



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 :

Simon Teyssède

le : lundi 14 novembre 2011

Titre :

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

JURY

Stéphane Robin
Sophie Danvy
Jean-Michel Elsen
Anne Ricard

Ecole doctorale :

Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)

Unité de recherche :

Station d'Amélioration Génétique des Animaux (INRA - SAGA)

Directeur(s) de Thèse :

Jean-Michel Elsen
Anne Ricard

Rapporteurs :

Carole Charlier
Françoise Clerget-Darpoux

REMERCIEMENTS

Je remercie tout d'abord tous les financeurs et partenaires du projet GENEQUIN sans qui ma thèse n'aurait été possible : l'Agence Nationale de la Recherche (ANR), le Fonds Éperon, la région Basse-Normandie, l'Institut Français du Cheval et de l'Équitation (IFCE), l'Ecole Nationale Vétérinaire d'Alfort (ENVA), la faculté de médecine vétérinaire et le centre de Médecine Sportive (CEMESPO) de Liège et l'Institut National de la Recherche Agronomique (INRA).

Je remercie Christèle Robert-Granié, directrice de la Station d'Amélioration Génétique des Animaux de l'INRA d'Auzeville pour les moyens mis à ma disposition durant cette thèse.

Je tiens à remercier Carole Charlier et François Clerget-Darpoux pour m'avoir fait l'honneur d'être rapporteurs et Stéphane Robin et Sophie Danvy pour avoir accepté d'être examinateurs.

Je remercie tout particulièrement mes deux encadrants : Anne Ricard et Jean-Michel Elsen. Ils m'ont suivi avec patience tout au long de cette thèse et ont su m'aider dans les moments difficiles. Ils m'ont beaucoup apporté et c'est avec beaucoup d'humilité que je parle d'eux. Leurs connaissances, leurs manières d'aborder les problèmes et leurs raisonnements n'ont cessé de m'éblouir chaque jour et vont beaucoup m'apporter dans ma vie professionnelle. Merci Anne et Jean-Michel.

Et puis je tiens à remercier tout le petit monde de la SAGA... Andrés tout d'abord pour sa gentillesse et son nombre d'heures incalculables à déboguer mes programmes ; Hugues, Julien, Cécile, Guillaume, Anne pour les petits moments détentes ! Le groupe des informaticiens (François, Gilbert, Edmond, Dudu, Alain) pour les pauses cafés et les discussions sportives du quotidien. Cela va beaucoup me manquer. Line, Valérie, Dounia, Nancy et Carine pour le côté administratif de cette thèse. Un grand merci à vous et à tous ceux qui n'ont pas été cités mais qui ont contribué au bon déroulement de ce travail.

Je tiens à dédier cette thèse à ma famille. A mes soeurs, Charlotte et Pauline, et surtout à mes parents, Claire et Christian, qui m'ont toujours soutenu, aidé, supporté et m'ont donné les moyens de réussir mes études. Je vous en serai à jamais reconnaissant.

RESUME

Diverses lésions ostéochondrales peuvent affecter les articulations des jeunes chevaux et réduire leurs futures performances en course. L'objectif de cette thèse est d'identifier les régions du génome, appelées locus à caractère quantitatif (QTL), associées avec des caractères mesurant l'ostéochondrose (OC) enregistrés dans le programme GENEQUIN sur une population de Trotteurs Français. Le génotypage a été réalisé à l'aide de la puce SNP Illumina BeadChip EquineSNP50, qui est dense et permet d'exploiter le déséquilibre de liaison par des analyses d'association. Ces analyses sont sujettes à certains problèmes en présence d'une structure familiale des données. Dans la première partie de la thèse, une comparaison de la puissance et de la robustesse d'un choix restreint de méthodes d'analyses est effectuée. L'originalité de ce travail réside dans la dérivation algébrique des moments des distributions des statistiques de test comparées, donnant ainsi plus de généralité à nos résultats et permettant une meilleure compréhension des différences. Les résultats peuvent notamment servir à l'optimisation du dispositif expérimental. La deuxième partie est consacrée à la cartographie des régions QTL des caractères mesurant l'OC en différents sites articulaires dans une population de 583 Trotteurs Français. Cette étude a permis de mettre en évidence plusieurs régions QTL d'effets moyens et faibles à un niveau significatif mais pas hautement significatif. Nous montrons que l'OC est un caractère polygénique et qu'aucun QTL, ayant un effet à la fois sur l'OC du jarret et l'OC du boulet, n'est détectable dans ce protocole QTL, ce qui infirme l'hypothèse simple d'une cause génétique commune de la sensibilité à cette maladie sur les différents sites anatomiques. Suite à ces travaux, l'identification des gènes candidats et des mutations causales devrait clarifier la physiopathologie moléculaire de l'OC et ainsi permettre de développer des stratégies efficaces pour l'évaluation des risques. Pendant ce temps, les marqueurs peuvent être utilisés dans un contexte de sélection assistée par marqueurs afin d'améliorer la santé et le bien-être du cheval.

Mots clés : analyse d'association, équine, locus à caractère quantitatif (QTL), ostéochondrose, Trotteurs Français, structure de population.

ABSTRACT

Osteochondral lesions are commonly observed in young horses and may be responsible for reduced performances in racing. The purpose of the PhD thesis was to identify genome regions, called quantitative trait loci (QTL), associated with various traits measuring osteochondrosis (OC) and recorded in the GENEQUIN program in a population of French Trotters horses. Genotyping was performed using the EquineSNP50 Illumina high density chip, which allows to exploit the linkage disequilibrium with genome-wide association studies. These analyses are subject to several problems in presence of family structure. We hence first proposed a comparison of power and robustness of a limited choice of models for this type of analysis. The originality of this work lies in the algebraic derivation of the distribution moments of the test statistics compared, making the outcome of this comparison more general and allowing a better understanding of differences. The results can be used to establish an experimental design. The second part was devoted to the QTL fine mapping of traits that measure OC in different joint sites. This study highlighted several significant QTL with low and medium effects but none of them were highly significant. We showed that OC is a polygenic trait and we were not able to identify QTL affecting both OC on the hock and the fetlock, rejecting the hypothesis of a single genetic determinism of susceptibility to this disease accross anatomical sites. Further studies will now focus on the identification of candidate genes and screening for mutation in an attempt to clarify the molecular physiopathology of OC and develop efficient strategies for risk assessment. Meanwhile, markers could be used in a marker-assisted selection context to improve horse health and welfare.

Keywords : Equine, French Trotters horses, genome-wide association, population structure, osteochondrosis, quantitative trait loci (QTL).

LISTE DES ABREVIATIONS

ACP : Analyse en composante principale
AOAJ : Affections Ostéo-Articulaires Juvéniles
cM : centi-Morgan
CSH : *Chromosome Segment Homozygosity*
df : degré de liberté
ECA : *Equus Callabus chromosome*
EM : *Expectation Maximization*
FBAT : *Family based association test*
FDR : Erreur de type 1
HAN : cheval hanovrien
HaploIBD : Modèle aléatoire incluant un effet haplotypique et un effet polygénique
HMM : Modèle de markov caché
HWE : Equilibre de Hardy-Weinberg
IBD : Identité par descendance
IBS : Identité par état
KOSC : Kystes osseux sous-chondraux
LA : Analyse de liaison
LD : Déséquilibre de liaison
LDA : Analyse d'association
LDLA : Analyse combinant liaison et association
LSG : Logarithme du Score Global
MAF : Fréquence de l'allèle mineur
Mb : Megabase
MCMC : Markov Chain Monte Carlo
MS : Microsatellites
NRL : Neuropathie laryngée récurrente
OC : Ostéochondrose
OCD : Osteochondrites disséquantes
OC RICT : Ostéochondrose du relief intermédiaire de la cochlée tibiale
QQ-Plot : *Quantile-Quantile Plot*
QTDT : *Quantitative Transmission disequilibrium test*

QTL : *Quantitative Trait Loci*

REML : Estimation du maximum de vraisemblance restreint

SG : Score Global de l'ostéochondrose

SNP : *Single Nucleotide Polymorphism*

SNPMixed : Modèle mixte incluant un effet du marqueur et un effet polygénique

TDT : *Transmission disequilibrium test*

TF : Trotteur Français

TN : Trotteur Norvégien



Table des matières

Introduction générale	13
Partie I La détection de QTL	15
1 Contexte bibliographique	17
1.1 Introduction	17
1.2 Analyses Préliminaires	18
1.2.1 L'équilibre d'Hardy-Weinberg (HWE)	19
1.2.2 Le déséquilibre de liaison	20
1.2.3 Données génotypiques manquantes	25
1.2.4 La reconstruction des phases	26
1.2.5 Informations sur les données	28
1.2.6 Conclusion : analyses préliminaires	29
1.3 Analyses d'association (LDA)	29
1.3.1 Analyse uni-QTL SNP par SNP	29
1.3.2 Lutter contre les effets d'une éventuelle stratification de la population	37
1.3.3 Analyses uni-QTL par haplotype	45
1.3.4 Analyses multi-QTL	49
1.3.5 Conclusion : méthodes LDA	50
1.4 Analyses combinant association et liaison (LDLA)	51
1.4.1 Tests d'association dans des familles nucléaires	51
1.4.2 La théorie de Meuwissen et Goddard	55
1.4.3 Conclusion : méthodes LDLA	60
1.5 Tests multiples	61
1.5.1 Correction de Bonferroni	62
1.5.2 Seuils par permutations	62
1.5.3 False Discovery Rate (FDR)	62
2 Comparaison algébrique de méthodes uni-QTL	65
2.1 Résumé de l'article	65
2.2 Article Méthodologique	68

2.3	Validation des résultats par simulations	124
2.3.1	Paramètres des simulations	124
2.3.2	Résultats	124
2.4	Bilan et perspectives	128
Partie II Analyse de données réelles		129
3	Contexte bibliographique	131
3.1	Introduction	131
3.2	L'Ostéochondrose	132
3.2.1	Définition	132
3.2.2	Pathologie	132
3.2.3	Clinique et diagnostic des AOAJ	133
3.2.4	Étiologie	134
3.3	Les études de QTL de l'ostéochondrose	138
3.3.1	Chez les chevaux	138
3.3.2	Dans d'autres espèces	140
3.4	Problèmes existants et enjeux de la thèse	140
4	Analyse GENEQUIN	145
4.1	Résumé de l'article	145
4.2	Article appliqué	147
4.3	Compléments sur les phénotypes	177
4.3.1	Description des données de base	177
4.3.2	Effets environnementaux	183
4.3.3	Description de la structure de parenté	185
4.4	Compléments sur les marqueurs et les génotypes utilisés	186
4.4.1	Description de la puce Illumina BeadChip EquineSNP50	186
4.4.2	Qualité des typages	186
4.5	Compléments sur les QTL détectés	190
4.5.1	Nouveaux caractères analysés	190
4.5.2	Compléments sur les méthodes d'analyses	190
4.5.3	Bilan des QTL détectés	192
4.5.4	Discussion	194
4.5.5	QTL de l'OC du jarret sur ECA 3 : description	196
4.5.6	QTL de l'OC du jarret sur ECA 3 : test de validation	198
4.6	Bilan et perspectives	200
Conclusion générale		205
Annexes		207
Annexes A : Complément de l'article méthodologique 2.2.2		207
Annexes B : Détails des lésions d'ostéochondrose		233

Annexes C : Article LDSO 237

Bibliographie **249**

Introduction générale

Depuis des années, l'activité économique du milieu équin est en constante augmentation et les bonnes performances sportives des chevaux de courses sont au centre de cet essor. Malheureusement, ces performances sont parfois diminuées par des affections. Parmi celles-ci, l'ostéochondrose (OC) et la neuropathie laryngée récurrente (NRL) sont les plus répandues et peuvent être responsables de problèmes de boiteries ou de troubles respiratoires. Diminuer l'incidence de ces affections constitue alors un double enjeu : améliorer le bien-être du cheval et réduire le fort impact économique qu'elles produisent sur les différents aspects de la valorisation du cheval.

Ces affections sont multifactorielles et ont toutes deux une composante héréditaire dont le déterminisme génétique demande à être précisé. Étudier ce déterminisme, soupçonné d'être polygénique pour l'OC et monogénique pour la NRL, en recherchant le ou les gènes causaux, permettrait de mettre en place des dispositifs visant à réduire l'incidence de ces affections au sein d'une population (sélection assistée par marqueurs, sélection génomique, ou dépistages individuels). Financé par l'Agence Nationale de la Recherche (ANR), le Fonds Éperon, la région Basse-Normandie et l'Institut Français du Cheval et de l'Équitation (IFCE), et mêlant un partenariat entre l'Institut National de la Recherche Agronomique (INRA), l'École Nationale Vétérinaire d'Alfort (ENVA), la faculté de médecine vétérinaire et le centre de Médecine Sportive (CEMESPO) de Liège, le projet GENEQUIN a débuté dans ce contexte en 2008 avec pour but de répondre à plusieurs objectifs :

1. Faire une étude épidémiologique afin de mieux comprendre la composante génétique de l'ostéochondrose et de la neuropathie laryngée récurrente.
2. Construire une cohorte appropriée (environ 600 de chevaux sains et atteints) pour chacune des maladies en vue de l'étape 3. Décrire précisément les phénotypes. Prélever les échantillons ADN des chevaux phénotypés et de certains de leurs parents pour les génotyper à l'aide de la puce Illumina BeadChip EquineSNP50.
3. Avec les outils statistiques appropriés, utiliser les informations génotypiques pour trouver des marqueurs moléculaires associés à la maladie et détecter les porteurs.
4. Aller des marqueurs aux gènes impliqués dans le développement de la maladie, dont l'objectif final serait de dénouer les mécanismes moléculaires sous-jacents.

Cette thèse s'inscrit dans ce projet et a pour objectif de répondre au point 3, uniquement pour la partie sur l'ostéochondrose. Néanmoins, des interactions régulières existent entre les différents points et entre les deux affections. Cette thèse est composée de deux grandes parties.

La première partie est consacrée aux outils statistiques qui permettront de réaliser l'analyse des données collectées. Ainsi, un premier chapitre introduit l'état de l'art des outils existants en essayant de développer le plus possible les qualités et les défauts de chacun d'eux. Le principal problème qui ressort est la robustesse des modèles en présence de structures familiales. Avec l'objectif d'une meilleure compréhension et par soucis d'exhaustivité, la robustesse et la puissance d'un choix restreint de ces modèles ont été comparées algébriquement dans un deuxième chapitre et font l'objet d'un article en préparation.

La deuxième partie est consacrée à l'étude de l'OC et à l'analyse des données collectées dans le cadre du projet. Un premier chapitre pose le contexte de l'étude en présentant l'OC et différentes études autour de cette affection. Le deuxième chapitre est consacré à l'analyse des données collectées sur la population de Trotteurs Français. Cette étude, qui a abouti à un article accepté dans *Journal of Animal Science*, est l'objet principal de cette thèse.

Première partie

La détection de QTL

Contexte bibliographique

1.1 Introduction

Depuis l'arrivée des séquences des génomes complets chez les animaux domestiques à partir de 2007 – 2008, de nouvelles perspectives, beaucoup plus favorables, se sont ouvertes pour identifier les gènes impliqués dans la variabilité génétique des caractères phénotypiques. Ces séquences ont permis l'identification de nombreux **marqueurs génétiques** et les avancées technologiques récentes nous offrent maintenant la possibilité d'utiliser des outils de génotypage à haut débit pour un coût relativement faible. Un marqueur génétique peut être défini comme un endroit du génome (locus) se présentant sous différentes formes dans une population (allèles). Les marqueurs **SNP** (Single Nucleotide Polymorphisms), qui sont des polymorphismes dus à la substitution, à l'insertion ou à la délétion d'un nucléotide, ont remplacé les marqueurs microsatellites, qui sont des répétitions de séquences très courtes. Les marqueurs SNP sont bi-alléliques et présentent les avantages d'être très nombreux sur le génome et facilement identifiables par les outils technologiques actuels à un coût raisonnable. A contrario, les marqueurs microsatellites sont beaucoup moins nombreux, ont un coût plus élevé, mais présentent l'avantage d'être très polymorphes.

L'utilisation des marqueurs de type SNP, présents en grand nombre, permet d'entrevoir une localisation plus précise du (ou des) gène(s) responsable(s) d'un phénotype. Deux principales approches sont envisageables dans la recherche des gènes causaux, l'approche gène candidat et la **cartographie des QTL**. L'approche gène candidat, qui nécessite des connaissances a priori sur l'implication de certains gènes sur le caractère étudié, consiste à étudier l'association entre un phénotype et des allèles de ces gènes candidats. La cartographie des QTL, beaucoup plus utilisée, explore, sans a priori, l'ensemble ou une partie du génome en testant de possibles associations entre le polymorphisme de marqueurs, tels que les SNP ou microsatellites, et la variabilité du phénotype. Cette approche permet de localiser des régions chromosomiques (**Quantitative Trait Loci ou QTL**) portant le (ou les) gène(s) dont le polymorphisme est impliqué dans la variabilité du caractère étudié.

En génétique animale, jusqu'en 2008, les QTL étaient détectés et localisés à l'aide de dispositifs expérimentaux familiaux comprenant au moins deux générations et de marqueurs microsatellites répartis sur tout le génome avec des distances entre marqueurs importantes > 20 Megabase (Mb). Pour ce type de données, la méthode de choix est **l'analyse de liaison** qui consiste à chercher une possible association intra-famille entre l'origine grand parentale des régions chromosomiques

transmises par les parents et les phénotypes des descendants. L'origine grand parentale est tracée par des marqueurs dont l'effet apparent peut donc varier d'une famille à l'autre. Plus la différence entre les moyennes des phénotypes des descendants ayant reçu l'haplotype paternel ou maternel du père est grande, plus le QTL est important. Néanmoins, l'origine grand parentale des segments chromosomiques proches reçus par un descendant étant le plus souvent identiques, la localisation du QTL est peu précise. Pour obtenir une localisation plus précise, il faut que des **recombinaisons** (échange de portions de chromosomes homologues durant la méiose) interviennent durant la constitution des gamètes. Plus il y a de recombinaisons dans une région de taille donnée, plus la localisation est précise mais plus les familles doivent être de grande taille. En pratique, les intervalles de confiance obtenus dans ce genre d'analyses sont très élevés (des dizaines de Mb qui peuvent contenir des centaines de gènes). On parle souvent de ces analyses comme des primo-localisations des QTL.

Les marqueurs SNP, qui se répartissent sur l'ensemble du génome à intervalle moyen de l'ordre de 40 – 50 kb (pour des puces d'environ 50000 marqueurs), permettent de localiser les QTL plus précisément en utilisant les **analyses d'association**. Ces analyses exploitent le phénomène de **déséquilibre de liaison**, association non aléatoire d'allèles en plusieurs locus dans la population des chromosomes (ou gamètes) présents dans la population. Plus deux locus sont proches, plus cette association est élevée. A l'inverse, 2 locus éloignés (notamment sur 2 chromosomes distincts) tendent à être en équilibre de liaison. Les études d'association entre un marqueur (ou un ensemble de marqueur) et un phénotype supposent donc que ce marqueur (ou cet ensemble) est en déséquilibre de liaison avec un QTL dont le polymorphisme cause une partie de la variabilité du caractère. Ces études se sont en premier développées en génétique humaine (les marqueurs SNP sont disponibles depuis une dizaine d'années) et sont devenues très courantes en génétique animale et végétale depuis 2008.

Ce chapitre a pour but de présenter l'état de l'art des étapes possibles et nécessaires à une bonne recherche de QTL. Il s'inspire en grande partie du plan et des travaux de Balding (2006), Laird and Lange (2006) et des cours que j'ai pu suivre du Pr. Ben Hayes. Leurs travaux m'ont permis de réaliser rapidement cet état de l'art des méthodes provenant de la génétique humaine et de la génétique animale. Cet éventail de méthode était particulièrement utile pour l'analyse d'une population dont les structures familiales sont intermédiaires entre celles des populations humaines et celles des animaux domestiques.

La première partie vise à expliquer les étapes préliminaires à la cartographie fine des QTL. Les autres parties sont consacrées à la cartographie fine des QTL à l'aide de méthodes de type analyse d'association ou de méthode combinant association et liaison. Leurs limites sont décrites.

1.2 Analyses Préliminaires

Avant toute recherche de QTL affectant un caractère, il est nécessaire de regarder en détail les données pour vérifier leur qualité, tirer d'elles de l'information et sélectionner la méthode la plus adaptée à l'analyse génétique. Cette section a pour but de présenter les analyses préliminaires à toute étude, notamment vérifier l'équilibre d'Hardy-Weinberg, savoir intégrer des données manquantes, reconstruire les phases d'haplotypes ou savoir calculer et tirer l'information qu'apporte le déséquilibre de liaison entre les marqueurs.

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

Définition et hypothèses

La théorie de l'équilibre d'Hardy-Weinberg fût proposée indépendamment par Hardy (1908) et Weinberg (1908). Elle stipule que les fréquences des allèles et du génotype d'un locus restent constantes de générations en générations (d'où la notion d'équilibre) si les hypothèses suivantes sont respectées :

- La population est de taille infinie (\sim population de grande taille, loi des grands nombres).
- Les espèces étudiées sont diploïdes et à reproduction sexuée.
- Il n'y a pas de migration.
- Il n'y a pas de sélection.
- Il n'y a pas de mutation.
- Le régime de reproduction est panmictique (les gamètes s'associent au hasard, ou les couples se forment aléatoirement)
- Les fréquences alléliques des mâles et des femelles sont identiques.

Soit le locus A possédant 2 allèles (type SNP), notés A_1 et A_2 , de fréquences respectives p et q ($p + q = 1$). Si on se place dans le cadre d'un équilibre d'Hardy-Weinberg, les fréquences génotypiques seront :

$$f_{A_1A_1} = p^2, \quad f_{A_1A_2} = 2pq, \quad f_{A_2A_2} = q^2.$$

Différents tests possibles

Il existe plusieurs tests permettant de voir si on dévie de l'équilibre d'Hardy-Weinberg. Le plus simple est le test de Pearson (plus connu sous le nom de "test du χ^2 ") dont la distribution sous l'hypothèse nulle suit asymptotiquement une loi du χ^2 . Le test de Pearson n'est pas optimal lorsque la fréquence d'un des génotypes présents est faible. Dans ces conditions, il est préférable d'utiliser un test exact de Fisher. On trouve facilement, dans la littérature, d'autres tests exacts comme ceux de Wigginton et al. (2005) ou Guo and Thompson (1992).

Exemple : Le test de Pearson

Soit un échantillon de n individus génotypés au locus A bi-allélique. Notons l le nombre d'allèles (ici 2) et m le nombre de génotypes possibles (ici 3). Notons également $n_{A_1A_1}$ (resp. $n_{A_1A_2}$ et $n_{A_2A_2}$) le nombre d'individus qui possèdent le génotype A_1A_1 (resp. A_1A_2 et A_2A_2). La somme des trois nombres faisant n . Les fréquences alléliques s'obtiennent de la façon suivante :

$$n_{A_1} = 2n_{A_1A_1} + n_{A_1A_2}, \quad n_{A_2} = 2n_{A_2A_2} + n_{A_1A_2}$$

Pour tester l'hypothèse nulle H_0 : "Le locus est en équilibre de Hardy-Weinberg" contre H_1 : "Le locus n'est pas en équilibre de Hardy-Weinberg", on utilise le test du χ^2 suivant :

$$X^2 = \sum_{i=1}^m \frac{(O_i - E_i)^2}{E_i} \quad \text{avec} \quad X^2 \sim \chi_{m-1}^2 \quad (\text{asymptotiquement})$$

Où O_i et E_i représentent respectivement les fréquences absolues observées et espérées de chaque génotype (Table 1.1).

TABLE 1.1 – Fréquences observées et espérées en fonction des génotypes (p et q sont les fréquences des allèles A_1 et A_2 dans la population)

	génotypes		
	A_1A_1	A_1A_2	A_2A_2
O_i	$n_{A_1A_1}$	$n_{A_1A_2}$	$n_{A_2A_2}$
E_i	np^2	$2npq$	nq^2

Visualisation et implication d'une déviation à HWE

Après avoir réalisé un test HWE sur l'ensemble des marqueurs disponibles sur le génome, il est important de bien interpréter les résultats. On utilise généralement un *quantile-quantile (QQ) P-value plot* ou le *log QQ P-value plot*. Le *QQ Plot* est un moyen graphique permettant de visualiser les différences entre deux distributions de probabilités. Les distributions sont d'autant plus similaires que les points sont alignés. L'utilisation du log permet de mettre l'accent sur les plus petites *P values*. On trace le logarithme négatif de la $i^{\text{ème}}$ plus petite *P-value* (P-valeur observée) contre le logarithme négatif de i/L (P-valeur attendue), avec L le nombre de SNP testés. La Figure 1.1 est un exemple de *QQ Plot*. Une déviation à la première bissectrice (la droite $y = x$) correspond à un locus qui dévie de l'hypothèse H_0 d'équilibre d'Hardy-Weinberg. Ces déviations peuvent être causées par un écart aux hypothèses sur la structure de la population ou par des erreurs de génotypages (tendance à classer des hétérozygotes comme homozygotes). Avec l'objectif de supprimer les marqueurs qui ont des erreurs de génotypages, on écarte les locus qui dépassent un certain seuil de significativité (généralement pris entre 10^{-4} et 10^{-8}). Cependant, dans des études cas-témoins, des auteurs comme Wittke-Thompson et al. (2005) ont montré qu'une déviation d'HWE pouvait aussi être le signe d'une association avec la maladie.

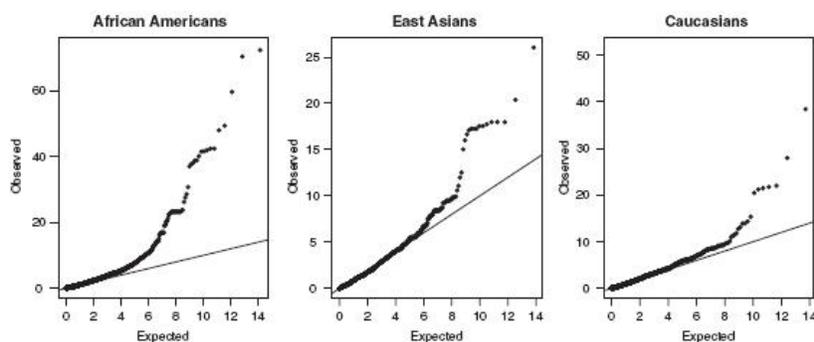


FIGURE 1.1 – *QQ Plots* d'un test statistique d'Hardy-Weinberg (Weir et al., 2004)

1.2.2 Le déséquilibre de liaison

Définition et Mesures

Le déséquilibre de liaison (LD pour *linkage disequilibrium*) est une notion très importante en génétique quantitative et est à la base des études d'association. On peut illustrer le LD en

considérant 2 marqueurs SNP A et B sur un même chromosome, avec A possédant les allèles (A_1, A_2) et B les allèles (B_1, B_2) . Quatre haplotypes sont possibles : A_1B_1 , A_1B_2 , A_2B_1 et A_2B_2 . Si chaque allèle a une fréquence de 0.5 et que les hypothèses HWE sont vérifiées, alors chaque haplotype devrait avoir une fréquence de 0.25. Ainsi, toute déviation à 0.25 de cette fréquence haplotypique signe un déséquilibre de liaison. Il existe de nombreuses mesures du LD entre les allèles à 2 locus qui, toutes, dépendent de la mesure suivante :

$$D_{A_1B_1} = f_{A_1B_1} - f_{A_1}f_{B_1}$$

c'est à dire de la différence entre la fréquence de l'haplotype A_1B_1 et le produit des fréquences de chacun des allèles A_1 et B_1 . Cette mesure porte sur les haplotypes et donc nécessite la connaissance des phases. Notons que le LD entre A_2 et B_2 , A_1 et B_2 ou A_2 et B_1 pourrait aussi être calculé. Lorsqu'on s'intéresse à des marqueurs bi-alléliques (type SNP), les résultats de ces différents calculs sont échangeables ($D_{A_1B_1} = -D_{A_1B_2} = -D_{A_2B_1} = D_{A_2B_2}$). Cependant, lorsque les marqueurs sont multi-alléliques (type Microsatellite), ce n'est plus le cas.

La mesure D définie précédemment n'est pas idéale car très dépendante des fréquences alléliques et non normalisée. D'autres mesures ont été dérivées et sont plus couramment utilisées :

- Le r^2 de Hill and Robertson (1968) :

$$r^2 = \frac{D^2}{f_{A_1}f_{A_2}f_{B_1}f_{B_2}}, \quad 0 \leq r^2 \leq 1$$

- Le D' de Lewontin (1964) :

$$D' = \frac{|D|}{D_{max}}, \quad D_{max} = \begin{cases} \min [f_{A_1}f_{B_2}, f_{A_2}f_{B_1}] & \text{si } D > 0 \\ \min [f_{A_1}f_{B_1}, f_{A_2}f_{B_2}] & \text{si } D < 0 \end{cases}$$

La mesure r^2 est préférée au D' pour deux raisons principales. Supposons qu'un locus soit un marqueur et l'autre un QTL (non génotypé) :

1. Le r^2 représente la proportion de la variance expliquée par le QTL observée au marqueur : i.e. si par exemple la taille des individus possède une variance due au QTL de 20cm et si le r^2 entre le marqueur et le QTL vaut 0.2, alors la variation observée au marqueur sera de 4cm.
2. Le r^2 est lié à la puissance de détection d'une analyse. La taille de l'échantillon doit être augmentée de $1/r^2$ pour avoir la même puissance qu'une expérience dans laquelle le QTL est observé directement (Pritchard and Przeworski, 2001).

LD sur plus de 2 locus

Le D' et le r^2 sont toutes deux des mesures de LD sur 2 locus. Cependant, il est intéressant de regarder le LD sur une région chromosomique. Pour cela, une première approche est de calculer une moyenne locale de LD pour des paires de locus. Des logiciels comme *GOLD* (Abecasis and Cookson, 2000) ou *Haplowiew* (Barret et al., 2005) permettent de voir le niveau de LD entre régions sur des graphiques. Une seconde approche pour mesurer le LD sur une région chromosomique est le *Chromosome Segment Homozygosity* (CSH) décrit par Hayes et al. (2003). Considérons la transmission d'un des chromosomes d'un des ancêtres de la population "actuelle". Au fil des générations, avec les recombinaisons, ce chromosome s'est cassé. Les descendants actuels ayant hérité d'un même élément

ponctuel du chromosome ancestral ne partagent qu'une petite région chromosomique autour de cet élément. Cette région est dite *Identique par descendance* (IBD pour *Identical By Descent*). Le CSH est la probabilité que deux segments chromosomiques de la même taille et du même endroit pris aléatoirement dans une population proviennent d'un ancêtre commun. Le Figure 1.2 propose une illustration de ce concept.

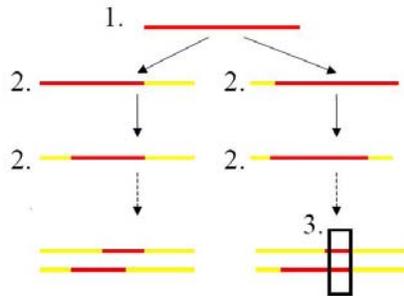


FIGURE 1.2 – Un ancêtre commun plusieurs générations auparavant (1) a des descendants (2). A chaque génération, le chromosome ancestral est cassé par des recombinaisons, jusqu'à la génération courante pour laquelle ne sont conservés que de très petits segments du chromosome ancestral (3) (d'après les cours du Pr. Ben Hayes).

Facteurs influençant le LD

De nombreux facteurs peuvent créer le LD et influencer son ampleur :

- **La dérive génétique** : ce phénomène décrit le changement de fréquences alléliques et haplotypiques dans une population entre générations, en raison de l'échantillonnage aléatoire qui se produit lors de la production d'un nombre fini de gamètes. Ces changements de fréquences sont fonction de la taille efficace de la population (N_e) : si N_e est petit, seul un petit nombre d'individus contribuent génétiquement à l'évolution de la population, et la dérive génétique est élevée ce qui induit une perte importante de variabilité génétique. Les haplotypes étant moins nombreux, le LD diminue peu au cours des générations. Un tel phénomène est bénéfique aux études d'association à une maladie car la dérive augmente la différence des fréquences alléliques entre les gènes de sensibilité à la maladie et le marqueur entre les cas et les témoins (Ardlie et al., 2002).
- **La croissance démographique** : une croissance rapide de la population diminue le LD en réduisant la dérive génétique.
- **Les mutations** génèrent du déséquilibre de liaison : Si une mutation apparaît dans un haplotype, elle est associée à celui-ci et aux allèles qu'il contient. A cet instant, le déséquilibre de liaison entre cette mutation et l'haplotype est total. Au cours du temps, cet haplotype pourra se transmettre aux descendants et donc gagner en fréquence.
- **La migration** : le LD peut être également créé par mélange de populations ou par migration. Dans un premier temps, le LD est proportionnel aux différences de fréquences alléliques entre les populations et il est indépendant de la distance entre marqueurs. Lors des générations suivantes, le déséquilibre de liaison décroît mais la décroissance diffère se-

lon que les locus sont liés ou non. Pour des locus indépendants, "l'étrange" LD possible entre ces marqueurs tend à disparaître rapidement alors que pour des locus liés il persiste plus longtemps pour finalement être cassé par les recombinaisons. Lorsqu'on fait une étude d'association à partir d'un échantillon d'individus qui dérive d'un mélange de populations, il faut faire très attention à l'existence de faux positifs dus à ce phénomène (voir chapitre suivant).

- **La sélection** peut induire une dérive génétique et une diminution du nombre de reproducteurs ce qui mène à une réduction du nombre d'haplotypes présents dans la population (donc une augmentation des fréquences haplotypiques).
- **La variabilité du taux de recombinaison** : le taux de recombinaison n'est pas constant au fil du génome, la plupart des recombinaisons se faisant dans les mêmes régions appelées hot spots. Donc le LD va être fort dans les régions non recombinantes et va se casser dans les hot spots.

Évolution, étendu et possibles interprétations du LD dans les populations

L'évolution du déséquilibre de liaison dans une population est fonction du taux de recombinaison entre les locus. Si on note c ce taux et D_0 le déséquilibre initial entre les deux locus dans une population de grande taille, et en supposant qu'il n'y ait pas de variation des fréquences alléliques entre deux générations, alors le LD entre les locus diminue d'un facteur $(1 - c)$ par génération. Donc on obtient :

$$D_t = (1 - c)^t D_0$$

Avec D_t le LD à la génération t . On voit bien que si c est faible, ce qui correspond à des locus proches, alors le déséquilibre de liaison ne diminue que très peu au cours des générations.

A l'équilibre (au sens de HWE), certains auteurs ont dérivé cette formule afin de relier l'espérance du LD à la taille effective de la population (N_e) et obtiennent ainsi l'approximation suivante (Sved, 1971; Hill, 1975; Tenesa et al., 2007) :

$$E(r^2) = \frac{1}{kN_e c + \alpha}$$

Avec $\alpha = 1$ en l'absence de mutation, $\alpha = 2$ pour prendre en compte les mutations, $k = 2$ pour le chromosome X et $k = 4$ pour les autosomes.

Reich et al. (2001) montrent que le LD sur de courtes distances reflète l'histoire ancienne de la taille effective de la population et le LD sur de longues distances reflète son histoire récente. Comme nous avons pu le voir, le LD est fonction de la taille effective de la population et du taux de recombinaison (approximé par la taille d'un segment chromosomique, en espérance 1 recombinaison par Morgan). Alors que la taille effective de la population humaine est de l'ordre de 10000 (Kruglyak, 1999), elle est beaucoup plus faible (de l'ordre de 50 à 500) chez les animaux de ferme du fait de la sélection et du fort apparentement entre individus. On s'attend donc à ce que le LD s'étende sur de plus longues distances : on y trouve en effet un LD modéré ($r^2 \geq 0.2$) sur des distances de l'ordre de $100kb$ (pour $N_e = 100$) contrairement aux humains où ce niveau de LD s'étend sur des régions d'environ $5kb$ (Tenesa et al., 2007).

La figure 1.3 montre, en prenant l'exemple des "Within modern equine breed", qu'il faut un marqueur environ tous les $0.2Mb = 0.2cM$ pour obtenir un r^2 moyen de 0.2. Comme le génome du cheval est d'environ $2,46Gb$, il faut de l'ordre de 12000 SNP bien repartis pour obtenir un

marqueur tous les 0.2Mb. En pratique, comme une répartition équitable n'est pas possible, il faut doubler le nombre de SNP pour obtenir le r^2 visé de 0.2. La puissance de détection d'un QTL en analyse d'association dépendant du LD et de la taille de la population et ce niveau de LD permet de réaliser une analyse d'association sur tout le génome avec une taille de population raisonnable (Luo, 1998; Ball, 2005) : une puissance de détection de 0.8 avec un r^2 moyen de 0.2 sera atteinte avec 1000 individus (resp. 2000) pour détecter un QTL qui explique 5% de variance phénotypique (resp. 2.5%). Pour plus de détails, le lecteur est amené à lire la section sur les études d'association.

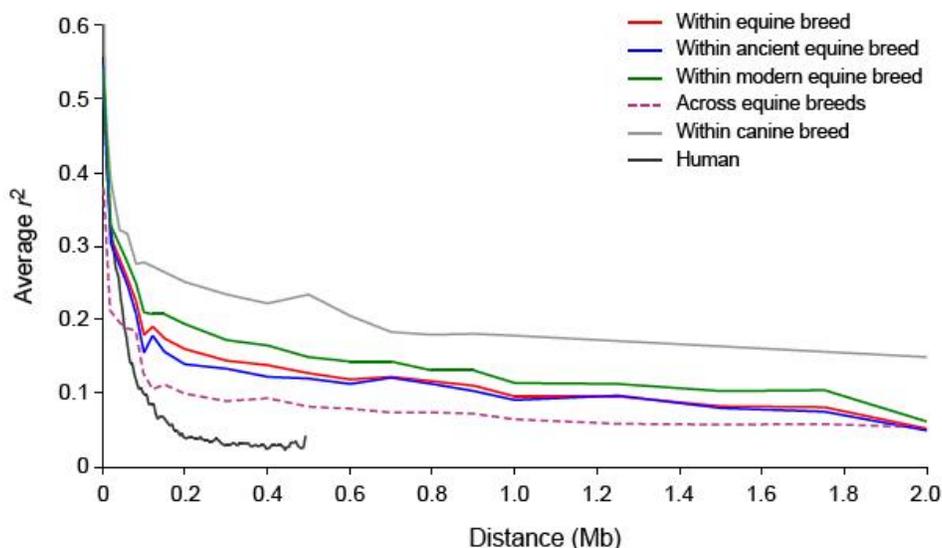


FIGURE 1.3 – Exemple d'une étude de l'étendu du LD en populations équines (Wade et al., 2009)

Haplotype blocks, recombination hotspots et SNP tagging

Avec l'utilisation de nombreux marqueurs (de l'ordre du million), de récentes études en génétique humaine ont montré l'existence de régions du génome avec de fort LD entre marqueurs (*haplotype block*) encadrées par des régions de faible LD voir sans LD (*recombination hot spots*). Une revue de ces études a été réalisée par Wall and Pritchard (2003). Il est nécessaire de prendre en compte cette information d'autant plus que l'on sait que 80% des recombinaisons ont lieu sur seulement 5 à 10% du génome (Kauppi et al., 2007). Il existe des logiciels qui permettent de trouver ce type de régions et l'un des plus connus est certainement *HOTSPOTTER* (Li and Stephens, 2003).

La découverte de blocs d'haplotypes a un intérêt majeur car elle permet de réduire le nombre de SNP à prendre en compte dans une analyse. En effet, les marqueurs étant en fort LD au sein d'un même bloc, la seule utilisation de quelques-uns de ces SNP permet d'expliquer au mieux la variabilité génétique de tout le bloc. La figure 1.4 montre une illustration de cette stratégie connue sous le nom de *SNP tagging* (Chapman et al., 2003).

Néanmoins, ces méthodes de blocs haplotypiques et de SNP tagging souffrent de quelques problèmes : il n'existe pas de méthodes optimales créant des blocs ; les blocs sont difficilement comparables entre différentes populations et ces méthodes nécessitent un grand nombre de marqueurs

(> 300000 SNP). Le dernier problème montre pourquoi il est encore difficile de réaliser ce genre d'études dans les populations animales. Cependant, du fait que ces stratégies ont fait leurs preuves chez les humains et que la technologie évolue chaque jour, les applications à l'animal ne sauraient tarder.

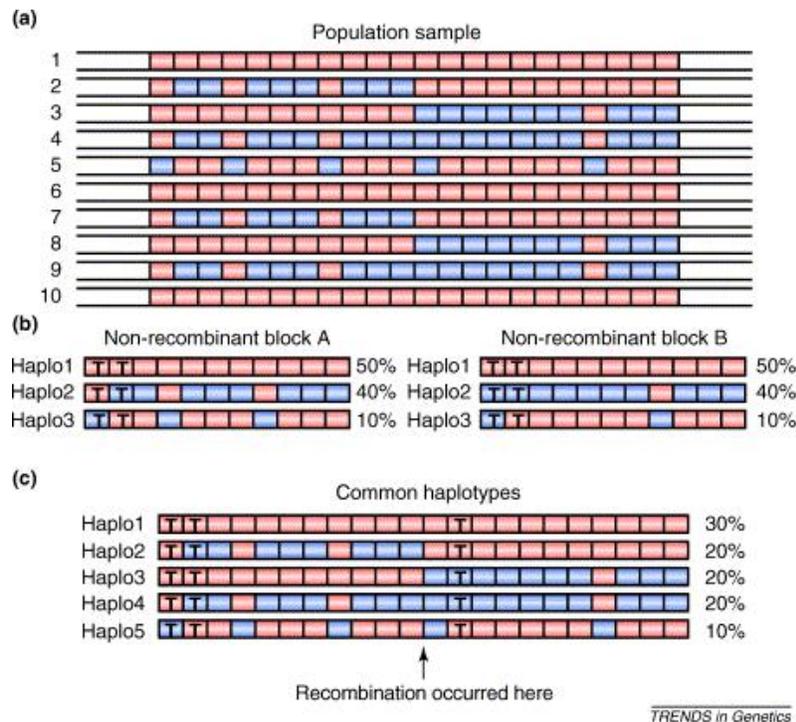


FIGURE 1.4 – Exemple de *SNP tagging* (Cardon and Abecasis, 2003)

1.2.3 Données génotypiques manquantes

Le génotypage des individus avec des SNP requiert des outils technologiques importants et certaines déconvenues se présentent, comme des erreurs de génotypages ou des génotypes manquants. On vient de voir que le test HWE permet de détecter certaines erreurs de génotypages, mais il ne permet pas de les corriger ou d'imputer les génotypes manquants. L'apparition de génotypes manquants est régulièrement due à une mauvaise séparation entre les "cluster" prédéfinis de la puce, qui permettent d'affilier chaque individu à un génotype. En général, un premier filtrage consiste à supprimer les SNP dont le pourcentage d'individus génotypés avec succès, appelé *call freq*, est inférieur à un seuil généralement fixé autour de 80%. De même, un individu dont le pourcentage de génotypes manquants sur l'ensemble des marqueurs est trop important (en général supérieur à 2%), appelé *call rate*, est supprimé des futures analyses. Néanmoins, pour tous les autres individus et marqueurs non supprimés, certains génotypes seront manquants. Quelle influence cela aura-t'il sur la détection des QTLs ? Des faux-positifs peuvent-ils apparaître ? De quels moyens disposons-nous pour résoudre ces problèmes ?

Dans une étude en cas-témoins, l'apparition de faux-positifs dus aux génotypes manquants est rare car les cas comme les témoins sont affectés de manière équitable. Cependant, lorsque les cas

et les témoins sont génotypés différemment (par exemple, les génotypages ont été réalisés dans des laboratoires différents ou plusieurs techniques ont été utilisées au coeur d'un même laboratoire), des faux-positifs peuvent apparaître (Clayton et al., 2005). Un moyen simple de repérer ce problème est de donner le code 1 aux génotypes observés, 0 aux génotypes manquants et de faire un test d'association sur cette variable entre cas et témoins (Balding, 2006). D'autres auteurs ont tenté de montrer l'impact des données manquantes sur une étude d'association. Hirschhorn and Daly (2005) ont montré, sur une étude de type *transmission disequilibrium test (TDT)*, que le nombre de faux-positif augmentait en fonction du nombre de génotypes hétérozygotes manquants (surtout pour des allèles rares).

Mineure dans des analyses uni SNP, la question des données manquantes est importante quand plusieurs SNPs sont analysés simultanément. Une analyse restreinte aux seules données complètes peut ne porter que sur une faible fraction de la population de départ. Une solution basique serait de remplacer les génotypes manquants par la moyenne des génotypes observés ou par le génotype le plus probable. Cependant, ces solutions modifieraient le déséquilibre de liaison avec les marqueurs proches et ainsi, des biais et une perte de puissance pourraient apparaître. L'idée est plutôt de remplacer les génotypes manquants par une valeur prédite basée sur les génotypes observés des SNP voisins. En général, les méthodes existantes, par maximum de vraisemblance ou bayésiennes, permettent d'affecter une valeur aux génotypes manquants et de reconstruire les phases simultanément (voir section suivante). Les méthodes les plus populaires sont intégrées dans les logiciels *PHASE* (Stephens et al., 2001), *fastPHASE* (Scheet and Stephens, 2006) et *IMPUTE* (Marchini et al., 2007). D'autres stratégies existent comme celles basées sur des méthodes de classification où le génotype manquant est copié des autres individus qui ont les mêmes génotypes aux marqueurs voisins, ou des méthodes de régression comme celle de Souverein et al. (2006) qui modélisent les génotypes manquants comme une fonction de génotypes d'autres marqueurs et de phénotypes dans une régression logistique polytomique.

