Ratio").

La statistique $\hat{F}$ est définie comme le rapport de deux variables de khi-deux indépendantes divisées par leurs degrés de liberté respectifs. Cette statistique est donc bien définie pour effectuer un test de rejet de H_0. En effet, on a $\hat{Y}-\hat{Y}_0 = (P_E - P_{E_0})Y = (I - P_{E_0})P_E Y$ par le théorème des trois perpendiculaires. Or on remarque que $(I - P_{E_0})P_E Y = P_{E\cap E_0^\perp} Y$. Sous H_0 on a donc $P_{E\cap E_0^\perp} Y \overset{H_0}{=} P_{E\cap E_0^\perp}(X_0\beta_0 + \varepsilon) = P_{E\cap E_0^\perp}\varepsilon$ car $X_0\beta_0 \in E_0$. Par le théorème de Cochran on a donc $\|P_{E\cap E_0^\perp}\varepsilon\|_2^2 = \|\hat{Y} - \hat{Y}_0\|_2^2 \sim \sigma^2\chi^2(q)$ car $tr(P_{E\cap E_0^\perp}) = tr(P_E) - tr(P_{E_0}) = rang(X) - rang(X_0) = k - k_0 = q$.

De même, sous H_0, on a $Y - \hat{Y} = (I - P_E)Y = P_{E^\perp}Y = P_{E^\perp}\varepsilon$, car $E_0 \subset E$, et $tr(P_{E^\perp}) = tr(I) - tr(P_E) = n - k$. Par application du théorème de Cochran on a donc $\|Y - \hat{Y}\|_2^2 \sim \sigma^2\chi^2(n-k)$. Pour montrer l'indépendance du numérateur et du dénominateur il suffit de remarquer que $\hat{Y} - \hat{Y}_0 \in E \cap E_0^\perp$ et $Y - \hat{Y} \in E^\perp$, c'est à dire que $\hat{Y} - \hat{Y}_0$ et $Y - \hat{Y}$ sont des éléments d'espaces orthogonaux. Donc leur produit scalaire, ou leur covariance, est nul. Puisque tout est gaussien on a que ces éléments sont bien indépendants. En divisant le numérateur et le dénominateur par leurs degrés de liberté respectifs on a que $\hat{F}$ suit bien, par définition, la loi de Fisher $F(q, n-k)$.

6.3.2 Les tests F pour l'association et la liaison

On a vu précédemment que le test F nécessite l'hypothèse de sphéricité (i.e. d'homoscédasticité) des résidus, c'est à dire que la matrice Ω de variance-covariance du vecteur des résidus soit de la forme $\Omega = \sigma^2 I_n$. Si l'hypothèse de sphéricité n'est pas respectée, comme pour le modèle de liaison hétéroscédastique par exemple, alors $\hat{F}$ ne suit pas une loi de Fisher pour la norme $\|.\|_2$ appliquée à $\hat{Y} - \hat{Y}_0$ et $Y - \hat{Y}$, mais seulement pour la norme $\|.\|_{\Omega^{-1}}$.

En effet, pour démontrer cette propriété il suffit d'appliquer une décomposition de Cholesky à Ω sachant que, par définition d'une matrice de variance-covariance, elle est semi-définie positive. On peut donc écrire $\Omega = \sigma^2 LL'$ où L est une matrice triangulaire inférieure de plein rang. Si $Y = X\beta + \epsilon$, où $Y \sim \mathcal{N}_n(X\beta, \Omega)$ et $\epsilon \sim \mathcal{N}_n(0, \Omega)$, alors on a $Y = X\beta + \epsilon \iff Y^* = X^*\beta + \epsilon^*$ où $Y^* = L^{-1}Y$, $X^* = L^{-1}X$ et $\epsilon^* = L^{-1}\epsilon$. On a donc $\mathbb{E}(Y^*) = X^*\beta$ et $Var(Y^*) = L^{-1}Var(\epsilon)(L^{-1})' = L^{-1}\sigma^2 LL'(L^{-1})' = \sigma^2 I_n$, c'est à dire que $Y^* \sim \mathcal{N}_n(X^*\beta, \sigma^2 I_n)$. Ainsi on se ramène au cas de résidus sphériques et on montre de même, sous H_0, que $Y^* \sim \mathcal{N}_n(X_0^*\beta_0, \sigma^2 I_n)$. Sachant que $vect(X_0^*) \subset vect(X_1^*)$, et en

appliquant le théorème de Cochran, on a :

$$\hat{F} = \frac{\|\hat{Y}^* - \hat{Y}_0^*\|_2^2/q}{\|Y^* - \hat{Y}^*\|_2^2/n-k} = \frac{\|X^*\hat{\beta} - X_0^*\hat{\beta}_0\|_2^2/q}{\|Y^* - X^*\hat{\beta}\|_2^2/n-k} \underset{H_0}{\sim} F(q, n-k)$$

où $\hat{\beta} = (X^{*'}X^*)^{-1}X^{*'}Y^* = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$ et $\hat{\beta}_0 = (X_0^{*'}X_0^*)^{-1}X_0^{*'}Y^* = (X_0'\Omega^{-1}X_0)^{-1}$ $\times X_0'\Omega^{-1}Y$ sont les estimateurs des moindres carrés généralisés, pour la métrique Ω, sous H_1 et H_0 respectivement. Or on remarque aisément que :

$$\begin{cases} \|\hat{Y}^* - \hat{Y}_0^*\|_2^2 = (L^{-1}(X\hat{\beta} - X_0\hat{\beta}_0))'(L^{-1}(X\hat{\beta} - X_0\hat{\beta}_0)) = \sigma^2(\hat{Y} - \hat{Y}_0)'\Omega^{-1}(\hat{Y} - \hat{Y}_0) \\ \|Y^* - \hat{Y}^*\|_2^2 = (L^{-1}(Y - X\hat{\beta}))'(L^{-1}(Y - X\hat{\beta})) = \sigma^2(Y - \hat{Y})'\Omega^{-1}(Y - \hat{Y}) \end{cases}$$

Il vient donc que $\hat{F}$ peut se ré-écrire de la façon suivante :

$$\hat{F} = \frac{(\hat{Y} - \hat{Y}_0)'\Omega^{-1}(\hat{Y} - \hat{Y}_0)/q}{(Y - \hat{Y})'\Omega^{-1}(Y - \hat{Y})/n-k} = \frac{\|\hat{Y} - \hat{Y}_0\|_{\Omega^{-1}}^2/q}{\|Y - \hat{Y}\|_{\Omega^{-1}}^2/n-k} \underset{H_0}{\sim} F(q, n-k)$$

Cette ré-écriture de $\hat{F}$ pour la norme $\|.\|_{\Omega^{-1}}$ a l'avantage, par exemple, de permettre de calculer $\hat{F}$ sans passer par la décomposition de Cholesky de Ω. L'hypothèse faite dans les modèles d'association et les modèles de liaison est que cette matrice est connue. Ω est toujours semi-définie positive dans le cadre des modèles statistiques présentés, ce qui justifie la construction du test F généralisé pour ces modèles.

En effet, Ω est semi-définie positive pour les analyses où elle est diagonale positive (i.e. les cas homoscédastiques et hétéroscédastique considérés). La matrice J_m intervenant dans l'analyse d'association corrigée est toujours semi-définie positive, car $x^{'}J_m x \geq 0$ pour tout vecteur x de taille m. Or, une matrice diagonale par blocs est semi-définie positive si et seulement si chacun des blocs est semi-défini positif, et la somme de matrices semi-définies positives est également semi-définie positive, d'où l'on montre la semi-définie positivité pour Ω dans le modèle d'association corrigé.

6.3.2.1 Le test F pour l'association

Les éléments du modèle (6.1) sous H_0 sont donnés par $X_{0A} = \mathbb{1}_n$ et $\beta_{0A} = \mu$. La statistique de test $\hat{F}_A$ associée au modèle (6.1) s'écrit donc :

$$\hat{F}_A = \frac{(X_A\hat{\beta}_A - X_{0A}\hat{\beta}_{0A})'V^{-1}(X_A\hat{\beta}_A - X_{0A}\hat{\beta}_{0A})/1}{(Y_A - X_A\hat{\beta}_A)'V^{-1}(Y_A - X_A\hat{\beta}_A)/n-2} \underset{H_0}{\sim} F(1, n-2)$$

où

$$\begin{cases} \hat{\beta}_A = (X_A'V^{-1}X_A)^{-1}X_A'V^{-1}Y_A \\ \hat{\beta}_{0A} = (X_{0A}'V^{-1}X_{0A})^{-1}X_{0A}'V^{-1}Y_A \end{cases}$$

6.3.2.2 Le test F pour la liaison

Les éléments du modèle (6.2) sous H_0 sont donnés par $X_{0T}' = (X_{T\mu}^{(1)'}, .., X_{T\mu}^{(pe)'})$ et $\beta_{0T}' = (\mu_1, .., \mu_{pe})$. La statistique de test $\hat{F}_T$ associée au modèle (6.2) s'écrit donc :

$$\hat{F}_T = \frac{(X_T\hat{\beta}_T - X_{0T}\hat{\beta}_{0T})'\tilde{V}^{-1}(X_T\hat{\beta}_T - X_{0T}\hat{\beta}_{0T})/pe}{(Y_T - X_T\hat{\beta}_T)'\tilde{V}^{-1}(Y_T - X_T\hat{\beta}_T)/ne-2pe} \underset{H_0}{\sim} F(pe, ne-2pe)$$

où

$$\begin{cases} \hat{\beta}_T = (X_T'\tilde{V}^{-1}X_T)^{-1}X_T'\tilde{V}^{-1}Y_T \\ \hat{\beta}_{0T} = (X_{0T}'\tilde{V}^{-1}X_{0T})^{-1}X_{0T}'\tilde{V}^{-1}Y_T \\ ne = \frac{mpe}{2} \end{cases}$$

6.3.3 Le seuil de rejet empirique et théorique

6.3.3.1 Le seuil de rejet de H_0 associé à un test

Lorsque la distribution sous H_0 d'une certaine statistique de test $\hat{ST}$ n'est pas connue il convient de pouvoir la simuler empiriquement. Une approche possible pour ce faire, en pratique, revient à permuter les éléments du vecteur Y des phénotypes tout en conservant la structure généalogique et les relations entre les marqueurs sur les chromosomes (Churchill et Doerge, 1994). Ainsi on peut calculer des valeurs de $\hat{ST}$ pour des permutations différentes, en se plaçant sous l'hypothèse nulle de non-association entre phénotype et marqueur, afin de produire une distribution empirique de $\hat{ST}$ sous H_0. Une fois cette distribution obtenue, et ordonnée, il suffit alors de calculer le quantile $\hat{ST}_{\underline{\alpha}}$ associé à $100(1-\underline{\alpha})$ % de la distribution (avec $\underline{\alpha} \in]0,1[$) afin de définir un seuil empirique de rejet

pour H_0. On rejettera donc H_0, avec $100\underline{\alpha}$ % de chance de se tromper, si la valeur de $\hat{ST}$ obtenue sans permutation des phénotypes est supérieure au seuil $\hat{ST}_{\underline{\alpha}}$.

6.3.3.2 Le seuil de rejet de H_0 associé au test F

Pour la statistique de test $\hat{F}$ il n'est pas nécessaire de simuler sa loi sous H_0 afin de connaître le seuil de rejet, sauf sous des conditions assez restrictives, car cette loi est pratiquement toujours connue. En effet, il existe un résultat théorique qui justifie l'utilisation du test F dans la majorité des cas (i.e. normalité ou non normalité des observations), pour des effectifs suffisamment grands en pratique (Huber, 1981 ; Azaïs et Bardet, 2006).

Ce résultat peut s'énoncer succinctement de la façon suivante. Lorsque les résidus (et les observations) sont gaussiens, $\hat{F}$ converge en loi asymptotiquement vers $\frac{1}{q}\chi^2(q)$ sous H_0, par application du théorème de Slutsky, car le dénominateur converge en probabilité vers la loi dégénérée constante et égale à 1. Or, cette loi limite sous H_0 est toujours la même, avec une condition peu restrictive sur le projecteur P_E, lorsque les résidus (et les observations) ne sont pas gaussiens. Ce résultat théorique, que l'on ne démontre pas ici, affirme que $\hat{F}$ n'a pas besoin de l'hypothèse de gaussianité, sauf cas extrêmes, pour être approximativement exact (Huber, 1981).

Afin d'illustrer ce résultat, Bonnet et Lansiaux (1992) ont étudié le comportement de $\hat{F}$ en analyse de la variance à un facteur à 2, 5 et 10 niveaux, en utilisant divers types de lois non normales, avec un effectif allant de 4 à 80 pour chaque expérience. La validité de $\hat{F}$ a été évaluée, dans leur étude, en comparant le niveau de signification réel au niveau théorique $\underline{\alpha}$ du test pour $\underline{\alpha}$ =0.1, 0.05 et 0.01. L'étude a montré que l'on observe une déviation au niveau théorique du test seulement lorsque toutes les conditions suivantes sont réunies :

- dispositifs déséquilibrés
- petits échantillons
- loi dissymétrique

Leur étude a montré que, dans tous les autres cas, tout se passe comme si les données étaient gaussiennes. Ainsi le seuil de rejet de H_0 pour $\hat{F}$ obtenu à partir de la distribution empirique sous H_0 coïncide, dans la majorité des cas, avec le seuil obtenu à partir de la distribution théorique car ces deux distributions coïncident. Le paragraphe suivant donne

des exemples de comparaison de distributions empiriques (simulées) de $\hat{F}$ sous H_0, avec la distribution théorique (tabulée) correspondante, pour des observations (que l'on suppose être des phénotypes) issues de lois normales et non normales.

6.3.3.2.1 Exemples de convergence de la distribution empirique vers la distribution théorique

6.3.3.2.1.1 Cas gaussien

La figure 6.1 illustre des distributions de phénotypes simulés pour 1000 individus (20 pères et 50 descendants par père) lors d'une simulation, que l'on utilise afin de calculer des distributions empiriques associées aux statistiques de test d'association et de liaison sous H_0 (figures 6.2 et 6.3). Les phénotypes ont été simulés sous H_0 comme un mélange de plusieurs gaussiennes, respectivement d'espérances nulles et de variances égales à 0.5, 0.25 et 1, selon un schéma polygénique que l'on verra dans le chapitre 7 (cf.7.2.2.1). Les génotypes des 1000 individus ont été simulés suivant un schéma classique sous HWE (cf. 7.2.1) avec des fréquences équilibrées au marqueur (i.e. $f_{a_1} = f_{a_2} = 0.5$).

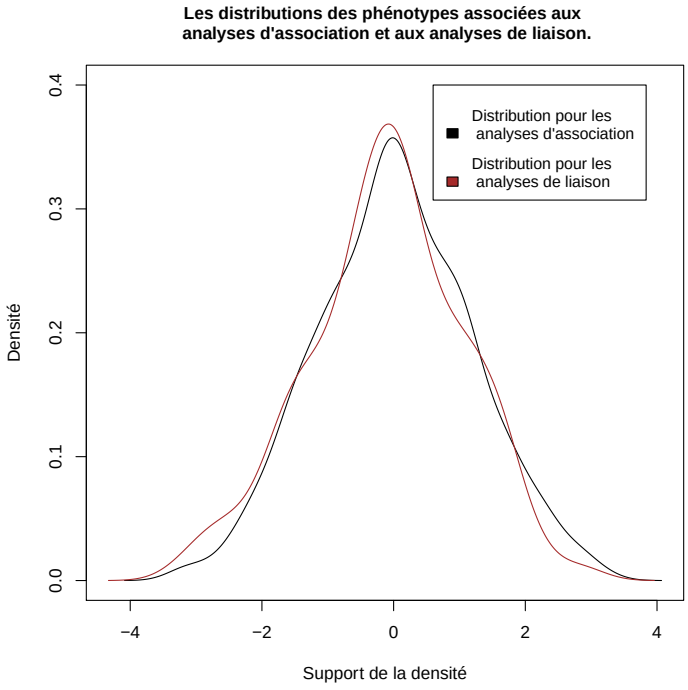


FIGURE 6.1 – Distributions générées comme un mélange de gaussiennes de moyennes nulles et de variances 0.5, 0.25 et 1.

Les phénotypes retenus pour la distribution de Fisher associée à l'analyse de liaison

homoscédastique, dans la figure 6.3, proviennent uniquement de descendants homozygotes, issus de pères hétérozygotes, parmi les 1000 individus (9 pères et 225 individus homozygotes). Les paramètres σ_u^2 et σ_ε^2 utilisés dans le calcul des statistiques de test sont supposés connus, et fixés ici à 0.5 et 1, selon un a priori par rapport au schéma de génération des phénotypes défini en 7.2.2.1. Le nombre de permutations utilisées pour approcher les distributions empiriques sous H_0 est fixé à 10000. A noter qu'ici les phénotypes n'ont pas été simulés en considérant une hétérogénéité de variances résiduelles inter familles, ces paramètres de variances n'interviennent donc pas dans les statistiques de test. On voit, dans les figures 6.2 et 6.3, que les distributions empiriques des statistiques de test convergent vers les distributions théoriques correspondantes par rapport aux degrés de libertés : $(1{,}1000-2)$ pour les analyses d'association et $(9,\ 225-18)$ pour l'analyse de liaison).

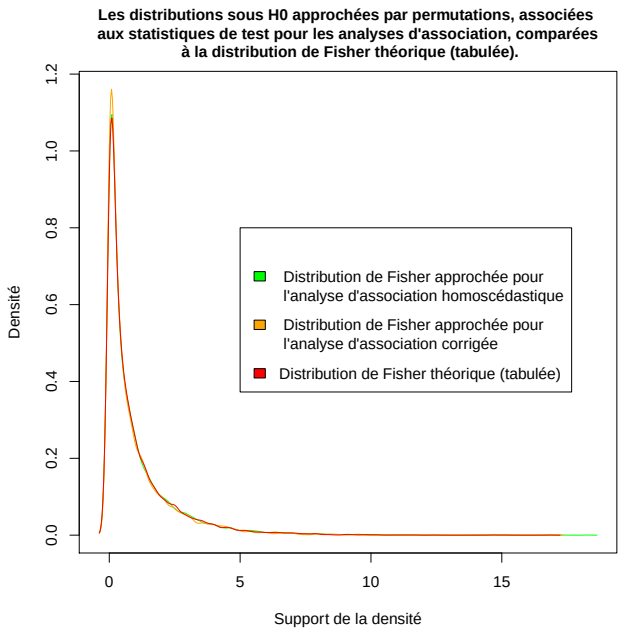


FIGURE 6.2 – Distributions empiriques et distribution théorique, pour les analyses d'association, dans le cas gaussien.

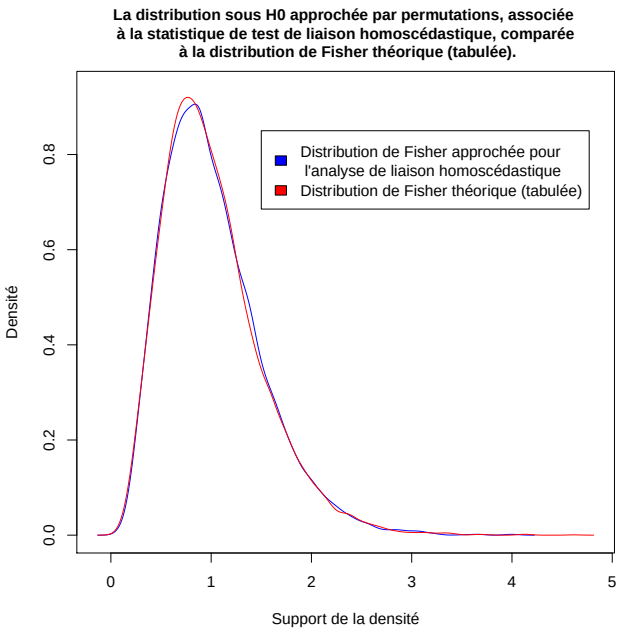


FIGURE 6.3 – Distribution empirique et distribution théorique, pour l'analyse de liaison homoscédastique, dans le cas gaussien.

6.3.3.2.1.2 Cas non-gaussien

La démarche et les simulations sont ici les mêmes que dans le cas gaussien, excepté que les phénotypes des 1000 individus sont générés dans une loi géométrique de paramètre égal à 0.5. La figure 6.4 illustre les distributions des phénotypes simulés et utilisés pour calculer des distributions empiriques associées aux statistiques de test sous H_0.

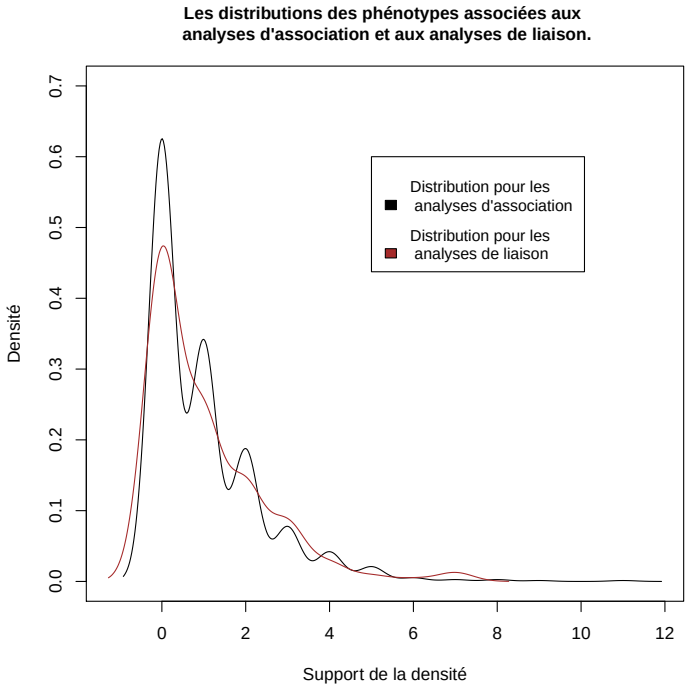


FIGURE 6.4 – Distributions générées à partir d'une loi géométrique de paramètre égale à 0.5.

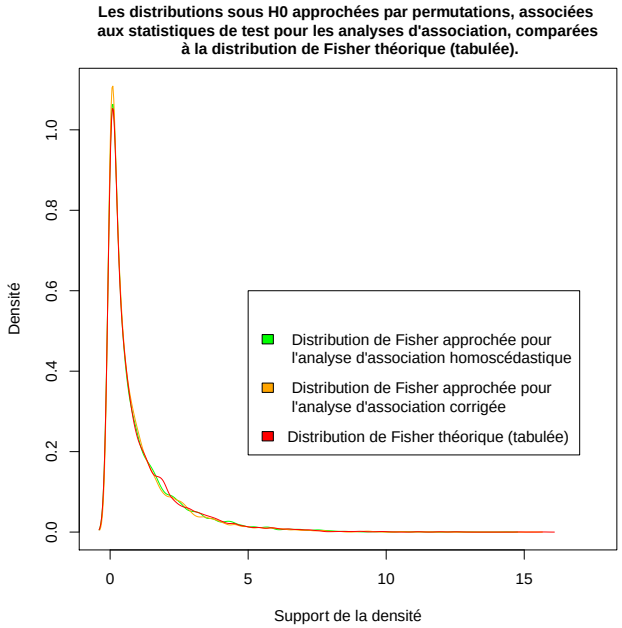


FIGURE 6.5 – Distributions empiriques et distribution théorique, pour les analyses d'association, dans le cas non-gaussien.

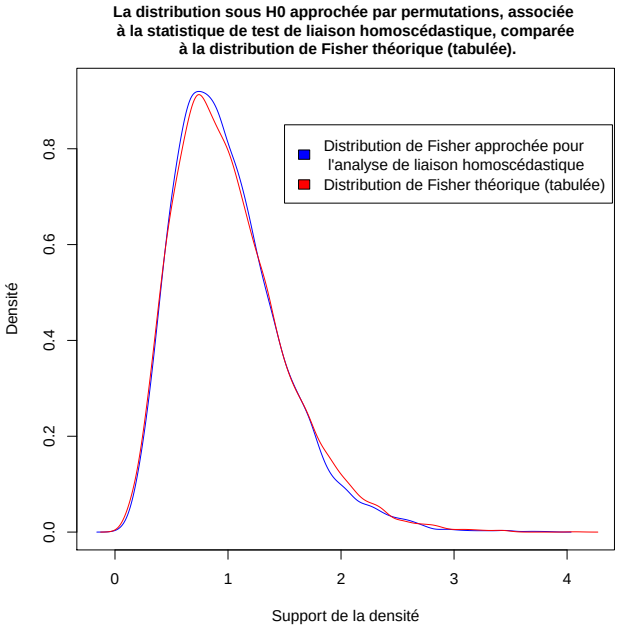


FIGURE 6.6 – Distribution empirique et distribution théorique, pour l'analyse de liaison homoscédastique, dans le cas non-gaussien.