1.2.4 La reconstruction des phases

L'utilisation d'haplotypes à la place de génotypes est généralement considérée comme un progrès majeur car elle permet une meilleure interprétation dans les analyses d'association. En effet, si on s'intéresse à des locus suffisamment proches pour qu'ils aient un taux de recombinaison très faible, les haplotypes seront transmis des parents aux descendants quasiment à l'identique. En conséquence, pour la cartographie de QTL comme pour les études d'évolution, où l'on suppose qu'un QTL est compris dans un haplotype, les haplotypes sont beaucoup plus informatifs que les simples marqueurs car ils permettent un meilleur suivi du QTL. Il existe différents moyens pour déterminer les haplotypes à partir des génotypes :

- Utiliser des méthodes biochimiques : il est possible, à l'aide de techniques de laboratoire, de reconstruire les haplotypes. Cependant, ces techniques sont lentes et souvent très coûteuses.
- Utiliser l'information familiale : si les génotypes des parents sont connus, les haplotypes des enfants peuvent être, mais pas toujours, déduits. Par exemple, supposons que le génotype d'un individu sur 3 marqueurs soit $A_2A_2 B_1B_2 C_1C_2$ et que ceux de ces parents soient $A_1A_2 B_1B_1 C_1C_2$ et $A_1A_2 B_1B_2 C_2C_2$, alors les haplotypes de cet individu sont $A_2B_1C_1/A_2B_2C_2$. Cependant, si les génotypes parentaux étaient $A_1A_2 B_1B_2 C_1C_2$ et $A_1A_2 B_1B_2 C_2C_2$, alors il serait impossible de conclure sur les phases haplotypiques de l'enfant. Pour un grand nombre de marqueurs ou lorsqu'on dispose d'un pedigree complexe avec des informations

manquantes, l'information familiale peut être insuffisante pour décrire la population des haplotypes.

- Utiliser des outils statistiques et des algorithmes de calcul sur les génotypes : l'idée est d'inférer les haplotypes présents dans une population à partir des génotypes d'un échantillon d'individus considérés comme non apparentés. On suppose que, même si un très grand nombre d'haplotypes étaient possibles dans cette population, dans la réalité, pour des locus étroitement liés (situation rencontrée quand on a un nombre important de SNP), seul un petit nombre de ces haplotypes vont être présents.

Différentes approches pour inférer les haplotypes à l'aide d'outils statistiques ou calculatoires ont été développées. Parmi les algorithmes de calculs, le plus connu est celui de Clark (1990) qui tente de réduire le nombre d'haplotypes dans la population dans une approche par parcimonie. Dans un premier temps, une liste tous les haplotypes connus est créée (de tels haplotypes de n marqueurs sont observables chez les individus qui sont homozygotes pour au moins $n - 1$ d'entre eux). Ensuite, les haplotypes de cette liste sont comparés avec les génotypes non phasés pour voir si ces génotypes ne peuvent pas être séparés en un haplotype présent dans la liste et un autre haplotype (nouveau ou présent également). Si ce dernier est nouveau, il est ajouté à la liste et le processus est réitéré. Cette approche souffre de deux problèmes : le premier, dont la probabilité augmente avec le nombre de SNPs, est la possibilité qu'aucun individu n'ait d'haplotypes connus. Dans cette situation, l'algorithme ne peut pas partir. Le second est que le résultat dépend de l'ordre de traitement des individus. Plusieurs travaux récents aboutissent à sélectionner la meilleure reconstruction possible (nombre d'haplotypes minimum quand la méthode est appliquée plusieurs fois sur le même jeu de données). Cela permet d'augmenter la puissance mais cette méthode reste cependant plus faible que celles décrites par la suite.

D'autres approches pour reconstruire les phases sont disponibles par maximum de vraisemblance. Utilisant l'algorithme *EM* (Expectation Maximization), les fréquences θ_i des haplotypes possibles dans la population sont initialisées de sorte que $\sum_i \theta_i = 1$. Ensuite, dans la phase *E* (*Expectation*), les probabilités des couples d'haplotypes possibles pour chaque individu (génotype) sont calculées :

$$\forall_{i,j} \forall_k, \quad P(h_i, h_j | g_k, \Theta^t), \quad 1 \leq i, j \leq H \text{ et } 1 \leq k \leq N$$

Où h_i, h_j sont les haplotypes possibles dans la population étudiée, g_k est le génotype de l'individu k observé et Θ^t représente l'ensemble des fréquences haplotypiques observées à l'itération t . L'étape suivante, *M* pour *Maximization*, consiste à calculer les nouvelles fréquences haplotypiques θ_i^{t+1} en prenant en compte les probabilités de l'étape *E* :

$$\theta_i^{t+1} = \frac{1}{2N} \left(\sum_{k=1}^N \left(\sum_{j=1, j \neq i}^H P(h_i, h_j | g_k, \Theta^t) \right) + 2P(h_i, h_i | g_k, \Theta^t) \right)$$

L'algorithme est alors réitéré T fois jusqu'à ce que les fréquences haplotypiques convergent.

Enfin, des approches bayésiennes ont été proposées pour reconstruire les haplotypes. Ces méthodes traitent les haplotypes inconnus comme des quantités aléatoires. Elles combinent la probabilité *a priori* ("croyance" sur la forme des haplotypes qu'on s'attend à voir dans la population) et la vraisemblance (l'information sur les données observées) dans le but de calculer la distribution *a posteriori* (la probabilité conditionnelle des haplotypes inconnus par rapport aux données génotypiques observées). Des algorithmes de type MCMC (Markov Chain Monte Carlo) permettent

d'utiliser ces approches. Elles sont souvent faciles à implémenter et précises mais exigent un temps de calcul assez conséquent.

De multiples méthodes et logiciels permettant d'inférer les phases haplotypiques se sont développés autour de ces 3 types d'approches. Le logiciel le plus connu, *PHASE* (Stephens et al., 2001), utilise une approche bayésienne où la probabilité *a priori* est calculée à partir de la théorie de la coalescence. Ce logiciel semble donner des estimations précises (Marchini et al., 2006) mais il est très coûteux en temps de calcul. Plus récemment, les logiciels *fastPHASE* (Scheet and Stephens, 2006) et *BEAGLE* (Browning and Browning, 2007) ont été créés à partir des chaînes de Markov cachées (HMM) et de l'algorithme *EM*. Ils donnent une précision aussi bonne que *PHASE* tout en étant plus rapides. Browning (2008) propose une très bonne revue des différentes méthodes existantes sur ce sujet.

Généralement, en population animale, les individus sont apparentés et les génotypes des pères et des descendants sont disponibles. Dans ce contexte, les méthodes de phasages citées précédemment ne sont pas optimales. Il existe des méthodes qui prennent en compte la partie transmission d'haplotypes entre parents et descendants et qui, de plus, utilisent le déséquilibre de liaison. On trouve parmi celles-ci *DualPHASE* et *DagPHASE* (Druet and Georges, 2010) qui sont basées sur *fastPHASE* et *BEAGLE*.

1.2.5 Informations sur les données

Données phénotypiques

Il convient avant de commencer la détection de QTL de bien définir le caractère que l'on étudie et de décrire le plus précisément possible le phénotype de chaque individu. En effet toute analyse faite à partir de phénotypes douteux mènerait à de faux résultats ou une perte importante de puissance. Certains caractères sont plus délicats à étudier que d'autres. Ainsi, les maladies peuvent être difficile à diagnostiquer alors que la taille d'un individu ne pose pas de problèmes. Trois grands types de caractères peuvent être rencontrés :

- **Les caractères binaires** : c'est le type de caractère le plus fréquemment étudié en génétique humaine. Les maladies sont souvent décrites par de tels caractères et les protocoles pour leur analyse sont généralement des études en cas-témoins (*case-control*). Dans l'idéal, ce type de protocole considère deux échantillons indépendants, un avec des individus atteints (cas, codés 1) et l'autre, supposé non relié aux individus atteints, avec des individus sains (témoins, codés 0). Ce type de protocole peut souffrir d'un problème de stratification des populations, largement décrit par la suite.
- **Les caractères catégoriels** : ce type de caractère s'exprime par une variable discrète qualitative ou quantitative avec plus de 2 valeurs possibles. Elles peuvent être ordonnées (par exemple petit, moyen, grand ou 1,2,3) ou non ordonnées (par exemple des sous types de maladies différents). Il peut être difficile de classer les individus dans une des catégories définies. C'est pourquoi, dans les catégories ordonnées, les chercheurs tendent souvent à donner un poids plus important aux individus extrêmes, certainement car leurs diagnostics sont plus évidents.
- **Les caractères quantitatifs** : une dernière catégorie de caractères que l'on peut rencontrer est celle des caractères quantitatifs. Ce sont des variables continues et non plus discrètes comme précédemment. Ils sont le plus fréquent dans le monde de l'élevage animal pour évaluer par exemple la production laitière, la taille d'un individu ou le nombre de

petits nés par an et par mère. Ceci dit il est très facile de passer d'un caractère quantitatif à un caractère binaire ou discret en choisissant par exemple dans la distribution des caractères les 5% à droite et les 5% à gauche comme cas et témoins.

Les différences entre ces types de caractères entraînent dans les analyses d'association l'utilisation de modèles mathématiques pouvant être différents. Le chapitre suivant donne les différents modèles adaptés pour ces types de données.

1.2.6 Conclusion : analyses préliminaires

Nous venons de voir au sein de ce chapitre l'importance des analyses préliminaires à la recherche de QTL affectant un caractère. Ces analyses préliminaires nous éclairent sur la qualité des données génotypiques, nous indiquent comment les améliorer lorsqu'elles manquent ou ne sont pas de bonne qualité, et ainsi, avec l'aide d'autres informations comme le déséquilibre de liaison ou l'information familiale disponible, nous donnent une voie méthodologique à suivre pour l'analyse de nos données. Lorsque l'on dispose d'une carte génétique dense (souvent le cas des marqueurs SNP) comme c'est le cas dans ce rapport, l'utilisation du déséquilibre de liaison dans la cartographie de QTL est essentiel. Il existe pour cela deux chemins possibles : les **analyses d'association (LDA)** et les **analyses combinant association et liaison (LDLA)**. Ces deux types d'analyses comprennent de nombreuses modélisations selon le type de caractère étudié, la structure familiale des données, ou selon le type d'analyse que l'on souhaite réaliser (i.e. analyse uni-SNP ou multi-SNP). Elles sont décrites en détail et séparément dans la suite de ce rapport.

1.3 Analyses d'association (LDA)

Cette section est consacrée aux méthodes pour la détection de QTLs basées sur l'utilisation du déséquilibre de liaison, couramment appelées méthodes d'association ou encore méthodes LDA. Le but de ces méthodes est de tester l'association entre un marqueur (ou plusieurs) et un phénotype dans une population d'individus supposés non apparentés. De nos jours, les puces SNP permettent d'avoir un grand nombre de marqueurs qui couvrent l'ensemble du génome et ces marqueurs sont suffisamment proches pour exploiter le LD qui existent entre ces marqueurs et entre ces marqueurs et d'éventuels QTL. De ce fait, lorsqu'une analyse d'association déclare un marqueur significatif, l'hypothèse sous-jacente est qu'il existe un QTL, proche du marqueur, qui est en LD avec lui.

Il est possible de faire des analyses LDA (on parlera plutôt d'analyse d'association) testant la présence d'un QTL confondu avec un marqueur, ou la présence d'un QTL en un point précis du génome connaissant les informations génotypique ou haplotypique qui l'entourent, ou tester toutes les positions à la fois. Nous verrons dans cette section le choix méthodologique le plus adapté à ces différentes possibilités mais également le plus adapté aux types de phénotypes possibles : binaires, catégoriels ou quantitatifs. Les protocoles pour la détection de QTLs en génétique animale sont souvent réalisés à partir d'individus apparentés, ce qui peut entraîner des problèmes de robustesse des analyses. Un gros paragraphe sera consacré à la leur résolution.

1.3.1 Analyse uni-QTL SNP par SNP

Dans cette section, on s'intéresse aux analyses d'association réalisées SNP par SNP. Les détails de ces méthodes seront discutés en fonction des différents types de caractères possibles. On parlera

des caractères binaires (cas-témoins), des caractères quantitatifs puis des caractères catégoriels en partant des modèles (ou tests) les plus basiques et en essayant d'aller vers des modèles plus poussés. Le problème de la robustesse et de la puissance seront abordées au fur et à mesure.

Différents tests en cas-témoins

Lorsqu'on s'intéresse à des caractères binaires et donc typiquement dans un protocole en cas-témoins, l'analyse la plus naturelle est de tester les génotypes du SNP contre le statut de l'individu. L'hypothèse nulle H_0 est donc : il n'y a pas d'association entre les lignes et les colonnes de la matrice 2×3 qui contient les nombres de génotypes en fonction du statut cas ou témoins de l'individu.

TABLE 1.2 – Exemple de matrice 2×3 contenant les génotypes (fréquences relatives) d'un SNP en fonction du statut de l'individu

	génotype			
	A_1A_1	A_1A_2	A_2A_2	
cas	10 (0.67)	4 (0.27)	1 (0.06)	15
témoins	5 (0.33)	5 (0.33)	5 (0.33)	15
	15 (0.50)	9 (0.30)	6 (0.20)	30

Avant de faire le test, on peut définir quelques termes qui nous serviront dans la suite du rapport. On appelle la *pénétrance* d'un génotype i , notée f_i , la probabilité d'être atteint sachant le génotype i .

$$f_i = P(\text{cas}|G_i) = \frac{\text{Nb de cas de génotype } i}{\text{Nb d'individus de génotype } i}$$

Avec $G_i, i = \{0, 1, 2\}$, le génotype i (0 correspond au génotype A_1A_1 , 1 au génotype A_1A_2 et 2 au dernier). Dans notre exemple, on obtient :

$$f_0 = \frac{10}{15} = 0.66, \quad f_1 = \frac{4}{9} = 0.44, \quad f_2 = \frac{1}{6} = 0.17.$$

On définit également les risques génotypiques relatifs (plus facile à interpréter) par $\lambda_i = f_i/f_0$ avec $i = 1, 2$ et ainsi on obtient dans cet exemple :

$$\lambda_1 = 0.67, \quad \lambda_2 = 0.25.$$

Les résultats dans cet exemple suggèrent que les risques génotypiques sont additifs car $\lambda_1 \sim (1 + \lambda_2)/2$ et que l'allèle A_1 est l'allèle responsable car $\lambda_1 < 1$. A partir de la Table 1.2, on peut réaliser un test χ^2 de Pearson à 2 degrés de liberté (df) ou un test exact de Fisher. On ne présente ici que les résultats obtenus avec un test du χ^2 car le test de Fisher est difficile à faire à la main pour des matrices supérieures aux matrices 2×2 .

$$\begin{aligned} X^2 &= \frac{(10 - 7.5)^2}{7.5} + \frac{(5 - 7.5)^2}{7.5} + \frac{(4 - 4.5)^2}{4.5} + \frac{(5 - 4.5)^2}{4.5} + \frac{(1 - 3)^2}{3} + \frac{(5 - 3)^2}{3} \\ &= 4.44 \end{aligned}$$

Cette statistique suit asymptotiquement une loi du χ^2 à 2 df (3 génotypes - 1). Ici, la *P-value* vaut 0.109 et donc, pour un seuil à 5%, on ne rejette pas H_0 (i.e. on accepte qu'il n'y ait pas d'effet

d'un génotype sur les cas-témoins). Néanmoins, ce résultat est fragile puisque, selon Cochran, il faut plus de 80% des cases remplies d'un nombre > 5 (sans qu'aucune ne soit vide) pour qu'il soit valide.

Ces tests sont assez puissants dans différents modèles génétiques (récessif, dominant, additif). Sous l'hypothèse fréquente que le risque génotypique est additif (i.e. le risque des hétérozygotes est l'intermédiaire des risques des homozygotes, figure 1.5), d'autres tests permettent d'obtenir plus de puissance. Le plus simple utilise les allèles plutôt que les génotypes ce qui diminue les degrés de liberté du test de Pearson de 2 à 1. Dans notre exemple, la table correspondante à ce test est présentée Table 1.3.

TABLE 1.3 – Exemple de matrice 2×2 contenant les allèles (fréquences relatives) d'un SNP en fonction du statut des individus

	Allèles		
	A_1	A_2	
cas	24 (0.8)	6 (0.2)	30
témoins	15 (0.5)	15 (0.5)	30
	39 (0.65)	21 (0.35)	60

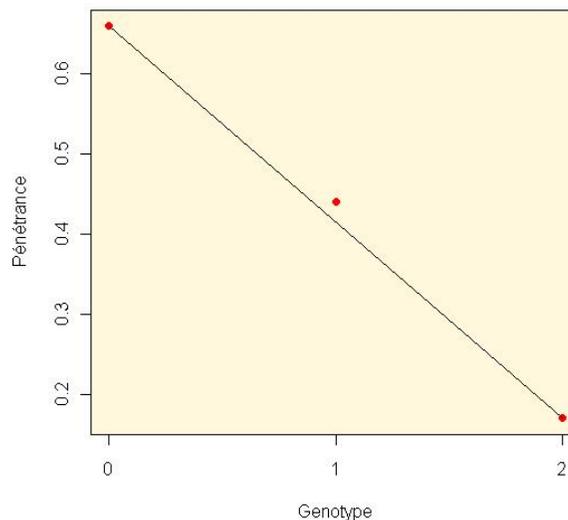


FIGURE 1.5 – **Évaluation du modèle génétique sous-jacent aux données de notre exemple.** Les points indiquent la pénétrance en fonction des trois génotypes possibles pour un SNP (ici codés 0,1,2). Ici les trois points sont quasiment alignés sur une droite, on peut donc supposer que le modèle génétique ou risque génotypique est additif.

Le test vaut maintenant :

$$\begin{aligned} X^2 &= \frac{(24 - 19.5)^2}{19.5} + \frac{(15 - 19.5)^2}{19.5} + \frac{(6 - 10.5)^2}{10.5} + \frac{(15 - 10.5)^2}{10.5} \\ &= 5.94 \end{aligned}$$

Sous H_0 , la statistique suit une loi du χ^2 à 1 df, donnant une P -value égale à 0.015. Conclusion, en fixant un seuil à 5%, l'hypothèse nulle H_0 est rejetée, i.e. on accepte qu'il y ait une différence significative entre les fréquences alléliques chez les cas et les témoins (ici l'allèle A_1 est significativement plus fréquent chez les cas que les témoins). Cet exemple illustre que le test de Pearson à 1 df d'un modèle génétique additif est plus puissant que le test de Pearson à 2 df. Cependant, Sasieni (1997), qui compare les différents tests d'association uni-SNP, montre qu'il faut faire attention à ces types de tests qui reposent sur des hypothèses d'HWE sur les cas et les témoins (la distribution du test allélique suit un χ^2 quand l'échantillon est en HWE, autrement elle peut être anti-conservative). Au contraire, le test de Cochran-Armitage (Armitage, 1955), également dit test Armitage ou trend test, s'affranchit de ces hypothèses HWE. En retour, ce test n'est pas robuste aux différents modèles génétiques et doit être utilisé uniquement lorsque celui-ci est connu. Si tel est le cas, il est asymptotiquement optimal.

Soit $T_x = T(x_0, x_1, x_2)$ la statistique du trend test, où $x_i \in [0, 1]$ correspond à un score affecté au génotype G_i servant à définir le modèle génétique sous-jacent. Généralement, on prend $x_0 = 0$, $x_2 = 1$ et on fait varier x_1 (0 pour un modèle récessif, 0.5 pour additif et 1 pour dominant). La statistique de test T_x correspondant au test Armitage s'écrit de la façon suivante :

$$T_x = \frac{(\frac{1}{n} \sum_{i=0}^2 x_i (sr_i - rs_i))^2}{(rs/n^2) \sum_{i=0}^2 n_i x_i (1 - x_i)}, \quad T_x \sim \chi_1^2$$

Avec les r_i correspondent aux nombres de génotypes i chez les cas ($\sum_i r_i = r$), s_i chez les témoins ($\sum_i s_i = s$) et n_i chez les deux ($\sum_i n_i = n$). La valeur de la statistique de test est ensuite comparée à une distribution du khi-deux à un degré de liberté. Dans notre exemple, le modèle est additif, on choisit donc $x_1 = 0.5$ et on obtient une statistique $T_x = 9.03$ qui donne une P -value de 0.003.

Il est difficile de choisir dans une étude un test relevant d'un modèle génétique généralement inconnu. Adopter un test Armitage implique une perte de puissance si le risque génotypique est loin de l'hypothèse faite, alors qu'on gagne beaucoup de puissance s'il est proche. Au contraire, le test de Fisher est robuste aux hypothèses sur le modèle génétique mais perd en puissance par rapport au test Armitage si celui-ci est bien adapté. Notons enfin qu'il existe d'autres tests qu'on peut qualifier d'intermédiaires qui pondèrent les trois modèles (récessif, dominant et additif) et choisissent le maximum des statistiques de tests de ces modèles (Freidlin et al., 2002).

Une autre approche des études cas-témoins est la **régression logistique**. Cette régression permet d'adapter aux caractères binaires les modèles linéaires (régression, ANOVA) développés pour des caractères quantitatifs.

L'idée est de trouver une fonction réelle monotone $g(f_i)$, appelée *fonction lien*, qui permette d'expliquer $f_i = P(\text{cas} | G_i)$ dans un modèle linéaire à l'aide de variables explicatives (e.g. $g(f_i) = x_i' \beta$). Plusieurs fonctions existent pour cela dont la plus connue est la fonction *logit* définie par :

$$\eta_i = g(f_i) = \text{logit}(f_i) = \ln \left(\frac{f_i}{1 - f_i} \right)$$

Avec :

$$\eta_i = \begin{cases} \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, & \text{(Modèle génotypique)} \\ \beta_0 + \beta_1 x_i, & \text{(Modèle allélique)} \\ \beta_0, & \text{(Sous } H_0) \end{cases}$$

Où les β_j sont les paramètres à estimer et les x_i et x_{ij} correspondent aux codes que l'on se donne pour les génotypes ou les allèles. Prenons par exemple le modèle génotypique, il existe 3 génotypes possibles (G_0 , G_1 , et G_2). Si un individu est de génotype G_0 , alors on lui affecte la valeur 0 dans x_{1i} et x_{2i} . Si il est de génotype G_1 alors on lui affecte les valeurs $x_{1i} = 1$ et $x_{2i} = 0$ et inversement pour le cas G_2 . Le test de l'hypothèse nulle $\beta_1 = \beta_2 = 0$ par le rapport de vraisemblance a 2 df et est équivalent à un test de Pearson à 2 df pour des échantillons de grandes tailles.

Si maintenant on s'intéresse au cas du modèle allélique, le codage de x dépend des hypothèses. Par exemple, pour un modèle additif, on affectera à x_i la valeur 0 (resp. 1, 2) si le génotype est G_0 (resp. G_1 , G_2). Pour un modèle récessif ou dominant, $x_i = \{0, 0, 1\}$ ou $x_i = \{0, 1, 1\}$ en fonction du génotypes des individus. Notons que le test du rapport de vraisemblance associé au modèle allélique additif est équivalent à un test Armitage et que tous les tests sur le modèle allélique ont 1 df.

La régression logistique est assez peu utilisée dans les analyses SNP par SNP, au profit des tests du Score, développés dans le but de générer des tests asymptotiquement équivalents aux tests de rapport de vraisemblance, car ils sont plus rapide à calculer. Les tests de Pearson et Armitage sont d'ailleurs des tests du Score sous les conditions du modèle de régression logistique décrites plus haut. Les avantages de la régression logistique (ou des tests du score associés) sont sa flexibilité permettant d'incorporer des covariables fixées (ex : sex, age) ou des interactions gènes-environnement ($G \times E$) et sa facilité d'extension du modèle, par exemple pour des analyses sur plusieurs SNP (voir la section sur les SNP multiples).

Les caractères catégoriels

Pour analyser ces types de caractères, on utilise les méthodes de la régression logistique polytomique (qui correspond à des variables dépendantes à $K > 2$ modalités). Pour des variables catégorielles désordonnées (sous classes de maladies), l'analyse via une régression logistique multinomiale semble être une bonne approche. En ce qui concerne les variables catégorielles ordonnées, diverses méthodes et modélisations ont été proposées comme le modèle *Logit adjacent à coefficients constants* ou le modèle *Logit odds-ratio cumulatifs proportionnels*. Ce dernier est le plus connu mais il est difficile de calculer une statistique du score correspondant à ce modèle. De ce fait, nous ne présentons ici que le modèle Logit adjacent.

On fait l'hypothèse que le risque de la catégorie k par rapport à la catégorie $k - 1$ est le même pour tous les k . Le principe est donc de calculer le Logit du passage d'une catégorie à l'autre :

$$\eta_{i,k} = \text{logit}(f_{i,k}) = \ln \left(\frac{f_{i,k}}{f_{i,k+1}} \right), \quad \text{avec } \sum_{k=1}^K f_{i,k} = 1$$

Avec (exemple du modèle génotypique) :

$$\eta_{i,k} = \beta_{0,k} + \beta_1 x_{1i} + \beta_2 x_{2i}, \quad 1 \leq k < K - 1$$

Notons également que la comparaison entre deux classes éloignées peut se faire :

$$\ln\left(\frac{f_{i,k}}{f_{i,1}}\right) = -\eta_{i,k-1} - \dots - \eta_{i,2} - \eta_{i,1}$$

A partir de là, un test de rapport de vraisemblance, basé sur un modèle multinomial, est effectué. La log-vraisemblance s'écrit de la manière suivante :

$$LL = \sum_i \sum_{k=1}^K y_{i,k} \ln(f_{i,k})$$

Avec $y_{i,k}$ le phénotype de l'individu i pour la catégorie k . D'autres tests sont possibles comme le test du score, qui est asymptotiquement équivalent au test du rapport de vraisemblance et qui, dans ce cas là, est une généralisation du test Armitage (Balding, 2006).

Les caractères quantitatifs

Les méthodes statistiques les plus naturelles pour réaliser une analyse d'association SNP par SNP pour des caractères quantitatifs sont les modèles d'analyse de la variance (ANOVA) et les modèles de régression linéaire.

Ainsi l'ANOVA permet d'expliquer la variable quantitative Y à partir d'un facteur qualitatif à plusieurs modalités ou niveaux, typiquement un génotype a $I = 3$ modalités ("11", "12" et "22"). Soit Y_{ij} la $j^{\text{ème}}$ répétition du niveau i avec $i = 1 \dots I$, $j = 1 \dots n_i$ et $\sum_{i=1}^I n_i = n$ qui correspond au nombre total d'individus. Le modèle ANOVA s'écrit :

$$y_{ij} \sim N(\mu_i, \sigma^2)$$

Ce qui est équivalent à :

$$\begin{aligned} y_{ij} &= \mu_i + e_{ij}, & \text{avec } e_{ij} &\sim N(0, \sigma^2) \\ &= \mu + \alpha_i + e_{ij}, & \text{avec } \sum_i \alpha_i &= 0 \quad (\text{Décomposition centrée}) \end{aligned}$$

Où α_i est l'effet du génotype i et e_{ij} le résidu aléatoire centré, les résidus étant supposés iid. Il est possible d'utiliser l'écriture matricielle $Y = X\beta + E$. L'estimation des paramètres se fait via la méthode des moindres carrés ordinaires :

$$\begin{aligned} \hat{\mu} = y_{\bullet\bullet} &= \frac{1}{n} \sum_i \sum_j y_{ij}, \quad \forall i : \hat{\alpha}_i = y_{i\bullet} - y_{\bullet\bullet} \\ \text{et } \hat{\sigma}^2 &= \frac{1}{n-I} \sum_i \sum_j \hat{e}_{ij}^2 = \frac{1}{n-I} \sum_i \sum_j (y_{ij} - y_{i\bullet})^2 \end{aligned}$$

Les logiciels SAS et R utilisent d'autres type de décompositions qualifiées de *niveau de référence*, respectivement $\alpha_I = 0$ et $\alpha_1 = 0$, qui mènent à des estimations des paramètres différentes.

Plusieurs tests correspondant à des hypothèses distinctes sont alors envisageables :

1. un test global : On teste s'il y a un effet du facteur génotype. H_0 s'écrit : $\{y_{ij} = \mu + e_{ij}\}$ contre $H_1 = \{y_{ij} = \mu + \alpha_i + e_{ij}\}$. Ce test correspond à un test de fisher $F_{I-1, n-I}$. En terme de paramètres, ce test s'écrit de la forme $H_0 = \{\forall i, \alpha_i = 0\}$ contre $H_1 = \{\exists i, \alpha_i \neq 0\}$.

2. un test spécifique aux paramètres : On teste s'il y a un effet spécifique d'un des niveaux du génotype. C'est à dire $H_0 = \{\alpha_i = 0\}$ contre $H_1 = \{\alpha_i \neq 0\}$. Sous H_0 , le test est effectué au moyen d'une statistique de student T_{n-I} . Attention à ces tests qui dépendent de la contrainte choisie au départ.

L'ANOVA est analogue à un test de Pearson à 2 df car elle compare l'hypothèse nulle qu'il n'y a pas d'association entre le génotype et le caractère étudié avec une alternative générale. Comme pour les caractères binaires, la régression permet de réduire le nombre de df à 1 en faisant l'hypothèse d'une relation linéaire entre la moyenne du caractère et le génotype (i.e. considérer un modèle additif et donc utiliser un modèle allélique). Les deux modèles nécessitent que la variance du caractère soit la même pour chaque génotype et que le caractère soit distribué selon une loi normale. Ceci dit, ces modèles sont assez robustes aux écarts à la normalité et il est toujours possible de transformer le caractère via le \log par exemple pour approximer celle ci.

Pour écrire le modèle de régression, les génotypes "11", "12" et "22" sont recodés 0, 1 et 2, et apparaissent alors comme le nombre d'allèles 2. Le modèle devient allélique et s'écrit :

$$\forall i, \quad y_i = \beta_0 + \beta_1 x_i + e_i, \quad \text{avec } e_i \sim N(0, \sigma^2), \quad \{e_i\}_{i=1\dots n} \text{ iid}$$

Avec β_0 la moyenne générale, β_1 l'effet additif ou allélique, x_i le nombre d'allèles 2 observé chez l'individu i et e_i les résidus. Ce modèle s'écrit matriciellement $Y = X\beta + E$. L'estimation des paramètres se fait aussi à l'aide des moindres carrés ordinaires et la statistique de test permettant de tester $H_0 = \{y_i = \beta_0 + e_i\}$ contre $H_1 = \{y_i = \beta_0 + \beta_1 x_i + e_i\}$, équivalent à $H_0 = \{\beta_1 = 0\}$ contre $H_1 = \{\beta_1 \neq 0\}$, suit approximativement une loi de Fisher $F_{1,n-2}$ ou une loi de student t_{n-2} .

TABLE 1.4 – Exemple de l'utilisation des deux modèles (ANOVA et régression linéaire) sur un petit jeu de données. L'ANOVA est réalisée avec les données "Génotype" et donne une valeur $F_{2,7} = 67.05$ correspondant à une P -valeur de $2.719e - 05$. Donc il y a bien un effet du génotype au marqueur sur le phénotype si l'on fixe un seuil à 5%. La régression linéaire est réalisée à partir des données "Allèle 2" et donne une $F_{1,8} = 150.7$ correspondant à une P -valeur de $1.802e - 06$. Il y a donc bien un effet de l'allèle 2 au marqueur sur le phénotype.

Animal	Phénotype	Génotype	Allèle 2
1	2.030502	1/1	0
2	2.335734	1/1	0
3	2.646192	1/1	0
4	3.542274	1/2	1
5	3.834241	1/2	1
6	3.407128	1/2	1
7	3.762855	1/2	1
8	3.689349	1/2	1
9	3.685757	1/2	1
10	4.871137	2/2	2

Ces modèles testent si un QTL est associé avec le marqueur. Cela se produit dans deux cas : soit le marqueur est le QTL, soit le marqueur est en déséquilibre de liaison avec le QTL. Il existe évidemment d'autres modèles pour exprimer cette association entre le marqueur et le QTL. Ceux

exprimés ici sont les plus basiques et sont souvent utilisés comme référence en comparaison avec d'autres méthodes (Grapes et al., 2004; Zhao et al., 2007a).

Facteurs affectant la puissance et limites de ces méthodes

On définit la puissance comme la probabilité pour que la valeur de la statistique de test soit dans la région critique quand H_1 est vraie :

$$\text{Puissance} = P(\text{Rejeter } H_0 | H_1) = 1 - \beta, \quad \text{avec } \beta = P(\text{erreur type II})$$

Donc, lorsqu'on fait une analyse d'association, la puissance est la probabilité de rejeter l'hypothèse nulle (pas d'association entre le marqueur et le QTL) quand un QTL existe réellement dans la population. Elle dépend de plusieurs choses :

1. Du r^2 entre le marqueur et le QTL. On a vu dans le premier chapitre que la taille de l'échantillon doit être augmentée par un facteur $1/r^2$ pour détecter un QTL non génotypé, comparé à la taille de l'échantillon pour tester le QTL lui même (Pritchard and Przeworski, 2001).
2. De la proportion de la variance phénotypique totale expliquée par le QTL, notée h_q^2 .
3. Du nombre N d'individus phénotypés.
4. De la fréquence allélique de l'allèle rare au marqueur. Difficile lorsque cette fréquence est inférieure à 5%.
5. Du seuil de significativité α fixé par l'utilisateur.

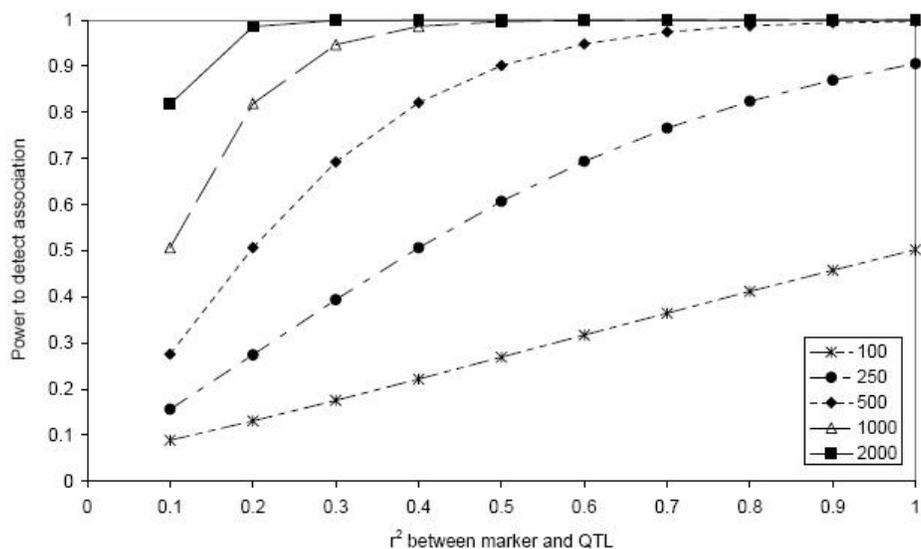


FIGURE 1.6 – Puissance de détection d'un QTL qui explique 5% de la variance phénotypique avec un marqueur. Expérience réalisée avec différents nombres d'individus présents dans la population, différents r^2 , une P -valeur à 5% et les fréquences de l'allèle rare du QTL et du marqueur sont toutes deux à 0.2 (d'après le cours du Pr. Ben Hayes)

En utilisant la figure 1.6 et la théorie sur l'étendue du LD, on peut faire des prédictions sur le nombre de marqueurs nécessaires pour détecter un QTL dans une étude d'association sur tout le génome.

Toutes les méthodes décrites dans cette section souffrent cependant d'un problème majeur. En effet, elles font toutes l'hypothèse que les individus sont non apparentés ou encore qu'il n'y a pas de structure de population. Cependant, dans les populations animales et dans moindre mesure chez les humains, ce n'est malheureusement pas le cas et la non prise en compte de cette information dans une analyse d'association entraîne de nombreux faux positifs (Cardon and Palmer, 2003; Clayton et al., 2005; Marchini et al., 2004). L'objet de la prochaine section est de décrire les méthodes qui permettent de lutter contre ces faux positifs.

1.3.2 Lutter contre les effets d'une éventuelle stratification de la population

De nombreux facteurs (sélection, goulot d'étranglement, structure familiale, pedigree, effets de milieu) affectent la structure de population, créant une hétérogénéité des fréquences alléliques entre les sous groupes de la population. Il est important de distinguer deux types de stratification : (1) une structure familiale de population se concrétisant par un apparentement entre individus ; (2) un mélange de populations.

Le problème de la stratification peut être illustré dans le cadre d'un mélange de populations dans un protocole cas témoins. Notons A_1 et A_2 les allèles au marqueur A et créons un mélange de deux populations X et Y :

Population X			Population Y			Population X+Y		
	A1	A2		A1	A2		A1	A2
Cas	160	160	Cas	160	40	Cas	320	200
Témoins	40	40	Témoins	160	40	Témoins	200	80

On obtient dans les populations X et Y une statistique $X^2 = 0$ concluant qu'il n'y a pas d'effet allélique sur le phénotype. Cependant pour la population $X+Y$, on obtient une statistique de test $X^2 = 7.81$ qui, lorsqu'on la compare à une loi du χ_1^2 à 5%, est significative. Ce faux positif est du uniquement au mélange de populations. Le tableau ci-dessous illustre l'apparition d'un faux négatif (la statistique pour la population X est 144, pour la population Y est 144 et pour la population $X+Y$ est 0).

Population X			Population Y			Population X+Y		
	A1	A2		A1	A2		A1	A2
Cas	160	40	Cas	40	160	Cas	200	200
Témoins	40	160	Témoins	160	40	Témoins	200	200

Pour lutter contre les problèmes statistiques liés à la stratification, l'idée est de créer une distance entre individus et d'intégrer cette information dans le modèle statistique. Pour cela, différentes méthodologies plus ou moins complexes se sont développées autour de deux axes qui utilisent soit l'information sur les parentés entre individus construite à l'aide du pedigree et/ou des génotypes aux marqueurs, soit l'information moléculaire (les SNP) sans prendre en compte l'information sur les parents.

Utiliser l'information sur les parentés

Pour lutter contre les effets d'une éventuelle stratification, l'information sur les parentés est utile de deux manières non exclusives : (1) elle permet de tracer la transmission des segments chromosomique entre générations et donc typiquement d'introduire l'analyse de liaison dans le modèle (LDLA) et, (2) elle permet de construire une matrice de parenté qui structure les dépendances entre phénotypes. Le premier point, largement documenté sera détaillé dans la section suivante (sur les méthodes LDLA).

Le second point revient à utiliser un modèle mixte :

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

Avec X la matrice d'incidence liant le génotype (ou les allèles) aux phénotypes de chaque individu, β l'effet génotypique ou allélique, Z est la matrice d'incidence des effets polygéniques liant chaque individu à son phénotype et u les effets polygéniques aléatoires avec $u \sim N(0, A\sigma_a^2)$. Tout l'intérêt de ce modèle vient du fait de traiter les effets polygéniques en aléatoire et ainsi se donner une structure de la matrice de covariance de ces effets. La matrice de parenté A est égale au double de la matrice des coefficients de parenté (au sens de Malécot) entre individus phénotypés ($A_{ij} = 2\phi_{ij}$ où ϕ_{ij} est le coefficient de parenté entre l'individu i et j). La Table 1.6 donne l'exemple de la matrice A correspondant à un pedigree fait d'une famille de demi-frères (Table 1.5).

TABLE 1.5 – Exemple d'un pedigree d'une famille de demi-frère

Animal	Père	Mère
1	0	0
2	0	0
3	0	0
4	1	2
5	1	2
6	1	3

TABLE 1.6 – Matrice A correspondant au pedigree de la Table 1.5

	Animal 1	Animal 2	Animal 3	Animal 4	Animal 5	Animal 6
Animal 1	1					
Animal 2	0	1				
Animal 3	0	0	1			
Animal 4	0.5	0.5	0	1		
Animal 5	0.5	0.5	0	0.5	1	
Animal 6	0.5	0	0.5	0.25	0.25	1

L'utilisation de ce type de modèle est courante et efficace. MacLeod et al. (2008) ont montré que le nombre de faux positifs est fonction de l'information disponible sur le pedigree (plus il y a d'informations valides sur le pedigree et moins il y aura de faux positifs). Cependant, cette information pedigree n'est pas toujours disponible et peut présenter parfois des erreurs. Une solution est alors de construire la matrice de parenté avec l'information génomique, c'est à dire l'information des marqueurs (décrite par la suite).

Utiliser l'information génomique

Plusieurs méthodes permettent de lutter contre les effets de la stratification de population sans nécessiter d'information familiale. Beaucoup d'entre elles reposent sur l'idée qu'un ensemble de marqueurs nuls (sans liaison avec des QTL) répartis sur tout le génome peuvent être utilisés pour apprécier une éventuelle structure sous-jacente de la population (Pritchard and Rosenberg, 1999). Leur raisonnement vient du fait que comme ces marqueurs sont indépendants de ceux qui affectent la maladie, et ne sont pas corrélés entre eux, alors ils doivent refléter les différences génétiques des populations sous-jacentes. Ainsi, il est possible de quantifier le niveau de structuration de populations et d'utiliser des méthodes pour le corriger. Un marqueur nul est typiquement un marqueur dont le test d'association classique (à 5%) n'a pas été significatif. On choisit régulièrement les marqueurs les moins significatifs et espacés sur tout le génome. D'après Balding (2006), il faut au minimum 100 SNP nuls bien répartis sur le génome pour identifier ce type de problème.

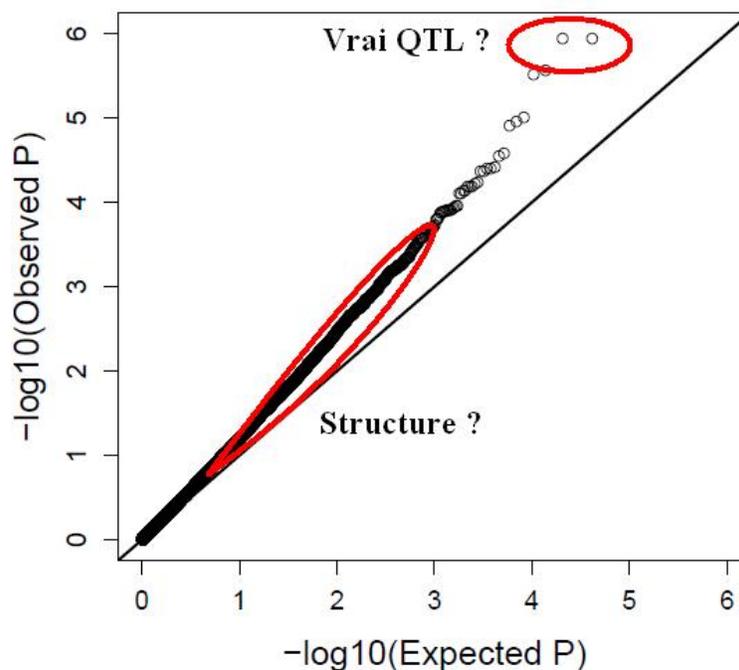


FIGURE 1.7 – Exemple d'inflation des statistiques de tests dans un *log QQ plot*. Elle est certainement due à la structure de la population

Le Contrôle Génomique :

Introduit par Devlin and Roeder (1999) pour des modèles cas-témoins, le contrôle génomique est l'une des toutes premières méthodes pour lutter contre les effets d'une éventuelle stratification de la population. L'idée très simple de ces auteurs est de faire tout d'abord une première analyse en cas-témoins avec un test Armitage sur chaque locus i , T_i . Dans un second temps, un facteur d'inflation λ sur les locus nuls (> 100) est calculé :

$$\lambda = \text{Mediane}(T_1, T_2, \dots, T_k) / 0.455$$

Où T_k représente le test Armitage au k^{iem} locus nuls et 0.455 est la médiane d'un χ_1^2 . L'idée ici est que, comme on s'attend à avoir peu ou pas du tout de SNP nuls associés avec le phénotype, une valeur de $\lambda > 1$ est certainement due à la stratification de population. Dans la figure 1.7, $\lambda = 1.11$ et donc il y a certainement un effet lié à la stratification de la population. En divisant les tests par ce facteur, on espère corriger pour la stratification de population. Il est possible de montrer que :

$$T_i \sim \lambda \chi_1^2$$

Il existe une stratégie identique pour des tests à 2 df proposée par Zheng et al. (2006). Plusieurs extensions, notamment une pour des caractères quantitatifs, ont été réalisées (Bacanu et al., 2002). Dans un premier temps, un test d'association allélique comme décrit dans la section précédente est réalisé ($y_i = \beta_0 + \beta_1 x_i + e_i$), puis une statistique de test de student ($T_k = \widehat{\beta}_1 / \sqrt{\text{var}(\widehat{\beta}_1)}$) est calculée pour les $k = 1, \dots, K$ locus nuls. Sous l'hypothèse nulle et pour de grande taille d'échantillons, T_k suit approximativement une distribution $N(0, \lambda)$ et donc $T_k^2 / \lambda \sim \chi_1^2$ avec λ défini comme précédemment.

Les gros avantages de cette méthode sont sa facilité à être mise en place et sa rapidité. Par contre, son défaut est que la correction est réalisée de la même manière pour tous les SNP. De plus, on peut montrer que λ est fonction de la taille de la population et que cette méthode est assez conservative pour des effectifs faibles et inversement pour de grands effectifs. De la même manière, elle peut être anti-conservative si un nombre insuffisant de SNP nuls sont pris en compte (Marchini et al., 2004). Son application en génétique animale est plus controversée puisque les populations animales étant fortement apparentées en général, les tests classiques qui ne corrigent pas pour la structure de population seront très inflatés (λ très grand). Dans ce cas, diviser l'ensemble des tests par ce même λ fait perdre une grosse partie de la puissance.

Structure d'Association :

L'idée de cette méthode est d'utiliser les marqueurs nuls pour affilier chaque individu à une sous-population et d'ensuite réaliser un test d'association classique au sein de ces sous-populations pour chacun des marqueurs considérés. C'est donc une analyse en deux parties :

1. Une première partie, pour laquelle le logiciel *STRUCTURE* a été développé par Pritchard et al. (2000a), a pour idée d'utiliser les marqueurs nuls pour apprendre sur la structure de la population et affilier chaque individu à une sous-population.

Considérons un échantillon d'individus non apparentés et génotypés à plusieurs locus non liés (i.e. pas en LD), que l'on note G . De plus, soit Z la population d'origine (inconnue) des individus et P les fréquences alléliques dans les sous-populations (inconnues). En utilisant un modèle bayésien, on cherche alors :

$$Pr(Z, P|G) \propto Pr(Z)Pr(P)Pr(G|Z, P)$$

La distribution exacte de cette distribution n'est pas facile à avoir mais il est possible de l'approximer en utilisant un algorithme MCMC (Markov Chain Monte Carlo). On fait l'hypothèse qu'il y a K sous-populations. Pritchard et al. (2000a) proposent deux modèles, un modèle sans et avec métissage de populations, c'est à dire qu'un individu peut être affilié à une ou K populations. Détaillons celui sans métissage. Les quantités $Pr(Z)$, $Pr(P)$ et $Pr(G|Z, P)$ sont définies par :

$$Pr(g_{il}^a = j|Z, P) = p_{z^{(i)}lj}$$

Avec g_{il}^a l'allèle parental a de l'individu i au locus l et $p_{z^{(i)}lj}$ la fréquence de l'allèle j au locus l dans la population d'origine $z^{(i)}$ de l'individu i . Si les populations d'origines sont équiprobables :

$$Pr(z^{(i)} = k) = 1/K$$

Enfin, les fréquences alléliques du marqueur l au sein de la sous-population k sont supposées distribuées dans une Dirichlet :

$$p_{kl\bullet} \sim \mathcal{D}(\lambda_1, \lambda_2), \quad \text{avec} \quad \sum_{j=1}^2 p_{klj} = 1 \quad \text{et on prend souvent} \quad \lambda_1 = \lambda_2 = 1.0$$

Ces trois équations permettent d'utiliser l'algorithme MCMC et de calculer les probabilités $Pr(P|G, Z)$ et $Pr(Z|G, P)$, ce qui donne approximativement la distribution souhaitée $Pr(Z, P|G)$. Dans le cas d'un métissage de populations, les vecteurs $Z^{(i)}$ sont remplacés par la matrice d'ascendance Q où q_{ik} est la probabilité qu'un gène de l'individu i provienne de la sous-population k .

2. La seconde partie, appelée *STRAT* et développée par Pritchard et al. (2000b), est un test d'association en chaque locus qui prend en compte les sous-populations. Dans *STRAT*, un test du rapport de vraisemblances $-2\log(\Lambda)$ est réalisé de l'hypothèses H_0 : "Les fréquences alléliques dans les sous-populations au locus testé ne dépendent pas du phénotype (i.e. \hat{p}_0)" contre H_1 : "Les fréquences alléliques dans les sous-populations au locus testé dépendent du phénotype (i.e. \hat{p}_1)" avec :

$$\Lambda = \frac{P_{H_1}(G|\hat{p}_1, \hat{Z}, Y)}{P_{H_0}(G|\hat{p}_0, \hat{Z}, Y)}$$

Avec Y le phénotype (cas ou témoins, i.e 0 ou 1), \hat{Z} estimé à l'aide de *STRUCTURE* et \hat{p}_0 , \hat{p}_1 estimés à l'aide de l'algorithme EM (Expectation-maximisation). La P -value de ce test est obtenue par simulations.

L'avantage principal de cette méthode est qu'elle donne des détails sur la structure de la population. Elle permet également d'intégrer un métissage de populations et peut aisément être étendue aux cas multi-locus ou haplotypiques. Cependant, cette méthode est lourde en temps de calcul, en particulier pour les SNPs. De plus, le modèle est construit avec un nombre de sous-populations fixé, et ne reflète donc pas vraiment la réalité. La question du nombre de sous-populations est une question qui n'est pas encore résolue. Cette analyse nécessite plus de marqueurs nuls que le contrôle génomique et souffre d'une perte de puissance lorsqu'il n'y a pas ou peu de structure de population. Notons enfin que d'autres auteurs se sont intéressés à cette approche en développant d'autres tests (Hoggart et al., 2003) ou en utilisant des approches par maximum de vraisemblance (Satten et al., 2001).

L'analyse par composante principale :

Price et al. (2006) ont proposé d'utiliser les marqueurs nuls dans une analyse en composante principale (ACP) pour diminuer le nombre de dimensions des données sur des axes continus de variation génétique (vecteurs propres) qui décrivent le mieux possible la variabilité entre les individus. Les axes ainsi définis, un ajustement des génotypes et des phénotypes est réalisé via la contribution de chaque individu sur un nombre d'axes fixé. Enfin, une dernière étape consiste à faire une analyse d'association classique SNP par SNP sur ces génotypes et phénotypes ajustés.

Soit g_{ji} le génotype pour le SNP j et de l'individu i (codé 0, 1, 2), avec $j = 1, \dots, M$ et $i = 1, \dots, N$. La première idée est de créer une matrice X centrée réduite des génotypes g_{ji} , i.e :

$$X(j, i) = \frac{g_{ji} - \mu_j}{\sqrt{p_j(1 - p_j)}}, \quad \text{avec} \quad \mu_j = \frac{\sum_i g_{ji}}{N} \quad \text{et} \quad p_j = \frac{\mu_j}{2}$$

Où μ_j représente la moyenne génotypique du SNP j et p_j représente la fréquence allélique du SNP j . On peut alors calculer la matrice de variance-covariance $\Psi = \frac{1}{M}X'X$ que l'on peut réécrire à l'aide de la décomposition en valeur singulière de X de la façon suivante :

$$\Psi = VS^2V^T$$

où V représente la matrice des vecteurs propres de Ψ et S^2 est la matrice des valeurs singulières au carré. La matrice V , $N \times N$, est particulièrement intéressante car chaque colonne de cette matrice représente un vecteur propre de Ψ et donc un axe de variation. On note alors a_{ik} la contribution sur l'axe k de l'individu i , i.e. c'est l'ancêtre du génotype de l'individu i . En fait, chaque axe crée une relation entre les génotypes des individus et donc une histoire entre chacun d'eux.

A partir de là, on peut ajuster les génotypes initiaux par rapport à ces axes et ainsi intégrer l'information des relations entre individus qui vont permettre de lutter contre la stratification de la population. L'ajustement se fait de la manière suivante :

$$g_{ji,adjusted} = g_{ji} - \sum_k \gamma_{jk} a_{ik}, \quad \text{avec} \quad \gamma_{jk} = \frac{\sum_i a_{ik} g_{ji}}{\sum_i a_{ik}^2}$$

Où γ_{jk} est le coefficient de régression prédit de l'ancêtre du génotype de l'individu i au SNP j au k^{iem} axe de variation (s'il n'y a pas de génotypes manquants, alors $\sum_i a_{ik}^2 = 1$). Le même ajustement est réalisé sur les phénotypes P_i . Cette démarche est identique à une régression multiple où les axes seraient des covariables.

La détection se fait à l'aide d'un test Armitage entre les génotypes et les phénotypes ajustés. La statistique de ce test, qui suit approximativement une loi de χ^2 , est équivalente à $(N - K - 1)$ fois le carré de la corrélation entre les génotypes ajustés et les phénotypes ajustés. Cette méthode est implémentée dans le logiciel *EIGENSTRAT*. La Figure 1.8 propose une illustration d'une analyse par ACP.

Patterson et al. (2006) ont montré que l'ACP peut être utilisée pour identifier des structures de population avec des données d'environ 100000 marqueurs et peut prendre en compte le LD entre les locus. Cependant, il faut faire attention à l'interprétation des vecteurs propres qui ne donnent pas uniquement de l'information sur la structure de population mais également une indication sur l'étendue du LD.

Approche génomique du Modèle mixte avec effet polygénique aléatoire :

Revenons au modèle mixte présenté précédemment. Le modèle est :

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

Avec pour rappel, X la matrice d'incidence liant génotypes ou allèles à chaque individu, β l'effet génotypique ou allélique, Z la matrice d'incidence des effets polygéniques liant chaque individu à son phénotype et u les effets polygéniques aléatoires avec $u \sim N(0, A\sigma_u^2)$. Quand seule l'information

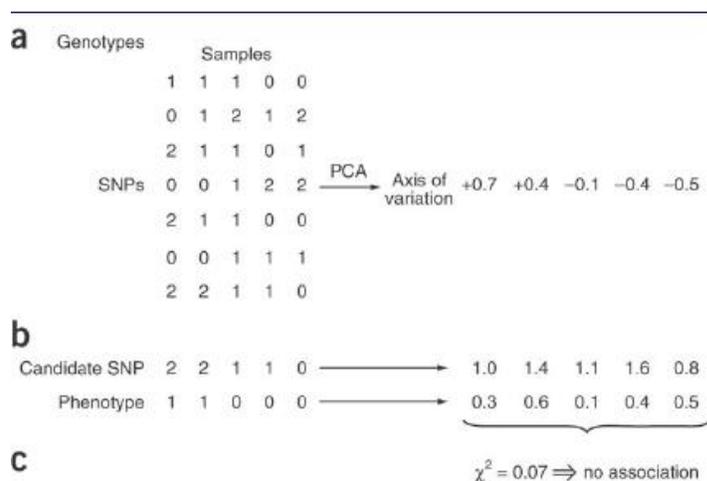


FIGURE 1.8 – Exemple de l'utilisation de la méthode d'association par ACP. (a) On a tout d'abord la matrice X dont on tire un axe de variation après avoir réalisé l'ACP. (b) Les génotypes et les phénotypes sont alors ajustés à partir de cet axe. (c) Enfin une analyse d'association est réalisée. From Price et al. (2006).

pedigree est utilisée, les éléments A_{ij} de A sont égaux à $2\phi_{ij}$ (avec ϕ_{ij} le coefficient de parenté entre l'individu i et j).

L'estimation de ces coefficients A_{ij} dépend de la quantité d'information disponible sur le pedigree (nombre de générations) et peut être affectée par les erreurs fréquentes de filiation. Les marqueurs SNP permettent de construire avec un degré de certitude contrôlable une matrice de parenté génomique G et non plus familiale A . La modélisation est identique et on cherche à écrire la matrice de variance-covariance G de $u \sim N(0, G\sigma_u^2)$. Diverses formules ont été proposées pour cela, nous en présentons ici trois :

1. Amin et al. (2007); Yang et al. (2010) proposent la mesure suivante :

$$G_{ij} = \frac{2}{m} \sum_{k=1}^m \frac{(g_{ik} - p_k)(g_{jk} - p_k)}{p_k(1 - p_k)}$$

Avec g_{ik} le génotype de l'individu i au SNP k (codé 0,1/2 et 1), p_k la fréquence de l'allèle majeur et m le nombre de SNP utilisés pour l'estimation de coefficient de parenté. Cette formule est également utilisée par Price et al. (2006) et Patterson et al. (2006) mais à un coefficient 2 près. A noter que les marqueurs monomorphes ne peuvent pas être pris en compte et que cette formule attribue un poids trop important aux marqueurs avec une faible fréquence de l'allèle mineur (MAF).

2. VanRaden (2008) propose une modification de la mesure précédente :

$$G_{ij} = 2 \frac{\sum_{k=1}^m (g_{ik} - p_k)(g_{jk} - p_k)}{\sum_{k=1}^m p_k(1 - p_k)}$$

La différence se fait sur la normalisation, ici on normalise sur l'ensemble des SNPs et non plus SNP par SNP.

3. Ritland (1996) et Hayes et al. (2007) proposent une approche par similarité. Pour un locus k donné, on calcule d'abord un indice de similarité entre deux individus i et j , et on obtient :

$$S_{ij} = \begin{cases} 1 & \text{si } g_{ik} = g_{jk} \\ \frac{1}{2} & \text{si } g_{ik} = 11 \text{ et } g_{jk} = 12 \text{ et inversement} \\ 0 & \text{si aucun allèle n'est en commun, exemple } g_{ik} = 11 \text{ et } g_{jk} = 22 \end{cases}$$

Ensuite, on calcule l'indice de similarité due au hasard par la formule $s_k = p^2 + q^2$ avec p et q les fréquences des allèles 1 et 2 au locus k . Le coefficient de similarité entre deux individus au locus k est alors défini par la formule :

$$R_{ijk} = \frac{S_{ijk} - s_k}{1 - s_k}$$

Les coefficients G_{ij} sont alors définis comme la moyenne des R_{ijk} sur tous les locus ($G_{ij} = \frac{1}{m} \sum_{k=1}^m R_{ijk}$). Notons que dans cette formulation, la diagonale de la matrice G vaut forcément 1 et donc on ne tient pas compte de la consanguinité.

On tire de ces formules les propriétés suivantes :

- Le coefficient de consanguinité d'un individu i est égal à $G_{ii} - 1$
- Le coefficient de parenté ϕ_{ij} (au sens de malécot) entre un individu i et j est

$$\phi_{ij} = \frac{1}{2} \frac{G_{ij}}{\sqrt{G_{ii}} \sqrt{G_{jj}}}$$

En l'absence de consanguinité, on obtient la formule classique $G_{ij} = 2\phi_{ij}$

Autres approches :

Il existe d'autres approches qui permettent de lutter contre les effets d'une éventuelle stratification de la population. Par exemple, Setakis et al. (2006) utilisent les marqueurs nuls comme covariables dans une régression logistique. Cette méthode est très rapide et flexible car la régression logistique permet d'intégrer des effets d'épistasie ou d'autres covariables dans le modèle. Epstein et al. (2007) proposent quant à eux une procédure en deux étapes : une première étape qui modélise la cote de la maladie sachant l'informativité des locus pour calculer un score de stratification ; dans un second temps, chaque individu se voit assigner à un groupe défini par le score de stratification, puis un test d'association entre la maladie et le locus au sein des groupes est réalisé.

Conclusion sur la prise en compte de la stratification de la population

Dans cette section, nous avons vu différentes méthodes qui permettent de prendre en compte les effets liés à la stratification de la population dans le modèle. Le meilleur choix parmi ces différentes méthodes dépend en pratique de deux facteurs.

Le premier facteur est le type de stratification que l'on suppose pour la population d'étude : dans le cas d'un mélange de populations, une analyse de type ACP ou STRUCTURE distingue mieux les groupes qu'un modèle mixte polygénique ; Si à l'inverse on pense que les individus de la (ou les) population(s) sont apparentés, la prise en compte de l'apparentement dans un modèle mixte avec effet polygénique en aléatoire est la solution. Certains auteurs proposent d'ailleurs de combiner ces différentes approches au sein d'un même modèle capable de faire face simultanément

à ces sources de stratification. Ainsi, Yu et al. (2006) ont proposé d'intégrer la matrice Q obtenue à l'aide de *STRUCTURE* dans un modèle mixte avec des effets polygéniques aléatoires et montrent que leur modèle est meilleur que le contrôle génomique et le modèle mixte polygénique standard dans plusieurs scénarios. Plus récemment, Zhao et al. (2007b) ont comparé le modèle mixte de Yu et al. (2006) avec d'autres modèles mixtes, notamment un modèle dans lequel la matrice Q est remplacée par une matrice P provenant d'une ACP. Ils montrent un léger gain de puissance de détection, pour des échantillons de petites tailles avec une faible densité de marqueurs. Cependant, si on en croit Zhang et al. (2008), les méthodes utilisant des techniques provenant de l'ACP semblent être plus robustes selon divers scénarios.

Le deuxième facteur est la disponibilité des logiciels permettant ces analyses et le temps de calcul de chacune d'elles. En effet, à ce jour, peu de logiciels bien carrossés sont disponibles pour réaliser facilement une détection de QTLs dans un modèle mixte et le fait d'estimer la variance polygénique pour tous les SNPs semblent être difficile quand le nombre de données (individus et marqueurs) est élevé. Dans ce contexte, certains auteurs ont développé des logiciels qui réalisent la détection de QTL via des approximations du modèle mixte complet. C'est le cas d'Aulchenko et al. (2007a) qui ont développé l'analyse GRAMMAR, qui fait une analyse d'association classique SNP par SNP sur les résidus calculés d'un modèle mixte polygénique, et dont la matrice de parenté est calculée à l'aide du pedigree. Cette analyse en deux étapes est beaucoup plus rapide qu'une analyse par un modèle mixte complet puisqu'on estime la variance polygénique une seule fois. Cependant, et on le verra dans le chapitre suivant, cette analyse est conservative et souffre d'une légère perte de puissance. Dans la même idée, Amin et al. (2007) proposent une extension de GRAMMAR qui permet de calculer la matrice de parenté via la génomique et utilisent en plus un contrôle génomique sur les tests. Le package R *GenABEL* développé par Aulchenko et al. (2007b) permet de réaliser ces analyses. Enfin, toujours dans la même idée, le logiciel EMMAX (Kang et al., 2010) permet de réaliser le modèle mixte complet mais en fixant la variance polygénique. Cette méthode semble robuste et puissante. Pour conclure, il ne faut probablement pas rejeter l'utilisation du modèle mixte complet dont le temps de calcul, certes plus long que celui des approximations GRAMMAR ou EMMAX, n'est pas une limite infranchissable.

1.3.3 Analyses uni-QTL par haplotype

Les analyses SNP par SNP sont assez problématiques car elles ne prennent pas en compte l'information provenant du LD entre les marqueurs. Cela a peu de conséquences lorsque les SNP sont éloignés les uns des autres puisqu'il y a peu de LD entre eux, ou lorsque la séquence complète est disponible car le variant causal sera alors typé. Cependant, dans la plupart des études, les densités de SNP sont entre ces deux extrêmes. Pour prendre en compte cette dépendance entre SNP successifs, on utilise des approches haplotypiques. Les analyses basées sur les haplotypes sont à priori plus cohérentes au sens biologique puisqu'elles permettent de mieux suivre la transmission d'un QTL situé entre plusieurs marqueurs. La connaissance des blocs haplotypiques (régions en fort LD avec peu de recombinaisons), discuté par Clark (2004), pourrait aussi être valorisée. Un haplotype permet également de mieux capturer l'information provenant du LD qu'un simple marqueur. Comme la puissance de détection d'une analyse d'association est fonction du LD entre marqueurs et QTL, les analyses haplotypiques devraient avoir une meilleure puissance que les analyses SNP par SNP (Akey et al., 2001). Nous verrons dans cette section que ces stratégies par haplotypes ont cependant des limites.

Analyses par haplotypes basiques

Pour les caractères binaires, comme en cas-témoins, la méthode la plus usuelle est l'utilisation d'une table de contingence $2 \times H$, avec H le nombre d'haplotypes possibles (égal au plus à 2^L où L est le nombre de locus dans l'haplotype), pour réaliser un test du χ^2 classique à $H - 1$ df. Ce test fait l'hypothèse que les paires d'haplotypes sont dans des proportions d'équilibre d'Hardy-Weinberg, ce qui implique que les génotypes doivent être en HWE à chaque locus et c'est rarement le cas pour des SNP proches à cause principalement du LD entre les locus (Schaid, 2004b).

Un autre moyen de tester l'effet d'un haplotype sur un caractère binaire est d'utiliser la régression logistique. Comme dans le cas de l'analyse SNP par SNP, différents modèles peuvent être utilisés. Soit par exemple un modèle à 2 locus $A = \{a, A\}$ et $B = \{b, B\}$ (la généralisation de ce modèle est ensuite triviale), il y a alors $2^L = 2^2 = 4$ haplotypes différents $h_i = \{ab, aB, Ab, AB\}$. En raisonnant en terme de génotype et en prenant comme référence l'haplotype ab , il y a alors 3 effets principaux à estimer correspondant aux homozygotes h_i/h_i et 6 interactions correspondant aux hétérozygotes h_i/h_j pour $i \neq j$. On écrit alors les modèles suivants :

$$\eta_i = \text{logit}(f_i) = \begin{cases} \beta_0 + \sum_{homo_j}^3 \beta_j x_{i,jj} + \sum_{hete_j}^6 \gamma_j x_{i,jk}, & \text{(Modèle génotypique)} \\ \beta_0 + \sum_{j=1}^4 \beta_j x_{i,j}, & \text{(Modèle haplotypique)} \\ \beta_0, & \text{(Sous } H_0) \end{cases}$$

Avec $f_i = P(\text{cas}|G_i)$, $x_{i,jj}$ l'indicatrice du génotype h_j/h_j de l'individu i , $x_{i,jk}$ l'indicatrice du génotype h_j/h_k , $x_{i,j}$ l'indicatrice de l'haplotype j de l'individu i et β_j, γ_j les effets des homozygotes et des hétérozygotes. Pour chacun de ces modèles, on utilise un test du rapport de vraisemblance (la vraisemblance suit une loi binomiale). Pour plus de détails sur ces modèles, il existe une excellente revue faites par Schaid (2004a). Notons qu'un test du score, qui approxime le rapport de vraisemblance d'un modèle haplotypique, est disponible (Schaid et al., 2002).

Pour les caractères quantitatifs, Zaykin et al. (2002) ont proposé d'utiliser une 2N-ANOVA similaire au modèle allélique pour un caractère binaire et donc au test de Pearson à 1 df (Sasieni, 1997). Tout individu ayant 2 haplotypes, le test du facteur haplotype dans une ANOVA exige de doubler les performances (donc la taille de la population analysée). Notons $y_{ij(d)}$ la $i^{\text{ème}}$ répétition du niveau de l'haplotype j , avec $j = 1 \dots J$ ($J \leq 2^L$ avec L le nombre de locus pris dans l'haplotype), $i = 1 \dots n_j$, $\sum_{j=1}^J n_j = N$ et tout ceci répété deux fois ($d = 1, 2$). La taille du vecteur phénotypique Y est $2N$. Le modèle d'ANOVA est :

$$\begin{aligned} y_{ij(d)} &= \mu_j + e_{ij}, & \text{avec } e_{ij} &\sim N(0, \sigma^2) \\ &= \mu + h_j + e_{ij}, & \text{avec } \sum_j h_j &= 0 \quad \text{(Décomposition centrée)} \end{aligned}$$

Où h_j est l'effet de l'haplotype j et e_{ij} le résidu aléatoire centré, les résidus étant supposés iid. Le test global qui découle d'un tel modèle, sous l'hypothèse nulle $H_0 = \{\forall j, h_j = 0\}$, correspond à un test de Fisher $F_{J-1, 2N-J}$. Les tests spécifiques qui correspondent aux hypothèses $H_0 = \{h_j = 0\}$, correspondent à des tests du Student T_{2N-J} (Attention : ils dépendent de la décomposition choisie, ici décomposition centrée). Cette ANOVA est analogue au test de Pearson à $H - 1$ df décrit plus haut.

Zaykin et al. (2002) propose également un test plus puissant basé la régression multiple dont

le modèle correspondant est :

$$\forall i, \quad y_i = \beta_0 + \sum_{h=1}^H \beta_h x_{ih} + e_i, \quad \text{avec } H = 2^L, \quad e_i \sim N(0, \sigma^2), \quad \{e_i\}_{i=1\dots n} \text{ iid}$$

Avec β_0 une constante, β_h l'effet de l'haplotype h , e_i les résidus et :

$$x_{ih} = \begin{cases} 1, & \text{Si l'individu } i \text{ est homozygote pour l'haplotype } h \\ \frac{1}{2}, & \text{Si l'individu } i \text{ est hétérozygote et inclus l'haplotype } h \\ 0, & \text{sinon} \end{cases}$$

Le test global associé à ce modèle, sous l'hypothèse nulle $H_0 = \{y_i = \beta_0 + e_i\}$, correspond à un test de Fisher $F_{H, N-H-1}$. Les tests spécifiques, sous $H_0 = \{\beta_h = 0\}$, correspondent quant à eux à des tests du Student T_{N-H-1} .

Problèmes avec les analyses basiques

Il existe certains problèmes avec les méthodes décrites ci-dessus :

- Que faire des haplotypes rares ? Cette question est directement reliée à la question du choix du nombre de marqueurs dans un haplotype. Plus il y a de marqueurs dans l'haplotype, plus on capture le fait que deux haplotypes pris au hasard dans une population dérivent d'un ancêtre commun (Identique par descendance ou IBD). Inversement, si le nombre de marqueurs est faible, on capte mal les probabilités d'être IBD (car possibilité d'être identique par état ou IBS). Malheureusement, plus grand est l'haplotype, moins celui-ci aura d'observations dans la population. Intégrer ces haplotypes rares dans l'analyse n'apporte que très peu d'information au prix de degrés de liberté et donc réduit la puissance de détection. Le choix du nombre de marqueurs dans l'haplotype, étudié dans des simulations par différents auteurs comme Grapes et al. (2006) ou Zhao et al. (2007a), est un compromis entre ces deux extrêmes. Leurs résultats ont montré que le nombre de marqueurs pris en compte dans un haplotype doit être compris entre 4 et 6 marqueurs.
- Comment prendre en compte les haplotypes similaires ? (i.e. des haplotypes IBS sur la quasi-totalité des marqueurs sauf un ou deux) Alors que deux haplotypes similaires semblent apporter la même information, les analyses basiques décrites ci-dessus ne prennent pas en compte cette similarité et estiment leurs effets séparément.

De multiples travaux se sont développés pour répondre à ces questions. Souvent basés sur des méthodes de classification, ils visent à créer une structure de covariance entre haplotypes pour diminuer le nombre d'haplotypes et ainsi augmenter leurs fréquences.

Méthodologies avancées

L'affirmation selon laquelle "2 haplotypes similaires semblent apporter la même information" reflète l'idée que la probabilité que les 2 copies d'un gène, affectant le caractère d'étude et localisé à proximité des marqueurs portant ces haplotypes, soient IBD (dérivent d'un ancêtre commun) sachant cette information haplotypique est d'autant plus grande que cette similarité est prononcée. Le cas d'école généralement avancé est celui d'une mutation du gène quantitatif, créant le QTL, apparu T générations avant l'observation, dans un haplotype particulier. T générations plus tard, les recombinaisons auront partiellement cassé cet haplotype, mais il restera plus de ressemblance (de

similarité haplotypique) entre chromosomes porteurs de l'allèle muté qu'entre deux chromosomes pris au hasard. Regrouper les haplotypes similaires prend donc un sens et permet d'augmenter le nombre d'informations de chacun d'eux et d'obtenir des estimations d'effets plus précises. La première étape pour prendre en compte cette similarité est de quantifier la ressemblance. Pour cela, divers auteurs ont proposé des indices de similarité heuristiques ou des méthodes d'estimation des probabilités de segments IBD calculées conditionnellement au haplotypes (Meuwissen and Goddard, 2001; Durrant et al., 2004; Li and Jiang, 2005). Ces derniers modèles sont plus complexes puisqu'ils font intervenir le concept d'ancêtre commun et sont généralement basés sur la théorie de la coalescence. Ces indices de similarité (ou ces probabilités d'IBD) définis, ils sont alors utilisés pour la détection de QTL selon différentes approches :

- *classification hiérarchique* : un arbre liant les haplotypes est construit en utilisant une distance dérivée de la mesure de similarité choisie. L'analyse LD est ensuite réalisée sur les classes d'haplotypes obtenus. Ce type d'approche est utilisée notamment par Li et al. (2006) et Waldron et al. (2006). La distance la plus couramment utilisée est $d(i, j) = 1 - \phi_{ij}$ où ϕ_{ij} représente la similarité entre l'haplotype i et j , mais on pourrait prendre d'autres transformations comme $d(i, j) = \log\left(\frac{1.01}{0.01 + \phi_{ij}}\right)$
- *Modèle mixte* : une deuxième approche consiste à utiliser ces indices ou probabilités pour construire une matrice de variance-covariance entre haplotypes. Cette matrice est nécessaire à un modèle mixte dans lequel les haplotypes sont des effets aléatoires. De ce fait, on utilise les vertus du modèle mixte qui permet de forcer les effets à suivre une loi normale et donc rétrécir (dans le sens "rendre moins importante") les estimations des effets des haplotypes avec un petit nombre d'observations. Le modèle s'écrit de la manière suivante :

$$Y = \mu + Zh + e$$

avec $h \sim N(0, H\sigma_h^2)$ et H représente la matrice de variance-covariance définie à partir des probabilités de ressemblance entre deux haplotypes. Ce modèle est notamment proposé par Meuwissen and Goddard (2000).

- *Modèle mixte et classification* : l'utilisation du modèle mixte ne suffit pas toujours pour autant. En effet, pour résoudre les équations du modèle mixte il faut que H soit inversible, et ce n'est pas toujours le cas. De plus, même si le modèle mixte permet d'obtenir l'estimation des effets des haplotypes avec un petit nombre d'observations, les estimations des variances avec les logiciels disponibles sont parfois très mauvaises. C'est par exemple le cas lorsqu'on travaille sur des données binaires avec peu de données et que certains haplotypes sont faiblement répandus. Il est souvent utile de regrouper les haplotypes dans la matrice H pour la rendre inversible et pouvoir estimer la variance correctement (Druet et al., 2008; Blott et al., 2003).
- *Autres approches* : Notons enfin que d'autres approches qui prennent en compte la ressemblance entre haplotypes existent. Zöllner and Pritchard (2005) proposent d'utiliser un modèle bayésien qui tient compte de la relation entre les haplotypes via la théorie de la coalescence. Mailund et al. (2006) proposent quant à eux une approche originale par phylogénie qui est proche de l'utilisation d'arbres de classification.

1.3.4 Analyses multi-QTL

Certains auteurs ont pensé utiliser les méthodes provenant de la sélection génomique afin de rechercher les QTLs (Onteru et al., 2011; Calus et al., 2011). L'idée est attrayante puisqu'elle permet de tester tous les SNPs simultanément et, de ce fait, de ne pas avoir à corriger pour la multiplicité des tests. Ces approches sont par définition multi-QTLs et se modélisent en considérant les effets des SNPs comme aléatoires, ce choix étant dicté pour deux raisons : mettre tous les SNPs en effets fixes demanderait un temps de calcul trop important ; elle valorise des *a priori* sur la distribution des effets des SNPs et permet d'intégrer facilement une corrélation entre eux. L'ensemble des méthodes existantes dans ce domaine se distinguent par la modélisation *a priori* des effets SNPs. Certains auteurs préfèrent une distribution gaussienne (type GBLUP ou RR-BLUP, VanRaden (2008)) quand d'autres utilisent plutôt une distribution exponentielle (type lasso, Park and Casella (2008); Tibshirani (1996)), une student (Bayes A, Meuwissen et al. (2001)) ou un mélange de distributions (Bayes B, Meuwissen et al. (2001); Bayes C π , Kizilkaya et al. (2010)). Prenons l'exemple du GBLUP, le modèle s'écrit :

$$y = \mu + Xa + e$$

Avec X la matrice d'incidence des génotypes (codés 0,1,2) de taille $n \times p$ (n étant le nombre d'individus et p le nombre de SNP), a le vecteur des effets SNPs supposés aléatoires avec $a \sim N(0, D\sigma_a^2)$ et e le vecteur des résidus avec $e \sim N(0, \sigma_e^2)$. En écrivant les équations classiques du BLUP (Henderson, 1973), on obtient les effets de tous les SNPs via :

$$\hat{a} = (X'X + D^{-1}\lambda)^{-1}X'(y - \hat{\mu}), \quad \text{avec } \lambda = \frac{\sigma_e^2}{\sigma_a^2}$$

Ces méthodes sont prometteuses et commencent à être utilisées pour la détection de QTLs. Ainsi, lors des récents congrès QTLMAS¹ XIII et XIV, certaines de ces méthodes ont été testées dans le cadre de la détection de QTL.

Les résultats du QTLMAS XIII, proposés par Maliepaard et al. (2010), montrent un gain de puissance et une meilleure localisation des QTL des analyses Bayes C π par rapport aux méthodes de type GRAMMAR et aux analyses LA (classiques et de type IBD comme présenté dans la section suivante). Un résultat étonnant au vu des simulations, qui comportaient des données de familles de pleins frères et une carte de marqueur à moyenne densité (distance moyenne de 1cM entre les deux marqueurs adjacents), et qui semblaient être plus propice aux analyses LA que LDA pour deux raisons principales : la distance entre les SNP de 1cM va difficilement permettre de trouver un marqueur en LD avec un QTL proche ; la forte structure familiale n'est à priori pas la meilleure chose pour les méthodes LDA et nécessite une correction dans le modèle. Le succès des méthodes Bayes C π laisse donc supposer que l'utilisation simultanée de l'ensemble des marqueurs permet de prendre en compte une partie de l'apparentement entre les individus, mais également capture une partie de l'information du LD entre les marqueurs et permet ainsi un gain de puissance et une localisation plus précise du QTL.

Les résultats du QTLMAS XIV, proposés par Mucha et al. (2011), vont dans le même sens et montrent un gain de puissance des analyses Bayes C π par rapport à d'autres méthodes multi-QTL

1. QTLMAS est une réunion annuelle de généticiens européens travaillant sur les animaux de ferme et les plantes, au cours de laquelle des méthodes sont comparées sur des données simulées

qui utilisent l'ensemble des marqueurs (non présentées dans ce rapport) et à l'analyse GRAMMAR. Les simulations comprenaient des familles de demi-frères de pères sur 4 générations et la distance entre marqueurs adjacents étaient de 0.05Mb (qui représente environ la distance moyenne existante avec les puce de 50k SNP). Pour les deux caractères simulés, un binaire et un quantitatif, les analyses Bayes C π ont détecté le plus de QTL tout en étant celles ayant donné le moins de faux positifs (une seule méthode a donné moins de faux positifs mais a obtenu des puissances bien inférieures). On peut cependant regretter le manque de comparaison avec des analyses de type modèle mixte LDA et LDLA, uni-SNP et haplotypiques, qui en pratique sont souvent utilisées pour ce type de données et ont fait leurs preuves.

Notons que les méthodes multi-QTL ne sont pas propres aux analyses LDA et à l'utilisation de génotypes. Meuwissen and Goddard (2004) proposent en effet un modèle multi-QTL LDLA et haplotypique. S'inspirant de leurs travaux, Calus et al. (2008) comparent 4 modèles haplotypiques par simulations, dont 2 modèles LD où les probabilités entre haplotypes sont calculées à partir de leurs statuts IBS et 2 modèles LDLA de type Meuwissen and Goddard (2004) où les probabilités IBD entre haplotypes sont calculées à partir du modèle de Meuwissen and Goddard (2001) (voir section suivante pour plus de détails sur ces probabilités). Le but de leur article n'étant pas de chercher des QTLs mais de bien prédire les valeurs génétiques des animaux, les conclusions qu'ils en tirent ne doivent pas être interprétées pour la détection de QTL. Néanmoins, elles nous offrent un critère de comparaison entre ces modèles et ont donc un réel intérêt. Ils concluent tout d'abord qu'apporter l'information du LA dans le modèle augmente sensiblement la précision (corrélation entre valeurs prédites et réelles) pour des fortes héritabilités (0.5) et est équivalente pour des faibles héritabilités (0.1). De plus, ils trouvent qu'augmenter le nombre de marqueurs dans l'haplotype de 2 à 10 apporte également un gain. Par contre, lorsque la densité de marqueurs augmente, l'ensemble des méthodes donnent sensiblement les mêmes résultats. Ils concluent que la méthode haplotypique LDLA est la meilleure solution ("la plus sûre") mais qu'elle nécessite un nombre très important de paramètres à estimer (qui est d'ailleurs fonction de la taille de l'haplotype choisit).

1.3.5 Conclusion : méthodes LDA

Notre inventaire des méthodes basées sur l'utilisation du déséquilibre de liaison pour la détection de QTLs a essentiellement exploré le problème de la stratification de la population, celui de l'information génétique exploitée (SNP ou haplotypes) et celui d'une analyse uni-QTL ou multi-QTL .

Le premier problème, discuté très extensivement par Price et al. (2010), conduit au choix de modéliser les effets de la stratification comme des effets fixes ou aléatoires. La modélisation d'effets découverts par des explorations préalables des données par ACP est très efficace dans les cas de mélanges de populations ou d'ancêtres communs mais pas adaptée aux cas de structure familiale (Patterson et al., 2006). A contrario, la modélisation de ces effets comme aléatoires permet de bien prendre en compte la structure familiale et la structure plus ancestrale. Combiner les deux informations dans un même modèle est une bonne approche pour corriger tout type de stratification de populations.

Le deuxième point concerne le choix de l'utilisation de SNP ou d'haplotypes dans l'analyse. Il n'y a pas réellement de consensus sur cette question. Les haplotypes présentent des avantages d'interprétations, sont plus proches de la réalité biologique (au sens de la transmission des parents de segments chromosomiques aux descendants) en permettant un meilleur suivi du QTL et une

meilleure modélisation du déséquilibre de liaison. Cependant, l'utilisation d'haplotypes nécessite la reconstruction des phases et aboutit généralement aux difficultés d'estimation des effets peu fréquents dans la population étudiée. De ce fait, on a souvent recours à des techniques de clusterisation et à l'utilisation du modèle mixte.

Le dernier point concerne le choix d'une analyse uni-QTL ou multi-QTL. Il semble que les analyses multi-QTL soient très prometteuses. Elles permettent de tester l'ensemble des points du génome en même temps et semblent bien capturer l'information du LD mais également corriger certains problèmes liés à la structure de population. Des effets polygéniques peuvent être rajoutés aux modèles pour être sûr de bien corriger ces derniers. Néanmoins, l'utilisation de ces méthodes perd de son intérêt lorsque les cartes sont denses. Des résultats récents (non publiés encore) ont montré sur plusieurs jeux de données réelles (chez les ovins, équins, bovins) que les corrélations entre les tests à chaque point de Bayes C π et un modèle mixte polygénique uni-QTL étaient de l'ordre de 0.9.

Pour conclure, retenons de cette section que l'utilisation du modèle mixte semble être une solution de choix pour pallier à ces divers problèmes, et fait l'objet d'un consensus croissant dans les divers mondes de la génétique (animale, végétale et plus récemment humaine).

1.4 Analyses combinant association et liaison (LDLA)

Pendant longtemps, l'analyse de liaison (LA) a été utilisée en génétique animale et humaine à l'aide de cartes de marqueurs à faible densité dans des protocoles familiaux (souvent familles de demi-frères en génétique animale et familles de trios en génétique humaine). Plus récemment, les puces SNPs à haut débit (entre 50k et 1000k) ont permis d'exploiter l'information apportée par le déséquilibre de liaison. Ces études d'associations sont plus puissantes (Risch and Merikangas, 1996) et donnent des régions d'intérêts plus précises et mieux localisées que les études LA. Elles ont donc renouvelé l'intérêt d'analyser la variabilité génétique des caractères quantitatifs et des maladies complexes même si à ce jour, peu de résultats après réplication ont été significatifs.

En principe, les analyses LA doivent être réalisées à partir de protocoles familiaux (trios en génétique humaine et familles plus grandes en génétique animale) et les analyses LD à partir d'échantillons d'individus non apparentés (possible en génétique humaine, beaucoup moins dans les populations animales). Mais avec les puces SNPs, les données familiales peuvent maintenant être analysées avec à la fois un côté LA qui assure une robustesse face à la structure de population et une partie LD pour garder la puissance de l'analyse. On parle alors d'études LDLA.

Dans cette section, seront traités successivement les études LDLA provenant de la génétique humaine, qui sont adaptées aux familles de petite taille (trios) et avec un apparentement entre fondateurs quasi nul, puis les études LDLA provenant de la génétique animale, adaptées aux des familles de grande taille avec un fort apparentement entre les fondateurs.

1.4.1 Tests d'association dans des familles nucléaires

Dans cette sous-section consacrée aux méthodes LDLA provenant de la génétique humaine, nous partirons du test du déséquilibre de transmission (*TDT : transmission disequilibrium test*) avec ses différentes extensions en allant vers une généralisation de ce type d'approche connues sous le nom d'études d'associations basées sur des familles (*Family-based association studies*).

Le "transmission disequilibrium test" (TDT)

Le TDT, proposé par Spielman et al. (1993), est le test le plus simple pour tester l'association dans un protocole familial à partir du génotype de trios constitués d'un descendant malade et de ses parents. L'idée du TDT est que, sous l'hypothèse nulle (pas de liaison entre le marqueur et le gène recherché), les lois mendéliennes déterminent la loi de transmission à l'enfant affecté des allèles aux marqueurs. Le TDT compare le nombre d'allèles au marqueur observé qui ont été transmis à son espérance sous H_0 . Ainsi, l'excès d'un même type d'allèle parmi les malades indique qu'un locus de susceptibilité pour la maladie étudiée est lié et associé avec ce locus.

Détails du modèle TDT Soient n trios avec n descendants malades et $2n$ parents. Supposons que le locus A testé, bi-allélique, possède les allèles $A1$ et $A2$. Le test du TDT est basé sur le tableau des fréquences entre les allèles transmis et non-transmis des parents aux descendants (Table 1.7). Les valeurs a , b , c et d du tableau s'interprètent comme le nombre de parents ayant le génotype $A1/A1$ (resp. $A1/A2$, $A2/A1$ et $A2/A2$) qui ont transmis l'allèle $A1$ (resp. $A1$, $A2$ et $A2$).

TABLE 1.7 – Tableau des effectifs des allèles transmis et non transmis des $2n$ parents aux n descendants

Allèles transmis	Allèles non transmis		Total
	$A1$	$A2$	
$A1$	a	b	$a + b$
$A2$	c	d	$c + d$
	$a + c$	$b + d$	$2n$

Seuls les parents hétérozygotes étant informatifs, on ne s'intéresse qu'au case b et c . A partir de là, on construit la statistique de test T qui compare la proportion des descendants recevant $A1$ à celle des descendants recevant $A2$:

$$T = \frac{(b - c)^2}{b + c}$$

Sous l'hypothèse nulle, T suit approximativement une distribution du khi-deux à un degré de liberté.

Quelques commentaires sur le TDT A l'origine, le TDT était utilisé pour tester la liaison en présence d'association. Cependant, comme la liaison et de l'association entre un locus et la maladie sont toutes deux nécessaires pour rejeter l'hypothèse nulle, le TDT est plutôt utilisé comme un test d'association (Laird and Lange, 2006). Cette nécessité est primordiale et prévient les faux positifs dues à une association sans liaison, situation se présentant lorsque la population possède une structure familiale ou est un mélange de populations. Le TDT est donc une méthode robuste à la stratification de la population et aux problèmes liés à une mauvaise classification phénotypique. Cependant, il est inutilisable lorsqu'on souhaite réaliser une analyse avec des parents manquants, un pedigree général, des phénotypes complexes ou des haplotypes. De très nombreuses extensions se sont développées (Schulze and McMahon, 2002) et ont fini par se généraliser sous la dénomination FBAT, pour "Family-based association tests" (Laird et al., 2000).

Généralisation du TDT : L'approche FBAT

L'idée des méthodes FBAT est d'étendre le TDT à un cadre plus général tout en gardant ses bonnes propriétés de robustesse. Le modèle FBAT le plus simple étend le TDT classique à la prise en compte de plusieurs descendants par famille et la possibilité d'avoir des descendants sains. Les tests FBAT se font non plus dans une famille de trio mais dans des familles nucléaires pouvant comprendre plusieurs descendants.

Description du modèle le plus simple Considérons n familles nucléaires indépendantes et notons n_i le nombre de descendants dans la famille i . On souhaite tester l'hypothèse H_0 : Il n'y a ni liaison ni association entre le marqueur et la maladie. Le calcul de la statistique de test se décompose en plusieurs étapes. On définit d'abord une statistique de test par famille à l'aide de la formule suivante :

$$S_i = \sum_{j=1}^{n_i} X_{ij} T_{ij}$$

Où X_{ij} est un certain codage du génotype de l'individu j dans la famille i et T_{ij} un certain codage du phénotype Y_{ij} de l'individu j dans la famille i . Mentionnons que X_{ij} peut être codé comme un scalaire en comptant par exemple le nombre d'allèle 1 au marqueur, ou comme un vecteur donnant à S_i une dimension supérieure à 1. Les valeurs des T_{ij} peuvent être binaires ou continues et peuvent être la représentation des Y_{ij} précorrégés par des covariables (ex : $T_{ij} = Y_{ij} - \mu$).

On définit ensuite une statistique U par :

$$U = \sum_i S_i - E(S_i)$$

Avec

$$\begin{aligned} E(S_i) &= \sum_j T_{ij} E(X_{ij}) \\ &= \sum_j T_{ij} \sum_{g \in G} X(g) P(g) \end{aligned}$$

Où G correspond à l'ensemble des génotypes possibles des enfants dans la famille i et $P(g)$ est la probabilité que l'enfant ait le génotype g dans la famille i .

Enfin, le statistique de test Z de la méthode FBAT est alors définie par :

$$\begin{aligned} Z &= \left(\frac{U}{\sqrt{V}} \right)^2 && \text{si } X \text{ est un scalaire} \\ &= U^T V^{-1} U && \text{si } X \text{ est un vecteur} \end{aligned}$$

Avec $V = Var(U)$ et Z suit un khi-deux à 1 ddl dans le cas où X est un scalaire et suit un khi-deux à ν ddl dans le cas où X est un vecteur, ν étant égal au rang de V .

Remarques et extensions possibles On peut noter que le TDT est un cas particulier des tests FBAT : les deux parents sont génotypés, le trait est binaire ($T = 0$: sain, $T = 1$: malade), X compte le nombre d'un allèle au marqueur, et un seul enfant par famille (malade). Dans ces conditions, les deux tests sont strictement équivalents. On peut également noter qu'il est possible

de tester d'autres hypothèses nulles comme par exemple $H_0 : \text{il y a liaison mais pas association}$. Dans ces cas, le test ne sera pas identique car la variance du test sera différente. De plus, en changeant la manière dont T est défini, on peut inclure des individus sains ou mettre plusieurs caractères. Changer la manière dont X est défini permet de tester d'autres modèles génétiques (récessivité ou dominance) et incorporer plusieurs marqueurs et plusieurs allèles à un marqueur. D'autres extensions du modèle classique FBAT existent et permettent d'introduire un pedigree ou des génotypes manquants.

Extensions du TDT par modèles de régression

Les méthodes basées sur des scores telles que les méthodes FBAT ne sont pas les seules extensions possibles du TDT. En effet, d'autres extensions ont été réalisées à partir des modèles de régression, notamment pour traiter les caractères quantitatifs. Cette théorie est partie de Allison (1997) qui ont proposé un test pour des trios basé sur la comparaison de deux modèles de régression. Plus tard, Fulker et al. (1999) et Abecasis et al. (2000), ont généralisé ces approches par modèle linéaire, qui permettent de tester association et liaison, et ont rendu possible de séparer les effets de liaison et d'association dans le modèle. Ce dernier est nommé *QTDT* pour "quantitative TDT".