On voit clairement, pour cet exemple caricatural, que les distributions associées aux phénotypes ne sont pas gaussiennes. On remarque toutefois, dans les figures 6.5 et 6.6,

que les distributions empiriques des statistiques de test convergent vers les distributions théoriques correspondantes (par rapport aux degrés de libertés) bien que les données ne suivent pas des lois normales.

6.3.3.3 Validité du seuil de rejet de H_0 associé au test F

L'utilisation du seuil théorique, ou empirique, pour le rejet de H_0 n'est valable que si $\hat{F}$ suit bien une loi de Fisher de facteur de décentrage nul sous H_0 (cf. 6.3.4). En effet, il peut arriver que ce facteur soit non nul, dû à une structure inhérente aux données (non prise en compte de la structure familiale, hétérogénéité de moyennes, dispositifs déséquilibrés...), bien que l'on soit réellement sous l'hypothèse nulle d'absence d'effet d'un QTL. Si tel est le cas on observe alors fréquemment une inflation du taux d'erreur de première espèce, par rapport au risque théorique $\underline{\alpha}$ fixé, dans l'évaluation de ce taux pour un ensemble de simulations.

Ce phénomène provient du fait que $\hat{F}$ dépassera le seuil de rejet défini sous H_0 avec une probabilité plus grande que $\underline{\alpha}$, lorsque le facteur de décentrage est non nul, bien que l'hypothèse H_1 ne soit pas vraie dans ce cas. L'exemple suivant illustre ce phénomène en reprenant l'exemple des phénotypes simulés sous H_0 comme une combinaison de plusieurs gaussiennes, vu précédemment en 6.3.3.2.1.1, mais avec une composante supplémentaire distribuée dans une $\mathcal{N}(0, 2)$. On impose de plus que la réalisation associée à cette composante soit fixée par famille et différente entre les familles. Ce schéma de simulation correspond à celui décrit en 7.2.2.3, que l'on verra dans le chapitre 7, pour lequel on simule une hétérogénéité de moyennes inter familles créée par ce schéma.

Les figures 6.7 et 6.8 illustrent les distributions des statistiques de test calculées, et théoriques, pour 10000 populations simulées sous H_0 dans le cadre de ce schéma de génération de données. Les statistiques de test sont calculées ici sur chacune des populations, sans notion de permutation, afin d'obtenir les distributions des $\hat{F}$ présentées dans ces figures. Ces figures illustrent également, par des traits verticaux pointillés, les $(1 - \underline{\alpha})$-quantiles associés à ces distributions pour $\underline{\alpha} = 0.01$. Comme précédemment, les paramètres σ_u^2 et σ_ε^2 sont non estimés et fixés à 0.5 et 1.

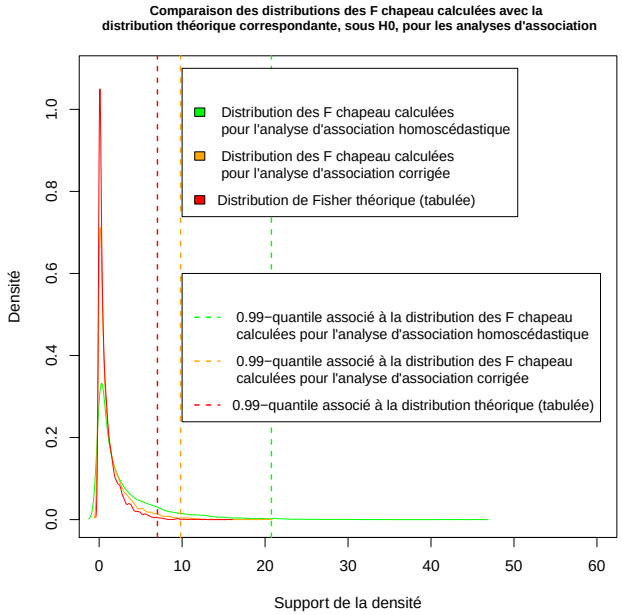


FIGURE 6.7 – Distributions des $\hat{F}$ calculées et la distribution théorique correspondante sous H_0, pour les analyses d'association, en cas d'hétérogénéité de moyennes.

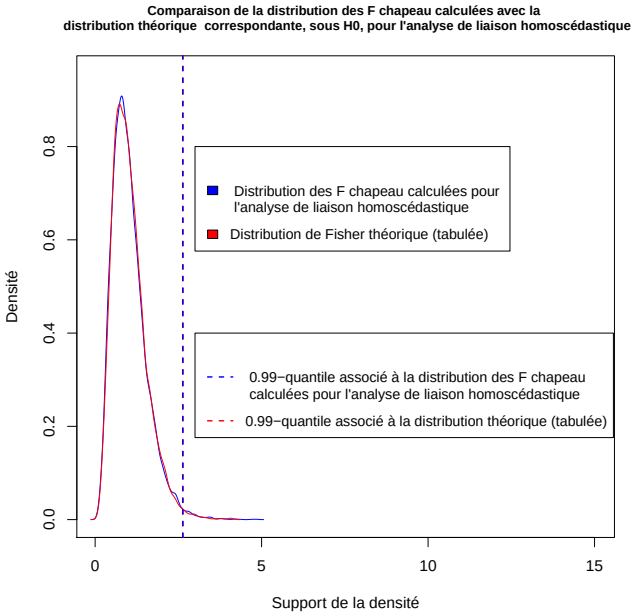


FIGURE 6.8 – Distribution des $\hat{F}$ calculées et la distribution théorique correspondante sous H_0, pour l'analyse de liaison homoscédastique, en cas d'hétérogénéité de moyennes.

On voit dans la figure 6.7 que les $\hat{F}$ associées aux analyses d'association ont une probabilité supérieure à 0.01, de dépasser le 0.99-quantile de la distribution théorique correspondante sous H_0, à cause d'un décentrement des distributions calculées. Le modèle d'analyse de liaison homoscédastique est quant à lui, par construction, en adéquation avec ce schéma de simulation générant une hétérogénéité de moyennes entre les familles. On voit dans la figure 6.8 que le 0.99-quantile de la distribution des $\hat{F}$ calculées coïncide avec celui de la distribution théorique et que l'on a donc un bon contrôle de l'erreur de première espèce pour ce cas de simulation.

On notera que la permutation des phénotypes, pour une population unique en pratique, sert à simuler la distribution empirique associée à une statistique de test dont la loi n'est pas connue sous H_0 (cf. 6.3.3.1). De plus, la permutation des données, à une population fixée, n'aidera pas à avoir un meilleur contrôle du taux d'erreur de première espèce, ici, pour les analyses d'association. En effet, la permutation des données place les statistiques de test sous H_0 qui, sous cette hypothèse, convergent en loi vers les lois théoriques respectives (cf. 6.3.3.2). Or, la valeur de $\hat{F}$ calculée sans permutation des phénotypes reste, dans le cadre d'une hétérogénéité de moyennes, une réalisation d'une loi

décentrée.

On verra dans le chapitre 7 qu'un choix plus judicieux que des valeurs choisies avec un a priori, telles que 0.5 et 1 pour les paramètres σ_u^2 et σ_ε^2, peut permettre de tenir compte de l'hétérogénéité de moyennes pour l'analyse d'association corrigée, en ayant une influence sur le facteur de décentrage associé, et que ce choix mène à un meilleur contrôle de l'erreur de première espèce et un même niveau de puissance (cf. 7.3.2, *iii*).

6.3.4 Le facteur de décentrage, la puissance et le risque associés au test F

Pour des résidus sphériques la loi de $\|\hat{Y} - \hat{Y}_0\|_2^2$ sous H_1 n'est plus $\sigma^2\chi^2(q)$, qui est un khi-deux centré, mais $\sigma^2\chi^2(q,\lambda)$ qui est un khi-deux décentré de facteur λ de décentrage donné par $\lambda = \frac{1}{\sigma^2}\|X\beta - P_{E_0}X\beta\|_2^2$ (Guyon, 2005). En effet, on a vu précédemment que $\|\hat{Y} - \hat{Y}_0\|_2^2 = \|P_{E\cap E_0^\perp}Y\|_2^2$ (cf 6.3.1). Sous H_1 on a donc $\|P_{E\cap E_0^\perp}Y\|_2^2 = Y'P_{E\cap E_0^\perp}Y$ où $Y \overset{H_1}{\sim} \mathcal{N}_n(X\beta, I_n\sigma^2)$. Or, par définition on sait que $\frac{Y'P_{E\cap E_0^\perp}Y}{\sigma^2}$ est une forme quadratique suivant sous H_1 une loi du khi-deux de facteur de décentrage égal à $\frac{1}{\sigma^2}(X\beta)'P_{E\cap E_0^\perp}X\beta$, et de degré de liberté égal à $tr(P_{E\cap E_0^\perp}) = tr(P_E - P_{E_0}) = q$ (cf. 6.3.1). Ce facteur de décentrage peut également se ré-écrire sous la forme $\frac{1}{\sigma^2}(X\beta)'P_{E\cap E_0^\perp}X\beta = \frac{1}{\sigma^2}(X\beta)'(P_E - P_{E_0})X\beta = \frac{1}{\sigma^2}((P_E - P_{E_0})X\beta)'(P_E - P_{E_0})X\beta = \frac{1}{\sigma^2}\|X\beta - P_{E_0}X\beta\|_2^2$ (car $X\beta \in E$), d'où l'on obtient la formulation énoncée plus haut. Sous H_1 la statistique $\hat{F}$ suit donc par définition une loi de Fisher décentrée $F(q, n-k, \lambda)$.

Le paramètre λ sera d'autant plus grand, lorsque H_1 sera vraie, que l'effet du QTL sera plus prononcé. La puissance $1 - \beta_{\hat{F}}$ associée à $\hat{F}$ sous H_1, et définie pour un seuil $f_{\underline{\alpha}}(q, n-k)$ de rejet de H_0, est donnée par :

$$1 - \beta_{\hat{F}} = P(\{\text{rejeter } H_0 | H_1 \text{ est vraie}\}) = P(\{\hat{F} \geq f_{\underline{\alpha}}(q, n-k)\})$$

où $f_{\underline{\alpha}}(q, n-k)$ est le $(1-\underline{\alpha})$-quantile de la loi de Fisher $F(q, n-k)$ sous H_0, et $\beta_{\hat{F}}$ est le risque d'erreur de deuxième espèce (faux négatif). La puissance associée à $\hat{F}$ est une fonction monotone croissante du paramètre λ. Il suffit donc de connaître le comportement de ce paramètre afin de caractériser le comportement de la puissance. Le paramètre λ peut donc aider, tel qu'on le verra en 7.3.1 , à comprendre l'influence de différents

facteurs (fréquences alléliques, paramètres de variance,...) sur la puissance. On notera que l'expression du facteur de décentrage dans le cas de non-sphéricité des résidus est obtenue, comme pour la construction des tests F, en considérant une décomposition de Cholesky de la matrice V ou $\tilde{V}$ (cf. 6.3.2). Dans ce cas on a $\lambda = \frac{1}{\sigma^2}\|X^*\beta - P_{E_0^*}X^*\beta\|_2^2$, où $P_{E_0^*} = X_0^*(X_0^{*'}X_0^*)^{-1}X_0^{*'}$ avec $V = \sigma^2 LL'$. Ce paramètre, après simplification, se ré-écrit également sous la forme $\lambda = \beta' X' V^{-1} X\beta - \beta' X' V^{-1} X_0 (X_0' V^{-1} X_0)^{-1} X_0' V^{-1} X\beta$. Cette formulation de λ sera utilisée en 7.3.1.

On remarque cependant que la formulation de la puissance, à partir de la loi décentrée, n'est pas toujours d'un intérêt pratique. En effet, le paramètre λ ne renseigne pas sur la façon dont les données sont générées et il ne permet donc pas, en l'occurrence, de connaître les situations favorables (ou non) aux différents modèles statistiques considérés. Ainsi la simulation de données, dans ce contexte, s'avère fondamentale et incontournable dans la comparaison de ces différents modèles. Par ailleurs, on remarque aussi qu'il n'est pas toujours évident d'avoir une estimation fiable de $1-\beta_{\hat{F}}$ pour un ensemble de simulations.

En effet, chaque réalisation de $\hat{F}$ à chaque simulation sous H_1, pour un modèle d'analyse de liaison par exemple, peut être obtenue à partir d'une loi de Fisher décentrée différente (ayant des degrés de libertés différents) dû au nombre de pères hétérozygotes qui change d'une simulation à l'autre. Pour ce genre de situation on peut néanmoins approximer $1-\beta_{\hat{F}}$ par une méthode de Monte-Carlo, car par définition on a :

$$1-\beta_{\hat{F}} = P(\{\text{rejeter } H_0 | H_1 \text{ est vraie}\}) = \mathbb{E}\Big[\mathbb{1}_{\{rejeter\ H_0|H_1\ est\ vraie\}}\Big]$$

On peut donc approximer $1-\beta_{\hat{F}}$ en approximant cette espérance. Il suffit donc de regarder le nombre de fois, sur l'ensemble des alternatives différentes, où $\hat{F}$ dépasse $f_{\underline{\alpha}}(q, n-k)$ afin de donner une approximation de la puissance. De la même façon, on pourra approximer le risque réel (i.e. le taux observé) d'erreur de première espèce en approximant $\mathbb{E}\Big[\mathbb{1}_{\{rejeter\ H_0|H_0\ est\ vraie\}}\Big]$.

Chapitre 7

Comparaison des puissances et des robustesses associées aux modèles uni-SNP d'association et uni-SNP de liaison

7.1 Cadre de l'étude

Ce chapitre compare les puissances et les robustesses des modèles d'association (corrigé et homoscédastique) avec celles des modèles de liaison (hétéroscédastique et homoscédastique) présentés dans le chapitre 6. Pour ce faire, on considère différents cadres de simulations pour les phénotypes et les génotypes au QTL. On note pour ce chapitre que, quelque soit le schéma de simulation des phénotypes pour une population, le schéma de simulation des génotypes au marqueur testé est toujours le même. On suppose classiquement HWE pour la simulation des génotypes au marqueur testé. Selon les situations ce marqueur est identifiable, ou non, au QTL.

7.2 Les schémas de simulation

7.2.1 Le schéma de simulation des génotypes

Le schéma que l'on utilise pour simuler les génotypes au sein des familles dans une population, basé sur les règles classiques de la transmission, est le suivant :

1. On fixe les fréquences alléliques f_{a_1} et $f_{a_2} = 1 - f_{a_1}$ au marqueur testé
2. On génère les génotypes des pères en supposant HWE. Soit $G^i(S) \to \{a_1a_1, a_1a_2, a_2a_2\}$ la variable aléatoire associée au génotype du père i ($i \in \{1, .., p\}$), où $S \sim \mathcal{U}([0, 1])$ dont

la réalisation est notée s. $G^i(S)$ est définie de la façon suivante :

$$G^i(S) = \begin{cases} a_1a_1 & \text{si } 0 \leq s \leq f_{a_1}^2 \\ a_1a_2 & \text{si } f_{a_1}^2 < s \leq 2f_{a_1}(1-f_{a_1}) + f_{a_1}^2 \\ a_2a_2 & \text{si } 2f_{a_1}(1-f_{a_1}) + f_{a_1}^2 < s \leq 1 \end{cases}$$

3. On génère les génotypes des descendants des pères de la façon suivante. Soit $G_k^i(W) \rightarrow \{a_1a_1, a_1a_2, a_2a_2\}$ la variable aléatoire associée au génotype du descendant k ($k \in \{1, ..., m\}$) du père i, où $W \sim \mathcal{U}([0,1])$ dont la réalisation est notée w. $G_k^i(W)$ est définie de la façon suivante :

i) Si $G^i(S) = a_1a_1$ alors :

$$G_k^i(W) = \begin{cases} a_1a_1 & \text{si } 0 \leq w \leq f_{a_1} \\ a_1a_2 & \text{si } f_{a_1} < w \leq 1 \end{cases}$$

ii) Si $G^i(S) = a_1a_2$ alors :

$$G_k^i(W) = \begin{cases} a_1a_1 & \text{si } 0 \leq w \leq \frac{1}{2}f_{a_1} \\ a_1a_2 & \text{si } \frac{1}{2}f_{a_1} < w \leq \frac{1}{2}(1-f_{a_1}) + \frac{1}{2}f_{a_1} + \frac{1}{2}f_{a_1} \\ a_2a_2 & \text{si } \frac{1}{2}(1-f_{a_1}) + \frac{1}{2}f_{a_1} + \frac{1}{2}f_{a_1} < w \leq 1 \end{cases}$$

iii) Si $G^i(S) = a_2a_2$ alors :

$$G_k^i(W) = \begin{cases} a_1a_2 & \text{si } 0 \leq w \leq f_{a_1} \\ a_2a_2 & \text{si } f_{a_1} < w \leq 1 \end{cases}$$

De cette façon, les génotypes des mères ne sont pas directement simulés. Dans la suite de ce travail, à l'exception des dérivations algébriques, on suppose que les fréquences alléliques sont équilibrées au marqueur/QTL pour les simulations.

7.2.2 Les schémas de simulation des phénotypes

Le modèle polygénique suppose que la valeur génétique u d'un individu est la somme des effets d'un grand nombre de gènes (Fisher, 1918). Sous cette hypothèse le phénotype

y de l'individu se décompose, classiquement, de la façon suivante :

$$y = u + \varepsilon = \frac{1}{2}p^{fa.} + \frac{1}{2}p^{mo.} + \phi + g + \varepsilon$$

où $p^{fa.}$ et $p^{mo.}$ sont respectivement les valeurs génétiques, ou polygéniques hors effet du QTL, du père et de la mère. On suppose donc que l'individu reçoit la moitié de la valeur génétique de chaque parent. Dans cette décomposition ϕ représente un aléas, dû à la méiose, qui explique que deux pleins frères peuvent ne pas être identiques. Cet aléa est supposé d'espérance nulle et de variance égale à la moitié de la variance génétique. Finalement g représente l'effet du génotype au QTL, s'il existe, et ε représente un résidu gaussien que l'on suppose centré.

Ce schéma de décomposition s'avère assez robuste pour le calcul des valeurs génétiques, en sélection animale, bien qu'il ne soit pas totalement biologiquement correct à cause des interactions sur le génome (Guillaume, 2009). Dans la suite de ce chapitre on utilise différents schémas de simulation pour les phénotypes, basés sur ce schéma polygénique, afin d'estimer les puissances et les taux d'erreur de première espèce des modèles uni-SNP d'association et uni-SNP de liaison, présentés dans le chapitre 6. On note $p_k^{mo.}$ la valeur polygénique propre à un individu k, transmise par une mère quelconque, et $p^{fa.,i}$ la valeur polygénique transmise par le père i à tous les individus dans une famille. Les variances polygénique et résiduelle sont respectivement fixées à 0.5 et 1 pour la simulation des phénotypes (i.e. $\sigma_u^2 = 0.5$ et $\sigma_\varepsilon^2 = 1$).

7.2.2.1 Effets alléliques au QTL identiques inter familles

Ce schéma correspond au cas où l'effet d'un génotype au QTL est le même intra et inter familles. Il se justifie par un allèle ayant le même effet chez tous les individus de la population. Ce schéma servira de base aux autres schémas de simulation. Il s'écrit :

$$y_k^i = \frac{1}{2}p^{fa.,i} + \frac{1}{2}p_k^{mo.} + \phi_k + g_k + \varepsilon_k$$

où ;

- y_k^i est le phénotype du descendant k dans la famille i
- p^i et $p_k^{mo.}$ sont des réalisations d'une loi $\mathcal{N}(0, 0.5)$

- ϕ_k et ε_k sont des réalisations des lois $\mathcal{N}(0, 0.25)$ et $\mathcal{N}(0, 1)$ respectivement
- $g_k = -2\alpha$ ou 0 ou 2α si le génotype au QTL est a_2a_2 ou a_1a_2 ou a_1a_1, avec $\alpha \in [0, 1]$

7.2.2.2 Variances résiduelles différentes inter familles

Ce schéma correspond au cas où l'effet d'un génotype au QTL est le même, intra et inter familles, et où la variance résiduelle est hétérogène entre les familles. Il se justifie par des environnements différents favorisant plus ou moins la dispersion des phénotypes dans les familles. Il peut aussi se justifier par un gène majeur dont les allèles sont en ségrégation dans les familles. Ce schéma de simulation s'écrit :

$$y_k^i = \frac{1}{2}p^{fa.,i} + \frac{1}{2}p_k^{mo.} + \phi_k + g_k + \varepsilon_k^{i,stand.}$$

où ;

- y_k^i est le phénotype du descendant k dans la famille i
- p^i et $p_k^{mo.}$ sont des réalisations d'une loi $\mathcal{N}(0, 0.5)$
- ϕ_k est une réalisation d'une loi $\mathcal{N}(0, 0.25)$
- $\varepsilon_k^{i,stand.} = \dfrac{\varepsilon_k^i}{\sigma_\varepsilon}$ tel que $\varepsilon_k^i \sim \mathcal{N}(0, \sigma^2_{\varepsilon^i})$, $\sigma^2_{\varepsilon^i} \sim Inv\text{-}\chi^2(1)$ et σ_ε est l'écart-type sur l'ensemble des résidus non standardisés (i.e. ε_k^i). On a donc des variances résiduelles hétérogènes entre les familles et une variance résiduelle valant 1, sur l'ensemble de la population, pour les résidus standardisés (i.e. $\varepsilon_k^{i,stand.}$)
- $g_k = -2\alpha$ ou 0 ou 2α si le génotype au QTL est a_2a_2 ou a_1a_2 ou a_1a_1, avec $\alpha \in [0, 1]$

7.2.2.3 Moyennes différentes inter familles

Ce schéma correspond au cas où l'effet d'un génotype au QTL est le même, intra et inter familles, et où l'on a une moyenne fixée pour tous les individus d'une même famille. Cet moyenne est supposée distribuée dans une loi de probabilité ayant une grande variance. Ce schéma se justifie par des différences distinctes dues à l'environnement familiale (e.g. alimentation nettement plus riche dans une famille que dans une autre). Ce schéma de simulation s'écrit :

$$y_k^i = \mu^i + \frac{1}{2}p^{fa.,i} + \frac{1}{2}p_k^{mo.} + \phi_k + g_k + \varepsilon_k$$

où ;

- y_k^i est le phénotype du descendant k dans la famille i
- μ^i est une réalisation d'une loi $\mathcal{N}(0, 2)$
- p^i et $p_k^{mo.}$ sont des réalisations d'une loi $\mathcal{N}(0, 0.5)$
- ϕ_k est une réalisation d'une loi $\mathcal{N}(0, 0.25)$
- ε_k est une réalisation d'une loi $\mathcal{N}(0, 1)$
- $g_k = -2\alpha$ ou 0 ou 2α si le génotype au QTL est a_2a_2 ou a_1a_2 ou a_1a_1, avec $\alpha \in [0, 1]$

7.2.2.4 Effets alléliques au QTL en interaction avec un locus

Ce schéma correspond au cas où l'effet d'un allèle au QTL s'exprime uniquement en présence d'un autre allèle à un autre locus. Ce schéma représente un cas d'épistasie et s'inspire du tableau 2 de l'article de Cordell (2002), où il existe seulement un allèle au QTL ayant un effet dès lors qu'un certain allèle est présent à un autre locus. Soient b_1, b_2 les allèles à ce locus et a_1, a_2 les allèles au QTL. Le schéma de simulation des génotypes au locus en interaction avec le QTL est le même que celui défini en 7.2.1. De plus, les allèles b_1, b_2 sont simulés de manière indépendante de a_1, a_2 (i.e. ils ne sont pas liés). Le schéma de simulation des phénotypes correspondant s'écrit :

$$y_k^i = \frac{1}{2}p^{fa.,i} + \frac{1}{2}p_k^{mo.} + \phi_k + g_k + \varepsilon_k$$

où ;

- y_k^i est le phénotype du descendant k dans la famille i
- p^i et $p_k^{mo.}$ sont des réalisations d'une loi $\mathcal{N}(0, 0.5)$
- ϕ_k et ε_k sont des réalisations des lois $\mathcal{N}(0, 0.25)$ et $\mathcal{N}(0, 1)$ respectivement
- $g_k = \alpha$ ou 2α (avec $\alpha \in [0, 1]$) si le génotype au QTL est a_1a_2 ou a_1a_1, respectivement, et que le génotype au locus en interaction a au moins l'allèle b_1 (i.e. b_1b_2 ou b_1b_1). Dans tous les autres cas, lorsque b_1 n'est pas présent, on a $g_k = 0$.