Présentation du QTDT Le modèle "orthogonal" de Abecasis et al. (2000) décompose l'effet du marqueur en un effet intra (β_b) et inter (β_w) familles. Il s'écrit de la manière suivante :

$$Y_{ij} = \mu + (P_{ij} - \bar{P})\beta_b + (M_{ij} - P_{ij})\beta_w + e_{ij}, \quad i = 1..n, \quad j = 1..n_i$$

Avec Y_{ij} le phénotype de l'individu j dans la famille i , μ la moyenne générale, M_{ij} le génotype du descendant (codé -1,0,1), $P_{ij} = \frac{P_{ij}^1 + P_{ij}^2}{2}$ la valeur espérée du génotype du descendant conditionnellement aux génotypes de ses parents et e_{ij} la résiduelle. De plus, on a $N = \sum_i n_i$. Matriciellement, ce modèle peut s'écrire :

$$Y = X\theta + e$$

Avec $\theta = (\mu, \beta_b, \beta_w)'$ le vecteur des effets et $X = (X_\mu, X_{\beta_b}, X_{\beta_w})$ la matrice d'incidence correspondante. Ce modèle est dit orthogonal car X_{β_b} et X_{β_w} sont asymptotiquement orthogonaux quand n_i tend vers l'infini (Boitard et al., 2010). Sous l'hypothèse nulle $H_0 : \text{Il n'y a pas d'association entre le marqueur et le caractère}$, qui se modélise par $H_0 : \beta_w = 0$, un test de student classique peut être réalisé :

$$T = \frac{\hat{\beta}_w}{\sqrt{\hat{\sigma}^2 e_3 (X'X)^{-1} e_3'}}$$

Avec $e_3 = (0, 0, 1)$ et $\hat{\beta}_w$ et $\hat{\sigma}^2$ les estimateurs des moindres carrés de β_w et σ^2 . Ce test suit asymptotiquement une distribution $N(0, 1)$. Une remarque importante faite par Abecasis et al. (2000) est que seule la partie "within" est robuste à la structure de population et au mélange de populations.

Conclusion sur les méthodes FBAT-QTDT

Les approches type FBAT et les approches type QTDT permettent de traiter les données dans des familles nucléaires. L'avantage principal de ces méthodes est qu'elles sont robustes, notamment à la structure de population et au mélange de populations. Néanmoins, ces approches FBAT

regroupent un nombre impressionnant de méthodes qui sont souvent spécifiques d'un problème particulier : génotypes manquants, pedigree, caractère quantitatif, multi-caractères, multi-marqueurs etc... Et, même si certaines d'entre elles sont adaptées à plusieurs de ces problèmes, aucune ne regroupe toutes ces bonnes caractéristiques. D'excellentes revues bibliographiques ont été réalisées et proposent un bon aperçu de ces approches (Laird and Lange, 2006, 2008).

Pour des caractères quantitatifs, les méthodes FBAT et QTDT sont assez similaires (Lange et al., 2002). Leurs différences (légères) viennent des hypothèses des modèles. Les méthodes FBAT utilisent des tests du score basés sur les allèles transmis aux descendants, conditionnellement à leurs phénotypes et aux génotypes parentaux. De plus, la distribution des tests statistiques sous l'hypothèse nulle est directement calculée à partir des lois de Mendel. Donc, la distribution est correcte tant que ces lois le sont, et ceci indépendamment de la distribution des phénotypes chez les descendants (Ewens et al., 2008). Les modèles de régression de type QTDT sont généralement plus puissants mais font des hypothèses sur la normalité des phénotypes rendant plus difficiles ses extensions possibles (haplotypes, multi-caractères etc...).

La bonne robustesse de ces méthodes est elle payée par une moindre puissance par rapport à d'autres méthodes d'associations? Boitard et al. (2010) et Lange et al. (2002) ont étudié d'un point de vue analytique la puissance des méthodes FBAT et QTDT sans toutefois la comparer à celles d'autres méthodes LD ou LDLA. De telles comparaisons ont été réalisées par simulations et ont souvent montré que les méthodes FBAT/QTDT étaient très robustes mais avaient une puissance plus faible que les autres (modèle mixte polygénique notamment). Par exemple, Aulchenko et al. (2007a) trouvent dans leur étude que le QTDT peut être jusqu'à deux fois moins puissant que GRAMMAR. Cette perte de puissance vient du fait que le modèle QTDT sépare l'effet du SNP en deux effets intra et inter familles, seul l'effet intra famille étant testé. Le chapitre suivant est consacré à ces études de puissance.

1.4.2 La théorie de Meuwissen et Goddard

Introduction

Les méthodes LDLA précédemment décrites sont robustes. Elles ont été développées en génétique humaine, dans un cadre où les données proviennent de familles de petites tailles et en général sous forme de trio (père-mère-descendant). De plus, les caractères étudiés sont le plus souvent des maladies pour lesquelles l'effet d'un gène majeur est régulièrement suspecté. En génétique animale, les familles sont souvent des grandes familles de demi-frères de pères (chez les ruminants, le cheval), ou un mélange de grandes familles de père et de mère (chez les autres monogastriques), avec une structure de population forte et les caractères étudiés sont en général des caractères quantitatifs. L'hypothèse la plus fréquente est que ces caractères sont contrôlés par plusieurs gènes. Dans ce contexte, l'utilisation d'un modèle mixte avec un effet polygénique en aléatoire est très pratique et efficace (cf 1.3.2). Allant plus loin, Meuwissen and Goddard (2000) proposent un modèle mixte avec deux effets aléatoires, l'effet polygénique et l'effet haplotypique. La matrice de corrélations entre les haplotypes (proportionnelle à la matrice des probabilités que deux segments chromosomiques homologues soient IBD conditionnellement aux informations marqueurs les entourant, que nous qualifierons plus loin matrice IBD) étant calculée à partir d'un modèle de coalescence qui tient compte de l'histoire de la population (décrite à partir du déséquilibre de liaison) et de la ressemblance entre les haplotypes.

Alors que la méthode de Meuwissen and Goddard (2000) ne travaille que sur le LD, d'autres

auteurs, comme George et al. (2000), proposent un modèle mixte dans lequel la matrice IBD entre haplotypes est estimée compte tenu de l'information sur la transmission des haplotypes parentaux et donc est une forme d'analyse de liaison (LA). Le modèle s'écrit de la même manière, c'est à dire :

$$Y = \mu + Z_v v + Zu + e$$

Avec Y le vecteur des performances, μ la moyenne des phénotypes, Z la matrice d'incidence reliant les animaux aux phénotypes, Z_v la matrice d'incidence reliant les haplotypes aux phénotypes, u le vecteur des effets polygéniques avec $u \sim N(0, A\sigma_a^2)$, v le vecteur des effets gamétiques avec $v \sim N(0, H\sigma_v^2)$ et e le vecteur des résidus. La matrice A étant la matrice d'apparentement entre les individus calculée à l'aide du pedigree et H étant la matrice de variance-covariance IBD au QTL conditionnellement à l'information des marqueurs flanquants. Les probabilités au sein de cette matrice H peuvent alors être calculées via l'approche de Pong-Wong et al. (2001), qui généralise celle utilisée par Fernando and Grossman (1989).

Détaillons un peu cette méthode en considérant un exemple avec uniquement 2 marqueurs M et N flanquants un QTL bi-allélique d'allèles Q et q . Intéressons nous de plus uniquement au chromosome reçu du père, le raisonnement étant le même pour celui reçu de la mère. Les deux allèles des marqueurs M et N du père sont respectivement (M_p, M_m) et (N_p, N_m) , les allèles du QTL dont on suppose qu'il est localisé au milieu de l'intervalle séparant les deux marqueurs sont (Q_p, Q_m) . M_p signifiant allèle du marqueur M sur le chromosome grand-paternel et M_m celui du chromosome grand-maternel. Soit enfin M_x, Q_x et N_x les trois allèles du chromosome paternel du descendant (Figure 1.9). On cherche alors à connaître les probabilités :

$$\begin{cases} P(Q_x = Q_p | M_x N_x M_p M_m N_p N_m) \\ P(Q_x = Q_m | M_x N_x M_p M_m N_p N_m) \end{cases}$$

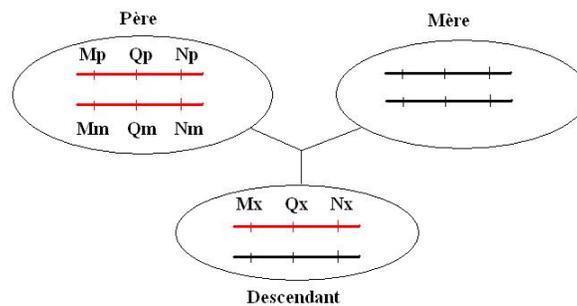


FIGURE 1.9 – Illustration des notations

Soit maintenant θ , θ_1 et θ_2 les taux de recombinaisons entre respectivement, M et N , M et Q , et, Q et N . Il est possible de calculer toutes les probabilités suivant le génotype du descendant (Table 1.8).

Décrivons le calcul par exemple pour $M_x = M_p$ et $N_x = N_p$. Le descendant a reçu Q_p sauf s'il y a

TABLE 1.8 – Calcul des probabilités via la méthode de Pong-Wong et al. (2001)

M_x	N_x	$P(Q_x = Q_p Marqueurs)$	$P(Q_x = Q_m Marqueurs)$
p	p	$(1 - \theta_1)(1 - \theta_2)/(1 - \theta)$	$\theta_1\theta_2/(1 - \theta)$
p	m	$(1 - \theta_1)\theta_2/\theta$	$\theta_1(1 - \theta_2)/\theta$
m	p	$\theta_1(1 - \theta_2)/\theta$	$(1 - \theta_1)\theta_2/\theta$
m	m	$\theta_1\theta_2/(1 - \theta)$	$(1 - \theta_1)(1 - \theta_2)/(1 - \theta)$
p	–	$1 - \theta_1$	θ_1
m	–	θ_1	$1 - \theta_1$
–	p	$1 - \theta_2$	θ_2
–	m	θ_2	$1 - \theta_2$
–	–	$1/2$	$1/2$

eu double recombinaison et donc :

$$P(Q_x = Q_p | M_x = M_p, N_x = N_p, M_m N_m) = \frac{(1 - \theta_1)(1 - \theta_2)}{(1 - \theta)}$$

$$P(Q_x = Q_m | M_x = M_p, N_x = N_p, M_m N_m) = \frac{\theta_1\theta_2}{(1 - \theta)}$$

De cette table découle la matrice IBD au QTL sachant les marqueurs flanquants. Cette table est donnée pour 2 marqueurs flanquants le QTL mais elle se construit de la même manière avec plus de marqueurs. D'autres méthodes de calcul de cette matrice IBD sont possibles, comme décrit dans la revue de la publication de George et al. (2000).

Modèle LDLA de Meuwissen et al. (2002)

Meuwissen et al. (2002) ont eu l'ingénieuse idée de combiner la méthode LD de Meuwissen and Goddard (2000) et la méthode LA de George et al. (2000) pour créer un modèle LDLA. Cette méthode a été développée pour des familles de demi-frères avec les génotypes paternels disponibles. Le modèle devient :

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

Avec Y le vecteur des performances, β le vecteur des effets fixes, X la matrice d'incidence des effets fixes, Z la matrice d'incidence reliant les animaux aux phénotypes, Z_h la matrice d'incidence reliant les haplotypes aux phénotypes, u le vecteur des effets polygéniques avec $u \sim N(0, A\sigma_a^2)$, h le vecteur des effets haplotypiques avec $h \sim N(0, H\sigma_h^2)$ et e le vecteur des résidus. La matrice A étant la matrice d'apparentement entre les individus calculée à l'aide du pedigree.

Toute l'astuce réside dans la construction de la matrice de variance-covariance entre haplotypes, ou matrice IBD, H . Cette matrice est calculée conditionnellement à l'information provenant du déséquilibre de liaison et de l'analyse de liaison. Pour cela, les auteurs distinguent les haplotypes paternels (PH) des haplotypes paternels (DPH) et maternels (DMH) hérités par le descendant. La matrice H est ensuite calculée par blocs. Un premier bloc est calculé via la méthode de Meuwissen and Goddard (2001) entre les PH et les DMH (dit haplotypes de bases ou fondateurs). Ensuite, en utilisant les règles de Pong-Wong et al. (2001), le deuxième bloc entre les DPH et les haplotypes de

bases est construit (Table 1.9). Si s est le nombre de pères génotypés et n le nombre de descendants, la matrice H est de taille $2(s+n) \times 2(s+n)$ et Z_h est de taille $n \times 2(s+n)$.

TABLE 1.9 – Forme de la matrice IBD en LDLA. a =bloc LD ; b =bloc LA

	PH	DMH	DPH
PH	a	a	b
DMH	a	a	b
DPH	b	b	b

A chaque position testée du génome, les composantes de la variance $\hat{\sigma}_u^2$, $\hat{\sigma}_h^2$ et $\hat{\sigma}_e^2$ sont estimées en maximisant la log-vraisemblance L :

$$L = -\frac{1}{2} \left[\ln(|V|) + \ln(|X'V^{-1}X|) + (Y - X\hat{\beta})'V^{-1}(Y - X\hat{\beta}) \right]$$

Avec $V = Var(Y) = [ZAZ'\sigma_u^2 + Z_hHZ_h'\sigma_h^2 + I\sigma_e^2]$ et $\hat{\beta}$ l'estimation des moindres carrés généralisés de β . En pratique, les estimations du maximum de vraisemblance restreint (REML) sont obtenues à l'aide de logiciels tels que AsREML (Gilmour et al., 2006) ou REMLF90 (Misztal et al., 2002). Les effets sont estimés sous l'hypothèse H_0 qu'il n'y a pas de QTL (i.e. sans la partie Z_hh) et sous l'hypothèse H_1 qu'il y a un QTL (i.e. le modèle complet). Le test du rapport de vraisemblance pour valider ou non la présence d'un QTL est établi :

$$\lambda = -2\ln\left(\frac{L(H_0)}{L(H_1)}\right)$$

Où $L(H_0)$ et $L(H_1)$ sont les vraisemblances maximales des observations quand leurs paramètres sont égaux à leurs estimations REML sous l'hypothèse H_0 et H_1 . La distribution du test n'est pas connue mais Self and Liang (1987) montrent qu'elle est proche de $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$.

Exemple des différences entre les matrices H provenant du LA, LD ou LDLA

Afin de bien comprendre les différences entre ces trois modèles LA, LD et LDLA, qui se différencient uniquement par la structure de la matrice H , prenons un exemple simple et détaillons le contenu de ces matrices. Soit une famille de demi-frères composée d'un père et de deux de ses descendants avec les haplotypes (sur 4 marqueurs) présentés dans la Table 1.10. Chacun des descendants ayant reçu un haplotype différent du père.

TABLE 1.10 – Exemple : Haplotypes d'une famille de 3 individus sur 4 marqueurs

	Haplo paternel	Haplo maternel
Père	1121	1222
Descendant 1	1121	1211
Descendant 2	1222	1212

On décompose tout d'abord les haplotypes paternels PH en PP et PM pour désigner les haplotypes paternels et maternels du père. De même, on décompose les DMH et DPH en $D(i)M$ et $D(i)P$ pour désigner les haplotypes maternels et paternels hérités par le descendant i .

Pour construire la matrice H d'un modèle pur LA, on suppose que les probabilités de transmission du descendant 1 ne sont pas connues avec une probabilité de 1 (mais 0.8) mais le sont pour le descendant 2. On obtient alors la matrice présentée dans la Table 1.11. La matrice H pour le seul LD est quant à elle de la forme décrite dans la Table 1.12 (c'est un exemple, les probabilités ne sont pas réelles ici)

TABLE 1.11 – Exemple : Matrice H LA

	PP	PM	D1M	D2M	D1P	D2P
PP	1					
PM	0	1				
D1M	0	0	1			
D2M	0	0	0	1		
D1P	0.8	0.2	0	0	1	
D2P	0	1	0	0	0.2	1

TABLE 1.12 – Exemple : Matrice H LD

	PP	PM	D1M	D2M	D1P	D2P
PP	1					
PM	0.3	1				
D1M	0.4	0.3	1			
D2M	0.2	0.6	0.1	1		
D1P	1	0.3	0.4	0.2	1	
D2P	0.3	1	0.3	0.6	0.3	1

Enfin, en prenant compte des Tables 1.11 et 1.12, on construit la matrice H LDLA (Tableau 1.13) de la façon suivante :

1. On construit la partie entre les haplotypes du père et les haplotypes des descendants hérités de la mère via l'information du DL (partie en bleu). Ces haplotypes forment les haplotypes de base.
2. On construit la partie entre les haplotypes de base et les haplotypes des descendants hérités du père via l'information provenant du LA et du LD (partie en rouge). Pour cela, on utilise la formule de Fernando and Grossman (1989) :

$$P(x, y) = rP_{IBD}(PP(x), y) + (1 - r)P_{IBD}(PM(x), y)$$

Avec x l'haplotype paternel du descendant, y un autre haplotype, r la probabilité que le descendant est hérité de l'haplotype paternel du père et $PP(x)$ (resp. $PM(x)$) l'haplotype paternel (resp. maternel) du père de l'individu qui a l'haplotype x . Par exemple, la probabilité entre D1P et PP se décompose de la manière suivante :

$$\begin{aligned} P(D1P, PP) &= rP_{IBD}(PP, PP) + (1 - r)P_{IBD}(PP, PM) \\ &= 0.8 \times 1 + 0.2 \times 0.3 \\ &= 0.86 \end{aligned}$$

TABLE 1.13 – Exemple : Matrice H LDLA

	PP	PM	D1M	D2M	D1P	D2P
PP	1					
PM	0.3	1				
D1M	0.4	0.3	1			
D2M	0.2	0.6	0.1	1		
D1P	0.86	0.44	0.38	0.28	1	
D2P	0.3	1	0.3	0.6	0.44	1

Ceci n'est qu'un simple exemple à partir d'hypothétiques probabilités LA et LD mais dont les probabilités LDLA sont bien calculées à partir de celles ci. On remarque que si les probabilités LA étaient optimales (égales à 0 ou 1), alors la matrice LDLA serait identique à matrice LD (dans le cas où la probabilité IBD (LD) entre un haplotype et lui même est bien égale à 1, ce qui n'est pas le cas dans la réalité). Combiner matrice LD et LA raffine les probabilités IBD entre haplotypes.

Extensions et modèles proches de celui de Meuwissen et al. (2002)

D'autres auteurs proposent des extensions ou des modèles proches de celui de Meuwissen et al. (2002). Par exemple, Druet et al. (2008) proposent d'utiliser exactement le même modèle mais de clusteriser la matrice H afin de la rendre inversible et d'augmenter les fréquences des haplotypes dans la population. Blott et al. (2003) proposent quant à eux un modèle proche dans lequel la partie LA est intégrée dans la matrice d'incidence des effets haplotypiques Z_h et la matrice H est construite sur les haplotypes de base. En reprenant l'exemple du paragraphe précédent, on aurait alors pour matrice H la forme décrite dans la Table 1.14 et pour produit $Z_h h$:

$$Z_h h = \begin{pmatrix} 0.8 & 0.2 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} PP \\ PM \\ D1M \\ D2M \end{pmatrix}$$

Blott et al. (2003) rajoutent une partie de clustering pour éviter les problèmes de non inversibilité de la matrice H et d'haplotypes en trop faible fréquence dans la population.

TABLE 1.14 – Exemple : Matrice H LD-Blott

	PP	PM	D1M	D2M
PP	1			
PM	0.3	1		
D1M	0.4	0.3	1	
D2M	0.2	0.6	0.1	1

1.4.3 Conclusion : méthodes LDLA

Dans cette section consacrée aux méthodes LDLA, nous avons vu deux types d'approches distinctes qui correspondent à des utilisations différentes. Une première approche concernait les

méthodes provenant de la génétique humaine, où les familles sont nucléaires et de petites tailles, les caractères traités sont le plus souvent des maladies avec des protocoles en cas-témoins. L'intérêt principal de ces méthodes est leur robustesse notamment à la structure de population et au mélange de populations. Un autre intérêt de ces approches est qu'elles permettent de séparer les effets intra et inter-familles et permettent ainsi de facilement tester plusieurs hypothèses nulles. Mais cette séparation des effets se fait au prix d'une moindre puissance. Elle fera l'objet d'une des parties du chapitre suivant.

Le deuxième type d'approches concernait les méthodes provenant de la génétique animale, où les familles sont de grandes tailles, généralement de demi-frères et où les caractères traités sont le plus souvent des caractères quantitatifs. Les méthodes se sont surtout développées autour de l'utilisation du modèle mixte et de l'utilisation des haplotypes (qui permettent d'exploiter plus facilement le côté transmission). Elles sont utilisées pour tester l'association entre des haplotypes et un caractère à une position donnée du génome, et ceci conditionnellement aux informations provenant du LD et de la transmission des haplotypes des pères aux descendants. L'objectif principal n'est donc pas le même que les méthodes FBAT/QTDT qui visent à bien distinguer les effets intra et inter familles. Ici, l'information sur la transmission n'est pas utilisée pour le test mais pour raffiner les probabilités d'IBD entre les haplotypes. De ce fait, ces méthodes, même si elles sont plus robustes à la structure de population (familiale) que de simples analyses d'association, ne sont pas censées l'être complètement et le sont encore moins pour des mélanges de populations. Il est donc nécessaire de prendre en compte la stratification de la population en introduisant dans le modèle des effets polygéniques et/ou des effets fixes (discuté dans la section sur la structure de population) pour rendre ces méthodes robustes. Par contre, elles profitent des bonnes propriétés du modèle mixte et de la précision des probabilités IBD entre haplotypes et sont ainsi plus puissantes que les méthodes FBAT/QTDT.

1.5 Tests multiples

Lors d'une analyse d'association classique, SNP par SNP ou haplotypique, on teste H_0 : "Il n'y pas d'association entre le marqueur (ou l'haplotype) et le phénotype" contre H_1 : "Le marqueur (ou haplotype) a un effet sur le phénotype". Ce test doit se faire en contrôlant l'erreur de première espèce à un niveau α (nous prendrons ici l'exemple $\alpha = 5\%$). En analyse d'association sur un ensemble de n tests SNP par SNP, cela signifie qu'on s'attend à avoir 5% de nos tests qui rejettent H_0 lorsque celle-ci est vraie où encore 5% de faux positifs (F_p) :

TABLE 1.15 – Scénarios possibles lors de n tests d'associations

	H_0 acceptée	H_0 rejetée	
H_0 vraie	V_n	F_p	V
H_0 fausse	F_n	V_p	F
	n-R	R	n

Ainsi dans une analyse SNP par SNP avec 50000 marqueurs, on fera $n = 50000$ tests au niveau 5% et on s'attend ainsi à avoir 2500 faux positifs! Ce nombre de faux positifs sera très largement supérieur au nombre de vrais positifs (V_p). La question du contrôle des erreurs dans les tests multiples est donc essentielle pour ces analyses. Nous présentons dans ce rapport trois

modalités pour maîtriser ces erreurs.

1.5.1 Correction de Bonferroni

L'objectif est de contrôler le nombre de faux positifs sur l'ensemble des N tests pour qu'il ne dépasse pas un nombre $N\alpha$ fixé. La probabilité α d'avoir au moins un test significatif par hasard sur l'ensemble des tests est appelée *Family-Wise Error-Rate* (FWER) :

$$\alpha = P_{H_0}(F_p > 0) = 1 - P_{H_0}(F_p = 0) = 1 - (1 - \alpha')^N$$

Où α' est le niveau de significativité pour chacun des N tests. L'idée est d'adapter le niveau de test pour chaque SNP au niveau global (de la famille de tests) α souhaité. Cette formule est à l'origine de deux corrections très connues, celle de Bonferroni (1936) et Sidak (1967). Sidak (1967) propose une application stricte de la formule, alors que Bonferroni (1936), notant que, pour α' petit $(1 - \alpha')^n \approx 1 - n\alpha'$, suggère de choisir $\alpha' = \alpha/n$ comme niveau de test individuel.

Cependant, ces corrections font l'hypothèse que les tests sont indépendants. Dans une étude d'association sur tout le génome, cette hypothèse est fautive pour plusieurs raisons dont la principale étant l'existence d'un déséquilibre de liaison entre locus. Ainsi, pour des locus en fort LD, ces corrections ont tendance à être assez conservatives.

1.5.2 Seuils par permutations

Pour rendre compte de la structure de déséquilibre de liaison observée entre les locus et approximer l'erreur de première espèce sur les N tests, une procédure de permutations peut être utilisée. L'idée est de ne pas modifier les génotypes et de permuter les phénotypes K fois. A chaque permutation, les N tests correspondants aux N SNP ou haplotypes sont faits. Les permutations des phénotypes, cassant leur associations avec les SNP, l'hypothèse de H_0 est surement satisfaites. En réalisant l'ensemble des tests pour chaque permutation, on obtient $N * K$ p -value. Le seuil α global à 5% est donné par le quantile $1 - \alpha$ de la distribution des p -values.

Cette procédure par permutations présentent l'avantage de créer un seuil directement à partir des données. Elle est malheureusement très gourmande en termes de temps de calculs. En effet, il faut environ 10000 permutations pour obtenir un seuil à 0.01% et 1000 permutations pour un seuil à 5% (Churchill and Doerge, 1994).

1.5.3 False Discovery Rate (FDR)

Une approche plus récente redéfinit le problème en remplaçant le α global par la notion de *false discovery rate* (FDR) qui représente la proportion du nombre de faux positifs sur le nombre total de tests rejetés. Mathématiquement, ce taux s'écrit :

$$FDR = E \left[\frac{F_p}{F_p + V_p} \right] = E \left[\frac{F_p}{R} \right]$$

Cependant, cette formule n'est pas définie pour $R = 0$, donc Benjamini and Hochberg (1995) ont proposé :

$$FDR = E \left[\frac{F_p}{R} | R > 0 \right] Pr(R > 0)$$

Benjamini and Hochberg (1995) donnent une procédure pour calculer le FDR :

1. Ordonner les p -values $p_{(1)} < \dots < p_{(m)}$ avec m le nombre de p -values.
2. Chercher le $\max(k)$ tel que $p_{(k)} < \alpha k/m$
3. Déclarer significatives les p -values inférieures au $\max(k)$

Cette mesure définit la proportion de fausses découvertes. Storey and Tibshirani (2003) voulant plutôt définir la proportion que les découvertes soient fausses, ont proposé d'utiliser une autre définition :

$$pFDR = E \left[\frac{F_p}{R} | R > 0 \right]$$

Ces deux définitions étant finalement assez proches dans des études d'associations avec de nombreux SNP, car on s'attend alors à avoir $Pr(R > 0) \approx 1$ et donc $FDR \approx pFDR \approx E[F_p]/E[R]$.

A la place d'ajuster les p -value comme dans les corrections de Bonferroni ou Sidak, de nouvelles valeurs, qui dérivent des $pFDR$ et qui sont appelées q -values, sont intégrées. L'idée est tout d'abord de calculer plusieurs $pFDR$ correspondant à plusieurs seuils α (où $0 < \alpha \leq 1$) de la façon suivante :

$$pFDR(\alpha) \approx \frac{E[F_p(\alpha)]}{E[R(\alpha)]}$$

Le calcul de $E[R(\alpha)]$ est simplement le nombre de p -values observées $\leq \alpha$. On définit la valeur $E[F_p(\alpha)] = V\alpha = \pi_0 m \alpha$ où π_0 , qui représente la part réelle d'hypothèse H_0 vraie, doit être estimé (Storey and Tibshirani, 2003). Ainsi, on obtient les q -values de la manière suivante :

$$q_i = \min_{\alpha \geq p_i} pFDR(\alpha) \quad i = 1, \dots, m$$

On note que pour un seuil α équivalent, $pFDR = \hat{\pi}_0 FDR$. Malgré la nécessité d'introduire certaines hypothèses dans l'estimation de π_0 , les méthodes FDR semblent être les plus adaptées pour les méthodes d'associations sur tout le génome (Manly et al., 2004).

Comparaison algébrique de méthodes uni-QTL

2.1 Résumé de l'article

Avec les marqueurs SNP, il est devenu possible d'exploiter le déséquilibre gamétique pour repérer et localiser les gènes ayant un effet sur les caractères de sensibilité aux maladies ou de productivité. Les premières méthodes ont visé à rechercher de tels déséquilibres gamétiques entre un ou des locus marqueurs et un locus putatif dont certains génotypes pourraient entraîner une prédisposition à une maladie. Néanmoins, comme nous l'avons vu précédemment, la question des associations erronées a été rapidement mise au centre des préoccupations. De telles erreurs se produisent quand la classification des membres d'une population selon des informations de marqueurs est confondue avec une autre source d'hétérogénéité, ayant elle-même un effet sur le caractère analysé. Les causes d'hétérogénéité sont souvent d'origines génétiques et liées à la stratification de la population, qui peut venir d'un mélange entre populations et/ou d'une structuration familiale.

Pour lutter contre ces problèmes, les analyses LDLA semblent être la solution mais cette option implique de construire un échantillonnage approprié, dans lequel la structuration en famille est voulue. Notamment parce que les contraintes pour constituer les échantillons y sont beaucoup moins fortes (Balding, 2006), les analyses d'association sont malgré tout très utilisées. Au départ, ces analyses étaient basées sur l'hypothèse que les individus de la population sont non apparentés, et les effets d'une éventuelle stratification n'étaient pas corrigés. Pour les caractères quantitatifs, un modèle très informatif pour prendre en compte ces effets est le modèle mixte dans lequel les performances sont décrites par un effet du SNP et un effet aléatoire polygénique dont la matrice de variance-covariance, également appelée matrice de parenté, peut être calculée à partir du pedigree ou de l'ensemble des génotypes marqueurs. Il semble qu'un consensus soit trouvé autour d'un tel modèle en analyse LDA (Yu et al., 2006; Amin et al., 2007; Zhang et al., 2008; Price et al., 2010).

Néanmoins, avec le modèle mixte, il faut, pour chaque marqueur, estimer les variances polygéniques et résiduelle et tester si l'effet du marqueur est significatif. Cette opération, répétée autant de fois qu'il y a de marqueurs testés peut s'avérer longue dans des protocoles de grande taille et des approches plus simples ont été proposées : l'approche GRAMMAR, de Aulchenko et al. (2007a) et Amin et al. (2007), qui réalise le modèle mixte en deux étapes, et l'approche EMMAX (pour Efficient Mixed-Model Association eXpedited), de Kang et al. (2010), qui estime la variance polygénique au préalable et la fixe dans l'analyse de tous les SNP.

Ces méthodes ont été évaluées par simulation. Aulchenko et al. (2007a) ont comparé GRAM-

MAR au modèle mixte complet, à la régression sans effet polygénique, au QTDT et à un FBAT simple sur des jeux de données simulées selon trois types de pedigree. Amin et al. (2007) ont comparé GRAMMAR et GRAMMAR-GC (avec un contrôle génomique). Price et al. (2010) ont comparé l'analyse en composante principale (EIGENSTRAT), le test Armitage, EMMAX avec ou sans ACP et ROADTRIPS de Thornton and McPeck (2010) qui traite les données génomique en aléatoire. Wu et al. (2011) ont mis en balance les approches basées sur l'ACP avec le contrôle génomique et la méthode ROADTRIPS. Erbe et al. (2011) ont comparé par simulation trois méthodes GWAS : la régression simple, GRAMMAR et un MTDT qui applique une analyse TDT à une estimation de l'aléa de méiose.

Dans l'ensemble, ces études numériques montrent que les approches intra famille sont moins puissantes que les analyses cas témoin dans les populations non apparentées (Aulchenko et al., 2007a; Zhang et al., 2008) et que parmi ces dernières il y a peu de différences de puissance quand on ne corrige pas pour une éventuelle structuration (Grapes et al., 2004). Elles démontrent très clairement l'absence de robustesse des méthodes de base telles que le test Armitage ou la simple régression (Yu et al., 2006; Wu et al., 2011; Erbe et al., 2011). Elles prouvent que les modèles les plus complets (Yu et al., 2006; Price et al., 2010) sont les plus robustes à tout type de stratification, et que les approximations comme GRAMMAR et EMMAX donnent de très bons résultats tant en termes de contrôle des erreurs en présence de structure familiale que de vitesse de calcul, au prix d'une perte de puissance dans certains cas.

Les comparaisons de méthodes par simulations pèchent par leur manque de généralité. Plusieurs travaux donnent des résultats algébriques dans des cas simples. Ainsi Fan and Xiong (2002) ont formalisé des analyses d'association par régression uni ou bi-marqueur, en dérivant leur puissance à partir d'un coefficient de décentrement de la statistique de test qui est fonction du LD entre marqueurs et QTL. Le test de Cochran Armitage est étudié par Freidlin et al. (2002), Guedj et al. (2006) et Li et al (2010). La puissance d'une analyse d'association par ANOVA ou régression, fonction des fréquences alléliques ou génotypiques aux marqueurs est dérivée par Ambrosius et al. (2004) repris récemment par Kozlitina et al. (2010). Pour le QTDT de Abecasis et al. (2000), des résultats algébriques ont été obtenus dans le cas de mélange de populations. Ainsi, Abecasis et al. (2000) donnent les espérances des effets intra et inter familles avec ou sans l'information des génotypes parentaux et Boitard et al. (2010) généralisent ces formules pour les variances et le test. Dans Lange et al. (2002), les formules algébriques pour la puissance des tests FBAT sont dérivées conditionnellement aux génotypes parentaux et aux génotypes des descendants.

L'article, qui forme le coeur de ce deuxième chapitre, a pour but d'aller plus loin dans la formalisation algébrique des puissances et erreurs de première espèce de quelques unes de ces méthodes : la régression simple, les méthodes approchées GRAMMAR et FASTA ainsi que le QTDT. L'objectif est plus précisément de formaliser l'effet de la structure génétique, en se focalisant sur les situations de structures génétiques cachées dues aux apparentements et non aux mélanges de sous populations. Les phénotypes dépendent donc à la fois de l'effet éventuel du gène ou marqueur testé et d'un fond polygénique. Le modèle de référence est un modèle mixte dépendant en particulier de la matrice des apparentements entre individus phénotypés. Cette analyse algébrique précise les conditions dans lesquelles ces méthodes sont utilisables et donne des pistes pour l'organisation des populations d'étude.

Toutes les méthodes comparées ici testent si la variabilité d'un caractère quantitatif y est associé ou non au polymorphisme d'un SNP. Le caractère y est supposé polygénique, c'est-à-dire sous l'influence de plusieurs QTL. Lorsqu'on teste l'association d'un SNP avec y , la variable aléatoire

y peut être décrite comme la somme des effets fixes putatifs β d'un QTL lié à ce SNP, un effet aléatoire polygénique u rassemblant les effets de tous les autres QTLs (non liés) et un effet aléatoire résiduel e ($y = \mu + x\beta + u + e$). Par la suite, ce modèle sera appelé le "vrai modèle". Les méthodes que nous mettons en comparaison sont utilisées pour estimer l'effet β en utilisant des modèles simplifiés. Nous indiquerons (i) ces modèles simplifiés. Pour tester l'effet du SNP dans chacun de ces modèles simplifiés (i), un test de student est construit de la manière suivante :

$$\tau^{(i)} = \frac{\hat{\beta}^{(i)} / \sqrt{V^{(i)}(\hat{\beta}^{(i)})}}{\sqrt{\hat{\sigma}_{e^{(i)}} / E^{(i)}(\hat{\sigma}_{e^{(i)}})}}$$

Comme y ne suit pas les modèles (i) mais le vrai modèle, ces tests ne suivent pas leurs distributions de student supposées. L'écrire du vrai test t dans ce modèle (i) serait :

$$t^{(i)} = \frac{\hat{\beta}^{(i)} / \sqrt{V(\hat{\beta}^{(i)})}}{\sqrt{\hat{\sigma}_{e^{(i)}} / E(\hat{\sigma}_{e^{(i)}})}}$$

Avec $E(\hat{\sigma}_{e^{(i)}})$ et $V(\hat{\beta}^{(i)})$ l'espérance et la variance de l'estimateur $\hat{\beta}^{(i)}$ lorsqu'on suppose que y suit le vrai modèle. Ce test suit asymptotiquement une distribution normale d'espérance $\frac{E(\hat{\sigma}_{e^{(i)}})}{\sqrt{V(\hat{\beta}^{(i)})}}$ et de variance 1. Le test $\tau^{(i)}$, tel que préconisé dans les modèles simplifiés, peut être réécrit de la manière suivante :

$$\tau^{(i)} = t^{(i)} \sqrt{\frac{V(\hat{\beta}^{(i)})}{V^{(i)}(\hat{\beta}^{(i)})}} \sqrt{\frac{E^{(i)}(\hat{\sigma}_{e^{(i)}})}{E(\hat{\sigma}_{e^{(i)}})}}$$

Ainsi, on montre que la distribution de $\tau^{(i)}$ suit une normale d'espérance $E(\tau^{(i)})$ et de variance $V(\tau^{(i)})$ avec :

$$\begin{aligned} E(\tau^{(i)}) &= \hat{\beta}^{(i)} \sqrt{\frac{1}{V^{(i)}(\hat{\beta}^{(i)})}} \sqrt{\frac{E^{(i)}(\hat{\sigma}_{e^{(i)}})}{E(\hat{\sigma}_{e^{(i)}})}} \\ V(\tau^{(i)}) &= \frac{V(\hat{\beta}^{(i)})}{V^{(i)}(\hat{\beta}^{(i)})} \frac{E^{(i)}(\hat{\sigma}_{e^{(i)}})}{E(\hat{\sigma}_{e^{(i)}})} \end{aligned}$$

Notre but est donc de donner les expressions de ces moments, pour chaque modèle (i), conditionnellement aux paramètres du vrai modèle de y , c'est-à-dire la matrice de parenté, le vecteur des génotypes et la variance polygénique. A partir de là, les erreurs de type-I et la puissance marginale des tests de chaque modèle (i) sont déterminées en fonction de la distribution d'échantillonnage des génotypes et des estimateurs de la variance polygénique sachant la matrice de parenté et les vrais paramètres des variances.

Nous montrons quelques exemples d'application directe de ces formules dans le cadre de familles de père avec 5, 30 et 60 descendants par père et une taille d'échantillon fixée à 600 individus, ce qui correspond à peu près à la taille des données GENEQUIN présentées dans le chapitre suivant. Au niveau de la robustesse, les méthodes QTDT et FASTA donnent de très bon résultats alors que le modèle de Régression et la méthode GRAMMAR sont respectivement très inflaté et légèrement conservatif quand la structure familiale et l'héritabilité augmentent. La puissance de GRAMMAR

et de FASTA dépend de la structure familiale pour des héritabilités moyennes comprises entre 0.1 et 0.3. La puissance pour le modèle de régression, quant à elle, diminue fortement avec l'augmentation de la structure familiale et de l'héritabilité. Enfin, la puissance du QTDT n'est pas impactée ni par la structure familiale ni par l'héritabilité, mais reste beaucoup plus faible que les autres méthodes. Parmi les modèles simplifiés étudiés, FASTA s'est montré être le plus robuste et le puissant.

Une étude sur le choix d'un dispositif expérimental est par la suite abordée à partir des résultats théoriques sur la puissance de la méthode FASTA et montre peu de différences en terme de nombre d'individus à phénotyper entre les différentes structures familiales envisagées mais ce nombre est très différent entre les effets du SNP supposés.

2.2 Article Méthodologique

Cet article, intitulé "Influence of population structure on power and robustness of current association mapping tests" a été soumis à *Genetics*. Les auteurs sont : S. Teyssède, J-M Elsen et A. Ricard. Les détails de certaines formules sont présentés en Annexe A.

1

2 **Statistical distributions of test statistics used for quantitative trait association mapping**
3 **in structured populations**

4 Simon Teyssevre*, Jean-Michel Elsen* and Anne Ricard[§]

5

6 ^{*}INRA, UR 631 Station d'Amélioration Génétique des Animaux, 31326 Castanet-Tolosan,
7 France.

8 [§]INRA, UMR 1313 Génétique Animale et Biologie Intégrative, 78352 Jouy-en-Josas, France.

9

10 **Running head:** Distributions of GWAS test statistics

11 **Keywords:** GWAS, structured population, type I error, power, families.

12 **Corresponding author:**

13 Anne Ricard

14 INRA-SAGA

15 Auzeville

16 B.P. 52627

17 31326 Castanet-Tolosan Cedex

18 France

19 phone number: 33 5 61 28 51 83

20 fax number: 33 5 61 28 53 53

21 email: anne.ricard@toulouse.inra.fr

22

23 **ABSTRACT**

24 Spurious associations between single nucleotide polymorphisms and phenotypes are a big
25 issue in genome wide association studies and have led to the underestimation of type I errors
26 and overestimation of the number of quantitative trait loci found. Many authors have
27 investigated the influence of the structure of populations on the robustness of methods using
28 simulations. The aim of this paper is to go further in the algebraic formalization of power and
29 first type errors for some of the classical statistical methods used: simple regression, two
30 approximate methods of mixed models involving the single nucleotide effect and the random
31 polygenic effect and the transmission/disequilibrium test for quantitative traits and nuclear
32 families. The expectation and variance of the test statistics are given in function of the
33 relationship matrix and heritability of the polygenic effect. These formulae attempted to
34 compute type 1 errors and the power of these methods for any kind of relationship matrix
35 between the phenotyped and genotyped data in any situation of heritability. They can
36 therefore be easily used to provide the correct threshold of type 1 errors and to calculate the
37 power in order to organize protocols.

38 **INTRODUCTION**

39 SNP information has made possible the use of gametic disequilibrium for the detection and
40 localization of loci affecting phenotypes. The first methods searched for such disequilibrium
41 between one or a few marker loci and loci provoking disease susceptibility. Case control
42 designs were used (Risch 2000). Typically, data were analyzed comparing the frequency of
43 marker alleles between healthy and diseased individuals, using for instance the Relative Risk
44 criterion (Woolf 1955). A similar approach for quantitative traits (including production traits
45 in animals or plants) was to model the expectation of their distribution as a linear combination
46 of marker genotypes, alleles or haplotype effects. Grapes *et al.* (2004) and Zhao *et al.* (2007a)

47 recently demonstrated that the single marker regression model is as powerful and precise as
48 other more sophisticated techniques (multiple regression, regression on haplotypes or the IBD
49 method proposed by Meuwissen and Goddard (2000)).

50 Spurious associations are a big issue that has been investigated by many authors. Such errors
51 occur when population classification based on marker information is confounded with another
52 source of heterogeneity, which affects the analyzed trait. The problem of genetic
53 heterogeneity has been particularly studied. Two non-exclusive situations can occur: (i) a
54 population consisting of genetically different subpopulations and (ii) a population consisting
55 of related individuals, the information about these relationships being, or not, recorded. It was
56 clearly shown for instance that neither the Relative Risk nor simple regression were robust to
57 a genetic stratification of the population or the mixture of different groups (breeds, lines, etc.)
58 or families (Pritchard and Rosenberg 1999, Cardon and Palmer 2003, Marchini *et al.* 2004,
59 Clayton *et al.* 2005).

60 To fight against these potential errors, many different approaches have been proposed. The
61 first was to restrict the analysis to within family comparisons, linking association analysis to
62 transmission studies. Within this framework, samples have to be carefully organized,
63 recruiting ad hoc families. The principle was to correlate, for each tested marker, the progeny
64 phenotype (for the studied trait) with the deviation, from its expectation, of the number of
65 alleles the progeny received from its heterozygous parent(s). This idea was first implemented
66 in the Transmission Disequilibrium Test (TDT) performed by Spielman *et al.* (1993), and
67 then largely developed by others. Ewens and Spielman (1995), when comparing the TDT and
68 a “Within Family Contingency Statistic” close to the Falk and Rubinstein (1987) Haplotype
69 Relative Risk, demonstrated the robustness of TDT in various subdivision and admixture
70 scenarios.

71 Two largely represented families of methods generalize these within-family comparisons to
72 quantitative traits: the “quantitative TDT” or QTDT (Allison 1997; Fulker *et al.* 1999;
73 Abecasis *et al.* 2000a; Abecasis *et al.* 2000b) and the Family-Based Association Tests or
74 FBAT (Rabinowitz 1997; Laird *et al.* 2000; Laird and Lange 2006; Laird and Lange 2008;).
75 All these methods are robust to population stratifications, of similar power (Lange *et al.* 2002;
76 Ewens *et al.* 2008) and more powerful than the first tests developed for family-based
77 association studies (Abecasis *et al.* 2000b).

78 If the limitation of spurious associations by using within-family analyses was very successful,
79 case–control association studies in populations consisting of individuals assumed to be
80 unrelated were nevertheless frequent, in particular because the recruitment of the
81 corresponding samples is much easier (Balding 2006). A number of techniques were derived
82 to limit false positives: the “genomic control” corrects the test statistic by the deviation
83 between the observed and expected medians in the absence of stratification (Devlin and
84 Roeder 1999, Bacanu *et al.* 2002), a structure effect may be added to the performance
85 description model (Pritchard *et al.* 2000a; Pritchard *et al.* 2000b; Satten *et al.* 2001; Zhu *et al.*
86 2002; Price *et al.* 2006; Zhu *et al.* 2008) and marker transmission between generations can be
87 generalized (Meuwissen and Goddard 2000; Meuwissen *et al.* 2002).

88 Concerning quantitative traits, known or hidden genetic structures are usefully modeled in
89 mixed models where the performance expectation is the sum of fixed effects, including the
90 tested marker effect and a random individual effect. Covariances between these individual
91 effects are proportional to the polygenic variance and coancestry coefficients which can be
92 estimated from pedigree or marker information (Ritland 1996; Hayes *et al.* 2007; Vanraden
93 2008; Yang *et al.* 2010). This model is a standard that has been used in animal genetics for
94 many years (Henderson 1975; Quaas and Pollak 1980) and more recently but successfully in
95 human genetics (Price *et al.* 2010; Zhang *et al.* 2010).