7.3 Principaux résultats de l'étude

Afin de simplifier la lecture, on réfère aux différents schémas définis en 7.2.2.1 à 7.2.2.4 par *i)* à *iv)* respectivement dans la suite de ce chapitre.

7.3.1 Facteurs de décentrage et puissances analytiques approchées

Cette sous-section donne les facteurs de décentrage et les puissances analytiques approchées, sous HWE au QTL, associés aux modèles statistiques décrits en 6.2.1 et 6.2.2. On rappelle que le facteur de décentrage, associé à un modèle statistique défini sous H_1, est uniquement connu sous l'hypothèse que ce modèle est "vrai". De ce fait, une puissance analytique approchée peut ne pas donner une bonne description de la vraie puissance espérée, en pratique, si le modèle statistique considéré est en mauvaise adéquation avec les processus qui génèrent les données (i.e. les génotypes et phénotypes). Remarquons que HWE repose sur plusieurs conditions, incluant parmi celles-ci que la population étudiée soit de taille infinie. Or, dans ce chapitre on considère que les populations sont de tailles finies et on verra dans la sous-section 7.3.2 que les fréquences espérées sous HWE ne sont pas toujours obtenues lors des simulations. Rappelons toutefois que le test F est asymptotiquement robuste, à la fois sous H_0 et H_1, pour de nombreuses situations (Huber, 1981 ; Tiku et Akkaya, 2004). L'intérêt de la puissance analytique est qu'elle permet de justifier la vraisemblance de la puissance obtenue par simulation, lorsqu'il y a une relativement bonne adéquation entre le modèle statistique considéré et les schémas (i.e. les processus) de génération des génotypes et phénotypes. Dans cette situation, la formulation analytique du facteur de décentrage permet éventuellement de connaître l'influence de différents facteurs tels que les fréquences, les paramètres de variances et autres sur la puissance espérée en pratique. Les dérivations algébriques associées aux facteurs de décentrage et aux estimateurs, pour les modèles statistiques comparés, sont données en 7.4.

7.3.1.1 Facteurs pour les modèles d'association

Les facteurs de décentrage associés aux modèles uni-SNP d'association (homoscédastique $\lambda_A^{homo.}$ et corrigé $\lambda_A^{corr.}$), décrits en 6.2.1, sont donnés par :

$$\begin{cases} \lambda_A^{homo.} = \dfrac{8mp\alpha^2 f_{a_1} f_{a_2}}{\sigma_\varepsilon^2} = \dfrac{8n\alpha^2 f_{a_1} f_{a_2}}{\sigma_\varepsilon^2} \\ \\ \lambda_A^{corr.} = \dfrac{8mp\alpha^2 f_{a_1} f_{a_2}}{\sigma_\varepsilon^2\left(1 + \dfrac{3\sigma_u^2}{4\sigma_\varepsilon^2}\right)} = \dfrac{8n\alpha^2 f_{a_1} f_{a_2}}{\sigma_\varepsilon^2\left(1 + \dfrac{3\sigma_u^2}{4\sigma_\varepsilon^2}\right)} \end{cases}$$

Le modèle corrigé coïncide, par définition, avec le modèle homoscédastique lorsque $\sigma_u^2 = 0$. On s'attend donc à avoir $\lambda_A^{corr.} = \lambda_A^{homo.}$ lorsque $\sigma_u^2 = 0$, ce qui est effectivement le cas.

7.3.1.2 Facteurs pour les modèles de liaisons

Les facteurs de décentrage associés aux modèles uni-SNP de liaison (homoscédastique $\lambda_T^{homo.}$ et hétéroscédastique $\lambda_T^{hétéro.}$), décrits en 6.2.2, sont donnés par :

$$\begin{cases} \lambda_T^{homo.} = \dfrac{1}{\sigma_\varepsilon^2} \displaystyle\sum_{i=1}^{pe} 2m\delta_i^2 f_{a_1} f_{a_2} \\ \\ \lambda_T^{hétéro.} = \displaystyle\sum_{i=1}^{pe} \dfrac{2m\delta_i^2 f_{a_1} f_{a_2}}{\sigma_{\varepsilon_i}^2} \end{cases}$$

où $pe = 2f_{a_1}f_{a_2}p$ est le nombre espéré de pères hétérozygotes (informatifs) pour l'analyse et $\forall i \in \{1, ..., pe\}$ $\delta_i = 2\alpha$ si l'effet α de l'allèle a_1 est le même dans toutes les familles. On sait que le modèle hétéroscédastique coïncide, par définition, avec le modèle homoscédastique lorsque $\sigma_{\varepsilon_1}^2 = \cdots = \sigma_{\varepsilon_{pe}}^2 = \sigma_\varepsilon^2$. On s'attend donc à avoir $\lambda_T^{hétéro.} = \lambda_T^{homo.}$ lorsque $\sigma_{\varepsilon_1}^2 = \cdots = \sigma_{\varepsilon_{pe}}^2 = \sigma_\varepsilon^2$, ce qui est en effet le cas.

7.3.1.3 Puissances analytiques approchées

Nous allons décrire la variance V_{g_k} expliquée par le génotype au QTL, qui est fonction de l'effet α de l'allèle a_1 et des fréquences alléliques, pour chaque schéma de simulation considéré. La variance totale des phénotypes pour les cas *i)*, *ii)* et *iv)* est donnée par $V_T = \sigma_u^2 + V_{g_k} + \sigma_\varepsilon^2$. Pour le cas *iii)* elle est donnée par $V_T = 2 + \sigma_u^2 + V_{g_k} + \sigma_\varepsilon^2$. On rappelle que α varie dans [0,1], plus précisément dans $]10^{-2}, 1]$ par pas de 10^{-2}. On suppose ici que α est le même dans toutes les familles, bien que ce ne soit pas vrai pour le schéma *iv)* par rapport aux autres schémas. Dans ce cas on a $\forall i \in \{1, ..., pe\}$ $\delta_i = 2\alpha$. Les puissances sont données pour $m = 15$ et 30 descendants par famille, un nombre de pères fixé à $p = 20$, et des fréquences alléliques équilibrées au QTL et au locus considéré en interaction pour le cas *iv)*. Le 0.99-quantile de la distribution de Fisher (centrée) correspondante a été fixé comme le seuil de rejet de H_0.

On suppose ici que les paramètres σ_u^2, σ_ε^2 et $(\sigma_{\varepsilon_i}^2)_{1 \leq i \leq pe}$ sont connus et sont les mêmes que ceux utilisés pour les schémas de génération des phénotypes. Autrement dit, on pose

$\sigma_u^2 = 0.5$, $\sigma_\varepsilon^2 = 1$ et les paramètres $(\sigma_{\varepsilon_i}^2)_{1\leq i\leq pe}$ sont les mêmes que ceux utilisés pour le schéma *ii)* de simulation (i.e. la variance résiduelle intra famille, standardisée sur la population, est donnée par $\sigma_{\varepsilon_i}^2 = \frac{\sigma_{\varepsilon^i}^2}{\sigma_\varepsilon^2}$). Les puissances analytiques approchées, associées aux facteurs de décentrage donnés précédemment en 7.3.1.1 et 7.3.1.2, sont décrites par les figures 7.1 à 7.8. Les puissances sont représentées en fonction du pourcentage de variance expliquée par le génotype au QTL, entre 0 et 10%, par rapport à la variance totale pour chacun des processus de génération de données décrits en 7.2.2. Cette représentation est nécessaire pour la comparaison de ces puissances avec celles estimées par Monte-Carlo, données dans la section 7.3.2 suivante. Remarquons que c'est le pourcentage de variance expliquée par le génotype au QTL qui change l'allure des courbes de puissance analytique, entre les cas étudiés, car les facteurs de décentrage sont supposés connus.

i) Pour le schéma avec des effets alléliques au QTL identiques inter familles (cf. 7.2.2.1) :

La variance expliquée sous HWE par le génotype au QTL pour ce schéma est donnée par :

$$\begin{aligned} V_{g_k} &= f_{a_1}^2(2\alpha)^2 + f_{a_2}^2(-2\alpha)^2 - \left(f_{a_1}^2.2\alpha - f_{a_2}^2.2\alpha\right)^2 = 8\alpha^2 f_{a_1} f_{a_2} \\ &= 2\alpha^2 \quad \text{si} \quad f_{a_1} = f_{a_2} = \frac{1}{2} \end{aligned}$$

Courbes de puissance analytique approchée :

■ : analyse d'association homoscédastique / ■ : analyse d'association corrigée

■ : analyse de liaison homoscédastique

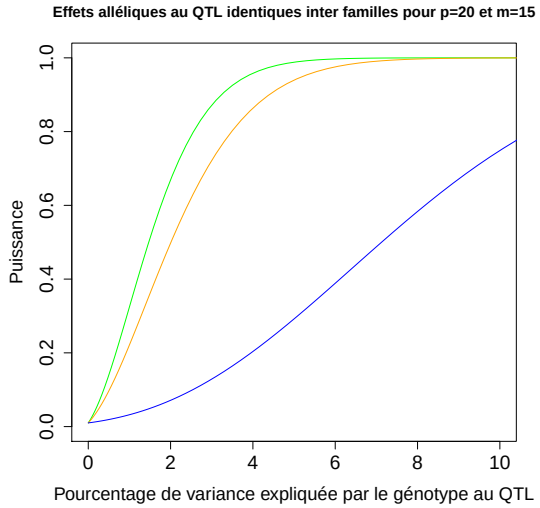


FIGURE 7.1 – Puissances analytiques approchées pour $m = 15$ dans le cas d'effets alléliques au QTL identiques inter familles.

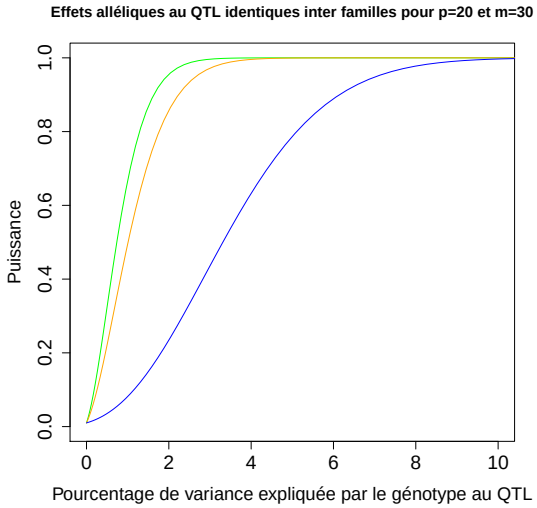


FIGURE 7.2 – Puissances analytiques approchées pour $m = 30$ dans le cas d'effets alléliques au QTL identiques inter familles.

Les figures 7.1 et 7.2 montrent une meilleure puissance pour les modèles d'association par rapport au modèle de liaison homoscédastique pour ce schéma. Cependant, on voit dans ces figures que le modèle de liaison homoscédastique peut donner la même puissance que les modèles d'association si la taille des familles augmente suffisamment. On voit également dans ces figures, de façon attendue, que les courbes associées au modèle d'association corrigé sont toujours en dessous de celles associées au modèle d'association homoscédastique.

ii) Pour le schéma avec des variances résiduelles différentes inter familles (cf. 7.2.2.2) :

La variance expliquée sous HWE par le génotype au QTL pour ce schéma est la même que pour le cas *i)*. La différence entre *i)* et *ii)* se situe sur les variances résiduelles qui sont différentes entre familles pour le cas *ii)*, bien que la variance totale des résidus soit égale à 1 pour ces deux cas. Les figures 7.3 et 7.4 montrent une meilleure puissance pour les modèles d'association par rapport au modèle de liaison hétéroscédastique pour ce schéma. Remarquons toutefois que le modèle de liaison hétéroscédastique donne ici une amélioration de la puissance, de 0.2 en moyenne, par rapport au modèle de liaison homoscédastique pour le cas *i)*.

Courbes de puissance analytique approchée :

■ : analyse d'association homoscédastique / ■ : analyse d'association corrigée

■ : analyse de liaison hétéroscédastique

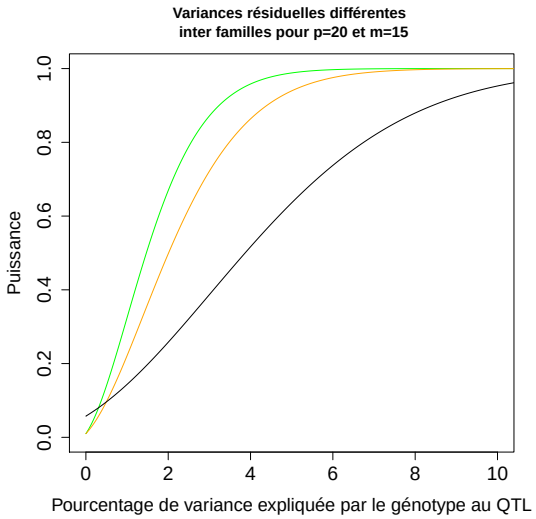


FIGURE 7.3 – Puissances analytiques approchées pour $m = 15$ dans le cas de variances résiduelles différentes inter familles.

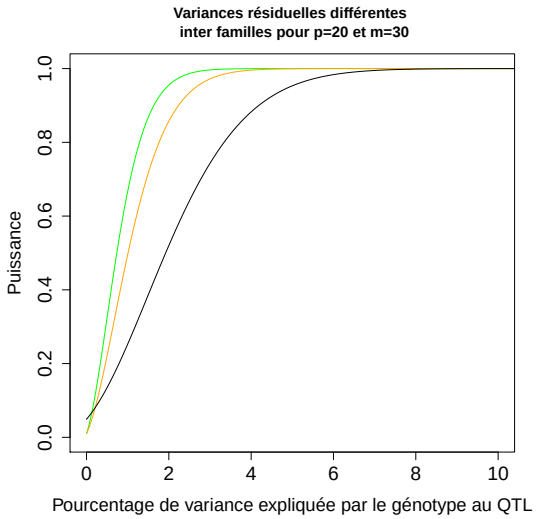


FIGURE 7.4 – Puissances analytiques approchées pour $m = 30$ dans le cas de variances résiduelles différentes inter familles.

Goffinet *et al.* (1999) ont montré qu'une statistique de test construite sur l'hypothèse d'hétéroscédasticité peut donner une amélioration de la puissance, par rapport à un test construit sur l'homoscédasticité, si le phénomène d'hétéroscédasticité est présent. Cependant, ils montrent que l'on observe seulement une légère amélioration de la puissance si ce phénomène n'est pas très prononcé. Remarquons finalement que la formulation de $\lambda_T^{hétéro.}$ explique probablement cette amélioration, car on a davantage de degrés de libertés sur le choix des différents paramètres de variances $(\sigma^2_{\varepsilon_i})_{1\leq i\leq pe}$, par rapport à $\lambda_T^{homo.}$, qui peuvent faire croître $\lambda_T^{hétéro.}$ si plusieurs de ces paramètres sont suffisamment petits.

iii) Pour le schéma avec des moyennes différentes inter familles (cf. 7.2.2.3) :

La variance expliquée sous HWE par le génotype au QTL pour ce schéma est la même que pour les cas *i)* et *ii)*. La différence avec ces cas *i)* et *ii)*, porte sur la contribution de la variance de la moyenne familiale à la variance totale. Les figures 7.5 et 7.6 montrent une meilleure puissance pour les modèles d'association par rapport au modèle de liaison homoscédastique, et une meilleure puissance générale par rapport aux cas *i)* et *ii)*, pour ce schéma.

Courbes de puissance analytique approchée :

■ : analyse d'association homoscédastique / ■ : analyse d'association corrigée

■ : analyse de liaison homoscédastique

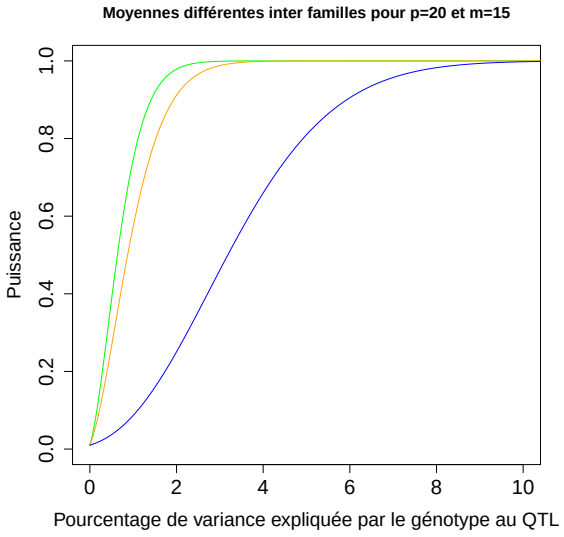


FIGURE 7.5 – Puissances analytiques approchées pour $m = 15$ dans le cas de moyennes différentes inter familles.

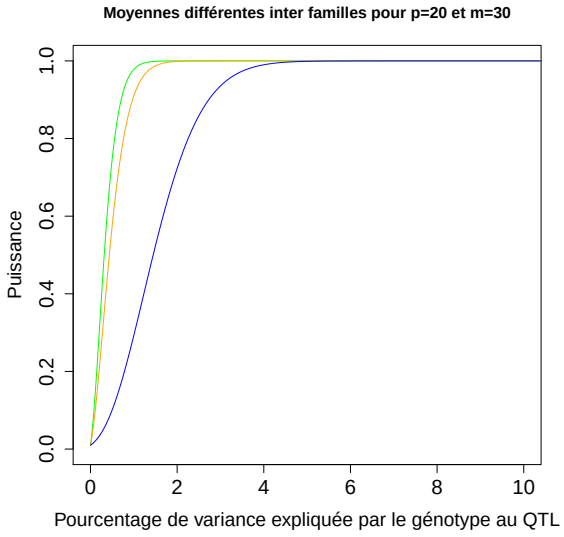


FIGURE 7.6 – Puissances analytiques approchées pour $m = 30$ dans le cas de moyennes différentes inter familles.

Cette meilleure puissance générale est dû à un effet d'échelle, dû au pourcentage de variance expliquée ici, sachant que les facteurs de décentrage sont les mêmes que ceux utilisés dans les cas *i)* et *ii)*. En effet, pour un α donné, ce pourcentage sera plus petit ici que pour les cas *i)* et *ii)*, à cause de la variance de la moyenne familiale intervenant au dénominateur du pourcentage. Cette différence d'échelle n'a cependant pas d'incidence sur la hiérarchie entre les puissances des modèles d'association et celles des modèles de liaison.

iv) Pour le schéma où le QTL est en interaction avec un locus (cf. 7.2.2.4) :

La variance expliquée sous HWE par le génotype au QTL pour ce schéma est donnée par :

$$\begin{aligned} V_{g_k} &= \left[f_{b_1}^2 f_{a_1}^2 (2\alpha)^2 + 2f_{b_1}f_{b_2}f_{a_1}^2(2\alpha)^2 + f_{b_1}^2 2f_{a_1}f_{a_2}(\alpha)^2 + 2f_{b_1}f_{b_2}.2f_{a_1}f_{a_2}(\alpha)^2\right] \\ &\quad - \left[f_{b_1}^2 f_{a_1}^2 (2\alpha) + 2f_{b_1}f_{b_2}f_{a_1}^2(2\alpha) + f_{b_1}^2 2f_{a_1}f_{a_2}(\alpha) + 2f_{b_1}f_{b_2}.2f_{a_1}f_{a_2}(\alpha)\right]^2 \\ &= \frac{3}{2}\alpha^2\left[2f_{a_1}^2 + f_{a_1}f_{a_2}\right] - \frac{9}{4}\alpha^2\left[f_{a_1}^2 + f_{a_1}f_{a_2}\right]^2 \ (\text{si} f_{b_1} = f_{b_2} = \frac{1}{2}) \\ &= \frac{9}{16}\alpha^2 \ (\text{si} f_{a_1} = f_{a_2} = \frac{1}{2}) \end{aligned}$$

Les figures 7.7 et 7.8 montrent une meilleure puissance pour les modèles d'association par rapport au modèle de liaison homoscédastique, et une meilleure puissance générale par rapport aux cas *i)* à *iii)*, pour ce schéma. Cette meilleure puissance générale est probablement due au fait que les modèles statistiques sont en mauvaise adéquation avec ce schéma de simulation, car les courbes de puissances sont peu fiables et semblent fausses ici. Autrement dit, on sur-estime les puissances ici en ne tenant pas compte des interactions dans les modèles statistiques pour la dérivation des facteurs de décentrage, bien que l'on tienne compte de ces interactions pour le pourcentage de variance expliquée au QTL dans la population. Ce phénomène d'échelle est le même que celui vu précédemment, dans le cas de moyennes différentes, mais en plus accentué à cause des interactions non considérés dans les modèles statistiques.

Courbes de puissance analytique approchée :

■ : analyse d'association homoscédastique / ■ : analyse d'association corrigée

■ : analyse de liaison homoscédastique

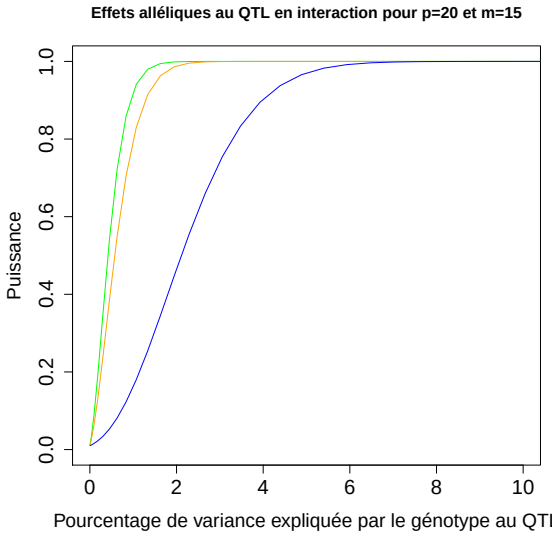


FIGURE 7.7 – Puissances analytiques approchées pour $m = 15$ dans le cas d'effets alléliques au QTL en interaction avec un locus.

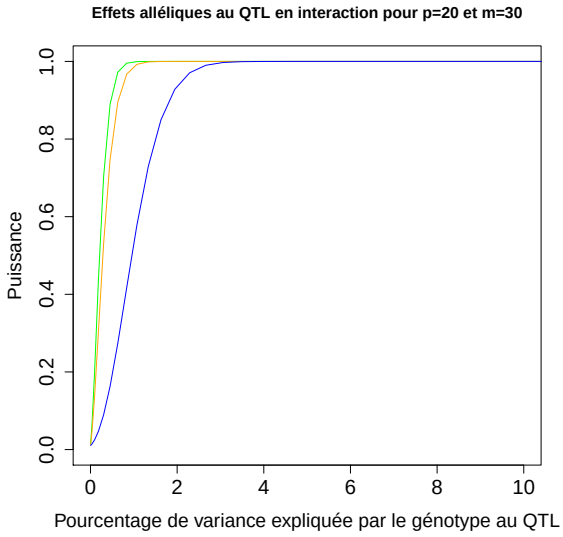


FIGURE 7.8 – Puissances analytiques approchées pour $m = 30$ dans le cas d'effets alléliques au QTL en interaction avec un locus.

7.3.2 Estimation par Monte-Carlo

7.3.2.1 Puissances et taux d'erreur de première espèce

Cette sous-section donne les puissances et les taux d'erreur de première espèce estimés par Monte-Carlo, dans le cadre des différents schéma proposés en 7.2.2, pour des jeux de simulations contenant de la variabilité dû à l'échantillonnage. Cette évaluation a été faite de manière indépendante de ce qui a été vu dans la sous-section 7.3.1 précédente, sans faire appel à la notion de facteur de décentrage. L'objet de cette sous-section est de savoir, entres autres, si les puissances analytiques approchées décrivent raisonnablement les puissances estimées sur des jeux de simulations contenant de la variabilité. Les statistiques de test d'association, et de liaison, ne sont plus calculées ici sur les nombres de pères et de descendants espérés sous HWE, mais sur les nombres réalisés à chaque simulation.

Le nombre de simulations d'une population pour un triplet (p, m, α) sous H_1 et sous H_0 (i.e. $\alpha = 0$) est fixé à 10000. Comme précédemment, les puissances sont données pour $m = 15$ et 30 descendants, un nombre de pères fixé à $p = 20$ et des fréquences équilibrées au marqueur testé. Les taux d'erreur de première espèce sont illustrés, dans les figures 7.12, 7.15, 7.18 et 7.20 en fonction du nombre de descendants qui varie de 10 à 50. Les puissances sont illustrées, dans les figures 7.9 à 7.24, en fonction du pourcentage de

variance expliquée (en moyenne) par le génotype au QTL sur l'ensemble des simulations. La variance V_{g_k} expliquée par le génotype au QTL a été évaluée ici de la manière suivante :

$$V_{g_k} = \frac{1}{n}\sum_{k=1}^{n}(g_k)^2 - (\frac{1}{n}\sum_{k=1}^{n} g_k)^2$$

où g_k est le génotype au QTL de l'individu k. Comme précédemment, le 0.99-quantile de la distribution de Fisher (centrée) correspondante a été fixé comme le seuil de rejet de H_0.

i) Pour le schéma avec des effets alléliques au QTL identiques inter familles (cf. 7.2.2.1) :

Les puissances estimées par Monte-Carlo dans les figures 7.9 et 7.10 décrivent les mêmes tendances générales que les puissances approchées dans les figures 7.1 et 7.2. Cependant les courbes analytiques sont respectivement plus hautes de 0.08, 0.05 et 0.11, en moyenne par rapport aux courbes estimées, pour l'association homoscédastique, l'association corrigée et la liaison homoscédastique.

Courbes de puissance estimée :

■ : analyse d'association homoscédastique / ■ : analyse d'association corrigée

■ : analyse de liaison homoscédastique

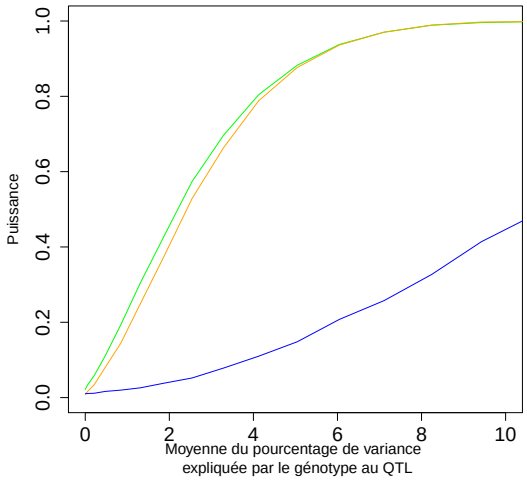


FIGURE 7.9 – Puissances estimées par Monte-Carlo pour $m = 15$ dans le cas d'effets alléliques au QTL identiques inter familles.

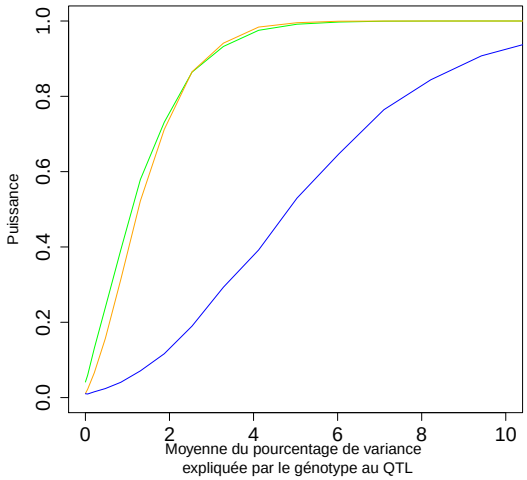


FIGURE 7.10 – Puissances estimées par Monte-Carlo pour $m = 30$ dans le cas d'effets alléliques au QTL identiques inter familles.

Remarquons que les fréquences espérées sous HWE ne sont pas vérifiées pour toutes les simulations sous H_1 et H_0. Cela explique, en partie des autres hypothèses des modèles

statistiques, les différences que l'on peut observer en général entre des puissances estimées par Monte-Carlo et des puissances analytiques approchées. La figure 7.11 donne un exemple des distributions du nombre de descendants homozygotes pour les 30 premières simulations associées au cas *i)* lorsque $p = 20$ et $m = 15$. Sous HWE on espère, à chaque simulation, observer $75 (= mpf_{a_1}^2)$ homozygotes de chaque classe. Or, la figure 7.11 montre que les fréquences espérées sous HWE ne sont pas toujours vérifiées en raison de l'aléa d'échantillonnage.

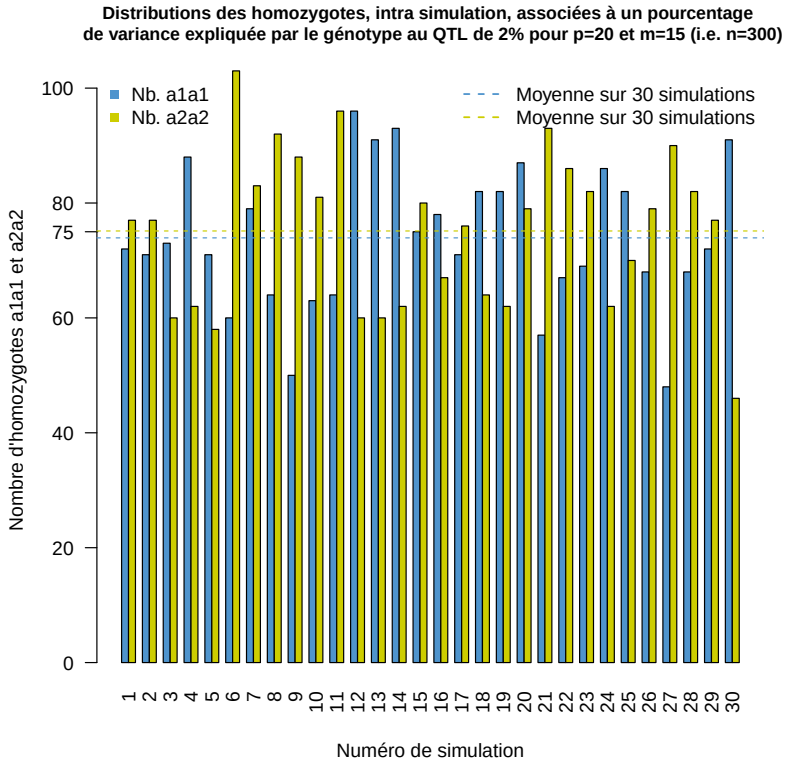


FIGURE 7.11 – Distributions du nombre de descendants homozygotes pour les 30 premières simulations associées au cas *i)* lorsque $p = 20$ et $m = 15$.

Les courbes de taux d'erreur de première espèce dans la figure 7.12 montrent un mauvais contrôle de cette erreur pour le modèle d'association homoscédastique, contrairement aux autres modèles d'analyses, lorsque le nombre de descendants augmente. Cette augmentation est très probablement due à la non prise en compte de la structure génétique dans le modèle d'association homoscédastique (Jung *et al.*, 2005 ; Manenti *et al.*,2009 ; Platt *et al.*, 2010). Remarquons que ces courbes d'erreur coïncident, par définition, avec celles que l'on obtiendrait pour le schéma *iv)* (i.e. le cas d'allèles en interaction) car les schémas *i)* et *iv)* sont les mêmes sous H_0.

Courbes de taux d'erreur de première espèce estimé :

■ : analyse d'association homoscédastique / ■ : analyse d'association corrigée

■ : analyse de liaison homoscédastique / ■ : taux d'erreur de première espèce fixé à 0.01

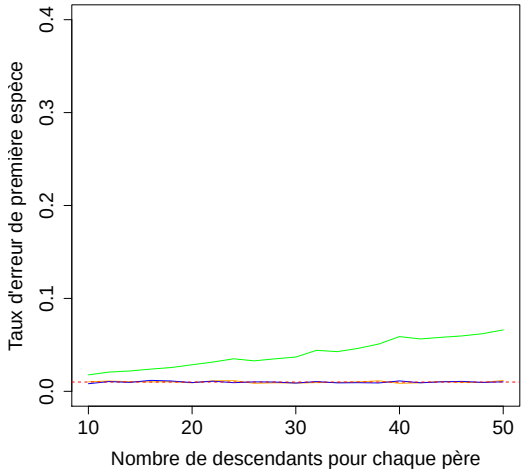


FIGURE 7.12 – Taux d'erreur de première espèce estimés par Monte-Carlo dans le cadre des schémas *i)* et *iv)* de simulation.

ii) Pour le schéma avec des variances résiduelles différentes inter familles (cf. 7.2.2.2) :

Les puissances estimées par Monte-Carlo, dans les figures 7.13 et 7.14, montrent les mêmes tendances générales que les puissances analytiques approchées dans les figures 7.3 et 7.4. Les courbes analytiques sont respectivement plus hautes de 0.10, 0.07 et 0.11, en moyenne par rapport aux courbes estimées, pour l'association homoscédastique, l'association corrigée et la liaison hétéroscédastique.

Courbes de puissance estimée :

■ : analyse d'association homoscédastique / ■ : analyse d'association corrigée

■ : analyse de liaison hétéroscédastique

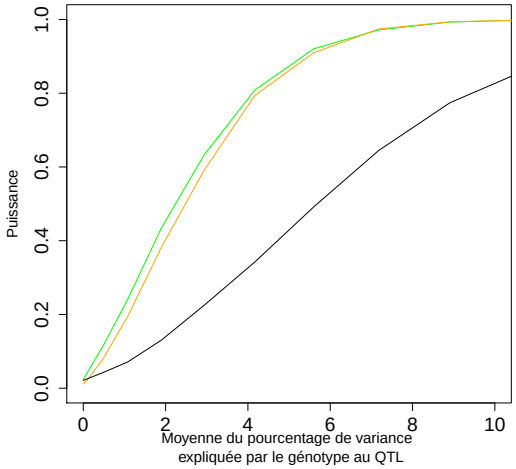


FIGURE 7.13 – Puissances estimées par Monte-Carlo pour $m = 15$ dans le cas de variances résiduelles différentes inter familles.

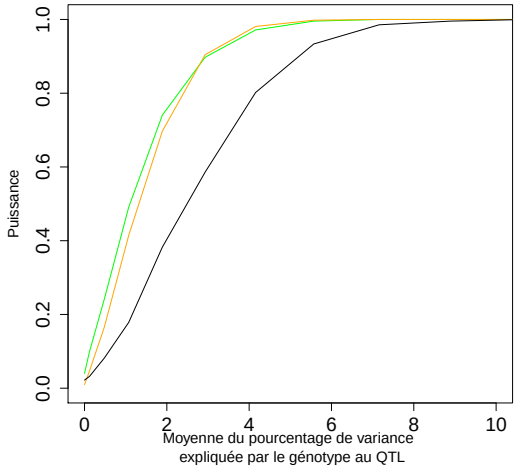


FIGURE 7.14 – Puissances estimées par Monte-Carlo pour $m = 30$ dans le cas de variances résiduelles différentes inter familles.

Courbes de taux d'erreur de première espèce estimé :

■ : analyse d'association homoscédastique / ■ : analyse d'association corrigée

■ : analyse de liaison hétéroscédastique / ■ : taux d'erreur de première espèce fixé à 0.01

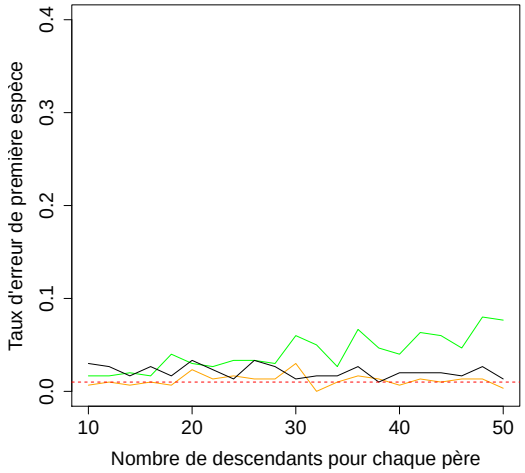


FIGURE 7.15 – Taux d'erreur de première espèce estimés par Monte-Carlo dans le cas de variances résiduelles différentes inter familles.

Les courbes de taux d'erreur de première espèce dans la figure 7.15 montrent un relativement bon contrôle de cette erreur pour le modèle d'association corrigé et respectivement une augmentation, et une légère inflation de ce taux, pour le modèle d'association homoscédastique et le modèle de liaison hétéroscédastique. Notons que l'augmentation de ce taux pour le modèle d'association homoscédastique, par rapport aux nombre de descendants, est comparable en grandeur à celle observée dans la figure 7.12.

iii) Pour le schéma avec des moyennes différentes inter familles (cf. 7.2.2.3) :

Les puissances estimées par Monte-Carlo dans les figures 7.16 et 7.17 montrent les mêmes tendances générales que les puissances analytiques approchées dans les figures 7.5 et 7.6, sauf pour le modèle d'association homoscédastique où l'on observe une baisse de puissance. Cette baisse de puissance est probablement due à la non prise en compte des effets de structure dus aux moyennes qui sont très différentes entre les familles. Les courbes analytiques sont respectivement plus hautes de 0.14, 0.08 et 0.13, en moyenne, pour l'association homoscédastique, l'association corrigée et la liaison homoscédastique par rapport aux courbes estimées. Remarquons que l'on observe aussi dans les figures 7.16 et 7.17 une amélioration générale des puissances estimées par rapport aux cas *i)* et *ii)*, ce qui correspond au phénomène déjà observé dans le cadre des puissances analytiques approchées.

Courbes de puissance et de taux d'erreur de première espèce estimé :

■ : analyse d'association homoscédastique / ■ : analyse d'association corrigée

■ : analyse de liaison homoscédastique / ■ : taux d'erreur de première espèce fixé à 0.01

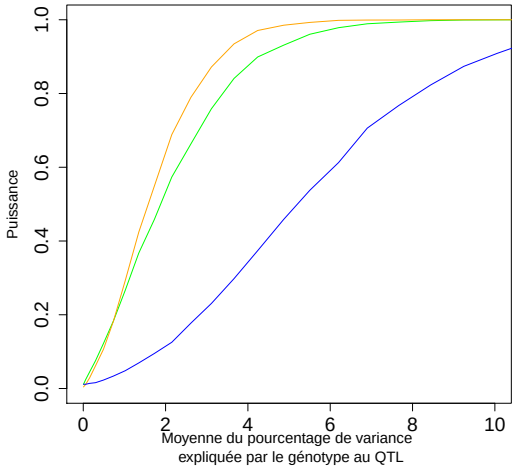


FIGURE 7.16 – Puissances estimées par Monte-Carlo pour $m = 15$ dans le cas de moyennes différentes inter familles.

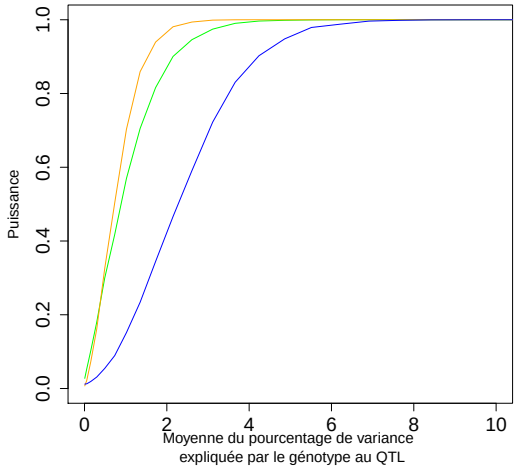


FIGURE 7.17 – Puissances estimées par Monte-Carlo pour $m = 30$ dans le cas de moyennes différentes inter familles.

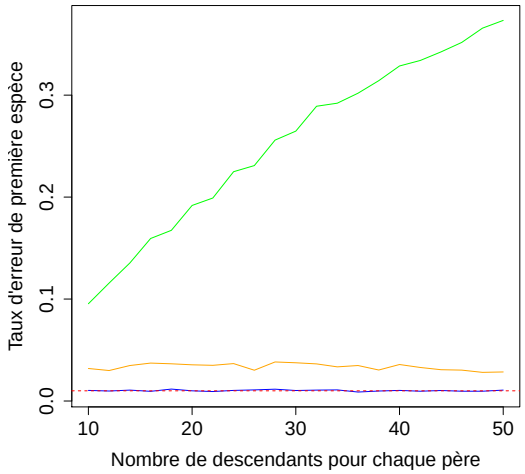


FIGURE 7.18 – Taux d'erreur de première espèce estimés par Monte-Carlo dans le cas de moyennes différentes inter familles.

Les courbes de taux d'erreur de première espèce montrent respectivement un mauvais contrôle et une inflation de ce taux pour le modèle d'association homoscédastique et le modèle d'association corrigé. Il est important de rappeler qu'on a supposé que $\sigma_u^2 = 0.5$ et $\sigma_\varepsilon^2 = 1$ pour cette étude. Ces valeurs ne sont donc pas estimées ici par maximisation de la vraisemblance, ou de la vraisemblance restreinte, comme proposé dans le chapitre 3. Or, on a vu en 6.3.3.3 que l'hétérogénéité de moyennes pouvait avoir une influence sur les facteurs de décentrage associés aux statistiques de test d'association sous H_0.

Afin de tenir compte de l'hétérogénéité de moyennes inter familles sous H_0 pour le modèle d'association corrigé, on peut remplacer σ_u^2 et σ_ε^2 par les variances inter et intra groupes obtenues de la décomposition de la variance totale. En notant respectivement $\underline{\hat{\sigma}}_u^2$ et $\underline{\hat{\sigma}}_\varepsilon^2$ les variances estimées inter et intra familles (groupes), et en admettant que les fréquences espérées sous HWE sont réalisées à chaque simulation, on montre que :

$$\begin{aligned}
\lambda^{corr.}_{A,\underline{\hat{\sigma}}_u^2,\underline{\hat{\sigma}}_\varepsilon^2} < \lambda^{corr.}_{A,\sigma_u^2=0.5,\sigma_\varepsilon^2=1} &\iff \frac{8n\alpha^2 f_{a_1} f_{a_2}}{\underline{\hat{\sigma}}_\varepsilon^2\left(1+\frac{3\underline{\hat{\sigma}}_u^2}{4\underline{\hat{\sigma}}_\varepsilon^2}\right)} < \frac{8n\alpha^2 f_{a_1} f_{a_2}}{\left(1+\frac{3(0.5)^2}{4}\right)} \\
&\iff \underline{\hat{\sigma}}_y^2 > \frac{1}{4}\underline{\hat{\sigma}}_u^2 + \frac{19}{16} \quad \text{où } \underline{\hat{\sigma}}_y^2 = \underline{\hat{\sigma}}_u^2 + \underline{\hat{\sigma}}_\varepsilon^2 \\
&\iff \underline{\hat{\sigma}}_y^2 > \phi(\underline{\hat{\sigma}}_u^2) \quad \text{où } \phi(\underline{\hat{\sigma}}_u^2) = \frac{1}{4}\underline{\hat{\sigma}}_u^2 + \frac{19}{16}
\end{aligned}$$

Sous HWE, cette inégalité permet donc de vérifier sur les données si $\underline{\hat{\sigma}}_u^2$ et $\underline{\hat{\sigma}}_\varepsilon^2$ font décroître le facteur de décentrage. Or, la figure 7.19 montre que cette inégalité est vérifiée pour les simulations effectuées sous H_0 dans le cadre du schéma de simulation *iii)* lorsque $m = 50$. En remplaçant σ_u^2 et σ_ε^2 par $\underline{\hat{\sigma}}_u^2$ et $\underline{\hat{\sigma}}_\varepsilon^2$ on obtient, dans les figures 7.20 à 7.22, de nouvelles courbes de taux d'erreur de première espèce et de puissances.

Courbes de puissance et de taux d'erreur de première espèce estimé :

■ : analyse d'association homoscédastique / ■ : analyse d'association corrigée

■ : analyse de liaison homoscédastique / ■ : taux d'erreur de première espèce fixé à 0.01

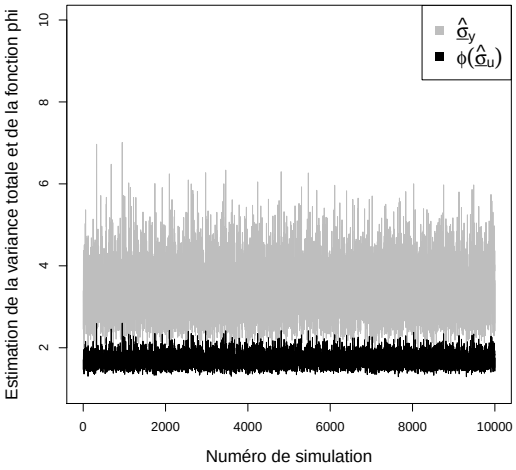


FIGURE 7.19 – Estimation de la variance totale et de la fonction ϕ pour les 10000 simulations sous H_0 dans le cadre du modèle 7.2.2.3

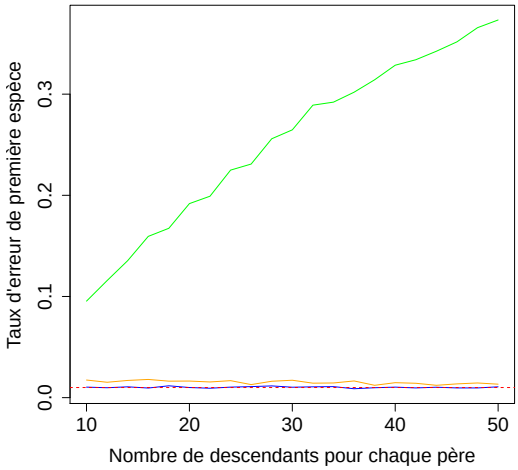


FIGURE 7.20 – Taux d'erreur de première espèce estimés par Monte-Carlo, en utilisant les variances inter et intra, dans le cas de moyennes différentes inter familles.

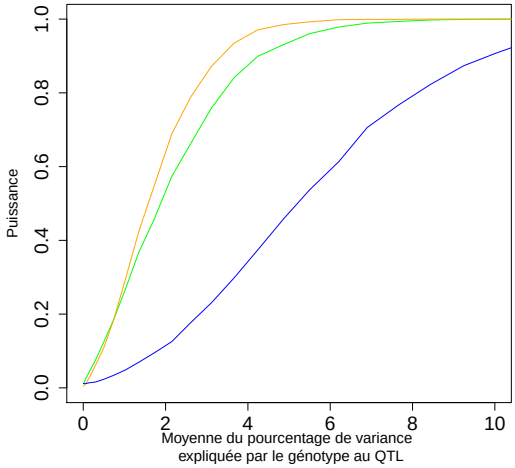


FIGURE 7.21 – Puissances estimées par Monte-Carlo pour $m = 15$, en utilisant les variances inter et intra dans le cas de moyennes différentes inter familles.

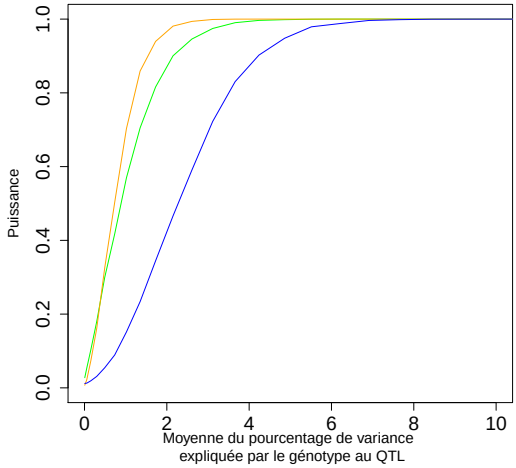


FIGURE 7.22 – Puissances estimées par Monte-Carlo pour $m = 30$, en utilisant les variances inter et intra dans le cas de moyennes différentes inter familles.

On voit dans les figures 7.20 à 7.22, par rapport aux figures 7.16 à 7.18 précédentes, que l'on a un meilleur contrôle de l'erreur de première espèce et un même niveau de puissance pour le modèle d'association corrigé. Le passage par l'inégalité ne constitue pas une méthode générale à appliquer en pratique, mais elle montre simplement que l'optimisation (EM, REML..) associée à l'estimation des paramètres σ_u^2 et σ_ε^2 intervient dans la robustesse du modèle d'association corrigé.

iv) Pour le schéma où le QTL est en interaction avec un locus (cf. 7.2.2.4) :

Les puissances estimées par Monte-Carlo dans les figures 7.23 et 7.24 ne décrivent pas les mêmes tendances générales que les puissances analytiques approchées, dans les figures 7.7 et 7.8, et sont beaucoup plus basses que ces dernières. Les courbes analytiques sont respectivement plus hautes de 0.31, 0.27 et 0.53, en moyenne, pour l'association homoscédastique, l'association corrigée et la liaison homoscédastique par rapport aux courbes estimées.

Remarquons également que l'on observe une baisse générale des puissances estimées par rapport aux cas *i)* à *iii)*. Cela montre que les modèles statistiques étudiés ne sont pas en bonne adéquation avec le processus de génération des données, ici, car ils ne tiennent pas compte des interactions par construction. Cependant, on voit dans la figure 7.8 que ces modèles peuvent donner une relativement bonne puissance lorsque la taille des familles augmente suffisamment (i.e. $m \geq 30$).