96 Following this modeling, polygenic and residual variances have to be estimated for each
97 marker before testing its significance. This estimation phase, to be repeated as many times as
98 there are markers tested, can be a limiting factor in large designs and simpler approaches have
99 been proposed. Aulchenko *et al.* (2007a); Aulchenko *et al.* (2007b) and Amin *et al.* (2007)
100 developed the GRAMMAR method which tests marker effects on phenotypes corrected by
101 the individual expected value estimated in a restricted model which is free of this effect. The
102 FASTA approach described by Chen and Abecasis (2007) is a score test equivalent to a BLUP
103 applied to the standard model, where variances are only estimated once in the restricted model
104 free of this effect.

105 Other approaches were proposed with the aim of accelerating computations (EMMA for
106 Efficient Mixed-Model Association, Kang *et al.* 2008; EMMAX for eXpedited, Kang *et al.*
107 2010; P3D for Population Parameters Previously Determined, Zhang *et al.* 2010). Finally, a
108 few models integrated spurious associations arising from subpopulations and family structures
109 (Yu *et al.* 2006; Amin *et al.* 2007; Zhao *et al.* 2007b; Zhang *et al.* 2009; Price *et al.* 2010).

110 These methods were evaluated by simulations. Aulchenko *et al.* (2007a) compared
111 GRAMMAR to the full mixed model, to the regression model without a polygenic effect, to
112 the QTDT and to a simple FBAT using simulated data sets corresponding to typical pedigrees.
113 Amin *et al.* (2007) compared the Genomic Control with GRAMMAR and GRAMMAR-GC.
114 Price *et al.* (2010) compared PCA (EIGENSTRAT), the Armitage test, EMMAX with or
115 without PCA and ROADTRIPS proposed by Thornton and Mcpeck (2010) in which genomic
116 data are modeled as random variables. Wu *et al.* (2011) compared PCA-based approaches
117 (Price *et al.* 2006, EIGENSTRAT; Zeggini *et al.* 2008, PCA-based Logistic Regression; Lee
118 *et al.* 2010, LAPSTRUCT which makes use of spectral graph theory to build principal
119 components) to the genomic control described by Devlin and Roeder (1999) and

120 ROADTRIPS. Erbe *et al.* (2011) compared three GWAS techniques: simple regression,
121 GRAMMAR and a “MTDT” which is a QTDT applied to Mendelian sampling terms.
122 On the whole, these numerical studies show that within-family approaches are less powerful
123 than case control analyses in populations of unrelated individuals (Aulchenko *et al.* 2007a;
124 Zhang *et al.* 2009) with no large differences between the latter (Grapes *et al.* 2004). They
125 demonstrate clearly the non-robustness of the simplest methods such as the Armitage test or
126 simple regression (Yu *et al.* 2006; Astle and Balding 2009; Erbe *et al.* 2011; Wu *et al.* 2011).
127 They show that more elaborate models are robust to any type of stratification (Yu *et al.* 2006;
128 Zhao *et al.* 2007b; Price *et al.* 2010), and that approximate techniques such as GRAMMAR
129 and EMMAX are very efficient in terms of error control when family structures exist, and of
130 computing speed, but show decreased power in some situations.

131 One of the main limits when comparing methods based on simulations is their lack of
132 generalization and only a few studies have provided algebraic results in simple situations. For
133 instance, Fan and Xiong (2002) formalized single or bi-marker association analyses by
134 regression deriving their power as a function of the non-centrality parameter of the test
135 statistic which depends on the LD between the markers and the QTL. In Ewens and Spielman
136 (1995) the Relative Risk, the Within-Family Contingency Statistic and the TDT were
137 compared algebraically considering a few admixture scenarios. The Cochran Armitage test
138 was studied by Freidlin *et al.* (2002); Guedj *et al.* (2007); Li *et al.* (2010). The power of
139 ANOVA or regression-based association analyses, as a function of allelic or genotypic
140 frequencies was derived by Ambrosius *et al.* (2004), and recently completed by Kozlitina
141 *et al.* (2010). Results were obtained in population mixture situations for the QTDT described by
142 Abecasis *et al.* (2000a). For instance, Abecasis *et al.* (2000a) provided the within- and
143 between-family expectations with and without parental information, and Boitard *et al.* (2010)
144 generalized the corresponding formulae for variances and tests. In Lange *et al.* (2002),

145 algebraic formulae were given for the power of FBAT, depending on parental and progeny
146 genotypes.

147 The aim of this paper is to go further in the algebraic formalization of the power and first type
148 errors for some of the aforementioned statistics: simple regression, the approximate methods
149 GRAMMAR (Amin *et al.* 2007; Aulchenko *et al.* 2007a) and FASTA (Chen and Abecasis
150 2007) and the QTDT described by Abecasis *et al.* (2000a). Our goal was to explore the effect
151 of genetic structure, focusing on hidden familial relationships and not on population mixtures.
152 In such situations, phenotypes are both under the influence of the QTL effect linked to tested
153 markers and a polygenic background. The model of reference used in this study was the
154 standard mixed model which includes the coancestry coefficients as parameters. This work
155 shows in which situations the methods studied here can be considered as correct and gives
156 some help for population sampling.

157

158 **MATERIELS AND METHODS**

159 ***Aim of the paper***

160 All the compared statistics test if, or not, the variability of a quantitative trait, y , is associated
161 to the polymorphism of SNPs considered one by one. The trait y is assumed polygenic, i.e.
162 under the influence of many QTL. When testing a particular SNP- y association, the y random
163 variable may be described as the sum of the putative fixed effect β of a QTL linked to this
164 SNP, a random polygenic effect u pooling the effects of all other (unlinked) QTLs and
165 random noise e ($\mathbf{y} = \mathbf{1}\mu + \mathbf{x}\beta + \mathbf{u} + \mathbf{e}$). Hereafter, this model is designated as the “true model”.
166 The approximated methods mentioned in the introduction, aim at estimating the β effect,
167 using simplified models. Generally, for each of these simplified models (i), the regression
168 coefficient of the SNP effect is estimated by the general least squares estimator $\hat{\beta}^{(i)}$. A

169 classical student test is then constructed to test the null hypothesis for the SNP effect. Let
 170 $E^{(i)}(\hat{\beta}^{(i)})$ and $V^{(i)}(\hat{\beta}^{(i)})$ be the expectation and variance of the estimator $\hat{\beta}^{(i)}$, $\hat{\sigma}_{e^{(i)}}^2$ and
 171 $E^{(i)}(\hat{\sigma}_{e^{(i)}}^2)$ an estimator of the residual variance and its expectation, all assuming model (i).
 172 The tests are:

$$173 \quad \tau^{(i)} = \frac{\hat{\beta}^{(i)} / \sqrt{V^{(i)}(\hat{\beta}^{(i)})}}{\sqrt{\hat{\sigma}_{e^{(i)}}^2 / E^{(i)}(\hat{\sigma}_{e^{(i)}}^2)}}$$

174 As a ratio between a normal distribution with unit variance and an independent square root
 175 χ^2 distribution, these tests are supposed to follow non-central t-distributions with a non-
 176 centrality parameter $E^{(i)}(\hat{\beta}^{(i)}) / \sqrt{V^{(i)}(\hat{\beta}^{(i)})}$. However, these tests do not in fact follow these
 177 distributions because \mathbf{y} does not follow the simplified models (i). The tests which follow a
 178 student distribution are the tests computed with expectations and variance of $\hat{\beta}^{(i)}$
 179 corresponding to the true model for \mathbf{y} . Let $E(\hat{\beta}^{(i)})$ and $V(\hat{\beta}^{(i)})$ be the expectation and
 180 variance of the estimator $\hat{\beta}^{(i)}$ and $E(\hat{\sigma}_{e^{(i)}}^2)$ the expectation of the estimator of residual variance
 181 assuming \mathbf{y} follows the true model. The valid student tests are:

$$182 \quad t^{(i)} = \frac{\hat{\beta}^{(i)} / \sqrt{V(\hat{\beta}^{(i)})}}{\sqrt{\hat{\sigma}_{e^{(i)}}^2 / E(\hat{\sigma}_{e^{(i)}}^2)}}$$

183 These student distributions tend to normal distributions when the number of animals involved
 184 in the analysis is sufficiently high (hundred animals). These normal distributions have mean
 185 $\frac{E(\hat{\beta}^{(i)})}{\sqrt{V(\hat{\beta}^{(i)})}}$ and variance 1 (Johnson and Kotz 1970). Because the test $\tau^{(i)}$ used instead of $t^{(i)}$

186 may be expressed as $\tau^{(i)} = t^{(i)} \sqrt{\frac{V(\hat{\beta}^{(i)})}{V^{(i)}(\hat{\beta}^{(i)})} \frac{E^{(i)}(\hat{\sigma}_{e^{(i)}}^2)}{E(\hat{\sigma}_{e^{(i)}}^2)}}$, the test $\tau^{(i)}$ will have normal
 187 distribution with mean:

188
$$E(\tau^{(i)}) = E(\hat{\beta}^{(i)}) \sqrt{\frac{1}{V^{(i)}(\hat{\beta}^{(i)})}} \sqrt{\frac{L \cdot V_{\sigma^2}^{(i)}}{E(\hat{\sigma}_{\sigma^2}^2)}} \quad (\text{I})$$

189 and variance:

190
$$V(\tau^{(i)}) = \frac{V(\hat{\beta}^{(i)})}{V^{(i)}(\hat{\beta}^{(i)})} \frac{E^{(i)}(\hat{\sigma}_{\sigma^2}^2)}{E(\hat{\sigma}_{\sigma^2}^2)}. \quad (\text{II}).$$

191 The aim of the paper was to express these moments as a function of the parameters of the true
 192 model for \mathbf{y} , i.e. the relationship matrix and polygenic variance. Then, the true type I error and
 193 power of the tests in each of model (i) were analytically determined. Under the null
 194 hypothesis ($H_0, \beta = 0$), the used tests $\tau^{(i)}$ were supposed to have expectation 0 and variance

195 1. For a given expected type I error α , the threshold for rejecting the hypothesis was chosen
 196 as $t_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ where Φ is the standardized cumulative normal distribution. With the
 197 same threshold, the knowledge of the true variance and expectation of the tests $\tau^{(i)}$ allowed

198 us to compute the actual true type 1 error $\alpha^{(i)} = 2 \left[1 - \Phi \left(\frac{t_{\alpha/2} - E_{\beta=0}(\tau^{(i)})}{\sqrt{V_{\beta=0}(\tau^{(i)})}} \right) \right]$ where $E_{\beta=0}(\tau^{(i)})$ is

199 the expectation of test statistic and $V_{\beta=0}(\tau^{(i)})$ the variance of test statistic assuming the null
 200 hypothesis. Under the alternative hypothesis ($H_1, \beta = b$), the statistical power was computed

201 as $P_{\alpha,b}^{(i)} = 1 - \Phi \left(\frac{t_{\alpha/2} - E_{\beta=b}(\tau^{(i)})}{\sqrt{V_{\beta=b}(\tau^{(i)})}} \right)$, with the same definition for the threshold and the true

202 regression coefficient b . The bias of the estimator of the regression coefficient of the SNP
 203 effect was also computed as $(E_{\beta=b}(\hat{\beta}^{(i)}) - b) / b$.

204 In the following, the true model and simpler models (i) were defined. Then, the expectation
 205 and variance of the used test $\tau^{(i)}$ were expressed as function of the parameters conditional to
 206 given genotypes and variance of polygenic effects. Finally, the marginal Type I error and

207 power were given by integrating the genotypes and polygenic variance estimators given the
 208 relationship matrix and true variance parameters. It should be noted that the power was
 209 calculated according to the SNP effect, not the effect of a QTL linked to the SNP. To
 210 calculate the power to detect a QTL effect assuming linkage disequilibrium r^2 between a SNP
 211 and the QTL, the regression coefficient of the QTL effect corresponded to the SNP effect
 212 divided by r .

213 **Models**

214 The true model was the following mixed model:

$$215 \quad \mathbf{y} = \mathbf{1}\mu + \mathbf{x}\beta + \mathbf{u} + \mathbf{e},$$

216 where \mathbf{y} was the vector of the observed trait (one performance by animal), μ the vector of the
 217 overall mean, β the regression coefficient of the fixed SNP effect, \mathbf{u} the vector of random
 218 additive genetic effects of the animals and \mathbf{e} the vector of random residuals. Let $E(\mathbf{u}) = \mathbf{0}$,
 219 $V(\mathbf{u}) = \mathbf{A}\sigma_u^2$ with \mathbf{A} the relationship matrix and σ_u^2 the additive polygenic variance, and
 220 $V(\mathbf{e}) = \mathbf{I}\sigma_e^2$ with σ_e^2 the residual variance. The heritability was defined as the ratio between
 221 the polygenic genetic variance and the sum of polygenic variance and residual variance:

$$222 \quad h^2 = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \text{ and we defined } \sigma_y^2 = \sigma_u^2 + \sigma_e^2. \text{ The vector } \mathbf{x} \text{ was the incidence vector of the SNP}$$

223 effect, defined as $\mathbf{x} = \mathbf{w} - \mathbf{1}\bar{w}$ (see for example Meuwissen *et al.* (2009)), where \mathbf{w} was
 224 $-2p/\sqrt{2pq}$ for genotype 11, $(1-2p)/\sqrt{2pq}$ for genotype 12, $2q/\sqrt{2pq}$ for genotype 22,
 225 with p the frequency of allele 2. So that $E(w) = 0$ and $V(w) = 1$. According to the definition
 226 of \mathbf{x} , the relation between the regression coefficient of the reference model and the allele
 227 substitution effect (difference between genotype 11 and 12 or 12 and 22) was:

$$228 \quad \beta_{allele} = \beta / \sqrt{2pq}. \text{ So, the same statistical power was obtained for different substitution allele}$$

229 effects according to the frequency of the allele (MAF, minimum allele frequency). For
230 simplicity no other fixed effect was added to the model.

231 We analysed 4 simpler models used instead of this true model to estimate the SNP effect. The
232 first three models were association analysis whereas the fourth was a linkage and association
233 analysis. The superscript (i), $i = 1, \dots, 4$ was added to identify the effects specific to each of the
234 four models.

235 1) The first model was a simple REGRESSION model without a polygenic effect:

$$236 \quad \mathbf{y} = \mathbf{1}\mu^{(1)} + \mathbf{x}\beta^{(1)} + \mathbf{e}^{(1)}. \quad (1)$$

237 2) The second model was the GRAMMAR method developed by Aulchenko *et al.* (2007a)
238 and Amin *et al.* (2007). The GRAMMAR method is a two-step method, first:

$$239 \quad \mathbf{y} = \mathbf{1}\mu^{(2a)} + \mathbf{u}^{(2a)} + \mathbf{e}^{(2a)}. \quad (2a)$$

240 Then the estimates of residuals are used to estimate the SNP effect:

$$241 \quad \hat{\mathbf{e}}^{(2a)} = \mathbf{1}\mu^{(2b)} + \mathbf{x}\beta^{(2b)} + \mathbf{e}^{(2b)}. \quad (2b)$$

242 3) The third model was the FASTA approach from Chen and Abecasis (2007). This model
243 uses a score which is equivalent to the full mixed model:

$$244 \quad \mathbf{y} = \mathbf{1}\mu^{(3)} + \mathbf{x}\beta^{(3)} + \mathbf{u}^{(3)} + \mathbf{e}^{(3)}. \quad (3),$$

245 but with variance components estimated from the same random model as used in the first step
246 of GRAMMAR analysis ($\mathbf{y} = \mathbf{1}\mu^{(2a)} + \mathbf{u}^{(2a)} + \mathbf{e}^{(2a)}$), so with $V(\mathbf{u}^{(3)}) = \mathbf{A}\hat{\sigma}_{\mu^{(2a)}}^2$ instead of
247 $V(\mathbf{u}^{(3)}) = \mathbf{A}\sigma_{\mu^{(3)}}^2$ and $V(\mathbf{e}^{(3)}) = \mathbf{I}\hat{\sigma}_{e^{(2a)}}^2$.

248 4) The fourth model was a linkage analysis and association method: the QTDT developed by
249 Abecasis *et al.* (2000a). Let $\mathbf{z} = \frac{\mathbf{x}_s + \mathbf{x}_d}{2}$ where \mathbf{x}_s and \mathbf{x}_d denote the genotype of the sire
250 (respectively dam) of the animal, the model was:

$$251 \quad \mathbf{y} = \mathbf{1}\mu^{(4)} + (\mathbf{z} - \mathbf{1}\bar{z})\beta_b^{(4)} + (\mathbf{x} - \mathbf{z})\beta_w^{(4)} + \mathbf{e}^{(4)} \quad (4),$$

252 where $\beta_b^{(4)}$ is the regression coefficient between families and $\beta_w^{(4)}$ the regression coefficient
 253 within families.

254 **Expectation and variance of the estimator of the SNP effect and of the test**
 255 **statistic.**

256 **Model 1, REGRESSION model:** Assuming model (1), the SNP effect was estimated by:

$$257 \quad \hat{\beta}^{(1)} = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{y}.$$

258 The user assumed that $V^{(1)}(\hat{\beta}^{(1)}) = (\mathbf{x}'\mathbf{x})^{-1} \sigma_{e^{(1)}}^2$ and estimated the residual variance with the sum
 259 of the square of residuals assuming $E^{(1)}(\hat{\mathbf{e}}^{(1)'}\hat{\mathbf{e}}^{(1)}) = (n-2)\sigma_{e^{(1)}}^2$. But in fact, when considering
 260 that \mathbf{y} followed the true model, the true expressions were as follows. The expectation of this
 261 estimator was:

$$262 \quad E(\hat{\beta}^{(1)}) = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'E(\mathbf{y}) = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'(\mathbf{1}\mu + \mathbf{x}\beta) = \beta,$$

263 because $\mathbf{x}'\mathbf{1} = (\mathbf{w} - \bar{\mathbf{w}})' \mathbf{1} = 0$. So the estimator of the SNP effect was still unbiased. The
 264 variance of the estimator was:

$$265 \quad V(\hat{\beta}^{(1)}) = \sigma_e^2 \left[(\mathbf{x}'\mathbf{x})^{-1} + \frac{h^2}{1-h^2} (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{A}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1} \right]. \quad (\text{III})$$

266 So that the variance of the estimator of the SNP effect included a function of heritability and
 267 relationship matrix in addition to the usual factor involving residual variance. The residual
 268 variance was estimated using the sum of squares of residuals. The expectation was:

$$269 \quad nE(\hat{\sigma}_{e^{(1)}}^2) = E(\hat{\mathbf{e}}^{(1)'}\hat{\mathbf{e}}^{(1)}) = \sigma_e^2 \left[(n-2) + \frac{h^2}{(1-h^2)} (\text{tr}(\mathbf{A}) - (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{A}\mathbf{x} - (\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}'\mathbf{A}\mathbf{1}) \right],$$

270 with n the number of animals involved in the analysis.

271 Finally, the mean and variance of the test statistic effectively used were:

$$272 \quad E(\hat{\tau}^{(1)}) = \frac{\sqrt{\mu' \sigma_y} \sqrt{\mathbf{A} \mathbf{A}'} }{\sqrt{1+h^2 \frac{\text{tr}(\mathbf{A}) - (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{A}\mathbf{x} - (\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}'\mathbf{A}\mathbf{1} - (n-2)}{(n-2)}}}$$

$$273 \quad V(\hat{\tau}^{(1)}) = \frac{1+h^2 (\mathbf{x}'\mathbf{A}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1} - 1)}{1+h^2 \frac{\text{tr}(\mathbf{A}) - (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{A}\mathbf{x} - (\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}'\mathbf{A}\mathbf{1} - (n-2)}{n-2}}$$

274 **Model 2, GRAMMAR model:** Assuming model (2b), the SNP effect was estimated by:

$$275 \quad \hat{\beta}^{(2b)} = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' \hat{\mathbf{e}}^{(2a)} = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' (\mathbf{y} - \mathbf{1} \hat{\mu}^{(2a)} - \hat{\mathbf{u}}^{(2a)}).$$

276 The user assumed that $V^{(2)}(\hat{\beta}^{(2b)}) = (\mathbf{x}'\mathbf{x})^{-1} \sigma_{e^{(2b)}}^2$ and $E^{(2)}(\hat{\mathbf{e}}^{(2b)} \hat{\mathbf{e}}^{(2b)'}) = (n-2) \sigma_{e^{(2b)}}^2$. To develop

277 the correct formulae, we needed to know the expectation and variance of estimators of the

278 polygenic effects in the random model (2a). The mixed model equation of model (2a) could

279 be noted as:

$$280 \quad \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}' \\ \mathbf{1} & \mathbf{I} + \lambda^{(2a)} \mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu}^{(2a)} \\ \hat{\mathbf{u}}^{(2a)} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{y} \end{bmatrix} \text{ where } \lambda^{(2a)} = \frac{\sigma_{e^{(2a)}}^2}{\sigma_{u^{(2a)}}^2} \text{ and } \begin{bmatrix} \hat{\mu}^{(2a)} \\ \hat{\mathbf{u}}^{(2a)} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{11}^{(2a)} & \mathbf{C}_{1u}^{(2a)} \\ \mathbf{C}_{u1}^{(2a)} & \mathbf{C}_{uu}^{(2a)} \end{bmatrix} \begin{bmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{y} \end{bmatrix}.$$

281 Then, assuming that \mathbf{y} followed the true model:

$$282 \quad E(\hat{\mathbf{u}}^{(2a)}) = (\mathbf{C}_{u1}^{(2a)} \mathbf{1}' + \mathbf{C}_{uu}^{(2a)}) E(\mathbf{y}) = (\mathbf{C}_{u1}^{(2a)} \mathbf{1}' + \mathbf{C}_{uu}^{(2a)}) (\mathbf{1}\mu + \mathbf{x}\beta) = \mathbf{C}_{uu}^{(2a)} \mathbf{x}\beta.$$

283 The estimates of the polygenic effects were biased, and:

$$284 \quad V(\hat{\mathbf{u}}^{(2a)}) = (\mathbf{C}_{u1}^{(2a)} \mathbf{1}' + \mathbf{C}_{uu}^{(2a)}) V(\mathbf{y}) (\mathbf{C}_{u1}^{(2a)} \mathbf{1}' + \mathbf{C}_{uu}^{(2a)})' \\ = \sigma_u^2 (\mathbf{A} - \lambda^{(2a)} \mathbf{C}_{uu}^{(2a)}) + (\sigma_e^2 - \lambda^{(2a)} \sigma_u^2) \mathbf{C}_{uu}^{(2a)} (\mathbf{I} - \lambda^{(2a)} \mathbf{A}^{-1} \mathbf{C}_{uu}^{(2a)})$$

285 So that when computing the expectation of estimator of the SNP effect:

$$286 \quad E(\hat{\beta}^{(2b)}) = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' E(\mathbf{y} - \mathbf{1} \hat{\mu}^{(2a)} - \hat{\mathbf{u}}^{(2a)}) = \beta - (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} \beta$$

287 The estimator of the SNP effect was biased,

$$288 \quad V(\hat{\beta}^{(2b)}) = \sigma_e^2 \left[(\mathbf{x}'\mathbf{x})^{-1} - (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} (\mathbf{x}'\mathbf{x})^{-1} \right] - (\sigma_e^2 - \lambda^{(2a)} \sigma_u^2) \lambda^{(2a)} (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{A}^{-1} \mathbf{C}_{uu}^{(2a)} \mathbf{x} (\mathbf{x}'\mathbf{x})^{-1} \\ 289 \quad \text{(IV)}$$

290 and the residual variance was estimated using the sum of squares of residuals:

$$\begin{aligned}
291 \quad E(\hat{\tau}^{(2b)}) &= \sigma_e^2 (n-2) \tau + (\mathbf{C}_{uu}^{(2a)})^{-1} (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x}' (\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}' \mathbf{C}_{uu}^{(2a)} \mathbf{1} \\
&+ \beta^2 \mathbf{x}' \mathbf{C}_{uu}^{(2a)} (\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}') \mathbf{C}_{uu}^{(2a)} \mathbf{x} \\
&- (\sigma_e^2 - \lambda^{(2a)} \sigma_u^2) \lambda^{(2a)} \text{tr}((\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}') \mathbf{C}_{uu}^{(2a)} \mathbf{A}^{-1} \mathbf{C}_{uu}^{(2a)}) \quad (V)
\end{aligned}$$

292 Hence,

$$293 \quad E(\tau^{(2b)}) = (\beta - (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} \beta) \frac{\sqrt{(n-2) \mathbf{x}'\mathbf{x}}}{\sqrt{E(\hat{\epsilon}^{(2b)'} \hat{\epsilon}^{(2b)})}}$$

$$294 \quad V(\tau^{(2b)}) = \left(\sigma_e^2 [1 - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} (\mathbf{x}'\mathbf{x})^{-1}] - (\sigma_e^2 - \lambda^{(2a)} \sigma_u^2) \lambda^{(2a)} \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{A}^{-1} \mathbf{C}_{uu}^{(2a)} \mathbf{x} (\mathbf{x}'\mathbf{x})^{-1} \right) \frac{(n-2)}{E(\hat{\epsilon}^{(2b)'} \hat{\epsilon}^{(2b)})}$$

295 **Model 3, FASTA model:** The only difference with the true model was the variance

296 components used. These variance components were the same as in the GRAMMAR model.

297 The mixed model equation for model (3) was:

$$298 \quad \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{x} & \mathbf{1}' \\ \mathbf{x}'\mathbf{1} & \mathbf{x}'\mathbf{x} & \mathbf{x}' \\ \mathbf{1} & \mathbf{x} & \mathbf{I} + \lambda^{(2a)} \mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu}^{(3)} \\ \hat{\beta}^{(3)} \\ \hat{\mathbf{u}}^{(3)} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{x}'\mathbf{y} \\ \mathbf{y} \end{bmatrix} \quad \text{with } \lambda^{(2a)} = \frac{\sigma_e^{2(2a)}}{\sigma_u^{2(2a)}}, \text{ from the first model (2a) used in}$$

299 GRAMMAR.

300 If we note

$$301 \quad \begin{bmatrix} \mathbf{C}_{11}^{(3)} & \mathbf{C}_{1\beta}^{(3)} & \mathbf{C}_{1u}^{(3)} \\ \mathbf{C}_{\beta 1}^{(3)} & \mathbf{C}_{\beta\beta}^{(3)} & \mathbf{C}_{\beta u}^{(3)} \\ \mathbf{C}_{u1}^{(3)} & \mathbf{C}_{u\beta}^{(3)} & \mathbf{C}_{uu}^{(3)} \end{bmatrix} \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{x} & \mathbf{1}' \\ \mathbf{x}'\mathbf{1} & \mathbf{x}'\mathbf{x} & \mathbf{x}' \\ \mathbf{1} & \mathbf{x} & \mathbf{I} + \lambda^{(2a)} \mathbf{A}^{-1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \mathbf{0} \\ 0 & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}$$

302 The estimator of the SNP effect was:

$$303 \quad \hat{\beta}^{(3)} = (\mathbf{C}_{\beta 1}^{(3)} \mathbf{1}' + \mathbf{C}_{\beta\beta}^{(3)} \mathbf{x}' + \mathbf{C}_{\beta u}^{(3)}) \mathbf{y}$$

304 The user assumed that $V^{(3)}(\hat{\beta}^{(3)}) = \mathbf{C}_{\beta\beta}^{(3)} \sigma_{\epsilon^{(3)}}^2$, and used the sum of products between

305 performances and residuals to estimate the residual variance as usual in mixed models

306 assuming $E^{(3)}(\mathbf{y}' \hat{\epsilon}^{(3)}) = (n-2) \sigma_{\epsilon^{(3)}}^2$. Then, the expectation and variance of the estimator of the

307 SNP effect, assuming true model for \mathbf{y} , were:

$$308 \quad E(\hat{\beta}^{(3)}) = E(\mathbf{C}_{\beta 1}^{(3)} \mathbf{1}' \mathbf{y} + \mathbf{C}_{\beta \beta}^{(3)} \mathbf{x}' \mathbf{y} + \mathbf{C}_{\beta u}^{(3)} \mathbf{y}) = \beta$$

$$309 \quad V(\hat{\beta}^{(3)}) = \sigma_e^2 \mathbf{C}_{\beta \beta}^{(3)} - (\sigma_e^2 - \lambda^{(2a)} \sigma_u^2) \lambda^{(2a)} \mathbf{C}_{\beta u}^{(3)} \mathbf{A}^{-1} \mathbf{C}_{u \beta}^{(3)}.$$

310 And, after some matrix algebra, knowing $\mathbf{C}_{\beta u}^{(3)} = -\mathbf{C}_{\beta \beta}^{(3)} \mathbf{x}' \mathbf{C}_{uu}^{(2a)}$, $\mathbf{C}_{\beta \beta}^{(3)} = [\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x}]^{-1}$ and

$$311 \quad \lambda^{(2a)} \mathbf{C}_{uu}^{(2a)} \mathbf{A}^{-1} = \mathbf{I} - \mathbf{C}_{u1}^{(2a)} \mathbf{1}' - \mathbf{C}_{uu}^{(2a)},$$

$$312 \quad V(\hat{\beta}^{(3)}) = \sigma_e^2 (\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x})^{-1} - (\sigma_e^2 - \lambda^{(2a)} \sigma_u^2) ((\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x})^{-1})^2 (\mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{C}_{uu}^{(2a)} \mathbf{x})$$

$$313 \quad E(\hat{\epsilon}^{(3)'} \mathbf{y}) = (n-2) \sigma_e^2 - \left(\sigma_e^2 - \sigma_u^2 \lambda^{(2a)} \right) \left(\text{tr}(\mathbf{C}_{uu}^{(2a)}) - \frac{1}{n} \mathbf{1}' \mathbf{C}_{uu}^{(2a)} \mathbf{1} - \frac{\left(\mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} + \frac{1}{n} \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{1} \mathbf{1}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{C}_{uu}^{(2a)} \mathbf{x} \right)}{(\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x})} \right)$$

314 Hence,

$$315 \quad E(\tau^{(3)}) = \beta \sqrt{\frac{\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x}}{E(\hat{\epsilon}^{(3)'} \mathbf{y})}} \frac{\sqrt{(n-2)}}{\sqrt{E(\hat{\epsilon}^{(3)'} \mathbf{y})}} \quad (\text{VI})$$

$$316 \quad V(\tau^{(3)}) = \left(\sigma_e^2 - (\sigma_e^2 - \lambda^{(2a)} \sigma_u^2) \frac{(\mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{C}_{uu}^{(2a)} \mathbf{x})}{(\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x})} \right) \left(\frac{n-2}{E(\hat{\epsilon}^{(3)'} \mathbf{y})} \right)$$

317 **Model 4, QTDT model:** Assuming model (4), there were two regression coefficients to be

318 estimated. They were:

$$319 \quad \begin{bmatrix} \hat{\mu}^{(4)} \\ \hat{\beta}_b^{(4)} \\ \hat{\beta}_w^{(4)} \end{bmatrix} = \begin{bmatrix} \mathbf{1}' \mathbf{1} & \mathbf{1}' (\mathbf{z} - \mathbf{1} \bar{z}) & \mathbf{1}' (\mathbf{x} - \mathbf{z}) \\ (\mathbf{z} - \mathbf{1} \bar{z})' \mathbf{1} & (\mathbf{z} - \mathbf{1} \bar{z})' (\mathbf{z} - \mathbf{1} \bar{z}) & (\mathbf{z} - \mathbf{1} \bar{z})' (\mathbf{x} - \mathbf{z}) \\ (\mathbf{x} - \mathbf{z})' \mathbf{1} & (\mathbf{x} - \mathbf{z})' (\mathbf{z} - \mathbf{1} \bar{z}) & (\mathbf{x} - \mathbf{z})' (\mathbf{x} - \mathbf{z}) \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}' \mathbf{y} \\ (\mathbf{z} - \mathbf{1} \bar{z})' \mathbf{y} \\ (\mathbf{x} - \mathbf{z})' \mathbf{y} \end{bmatrix}.$$

$$320 \quad \text{If we note } \hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\mu}^{(4)} \\ \hat{\beta}_b^{(4)} \\ \hat{\beta}_w^{(4)} \end{bmatrix} = (\mathbf{Q}' \mathbf{Q})^{-1} \mathbf{Q}' \mathbf{y} \text{ with } \mathbf{Q} = [\mathbf{1} \quad (\mathbf{z} - \mathbf{1} \bar{z}) \quad (\mathbf{x} - \mathbf{z})]$$

321 The variance of the estimators assumed by the user were $V^{(4)}(\hat{\boldsymbol{\theta}}) = (\mathbf{Q}' \mathbf{Q})^{-1} \sigma_{\epsilon^{(4)}}^2$ and

$$322 \quad E^{(4)}(\hat{\boldsymbol{\epsilon}}^{(4)'} \hat{\boldsymbol{\epsilon}}^{(4)}) = (n-3) \sigma_{\epsilon^{(4)}}^2$$

323 The expectation and variance of estimates of the regression coefficient were in fact:

$$324 \quad E(\hat{\theta}) = (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'E(\mathbf{y}) = (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'(\mathbf{x}\beta + \mathbf{1}\mu)$$

$$325 \quad V(\hat{\theta}) = (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{A}\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\sigma_u^2 + (\mathbf{Q}'\mathbf{Q})^{-1}\sigma_e^2$$

326 And the sum of squares of residuals:

$$327 \quad \begin{aligned} E(\hat{\mathbf{e}}^{(4)'}\hat{\mathbf{e}}^{(4)}) &= tr((\mathbf{I} - \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}')V(\mathbf{y})) + E(\mathbf{y})'(\mathbf{I} - \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}')E(\mathbf{y}) \\ &= (n-3)\sigma_e^2 + (tr(\mathbf{A}) - tr(\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{A}))\sigma_u^2 + \mu^2(n - \mathbf{1}'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{1}) \\ &\quad + \beta^2(\mathbf{x}'\mathbf{x} - \mathbf{x}'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{x}) - \mu\beta(\mathbf{1}'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{x} + \mathbf{x}'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{1}) \end{aligned}$$

328 So that, the test on $\hat{\beta}_w^{(4)}$, the within-family regression, was:

$$329 \quad E(\tau^{(4)}) = \frac{[(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'(\mathbf{x}\beta + \mathbf{1}\mu)]_{3,1}}{\sqrt{[(\mathbf{Q}'\mathbf{Q})^{-1}]_{3,3}}} \frac{\sqrt{(n-3)}}{\sqrt{E(\hat{\mathbf{e}}^{(4)'}\hat{\mathbf{e}}^{(4)})}}$$

$$330 \quad V(\tau^{(4)}) = \frac{[(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{A}\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\sigma_u^2 + (\mathbf{Q}'\mathbf{Q})^{-1}\sigma_e^2]_{3,3}}{[(\mathbf{Q}'\mathbf{Q})^{-1}]_{3,3}} \frac{(n-3)}{E(\hat{\mathbf{e}}^{(4)'}\hat{\mathbf{e}}^{(4)})}$$

331 Where $[\mathbf{M}]_{3,3}$ denotes the coefficient of line 3 and column 3 of matrix \mathbf{M}

332 **True model:** With the true model, the classical formulae were:

$$333 \quad E(\hat{\beta}) = \beta,$$

$$334 \quad V(\hat{\beta}) = \sigma_e^2 \mathbf{C}_{\beta\beta},$$

$$335 \quad E(\hat{\mathbf{e}}'\mathbf{y}) = (n-2)\sigma_e^2,$$

$$336 \quad E(\tau) = \frac{\beta}{\sigma_e \sqrt{\mathbf{C}_{\beta\beta}}}$$

$$337 \quad V(\tau) = 1$$

338 ***Marginal expectation and variance of test statistics according to the***
 339 ***distribution of genotypes and estimators of variance components***

340 The above formulae gave the conditional expectation of the estimators of the SNP effects and
 341 the conditional expectation and variance of test statistics based on specific data, that is given
 342 \mathbf{w} , the genotypes of the data (or \mathbf{x} , the centred genotypes defined in the true model) and given
 343 the known variance components of the polygenic effects. They may be applied to any kind of
 344 data.

345 The aim of this part of the study was to calculate the marginal expectation and variance of test
 346 statistics, by integrating over the distribution of genotypes and of variance components of
 347 random polygenetic effects given the relationship matrix and variance components of true
 348 model. In that case, the quadratic forms involving \mathbf{x} and \mathbf{z} and the variance components of the
 349 random model (2) were replaced by their expectation. If E_x denotes these expectations and a_{ij}
 350 is defined as the relationship coefficient between animals i and j , then the coefficient of
 351 Mendelian Sampling variance d_{ii} can be defined as:

$$352 \quad d_{ii} = a_{ii} - \left(\frac{1}{4} a_{s_i s_i} + \frac{1}{4} a_{d_i d_i} + \frac{1}{2} a_{s_i d_i} \right) \text{ where } s_i \text{ the sire of animal } i \text{ and } d_i \text{ the dam.}$$

353 \mathbf{D} is the diagonal matrix with elements d_{ii} . We know that (Habier et al., 2007), assuming
 354 Hardy Weinberg equilibrium,

$$355 \quad E_x(w_i w_j) = a_{ij}, \quad E_x(w_i w_i) = a_{ii} \quad \text{and} \quad E_x(w_i z_i) = E_x(z_i z_i) = a_{ii} - d_{ii} \quad \text{and} \quad E_x(z_i z_j) = a_{ij},$$

356 when the genotype, \mathbf{w} , was expressed in a standardized form as shown in introduction. So
 357 that:

$$358 \quad E_x(\mathbf{x}'\mathbf{x}) = tr(\mathbf{A}) - \frac{1}{n} \mathbf{1}'\mathbf{A}\mathbf{1}$$

$$359 \quad E_x(\mathbf{x}'\mathbf{A}\mathbf{x}) = tr(\mathbf{A}\mathbf{A}) - \frac{2}{n} \mathbf{1}'\mathbf{A}\mathbf{A}\mathbf{1} + \frac{1}{n^2} \mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1}$$

$$360 \quad E_x(\mathbf{x}'\mathbf{C}_{uu}^{(2a)}\mathbf{x}) = tr(\mathbf{A}) - \lambda^{(2a)}tr(\mathbf{C}_{uu}^{(2a)}) + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2a)}\mathbf{1} - \frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2a)}\mathbf{1}$$

$$361 \quad E_x(\mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{x}) = tr(\mathbf{A}) - \lambda^{(2)}tr(\mathbf{C}_{uu}^{(2)}) - \lambda^{(2a)}tr(\mathbf{C}_{uu}^{(2a)}\mathbf{C}_{uu}^{(2a)}) + \frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2a)}\mathbf{1} - \frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2a)}\mathbf{C}_{uu}^{(2a)}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2a)}\mathbf{C}_{uu}^{(2a)}\mathbf{1}$$

$$362 \quad E_x(\mathbf{x}'\mathbf{C}_{uu}^{(2a)}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2a)}\mathbf{x}) = \mathbf{1}'\mathbf{A}\mathbf{1} - \lambda^{(2a)}\mathbf{1}'\mathbf{C}_{uu}^{(2a)}\mathbf{1} - \lambda^{(2a)}\mathbf{1}'\mathbf{C}_{uu}^{(2a)}\mathbf{C}_{uu}^{(2a)}\mathbf{1} + \frac{1}{n}\lambda^{(2a)}\mathbf{1}'\mathbf{C}_{uu}^{(2a)}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2a)}\mathbf{1}$$

363 and for the sums involved in QTDT (as in Abecasis *et al.* 2000a):

$$364 \quad E_x(\mathbf{Q}'\mathbf{Q}) = \begin{bmatrix} n & 0 & 0 \\ 0 & tr(\mathbf{A}) - \frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{1} - tr(\mathbf{D})\frac{n-1}{n} & 0 \\ 0 & 0 & tr(\mathbf{D}) \end{bmatrix}$$

$$365 \quad E_x(\mathbf{Q}'\mathbf{x}) = \begin{bmatrix} 0 \\ tr(\mathbf{A}) - \frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{1} - tr(\mathbf{D})\frac{n-1}{n} \\ tr(\mathbf{D})\frac{n-1}{n} \end{bmatrix}$$

$$366 \quad E_x(\mathbf{Q}'\mathbf{1}) = \begin{bmatrix} n \\ 0 \\ 0 \end{bmatrix}$$

$$367 \quad E_x(\mathbf{Q}'\mathbf{A}\mathbf{Q}) = \begin{bmatrix} \mathbf{1}'\mathbf{A}\mathbf{1} & 0 & 0 \\ 0 & tr(\mathbf{A}'\mathbf{A}) - tr(\mathbf{A}\mathbf{D}) - \frac{2}{n}(\mathbf{1}'\mathbf{A}'\mathbf{A}\mathbf{1} - \mathbf{1}'\mathbf{A}\mathbf{D}\mathbf{1}) + \frac{1}{n^2}(\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1} - \mathbf{1}'\mathbf{A}\mathbf{1}tr(\mathbf{D})) & 0 \\ 0 & 0 & tr(\mathbf{A}\mathbf{D}) \end{bmatrix}$$

368

369 These expectations replaced their corresponding terms in the preceding formulae in order to
 370 express all expectations and variances of the tests used for detection of the SNP effect in
 371 function of heritability and the relationship matrix. To do so, an approximation was made: the
 372 replacement of expectation in quotients and products by quotients and products of
 373 expectations. Concerning the expectation of variance components given relationships, the
 374 following expectations were used:

$$375 \quad E_x(\sigma_{u^{(2)}}^2) = \sigma_u^2 + \beta^2,$$

$$376 \quad E_x(\sigma_{e^{(2)}}^2) = \sigma_e^2,$$

$$377 \quad E_x(\lambda^{(2)}) = \frac{\sigma_e^2}{\sigma_u^2 + \beta^2}.$$

378 **Application**

379 The formulae may be applied to any kind of data when the relationship matrix is known,
 380 without any simulations. The results presented here were an illustration assuming 600
 381 recorded and genotyped progenies belonging to 120, 20 and 10 families of respectively $n_d=5$,
 382 30 and 60 half sibs as often occurs in animal analysis. The power was calculated for a
 383 regression coefficient of 0.14 in phenotypic standard deviation (or 2% of phenotypic variance)
 384 which was equivalent to an allele substitution effect of 0.20 for a minimum allele frequency
 385 (MAF) of 50% or an effect of 0.33 for a MAF of 10%. Then, the variation of some of the
 386 parameters used in the example was analyzed: total number of animals, estimates of variance
 387 components used in GRAMMAR and FASTA.

388 For families of half sibs, the preceding formulae were calculated using:

$$389 \quad c_{ii} = \frac{h^2 [16 + h^2(3n_d - 7)]}{(4 - h^2)(4 + h^2(n_d - 1))} + \frac{(h^2(n_d + 3))^2}{4n(1 - h^2)(4 + h^2(n_d - 1))}, \quad \text{the diagonal term of } \mathbf{C}_{uu}^{(2)},$$

$$390 \quad c_{ij} = \frac{4(1 - h^2)h^2}{(4 - h^2)(4 + h^2(n_d - 1))} + \frac{(h^2(n_d + 3))^2}{4n(1 - h^2)(4 + h^2(n_d - 1))}, \quad \text{the off-diagonal term of } \mathbf{C}_{uu}^{(2)} \text{ between}$$

$$391 \quad \text{half sibs, } c_{ij} = \frac{(h^2(n_d + 3))^2}{4n(1 - h^2)(4 + h^2(n_d - 1))} \text{ the off diagonal term of } \mathbf{C}_{uu}^{(2)} \text{ between animals from}$$

392 different families. The diagonal coefficients of the relationship matrix \mathbf{A} were 1, off-diagonal
 393 coefficients were $\frac{1}{4}$ between half sibs and 0 elsewhere. The matrix \mathbf{D} was diagonal with
 394 diagonal coefficients $\frac{1}{2}$. It should be noted that

$$395 \quad \lambda^{(2)} \mathbf{C}_{uu}^{(2)} \mathbf{A}^{-1} \mathbf{C}_{uu}^{(2)} = \mathbf{C}_{uu}^{(2)} - \mathbf{C}_{uu}^{(2)} \mathbf{C}_{uu}^{(2)} + \frac{1}{n} \mathbf{C}_{uu}^{(2)} \mathbf{1} \mathbf{1}' \mathbf{C}_{uu}^{(2)} = \mathbf{C}_{uu}^{(2)} - \mathbf{C}_{uu}^{(2)} \mathbf{C}_{uu}^{(2)} \text{ with families of equal sizes.}$$

396

RESULTS

397 **Robustness**

398 For an assumed type 1 error of 1%, the expected true type 1 error is plotted in figures *1a* to *1d*
 399 according to the heritability of the polygenic effect and the number of half-sibs in each family
 400 for the same overall number of genotyped animals (600). For the REGRESSION model, the
 401 type 1 error increased to a large extent with heritability and family size: 12% with $h^2 = 0.50$
 402 and families of 60 half sibs. For the GRAMMAR model, the type 1 error decreased with
 403 heritability and family size. FASTA and QTDT models were practically unaffected by
 404 polygenic variance and relationships in the sample analyzed.

405 **Power**

406 Figures *2a* to *2d* show the power of the methods. For the REGRESSION method, the power
 407 decreased with heritability and family size. For both the FASTA and GRAMMAR methods, the
 408 power first decreased to a minimum at a heritability of about 0.30 and then increased with
 409 heritability to tend to a value equal to the power obtained with a heritability of 0. The power
 410 was always higher with lower family sizes. The power of the QTDT method was unaffected by
 411 population structure and a polygenic effect but was very low compared to the other models.
 412 The power was also calculated for the same true type I error (figure 3). In that case, the power
 413 of the REGRESSION model was always lower than that of the FASTA model which was equal to
 414 the power of the GRAMMAR model. In the figure the power of the true mixed model was not
 415 represented because it was almost confounded with power of the FASTA model except for very
 416 low heritability values and high family sizes (for example, for $h^2=0.10$ and a family size of 60

417 half sibs, the power for the FASTA model and the true mixed model were 73.2% and 73.3%,
418 respectively).

419 ***Bias***

420 The only biased estimator was the one used with the GRAMMAR model and is plotted in figure
421 4. The value of the SNP effect was underestimated and the bias grew highly as heritability
422 increased (-56% for $h^2 = 0.50$ and families of 60 half sibs).