Courbes de puissance estimée :

■ : analyse d'association homoscédastique / ■ : analyse d'association corrigée

■ : analyse de liaison homoscédastique

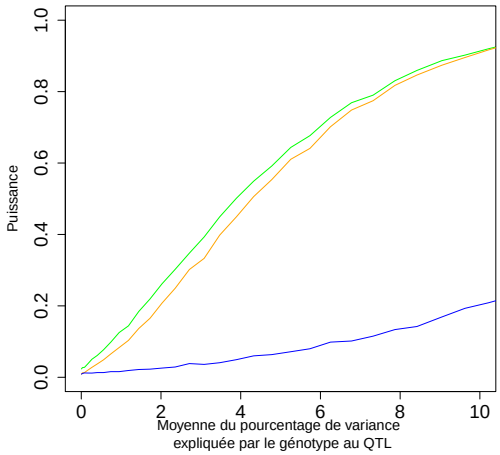


FIGURE 7.23 – Puissances estimées par Monte-Carlo pour $m = 15$ dans le cas d'effets alléliques au QTL en interaction avec un locus.

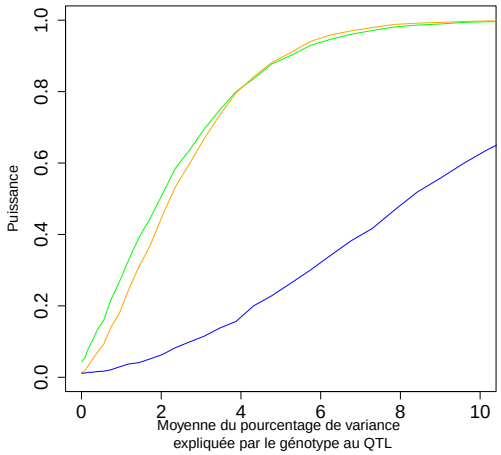


FIGURE 7.24 – Puissances estimées par Monte-Carlo pour $m = 30$ dans le cas d'effets alléliques au QTL en interaction avec un locus.

Tableau récapitulatif : différences moyennes entre les puissances

Le tableau 7.1 donne un récapitulatif des différences moyennes entre les courbes de puissances analytiques approchées et celles estimées par Monte-Carlo pour les différents schémas de simulation définis en 7.2.2.

Tableau de différences moyennes entre les puissances :

■ : analyse d'association homoscédastique / ■ : analyse d'association corrigée

■ : analyse de liaison homoscédastique / ■ : analyse de liaison hétéroscédastique

Schéma	Différences moyennes
i)	■ 0.08 ■ 0.05 ■ 0.11
ii)	■ 0.10 ■ 0.07 ■ 0.11
iii)	■ 0.14 ■ 0.08 ■ 0.13
iv)	■ 0.31 ■ 0.27 ■ 0.53

TABLEAU 7.1 – Différences moyennes entre les puissances analytiques approchées et les puissances estimées par Monte-Carlo

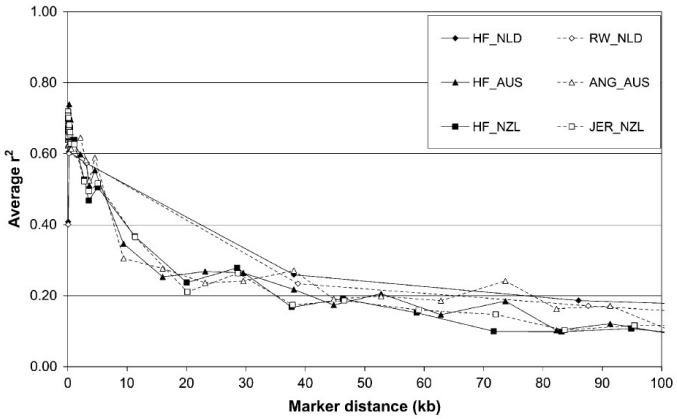
On voit dans le tableau 7.1 que, quelque soit le schéma de simulation, la plus petite différence moyenne entre les courbes de puissance est toujours donnée par l'analyse d'association corrigée pour la structure familiale. Le modèle d'association corrigé est donc celui qui est le moins influencé par la levée de ses hypothèses, selon les différents schémas de simulation considérés et la variance d'échantillonnage lors des simulations (i.e.

la déviation aux fréquences espérées sous HWE). La plus grande différence moyenne est quant à elle donnée, dans la majorité des cas, par l'analyse de liaison homoscédastique ou hétéroscédastique. Notons cependant que les modèles de liaison sont certainement plus sensibles que les modèles d'association à la déviation aux fréquences espérées sous HWE, lors des simulations, à cause du nombre de pères hétérozygotes intervenant dans les analyses de liaison. Les grandes différences moyennes pour le schéma *iv)* montrent que les modèles ne sont pas en adéquation avec ce schéma. Le schéma *iv)* n'est pas purement additif au QTL, contrairement aux autres schémas, car il tient compte d'une autre variable non modélisée par les modèles discriminés qui est la valeur de l'allèle au locus en interaction avec le QTL. Pour ce genre de situation le test F perd sa robustesse à taille d'échantillon fini sous H_1, ce qui peut s'observer empiriquement par simulations, et les puissances analytiques approchées sont en conséquence fausses.

7.3.2.2 Puissances avec prise en compte du LD

On suppose dans cette sous-section qu'il n'y a plus confusion entre le marqueur testé et le QTL. Autrement dit, on suppose que ces derniers sont deux positions distinctes et on note b_1, b_2 les allèles au marqueur testé et a_1, a_2 les allèles au QTL. L'objet de cette sous-section est de caractériser l'influence du LD, entre le marqueur testé et le QTL, sur les puissances associées aux différents modèles d'analyse dans le cadre du schéma *i)* de simulation (i.e. le schéma polygénique de base). La mesure du LD utilisée ici est le r^2 de Hill et Robertson (cf. 1.1.6.1). Le nombre de pères et de descendants sont respectivement fixés à $p = 20$ et $m = 30$. On suppose également que chaque allèle au marqueur testé, intra famille, est complètement associé à un des allèles au QTL. Autrement dit, on suppose qu'il existe un LD total intra famille entre les allèles au marqueur testé et les allèles au QTL.

La figure 7.25 illustre le r^2 moyen dans 6 races bovines en fonction de la distance physique entre deux SNP (De Roos *et al.*, 2008). Ce graphique montre qu'un r^2 moyen de 0.1 correspond, approximativement, à une distance physique de 10^5 bases entre deux marqueurs. Une telle distance entre deux SNP correspond à un faible taux de recombinaison de 0.1%, ce qui montre la vraisemblance de l'hypothèse de non-recombinaison dans les simulations pour un r^2 au moins égal à 0.1. Cette situation est réaliste lorsque deux loci sont proches sur le génome car les événements de recombinaison deviennent rares.



Average linkage disequilibrium (r^2) as a function of average genomic distance for Dutch black-and-white Holstein–Friesian bulls (HF_NLD), Dutch red-and-white Holstein–Friesian bulls (RW_NLD), Australian Holstein–Friesian bulls (HF_AUS), Australian Angus animals (ANG_AUS), New Zealand Friesian cows (HF_NZL), and New Zealand Jersey cows (JER_NZL) for distances between 0 and 100 kb. Each data point was based on 200 marker pairs, resulting in standard errors ≤0.03.

FIGURE 7.25 – r^2 moyen dans 6 races bovines en fonction de la distance physique entre deux SNP, d'après De Roos *et al.*, 2008

Remarquons que le LD mesuré sur l'ensemble d'une population étudiée peut être nul, bien que le LD mesuré dans chacune des familles constituant celle-ci soit maximal. Considérons l'exemple simple de deux familles, i et j, composées chacune de deux individus k et l. Soient $\{b_1a_1, b_2a_2\}^{i,k}$, $\{b_1a_1, b_2a_2\}^{i,l}$, $\{b_1a_2, b_2a_1\}^{j,k}$ et $\{b_1a_2, b_2a_1\}^{j,l}$ les haplotypes portés par les quatre individus sur l'ensemble de ces deux familles. On voit pour cet exemple simple que le LD mesuré sur l'ensemble des familles sera nul bien que le LD mesuré dans chacune d'elles sera total. Les figures 7.26 à 7.28 illustrent les puissances obtenues pour quatre niveaux de LD entre le marqueur testé et le QTL.

Courbes de puissance estimée :

■ : analyse d'association homoscédastique / ■ : analyse d'association corrigée

■ : analyse de liaison homoscédastique

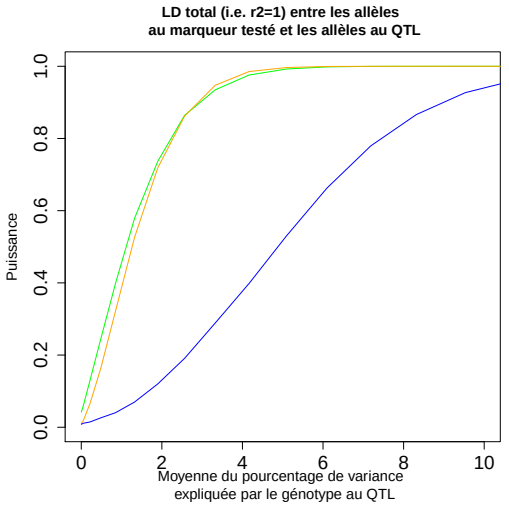


FIGURE 7.26 – Puissances estimées par Monte-Carlo dans le cadre du schéma polygénique pour $p = 20$, $m = 30$ et $r^2 = 1$.

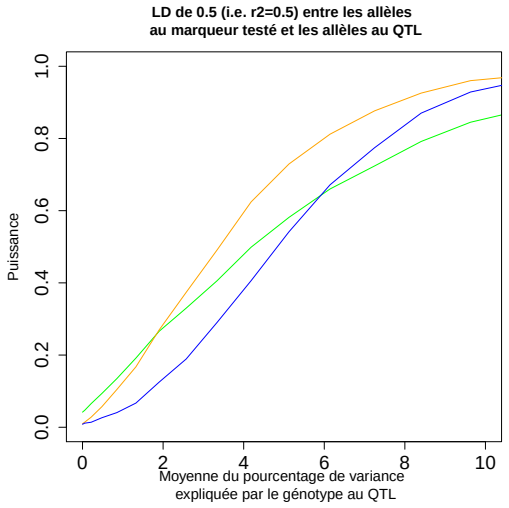


FIGURE 7.27 – Puissances estimées par Monte-Carlo dans le cadre du schéma polygénique pour $p = 20$, $m = 30$ et $r^2 = 0.5$.

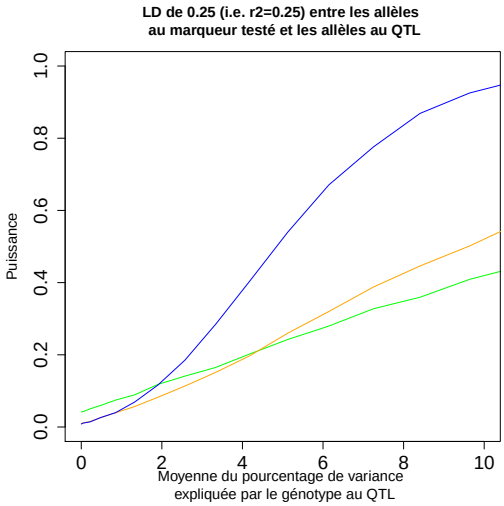


FIGURE 7.28 – Puissances estimées par Monte-Carlo dans le cadre du schéma polygénique pour $p = 20$, $m = 30$ et $r^2 = 0.25$.

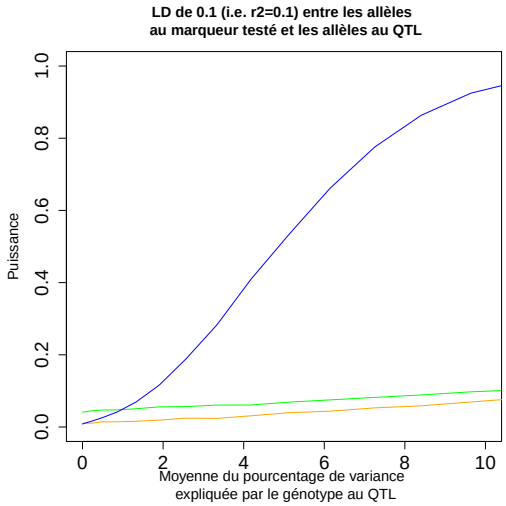


FIGURE 7.29 – Puissances estimées par Monte-Carlo dans le cadre du schéma polygénique pour $p = 20$, $m = 30$ et $r^2 = 0.10$.

Les figures 7.26 à 7.29 montrent respectivement, pour les modèles d'association par rapport au modèle de liaison, une meilleure et moins bonne puissance lorsque la mesure du r^2 est supérieure et inférieure à 0.5. Sous l'hypothèse de non recombinaison, on voit également dans ces figures que la puissance associée au modèle de liaison n'est pas influencée par le LD mesuré sur l'ensemble de la population, entre le marqueur testé et le QTL, lorsque celui-ci diminue progressivement. Ce phénomène risque donc d'être observé en pratique pour un faible taux de recombinaison, et il est conforme à l'hypothèse énoncé en 2.3.1 qui est que les modèles LA peuvent donner de la puissance, contrairement aux modèles LDA, lorsque l'on dispose d'une carte génétique de faible densité. Autrement dit, les modèles LA peuvent permettre de détecter un QTL si on dispose d'une carte génétique de faible densité, alors que ce n'est pas forcément le cas pour les modèles d'association.

7.4 Dérivation des facteurs de décentrage et des estimateurs associés aux modèles discriminés

7.4.1 Dérivation pour les modèles uni-SNP d'association

7.4.1.1 Modèle homoscédastique

i) Le facteur de décentrage dans le cas homoscédastique est donné par :

$$\lambda_A^{homo.} = \frac{1}{\sigma_\varepsilon^2} \| X_A \beta_A - P_{E_0} X_A \beta_A \|_2^2$$

où $P_{E_0} = X_{0A}(X'_{0A}X_{0A})^{-1}X'_{0A} = \frac{1}{n}J_n$ et on rappelle que J_n est la matrice n x n remplie de 1. Le terme $P_{E_0}X_A\beta_A$ est donc égal à :

$$P_{E_0}X_A\beta_A = \frac{1}{n}\begin{pmatrix} mp\mu + 2mpf_{a_1}^2\alpha - 2mpf_{a_2}^2\alpha \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ mp\mu + 2mpf_{a_1}^2\alpha - 2mpf_{a_2}^2\alpha \end{pmatrix} = \begin{pmatrix} \mu + 2\alpha(f_{a_1} - f_{a_2}) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \mu + 2\alpha(f_{a_1} - f_{a_2}) \end{pmatrix}, \text{ car } n = mp$$

Il vient donc que $X_A\beta_A - P_{E_0}X_A\beta_A$ est donné par :

$$X_A\beta_A - P_{E_0}X_A\beta_A = \begin{pmatrix} 2\alpha(1 - (f_{a_1} - f_{a_2})) \\ \vdots \\ 2\alpha(1 - (f_{a_1} - f_{a_2})) \\ -2\alpha(f_{a_1} - f_{a_2}) \\ \vdots \\ -2\alpha(f_{a_1} - f_{a_2}) \\ -2\alpha(1 + (f_{a_1} - f_{a_2})) \\ \vdots \\ -2\alpha(1 + (f_{a_1} - f_{a_2})) \end{pmatrix} = \begin{pmatrix} 4\alpha f_{a_2} \\ \vdots \\ 4\alpha f_{a_2} \\ -2\alpha(1 - 2f_{a_2}) \\ \vdots \\ -2\alpha(1 - 2f_{a_2}) \\ -4\alpha f_{a_1} \\ \vdots \\ -4\alpha f_{a_1} \end{pmatrix}$$

Le paramètre $\lambda_A^{homo.}$ se réduit finalement (après simplification) à :

$$\boxed{\lambda_A^{homo.} = \frac{1}{\sigma_\varepsilon^2}\Big(16n\alpha^2 f_{a_1}^2 f_{a_2}^2 + 8n\alpha^2(1 - 2f_{a_1})^2 f_{a_1} f_{a_2} + 16n\alpha^2 f_{a_1}^2 f_{a_2}^2\Big) = \frac{1}{\sigma_\varepsilon^2}\Big(8n\alpha^2 f_{a_1} f_{a_2}\Big)}$$

ii) L'estimateur associé à β_A est donné par :

$$\hat{\beta}_A = \begin{pmatrix} \hat{\mu} \\ \hat{\alpha} \end{pmatrix} = (X'_AX_A)^{-1}X'_AY_A = \begin{pmatrix} n & 2n(f_{a_1} - f_{a_2}) \\ 2n(f_{a_1} - f_{a_2}) & 4n(f_{a_1}^2 + f_{a_2}^2) \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{p}(Y_{i2.} + Y_{i0.} + Y_{i^-2.}) \\ 2\sum_{i=1}^{p}(Y_{i2.} - Y_{i^-2.}) \end{pmatrix}$$

$$= \frac{1}{8n^2 f_{a_1} f_{a_2}} \begin{pmatrix} 4n(f_{a_1}^2 + f_{a_2}^2) & 2n(f_{a_2} - f_{a_1}) \\ 2n(f_{a_2} - f_{a_1}) & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{p}(Y_{i2.} + Y_{i0.} + Y_{i^-2.}) \\ 2\sum_{i=1}^{p}(Y_{i2.} - Y_{i^-2.}) \end{pmatrix}$$

$$\boxed{\Longrightarrow \hat{\alpha} = \frac{1}{4n f_{a_1} f_{a_2}} \left[\sum_{i=1}^{p} \Big[Y_{i2.} - Y_{i^-2.} + (f_{a_2} - f_{a_1}) \Big(Y_{i2.} + Y_{i0.} + Y_{i^-2.} \Big) \Big] \right]}$$

où $Y_{i2.} = \sum_{k=1}^{mf_{a_1}^2} Y_{i2k}$, $Y_{i0.} = \sum_{k=1}^{2mf_{a_1}f_{a_2}} Y_{i0k}$ et $Y_{i^-2.} = \sum_{k=1}^{mf_{a_2}^2} Y_{i^-2k}$. Pour des fréquences équilibrées, $\hat{\alpha}$ correspond donc à la différence de moyennes entre les individus porteurs de a_1a_1 et ceux porteurs de a_2a_2, i.e. $\hat{\alpha} = \dfrac{\sum_{i=1}^{p} Y_{i2.} - \sum_{i=1}^{p} Y_{i^-2.}}{n}$.

7.4.1.2 Modèle corrigé

i) Le facteur de décentrage pour le modèle d'association corrigé est donné par :

$$\lambda_A^{corr.} = \underbrace{\beta_A' X_A' V^{-1} X_A \beta_A}_{(1)} - \underbrace{\beta_A' X_A' V^{-1} X_{0A} (X_{0A}' V^{-1} X_{0A})^{-1} X_{0A}' V^{-1} X_A \beta_A}_{(2)}$$

où l'on rappelle que $V = \sigma_\varepsilon^2 (I_n + \rho A)$, avec $A = \bigoplus_{i=1}^{p} S_i$ tel que $\forall i \in \{1, .., p\}$ $S_i = \frac{3}{4} I_m + \frac{1}{4} J_m$, et $\rho = \dfrac{\sigma_u^2}{\sigma_\varepsilon^2}$. On rappelle aussi que A est une matrice diagonale par blocs dont ces derniers sont les S_i. On commence par calculer l'inverse de V afin de calculer $\lambda_A^{corr.}$. Or, on remarque que la matrice V peut aussi se ré-écrire de la façon suivante :

$$V = \sigma_\varepsilon^2 \begin{pmatrix} B_1 & & & \\ & \ddots & 0 & \\ & & B_i & \\ & 0 & & \ddots \\ & & & & B_p \end{pmatrix} \Longrightarrow V^{-1} = \sigma_\varepsilon^{-2} \begin{pmatrix} B_1^{-1} & & & \\ & \ddots & 0 & \\ & & B_i^{-1} & \\ & 0 & & \ddots \\ & & & & B_p^{-1} \end{pmatrix}$$

où $B_i = I_m + \rho S_i = I_m + \rho\Big(\frac{3}{4} I_m + \frac{1}{4} J_m\Big) = \frac{4+3\rho}{4} I_m + \frac{\rho}{4} J_m = \gamma I_m + l J_m$ (i.e. on pose $\gamma = \dfrac{4+3\rho}{4}$ et $l = \dfrac{\rho}{4}$). En utilisant l'identité d'inversion de Sherman-Morisson on a donc

$$B_i^{-1} = (\gamma I_m + l J_m)^{-1} = \frac{1}{\gamma} I_m - \frac{l}{\gamma(\gamma + lm)} J_m.$$

Calcul du terme (1) de $\lambda_A^{corr.}$:

On remarque que $X_A' V^{-1} X_A = \frac{p}{\sigma_\varepsilon^2} X_A^{\prime(i)} B_i^{-1} X_A^{(i)}$, où $X_A^{(i)}$ est décrit en 6.2.1. Le terme (1) vaut donc :

$$\beta_A' X_A' V^{-1} X_A \beta_A = \beta_A' \left[\frac{p}{\sigma_\varepsilon^2} X_A^{\prime(i)} B_i^{-1} X_A^{(i)} \right] \beta_A = \frac{p}{\sigma_\varepsilon^2} \beta_A' \left[\frac{1}{\gamma} X_A^{\prime(i)} X_A^{(i)} - \frac{l}{\gamma(\gamma + lm)} X_A^{\prime(i)} J_m X_A^{(i)} \right] \beta_A$$

$$= \frac{p}{\sigma_\varepsilon^2} \beta_A' \left[\frac{1}{\gamma} \begin{pmatrix} m & 2m(f_{a_1} - f_{a_2}) \\ 2m(f_{a_1} - f_{a_2}) & 4m(f_{a_1}^2 + f_{a_2}^2) \end{pmatrix} - \frac{l}{\gamma(\gamma + lm)} \begin{pmatrix} m^2 & 2m^2(f_{a_1} - f_{a_2}) \\ 2m^2(f_{a_1} - f_{a_2}) & 4m^2(f_{a_1} - f_{a_2})^2 \end{pmatrix} \right] \beta_A$$

$$= \frac{mp}{\sigma_\varepsilon^2(\gamma + lm)} \beta_A' \begin{pmatrix} 1 & 2(f_{a_1} - f_{a_2}) \\ 2(f_{a_1} - f_{a_2}) & \dfrac{4[\gamma + 2f_{a_1} f_{a_2}(lm - \gamma)]}{\gamma} \end{pmatrix} \beta_A$$

$$= \frac{mp}{\sigma_\varepsilon^2(\gamma + lm)} \left[\mu^2 + 4\mu\alpha(f_{a_1} - f_{a_2}) + \frac{4\alpha^2[\gamma + 2f_{a_1} f_{a_2}(lm - \gamma)]}{\gamma} \right]$$

Calcul du terme (2) de $\lambda_A^{corr.}$:

On calcule d'abord le terme $X_A' V^{-1} X_{0A}$ $(= (X_{0A}' V^{-1} X_A)')$, on a :

$$X_A' V^{-1} X_{0A} = \frac{p}{\sigma_\varepsilon^2} \left[\frac{1}{\gamma} \begin{pmatrix} m \\ 2m(f_{a_1} - f_{a_2}) \end{pmatrix} - \frac{l}{\gamma(\gamma + lm)} \begin{pmatrix} m^2 \\ 2m^2(f_{a_1} - f_{a_2}) \end{pmatrix} \right]$$

$$= \frac{mp}{\sigma_\varepsilon^2(\gamma + lm)} \begin{pmatrix} 1 \\ 2(f_{a_1} - f_{a_2}) \end{pmatrix}$$

Le terme (2) vaut donc :

$$\beta_A' X_A' V^{-1} X_{0A} (X_{0A}' V^{-1} X_{0A})^{-1} X_{0A}' V^{-1} X_A \beta_A$$
$$= \beta_A' \frac{mp}{\sigma_\varepsilon^2(\gamma + lm)} \begin{pmatrix} 1 \\ 2(f_{a_1} - f_{a_2}) \end{pmatrix} \left[\frac{p}{\sigma_\varepsilon^2} X_{0A}^{\prime(i)} B_i^{-1} X_{0A}^{(i)} \right]^{-1} \frac{mp}{\sigma_\varepsilon^2(\gamma + lm)} \begin{pmatrix} 1 & 2(f_{a_1} - f_{a_2}) \end{pmatrix} \beta_A$$

$$= \beta'_A \frac{mp}{\sigma^2_\varepsilon(\gamma+lm)} \begin{pmatrix} 1 \\ 2(f_{a_1}-f_{a_2}) \end{pmatrix} \Big[\frac{mp}{\sigma^2_\varepsilon(\gamma+lm)}\Big]^{-1} \frac{mp}{\sigma^2_\varepsilon(\gamma+lm)} \begin{pmatrix} 1 & 2(f_{a_1}-f_{a_2}) \end{pmatrix} \beta_A$$

$$= \frac{mp}{\sigma^2_\varepsilon(\gamma+lm)} \Big[\mu^2 + 4\mu\alpha(f_{a_1}-f_{a_2}) + 4\alpha^2(f_{a_1}-f_{a_2})^2\Big]$$