423 ***Effect of the total size of the sample on robustness***

424 The total sample size did not cause the robustness to deviate greatly from assumed robustness
425 when compared to the effect of family size. This small effect emphasized the underestimation
426 of the type 1 error for the REGRESSION method and the overestimation of the type 1 error
427 for the GRAMMAR method. For example, for $h^2 = 0.50$, families of 60 half sibs and an
428 assumed type 1 error of 1%, a total sample of 600 animals gave a 11.9% true type 1 error for
429 the regression method compared to 12.6% with a total sample of 6000 animals. With the same
430 structure, the type 1 error for the GRAMMAR method was 0.38% with 600 animals and
431 0.35% with 6000 animals.

432 ***Effect of the variance components introduced in the GRAMMAR and FASTA*** 433 ***methods***

434 The final models in both the GRAMMAR and FASTA methods use variance components that
435 are estimated with the same simple random model. Results presented in previous section were
436 marginal expectations using the distribution of the estimator of the variance components. But
437 what happens if the variance components are not estimated on the same sample as used to
438 estimate the SNP effect? Heritability may be introduced in the model if it is considered that
439 better evaluation was obtained with other data. Figure 5 presents the type 1 error for the
440 GRAMMAR and FASTA methods assuming the true heritability was 0.30 but that the

441 heritability used in model (2b) and (3) was under- or overestimated from 0.05 to 0.55. In the
442 case of underestimated heritability, the type 1 error increased as the heritability decreased.
443 Even with the very conservative test of the GRAMMAR model, the type 1 error reached 1.9%
444 with a high family size (60 half sibs) when heritability was much smaller than the true value
445 (0.05 against 0.30) when 1% was assumed. The type 1 error of the FASTA test, normally
446 practically unaffected, also deviated from the expected value with an error of 2.5% for the
447 same parameters.

448 **Protocol design**

449 Our algebraic results are a tool for designing populations or estimating the success of a given
450 design before starting the genotyping process. FASTA statistics, which are not subject to first
451 type errors due to genetic stratification, should be retained for this purpose.

452 As shown in figure 6, the method's power is largely dependent on population size, or the total
453 number of observed individuals. Even if it is only marginally influenced by the way that the
454 population is organized in families and the heritability of the trait, the experimenter may be
455 limited (by, for example, budgetary reasons) to a fixed total size and may only adjust family
456 structure. Figure 7 shows how the total population size should be adjusted according to family
457 structure for a moderate SNP effect of 2% of phenotypic variance, showing a difference of
458 183 individuals between the least and most favorable situations in order to obtain a power of
459 80%. Even if this difference seems moderate, 183 individuals represented a fifth of the
460 genotyping costs.

461 **Link to Genomic Control**

462 A very common measure of the deviation of a test's empirical distribution to its theoretical
463 distribution in association studies is the Genomic Control (GC, Devlin and Roeder 1999). As
464 pointed out by Bacanu *et al.* (2002), the GC inflation factor may be interpreted in the case of

465 multiple Student T tests on quantitative traits as the variance of the normal distribution
466 approximately followed the T test. Even if, as presented here, the test's distribution
467 expectation is influenced, under H1, by structures in the data, its variance is closely related to
468 the GC measure. Figure 8 presents, as a function of heritability and family structure, the GC
469 measure as approximated by this variance. It clearly shows that inflation is very limited for
470 the GRAMMAR method, and may be important for the REGRESSION method when families
471 are large and heritability high.

472

DISCUSSION

473 ***Correctness of algebra***

474 Details of algebra were given in supplementary material 1. Several approximations were used
475 in algebra, notably:

- 476 - ignoring the variance of the estimator of the SNP effect due to the estimation of the
477 variance component instead of true parameters (Kenward and Roger 1997),
- 478 - replacing quadratic forms by their expectations in products and ratios.

479 Therefore, simulations were first performed to validate the formulae (supplementary material
480 2). All simulations showed a very good coherence with analytical results, so that all results in
481 the paper were given strictly from the analytical formulae. An R program was given in
482 supplementary material 3 to compute type I error and power for the 4 methods with any
483 relationship matrix and heritability.

484 ***How does each method work?***

485 The formulae given in material and method of this paper were not easy to interpret. In that
486 section, we investigated which parameters principally influenced the variation of expectation
487 and variance of the test statistic in each of the four models and then produced the variation of

488 type 1 error and power in function of sample relationships and heritability of the trait. To do
 489 so, the formulae were approximated to their main components and the matrices in the
 490 formulae were translated into common meaning terms.

491 As presented in materiel and methods, the difference between the actual and expected type I
 492 errors at the choice of a threshold was due first, to deviation of the expectation of the test used
 493 $E(\tau^{(i)})$ from the assumed expectation $E(t^{(i)})$ and second, to deviation of the variance of the
 494 test used $V(\tau^{(i)})$ from the assumed variance which was equal to 1. In the explanations, we
 495 used the equations I and II which gave $E(\tau^{(i)})$ and $V(\tau^{(i)})$. According to equation II, the
 496 deviation of $V(\tau^{(i)})$ from 1 was due to the product of the ratio between the true variance of $\hat{\beta}$
 497 ($V(\hat{\beta}^{(i)})$) and the assumed variance ($V^{(i)}(\hat{\beta}^{(i)})$) with the ratio between the assumed
 498 expectation of $\hat{\sigma}_e^2$ ($E^{(i)}(\hat{\sigma}_{e^{(i)}}^2)$) and true expectation ($E(\hat{\sigma}_{e^{(i)}}^2)$).

499 **In the regression method:** under H_0 ($\beta = 0$), the expectation of the test $E(\tau^{(i)})$ was equal to
 500 0 as assumed by the user. But the variance of the test $V(\tau^{(i)})$ was greater than 1 principally
 501 because the variance of $\hat{\beta}$ was higher than that assumed ($(\mathbf{x}'\mathbf{x})^{-1}\sigma_{e^{(i)}}^2$). The difference between
 502 the true and assumed expectation of $\hat{\sigma}_e^2$ ($E(\hat{\sigma}_{e^{(i)}}^2)$ and $E^{(i)}(\hat{\sigma}_{e^{(i)}}^2)$) was negligible. The fact
 503 that the variance of the test was greater than one explained why the type 1 error was higher
 504 than that expected. The high $V(\hat{\beta})$ was due to the probability for 2 half sibs to share the same
 505 SNP because of their relationship. If a polygenic effect occurs, this local similarity at the SNP
 506 and the similarity between performances due to the polygenic effect were confounded with a
 507 true SNP effect. This only happens by chance so the test was not biased but the variance of $\hat{\beta}$
 508 was increased. This influence was directly linked to the variance of the coefficients of the
 509 relationship matrix. The term involved in the increase of $V(\hat{\beta})$ (see III) was $E_x(\mathbf{x}'\mathbf{A}\mathbf{x})$,

510 which is related to the variance of the coefficients of the relationship matrix **A**. Therefore,
 511 high relationships between genotyped animals were not a factor of increase of the type 1
 512 error, it was the mixture of high and low relationships which impacted the type 1 error. This is
 513 the case with independent large families (half sibs, full sibs). This effect on the variance of the
 514 test was proportional to the ratio between the polygenic variance and residual variance ($\frac{h^2}{1-h^2}$
 515 , III), so increased exponentially with heritability. The increase of the type 1 error with
 516 heritability and family size did not systematically imply an increase of power. Under the H1
 517 hypothesis ($\beta = b$), the variance of the test was still higher than 1 and increased with
 518 heritability while the expectation of the test did not vary greatly with heritability (only
 519 accordingly to a slight variation in the expectation of the estimation of error variance). So the
 520 abscissa of the normal curve used to calculate the power ($\frac{t_{\alpha/2} - E_{\beta=b}(\tau^{(1)})}{\sqrt{V_{\beta=b}(\tau^{(1)})}}$) increased with
 521 heritability and the variance of relationships when the threshold chosen for type I error ($t_{\alpha/2}$)
 522 was lower than the expectation of the test ($E_{\beta=b}(\tau^{(1)})$), (power over 50%), or in other terms,
 523 when this value was negative. In this case, the lower part of the normal curve should be higher
 524 than the chosen threshold as heritability increases. This explained why the power decreased
 525 with heritability and the variance of relationships.

526 **In the grammar model:** under the H0 hypothesis ($\beta = 0$), the expectation of the test
 527 $E(\tau^{(2b)})$ was equal to 0 even if the test was biased, as expected by the user. But the variance
 528 of the test was less than 1 and decreased when heritability and family size increased. The
 529 variance of $\hat{\beta}$ and the expectation of $\hat{\sigma}_e^2$ both decreased when heritability increased, and, as
 530 they acted as a ratio (II), this decrease in the variance of the test was due to the relatively
 531 higher decrease of the variance of $\hat{\beta}$ compared to the decrease of expectation of $\hat{\sigma}_e^2$. The

532 decrease of $E(\hat{\sigma}_e^2)$ could be explained in the following way: The GRAMMAR method uses
 533 performances corrected for the estimates of polygenic effects. So the variance of this new
 534 phenotype was the residual variance of performance minus the part of genetic variance not
 535 explained by the estimates of polygenic effects. If polygenic effects were known (reliability of
 536 1), this variance would be the total residual variance, but since the reliability of the polygenic
 537 effect was not 1, the residual variance was decreased by (1-reliability) the genetic variance.
 538 The reliability was here defined as the square correlation between true and estimated
 539 polygenic effects. This impacted directly the expectation of the sum of squared residuals used
 540 to estimate σ_e^2 (see V) and was the meaning of
 541 $E(\hat{\mathbf{e}}^{(2b)} \cdot \hat{\mathbf{e}}^{(2b)}) = \sigma_e^2 (n-2 - tr(\mathbf{C}_{uu}^{(2a)})) \approx (n-2) (\sigma_e^2 - (1-reliability)\sigma_a^2)$ which could be
 542 expressed as: $E(\hat{\mathbf{e}}^{(2b)} \cdot \hat{\mathbf{e}}^{(2b)}) \approx (n-2)\sigma_e^2 \left(1 - \frac{h^2}{1-h^2} (1-reliability) \right)$ (where *reliability* is the mean
 543 of the reliabilities of the estimates of polygenic values of all individuals). Compared to this
 544 major influence, the remaining terms in $E(\hat{\mathbf{e}}^{(2b)} \cdot \hat{\mathbf{e}}^{(2b)})$ (see V), i.e. mean terms and under H1
 545 hypothesis, the effect of error on the estimation of variance components due to the use of a
 546 pure random model in the estimation of heritability (factor of $(\sigma_e^2 - \lambda^{(2a)}\sigma_u^2) = \sigma_u^2(\lambda - \lambda^{(2a)})$)
 547 and the effect of SNP variance (factor of β^2) were negligible on the expectation of $\hat{\sigma}_e^2$. So,
 548 as the heritability increased, $E(\hat{\mathbf{e}}^{(2b)} \cdot \hat{\mathbf{e}}^{(2b)})$ decreased (as (1-reliability) was always positive)
 549 compared to the assumed $(n-2)\sigma_e^2$. For the variance of $\hat{\beta}$, the decrease in function of
 550 heritability was also due to the decrease of new phenotypic variance but not only. If the
 551 variance of $\hat{\beta}$ was only influenced by the variance of the new phenotype, this expected
 552 variance would be this new residual variance divided by the sum of squared genotypes, i.e. n
 553 in expectation when genotypes are standardized. And the variance of the test would be 1. But,

554 in fact, the quadratic form $\mathbf{x}'\mathbf{C}_{uu}^{(2a)}\mathbf{x}$ appeared in $V(\hat{\beta})$ (IV). For the expectation, this quadratic
 555 form was equal to $tr(\mathbf{C}_{uu}^{(2a)}V(\mathbf{x}))$ and nearly equal to $tr(\mathbf{C}_{uu}^{(2a)}\mathbf{A})$ and so, because this
 556 relationship matrix \mathbf{A} was the same as the one used for the estimation of polygenic effects,
 557 due to mixed model equations, we obtained $tr(\mathbf{C}_{uu}^{(2a)}\mathbf{A}) \approx tr(\mathbf{A}) - \lambda^{(2a)}tr(\mathbf{C}_{uu}^{(2a)}) \approx n(\text{reliability})$
 558 . This can be explained as follows: as there are covariances between the genotypes of related
 559 animals and, for the same animals, covariances between prediction errors of polygenic values,
 560 this quadratic form was greater than the simple sum of the variance of prediction errors and
 561 was equivalent to the sum of reliabilities. Hence, $V(\hat{\beta}) = \frac{1}{n}\sigma_e^2(1 - \text{reliability})$ different from
 562 the assumed $\frac{1}{n}\sigma_e^2$ (with $E_x(\mathbf{x}'\mathbf{x}) \approx tr(\mathbf{A}) = n$). Finally, the variance of the test was
 563 approximately, for moderate heritability, $V(\tau^{(2b)}) = \frac{1 - \text{reliability}}{1 - \frac{h^2}{1 - h^2}(1 - \text{reliability})}$ which finally
 564 explained why the variance of the test was lower than 1, and why the true type 1 error was
 565 lower than expected and decreased with heritability. For high heritability values, all terms
 566 must be taken into account to calculate $V(\tau^{(2b)})$ because both $V(\hat{\beta})$ and $E(\hat{\sigma}_e^2)$ tended to zero
 567 and the ratio was really instable. Concerning the power of the test, its decrease with
 568 heritability was more due to the variation of the expectation of the test under the H1
 569 hypothesis than to its variance. This is because the variation in the abscissa of the normal
 570 curve influenced more the integral used to calculate the power at this scale than the variance.
 571 The test was greatly biased and the reduction in the expectation of $\hat{\beta}$ was directly linked to
 572 (1-reliability). However this considerable bias did not have much impact on the power
 573 because fortunately, as mentioned previously, the residual variance of the new phenotype also
 574 decreased: if this was not the case the test would be very inefficient. But, as stated, this

575 decrease was function of $\left(1 - \frac{n}{1-h^2}(1-reliability)\right)$ which explained the decrease of the
576 expectation of the test with heritability and therefore the decrease in power. After decreasing
577 for high heritability values, the power then increased and reached a constant because of the
578 simultaneous decrease of the variance and expectation of the test τ . To summarize, the
579 variation of the type 1 error and power in respect to heritability was due to the relationships
580 between animals that were used simultaneously for estimation of the polygenic effect and of
581 the SNP effect, but not strictly to the existence of a polygenic effect. If most of the
582 phenotypes used to estimate the polygenic value of the animal belonged to animals without
583 genotypes and without relationships with the other genotyped animals, the GRAMMAR test
584 would not show these type I error and power patterns. This may be the case when analysing
585 unrelated genotyped sires whose phenotypes would be means of performances of
586 ungenotyped progeny. The estimator of the SNP effect would still be biased downwards but
587 the type 1 error and power would be practically unaffected by heritability and family
588 structure.

589 **The fasta model:** In this model the only difference with the true mixed model was the error
590 made on variance components since they were estimated with a pure random model.
591 Therefore, under the H0 hypothesis, i.e. without a SNP effect, the variance components were
592 the same and the type 1 error was not affected by the heritability of the trait or relationships
593 within the sample. Under the H1 hypothesis the influence of heritability on power was only
594 moderate and reserved to low to medium heritability values. In these cases, the effect of
595 heritability on power was mainly due to its effect on the expectation of test rather than on its
596 variance which was very close to 1. The expectation of the test varied with heritability not
597 only for the test used, $\tau^{(3)}$, but it also varied in a similar manner for the valid test $t^{(3)}$ and

598 depended on the assumed variance of $\hat{\beta}$, i.e. $V^{(3)}(\hat{\beta}^{(3)}) = C_{\beta\beta}^{(3)}\sigma_{e^{(3)}}^2$. Since the covariances
 599 between the genotypes of related animals and the coefficient representing the prediction error
 600 for the polygenic effect were the same as in the GRAMMAR model, the expectation of
 601 $C_{\beta\beta}^{(3)} = [\mathbf{x}'\mathbf{x} - \mathbf{x}'\mathbf{C}_{uu}^{(2a)}\mathbf{x}]^{-1}$ was $[n(1 - reliability)]^{-1}$. The expectation of $\mathbf{y}'\hat{\mathbf{e}}^{(3)}$ used to estimate
 602 the residual variance was very close to the assumed value $(n-2)\sigma_{e^{(3)}}^2$. Hence, according to
 603 (VI), and for low to moderate heritability values, $E(\tau^{(3)}) \approx (\beta / \sigma_y) \sqrt{\frac{n(1 - reliability)}{1 - h^2}}$. So,
 604 when the reliability was higher than the heritability, the expectation and the power of the test
 605 both decreased. This was particularly important when half sibs influenced the reliability so
 606 when heritability was low. As the heritability increased, the reliability tended to heritability so
 607 the power became less sensitive to variations in heritability and equalled the power observed
 608 without a polygenic effect. The error made on the estimation of heritability via the two-step
 609 procedure had only a very slight effect on these variations and was perceptible only at low
 610 heritability values for which the error made on the estimation of heritability was higher. This
 611 effect corresponds to a decrease of expectation due to an increase of the variance of $\hat{\beta}$
 612 imputable to an overestimation of h^2 .

613 **QTDT method:** As the QTDT method used the within families information, the variance of
 614 $\hat{\beta}_w$ was not affected by the relationships that existed between phenotyped animals in the data
 615 set. It depended solely on the trace of Mendelian sampling variance matrix and so eventually
 616 on inbreeding within in the data but not on relationships between phenotyped animals
 617 (whatever the form of families, assuming the genotypes of the parents were known). In
 618 expectation, integrating over genotypes given relationship matrix as developed in
 619 supplementary material 1, without inbreeding, $V(\hat{\beta}_w^{(4)}) = \frac{1}{n/2}(\sigma_u^2 + \sigma_e^2) = \frac{2}{n}\sigma_y^2$. So, the

620 polygenic effect had no influence on it. The only deviation from a model without a polygenic
621 effect concerned the expectation of residual variance. For the same reasons, residual variance
622 was more or less equal to the phenotypic variance and was only slightly reduced by
623 relationships. Hence, the type I error was practically unaffected by heritability and
624 relationships. For the same reasons, the power was unaffected by heritability and the
625 relationship matrix but was very much lower than in the other models because the test used
626 half the genetic variance (only the Mendelian sampling variance). This divided by a factor
627 $\sqrt{2}$ the expectation of the test: $E(\tau^{(4)}) \approx \sqrt{n/2}\beta / \sigma$, and decreased consequently the power.

628 ***Comparison between methods***

629 The type I error increased with relationships and heritability for the REGRESSION method,
630 decreased for the GRAMMAR method and was unaffected for the FASTA and QTDT methods.
631 The power calculated with a similar assumed type I error (and not an real type I error) was
632 higher for the REGRESSION method than for the FASTA method for low (large family sizes) to
633 moderate (small family sizes) heritability values. The power obtained with the GRAMMAR
634 method was always lower than with the FASTA method. But for a true identical type I error
635 (with the threshold chosen to reach the same true type I error), the power of the REGRESSION
636 method was always lower than that of the FASTA method and decreased very rapidly with
637 heritability and family size. In this situation, the power of the GRAMMAR method was exactly
638 the same as that of the FASTA method. Thus, these two latter methods, when corrected for the
639 error made on the calculation of the type I error, have the same power. The power of these
640 two methods was also practically identical to that of the true mixed model, except for very
641 low heritability values for which a very slight difference was observed between the FASTA and
642 true mixed models.

643 These results are in general agreement with the few papers found in the litterature. Aulchenko
644 *et al.* (2007a) demonstrated on a simple example with three pedigrees that the type 1 error of
645 the REGRESSION method increases with heritability and family size (from unrelated small
646 nuclear families to a mixture of half and full sib families in pig-type pedigrees) while the
647 contrary was observed with the GRAMMAR method, an observation fully consistent with our
648 figures 1a and 1b. The authors also classed the methods in respect to decreasing empirical
649 power: FASTA > GRAMMAR > REGRESSION > TDT, and found very limited differences in power
650 between the mixed model and the FASTA method. In a range of family sizes limited from 1 to
651 4, Zhang *et al.* (2009) found that the power of the QTDT method increases with this parameter,
652 a result compatible with the slight increase we observed in the range 5 to 60. Erbe *et al.*
653 (2011) confirmed that the GRAMMAR method provides a better control of the first type error
654 than the REGRESSION method, and found that in a populaiton of 500 progenies this error was
655 higher when they came from 25 rather than 250 sires.

656 Therefore, as a general result, we would not recommend the REGRESSION and GRAMMAR
657 models but do recommend the FASTA method which is very close to the full mixed model and
658 expected to be computationally faster. Nevertheless, when should the first two methods be
659 used? And when might use of the FASTA method become dangerous? To answer the first
660 question, it should be noted that the advantage of the REGRESSION model is that no heritability
661 is needed, so it may be interesting when the heritability is unknown or when the number of
662 animals is too low to estimate it in the sample. Another point is that this method may be
663 interesting in situations when a large type I error is not a problem. For example if the
664 objective is to first select markers before performing another type of analysis, the aim in this
665 kind of situation being to keep only the good markers, whatever the number of bad ones. The
666 advantage of the GRAMMAR method is that is shows the same power as the FASTA method
667 when corrected for type I error underestimation. This correction may be performed easily

668 using analytical formulae or by analyzing the QQ plot – in this case analysis would be really
669 fast, faster than the FASTA model. Moreover, when the GRAMMAR method uses a polygenic
670 effect obtained from another experiment and from animals without relationships with other
671 genotyped animals, the method is as efficient as the FASTA method. To answer the second
672 question, use of the FASTA (and GRAMMAR) method depends on the variance components that
673 are introduced. The difference between the expectation of heritability estimated in the pure
674 random model and the true one was small when the fixed SNP effect was small so that the
675 final influence was not important (the low performances of GRAMMAR were due to the use of
676 residuals, not the error on heritability). This explains why the FASTA method is close to the
677 full mixed model (type I and power). But what might happen if a variance component other
678 than the one estimated in the sample was used or if fortuitously the sample gave a variance
679 component very different from true one? What happens to the conditional distribution of the
680 test given false heritability? In this case, the coefficient in the GRAMMAR method involving
681 the difference in heritability is important and increases the variance of the test. The difference
682 between true and used heritabilities produces a high coefficient for low heritability and
683 increased variance and increased type I error. Since the GRAMMAR method is supposed to be a
684 very conservative method, the difference between the expected and obtained values may be
685 surprising. The FASTA method behaved similarly, but in any case, only a considerably
686 underestimated heritability produced only a moderate increase of the type 1 error. The true
687 power (for true type I error) was in this case reduced when h^2 was underestimated (-4% when
688 h^2 was supposed to be 0.10 instead of 0.30) but the decrease remained limited. Therefore it
689 would seem that the FASTA method works whatever the situation. When using the FASTA
690 method, underestimating the heritability was actually more risky than overestimating it (in
691 terms of type I error and power). But it should be kept in mind that the power of even the true

692 mixed model is lower for moderate heritability values than with 0 or 1, whatever the method
693 used.

694 **CONCLUSION**

695 The analytical formulae of distribution of the tests used to detect the SNP effect in four of the
696 most common models were given in the case of structured populations due to relationships
697 between individuals. These formulae attempted to compute the type 1 errors and power of
698 these methods given any kind of relationship matrix between phenotyped and genotyped data
699 in any situation of heritability of a polygenic effect, the aim of the study being that they can
700 be easily used to give the correct threshold of type 1 error and to calculate the power in order
701 to organize protocols. An R program was given in supplementary material 3. This paper also
702 provided general results on the efficacy of each method. The type 1 error increased with the
703 variability of relationships and heritability for the REGRESSION method, decreased for the
704 GRAMMAR method and was unaffected for the FASTA and QTDT methods. Power was low for
705 QTDT. In conclusion, we do not recommend the REGRESSION and GRAMMAR models but
706 recommend the FASTA method which is very close to the full mixed model.

707 **LITTERATURE CITED**

- 708 Abecasis, G. R., L. R. Cardon and W. O. C. Cookson, 2000a A general test of association for
709 quantitative traits in nuclear families. *Am. J. Hum. Genet.* 66: 279-292.
- 710 Abecasis, G. R., W. O. C. Cookson and L. R. Cardon, 2000b Pedigree tests of transmission
711 disequilibrium. *Eur. J. Hum. Genet.* 8: 545-551.
- 712 Allison, D. B., 1997 Transmission-disequilibrium tests for quantitative traits. *Am. J. Hum.*
713 *Genet.* 60: 676-690.

- 714 Ambrosius, W. T., E. M. Lange and C. D. Langefeld, 2004 Power for genetic association
715 studies with random allele frequencies and genotype distributions. *Am. J. Hum. Genet.*
716 74: 683-693.
- 717 Amin, N., C. M. Van Duijn and Y. S. Aulchenko, 2007 A Genomic Background Based
718 Method for Association Analysis in Related Individuals. *PLoS One* 2.
- 719 Astle, W., and D. J. Balding, 2009 Population Structure and Cryptic Relatedness in Genetic
720 Association Studies. *Stat. Sci.* 24: 451-471.
- 721 Aulchenko, Y. S., D. J. De Koning and C. Haley, 2007a Genomewide rapid association using
722 mixed model and regression: A fast and simple method for genomewide pedigree-
723 based quantitative trait loci association analysis. *Genetics* 177: 577-585.
- 724 Aulchenko, Y. S., S. Ripke, A. Isaacs and C. M. Van Duijn, 2007b GenABEL: an R library
725 for genome-wide association analysis. *Bioinformatics* 23: 1294-1296.
- 726 Bacanu, S. A., B. Devlin and K. Roeder, 2002 Association studies for quantitative traits in
727 structured populations. *Genet. Epidemiol.* 22: 78-93.
- 728 Balding, D. J., 2006 A tutorial on statistical methods for population association studies. *Nat.*
729 *Rev. Genet.* 7: 781-791.
- 730 Boitard, S., B. Mangin and J. M. Azais, 2010 Asymptotic Distribution of the "Orthogonal"
731 Quantitative Transmission Disequilibrium Test in a Structured Population: Exact
732 Formula. *Stat. Appl. Genet. Mol. Biol.* 9: 11.
- 733 Cardon, L. R., and L. J. Palmer, 2003 Population stratification and spurious allelic
734 association. *Lancet* 361: 598-604.
- 735 Chen, W. M., and G. R. Abecasis, 2007 Family-based association tests for genomewide
736 association scans. *Am. J. Hum. Genet.* 81: 913-926.

- 737 Clayton, D. G., N. M. Walker, D. J. Smyth, R. Pask, J. D. Cooper *et al.*, 2005 Population
738 structure, differential bias and genomic control in a large-scale, case-control
739 association study. *Nature Genet.* 37: 1243-1246.
- 740 Devlin, B., and K. Roeder, 1999 Genomic control for association studies. *Biometrics* 55: 997-
741 1004.
- 742 Erbe, M., F. Ytournal, E. C. G. Pimentel, A. R. Sharifi and H. Simianer, 2011 Power and
743 robustness of three whole genome association mapping approaches in selected
744 populations. *J. Anim. Breed. Genet.* 128: 3-14.
- 745 Ewens, W. J., M. Y. Li and R. S. Spielman, 2008 A Review of Family-Based Tests for
746 Linkage Disequilibrium between a Quantitative Trait and a Genetic Marker. *PLoS*
747 *Genet.* 4: e1000180.
- 748 Ewens, W. J., and R. S. Spielman, 1995 The transmission/disequilibrium test: history,
749 subdivision, and admixture. *Am. J. Hum. Genet.* 57: 455-464.
- 750 Falk, C. T., and P. Rubinstein, 1987 Haplotype relative risks: an easy reliable way to construct
751 a proper control sample for risk calculations. *Ann. Hum. Genet.* 51: 227-233.
- 752 Fan, R. Z., and M. M. Xiong, 2002 High resolution mapping of quantitative trait loci by
753 linkage disequilibrium analysis. *Eur. J. Hum. Genet.* 10: 607-615.
- 754 Freidlin, B., G. Zheng, Z. H. Li and J. L. Gastwirth, 2002 Trend tests for case-control studies
755 of genetic markers: Power, sample size and robustness. *Hum. Hered.* 53: 146-152.
- 756 Fulker, D. W., S. S. Cherny, P. C. Sham and J. K. Hewitt, 1999 Combined linkage and
757 association sib-pair analysis for quantitative traits. *Am. J. Hum. Genet.* 64: 259-267.
- 758 Grapes, L., J. C. M. Dekkers, M. F. Rothschild and R. L. Fernando, 2004 Comparing linkage
759 disequilibrium-based methods for fine mapping quantitative trait loci. *Genetics* 166:
760 1561-1570.

- 761 Guedj, M., E. Della-Chiesa, F. Picard and G. Nuel, 2007 Computing power in case-control
762 association studies through the use of quadratic approximations: application to meta-
763 statistics. *Ann. Hum. Genet.* 71: 262-270.
- 764 Habier, D., R. L. Fernando and J. C. M. Dekkers, 2007 The impact of genetic relationship
765 information on genome-assisted breeding values. *Genetics* 177: 2389-2397.
- 766 Hayes, B. J., A. J. Chamberlain, H. Mcpartlan, I. Macleod, L. Sethuraman *et al.*, 2007
767 Accuracy of marker-assisted selection with single markers and marker haplotypes in
768 cattle. *Genet. Res.* 89: 215-220.
- 769 Henderson, C. R., 1975 Comparison of alternative sire evaluation methods. *J. Anim. Sci.* 41:
770 760-770.
- 771 Johnson, N. L., and S. Kotz, 1970 *Distributions in statistics: continuous univariate*
772 *distributions*. Wiley, New York.
- 773 Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong *et al.*, 2010 Variance
774 component model to account for sample structure in genome-wide association studies.
775 *Nature Genet.* 42: 348-U110.
- 776 Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman *et al.*, 2008 Efficient
777 control of population structure in model organism association mapping. *Genetics* 178:
778 1709-1723.
- 779 Kenward, M. G., and J. H. Roger, 1997 Small sample inference for fixed effects from
780 restricted maximum likelihood. *Biometrics* 53: 983-997.
- 781 Kozlitina, J., C. Xing, A. Pertsemlidis and W. R. Schucany, 2010 Power of Genetic
782 Association Studies with Fixed and Random Genotype Frequencies. *Ann. Hum.*
783 *Genet.* 74: 429-438.
- 784 Laird, N. M., S. Horvath and X. Xu, 2000 Implementing a unified approach to family-based
785 tests of association. *Genet. Epidemiol.* 19: S36-S42.

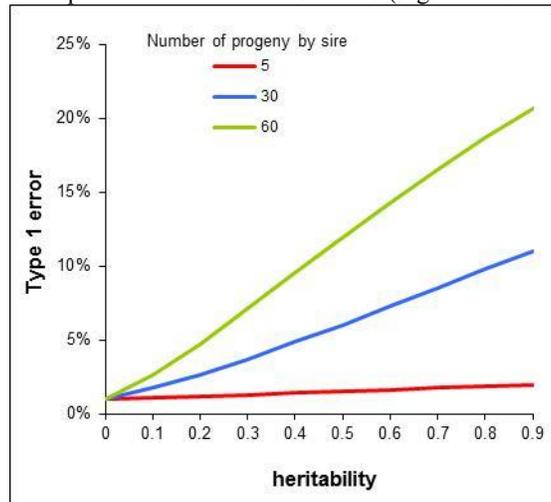
- 786 Laird, N. M., and C. Lange, 2006 Family-based designs in the age of large-scale gene-
787 association studies. *Nat. Rev. Genet.* 7: 385-394.
- 788 Laird, N. M., and C. Lange, 2008 Family-Based Methods for Linkage and Association
789 Analysis, pp. 219-252 in *Genetic Dissection of Complex Traits, 2nd Edition*, edited by
790 D. C. Rao and C. C. Gu.
- 791 Lange, C., D. L. Demeo and N. M. Laird, 2002 Power and design considerations for a general
792 class of family-based association tests: Quantitative traits. *Am. J. Hum. Genet.* 71:
793 1330-1341.
- 794 Lee, A. B., D. Luca, L. Klei, B. Devlin and K. Roeder, 2010 Discovering Genetic Ancestry
795 Using Spectral Graph Theory. *Genet. Epidemiol.* 34: 51-59.
- 796 Li, T. F., Z. H. Li, Z. L. Ying and H. Zhang, 2010 Influence of population stratification on
797 population-based marker-disease association analysis. *Ann. Hum. Genet.* 74: 351-360.
- 798 Marchini, J., L. R. Cardon, M. S. Phillips and P. Donnelly, 2004 The effects of human
799 population structure on large genetic association studies. *Nature Genet.* 36: 512-517.
- 800 Meuwissen, T. H. E., and M. E. Goddard, 2000 Fine mapping of quantitative trait loci using
801 linkage disequilibria with closely linked marker loci. *Genetics* 155: 421-430.
- 802 Meuwissen, T. H. E., A. Karlsen, S. Lien, I. Olsaker and M. E. Goddard, 2002 Fine mapping
803 of a quantitative trait locus for twinning rate using combined linkage and linkage
804 disequilibrium mapping. *Genetics* 161: 373-379.
- 805 Meuwissen, T. H. E., T. R. Solberg, R. Shepherd and J. A. Woolliams, 2009 A fast algorithm
806 for BayesB type of prediction of genome-wide estimates of genetic value. *Genet. Sel.*
807 *Evol.* 41.
- 808 Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick *et al.*, 2006
809 Principal components analysis corrects for stratification in genome-wide association
810 studies. *Nature Genet.* 38: 904-909.

- 811 Price, A. L., N. A. Zaitlen, D. Reich and N. Patterson, 2010 New approaches to population
812 stratification in genome-wide association studies. *Nat. Rev. Genet.* 11: 459-463.
- 813 Pritchard, J. K., and N. A. Rosenberg, 1999 Use of unlinked genetic markers to detect
814 population stratification in association studies. *Am. J. Hum. Genet.* 65: 220-228.
- 815 Pritchard, J. K., M. Stephens and P. Donnelly, 2000a Inference of population structure using
816 multilocus genotype data. *Genetics* 155: 945-959.
- 817 Pritchard, J. K., M. Stephens, N. A. Rosenberg and P. Donnelly, 2000b Association mapping
818 in structured populations. *Am. J. Hum. Genet.* 67: 170-181.
- 819 Quaas, R. L., and E. J. Pollak, 1980 Mixed model methodology for farm and ranch beef cattle
820 testing programs. *J. Anim. Sci.* 51: 1277-1287.
- 821 Rabinowitz, D., 1997 A transmission disequilibrium test for quantitative trait loci. *Hum.*
822 *Hered.* 47: 342-350.
- 823 Risch, N. J., 2000 Searching for genetic determinants in the new millennium. *Nature* 405:
824 847-856.
- 825 Ritland, K., 1996 Estimators for pairwise relatedness and individual inbreeding coefficients.
826 *Genet. Res.* 67: 175-185.
- 827 Satten, G. A., W. D. Flanders and Q. H. Yang, 2001 Accounting for unmeasured population
828 substructure in case-control studies of genetic association using a novel latent-class
829 model. *Am. J. Hum. Genet.* 68: 466-477.
- 830 Spielman, R. S., R. E. Mcginnis and W. J. Ewens, 1993 Transmission test for linkage
831 disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus
832 (IDDM) *Am. J. Hum. Genet.* 52: 506-516.
- 833 Thornton, T., and M. S. Mcpeek, 2010 ROADTRIPS: Case-Control Association Testing with
834 Partially or Completely Unknown Population and Pedigree Structure. *Am. J. Hum.*
835 *Genet.* 86: 172-184.

- 836 Vanraden, P. M., 2008 Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* 91:
837 4414-4423.
- 838 Woolf, B., 1955 On estimating the relation between blood group and disease. *Ann. Hum.*
839 *Genet.* 19: 251-253.
- 840 Wu, C. Q., A. Dewan, J. Hoh and Z. H. Wang, 2011 A Comparison of Association Methods
841 Correcting for Population Stratification in Case-Control Studies. *Ann. Hum. Genet.*
842 75: 418-427.
- 843 Yang, J. A., B. Benyamin, B. P. Mcevoy, S. Gordon, A. K. Henders *et al.*, 2010 Common
844 SNPs explain a large proportion of the heritability for human height. *Nature Genet.*
845 42: 565-569.
- 846 Yu, J. M., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki *et al.*, 2006 A unified mixed-
847 model method for association mapping that accounts for multiple levels of relatedness.
848 *Nature Genet.* 38: 203-208.
- 849 Zeggini, E., L. J. Scott, R. Saxena, B. F. Voight, J. L. Marchini *et al.*, 2008 Meta-analysis of
850 genome-wide association data and large-scale replication identifies additional
851 susceptibility loci for type 2 diabetes. *Nature Genet.* 40: 638-645.
- 852 Zhang, L., J. Li, Y. F. Pei, Y. J. Liu and H. W. Deng, 2009 Tests of Association for
853 Quantitative Traits in Nuclear Families Using Principal Components to Correct for
854 Population Stratification. *Ann. Hum. Genet.* 73: 601-613.
- 855 Zhang, Z. W., E. Ersoz, C. Q. Lai, R. J. Todhunter, H. K. Tiwari *et al.*, 2010 Mixed linear
856 model approach adapted for genome-wide association studies. *Nature Genet.* 42: 355-
857 360.
- 858 Zhao, H. H., R. L. Fernando and J. C. M. Dekkers, 2007a Power and precision of alternate
859 methods for linkage disequilibrium mapping of quantitative trait loci. *Genetics* 175:
860 1975-1986.

- 861 Zhao, K. Y., M. J. Aranzana, S. Kim, C. Lister, C. Shindo *et al.*, 2007b An Arabidopsis
862 example of association mapping in structured samples. PLoS Genet. 3.
- 863 Zhu, X. F., S. C. Li, R. S. Cooper and R. C. Elston, 2008 A unified association analysis
864 approach for family and unrelated samples correcting for stratification. Am. J. Hum.
865 Genet. 82: 352-365.
- 866 Zhu, X. F., S. L. Zhang, H. Y. Zhao and R. S. Cooper, 2002 Association mapping, using a
867 mixture model for complex traits. Genet. Epidemiol. 23: 181-196.
- 868

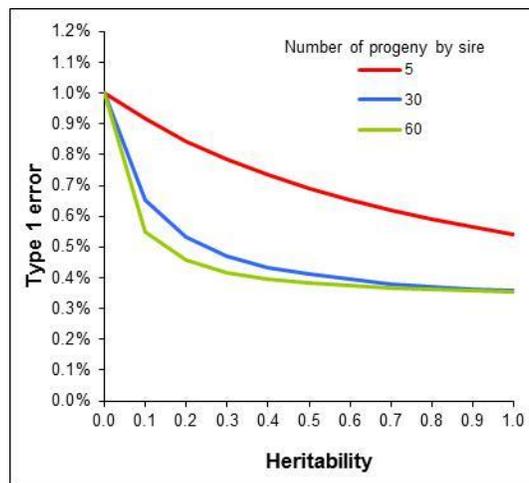
869 Figure 1a – True type 1 error for an assumed type 1 error of 1% with equal half sib families in
870 a sample of 600 animals for model 1 (regression model).



871

872

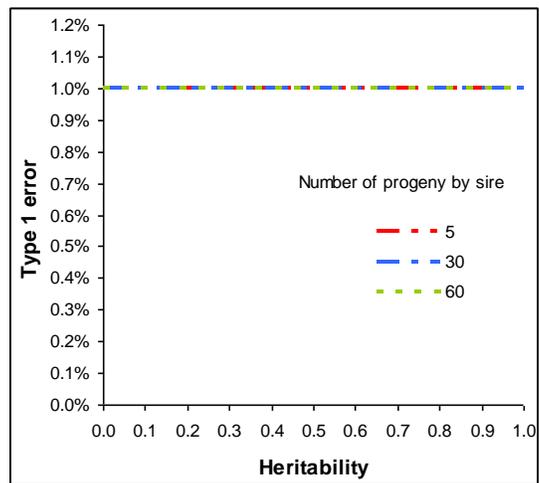
873 Figure 1b – True type 1 error for an assumed type 1 error of 1% with equal half sib families in
874 a sample of 600 animals for model 2 (grammar model).



875

876

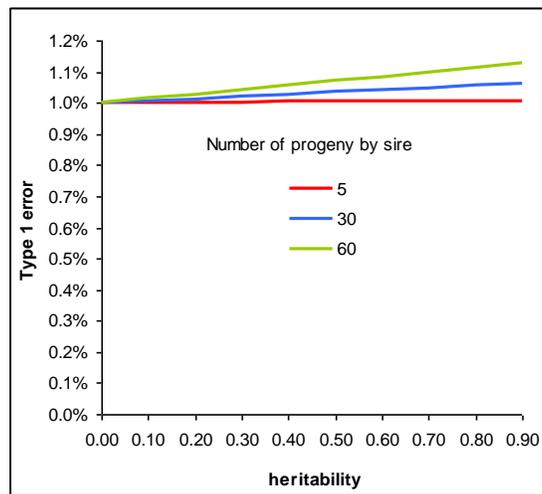
877 Figure 1c – True type 1 error for an assumed type 1 error of 1% with equal half sib families in
878 a sample of 600 animals for model 3 (fasta model).



879

880

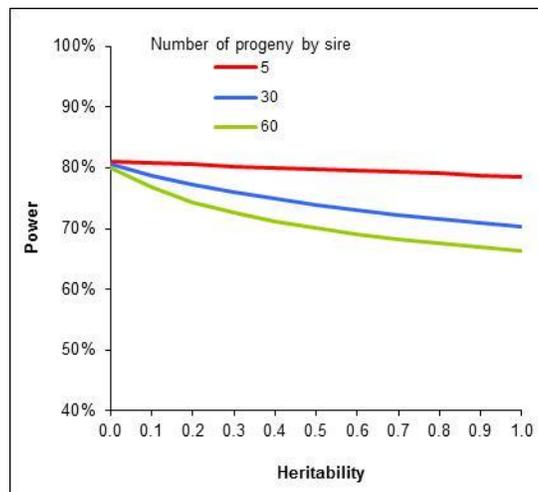
881 Figure 1d – True type 1 error for an assumed type 1 error of 1% with equal half sib families in
882 a sample of 600 animals for model 4 (QTDT model).



883

884

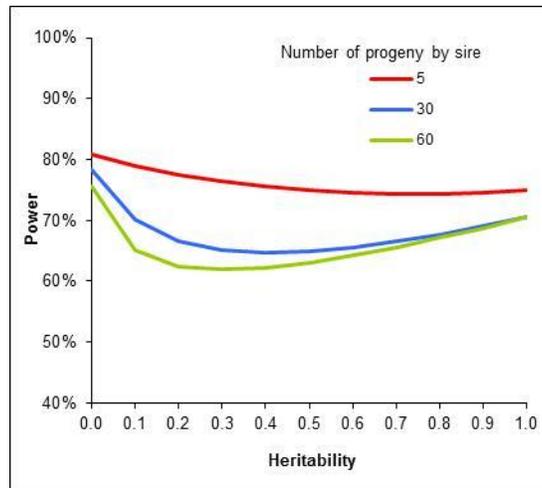
885 Figure 2a – Power for model 1 (regression model) with an assumed type 1 error of 1% in the
886 case of equal half sib families in a sample of 600 animals for a true regression coefficient of
887 $0.20 \sigma_y$ (MAF 50%) or $0.33 \sigma_y$ (MAF 10%) equivalent to 2% of phenotypic variance.



888

889

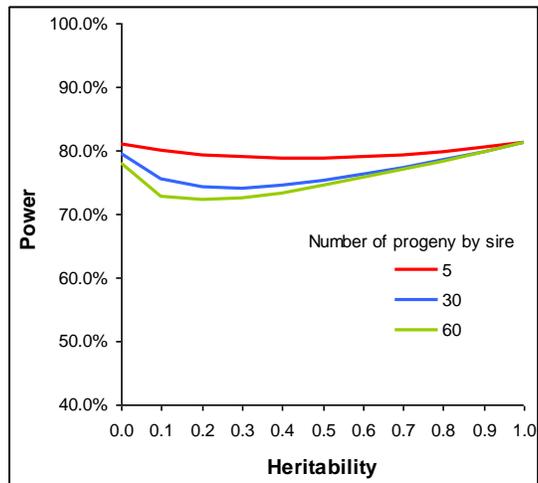
890 Figure 2b – Power for model 2 (grammar model) with an assumed type 1 error of 1% in the
891 case of equal half sib families in a sample of 600 animals for a true regression coefficient of
892 $0.20 \sigma_y$ (MAF 50%) or $0.33 \sigma_y$ (MAF 10%) equivalent to 2% of phenotypic variance.



893

894

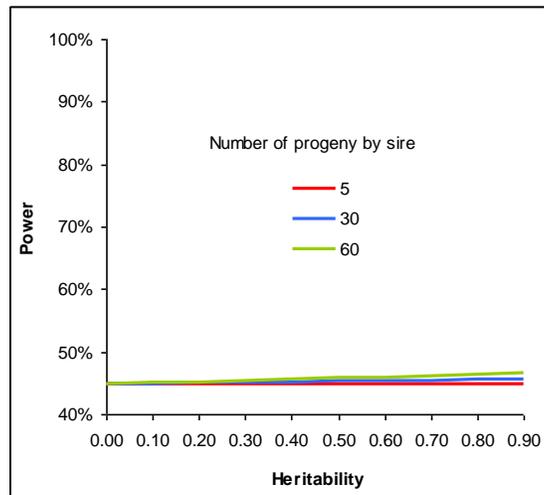
895 Figure 2c – Power for model 3 (fasta model) with an assumed type 1 error of 1% in the case
896 of equal half sib families in a sample of 600 animals for a true regression coefficient of 0.20
897 σ_y (MAF 50%) or $0.33 \sigma_y$ (MAF 10%) equivalent to 2% of phenotypic variance..