Finalement le paramètre $\lambda_A^{corr.}$ se réduit (après simplification) à :

$$\lambda_A^{corr.} = \text{terme (1) - terme (2)} \; = \frac{mp}{\sigma^2_\varepsilon(\gamma+lm)} \Big[\frac{4\alpha^2}{\gamma}\Big[\gamma + 2f_{a_1}f_{a_2}(lm-\gamma) - \gamma(f_{a_1}-f_{a_2})^2\Big]\Big]$$

$$\boxed{\lambda_A^{corr.} = \frac{8mp\alpha^2 f_{a_1}f_{a_2}}{\sigma^2_\varepsilon\Big(1+\dfrac{3\sigma^2_u}{4\sigma^2_\varepsilon}\Big)} = \frac{8n\alpha^2 f_{a_1}f_{a_2}}{\sigma^2_\varepsilon\Big(1+\dfrac{3\sigma^2_u}{4\sigma^2_\varepsilon}\Big)}}$$

ii) L'estimateur associé à β_A est donnée par :

$$\hat{\beta}_A = \begin{pmatrix} \hat{\mu} \\ \hat{\alpha} \end{pmatrix} = (X'_A V^{-1} X_A)^{-1} X'_A V^{-1} Y_A$$

$$= \frac{\sigma^2_\varepsilon(\gamma+lm)}{mp} \begin{pmatrix} 1 & 2(f_{a_1}-f_{a_2}) \\ 2(f_{a_1}-f_{a_2}) & \dfrac{4[\gamma + 2f_{a_1}f_{a_2}(lm-\gamma)]}{\gamma} \end{pmatrix}^{-1} \left(\frac{1}{\sigma^2_\varepsilon}\sum_{i=1}^{p} X_A'^{(i)} B_i^{-1} Y_A^{(i)}\right)$$

$$= \frac{\sigma^2_\varepsilon \gamma}{8nf_{a_1}f_{a_2}} \begin{pmatrix} \dfrac{4[\gamma + 2f_{a_1}f_{a_2}(lm-\gamma)]}{\gamma} & 2(f_{a_2}-f_{a_1}) \\ 2(f_{a_2}-f_{a_1}) & 1 \end{pmatrix} \times \frac{1}{\sigma^2_\varepsilon} \begin{pmatrix} \dfrac{1}{(\gamma+lm)}\displaystyle\sum_{i=1}^{p}(Y_{i2.} + Y_{i0.} + Y_{i-2.}) \\ \dfrac{2}{\gamma}\displaystyle\sum_{i=1}^{p}\Big[(Y_{i2.} - Y_{i-2.}) + \dfrac{lm(f_{a_2}-f_{a_1})}{(\gamma+lm)}(Y_{i2.} + Y_{i0.} + Y_{i-2.})\Big] \end{pmatrix}$$

$$\boxed{\Longrightarrow \hat{\alpha} = \frac{1}{4nf_{a_1}f_{a_2}}\left[\sum_{i=1}^{p}\Big[Y_{i2.} - Y_{i-2.} + (f_{a_2}-f_{a_1})\Big(Y_{i2.} + Y_{i0.} + Y_{i-2.}\Big)\Big]\right]}$$

où $Y_{i2.} = \displaystyle\sum_{k=1}^{mf_{a_1}^2} Y_{i2k}$, $Y_{i0.} = \displaystyle\sum_{k=1}^{2mf_{a_1}f_{a_2}} Y_{i0k}$ et $Y_{i-2.} = \displaystyle\sum_{k=1}^{mf_{a_2}^2} Y_{i-2k}$.

7.4.2 Dérivation pour les modèles uni-SNP de liaison

7.4.2.1 Modèle homoscédastique

i) Le facteur de décentrage dans le cas homoscédastique est donné par :

$$\lambda_T^{homo.} = \frac{1}{\sigma_\varepsilon^2}\|X_T\beta_T - P_{E_0}X_T\beta_T\|_2^2$$

où $P_{E_0} = X_{0T}(X'_{0T}X_{0T})^{-1}X'_{0T} = X_{0T}\Big[\sum_{i=1}^{pe} X_{T\mu}^{\prime(i)}X_{T\mu}^{(i)}\Big]^{-1}X'_{0T} = X_{0T}\Big[\frac{m}{2}I_{pe}\Big]^{-1}X'_{0T}$ car tous les éléments de $X_{T\mu}^{\prime(i)}X_{T\mu}^{(i)}$ sont nuls sauf l'élément de la $i^{\text{ème}}$ ligne et de la $i^{\text{ème}}$ colonne de ce produit qui vaut $\frac{m}{2}$. Il vient donc que P_{E_0} est égal à :

$$P_{E_0} = \frac{2}{m}X_{0T}X'_{0T} = \frac{2}{m}\begin{pmatrix} X_{T\mu}^{(1)}X_{T\mu}^{\prime(1)} & & & & \\ & \ddots & & 0 & \\ & & X_{T\mu}^{(i)}X_{T\mu}^{\prime(i)} & & \\ & 0 & & \ddots & \\ & & & & X_{T\mu}^{(pe)}X_{T\mu}^{\prime(pe)} \end{pmatrix}$$

où $\forall i \neq j$ on a $X_{T\mu}^{(i)}X_{T\mu}^{\prime(j)} = 0_{\frac{m}{2}}$ et on remarque que $X_{T\mu}^{(i)}X_{T\mu}^{\prime(i)} = J_{\frac{m}{2}}$. Il vient donc que $X_T\beta_T - P_{E_0}X_T\beta_T$ (en posant $Z = X_T\beta_T$) est donné par :

$$X_T\beta_T - P_{E_0}X_T\beta_T = \begin{pmatrix} Z^{(1)} \\ \vdots \\ Z^{(i)} \\ \vdots \\ Z^{(pe)} \end{pmatrix} - \frac{2}{m}\begin{pmatrix} J_{\frac{m}{2}} & & & & \\ & \ddots & & 0 & \\ & & J_{\frac{m}{2}} & & \\ & 0 & & \ddots & \\ & & & & J_{\frac{m}{2}} \end{pmatrix}\begin{pmatrix} Z^{(1)} \\ \vdots \\ Z^{(i)} \\ \vdots \\ Z^{(pe)} \end{pmatrix}, \text{ où } Z^{(i)} = \begin{pmatrix} \mu_i + \delta_i \\ \vdots \\ \mu_i + \delta_i \\ \mu_i - \delta_i \\ \vdots \\ \mu_i - \delta_i \end{pmatrix}$$

Le terme intermédiaire $Z^{(i)} - \frac{2}{m}J_{\frac{m}{2}}Z^{(i)}$ (après quelques simplifications) se réduit à :

$$Z^{(i)} - \frac{2}{m}J_{\frac{m}{2}}Z^{(i)} = \begin{pmatrix} 2\delta_i f_{a_2} \\ \vdots \\ 2\delta_i f_{a_2} \\ -2\delta_i f_{a_1} \\ \vdots \\ -2\delta_i f_{a_1} \end{pmatrix} \begin{matrix} \Big\} \frac{mf_{a_1}}{2} \\ \\ \Big\} \frac{mf_{a_2}}{2} \end{matrix} = U^{(i)}$$

Finalement le paramètre $\lambda_T^{homo.}$, après simplification, est donné par :

$$\boxed{\lambda_T^{homo.} = \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{pe} U'^{(i)}.U^{(i)} = \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{pe} 2m\delta_i^2 f_{a_1} f_{a_2}}$$

ii) L'estimateur associé à β_T est donné par :

$$\hat{\beta}_T = \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_{pe} \\ \hat{\delta}_1 \\ \vdots \\ \hat{\delta}_{pe} \end{pmatrix} = (X_T' X_T)^{-1} X_T' Y_T = \left[\sum_{i=1}^{pe} X_T'^{(i)} X_T^{(i)}\right]^{-1} \left[\sum_{i=1}^{pe} X_T'^{(i)} Y_T^{(i)}\right]$$

$$\Longleftrightarrow \hat{\beta}_T = \left[\sum_{i=1}^{pe} \begin{pmatrix} X_{T\mu}'^{(i)} X_{T\mu}^{(i)} & X_{T\mu}'^{(i)} X_{T\delta}^{(i)} \\ X_{T\delta}'^{(i)} X_{T\mu}^{(i)} & X_{T\delta}'^{(i)} X_{T\delta}^{(i)} \end{pmatrix}\right]^{-1} \left[\sum_{i=1}^{pe} \begin{pmatrix} X_{T\mu}'^{(i)} Y_T^{(i)} \\ X_{T\delta}'^{(i)} Y_T^{(i)} \end{pmatrix}\right]$$

où $X_{T\mu}'^{(i)} X_{T\mu}^{(i)} = X_{T\delta}'^{(i)} X_{T\delta}^{(i)} = \begin{pmatrix} 0 & & \\ & \ddots & \\ & & 0 \end{pmatrix}_{\frac{m}{2}(i,i)}$, $X_{T\mu}'^{(i)} X_{T\delta}^{(i)} = X_{T\delta}'^{(i)} X_{T\mu}^{(i)} = 0_{pe\times pe}$

, $X_{T\mu}'^{(i)} Y_T^{(i)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ Y_{i1.} + Y_{i^-1.} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ et $X_{T\delta}'^{(i)} Y_T^{(i)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ Y_{i1.} - Y_{i^-1.} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

On a donc que $\hat{\delta}_i = \frac{2}{m}\big(Y_{i1.} - Y_{i^-1.}\big)$, où $Y_{i1.} = \sum_{s=1}^{\frac{mf_{a_1}}{2}} Y_{i1s}$ et $Y_{i^-1.} = \sum_{s=1}^{\frac{mf_{a_2}}{2}} Y_{i^-1s}$. Remarquons que les individus porteurs de a_1 et a_2 sont respectivement homozygotes a_1a_1 et a_2a_2 au génotype. Il vient donc que $\hat{\delta}_i$ se ré-écrit également de la façon suivante :

$$\boxed{\hat{\delta}_i = \frac{2}{m}\big(Y_{i1.} - Y_{i^-1.}\big) = \frac{2}{m}\Big(\sum_{s=1}^{\frac{mf_{a_1}}{2}} Y_{i2s} - \sum_{s=1}^{\frac{mf_{a_2}}{2}} Y_{i^-2s}\Big)}$$

Autrement dit, $\hat{\delta}_i$ est égal à deux fois l'estimateur de α au sein de la famille i lorsque les fréquences sont équilibrées.

7.4.2.2 Modèle hétéroscédastique

i) Le facteur de décentrage pour le modèle de liaison hétéroscédastique est donné par :

$$\lambda_T^{hétéro.} = \underbrace{\beta_T' X_T' \tilde{V}^{-1} X_T \beta_T}_{(1)} - \underbrace{\beta_T' X_T' \tilde{V}^{-1} X_{0T} (X_{0T}' \tilde{V}^{-1} X_{0T})^{-1} X_{0T}' \tilde{V}^{-1} X_T \beta_T}_{(2)}$$

où l'on rappelle que $\tilde{V} = \bigoplus_{i=1}^{pe} \sigma_{\varepsilon_i}^2 I_{\frac{m}{2}}$.

Calcul du terme (1) de $\lambda_T^{hétéro.}$:

En posant $Z = X_T \beta_T$, le terme (1) de cette expression vaut :

$$\beta_T' X_T' \tilde{V}^{-1} X_T \beta_T = \begin{pmatrix} Z'^{(1)} & \dots & Z'^{(i)} & \dots & Z'^{(pe)} \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma_{\varepsilon_1}^2} Z'^{(1)} \\ \vdots \\ \frac{1}{\sigma_{\varepsilon_i}^2} Z'^{(i)} \\ \vdots \\ \frac{1}{\sigma_{\varepsilon_{pe}}^2} Z'^{(pe)} \end{pmatrix}, \text{ où } Z^{(i)} = \begin{pmatrix} \mu_i + \delta_i \\ \vdots \\ \mu_i + \delta_i \\ \mu_i - \delta_i \\ \vdots \\ \mu_i - \delta_i \end{pmatrix} \begin{matrix} \\ \\ \left.\begin{matrix} \\ \\ \end{matrix}\right\} \frac{mf_{a_1}}{2} \\ \left.\begin{matrix} \\ \\ \end{matrix}\right\} \frac{mf_{a_2}}{2} \end{matrix}$$

$$= \frac{m}{2} \sum_{i=1}^{pe} \frac{1}{\sigma_{\varepsilon_i}^2} \left[\mu_i^2 + 2\mu_i \delta_i (f_{a_1} - f_{a_2}) + \delta_i^2 \right] \text{ (après quelques simplifications)}$$

Calcul du terme (2) de $\lambda_T^{hétéro.}$:

L'expression $X_{0T}(X_{0T}' \tilde{V}^{-1} X_{0T})^{-1} X_{0T}'$ du terme (2) s'écrit :

$$X_{0T}(X_{0T}' \tilde{V}^{-1} X_{0T})^{-1} X_{0T}' = X_{0T} \begin{pmatrix} \frac{m}{2\sigma_{\varepsilon_1}^2} & & & & \\ & \ddots & & 0 & \\ & & \frac{m}{2\sigma_{\varepsilon_i}^2} & & \\ & 0 & & \ddots & \\ & & & & \frac{m}{2\sigma_{\varepsilon_{pe}}^2} \end{pmatrix}^{-1} X_{0T}'$$

Pour les raisons évoquées précédemment, en 7.4.2.1 sur le produit de $X_{0T}X'_{0T}$, on a :

$$X_{0T}(X'_{0T}\tilde{V}^{-1}X_{0T})^{-1}X'_{0T} = \frac{2}{m}\begin{pmatrix} \sigma^2_{\varepsilon_1}J_{\frac{m}{2}} & & & \\ & \ddots & 0 & \\ & & \sigma^2_{\varepsilon_i}J_{\frac{m}{2}} & \\ & 0 & & \ddots \\ & & & \sigma^2_{\varepsilon_{pe}}J_{\frac{m}{2}} \end{pmatrix}$$

Après quelques simplifications, il vient que le terme (2) est donné par :

$$\begin{aligned}\beta'_T X'_T\tilde{V}^{-1}X_{0T}(X'_{0T}\tilde{V}^{-1}X_{0T})^{-1}X'_{0T}\tilde{V}^{-1}X_T\beta_T &= \frac{2}{m}\sum_{i=1}^{pe}\frac{1}{\sigma^2_{\varepsilon_i}}Z'^{(i)}J_{\frac{m}{2}}Z^{(i)} \\ &= \frac{m}{2}\sum_{i=1}^{pe}\frac{1}{\sigma^2_{\varepsilon_i}}\Big[\mu_i + \delta_i(f_{a_1} - f_{a_2})\Big]^2\end{aligned}$$

Finalement le paramètre $\lambda_T^{hétéro.}$ est donné par :

$$\boxed{\lambda_T^{hétéro.} = \text{ terme (1) -terme (2) } = \sum_{i=1}^{pe}\frac{2m\delta_i^2 f_{a_1}f_{a_2}}{\sigma^2_{\varepsilon_i}}}$$

ii) L'estimateur associé à β_T est donné par :

$$\hat{\beta}_T = \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_{pe} \\ \hat{\delta}_1 \\ \vdots \\ \hat{\delta}_{pe} \end{pmatrix} = (X'_T\tilde{V}^{-1}X_T)^{-1}X'_T\tilde{V}^{-1}Y_T = \left[\sum_{i=1}^{pe}\frac{1}{\sigma^2_{\varepsilon_i}}X_T'^{(i)}X_T^{(i)}\right]^{-1}\left[\sum_{i=1}^{pe}\frac{1}{\sigma^2_{\varepsilon_i}}X_T'^{(i)}Y_T^{(i)}\right]$$

$$\iff \hat{\beta}_T = \left[\sum_{i=1}^{pe}\frac{1}{\sigma^2_{\varepsilon_i}}\begin{pmatrix} X_{T\mu}'^{(i)}X_{T\mu}^{(i)} & X_{T\mu}'^{(i)}X_{T\delta}^{(i)} \\ X_{T\delta}'^{(i)}X_{T\mu}^{(i)} & X_{T\delta}'^{(i)}X_{T\delta}^{(i)} \end{pmatrix}\right]^{-1}\left[\sum_{i=1}^{pe}\frac{1}{\sigma^2_{\varepsilon_i}}\begin{pmatrix} X_{T\mu}'^{(i)}Y_T^{(i)} \\ X_{T\delta}'^{(i)}Y_T^{(i)} \end{pmatrix}\right]$$

où les produits $X_{T\mu}'^{(i)}X_{T\mu}^{(i)}$, $X_{T\mu}'^{(i)}X_{T\delta}^{(i)}$, $X_{T\delta}'^{(i)}X_{T\mu}^{(i)}$, $X_{T\delta}'^{(i)}X_{T\delta}^{(i)}$, $X_{T\mu}'^{(i)}Y_T^{(i)}$ et $X_{T\delta}'^{(i)}Y_T^{(i)}$ sont les mêmes que ceux calculés pour le cas du modèle de liaison homoscédastique. En se reportant aux résultats obtenus pour le modèle de liaison homoscédastique (matrices diagonales

et nulles) on a :

$$\boxed{\hat{\delta}_i = \frac{2}{m}\big(Y_{i1.} - Y_{i^-1.}\big)}$$

où $Y_{i1.} = \sum_{s=1}^{\frac{mfa_1}{2}} Y_{i1s}$ et $Y_{i^-1.} = \sum_{s=1}^{\frac{mfa_2}{2}} Y_{i^-1s}$.

7.5 Discussion et perspectives de l'étude

J'ai comparé dans cette partie les puissances et les robustesses associées à des modèles uni-SNP de liaison avec celles de modèles uni-SNP d'association. L'étude a montré que les modèles de liaison considérés et le modèle d'association corrigé pour la structure génétique sont robustes aux faux positifs, pour plusieurs cas de simulation, conformément à ce qui est cité dans la littérature (cf. 2.3.1 et 2.3.2). Cependant, j'ai montré dans cette étude que la robustesse des modèles d'association dépend aussi parfois de l'optimisation (EM, REML..) associée à l'estimation des paramètres de ces modèles, en plus de la nécessité de la prise en compte de la structure génétique. Ce travail a également montré que les modèles uni-SNP de liaison considérés sont moins puissants que les modèles uni-SNP d'association pour plusieurs scénarios de simulation. Les modèles de liaison sont décrits fréquemment dans la littérature comme étant peu puissants pour la détection de QTL expliquant une faible proportion de la variance phénotypique (Risch, 2000 ; Onkamo *et al.*, 2002 ; McQueen *et al.*, 2005).

Sham *et al.* (2000) ont montré que la puissance associée à une analyse d'association, dans le cadre d'un modèle général proposé par Fulker *et al.* (1999), est liée linéairement à la proportion de variance expliquée par le QTL. Ils ont montré que la puissance associée à une analyse de liaison, définie dans ce même cadre, est liée quadratiquement à la proportion de variance expliquée par le QTL. Ils montrent ainsi, dans leur étude, que l'analyse de liaison n'a d'intérêt que lorsque le QTL explique au moins 10% de la variance totale. Le nombre de pères hétérozygotes est également un facteur limitant dans les analyses de liaison. Par exemple, pour des fréquences alléliques très déséquilibrées au QTL (cas d'allèles rares) l'application des modèles de liaison devient difficile à cause de la diminution considérable du nombre espéré de pères hétérozygotes. J'ai supposé que les fréquences alléliques au QTL étaient équilibrées dans ce travail, ce qui correspond à une situation optimale pour

le nombre espéré de pères hétérozygotes.

Un autre facteur limitant dans les analyses de liaison est la prise en compte de descendants uniquement homozygotes lorsque les génotypes des mères ne sont pas connus. Il est toutefois possible de connaître la transmission des allèles d'un père hétérozygote à des descendants hétérozygotes, lorsque la densité de marqueurs augmente et que l'on reconstitue les phases chez le père, ou quand les mères sont génotypées. Dans cette situation, la plupart des descendants du père hétérozygote sont informatifs. Cependant, l'utilisation de tous les descendants d'un père hétérozygote, dans le cadre du travail réalisé, ne rend pas les modèles de liaison plus puissants que les modèles d'association. Cela peut se comprendre, en outre du travail de Sham *et al.* (2000), par le fait que seulement une moitié des pères sont hétérozygotes, en espérance, lorsque les fréquences alléliques sont équilibrées au QTL. La figure 7.30 donne un exemple de la puissance analytique approchée pour le modèle de liaison homoscédastique, dans le cadre du schéma polygénique de base, lorsque tous les descendants sont considérés informatifs. Cet exemple montre que ce modèle de liaison reste moins puissant que les modèles d'association, bien que l'on utilise tous les pères hétérozygotes et descendants informatifs.

Courbes de puissance analytique approchée :

■ : analyse d'association homoscédastique / ■ : analyse d'association corrigée

■ : analyse de liaison homoscédastique

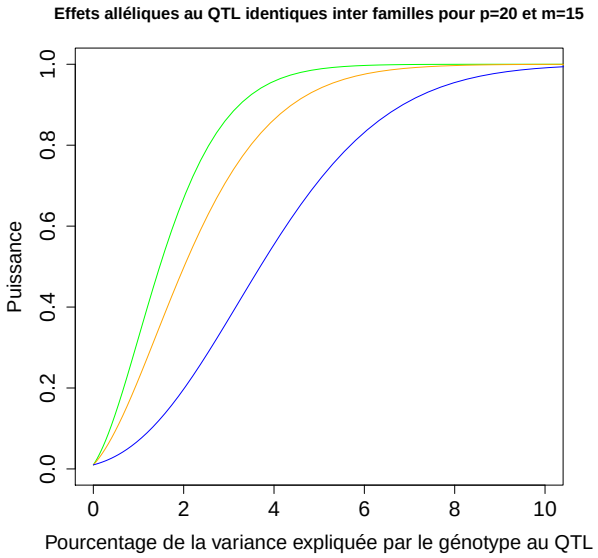


FIGURE 7.30 – Puissances analytiques approchées pour $p = 20$ et $m = 15$ dans le cas d'effets alléliques au QTL identiques inter familles, en considérant l'ensemble des descendants.

Si beaucoup de populations animales sont composées de familles de demi-frères, ce qui

justifie l'hypothèse faite, les résultats obtenus ne l'ont été que dans ce cadre. La levée de cette hypothèse, en considérant une population structurée autrement, pourrait changer les résultats obtenus. Plus d'études devraient donc être menées sur d'autres structures afin d'apprécier la puissance, et la robustesse, des modèles d'association et des modèles de liaison.

L'étude réalisée a montré que les différences moyennes entre les puissances analytiques approchées et les puissances estimées sont plus petites pour le modèle d'association corrigé pour la structure, par rapport aux autres modèles comparés. Cela montre que ce modèle est le moins influencé par la levée de ses hypothèses selon les différents schémas de génération et la variance d'échantillonnage lors des simulations (i.e. la déviation aux fréquences espérées sous HWE). Bien que ces différences ont été enflées par la variance d'échantillonnage, on peut supposer que ce modèle est en relativement bonne adéquation avec les schémas *i)* à *iii)*. Les figures 7.31 à 7.33 donnent les courbes de puissance analytique et estimée superposées pour les schémas *i)* à *iii)* lorsque $p = 20$ et $m = 30$. On voit dans ces figures que les courbes analytiques et estimées sont relativement proches pour le modèle d'association corrigé. Dans ces figures, on remarque également que les courbes analytiques justifient la vraisemblance des courbes obtenues par simulations bien que l'on observe certaines différences entre ces courbes.

Courbes de puissance analytique (−) et estimée (−−) :

■ : analyse d'association homoscédastique / ■ : analyse d'association corrigée

■ : analyse de liaison homoscédastique / ■ : analyse de liaison hétéroscédastique

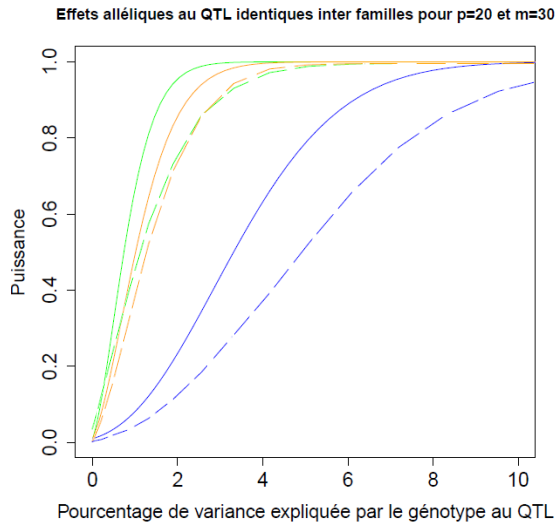


FIGURE 7.31 – Puissances pour $m = 30$ dans le cas d'effets alléliques au QTL identiques inter familles (schéma *i)*).