898

899

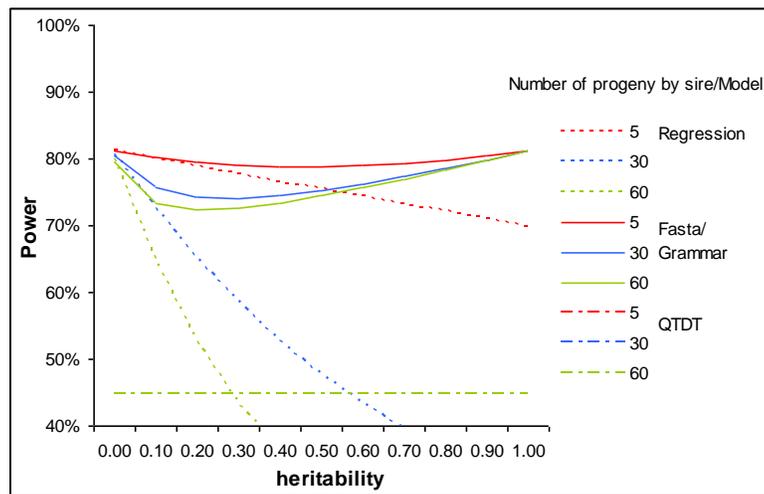
900 Figure 2d – Power for model 4 (QTDT model) with an assumed type 1 error of 1% in the case
901 of equal half sib families in a sample of 600 animals for a true regression coefficient of 0.20
902 σ_y (MAF 50%) or $0.33 \sigma_y$ (MAF 10%) equivalent to 2% of phenotypic variance.



903

904

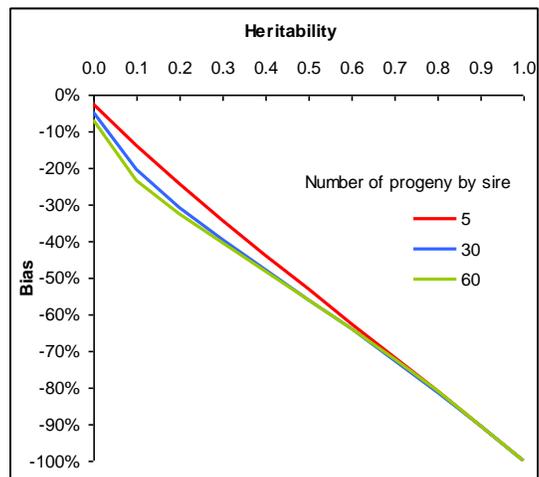
905 Figure 3 – Power for model 1 (regression model), model 2 (grammar model), model 3 (fasta
 906 model) and model 4 (QTDT) with a true type 1 error of 1% in the case of equal half sib
 907 families in a sample of 600 animals for a true regression coefficient of $0.20 \sigma_y$ (MAF 50%)
 908 or $0.33 \sigma_y$ (MAF 10%) equivalent to 2% of phenotypic variance.



909

910

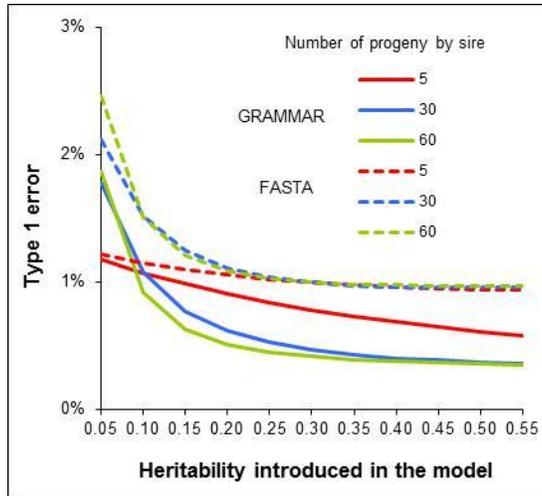
911 Figure 4 – Bias for model 2 (grammar model) as a percentage of true regression coefficient in
912 the case of equal half sib families in a sample of 600 animals as a proportion of true effect.



913

914

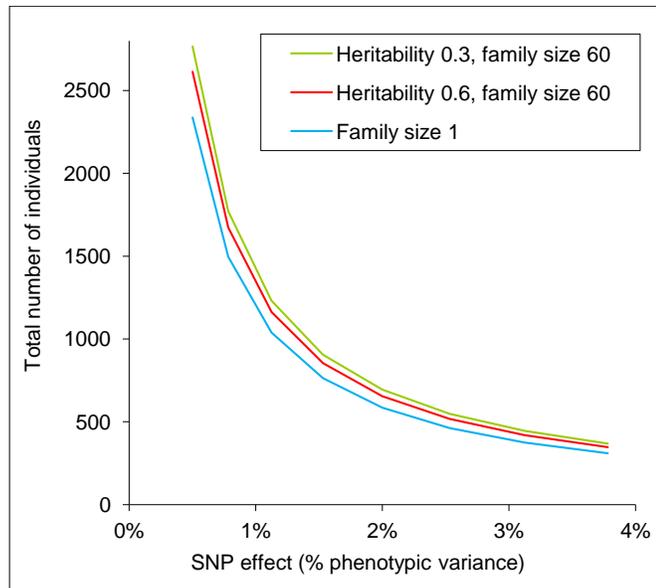
915 Figure 5 – True type 1 error for an assumed type 1 error of 1% with equal half sib families in
916 a sample of 600 animals for model 2 (GRAMMAR model) and 3 (FASTA model) with
917 erroneous heritability introduced in the models and a true heritability of 0.30.



918

919

920 Figure 6 –Population size required to reach a 80% power with a 1% first type error, as a
921 function of the SNP effect and heritability (half sib families of 60 individuals).
922



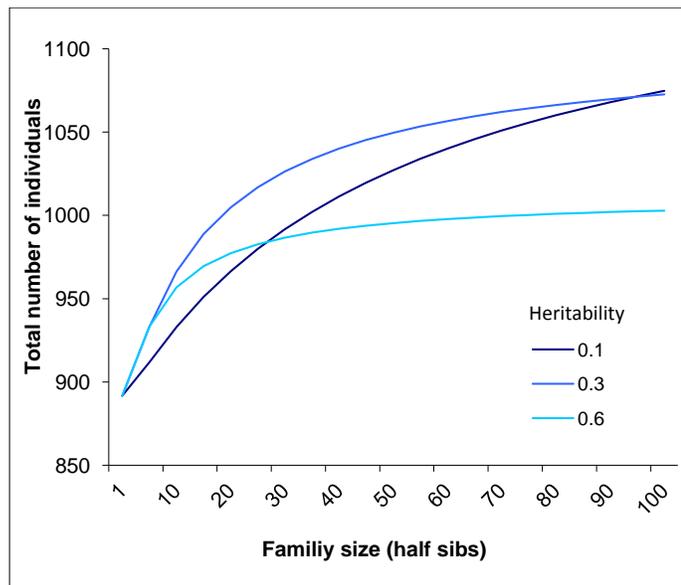
923

924

925

926 Figure 7 –Population required to reach a 80% power with a 1% first type error, with the SNP
927 effect explaining 2% of phenotypic variance as a function of family size with different
928 heritabilities of the trait.

929

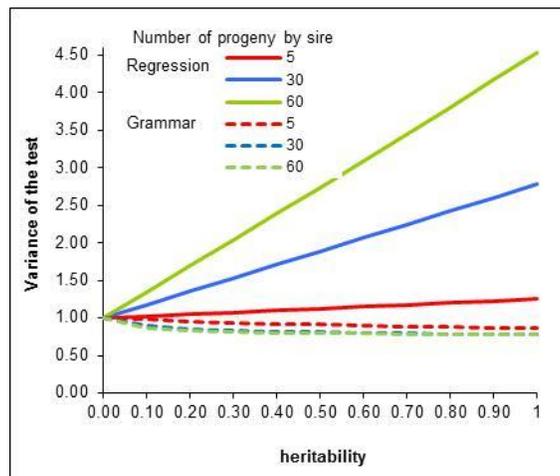


930

931

932 Figure 8 – Variance of the test, equivalent to Control Genomic measure, for model 1
933 (regression model) and model 2 (grammar model) under the H0 hypothesis (without a SNP
934 effect) in the case of equal half sib families in a sample of 600 animals.

935



936

937

2.3 Validation des résultats par simulations

2.3.1 Paramètres des simulations

Afin de valider les formules algébriques décrites dans l'article précédent, certaines simulations ont été réalisées. Toutes les méthodes, à savoir, la Régression, le QTDT, GRAMMAR et FASTA ont été testées. Cette validation a été restreinte aux structures familiales et héritabilités choisies dans l'application de l'article. Nous nous intéressons donc à une population de 600 descendants génotypés issus de 120, 20 et 10 pères qui ont respectivement 5, 30 et 60 descendants. Pour ce faire, les génotypes d'un SNP pour les pères et les mères sont simulés avec une MAF de 0.5 et ceux des descendants sont alors déduits de ses parents. Ensuite, les valeurs polygéniques des pères et des descendants et les phénotypes des descendants sont calculés sans et avec l'effet d'un QTL correspondant à un effet de substitution d'un allèle de 0.20 (équivalent à un coefficient de régression de 0.141 en écart-type phénotypique ou encore un effet du QTL expliquant 2% de la variance phénotypique). La robustesse et la puissance de chaque méthode sont alors évaluées sur ces deux phénotypes (avec ou sans QTL) pour un seuil de significativité de 5% (à différencier au seuil de 1% pris dans l'article). Les simulations sont réalisées pour des héritabilités allant de 0 à 1 par pas de 0.1. Pour chaque scénario, 10000 simulations sont réalisées. Au total, 1320000 simulations ont été effectuées.

Pour les analyses GRAMMAR et FASTA, le logiciel ASREML (Gilmour et al., 2006) a été utilisé afin d'estimer les composantes de la variance. Notons également que pour ces deux méthodes, la matrice de parenté utilisée provient de l'information pedigree et non de l'information génomique.

2.3.2 Résultats

Tous les pourcentages entre les valeurs théoriques et simulées de cette section sont donnés sur l'échelle du type-1 erreur et du type-2 erreur (robustesse et puissance), qui s'expriment chacun en pourcentage. Le calcul des valeurs théoriques a été réalisé à partir d'un programme R appelé RobPower.R (disponible sur demande à teyssedre.simon@voila.fr et prochainement sur un site internet).

Modèle de régression

La Figure 2.1 présente les résultats obtenus par simulations (courbes en pointillées) et les résultats théoriques (courbes continues) pour le modèle de régression. Pour la robustesse, la valeur absolue de l'écart entre les courbes est en moyenne de 0.26% (± 0.25) et atteint son maximum (1.09%) pour 60 descendants par familles et une héritabilité de 1. L'écart est en moyenne plus élevé quand la structure en famille est importante (0.46% pour 60 descendants par famille contre 0.11% pour 5 descendants par famille). Pour la puissance, la valeur absolue de l'écart entre les courbes est en moyenne de 0.37% (± 0.17) et atteint son maximum (0.76%) pour 60 descendants par familles et une héritabilité de 0.3. L'écart est en moyenne plus élevé pour la structure familiale avec 60 descendants par famille (0.46%).

Globalement, les écarts sont faibles et nous permettent de conclure à la validité des résultats théoriques pour le modèle de régression.

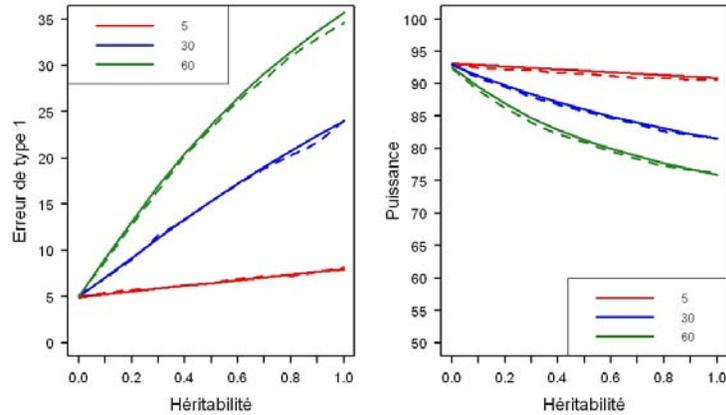


FIGURE 2.1 – Robustesse et puissance du modèle de régression

Modèle QTDT

La Figure 2.2 présente les résultats obtenus par simulations (courbes en pointillées) et les résultats théoriques (courbes continues) pour le modèle QTDT. Pour la robustesse, la valeur absolue de l'écart entre les courbes est en moyenne de 0.28% (± 0.22) et atteint son maximum (0.74%) pour 60 descendants par familles et une héritabilité de 0.7. L'écart est en moyenne plus élevé quand la structure en famille est importante (0.52% pour 60 descendants par famille contre 0.11% pour 5 descendants par famille). Pour la puissance, la valeur absolue de l'écart entre les courbes est en moyenne de 1.12% (± 0.63) et atteint son maximum (0.76%) pour 60 descendants par familles et une héritabilité de 0.3. L'écart est en moyenne plus élevé quand la structure en famille est importante (1.81% pour 60 descendants par famille contre 0.47% pour 5 descendants par famille).

Globalement, les écarts sont relativement faibles et nous permettent de conclure sur la validité des résultats théoriques pour le modèle QTDT.

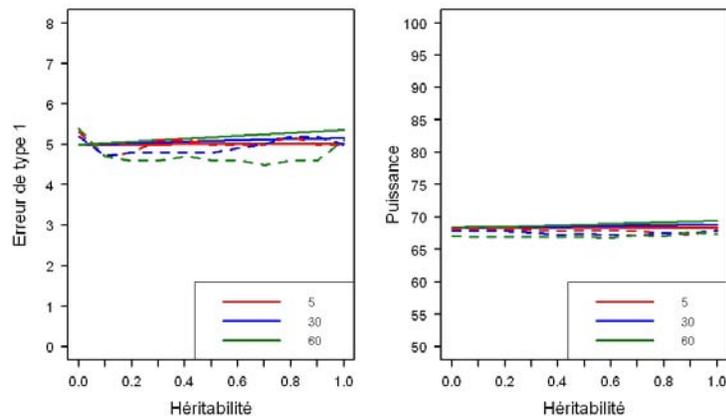


FIGURE 2.2 – Robustesse et puissance du modèle QTDT

Modèle GRAMMAR

La Figure 2.3 présente les résultats obtenus par simulations (courbes en pointillées) et les résultats théoriques (courbes continues) pour la méthode GRAMMAR. Pour la robustesse, la valeur absolue de l'écart entre les courbes est en moyenne de 0.22%(±0.17) et atteint son maximum (0.7%) pour 60 descendants par familles et une héritabilité nulle. Pour la puissance, la valeur absolue de l'écart entre les courbes est en moyenne de 0.58%(±0.71) et atteint son maximum (2.97%) pour 60 descendants par familles et une héritabilité de 0.9. L'écart est en moyenne plus élevé quand la structure en famille est importante (1.16% pour 60 descendants par famille contre 0.11% pour 5 descendants par famille). On note que les écarts semblent augmenter avec la valeur de l'héritabilité. Une explication possible réside dans le fait que l'estimation de l'héritabilité obtenue avec ASREML est biaisée (sous estimée) pour les fortes valeurs d'héritabilité simulées. La Figure 2.4 donne la distribution entre les héritabilités espérées (théoriques) et observées (simulées).

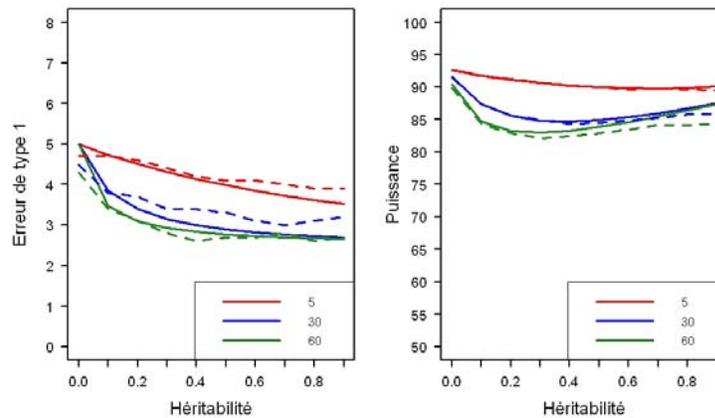


FIGURE 2.3 – Robustesse et puissance pour GRAMMAR

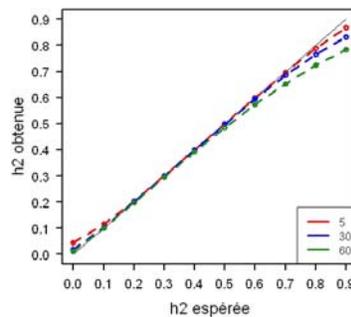


FIGURE 2.4 – Biais entre héritabilité espérée et observée pour la méthode GRAMMAR

Modèle FASTA

La Figure 2.5 présente les résultats obtenus par simulations (courbes en pointillées) et les résultats théoriques (courbes continues) pour la méthode FASTA. Pour la robustesse, la valeur absolue de l'écart entre les courbes est en moyenne de 0.18% (± 0.14) et atteint son maximum (0.5%) pour 30 descendants par familles et une héritabilité de 0.6. Pour la puissance, la valeur absolue de l'écart entre les courbes est en moyenne de 0.90% (± 0.32) et atteint son maximum (1.74%) pour 60 descendants par familles et une héritabilité de 1. L'écart est en moyenne plus élevé quand la structure en famille est importante (1.25% pour 60 descendants par famille contre 0.59% pour 5 descendants par famille). De la même manière que pour GRAMMAR, on note que les écarts semblent augmenter avec la valeur de l'héritabilité et sont possiblement dues au biais entre les héritabilités espérées et observées. La Figure 2.6 donne la distribution entre ces héritabilités.

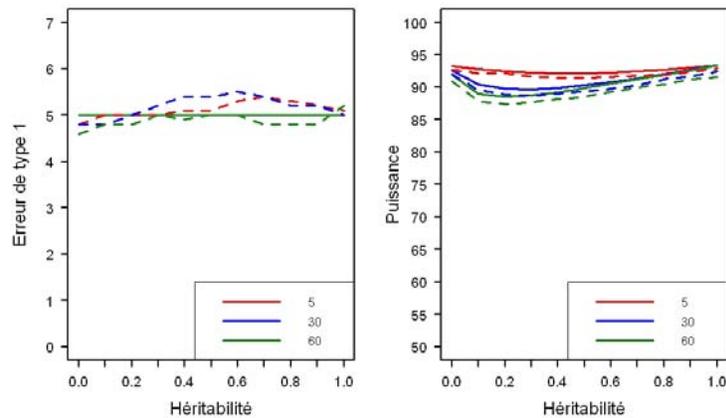


FIGURE 2.5 – Robustesse et puissance pour FASTA

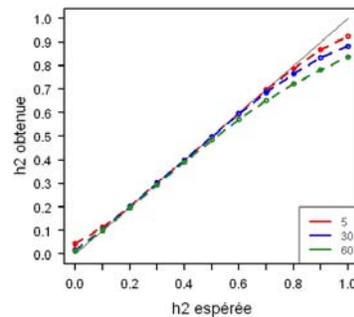


FIGURE 2.6 – Biais entre héritabilité espérée et observée pour la méthode FASTA

2.4 Bilan et perspectives

Dans ce chapitre, la robustesse et la puissance de méthodes uni-SNP simplifiant le vrai modèle mixte ont été évaluées dans le cas d'un caractère quantitatif. L'originalité de ce travail réside dans le fait d'avoir évalué ces paramètres à l'aide de formules analytiques et non par simulations. Parmi les méthodes simplifiées testées, nous montrons que l'approximation du modèle mixte FASTA, qui fixe la variance polygénique, est la plus puissante et la plus robuste à l'existence de structures familiales et a des performances très proches que celle du vrai modèle mixte, tout en étant plus rapide à calculer.

Ces dérivations algébriques des formules permettent une généralisation des résultats, alors que les simulations sont toujours réalisées dans un cadre précis. Elles permettent d'évaluer la robustesse et la puissance espérées de n'importe quel dispositif défini par, un nombre d'individus, une matrice de parenté et une héritabilité. Elles permettent aussi de planifier le dispositif expérimental pour obtenir une puissance suffisante pour détecter un QTL d'effet donné sur la variance du caractère étudié.

Ce travail était focalisé sur les problèmes liés à la structure familiale des données. Une extension à envisager serait l'étude de mélanges entre populations. De plus, d'autres modèles ou méthodes pourraient être testées de la même manière. Les méthodes présentées ici sont toutes uni-SNP et de ce fait uni-QTL. Il faudrait envisager l'extension vers des méthodes uni-QTL haplotypiques et multi-QTL. Il s'agit là d'un travail long qui n'était pas envisageable dans la durée de cette thèse.

Deuxième partie

Analyse de données réelles

Contexte bibliographique

3.1 Introduction

Les lésions ostéochondrales sont couramment observées chez les jeunes chevaux et peuvent être responsables de boiteries et d'une réduction des performances en course. Les lésions les plus fréquentes, groupées sous le nom générique d'ostéochondrose (OC), sont les lésions d'ostéochondrites disséquantes (OCD) et les kystes osseux (Jeffcott, 1991; Trotter and McIlwraith, 1981). Les sites du boulet, du jarret et du grasset sont les plus affectés. Les manifestations d'OC semblent avoir une origine multifactorielle (Jeffcott, 1991; Philipsson et al., 1993; Hurtig and Pool, 1996) et plusieurs facteurs incluant des prédispositions génétiques, la nutrition, l'exercice et d'autres effets environnementaux jouent un rôle dans sa pathogénie. Cependant, l'étiologie et la physiopathologie de l'OC ne sont pas complètement comprises.

La prévalence de l'OC varie en fonction des races et des sites étudiés. La comparaison des prévalences entre les études est rendue difficile par les différences entre les définitions de l'OC, les races étudiées, les mesures de l'OC, les examens et le nombre de clichés radiographiques effectués. C'est pourquoi on trouve une grande disparité entre les précédentes études sur la prévalence de l'OC allant d'environ 10% à 70% (Denoix et al., 1996, 2000; Grondahl and Dolvik, 1993; Philipsson et al., 1993; Pieramati et al., 2003; Ricard, 2002; Sandgren et al., 1993b; Schougaard et al., 1990; Stock et al., 2006; Wittwer et al., 2006). La prévalence, majoritairement étudié sur les sites du boulet et du jarret, est en général plus élevée sur les articulations du boulet (de 8% à 24%) que celles du jarret (de 0% à 16%) (VanGrevenhof et al., 2009).

Outre les aspects du bien-être et de la santé du cheval, l'OC a un impact économique fort puisqu'elle touche directement différents aspects de la valorisation du cheval. Le premier point concerne les conséquences sportives éventuelles car si un cheval est atteint, ses performances risquent de diminuer et ainsi il rapporte moins d'argent. Un deuxième point intervient lors de la possible vente du cheval, un cheval atteint voit sa valeur marchande diminuer fortement et pouvant même faire l'objet de refus dans certains cas. L'éleveur peut également envisager une opération pour soigner le cheval et là encore, le coût et les risques ne sont pas négligeables. Enfin, l'OC étant héréditaire, un étalon atteint ne sera pas utilisé pour la reproduction.

Ce chapitre a pour but de présenter l'OC avant d'aborder ce qui nous intéresse particulièrement, i.e. la partie génétique en cherchant les QTL responsables de l'OC. C'est pourquoi, en s'inspirant des travaux de Leupeule (2007), la première partie fait un bilan de ce qu'est et ce

qui entoure l'OC : définition, pathologie, diagnostics et étiologie. La deuxième partie présente les différentes études de QTL qui ont été réalisées chez les chevaux et les autres espèces.

3.2 L'Ostéochondrose

3.2.1 Définition

Le premier cas d'ostéochondrite disséquante (OCD) fût rapporté chez les humains par Paget en 1880 et nommé OCD en 1887 par König (Hurtig and Pool, 1996). Ils faisaient l'hypothèse qu'une inflammation de l'os et du cartilage était à l'origine d'une "dissection" d'un fragment ostéo-cartilagineux dans l'articulation. Chez le cheval, c'est Nilson en 1947 qui fût le premier à parler d'OC dans le grasset d'un cheval scandinave (VanWeeren and Barneveld, 1999; Jeffcott, 1997). Plus tard, plusieurs auteurs décrivent l'OC comme plusieurs entités qui dérivent de l'hypothèse de la survenue d'un déficit dans le processus d'ossification endochondrale sur l'ensemble des articulations. La première description détaillée de l'OC chez le cheval a été donnée par Rejno et Strömberg en 1978. En 1986, un groupe d'américains ont rassemblés, sous la dénomination de maladie orthopédique du développement (DOD pour *developmental orthopaedic disease*), les affections telles que l'ostéochondrite, les déviations angulaires, les kystes osseux, l'épiphysite, les malformations des os carpaux et tarsaux et certaines affections dégénératives juvéniles (McIlwraith, 1986). En France, les DOD sont connus sous l'appellation AOAJ pour Affections Ostéo-Articulaires Juvéniles (Denoux et al., 1996; McIlwraith, 2004).

L'OC, au sens strict, est définie comme la lésion primaire qui se crée lors de la transformation du cartilage en os. Cette lésion primaire peut par la suite se manifester sous les différentes entités qui constituent les AOAJ, telles que les kystes osseux sous-chondraux (KOSC), les épiphysites, etc. C'est pourquoi, par la suite, nous parlerons d'OC en faisant référence à l'OC au sens large, c'est-à-dire avec ses manifestations. Cependant, les causes initiales de l'OC restent inconnues. En effet, certains auteurs pensent que les AOAJ peuvent-être attribuées à d'autres lésions que des lésions d'ostéochondrose (Jeffcott, 1991; McIlwraith, 2004; VanWeeren, 2006b). De plus, même si elles sont étudiées depuis près de 20 ans, il n'existe pas de consensus sur la classification des AOAJ et des entités à y inclure (Jeffcott, 1991; Hurtig and Pool, 1996; McIlwraith, 2004).

3.2.2 Pathologie

L'OC provient à la base d'un défaut d'ossification endochondrale dont les causes initiales restent inconnues (Jeffcott, 1991; McIlwraith, 2004). Ce défaut implique qu'il y ait une rétention de cartilage et celle-ci mène à un épaississement de ce cartilage au niveau de la surface articulaire. Il apparaît alors une zone de fragilité qui sera plus sensible aux traumatismes externes et internes.

Cette zone de fragilité du cartilage va ensuite évoluer et certaines lésions peuvent même disparaître et donc passer inaperçues à l'âge adulte. Ces rémissions spontanées induisent souvent une minoration de l'incidence de l'ostéochondrose chez le cheval. L'OC peut évoluer en plusieurs entités d'AOAJ en fonction de la fragilité de la zone et du site considéré, car les forces biomécaniques exercées sur différents sites ne sont pas les mêmes. Il peut exister des forces de cisaillement, de compression ou de tension. Les forces de cisaillement sont plutôt localisées sur les sites du jarret et du boulet et induisent fréquemment des manifestations d'OC de la forme de nodules ostéochondraux péri-articulaires (NOCPA) ou des fragmentations ostéochondrales péri-articulaires

(FOCPA). Les forces de compression se retrouvent un peu partout mais plus spécifiquement sur le grasset et le boulet avec notamment des manifestations d'OC sous forme de KOSC, de fragmentations ostéochondrales de surface articulaire (FOCSA) ou de dysplasies. Enfin, les forces de tension se retrouvent plutôt dans le boulet postérieur et dans le carpe avec des manifestations d'OC sous forme de nodules ostéochondraux d'avulsion ligamentaire (NOCAL) et de fragmentations ostéochondrales d'avulsion ligamentaire (FOCAL). A ces trois groupes regroupant des manifestations de l'OC différentes en fonction de forces biomécaniques, on peut rajouter l'arthropathie juvénile que l'on retrouve principalement dans le boulet.

Contrairement à ce que certains auteurs pensaient, l'OC se développe dans les premiers mois après la naissance (Carlson et al., 1995; Dik et al., 1999). Cependant, les AOAJ ou les formes de manifestations de l'OC n'apparaissent pas au même âge, les périodes d'expression sont propres à chacune et ces manifestations résultent d'un processus dynamique. Les lésions surviennent à un âge précoce, mais une réponse est créée immédiatement et la plupart des lésions disparaissent dans les mois suivants. Par contre, certaines lésions restent et deviennent alors permanentes (Carlsten et al., 1993; Dik et al., 1999). La capacité de réparation du cartilage articulaire diminue rapidement après la naissance, avec une capacité de réparation à l'âge mûr qui est pratiquement nulle à cause de la difficulté qu'a le corps à produire du collagène au fil du temps (Bertone et al., 2005). Tout ceci signifie qu'il existe un point de non retour, après lequel toute lésion existante ne sera plus réparée. Ce point de non retour varie entre les sites et les espèces. Chez le cheval, il est estimée entre les 6 mois et 1 an après la naissance mais reste assez variable (Dik et al., 1999; VanWeeren, 2006a).

3.2.3 Clinique et diagnostic des AOAJ

L'ensemble des lésions d'OC ne se manifeste pas par une symptomatologie, seules quelques unes font apparaître des signes cliniques. Les signes cliniques sont difficiles à définir à cause des différentes manifestations d'OC, souvent propres à différents sites articulaires. Néanmoins, parmi ces signes cliniques, on retrouve fréquemment une augmentation du volume d'une articulation, de la raideur et de la boiterie. Dans les cas les plus graves peuvent se rajouter de la douleur et une baisse des performances.

Lorsqu'un examen clinique révèle une forte suspicion de manifestations d'OC chez un cheval, deux techniques sont généralement employées afin de confirmer ce diagnostic : la radiographie et l'échographie. La radiographie est la technique la plus utilisée et reste la technique de choix pour confirmer des manifestations d'OC (Denoix et al., 2002), principalement à cause de son faible coût et de certaines de ces spécificités d'interprétations. Néanmoins, étant donné que les AOAJ se localisent sur toutes les articulations du cheval, il n'est pas toujours possible ou même nécessaire de radiographier un cheval sur toutes ses articulations pour des raisons de coût ou de puissance des appareils (notamment pour les vertèbres et l'épaule). De ce fait, les études rapportant des diagnostics radiographiques diffèrent par le fait d'être spécifiques à une ou plusieurs articulations ou même sur un site précis d'une articulation. Les articulations les plus souvent radiographiées sont celles du boulet et du jarret (Carlsten et al., 1993; Sandgren et al., 1993a; Stock et al., 2006; Wittwer et al., 2006).

Il est possible que l'examen radiographique ne soit pas suffisant, notamment sur des cas précoces où les lésions sont limitées au cartilage de croissance. L'échographie est alors un bon complément des clichés radiographiques. Par exemple, dans le cas de lésions type OCD, elle permet de situer précisément le fragment s'il est libre, d'évaluer sa taille et de visualiser les tissus mous

afin d'orienter le chirurgien dans une possible intervention (Recht et al., 2005). L'échographie n'est pas le seul complément d'un examen radiographique, d'autres techniques comme la scintigraphie ou l'imagerie par résonance magnétique peuvent être utilisées. Néanmoins, toutes ces techniques sont très onéreuses et sont de ce fait rarement utilisées.

3.2.4 Étiologie

Dans cette section, nous considérons l'OC par l'ensemble de ces manifestations qui composent les AOAJ et non comme la cause initiale à ces différentes entités (lésion primaire lors du processus d'ossification). Bien que la cause initiale de l'OC soit inconnue, certains auteurs ont montré que les manifestations d'OC (donc les AOAJ) ont toutes une étiopathogénie complexe qui varie selon l'entité lésionnelle considérée (McIlwraith, 2004) et ont un déterminisme multifactoriel (Jeffcott, 1991; Philipsson et al., 1993; Hurtig and Pool, 1996). Il existe donc plusieurs facteurs, qualifiés de facteurs de risque, qui peuvent provoquer la maladie. Tous ces facteurs ne sont pas connus et agissent à des degrés différents sur la maladie. Néanmoins la plupart des facteurs connus sont maîtrisables et présentent donc un intérêt majeur puisqu'il est possible d'agir directement sur eux afin de diminuer l'incidence et la prévalence de la maladie. Nous présentons ici les facteurs de risque les plus connus et les plus étudiés dans la recherche, à savoir des prédispositions génétiques, l'alimentation, les pratiques d'élevage et la vitesse de croissance.

Hypothèse héréditaire et prédisposition génétique

Rapidement, les vétérinaires et les éleveurs se sont aperçus que certains étalons avaient des produits plus fréquemment atteints que d'autres et Rejnö and Strömberg (1978) le notent d'ailleurs dès 1978. Dès lors, des prédispositions génétiques ont été suspectées et certaines études sur l'héritabilité de l'OC ont vu naissance. L'héritabilité est définie comme la part de la variation phénotypique qui est attribuable à la variation génétique. En pratique, la valeur de l'héritabilité détermine donc la relation entre la présence d'OC chez un étalon et dans sa descendance. Ainsi, si l'héritabilité est nulle, alors l'apparition de l'OC n'est due qu'aux conditions environnementales et le fait qu'un étalon soit atteint d'OC ne permettra en aucun cas de prédire la fréquence d'apparition d'OC dans sa descendance. Dans les autres cas, le fait de connaître l'héritabilité peut permettre de prévoir une augmentation ou une diminution de la prévalence dans la descendance d'un étalon. Ainsi, pour un étalon atteint de l'OC dans une population où la prévalence est de 30% et l'héritabilité de 0.2, alors la prévalence de la descendance de cet étalon augmentera de 4% par rapport à la prévalence de la population (Ricard, 2007). Notons que ce genre de prédictions ne peuvent être réalisées qu'avec la connaissance conjointe de la prévalence et de l'héritabilité. La valeur de l'héritabilité permet également de conditionner l'efficacité d'une sélection de reproducteurs contre l'apparition d'OC.

Pour comparer différentes études d'héritabilité, l'idéal serait qu'elles utilisent les mêmes phénotypes, avec des effectifs de grandes tailles en famille de pères, des échantillonnages aléatoires dans la descendance des pères et qu'une même race soit étudiée. Or en pratique, ce n'est pas du tout le cas. Au niveau des phénotypes, certains auteurs optent pour définir des héritabilités des différentes entités AOAJ, d'autres utilisent l'OC sous toutes ses manifestations et d'autres définissent les héritabilités en fonction d'un site radiographique. Néanmoins, la comparaison reste possible entre les études par site et par entité puisque généralement, les lésions observées sur un site particulier correspondent à une entité AOAJ majoritaire. En général, du fait de cette multiplicité des phénotypes, un biais existe dans la comparaison, ce qui peut être un facteur des grandes disparités

entre les héritabilités observées dans diverses études que nous verrons dans le paragraphe suivant. Au niveau des effectifs, les études sont réalisées en pratique à partir d'échantillons de petite taille ou de moyenne taille, c'est à dire entre 500 et 1000 chevaux (Ricard, 2002; Pieramati et al., 2003; Schober et al., 2003; Grondahl and Dolvik, 1993; Philipsson et al., 1993; Schougaard et al., 1990; VanGrevenhof et al., 2009), ce qui implique des estimations peu précises car plus l'effectif est grand, plus l'intervalle de confiance autour de la vraie valeur de l'héritabilité est petit. Récemment, certaines études avaient des effectifs beaucoup plus élevés, et comprenaient entre 2000 et 5000 chevaux (Der Kinderen, 2005; Stock et al., 2005; Stock and Distl, 2006). Au niveau de l'échantillonnage aléatoire dans la descendance des pères, les chevaux ne sont quasi-jamais choisis strictement au hasard. Une sélection est souvent réalisée par la limitation du nombre de descendants par étalon ou lorsque la collecte des chevaux résulte d'une vente aux enchères (souvent le cas dans les études avec beaucoup de chevaux). De ce fait, l'héritabilité est alors sur- ou sous-estimée sans savoir dans quel sens (Ricard, 2007). Enfin, un dernier point qui rend la comparaison entre les estimations d'héritabilités difficiles est la diversité des races qui sont étudiées. On retrouve aussi bien des races de trotteurs (français, norvégiens ou standardbred) que des races de chevaux de sport (allemands, français, hanovriens etc...) et les résultats d'héritabilité entre ces races n'ont aucune raison d'être similaires.

L'estimation de l'héritabilité est souvent réalisée à partir d'un modèle père ou un modèle polygénique et certains effets du milieu doivent être corrigés. En effet, parfois la répartition entre les effets génétiques et les effets environnementaux est déséquilibrée et la non prise en compte simultanée de ces effets induit une mauvaise estimation de la part génétique du caractère. Parmi les effets environnementaux pris en compte, on retrouve principalement l'âge et le sexe des chevaux. Cependant, l'influence du sexe sur l'OC porte à des discordances. Certains auteurs trouvent des prévalences plus élevées chez les mâles (Sandgren et al., 1993a; Geffroy et al., 1997), ou chez les femelles (Wittwer et al., 2006), quand d'autres aboutissent à des prévalences équivalentes (Grondahl, 1991; Denoix et al., 2000; Stock et al., 2006). L'âge quand à lui, joue un rôle crucial car comme on a pu le voir dans les paragraphes précédents, les lésions évoluent durant les premiers mois et peuvent même disparaître. Les études portent souvent sur les chevaux de deux à trois ans, c'est à dire, après le supposé point de non retour et avant l'entraînement du cheval. Il est difficile de savoir s'il vaut mieux radiographier les chevaux avant ce point de non retour pour permettre d'observer des lésions qui se seraient résorbées ou après. Une chose est certaine, c'est que plus les chevaux sont âgés, plus les facteurs environnementaux, comme par exemple l'entraînement du cheval, risquent de biaiser l'estimation.

Même si la comparaison entre les études est rendue difficile à cause des problèmes cités précédemment, il reste intéressant de regarder les différents résultats obtenus dans ces études. Pour l'OC du jarret, une grande variabilité est observée puisque les résultats varient entre 0 et 0.52 selon les races et les méthodes utilisées (Ricard, 2002; Ricard et al., 2010; Schober et al., 2003; Grondahl and Dolvik, 1993; Philipsson et al., 1993; Schougaard et al., 1990; VanGrevenhof et al., 2009; Der Kinderen, 2005; Stock et al., 2005; Stock and Distl, 2006). Pour l'OC du boulet, les études donnent des valeurs d'héritabilité comprises entre 0.10 et 0.29 (Ricard, 2002; Ricard et al., 2010; Schober et al., 2003; Grondahl and Dolvik, 1993; Philipsson et al., 1993; VanGrevenhof et al., 2009; Stock et al., 2005; Stock and Distl, 2006). Pour l'OC du grasset, les valeurs sont comprises entre 0 et 0.09 (Ricard, 2002; Ricard et al., 2010; Der Kinderen, 2005). Le lecteur est amené à voir la Table 1 de VanGrevenhof et al. (2009) et la Figure "héritabilité des affections ostéo-articulaires en différents sites" de Ricard (2007) pour plus de détails sur les différentes valeurs d'héritabilité obtenues dans

la littérature. En France, les travaux effectués par Ricard et al. (2010) sur les Trotteurs Français montrent des valeurs d'héritabilité égales à 0.29 pour les lésions du boulet, 0.19 pour celles du jarret et 0 ailleurs.

Quelques études ce sont intéressées à la corrélation génétique entre les sites héréditaires, c'est à dire principalement entre les lésions d'OC du boulet et du jarret (Grondahl and Dolvik, 1993; Stock et al., 2005; Ricard et al., 2010; Stock and Distl, 2006). Ces études sont peu nombreuses car elles nécessitent un grand nombre d'individus. Cependant, les estimations vont toutes dans le même sens de corrélations génétiques très faibles. Les estimations varient entre -0.27 (Stock et al., 2005) et 0.26 (Ricard et al., 2010) avec des écarts-types d'erreurs importants, respectivement 0.14 et 0.22. Ceci implique qu'il existe une certaine forme d'indépendance génétique entre des anomalies situées dans différents sites articulaires. Plusieurs hypothèses sont envisageables pour expliquer ces résultats mais ne sont pas démontrées. Parmi ces hypothèses possibles, on trouve celle que la cause commune aux différents sites est compromise, ou celle que les manifestations de l'OC résultent d'un autre processus génétique qui n'est pas commun aux différents sites.

Croissance

La vitesse de croissance, staturale et/ou pondérale, peut être un facteur de manifestations d'OC chez le cheval (Jeffcott, 1991; Watkins et al., 1999; McIlwraith, 2004). En effet, une croissance rapide augmenterait les forces de pression sur les articulations portantes et ainsi favoriserait l'apparition d'AOAJ. Cette croissance rapide est régulièrement observée puisque le cheval est à croissance rapide et réalise 70% de son poids adulte et de sa hauteur au garrot avant l'âge de 1 an (Martin-Rosset, 2001), période également préférentielle aux manifestations d'OC (Carlson et al., 1995; Dik et al., 1999). Les races les plus touchées seraient donc les races à vitesse de croissance élevée, c'est à dire avec un gain moyen quotidien (GMQ) en hauteur au garrot et poids supérieurs aux autres. Certaines études le montrent et trouvent une relation entre les individus atteints d'OC et le GMQ mais les résultats ne sont que légèrement significatifs (Thompson et al., 1988; Pagan and Jackson, 1996; VanWeeren et al., 1999; Donabedian et al., 2006; Stock et al., 2006; Lepeule et al., 2009, 2011), et parfois même des auteurs trouvent des résultats non significatifs (Alvarado et al., 1990; Sandgren et al., 1993b; Jelan et al., 1996). Les causes de GMQ élevées sont principalement génétique et alimentaire.

Alimentation et rationnement

L'alimentation est un facteur de risque des AOAJ car elle peut influencer le processus d'ossification ou conditionner la vitesse de croissance. De par son accessibilité par un ensemble de professions et sa facilité à être contrôlée, l'alimentation est l'un des points qui a engendré la plus abondante littérature. Par contre, l'étude de l'alimentation est rendue difficile par de nombreux paramètres souvent très corrélés entre eux et par une connaissance des apports journaliers individuels souvent approximative. De nombreux paramètres alimentaires peuvent être à l'origine de manifestations d'OC (Wolter, 1996). Ici, nous verrons quelques uns d'entre eux, à savoir une suralimentation énergétique et azotée, des déséquilibres ioniques et certains autres paramètres alimentaires.

La suralimentation énergétique est l'un des facteurs les plus importants dans l'apparition de lésions d'OC. Elle entraîne une surcharge pondérale et grasseuse ainsi que des articulations plus fragiles à cause du manque d'ossification. De ce fait, sous la contrainte d'effets biomécaniques, ces zones fragiles en ossification ont plus de risque de voir apparaître des entités AOAJ (Sandgren et al.,

1993a; Watkins et al., 1999; Savage et al., 1993a). Au niveau des apports excessifs de protéines, si ceux-ci sont apportés de façon isolée, alors ils ne semblent pas générer de conséquences sur la croissance et donc sur le développement d'AOAJ. Par contre, en association avec une ration trop riche, les risques de manifestations d'OC sont augmentés (Savage et al., 1993a; Lienasson, 2005).

Les principaux déséquilibres ioniques identifiés comme facteurs potentiels de manifestations d'OC sont le calcium (Ca), le Phosphore (P), le cuivre (Cu) et le Zinc (Zn). Le cuivre intervient dans le processus d'ossification et une carence en cuivre serait à l'origine d'une déficience dans celui-ci en rendant le cartilage plus fragile et à même à se fracturer, menant de ce fait à la possible apparition d'AOAJ (Bridges and Harris, 1988). En revanche, même si sa validité semble être limitée à certaines périodes (Pearce et al., 1998), une supplémentation en Cu des poulains et des poulinières aiderait à prévenir et pourrait même minorer l'ostéochondrose (Knight et al., 1990). Les excès de Zn sont quant à eux délétères et peuvent induire des manifestations d'OC, probablement par leur effet antagoniste sur l'absorption du Cu. De ce fait, certains auteurs comme McIlwraith (2004) proposent des recommandations sur le rapport Zn/Cu de l'ordre de 4 à 5. Il est également intéressant de noter que le déficit en Zn est lui aussi délétère et peut générer des problèmes ostéo-cartilagineux. Enfin, les excès de Ca et de P, qui sont les deux principaux composants ioniques des structures osseuses, peuvent aussi induire des manifestations d'OC (Savage et al., 1993b). L'excès de Ca étant un facteur nuisible à l'assimilation des autres éléments (phosphore, cuivre et zinc) dont les rôles dans le bon développement osseux et articulaire sont essentiels.

D'autres paramètres alimentaires ont un rôle dans le processus d'ossification endochondrale comme notamment les acides gras essentiels, l'acidose métabolique, les oligo-éléments, les vitamines A, C, D et les compléments de type collagène d'origine marine. Pour autant, leurs rôles n'ont pas vraiment été prouvés puisque leurs études, qui consistent à estimer et faire varier les quantités ingérées, semblent réellement difficiles.

Pratiques d'élevage

L'exercice (ou l'activité physique d'une manière générale) est souvent une période révélatrice des manifestations cliniques de l'OC (boiteries, irrégularités...). En effet, dans leurs études portant sur l'exercice du cheval, Barneveld and VanWeeren (1999) et VanWeeren and Barneveld (1999) pensent que l'exercice a un effet délétère sur le cartilage articulaire et sa maturation le rendant moins apte à se régénérer en cas de lésion. Durant la période de maturation (i.e. avant la période de non retour évoquée précédemment), l'exercice est donc un risque dans le développement des lésions et tout traumatisme pendant cette période serait alors un agent étiologique de l'OC. D'autres auteurs montrent cependant que l'exercice a des effets positifs et soit même essentiel au bon développement ostéo-articulaire (Jeffcott, 1991; Martin-Rosset, 2001). Au final, l'exercice est bon pour le développement du cheval mais doit être contrôlé et dosé pour ne pas être délétère.

Les effets des pratiques d'élevage sont souvent évoqués comme étant responsable de manifestations d'OC (Jeffcott, 1993; Martin-Rosset, 2001; McIlwraith, 2004; Lepeule et al., 2011). Parmi ces pratiques, on note surtout la taille de l'aire de l'exercice, le type de logement, la qualité des sols et les conditions sanitaires. Certaines d'entre elles ont même été mises en évidence dans des études expérimentales (Caure et al., 1998; VanWeeren and Barneveld, 1999; Wilke et al., 2003; Lepeule et al., 2011).

3.3 Les études de QTL de l'ostéochondrose

3.3.1 Chez les chevaux

Étant donné qu'il ne fait plus doute qu'existe une prédisposition génétique à l'OC, plusieurs études récentes ont visé à trouver les marqueurs moléculaires ou QTL associés à l'OC.