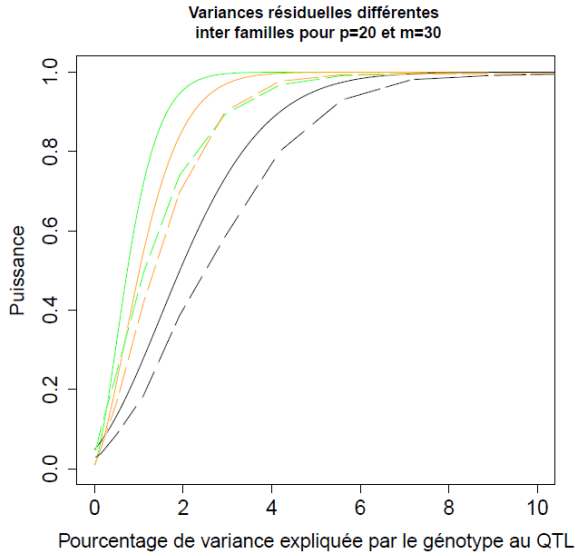


FIGURE 7.32 – Puissances pour $m = 30$ dans le cas de variances résiduelles différentes inter familles (schéma *ii)*).

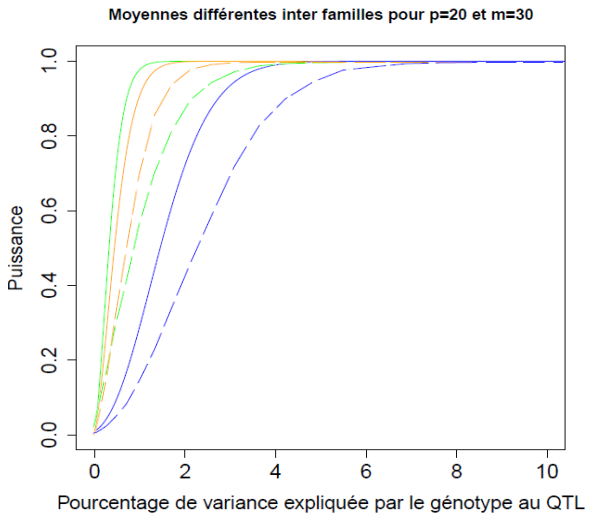


FIGURE 7.33 – Puissances pour $m = 30$ dans le cas de moyennes différentes inter familles (schéma *iii)*).

Remarquons que la vraisemblance des résultats obtenus par Monte-Carlo sous H_1 justifie la vraisemblance de ceux obtenus sous H_0 car les phénotypes sont simulés sous les mêmes conditions, dans les deux situations, à l'exception des effets alléliques au QTL qui ne sont pas présents lorsque l'on se place sous H_0.

Au vu des puissances analytiques et estimées pour le schéma *iv)* considérant l'interaction, il est clair que les modèles discriminés ne sont pas en adéquation avec ce schéma. De manière générale on a $\mathbb{E}[\hat{\beta}] = \beta$ lorsqu'un modèle statistique est en adéquation avec les processus qui génèrent les données, où β est une grandeur d'intérêt intervenant dans la génération des dernières. Pour le schéma *iv)* on montre que $\mathbb{E}[\hat{\alpha}] = \frac{3}{32}\alpha \neq \alpha$ pour les modèles d'association et $\mathbb{E}[\hat{\delta}_i] = \frac{3}{16}\alpha \neq 2\alpha = \delta_i$ pour les modèles de liaison (cf. compléments en Annexe A). En d'autres termes, les modèles statistiques considérés sont biaisés pour l'estimation des effets alléliques au QTL. Ce cas montre la nécessité d'avoir une connaissance approfondie des processus générant les phénotypes afin de construire des modèles plus robustes sous l'alternative. On peut également montrer pour les schémas *i)* à *iii)* que $\mathbb{E}[\hat{\alpha}] = \alpha$ pour les modèles d'association et $\mathbb{E}[\hat{\delta}_i] = 2\alpha = \delta_i$ pour les modèles de liaison.

Toutefois, cette étude a montré que la situation où les modèles de liaison ont un clair avantage par rapport aux modèles d'association, est celle où il existerait un faible LD populationnel entre un marqueur et un QTL. Autrement dit, cette situation correspond au cas où la carte génétique utilisée ne serait pas suffisamment dense pour bien capter le LD populationnel entre un marqueur testé et un QTL. Dans ce travail, les modèles d'association ont montré moins de puissance que les modèles de liaison lorsque le LD entre

le marqueur testé et le QTL est inférieur à 0.5. Ce résultat est en accord avec celui cité dans Sham *et al.* (2000), qui est que l'on ne détecte pas correctement le LD populationnel entre un marqueur et un QTL à plus de 20Kb. A une telle distance physique on peut observer un r^2 moyen inférieur à 0.5 (cf. figure 7.25) dans les populations animales avec les puces couramment utilisées.

La capacité des modèles de liaison à détecter des QTL pour des cartes à faible densité provient de leur construction qui leur permet de toujours s'affranchir du LD populationnel. En effet, on a vu que le LD intra famille exploité par les modèles LA peut être maximal, quand celui mesuré sur l'ensemble de la population est nul. Cependant, cette construction est en contrepartie responsable du manque de précision des analyses de liaison pour la cartographie de QTL, car le LD intra famille est un phénomène toujours présent et qui s'étend sur de grandes distances. Les analyses de liaison ne sont donc pas la méthode de choix pour la cartographie fine de QTL.

Des modèles dits LDLA ("Linkage Disequilibrium Linkage Analysis") ont été développés afin de concilier les avantages des approches LDA et LA. Le modèle LDLA le plus connu et utilisé en pratique est celui de Meuwissen *et al.* (2002). Ce modèle fait partie de la classe des modèles mixtes définie par l'équation (2.3) en 2.2.1.2, i.e. c'est un modèle prenant en compte des effets haplotypiques et polygéniques en aléatoires. Les probabilités d'IBD entre haplotypes contenues dans la matrice H de ce modèle sont calculées par la méthode proposée dans Meuwissen et Goddard (2001). Les probabilités d'IBD entre les individus non fondateurs du pedigree sont calculées par la transmission, ce qui correspond à la partie LA, et les probabilités d'IBD entre les fondateurs du pedigree sont calculées par la fonction $P(IBD)$ décrite en 4.2.3, ce qui correspond à la partie LD telle que décrite par les auteurs (Uleberg et Meuwissen, 2007).

Or, on a vu que les prédictions non discrètes dans [0,1] (i.e. les éléments de la matrice H) mènent à une détérioration de la prise en compte du LD par l'utilisation d'haplotypes. Ainsi, les modèles LDLA ne sont peut-être pas adaptés pour une cartographie fine par l'exploitation du LD. Legarra et Fernando (2009) ont proposé de simplifier cela en réduisant les probabilités d'IBD dans cette stratégie à l'observation des états IBS entre les haplotypes fondateurs. Hayes *et al.* (2006) ont montré que le modèle LDLA proposé par Meuwissen *et al.* (2002) perd en précision de cartographie, et gagne en puissance, lorsqu'il

y a peu de recombinaisons. La raison évoquée par ces auteurs à ce phénomène est que le LD capté par le modèle, entre les haplotypes et le QTL, n'est pas suffisamment local lorsqu'il y a peu de recombinaisons.

Finalement, il peut être intéressant de conserver les approches LA pour une détection en cas d'une faible densité de marquage, ou pour renforcer la validité des résultats de cartographie obtenus par les analyses d'association. Cette logique s'explique par le fait que le LD intra-famille est toujours présent et assez fort entre une position testée et un QTL, mais ce LD sera aussi d'autant plus fort lorsque le LD populationnel sera très fort entre ces positions. Ainsi un QTL détecté par une analyse d'association devrait, en principe, avoir des chances d'être détecté par une analyse de liaison alors que la réciproque n'est pas forcément vraie (Greenberg, 1993). Cependant, Greenberg (1993) évoque des situations réelles où l'association détectée entre des allèles et un phénotype ne mène pas à une liaison détectée entre ces derniers. L'une des raisons qu'il évoque pour ces situations est une faible proportion de la variance phénotypique expliquée par ces allèles, que l'on ne peut détecter que par des analyses d'association. Cette faible proportion de variance étant fréquemment le résultat d'interactions entre loci sur le génome. Une autre raison possible est qu'il arrive aussi, parfois, qu'il n'y ait aucun père hétérozygote pour l'analyse de liaison (Frésard *et al.*, 2012)

Bilan et perspectives

Rappel sur les objectifs et le contexte de la thèse

Les objectifs de cette thèse étaient de discriminer entre des modèles utilisés en routine en cartographie de QTL et d'apporter des éclaircissements sur la meilleure façon d'exploiter le LD, via l'utilisation des données omiques, afin d'optimiser les méthodes de cartographie fine. Parmi les nombreux modèles proposés dans la littérature, nous avons distingué deux catégories de modèles : le LA et le LDA, qui sont établis dans leur utilisation routinière en cartographie de QTL. Le LDA est devenue l'approche de choix et l'utilité des modèles LA est remise en question avec le génotypage à haut débit.

La partie III de cette thèse vise à donner des réponses à cette question. Dans cette partie, des modèles uni-SNP de liaison sont comparés avec des modèles uni-SNP d'association, par rapport à leurs puissances et robustesses statistiques en cartographie de QTL. Cependant, les approches uni-SNP comportent un défaut majeur qui est leur sensibilité aux associations à longue distance dû au LD biallélique qui peut s'étendre assez loin, voir exister entre des chromosomes. Les réponses données dans cette partie ne tiennent pas compte de ce LD entre chromosomes, sachant que ce dernier ne s'observe pas systématiquement sur le génome (Weiss et Clark, 2002).

Afin de pallier le caractère non systématiquement local du LD biallélique, des approches haplotypiques ont été proposées. Outre leur meilleure exploitation locale du LD, ces méthodes permettent d'effectuer une inférence sur les allèles non observés en des QTL putatifs. Diverses méthodes haplotypiques ont été proposées dans ce cadre. Aucune ne semble faire l'objet d'un consensus auprès de la communauté des chercheurs en cartographie de QTL. L'objectif de la partie II a été de proposer une méthode de discrimination

entre ces approches haplotypiques et de traduire algébriquement leur comportement par rapport au LD multiallélique. L'objectif de l'étude algébrique a été de comprendre les mécanismes sous-jacents à la cartographie de QTL par l'exploitation du LD multiallélique capté par les méthodes haplotypiques. Cette étude a été nécessaire afin d'apporter des éclaircissements sur la meilleure façon d'exploiter le LD via l'utilisation des données omiques.

Résultats obtenus par rapport aux objectifs

J'ai proposé dans la première partie une méthode, basée sur une distance matricielle, qui aide à mieux discriminer entre les méthodes haplotypiques effectuant une inférence sur les allèles non observés à une position testée sur des chromosomes (i.e. une prédiction d'identité allélique). Cette méthode a permis de valider des résultats de cartographie et a mis en évidence qu'une méthode simple, telle que l'IBS entre haplotypes, suffit pour cartographier finement un QTL par l'utilisation du LD multiallélique. Ce dernier résultat est tout de même conditionné par la structure des données, à savoir que l'on doit disposer d'une carte génétique suffisamment dense et qu'il n'y ait pas un niveau important d'erreur de phasage des chromosomes. Sous réserve d'existence de variants causaux, un marquage très dense devrait permettre de capter le LD multiallélique entre des haplotypes et ces allèles cachés.

J'ai également montré dans cette partie qu'il peut être difficile de construire un prédicteur d'identité allélique en des loci non observés qui ne soit pas uniquement basé sur la similarité totale entre des haplotypes et qui tienne bien compte du LD entre ces derniers et les allèles au QTL. En effet, j'ai pu montrer que la prise en compte de la similarité partielle entre deux haplotypes, outre la similarité totale, détériore l'exploitation du LD entre ces derniers et les allèles au QTL. Ce résultat, bien qu'il ne soit démontré que dans un cas particulier, met en défaut l'hypothèse généralement avancée dans la littérature (Meuwissen et Goddard, 2001 ; Meuwissen *et al.*, 2002 ; Li et Jiang, 2005), qui est que l'on exploite mieux le LD entre les allèles au QTL et des haplotypes si l'on tient compte de similarité partielle et totale entre ces derniers. Or, ce LD multiallélique est celui que l'on cherche à exploiter au mieux pour la cartographie de QTL.

A la lumière des résultats algébriques de cette partie, il est possible que l'IBS entre haplotypes soit la méthode de choix pour l'analyse d'un marquage à haute densité exhibant des forts niveaux de LD multiallélique localement. Or, avec le séquençage complet des individus cette possibilité deviendra sans doute une réalité. Les résultats algébriques ont montré que l'IBS entre haplotypes possède de bonnes propriétés théoriques par rapport au LD multiallélique.

J'ai étudié dans la seconde partie les puissances et les robustesses associées à des modèles uni-SNP d'association et des modèles uni-SNP de liaison, sous l'hypothèse d'équilibre de Hardy-Weinberg au QTL. Cette étude a été faite à la fois sur le plan théorique et par simulation. Les résultats obtenus dans cette partie correspondent globalement à ceux cités dans littérature. L'étude a montré que les modèles de liaison, et les modèles d'association tenant compte de la structure génétique, sont robustes aux faux positifs pour l'essentiel des cas de simulation étudiés. Cependant, j'ai aussi montré dans cette partie que les modèles d'association sont robustes aux faux positifs si les paramètres (variances) sont correctement estimés et que la structure des données est correctement prise en compte.

L'étude a également montré que la différence entre la puissance analytique approchée et la puissance estimée du test de Fisher est plus petite pour le modèle d'association corrigé. Elle montre ainsi que ce modèle est le moins influencé par la levée de ses hypothèses selon les différents schémas de simulation considérés. Finalement, les simulations dans cette partie ont montré que les modèles de liaison comportent un réel intérêt, contrairement aux modèles d'association, lorsque les cartes génétiques ne sont pas suffisamment denses. La capacité des modèles LA à détecter des QTL pour des cartes à faible densité provient de leur construction, qui les rend antagonistes, dans une certaine mesure, aux modèles LDA. Cependant, il semblerait que les analyses effectuées en pratique se tournent davantage vers les analyses d'association avec des puces d'au moins 50K dans beaucoup d'espèces (porcins, caprins, ovins, bovins..).

Au vu des résultats obtenus dans cette partie, il peut être intéressant de conserver les approches LA afin d'effectuer une détection de QTL dans les zones de faible LD. Les modèles LA peuvent également aider à renforcer, dans certaines situations, les résultats de cartographie obtenus par les approches LDA. En effet, les QTL détectés par une approche LDA auront des chances d'être détectés par une approche LA alors que la situation inverse

sera moins courante. Cette mécanique s'explique par le fait que le LD intra-famille est toujours présent et assez fort entre une position testée et un QTL, mais ce LD sera aussi d'autant plus fort lorsque le LD populationnel sera très fort entre ces positions. Cependant, il arrive parfois que l'on ne détecte pas des QTL par des approches LA bien que ce soit la situation inverse pour des approches LDA. Il existe diverses raisons expliquant ce cas de figure telles que l'interaction entre loci et des QTL ayant un faible effet sur le phénotype.

Toutefois, bien que les approches LA puissent aider à détecter des associations, ces approches sont peu précises pour la cartographie fine de QTL et elles ne comportent donc pas d'intérêt dans ce sens. De plus, ces approches sont fastidieuses à mettre œuvre dans certaines espèces, telle que chez l'homme par exemple, car elles nécessitent de construire des familles adaptées, ce qui rend l'utilisation des modèles LA moins évidente que les modèles LDA.

Conclusions générales de la thèse

A la lumière de l'ensemble des résultats et perspectives des parties traitées dans le cadre de cette thèse, on peut conclure que les approches d'association utilisant les haplotypes et corrigées pour la structure des données constituent un axe de recherche à privilégier. En effet, outre leur capacité à bien décrire un LD local, ces approches sont cohérentes d'un point de vue biologique. Elles peuvent par exemple rendre compte d'un multiallélisme sous-jacent à un QTL, quelque soit le nombre d'allèles non observés au QTL. Cependant l'application de ces approches nécessite la reconstruction des phases des chromosomes. Cette reconstruction peut être parfois difficile et constitue un facteur limitant dans l'application des modèles utilisant des haplotypes. Le choix de la taille des haplotypes en nombre de marqueurs, selon la densité de marquage, est également un paramètre à déterminer pour l'utilisation de ces modèles.

L'étape suivante à ce travail de thèse conduirait naturellement à l'étude de la puissance et de la robustesse des modèles haplotypiques d'association corrigés pour la structure. Cette étude est importante si l'on veut comprendre les différents facteurs influençant la maîtrise de l'erreur de première espèce et la puissance associées à ces modèles d'analyse. La démarche et les dérivations algébriques nécessaires à cette étude ont été faites dans le

cadre des approches uni-SNP, en effets fixes, qui sont des cas particuliers des approches haplotypiques. Ces éléments peuvent donc être utilisés afin de généraliser l'étude de la puissance et de la robustesse aux modèles utilisant des haplotypes en effets fixes.

Bibliographie

[1] Abdallah J, Goffinet B, Cierco-Ayrolles C, and Pérez-Enciso M. **Linkage disequilibrium fine mapping of quantitative trait loci : a simulation study**. *Genet Sel Evol*, 35 :513–532, 2003.

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

[3] Aranzana M J, Kim S, Zhao K, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang C, Toomajian C, Traw B, Zheng H, Bergelson J, Dean C, Marjoram P, and Nordborg M. **Genome-Wide Association Mapping in Arabidopsis Identifies Previously Known Flowering Time and Pathogen Resistance Genes**. *PLoS Genet*, 1(5) :e60, 2005.

[4] Azaïs J M and Bardet J M. **Le modèle linéaire par l'exemple**. *Dunod, Sciences Sup*, 2006.

[5] Bercovici S, Meek C, Wexler Y, and Geiger D. **Estimating genome-wide IBD sharing from SNP data via an efficient hidden Markov model of LD with application to gene mapping**. *Bioinformatics*, 26(12) :i175–i182, 2010.

[6] Bodmer W F. **Human genetics : the molecular challenge**. *BioEssays*, 7 :41–45, 1987.

[7] Boehnke M. **Limits of resolution of genetic linkage studies : implications for the positional cloning of human disease genes**. *Am J Hum Genet*, 55 :379, 1994.

[8] Boer M P, Wright D, Feng L, Podlich D W, Luo L, Cooper M, and van Eeuwijk F A. **A Mixed-Model Quantitative Trait Loci (QTL) Analysis for Multiple-Environment Trial Data Using Environmental Covariables for QTL-by-Environment Interactions, With an Example in Maize**. *Genetics*, 177 :1801–1813, 2007.

[9] Boitard S. **Cartographie de gènes à caractères quantitatifs par déséquilibre de liaison**. *Thèse, Université Paul Sabatier - Toulouse III*, page 12, 2006.

[10] Boleckova J, Christensen O F, Sørensen P, and Sahana G. **Strategies for haplotype-based association mapping in a complex pedigreed population**. *Czech J Anim Sci*, 1 :1–9, 2012.

[11] Bonnet P and Lansiaux G. **Mémoire de maitrise**. *Université Paul Sabatier*, 1992.

[12] Borman S. **Topics in multiframe superresolution restoration**. *PhD Thesis, University of Notre Dame, Indiana, USA*, pages 225–234, 2004.

[13] Browning S R. **Multilocus association mapping using variable-length Markov chains**. *Am J Hum Genet*, 78 :903–913, 2006.

[14] Browning B L and Browning S R. **Efficient multilocus association testing for whole genome association studies using localized haplotype clustering**. *Genet Epidemiol*, 31 :365–375, 2007.

[15] Browning B L and Browning S R. **Haplotypic analysis of Wellcome Trust Case Control Consortium data**. *Human Genet*, 123 :273–280, 2008.

[16] Bush W S and Moore J H. **Genome-wide association studies**. *PLoS computational biology*, 8(12) :e1002822, 2012.

[17] Calus M P L, Meuwissen T H E, Windig J J, knol E F, Schrooten C, Vereijken A L J, and Veerkamp R F. **Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values**. *Genet Sel Evol*, 41 :11, 2009.

[18] Cardon L R and Abecasis G R. **Using haplotype blocks to map human complex trait loci**. *Trends in Genet*, 19 :136–140, 2003.

[19] Charlesworth B. **Fundamental concepts in genetics : Effective population size and patterns of molecular evolution and variation**. *Nature Reviews Genetics*, 10 :195–205, 2009.

[20] Chen Y, Li X, and Li J. **A novel approach for haplotype-based association analyzis using family data**. *BMC Bioinformatics*, 11(Suppl 1) :S45, 2010.

[21] Chen T and Martin E. **Bayesian linear regression and variable selection for spectroscopic calibration**. *Analytica chimica acta*, 631 :13–21, 2009.

[22] Churchill G A and Doerge R W. **Empirical threshold values for quantitative trait mapping.** *Genetics*, 138 :963–971, 1994.

[23] Cierco-Ayrolles C, Abdallah J, Boitard S, Chikhi L, de Rochambeau H, Tsitrone A, Veyrieras J B, and Mangin B. **On linkage disequilibrium measures : methods and applications, in Recent Research Developments in Genetics and Breeding**. *Research SignPost, Kerala, India*, pages 151–180, 2004.

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

[25] Couvreur C. **The EM Algorithm : A Guided Tour**. *In Proc. 2d IEEE European Workshop on Computationaly Intensive Methods in Control and Signal Processing, Pragues, Czech Republik*, 1996.

[26] Cordell H J. **Epistasis : what it means, what it doesn't mean, and statistical methods to detect it in humans**. *Human molecular genetics*, 11 :2463–2468, 2002.

[27] de Bakker P I, Yelensky R, Pe'er I, Gabriel S B, Daly M J, and Altshuler D. **Efficiency and power in genetic association studies**. *Nature Genetics*, 37 :1217–1223, 2005.

[28] Dempster A P, Laird N M, and Rubin D B. **Maximum likelihood from in-complete data via the em algorithm**. *J R Statist Soc*, Series B, 39(1) :1–38, 1977.

[29] De Roos APW, Hayes B J, Spelman R J, and Goddard M E. **Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle**. *Genetics*, 179 :1503–1512, 2008.

[30] Devlin B and Risch N. **A comparison of linkage disequilibrium measures for fine-scale mapping**. *Genomics*, 29(2) :311–22, 1995.

[31] Druet T and Georges M. **A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and Quantitative Trait Locus Fine Mapping**. *Genetics*, 184 :789–798, 2010.

[32] Druet T, Fritz S, Boussaha M, Ben-Jemaa S, Guillaume F, Derbala D, Zelenika D, Lechner D, Charon C, Boichard D, Gut IG, Eggen A, and Gautier M. **Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map**. *Genetics*, 178(4) :2227–35, 2008.

[33] Durrant C, Zondervan K T, Cardon L R, Hunt S, Deloukas P, and Morris A P. **Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes**. *Am J Hum Genet*, 75 :35–43, 2004.

[34] Elsen JM, Mangin B, Goffinet B, Boichard D, and Le Roy P. **Alternative models for QTL detection in livestock. I General introduction**. *Genet Sel Evol*, 31 :213–224, 1999.

[35] Falconer D and Mackay T. **Introduction to quantitative genetics**. *Longman*, 4, 1996.

[36] Fan R and Xiong M. **High resolution mapping of quantitative trait loci by linkage disequilibrium analysis**. *Eur J Hum Genet*, 10(10) :607–15, 2002.

[37] Fan R, Jung J, and Jin L. **High-Resolution Association Mapping of Quantitative Trait Loci : A Population-Based Approach**. *Genetics*, 172 :663–686, 2006.

[38] Farnir F, Coppieters W, Arranz J J, Berzi P, Cambisano N, Grisart B, Karim L, Marcq F, Moreau L, Mni M, Nezer C, Simon P, Vanmanshoven P, Wagenaar D, and George M. **Extensive genome-wide linkage disequilibrium in cattle**. *Genome Res*, 10 :220–227, 2000.

[39] Fisher R A. **The correlation between relatives on the supposition of Mendelian inheritance**. *Trans Roy Soc Edinb*, 52 :399–433, 1918.

[40] Fisher R A. **On the dominance ratio**. *Proc. Roy. Soc. Edinb.*, 52 :312–341, 1922.

[41] Fisher R A. **The distribution of gene ratios for rare mutations**. *Proc. Roy. Soc. Edinb.*, 50 :205–220, 1930.

[42] Flint-Garcia S A, Thuillet A C, Yu J, Pressoir G, Romero S M, Mitchell S E, Doebley J, Kresovich S, Goodman M M, and Buckler E S. **Maize association population : a high-resolution platform for quantitative trait locus dissection**. *The Plant Journal*, 44 :1054–1064, 2005.

[43] Foster S. **The LASSO Linear Mixed Model for Mapping Quantitative Trait Loci**. *PhD Thesis, University of Adelaide, Australia*, page 60, 2006.

[44] Foulley J L. **EM algorithm : theory and application to the mixed model**. *Journal de la Société Française de Statistique*, 143(3-4) :57–109, 2002.

[45] Foulley J L. **A simple argument showing how to derive restricted maximum likelihood**. *Journal of Dairy Science*, 76 :2320–2324, 1993.

[46] Frésard L, Leroux S, Dehais P, Servin B, Gilbert H, Bouchez O, Klopp C, Cabau C, Vignoles F, Feve K, et al. **Fine mapping of complex traits in non-model species : using next generation sequencing and advanced intercross lines in Japanese quail**. *BMC genomics*, 13 :551, 2012.

[47] Fulker D W, Cherny S S, Sham P C, and Hewitt J K. **Combined linkage and association sib-pair analysis for quantitative traits**. *Am J Hum Genet*, 64 :259–267, 1999.

[48] Gianola D, Foulley J L, and Fernando R. **Prediction of breeding values when variances are not known**. *Genet Sel Evol*, 18 :485–498, 1986.

[49] Gianola D, Campos G De Los, Hill W G, Manfredi E, and Fernando R. **Additive Genetic Variability and the Bayesian Alphabet**. *Genetics*, 183 :347–363, 2009.

[50] Goffinet B, Le Roy P, Boichard D, Elsen J M, and Mangin B. **Alternative models for QTL detection in livestock. III. Heteroskedastic model and models corresponding to several distributions of the QTL effect**. *Genet Sel Evol*, 31 :341–350, 1999.

[51] Gourieroux C and Montfort A. **Statistique et modèles économétriques**. *Economica, France*, 1989.

[52] Grapes L, Firat M Z, Dekkers J C M, Rothschild M F, and Fernando R L. **Optimal haplotype structure for linkage disequilibrium-based fine mapping of quantitative trait loci using identity by descent**. *Genetics*, 172 :1955–1965, 2005.

[53] Grapes L, Dekkers J C M, Rothschild M F, and Fernando R L. **Comparing Linkage Disequilibrium-Based Methods for Fine Mapping Quantitative Trait Loci**. *Genetics*, 166 :1561–1570, 2004.

[54] Greenberg D A. **Linkage analysis of "necessary" disease loci versus "susceptibility" loci**. *Am J Hum Genet*, 52 :135, 1993.

[55] Guillaume F. **Intégration de l'information moléculaire dans l'évaluation génétique**. *Thèse, AgroParisTech*, page 14, 2009.

[56] Guyader A. **Régression linéaire**. *Université Rennes 2*, pages 60–61, 2011.

[57] Guyon X. **Le modèle linéaire et ses généralisations**. *Université Paris 1 - Statistique Appliquée Modélisation Stochastique (SAMOS)*, page 11, 2005.

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

[59] Haley C S and Knott S A. **A simple regression method for mapping quantitative trait loci in line crosses using flanking markers**. *Heredity*, 69 :315–324, 1992.

[60] Harville D A. **Bayesian inference for variance components using only error contrasts**. *Biometrika*, 61 :383–385, 1974.

[61] Harville D A. **Maximum likelihood approaches to variance component estimation and to related problems**. *J Amer Statist Assoc*, 72 :320–340, 1977.

[62] Hayes B J, Gjuvsland A, and Omholt S. **Power of QTL mapping experiments in commercial Atlantic salmon populations, exploiting linkage and linkage disequilibrium and effect of limited recombination in males**. *Heredity*, 97 :19–26, 2006.

[63] Hayes B J, Pryce J, Chamberlain A J, Bowman P J, and Goddard M E. **Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction : Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits**. *PLoS Genet*, 6(9) :e1001139. doi :10.1371/journal.pgen.1001139, 2010.

[64] He W, Fernando R L, Dekkers J C M, and H H, Gilbert. **A gene frequency model for QTL mapping using Bayesian inference**. *Genet Sel Evol*, 42 :1–21, 2010.

[65] Hedrick P W and Thomson G. **A two-locus neutrality test : applications to humans, *E. coli* and Lodgepole pine**. *Genetics*, 112 :135–156, 1985.

[66] Hedrick P W. **Gametic Disequilibrium Measures : Proceed With Caution**. *Genetics*, 117 :331–341, 1987.

[67] Henderson C R. **Estimation of genetic parameters (abstract)**. *Ann Math Statist*, 21 :309–310, 1950.

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

[69] Henderson C R. **Best linear unbiased estimation and prediction under a selection model**. *Biometrics*, 31 :423–447, 1975.

[70] Henderson C R. **Applications of Linear Models in Animal Breeding**. *University of Guelph, Guelph, Third Edition*, 1984.

[71] Hill W G and Robertson A. **Linkage disequilibrium in finite populations**. *Theor Appl Genet*, 38 :226–231, 1968.

[72] Hoerl A E and Kennard R W. **Ridge Regression : Applications to Nonorthogonal Problems**. *Technometrics*, 12 :55–68, 1970.

[73] Huber P J. **Robust statistics**. page 157, 1981.

[74] Jorde L B. **Linkage disequilibrium and the search for complex disease genes**. *Genome Res*, 10 :1435–1444, 2000.

[75] Jung J, Fan R, and Jin L. **Combined linkage and association mapping of quantitative trait loci by multiple markers**. *Genetics*, 170 :881–898, 2005.

[76] Kao C H. **On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci**. *Genetics*, 156 :855–865, 2000.

[77] Knott S A, Elsen J M, and Haley C S. **Methods for multiple-marker mapping of quantitative trait loci in half-sib populations**. *Theor Appl Genet*, 93 :71–80, 1996.

[78] Knüppel S, Esparza-Gordillo J, Marenholz I, Holzhütter H G, Bauerfeind A, Ruether A, Weidinger S, Lee Y A, and Rohde K. **Multi-locus stepwise regression : a haplotype based algorithm**

for finding genetic associations applied to atopic dermatitis. *BMC Medical Genetics*, 13 :8, 2012.

[79] Kolbehdari D, Jansen G B, Schaeffer L R, and Allen B O. **Power of QTL detection by either fixed or random models in half-sib designs**. *Genet Sel Evol*, 37 :601–614, 2005.

[80] Korte A and Farlow A. **The advantages and limitations of trait analysis with GWAS : a review**. *Plant methods*, 9 :29, 2013.

[81] LaMotte L R. **A direct derivation of the REML likelihood function**. *Statistical Papers*, 48 :321–327, 2007.

[82] Lander E S and Botstein D. **Mapping mendelian factors underlying quantitative traits using RFLP linkage maps**. *Genetics*, 121 :185–199, 1989.

[83] Legarra A and Fernando R L. **Linear models for joint association and linkage QTL mapping**. *Genet Sel Evol*, 41 :43–59, 2009.

[84] Le Roy P P, Elsen J M, Boichard D, Mangin M, Bidanel JP, and Goffinet B. **An algorithm for QTL detection in mixture of full and half sib families**. *In 6th WCGALP. Armindale*, 26 :257–260, 1998.

[85] Lewontin R C. **On measures of gametic disequilibrium**. *Genetics*, 49 :49–67, 1964.

[86] Li J and Jiang T. **Haplotype-based linkage disequilibrium mapping via direct data mining**. *Bioinformatics*, 21 :4384–4393, 2005.

[87] Li M, Wing H W, and Art B O. **A Sparse Transmission Disequilibrium Test for Haplotypes Based on Bradley-Terry Graphs**. *Hum Hered*, 73 :52–61, 2012.

[88] Lin W Y, Yi N, Zhi D, Zhang K, Gao G, Tiwari H K, and Liu N. **Haplotype-Based Methods for Detecting Uncommon Causal Variants With Common SNPs**. *Genet Epidemiol*, 36 :572–582, 2012.

[89] Lindley D V and Smith A FM. **Bayes estimates for the linear model**. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–41, 1972.

[90] Long A D and Langley C H. **The Power of Association Studies to Detect the Contribution of Candidate Genetic Loci to Variation in Complex Traits**. *Genome Res*, 9 :720–731, 1999.

[91] Mailund T, Besenbacher S, and Schierup M. **Whole genome association mapping by incompatibilities and local perfect phylogenies**. *BMC Bioinformatics*, 7 :454, 2006.

[92] Manenti G, Galvan A, Pettinicchio A, Trincucci G, Spada E, Zolin A, Milani S, Gonzalez-Neira A, and A. Dragani T. **Mouse Genome-Wide Association Mapping Needs Linkage Analysis to Avoid False-Positive Loci**. *PLoS Genet*, 5(1) :e1000331. doi :10.1371/journal.pgen.1000331, 2009.

[93] Maruyama T. **Stochastic integrals and their application to population genetics**. *In : M. Kimura (Ed.), Molecular Evolution, Protein Polymorphism and the Neutral Theory. Springer-Verlag, Berlin.*, pages 151–166, 1982.

[94] Martinez O and Curnow R N. **Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers**. *Theor Appl Genet*, 85 :480–488, 1992.

[95] Mathews K L, Malosetti M, Chapman S, McIntyre L, Reynolds M, Shorter R, and van Eeuwijk F. **Multi-environment QTL mixed models for drought stress adaptation in wheat**. *Theor Appl Genet*, 117(7) :1077–1091, 2008.

[96] Maurer H P, Knaak C, Melchinger A E, Ouzunova M, and Frisch M. **Linkage disequilibrium between SSR markers in six pools of elite lines of an european breeding program for hybrid maize**. *Maydica*, 51 :269–279, 2006.

[97] McQueen M B, Murphy A, Kraft P, Su J, Lazarus R, Laird N M, Lange C, and Van Steen K. **Comparison of linkage and association strategies for quantitative traits using the COGA dataset**. *BMC genetics*, 6 :S96, 2005.

[98] Meuwissen T H E and Goddard M E. **Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci**. *Genetics*, 155 :421–430, 2000.

[99] Meuwissen T H E and Goddard M E. **Prediction of identity by descent probabilities from marker haplotypes**. *Genet Sel Evol*, 33 :605–634, 2001.

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

[101] Meuwissen T H E, Karlsen A, Lien S, Olsaker I, and Goddard M E. **Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping**. *Genetics*, 161 :373–379, 2002.

[102] Minichiello M J and Durbin R. **Mapping trait loci by use of inferred ancestral recombination graphs**. *Am J Hum Genet*, 79 :910–922, 2006.

[103] Newman DL, Abney M, McPeek MS, Ober C, and Cox NJ. **The Importance of Genealogy in Determining Genetic Associations with Complex Traits**. *Am J Hum Genet*, 69(5) :1146–8, 2001.

[104] Onkamo P, Ollikainen V, Sevon P, Toivonen H, Mannila H, and Kere J. **Association analysis for quantitative traits by data mining : QHPM**. *Annals of human genetics*, 66 :419–429, 2002.

[105] Patterson H D and Thompson R. **Recovery of inter-block information when block sizes are unequal**. *Biometrika*, 58 :545–554, 1971.

[106] Platt A, Vilhjalmsson B J, and Nordborg M. **Conditions Under Which Genome-Wide Association Studies Will be Positively Misleading**. *Genetics*, 186 :1045–1052, 2010.

[107] Pong-Wong R, George A W, Woolliams J A, and Haley C S. **A simple and rapid method for calculating identity-by-descent matrices using multiple markers**. *Genet Sel Evol*, 33 :453–471, 2001.

[108] Ramos A M, Crooijmans R P, Affara N A, Amaral A J, Archibald A L, Beever J E, Bendixen C, Churcher C, Clark R, Dehais P, Hansen M S, Hedegaard J, Hu Z L, Kerstens H H, Law A S, Megens

H J, Milan D, Nonneman D J, Rohrer G A, Rothschild M F, Smith T P, Schnabel R D, Van Tassell C P, Taylor J F, Wiedmann R T, Schook L B, and Groenen M A. **Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology**. *PLoS ONE*, 4(8) :e6524. doi :10.1371/journal.pone.0006524, 2009.

[109] Rao C R. **Linear Statistical Inference and its Applications**. 2^{nd} *edition,Wiley, New-York*, 1973.

[110] Rao C R and Kleffe J. **Estimation of variance components and applications**. *North Holland series in statistics and probability, Elsevier, Amsterdam*, pages 61–63, 1988.

[111] Remington D L, Thornsberry J M, Matsuoka Y, Wilson L M, Whitt S R, Doebley J, Kresovich S, Goodman M M, and Buckler E S. **Structure of linkage disequilibrium and phenotypic associations in the maize genome**. *Proc Natl Acad Sci*, 98 :11479–11484, 2001.

[112] Risch N and Merikangas K. **The future of genetic studies of complex human diseases**. *Science*, 273 :1516–1517, 1996.

[113] Risch N J. **Searching for genetic determinants in the new millennium**. *Nature*, 405 :847–856, 2000.

[114] Roldan D L, Gilbert H, Henshall J M, Legarra A, and Elsen J M. **Fine-mapping quantitative trait loci with a medium density marker panel : efficiency of population structures and comparison of linkage disequilibrium linkage analysis models**. *Genet Res Camb*, 94(4) :223–234, 2012.

[115] Ron D, Singer Y, and Tishby N. **On the learnability and usage of acyclic probabilistic finite automata**. *J Comp Syst Sci*, 56 :133–152, 1998.

[116] Rosenthal J S. **Rates of convergence for Gibbs sampling for variance component models**. *The Annals of Statistics*, 23 :740–761, 1995.

[117] Schaid D J. **Evaluating associations of haplotypes with traits**. *Genet Epidemiol*, 27 :348–364, 2004.

[118] Searle S R. **Linear Models for Unbalanced Data**. *John Wiley & Sons, New York, NY*, 1987.

[119] Searle S R. **Variance components — some history and a summary account of estimation methods**. *J Anim Breed Genet*, 106(1-6) :1–29, 1989.

[120] Searle S R, Casella G, and McCulloch C E. **Variance components**. *John Wiley & Sons*, 1992.

[121] Self S G and Liang K Y. **Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions**. *J Amer Statist Assoc*, 82 :605–610, 1987.

[122] Sham P C, Cherny S S, Purcell S, and Hewitt J K. **Power of Linkage versus Association Analysis of Quantitative Traits, by Use of Variance-Components Models, for Sibship Data**. *Am J Hum Genet*, 66 :1616–1630, 2000.

[123] Slatkin M. **Disequilibrium mapping of a quantitative-trait locus in an expanding population**. *Am J Hum Genet*, 64 :1765–1773, 1999.

[124] Stephens M, Smith N J, and Donnelly P. **A new statistical method for haplotype reconstruction from population data**. *Am J Hum Genet*, 68 :978–989, 2001.

[125] Scheet P and Stephens M. **A fast and flexible statistical model for large-scale population genotype data : applications to inferring missing genotypes and haplotypic phase**. *Am J Hum Genet*, 78 :629–644, 2006.

[126] Smith A B, Cullis B R, and Thompson R. **The analysis of crop cultivar breeding and evaluation trials : An overview of current mixed model approaches**. *Journal of Agricultural Science*, 143 (6) :449–462, 2005.

[127] Strachan T and Read A P. **Human Molecular Genetics. 2nd edition**. *New York : Wiley-Liss*, 1999.

[128] Stranger B E, Stahl E A, and Towfique R. **Progress and promise of genome-wide association studies for human complex trait genetics**. *Genetics*, 187 :367–383, 2011.

[129] Terwilliger J D and Weiss K M. **Linkage disequilibrium mapping of complex disease : fantasy or reality ?** *Curr Opin Biotechnol*, 9 :578–594, 1998.

[130] Teyssèdre S. **Dissection génétique des caractères par analyse de liaison et d'association : aspects méthodologiques et application à la sensibilité à l'ostéochondrose chez les Trotteurs Français**. *Thèse, Université Paul Sabatier - Toulouse III*, page 47, 2011.

[131] Tibshirani R. **Regression shrinkage and selection via the lasso**. *Journal of the Royal Statistical Society*, 58 :267–288, 1996.

[132] Tiku M L and Akkaya A D. **Robust estimation and hypothesis testing**. *New Age International*, 2004.

[133] Uleberg E and Meuwissen T H E. **Fine mapping of multiple QTL using combined linkage and linkage disequilibrium mapping–A comparison of single QTL and multi QTL methods**. *Genetics Selection Evolution*, 39 :285–299, 2007.

[134] van Eeuwijk F A, Boer M, Totir L R, Bink M, Wright D, Winkler C R, Podlich D, Boldman K, Baumgarten A, Smalley M, Arbelbide M, ter Braak C J F, and Cooper M. **Mixed model approaches for the identification of QTLs within a maize hybrid breeding program**. *Theor Appl Genet*, 120 :429–440, 2010.

[135] Wang D L, Zhu J, Li Z K, and Paterson A H. **Mapping QTLs with epistatic effects and QTL×environment interactions by mixed linear model approaches**. *Theor Appl Genet*, 99 :1255–1264, 1999.

[136] Weiss K M and Clark A G. **Linkage disequilibrium and the mapping of complex human traits**. *Trends in Genet*, 18 :19–24, 2002.

[137] Wright S. **Evolution in Mendelian populations**. *Genetics*, 16 :97–159, 1931.

[138] Yi Y and Wang X. **Comparison of Wald, Score, and Likelihood Ratio Tests for Response Adaptive Designs**. *Journal of Statistical Theory and Applications*, 10 :553–569, 2011.

[139] Ytournel F, Gilbert H, and Boichard D. **Concordance between IBD probabilities and linkage disequilibrium**. *EAAP, Dublin*, Abstract number :1248, 2007.

[140] Ytournel F, Boichard D, Gilbert H, and Legarra A. **LDSO : A complete program for the simulation of pedigrees and molecular information under various evolutionary forces**. *J Anim Breed Genet*, 129 :417–421, 2012.

[141] Yu J, Pressoir G, Briggs W H, Vroh Bi I, Yamasaki M, Doebley J F, McMullen M D, Gaut B S, Nielsen D M, Holland J B, Kresovich S, and Buckler E S. **A unified mixed model method for association mapping that accounts for multiple levels of relatedness**. *Nat Genet*, 38 :203–8, 2005.

[142] Zhang Z, Buckler E S, Casstevens T M, and Bradbury P J. **Software engineering the mixed model for genome-wide association studies on large samples**. *Briefings in bioinformatics*, 10(6) :664–675, 2009.

[143] Zou H and Hastie T. **Regularization and variable selection via the elastic net**. *J R Statist Soc B*, 67, Part 2 :pp. 301–320, 2005.

Annexes

Annexe A : Compléments sur l'espérance des estimateurs pour la partie III

Dans le cadre des schémas de simulation décrits en 7.2.2.1 à 7.2.2.4, $\mathbb{E}[\hat{\alpha}]$ et $\mathbb{E}[\hat{\delta}_i]$ sont données par :

i) Pour le schéma avec des effets alléliques identiques inter familles (cf. 7.2.2.1) :

$$\begin{aligned}\mathbb{E}[\hat{\alpha}] &= \frac{1}{4nf_{a_1}f_{a_2}}\left[\sum_{i=1}^{p}\Big[mf_{a_1}^2\mathbb{E}[Y_{i2k}] - mf_{a_2}^2\mathbb{E}[Y_{i-2k}] + (f_{a_2}-f_{a_1})(mf_{a_1}^2\mathbb{E}[Y_{i2k}] + mf_{a_2}^2\mathbb{E}[Y_{i-2k}])\Big]\right] \\ &= \frac{1}{4nf_{a_1}f_{a_2}}\left[\sum_{i=1}^{p}\Big[mf_{a_1}^2 2\alpha + mf_{a_2}^2 2\alpha + (f_{a_2}-f_{a_1})(mf_{a_1}^2 2\alpha - mf_{a_2}^2 2\alpha)\Big]\right] = \alpha\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\hat{\delta}_i] &= \frac{2}{m}\Big(\frac{mf_{a_1}}{2}\mathbb{E}[Y_{i1s}] - \frac{mf_{a_2}}{2}\mathbb{E}[Y_{i-1s}]\Big) = f_{a_1}\mathbb{E}[Y_{i1s}] - f_{a_2}\mathbb{E}[Y_{i-1s}] = f_{a_1}\mathbb{E}[Y_{i2k}] - f_{a_2}\mathbb{E}[Y_{i-2k}] \\ &= 2\alpha(f_{a_1}+f_{a_2}) = 2\alpha = \delta_i\end{aligned}$$

Remarque : les individus ayant reçu l'allèle a_1 d'un père hétérozygote doivent impérativement être homozygotes a_1a_1, ici, pour l'analyse de liaison car les génotypes des mères ne sont pas connus, d'où l'on sait que $\mathbb{E}[Y_{i1s}] = \mathbb{E}[Y_{i2k}]$ (idem pour a_2).

ii) Pour le schéma avec des variances résiduelles différentes inter familles (cf. 7.2.2.2) :

$\mathbb{E}[\hat{\alpha}] = \alpha$ (car le calcul est le même que pour le schéma défini en *i)*

$\mathbb{E}[\hat{\delta}_i] = 2\alpha = \delta_i$ (idem)

iii) Pour le schéma avec des moyennes différentes inter familles (cf. 7.2.2.3) :

$$\mathbb{E}[\hat{\alpha}] = \frac{1}{4nf_{a_1}f_{a_2}}\left[\sum_{i=1}^{p}\Big[mf_{a_1}^2(\mu^i+2\alpha) - mf_{a_2}^2(\mu^i-2\alpha) + (f_{a_2}-f_{a_1})(mf_{a_1}^2(\mu^i+2\alpha) + 2mf_{a_1}f_{a_2}\mu^i + mf_{a_2}^2(\mu^i-2\alpha))\Big]\right] = \alpha$$

$$\mathbb{E}[\hat{\delta}_i] = f_{a_1}\mathbb{E}[Y_{i2k}] - f_{a_2}\mathbb{E}[Y_{i-2k}] = f_{a_1}(\mu^i+2\alpha) - f_{a_2}(\mu^i-2\alpha) = \mu^i(f_{a_1}-f_{a_2}) + 2\alpha$$
$$= 2\alpha = \delta_i \quad \text{si} \quad f_{a_1} = f_{a_2} = \frac{1}{2}$$

iv) Pour le schéma avec des effets alléliques au QTL en interaction avec un locus (cf. 7.2.2.4) :

D'abord remarquons que dans le cadre de ce schéma on a $\mathbb{E}[Y_{i2k}] = f_{b_1}^2 f_{a_1}^2 2\alpha + 2f_{b_1}f_{b_2}f_{a_1}^2 2\alpha + f_{b_2}^2 f_{a_1}^2.0 = 2f_{a_1}^2\alpha(1-f_{b_2}^2)$, $\mathbb{E}[Y_{i0k}] = f_{b_1}^2.2f_{a_1}f_{a_2}\alpha + 2f_{b_1}f_{b_2}.2f_{a_1}f_{a_2}\alpha + 0 = 2f_{a_1}f_{a_2}\alpha(1-f_{b_2}^2)$ et $\mathbb{E}[Y_{i-2k}] = 0$. En supposant que $f_{b_1} = f_{b_2} = \frac{1}{2}$, on montre (après simplification) que :

$$\mathbb{E}[\hat{\alpha}] = \frac{3}{4}\Big[f_{a_1}\alpha(f_{a_1}^2 + f_{a_2}^2 - f_{a_1}f_{a_2})\Big] = \frac{3\alpha}{32} \neq \alpha \text{ pour } f_{a_1} = f_{a_2} = \frac{1}{2}$$
$$\mathbb{E}[\hat{\delta}_i] = \frac{3}{2}f_{a_1}^3\alpha = \frac{3\alpha}{16} \neq 2\alpha \text{ pour } f_{a_1} = f_{a_2} = \frac{1}{2}$$

Annexe B : Liste des équations et liste des définitions, propositions et théorèmes

Liste des équations

Liste de définitions, propositions et théorèmes

Annexe C : Liste des tableaux et table des figures

Liste des tableaux

Table des figures