La première étude, proposée par Dierks et al. (2007), s'est intéressée aux QTLs responsables de l'OC dans une population de chevaux hanovriens (HAN). L'étude portait sur les manifestations de l'OC et de l'OCD sur le boulet et/ou le jarret. Pour ce faire, Dierks et al. (2007) ont utilisé un échantillon de 104 HAN issus de 14 étalons, tous génotypés sur 260 Microsatellites (MS) couvrant le génome. Les résultats d'une analyse de liaison multipoint via Merlin (Abecasis et al., 2002) ont montré, au seuil de significativité empirique du chromosome à 5%, 8 régions impliquées dans les manifestations d'OC ou de type OCD sur le boulet et le jarret ensemble, 9 pour le boulet seul et 9 pour le jarret seul. Pour le boulet et le jarret, la région présentant le plus fort intérêt pour l'OC (resp. pour l'OCD) a été localisée au alentour des 27cM (resp. 42cM) sur le chromosome 2. Pour le boulet, la région présentant le plus fort intérêt pour l'OC (resp. pour l'OCD) a été localisée au alentour des 31cM (resp. 41cM) sur le chromosome 2. Pour le jarret, la région présentant le plus fort intérêt pour l'OC (resp. pour l'OCD) a été localisée au alentour des 49cM (resp. 45cM) sur le chromosome 2 (resp. 16). Le pic le plus fort étant celui trouvé pour l'OCD du jarret sur le chromosome 16. Dans leur discussion, Dierks et al. (2007) semblaient porter un intérêt particulier sur les régions des chromosomes 2, 4, 5 et 16 et ont alors commencé à parler de gènes sous-jacents à ces régions en vue d'un possible *fine mapping*. D'ailleurs, récemment, Dierks et al. (2010) ont raffiné la région du chromosome 2 à l'aide de marqueurs SNP sur la même population d'HAN.

Une autre détection de QTLs a été réalisée par Wittwer et al. (2007) à partir de marqueurs Microsatellites. Cette étude portait sur un échantillon de 117 chevaux de trait d'Allemagne du Sud (le bavarois) issus de 9 étalons. Les auteurs ont étudié les manifestations d'OC et d'OCD sur les sites du boulet et/ou du jarret également à l'aide d'une analyse LA via Merlin. Ils ont montré, au seuil de significativité empirique du chromosome à 5%, 9 régions impliquées dans des manifestations d'OC ou d'OCD développées sur le boulet et le jarret, 23 dans le boulet ou dans un site particulier du boulet et 4 dans le jarret. Pour le boulet et le jarret, la région présentant le plus fort intérêt pour l'OC a été localisée au alentour des 46cM sur le chromosome 17. Pour le boulet, la région présentant le plus fort intérêt pour l'OC (resp. pour l'OCD) a été localisée au alentour des 156cM (resp. 46cM et 70cM) sur le chromosome 1 (resp. chromosome 18 et 4). Pour le jarret, la région présentant le plus fort intérêt pour l'OC a été localisée au alentour des 78cM sur le chromosome 18. Le pic le plus fort étant celui trouvé pour l'OCD du boulet sur le chromosome 4. Wittwer et al. (2007) se sont particulièrement intéressés aux régions des chromosomes 4 et 18 et un *fine mapping* à partir de marqueurs SNP a permis de valider et réduire la taille de ces régions (Wittwer et al., 2008, 2009).

La première étude recensée utilisant les marqueurs SNP de la puce Illumina BeadChip EquineSNP50 sur tout le génome pour la détection de QTLs de l'OC est celle de Lampe et al. (2009b). Elle consistait en l'étude de 154 HAN, dont 40 provenaient de la même population de HAN que Dierks et al. (2007). Les caractères étudiés étaient identiques à la précédente étude sur les HAN, à savoir les manifestations de l'OC et de l'OCD sur le boulet et/ou le jarret. Afin de détecter les QTLs, un test d'association cas/témoins corrigé pour l'apparentement entre les individus a été utilisé. Une région était déclarée comme un QTL lorsque la valeur $-\log_{10}(P)$ était supérieure à 2.5 pour au moins 3 SNP consécutifs. Ils ont trouvé 7 QTLs pour l'OC et l'OCD du boulet et du

jarret, 10 sur le boulet et 7 sur le jarret. Pour les manifestations de l'OC (resp. OCD) du boulet et du jarret, ils trouvent le pic le plus fort en position 36cM (resp 65cM) sur le chromosome 18 (resp. 3). Pour les manifestations de l'OC et de l'OCD du boulet, ils trouvent le pic le plus fort en position 12cM sur le chromosome 30. Pour les manifestations de l'OC et de l'OCD du jarret, ils trouvent le pic le plus fort en position 44cM sur le chromosome 1.

Plus récemment, une autre analyse via les marqueurs SNP de la puce Illumina BeadChip EquineSNP50 a été réalisée par Lykkjen et al. (2010) sur une population de Trotteurs Norvégiens (TN). L'étude portait sur 162 Trotteurs, issus de 22 familles de père, génotypés et phénotypés pour une entité AOAJ du jarret, à savoir l'ostéochondrose du relief intermédiaire de la cochlée tibiale (OC RICT). Cette entité étant de loin la plus fréquente au niveau du jarret, le caractère étudié peut facilement être comparé au caractère sains/atteints sur le jarret. Les auteurs ont utilisé plusieurs méthodes pour analyser les données, toutes uni-SNP, telles qu'un test Armitage, une régression logistique ou un modèle mixte dont la matrice de variance-covariance était calculée à partir de l'information du pedigree. Ce dernier étant utilisé comme référence dans leur discussion. Au seuil $p < 5 \times 10^{-5}$, ils ont montré 4 QTLs détectés avec le modèle mixte sur les chromosomes 5, 10, 27 et 28 mais également 6 autres QTLs détectés avec les autres méthodes sur les chromosomes 1, 3, 4, 5, 9 et 18. Le pic le plus élevé avec le modèle mixte étant atteint autour des 80cM sur le chromosome 10 (pvalue 1.19×10^{-5}).

Il est difficile de comparer les QTLs détectés dans ces différentes études puisque les races, les phénotypes étudiés, ainsi que les densités de marqueurs ne sont pas toujours les mêmes. La liste des QTLs trouvés dans les différentes études est donnée dans la Table 3.1. Néanmoins, il est intéressant de regarder si une région ne serait pas propre à différentes races et n'agirait pas sur un même site ou une même entité AOAJ. La comparaison la plus facile est celle entre les études de Dierks et al. (2007) et Lampe et al. (2009b) sur les HAN, l'une utilisant des marqueurs MS, l'autre des SNPs et les deux études portant sur le même descriptif des phénotypes. Les résultats sont étonnants et peu encourageants. En effet, une vingtaine de QTLs avaient été trouvés dans les deux études, mais seuls 3 d'entre eux semblent correspondre sur ECA 3 pour l'OCD du boulet, sur ECA 4 pour l'OC du boulet et du jarret et sur ECA 16 pour l'OC du jarret. Il semblerait donc que la majeure partie des QTLs trouvés dans ces études soient des faux positifs qui pourraient être liés au trop faible nombre d'individus pris en compte dans ces études (117 et 159). Entre races, très peu de QTLs semblent être en commun, les études de Dierks et al. (2007) et Wittwer et al. (2007) montrant néanmoins 3 régions possiblement communes, deux sur ECA 4 et une sur ECA 16 pour l'OC du boulet, mais à des distances de l'ordre des 5cM et avec des effectifs de petite taille. Récemment, Lampe et al. (2009a) ont évoqué la possibilité qu'un QTL détecté sur ECA 18 chez les bavares puissent également être impliqué chez les HAN. L'étude de Lykkjen et al. (2010) chez les Trotteurs Norvégiens va dans le même, à savoir que les QTLs trouvés chez ces trotteurs diffèrent des précédentes études et finalement un seul QTL serait susceptible d'être en commun à l'étude de Dierks et al. (2007) sur ECA 5. Si l'on compare les deux études SNP sur les races Trotteurs Norvégiens et HAN pour l'OC du jarret, aucun QTL n'est susceptible d'être en commun. Ceci peut être expliqué par une disparité trop importante entre les races ou par le fait que les analyses se fassent sur des effectifs trop faibles, ce qui mène parfois à l'apparition de faux positifs.

3.3.2 Dans d'autres espèces

L'ostéochondrose est commune à plusieurs espèces, comme notamment les chevaux, les chiens, les vaches, les porcs et les humains. Chez les espèces domestiques, la prévalence de l'OC est la plus élevée chez les chevaux et les porcs. De ce fait, et combiné à de multiples enjeux économiques ou sportifs, la plupart des études se sont portées sur l'OC chez les chevaux (comme vu précédemment) et les porcs (Grondalen, 1974; Crenshaw, 2006).

Andersson-Eklund et al. (2000) ont proposé la première analyse de QTLs de l'OC chez le porc. L'étude portait sur une population de 195 F2 croisés *Wild boar* × *Large White* et une analyse LA a été réalisée à l'aide de marqueurs MS. Ils ont montré 3 QTLs sur les *sus scrofa chromosomes* (SSC) 5, 13 et 15. Les auteurs se sont particulièrement intéressés à la région sur le SSC 5 car elle est homologue à la région du chromosome humain HSA 12q14-q24, région qui contient le possible gène candidat *Cartilage homeoprotein 1* (CART1 ou ALX1). Notons que cette région correspond à la région entre 12 et 13cM sur ECA 28. Un peu plus tard, Lee et al. (2003) ont étudié 302 F2 croisés *Meishan* × *Large White* à l'aide de 711 MS sur tout le génome. Ils ont trouvé deux QTLs sur les SSC 7 et 16 mais ces QTLs ne dépassaient pas le seuil des 5% sur le chromosome, ce qui est très faible. L'étude la plus intéressante est sans aucun doute celle de Christensen et al. (2010). En effet, ils ont étudié une population de 7172 individus issus d'un croisement entre des *duroc* × (*Large White* × *Landrace*), ce qui est considérable par rapport aux études précédentes. L'analyse, réalisée à partir du modèle LA de Fernando and Grossman (1989) sur 462 SNPs, a mis en évidence la présence de 11 QTLs situés sur les SSC 1, 4, 7, 10 et 15, qui dépassaient un seuil de 0.1% du génome. Les auteurs notent qu'un QTL, détecté sur SSC 5 à un seuil moindre, est proche du QTL détecté préalablement par Andersson-Eklund et al. (2000) même si la définition du phénotype est différente (l'un s'intéresse à l'OC d'un site en particulier et l'autre au niveau global). Par contre, les QTLs sur SSC 7 et 15 ne semblent pas correspondre aux précédentes études. Cette étude est sans contexte la plus grosse étude réalisée pour la détection de QTL de l'OC du fait de sa grande taille d'échantillon. Néanmoins, on peut regretter le faible nombre de marqueurs (environ 1 SNP tous les 4cM) qui ne permet pas l'utilisation de méthodes de type LD ou LDLA, qui auraient pu raffiner ces régions avec de gros intervalles de confiance (de l'ordre de 20cM).

3.4 Problèmes existants et enjeux de la thèse

L'analyse de la bibliographie disponible sur la détection de QTL de l'OC chez les chevaux donne une liste importante de QTL. Néanmoins, peu d'entre eux sont communs aux différentes études. Une explication possible est que l'OC en un site articulaire peut être associée à différents locus selon la race. Une autre explication envisageable est un manque d'homogénéité entre les mesures phénotypiques de l'OC des différentes études, notamment le nombre de clichés radiographiques pour les construire. Mais une des explications les plus probables est certainement la faible puissance des protocoles, en général de petite taille. Au maximum, ce nombre était de 164 pour les études portant sur l'utilisation de la puce Illumina BeadChip EquineSNP50 pour l'étude de Lykkjen et al. (2010). La puissance de détection en analyse d'association étant fonction du nombre d'individus, ces études n'étaient pas adaptées pour permettre de détecter des QTL de moyens et faibles effets (part de la variance phénotypique expliquée par le QTL < 7%). Or, il semble que l'OC soit une maladie polygénique, avec plusieurs QTL de faibles et moyens effets contrôlant ce caractère.

La recherche des QTL et des gènes sous-jacents étant réellement importante pour envisager de réduire l'incidence de la maladie dans la population, il y a donc une nécessité de réaliser une étude avec un plus grand nombre d'individus phénotypés. Le chapitre suivant est consacré à une étude que nous avons réalisé sur les Trotteurs Français (TF) où 525 puis 583 chevaux ont été analysés dans le but de trouver les QTL causaux.

TABLE 3.1: Détails des QTLs de l'OC détectés dans plusieurs études

QTLs DÉTECTÉS POUR L'OC								
ECA ¹	Dierks et al. (2007) 104 HAN, 260 MS		Wittwer et al. (2007) 117 bavarois, 157 MS		Lykkjen et al. (2010) 162 TN, 54k SNP		Lampe et al. (2009b) 154 HAN, 54k SNP	
	Pos ²	Trait ³	Pos ²	Trait ³	Pos ²	Trait ³	Pos ²	Trait ³
1							43-44	FH-OD, F-O, H-OD
1			115-126	FP				
1			132-138	FP	139	H-D		
1			150-156	F-OD				
1			193-194	F-OD				
2	27	FH-O, F-OD						
2	42-50	FH-OD, F-OD, H-O					104	F-OD
2								
3	20-30	FH-D, F-D					27	F-D
3							64	FH-D
3					113	H-D		
4	7	FH-O	7-10	FP				
4	24	FH-O, F-O						
4	46	FH-O, F-O					41	FH-OD, H-OD
4			58	FP				
4	66	F-O	70-74	FP				
4					76	H-D		
5	44-50	H-D	40-44	FH-O, F-O	42	H-D		
5	73-79	F-OD			78	H-D		
5	100	F-OD						
6							47	H-D
6								
8			79	FP				
8			109	FP				
9								
9					18	H-D		
10					60	H-D		
10					80	H-D		
13			0	F-D				
13							15	F-D
14							57	FH-O, H-O
15			24-27	H-O				
15			37	FH-O, F-O, H-O				
15	63	H-O						
16	3-8	F-OD, H-OD						
16	33	FH-OD, F-O, H-OD	33-39	F-O				
16	42-45	H-OD						
16	59	H-OD						
16	87-89	H-O					81-82	FH-OD, F-O, H-O
17			46	F-O				
18							36	FH-OD, F-O, H-O
18			45-54	F-D				
18					58	H-D		
18			78	FP, H-O				
19	0-2	FH-D						
21	0	H-D						
21	16-24	H-OD						
22							42	F-D
22			57	FH-O				
22			65	FH-O				
22			79	FH-O				
23			44	FH-O, F-O				
25			0	FH-O, F-OD				
26			7	FP				
26							27	F-O
27			0	F-O				
27					38	H-D		

Suite page suivante

ECA ¹	Dierks et al. (2007) 104 HAN, 260 MS		Wittwer et al. (2007) 117 bavarois, 157 MS		Lykkjen et al. (2010) 162 TN, 54k SNP		Lampe et al. (2009b) 154 HAN, 54k SNP	
	Pos ²	Trait ³	Pos ²	Trait ³	Pos ²	Trait ³	Pos ²	Trait ³
28			7	FH-O, F-O				
28					42	H-D		
29							16	H-OD
30							8-12	FH-OD, F-OD, H-OD
31			47	H-O				

1. ECA : *Equus Callabus* chromosome

2. Pos : Position en MegaBase (Mb)

3. Trait : F=boulet (fetlock), H=jarret (hock), FP=fragments ostéochondraux palmaire/plantaire du boulet (POF), O=OC, D=OCD, OD=OC et OCD

Analyse GENEQUIN

4.1 Résumé de l'article

Les lésions ostéochondrales sont couramment observées chez les jeunes chevaux et peuvent être responsables de mauvaises performances en course. Les lésions les plus fréquentes, groupées sous le nom générique d'"ostéochondrose" (OC), sont les lésions d'ostéochondrites disséquantes et les kystes osseux (Jeffcott, 1991; Trotter and McIlwraith, 1981). Les sites du boulet, du jarret et du grasset sont les plus affectés. Les manifestations d'OC semblent avoir une origine multifactorielle et plusieurs facteurs incluant des prédispositions génétiques, la nutrition, l'exercice et d'autres effets environnementaux jouent un rôle dans sa pathogénie. Cependant, l'étiologie et la physiopathologie de l'OC ne sont pas complètement comprises. La prévalence de l'OC varie de 10% à 25% entre les races (Grondahl and Dolvik, 1993; Philipsson et al., 1993) et l'héritabilité entre 0.17 et 0.52 chez les trotters entre les races et les sites étudiés (Schober et al., 2003; Stock et al., 2005; Stock and Distl, 2006). Différentes détections de QTLs pour l'OC sur tout le génome ont déjà été réalisées (voir chapitre précédent) sur différentes races (Dierks et al., 2007; Wittwer et al., 2007; Lykkjen et al., 2010). Cependant, peu de QTLs ont été trouvés en commun à ces études et une seule d'entre elles utilisait des marqueurs type SNP. L'objectif de notre étude était de réaliser une détection de QTL pour l'OC sur tout le génome à partir de la puce EquineSNP50 dans une population de Trotteurs Français (TF).

Les données provenaient du programme ANR GENEQUIN, programme qui visait à étudier l'OC dans une population de TF. Un total de 525 TF a été phénotypé. Le phénotypage, principalement réalisé au CIRALE, a été réalisé par deux vétérinaires spécialistes en orthopédie équine et consistait en la radiographie d'au moins 10 images prises sur l'ensemble des sites. A partir de ces clichés, diverses mesures de l'OC furent créées : un score global sur l'ensemble des clichés (GM, voir article pour détail), la présence ou l'absence d'OC sur le boulet (FM), le jarret (HM) et ailleurs que sur le boulet et le jarret (OM). Les chevaux phénotypés provenaient de 161 familles de pères et étaient âgés en moyenne de 3 ans. Le génotype était disponible pour tous les descendants et les pères avec au moins deux descendants. Après divers filtrages sur la qualité des marqueurs, l'analyse a été réalisée à partir de 41249 SNPs répartis sur l'ensemble des 31 chromosomes autosomes.

Une étude théorique sur la puissance en analyse d'association de notre protocole a été réalisée en suivant Ball (2005) et Luo (1998). Cette étude nous a permis de connaître la puissance de détection disponible via notre protocole pour des QTLs qui expliquaient 3%, 5% et 7% de la variance

phénotypique. Les résultats ont donné respectivement des puissances de 55%, 78% et 91% pour des QTL en LD de 0.35 avec un SNP (ce chiffre est le LD moyen pour un QTL situé à mi-distance entre 2 SNPs), montrant ainsi qu'un QTL expliquant plus de 7% de la variance phénotypique devait être détecté. Deux analyses distinctes ont été utilisées afin de détecter les QTLs sur l'ensemble du génome. La première, appelée SNPMixed, était une analyse SNP par SNP dans un modèle mixte où l'effet de la structure de population était corrigée par un effet aléatoire polygénique. Ce premier modèle utilisait uniquement l'information provenant du déséquilibre de liaison (LD). Le deuxième modèle était quant à lui une analyse par haplotype et utilisait l'information conjointe du LD et de la transmission des gamètes des pères aux descendants (LA).

Afin de lutter contre les problèmes de tests multiples, les seuils de significativité ont été choisis pour des P-valeurs égales à 5.10^{-4} , 5.10^{-5} et 5.10^{-6} . Le seuil le plus strict, 5.10^{-6} correspondait à un seuil de Bonferroni au niveau $\alpha = 5\%$ avec 10000 tests indépendants. Le seuil de 5.10^{-5} a été utilisé pour montrer des associations modérées (Burton et al., 2007). Enfin, le seuil de 5.10^{-4} a été utilisé afin de pouvoir comparer les QTLs entre les différents caractères étudiés. A cause du LD, plusieurs QTLs peuvent être identifiés dans la même région, c'est pourquoi les statistiques de score proposées par Guedj et al. (2006) furent utilisées pour créer des régions dans lesquelles on faisait l'hypothèse de la présence d'un QTL unique.

L'héritabilité a été estimée à 0.32 (± 0.14) pour GM, 0.27 (± 0.13) pour FM, 0.45 (± 0.15) pour HM et 0.13 (± 0.11) pour OM. En moyenne, plus de QTLs ont été détectés avec SNPMixed qu'avec HaploIBD pour les seuils $P < 5.10^{-4}$ et $P < 5.10^{-5}$. Au seuil $P < 5.10^{-4}$, 4 QTLs ont été détectés en commun avec les deux méthodes pour GM, 2 pour FM et 4 pour HM et OM. Au seuil $P < 5.10^{-5}$, 1 QTL sur ECA 13 a été détecté avec les deux méthodes pour GM, 2 QTLs sur ECA 3 et 14 l'ont été pour HM et 1 QTL sur ECA 15 l'a été pour OM. Par contre, aucun QTL n'a été détecté à ce seuil avec les deux méthodes pour FM. Au seuil le plus strict $P < 5.10^{-6}$, un seul QTL sur ECA 3 a été détecté avec HaploIBD pour HM à la position 105.05 Mb. Ce QTL, dont la P-valeur du test associé était de $-\log(P) = 5.52$ expliquait 7% de la variance phénotypique et représentait notre plus fort QTL.

La comparaison des QTLs entre les caractères a été réalisée à partir des QTLs détectés avec les deux méthodes au seuil $P < 5.10^{-4}$. A ce seuil, 2 des 4 QTLs détectés pour GM étaient proches de deux QTLs détectés pour FM (sur ECA 13 et 15), et les 2 autres étaient proches de deux QTLs détectés pour HM (sur ECA 3 et 14). Par contre, aucun QTL n'a été trouvé en commun entre les caractères FM, HM et OM.

Cette étude a permis de mettre en évidence la présence d'un fort QTL sur ECA 3 pour l'OC du jarret. De plus, plusieurs autres QTLs pour d'autres caractères semblaient également intéressants sur les chromosomes ECA 13 pour GM, 14 pour HM et 15 pour OM. Aucun QTL n'a été trouvé en commun entre HM et FM, ce qui n'était pas surprenant puisqu'en accord avec les faibles corrélations génétiques trouvées entre l'OC du boulet et du jarret (Grondahl and Dolvik, 1993; Stock and Distl, 2006). Tous les QTLs détectés pour GM au seuil $P < 5.10^{-4}$ étaient également détectés pour HM ou FM. Donc, comme on l'attendait de par sa définition, GM combine bien FM et HM et révèle de QTLs spécifiques. En comparaison avec les QTLs trouvés dans d'autres races, seulement 2 QTLs semblaient être proches des précédentes études : un sur ECA 3 (Lykkjen et al., 2010) et un sur ECA 13 (Lampe et al., 2009b). Néanmoins, notre étude portait sur une taille de population de 525 chevaux, ce qui est 3 fois supérieur aux précédentes études sur l'OC. De ce fait, notre étude est plus puissante et plus robuste.

En conclusion, plusieurs QTLs associés avec l'OC sur différents sites anatomiques ont été

révélés dans cette étude. Nous montrons que l'OC, maladie multifactorielle, est influencée par plusieurs gènes et nous n'avons pas trouvé pas que le déterminisme génétique soit identique pour l'OC développée au niveau du jarret et au niveau du boulet. Des études complémentaires vont maintenant se concentrer sur l'identification de gènes candidats et le dépistage des mutations dans une tentative de clarifier la physiopathologie de l'OC au niveau moléculaire et de développer des stratégies efficaces pour l'évaluation des risques. Pendant ce temps, les marqueurs pourraient être utilisés dans un contexte de sélection assistée par marqueurs pour améliorer la santé et le bien-être des chevaux.

4.2 Article appliqué

Cet article a été accepté dans *Journal of Animal Science*.

JOURNAL OF ANIMAL SCIENCE

The Premier Journal and Leading Source of New Knowledge and Perspective in Animal Science

Genome-wide association studies for osteochondrosis in French Trotters
S. Teyssèdre, M. C. Dupuis, G. Guérin, L. Schibler, J. M. Denoix, J. M. Elsen and A.
Ricard

J ANIM SCI published online August 12, 2011

The online version of this article, along with updated information and services, is located on
the World Wide Web at:

<http://jas.fass.org/content/early/2011/08/12/jas.2011-4031>



American Society of Animal Science

www.asas.org

Genome-wide association studies for osteochondrosis in French Trotters

S. Teyssède^{*2}, M.C. Dupuis^{‡2}, G. Guérin[†], L. Schibler[†], J.M. Denoix[‡], J.M. Elsen^{*} and A. Ricard[†]

^{*}Institut National de la Recherche Agronomique, UR 631, 31326 Castanet-Tolosan, France.

[†]Institut National de la Recherche Agronomique, UMR 1313, 78352 Jouy-en-Josas, France.

[‡]Ecole Nationale Vétérinaire d'Alfort, Centre d'Imagerie et de Recherche sur les Affections Locomotrices Equines, 14430 Goustranville, France.

¹The GENEQUIN program was supported by the French National Research Agency (ANR), by the Fonds Eperon, by the French Institute for the Horse and Equitation (IFCE) and the Basse-Normandie regional council.

²These authors have equally contributed to this article.

Corresponding author: Simon Teyssède

simon.teyssede@toulouse.inra.fr

ABSTRACT: A genome-wide association study (GWAS) for osteochondrosis (OC) in French Trotters horses (FT) was carried out to detect quantitative trait loci (QTL) using genotype data from the Illumina EquineSNP50 BeadChip assay. Analysis data came from 161 sire families of FT with 525 progeny and family sizes ranging from 1 to 20. Genotypes were available for progeny (525) and sires with at least two progeny (98). Radiographic data were obtained from progeny using at least ten views to reveal OC. All radiographic findings were described **by at least two veterinary experts in equine orthopedics** and severity indices (scores) were assigned based on the size and location of the lesion. Traits used were a global score, the sum of all severity scores lesions (GM, quantitative measurement), and the presence or absence of OC on the fetlock (FM), hock (HM) and other sites (OM). Data were analysed using two mixed models including fixed effects, polygenic effects and SNP or haplotype cluster effects. By combining results with both methods at moderate evidence of association threshold $P < 5 \times 10^{-5}$, this GWAS displayed one region for GM on *Equus caballus* chromosome (ECA) 13, two for HM on ECA 3 and 14, and one for OM on ECA 15. One region on ECA 3 for HM represented **ed** the most significant hit ($P = 3 \times 10^{-6}$). By comparing QTLs between traits at a lower threshold ($P < 5 \times 10^{-4}$), the four QTLs detected for GM were associated to a QTL detected for FM or HM but never both. Another interesting result was that no QTLs were found in common between HM and FM.

Key words: equine, French Trotters horses, genome-wide association, osteochondrosis, quantitative trait loci.

INTRODUCTION

Osteochondral lesions are commonly observed in young horses and may be responsible for reduced performances in racing. The most frequent lesions, grouped under the generic name of “osteochondrosis” (OC), are osteochondritis dissecans lesions and bone cysts (Jeffcott and Henson, 1998; Trotter and McIlwraith, 1981). Fetlock, hock and stifle joints are mainly affected. Osteochondrosis manifestations appear to be of multifactorial origin and many factors including genetic predisposition (Jeffcott, 1991; Philipsson et al., 1993; Stock and Distl, 2006a; van Weeren, 2006), nutrition, exercise and other environmental effects seem to play a role in its pathogenesis. However, the etiology and physiopathology of OC are not fully understood. Likewise, the molecular mechanisms involved in OC are still unknown.

The prevalence of OC in warmblood, coldblood, thoroughbred and trotter horses ranges from 10% to 25% across different breeds (Grøndahl and Dolvik, 1993; Philipsson et al., 1993; Stock et al., 2005b; van Grevenhof et al., 2009). Heritability estimates for OC fall within the ranges $h^2 = 0.17-0.52$ in trotters (Grøndahl and Dolvik, 1993; Philipsson et al., 1993) and $h^2 = 0-0.37$ in warmbloods (Schober et al., 2003; Stock and Distl, 2006b; Stock et al., 2005a).

Various whole-genome scans using microsatellite markers revealed quantitative trait loci (QTL) in Hanoverian warmblood breeds and South German coldbloods (Dierks et al., 2007; Wittwer et al., 2008; Wittwer et al., 2009; Wittwer et al., 2007). Those **studies** only shared one QTL on *equus caballus* (ECA) 16. More recently, using the Illumina Bead-Chip EquineSNP50, two genome-wide analyses were performed on Hanoverian warmblood (Lampe et al., 2010) and standardbred Norwegian horses (Lykkjen et al., 2010) and revealed several QTLs in these two breeds.

The aim of this study was to carry out a genome-wide association study using the Illumina BeadChip EquineSNP50 to identify QTLs associated with OC in French trotters (FT).

MATERIALS AND METHODS

Sample characteristics

Data were from the French GENEQUIN program, a study based on a population of French trotters (FT). A total of 525 FT were recorded. The sample came from 161 sire families with family sizes ranging from 1 to 20 with an average of 3.26 progeny per sire (63 sires had only 1 progeny). About 80% of the horses were less than 4 years old at the time of examination (mean age 2.8 ± 1.9). About 45% of the sample horses were females and 55% males (with 23% geldings). Sixty-nine percent of the horses were in training when screened radiographically and 40% had participated in at least one race. Genotypes were available for all progeny and for the 98 sires of 462 progeny (mean 4.7 ± 3.6 progeny per sire). More precisely, genotypes were obtained for all progeny and from sires with at least 2 progeny. Phenotypes including all measurements were available for all progeny only.

Radiography

Horses were recruited at CIRALE, the French imaging center and research on equine locomotor disorders, and a few veterinary clinics. Additional screening was performed on farms with high OC prevalence in foals and yearlings in order to increase family sizes and collect DNA samples from dams when available. Information on clinical signs, medical history (to exclude cases having already undergone surgery) and horse origins were requested. Radiographies and all reports were examined by at least two veterinary experts in equine orthopedics.

The radiologic screening consisted in at least ten views including the lateromedial projections of the frontlimb and hindlimb digits (foot, pastern and fetlock joints), dorsopalmar projections of the carpi, and lateromedial projections of the hocks and stifles.

Phenotypic Measurements

All radiographic findings were described accurately and severity indices (scores) were assigned based on the size and location of the lesion, as well as associated bone remodeling (Denoix et al., 2000). **These scores were calculated as the power 2 of four severity degrees (0 to 3): the 0 degree represented a very small size lesion and corresponded to a score of 1 ($2^0 = 1$); the 1 degree represented a small size lesion and corresponded to a score of 2 ($2^1 = 2$); the 2 degree represented a medium size lesion and corresponded to a score of 4; the 3 degree represented a consistent size lesion and corresponded to a score of 8. The frequency of lesions of degree 0 (respectively 1, 2 and 3) was 25% (respectively 55%, 19% and 1%).** A global score, the sum of the severity scores for all lesions **on all radiographic sites** was attributed to each individual: horses with a global score of 0 were considered healthy; horses with a global score >2 were considered affected; horses with a global score of one, or with a global score of zero or two but displaying unclear radiographic findings were considered intermediate. To avoid erroneous results due to misclassification, intermediate horses were excluded from the analysis. Finally, based on global scores in our sample, 263 healthy horses and 262 affected horses were included in the analysis.

Different traits were used for data analysis. **To obtain a phenotypic distribution as close as possible to a normal distribution,** a global measurement (GM) was defined as the log-transformed of **one plus** global score (**$GM = \log(1 + \text{global score})$**) and was analyzed as a continuous phenotype (Figure 1: mean \pm sd 0.78 ± 0.85). Site-specific analyses were also carried out following case-control studies based on the presence or absence of OC on the fetlock (FM), hock (HM) and other sites (OM). **The frequency of cases and controls were shown in Table 1.** Pearson correlations between site-specific traits were low (at the most 0.13) whereas GM was, by definition, highly correlated to the site-specific traits (Table 2).

Markers and genotype quality control

Horse genotyping was performed using the Illumina Equine SNP50 BeadChip assay at Labogena, according to the manufacturer's standard procedures. This array includes 54,602 evenly distributed SNPs throughout the genome. Markers were from all 31 equine autosomes and the X chromosome and the average distance between

two SNPs was 0.043 ± 0.055 Mb. The X chromosome was not included in analysis. All data were subject to quality control procedures. Firstly, only samples showing a minimum call rate of 98% (percentage of SNPs genotyped for an individual) were included in the study. Moreover, poor quality markers were discarded based on three criteria: markers genotyped in less than 80% of the samples (call freq < 80%), or having a minor allele frequency (MAF) under 5%, or deviating from **Hardy**-Weinberg equilibrium in cases and controls ($P < 10^{-8}$) were rejected. Finally, 41,249 SNPs were analyzed.

Statistical Analyses

Linkage disequilibrium and expected power. Average linkage disequilibrium (LD) was computed every 0.02 Mb across intervals of 0 to 1.6 Mb using the r^2 measure (Hill, 1974). The expected extent of LD assuming an effective population size of 150 and 1,000 individuals was also estimated using the method described by Tenesa et al. (2007). According to Ball (2005) and Luo (1998), the theoretical power of an association study performed on our sample size to detect a QTL in LD with a SNP was also estimated using R Package ldDesign software.

Single-SNP analyses. To model the polygenic part and the SNP effect simultaneously, we performed the following mixed model:

$$y = 1\mu + X\beta + Zu + e$$

with y is the vector of phenotypes, μ is the overall mean, β is the vector of the fixed effects **with $\beta' = (\beta_{AGE}, \beta_{SNP})$** , X is the incidence matrix of β **with the corresponding level of age at control and** the genotypes of all individuals (coded 0, 1 and 2), u is the vector of random polygenic effects with $u \sim N(0, A\sigma_u^2)$, Z is the incidence matrix for u and e is the vector of random residual effects with $e \sim N(0, I\sigma_e^2)$. **The** fixed effect “age at control”, with 2 levels (individuals controlled at two years or less, more than two years), was included in this analysis for GM and HM. This effect was not significant for the others traits. The A matrix is the relationship matrix based on the available pedigree information which included 2,796 horses. Parameters of this model were estimated using ASReml (Gilmour et al., 2006) for each SNP on GM, HM, FM and OM. A Student’s test of the

null hypothesis (no QTL, i.e. $\beta_{\text{SNP}} = 0$) against the alternative (there is a QTL, i.e. $\beta_{\text{SNP}} \neq 0$) was performed for each SNP. We called this method SNPMixed. Heritabilities for all traits were also estimated with this model without the SNP effects.

Haplotype Analyses. We used the Druet et al. (2008) method which combined linkage analysis (LA) and linkage disequilibrium analysis (LDLA). This method was based on the one described by Meuwissen et al. (2002) and derived from the original method proposed by Meuwissen and Goddard (2000). This is a variance component (VC) mapping method which includes information from LD between haplotypes and the transmission of haplotypes across generations. In order to provide complete information, the full procedure is described below. At first, we used the program DualPhase (Druet and Georges, 2010) to phase the haplotypes of all genotyped individuals. **The sire haplotypes and the maternally inherited haplotypes of the sons were then considered as base haplotypes. When a sire of a progeny was not genotyped, base haplotypes were then progeny haplotypes.** At each tested position the following procedure was applied:

1. Probabilities of transmission p_{ij} were computed to determine which base haplotype corresponded **to which paternally inherited haplotype**. The rules applied when computing these probabilities using the closest informative bracket can be found in Table 1 in Pong-Wong et al (Pong-Wong et al., 2001). Linkage disequilibrium information was not taken into account at this step.
2. Identity-by-descent (IBD) probabilities (Φ_p) were estimated among each pair of base haplotypes conditionally on the identity-by-state (IBS) status of the neighboring markers using windows of 6 flanking markers (Meuwissen and Goddard, 2001).
3. Base haplotypes were grouped with a clustering algorithm with SAS proc CLUST using $(1 - \Phi_p)$ as a measure of distance. Base haplotypes were grouped if Φ_p exceeded 0.6. Two additional rules were applied to assign haplotypes to clusters: i) when the two haplotypes of a sire were grouped in the same cluster, the paternally inherited haplotypes of all his sons were then grouped in this cluster ii) when the probability of transmission between a base haplotype and a haplotype was greater than 0.95 (it was grouped to the corresponding cluster).

4. The performances were modeled as follows:

$$y = X\beta + Zu + Z_h h + e$$

where y is the vector of phenotypes, **β is the vector of fixed effects including the overall mean and the effect “age at control”, X is the incidence matrix of the fixed effects,** u is the vector of random polygenic effects assumed to be normally distributed $u \sim N(0, A\sigma_u^2)$ with A the relationship matrix identical to the A matrix of SNPMixed, and h is a vector of random QTL effects corresponding to the haplotype clusters assumed to be normally distributed $h \sim N(0, H\sigma_h^2)$ with H the IBD matrix between haplotype clusters. Z and Z_h are the design matrices relating phenotypes respectively to corresponding animal effects and haplotype clusters. **The fixed effect “age at control” was only included in the model for GM and HM.**

Maximum likelihood estimations of genetic parameters were obtained using an expectation maximization-restricted maximum likelihood (EM-REML). The REMLF90 software (Misztal et al., 2002) was modified by Druet et al. (2008) to incorporate relationship matrices among QTL allelic effects. The presence or absence of a QTL at a given position was tested with the likelihood-ratio test statistic:

$$\lambda = -2 \ln \left(\frac{L(H_0)}{L(H_1)} \right)$$

where $L(H_0)$ and $L(H_1)$ are the likelihood of the observations when parameters are equal to their REML estimations values under the polygenic model with no QTL fitted (H_0) and the general model with the QTL (H_1) respectively. The distribution of the test is not known but was previously shown to be close to half of a 0-df plus half of a 1-df chi-square distribution for a single position (Self and Liang, 1987). P-values were computed using this distribution.

We applied this method to each SNP position on GM, HM, FM and OM. We called this method HaploIBD.

Criteria for selecting regions of interest. Results obtained after a genome-wide association study with a dense marker map are made obscure by issues related to multiple testing and the high correlations between close markers due to LD. Multiple testing was controlled by applying stringent thresholds corresponding to P-values

of 5×10^{-4} , 5×10^{-5} or 5×10^{-6} . **The 41,249 tests performed were not independent due to the LD between SNPs and the most stringent threshold (5×10^{-6}) was chosen as an approximation of 10,000 independent tests corrected with Bonferroni (1936). The threshold at 5×10^{-5} was used to provide moderate evidence of association (Burton et al., 2007) and the threshold at 5×10^{-4} was used to describe and compare QTLs between traits.** Many of the markers with their P-values that exceeded the thresholds were actually correlated and could be grouped together. To avoid false detections resulting from the high correlation between markers in LD, the local score statistics proposed by Guedj et al. (2006) were used. The idea is to identify the chromosomal segments where the non-independent tests were on average higher than a threshold due to the presence of a unique QTL. The higher the threshold, the smaller the segment size but the greater the number of positive segments at the genome level. Since we wanted to take into account the error related to the detection of spurious QTLs within a given chromosomal segment (resulting from genotype correlation at a QTL and markers in LD with this QTL), we chose a low threshold (corresponding to 20% of non-zero tests). This step gave a limited number of chromosomal regions where all positive signals could be explained by a unique QTL. Within each detected region with several correlated signals, we chose to keep only the highest signal and identified it as the unique QTL.

RESULTS

Heritability and linkage disequilibrium

Heritabilities were estimated at 0.32 (sd \pm 0.14) for GM, 0.27 (sd \pm 0.13) for FM, 0.45 (sd \pm 0.15) for HM and 0.13 (sd \pm 0.11) for OM.

Figure 2 shows the estimated extent of LD (r^2). The FT curve shows the estimated LD in our population of French Trotters, while the other two curves (Ne150 and Ne1000) show the theoretical extent of LD in populations for which the effective sizes are respectively 150 and 1,000 individuals. The FT curve was closer to the Ne1000 curve for short distances and closer to the Ne150 curve for longer distances. This behavior suggests

that the effective population size of FT decreased over generations. Using the Bead-Chip EquineSNP50, the mean distance between two SNPs is 0.043Mb which corresponds to an average LD of 0.25 (Figure 2). In this mean situation, at worst, a QTL is at a distance of about 0.02Mb from its two flanking SNPs which corresponds to a 0.35 LD. In this situation, considering bi-allelic markers and QTLs, the detection power of our 525-individual design should reach **55, 78 and 91% if QTLs respectively explain 3, 5 and 7%** of the total phenotypic variance (Figure 3). Beyond this mean situation, set so that any present QTLs could be detected, it should be noted that the LD between two SNPs remains highly variable.

QTL detection

QQPlots. QQPlots informed us about the validity of the obtained P-values and the presence or absence of a population structure that might not have been taken into account in our models. Figure 4 shows the QQPlots obtained with SNPMixed and HaploIBD methods for GM, FM, HM and OM. Only the HaploIBD test on FM is slightly conservative, the others seem to follow the correct distribution.

Detected QTLs. The results of QTL detection with rejection thresholds of $P < 5 \times 10^{-4}$, $P < 5 \times 10^{-5}$ and $P < 5 \times 10^{-6}$ are presented in Table 3.

When the threshold was set at $P < 5 \times 10^{-4}$, the SNPMixed method gave on average a larger number of QTLs than HaploIBD (resp. 14.5 and 4.75). With this threshold, most of the QTLs detected with HaploIBD were also detected with SNPMixed. We obtained 4 QTLs in common to both methods for GM, 2 for FM, and 4 for HM and OM.

When the threshold was set at $P < 5 \times 10^{-5}$, more QTLs were again detected with SNPMixed than HaploIBD (resp. 4.25 and 1.25). Upon analysis of these QTLs, all except for one detected on OM were detected with both with the HaploIBD and SNPMixed methods. The QTL found in common to both methods for GM was located on ECA 13 (**Associated SNPs were BIEC2-208655 and BIEC2-208753 with respectively SNPMixed and HaploIBD**), the 2 QTLs found in common for HM were located on ECA 3 (**BIEC2-808617 and BIEC2-**

808442) and 14 (**BIEC2-265953 and BIEC2-265956**) and, the QTL found in common for OM was located on ECA 15 (**BIEC2-320636 and BIEC2-320532**).

When the threshold was set at $P < 5 \times 10^{-6}$, only one QTL, influencing HM, was detected on ECA 3 with HaploIBD. It was located between 100.39 Mb and 107.92 Mb with a maximum at 105.05 Mb. At this position, the P-value was $-\log(P) = 5.52$ and it was estimated that the QTL explained 7% of the total phenotypic variance. With the SNPMixed method, the same QTL was not significant at $P < 5 \times 10^{-6}$ but was significant at $P < 5 \times 10^{-5}$. It was located with this method between 102.03 Mb and 107.37 Mb with a maximum at 105.88 Mb. At this position, the P-value was $-\log(P) = 4.94$ and the SNP effect accounted for 0.28 of the standard phenotypic variation.

Manhattan plots obtained for GM, HM, FM and OM with HaploIBD are shown in Figure 5, 6, 7 and 8.

Detected QTLs between traits. QTLs that control distinct traits may be located closely, suggesting a pleiotropic effect. When the threshold was set at $P < 5 \times 10^{-4}$, two of the four QTLs detected for GM with both methods (SNPMixed and HaploIBD) were close to QTLs detected for FM (on ECA 13 and 15), and the two others were close to QTLs detected for HM (on ECA 3 and 14) (Table 4). With this threshold, no QTL regions were shared by FM, HM and OM.

DISCUSSION

Animal populations are often composed of related individuals and this is the case in our sample of FT. Association studies are sensitive to such population structures and it is important to take them into account in the models. The two association methods, SNPMixed and HaploIBD, used in this study were robust to population structure (relationship structure), firstly by the use of a relationship matrix, and secondly with the use of linkage analysis for HaploIBD (Balding, 2006). This ensured a better control of false positives in the analysis. **The QQPlots indicate that population structure had been accounted for by the analyses.** Multiple tests could also create many false positives. Indeed around 40,000 individual tests were performed. The

classical Bonferroni correction (Bonferroni, 1936) was too strict because it assumes independence between each test, a hypothesis not fulfilled in our case. Tests were correlated, especially when using the haplotypes for 6 markers position after position. This is the reason why we considered the use of different thresholds that were less stringent than Bonferroni's correction. There is no consensus on the choice of an ideal method or concerning SNP versus haplotype analysis. By using both of these approaches and combining their results, we expected to obtain a better accuracy and fewer false positives.

In this study, we detected one QTL on ECA 3 associated with osteochondrosis of the hock at the $P < 5 \times 10^{-6}$ significant level using a method which combined linkage analysis and linkage disequilibrium and the use of haplotypes. Moreover, other regions were also associated with suggestive QTLs for different traits but at lower thresholds.

No QTLs were found in common between HM and FM, an observation consistent with the low genetic correlations between OC of the hock and fetlock (Grøndahl and Dolvik, 1993; Stock and Distl, 2006b). All QTLs detected for the global measurement of OC with the threshold $P < 5 \times 10^{-4}$ were associated to a QTL detected for HM or FM but never both. These QTLs were located on ECA 3 and 14 for HM and on ECA 13 and 15 for FM. As we expected from its definition, GM combines FM and HM and thus reveals specific QTLs. Consistently with the higher incidence of cases observed for the fetlock than for the hock, one could expect that the global measurement GM be more indicative of OC on the fetlock. However, the average score per lesion was higher for the hock than for the fetlock (**resp. 2.62 and 2.05 before the log-transformation**), reversing thus the tendency.

Other QTL detection studies for OC were conducted in different breeds. Dierks et al. (2007) worked on Hanoverian warmblood horses and Wittwer et al. (2008) on South German coldblood horses. Both studies were performed using microsatellite markers in samples of approximately 200 horses. Most of the QTL regions detected in these studies were different and in fact only a QTL on ECA 16 was found to be significant for FM between these breeds. The recent availability of the Illumina Bead-Chip EquineSNP 50 has allowed for genome-wide studies of OC susceptibility. As a result, Lampe et al. (2010) have reanalyzed the data of Dierks et al. (2007) in the Hanoverian warmblood sample, and mainly evidenced different QTLs from those previously

found with microsatellite markers. More recently, Lykkjen et al. (2010) studied 162 Norwegian SB horses with SNPs and detected 9 regions of interest. No QTLs were found in common to the two previous studies. However, two QTLs on ECA 3 and 13 were found both in our study and one of the two previous ones, but at genome distances of about 5 Mb. Our study was based on a sample of 525 French Trotters, a population that is 2 to 3 times larger than in the two other OC studies using SNPs. As a result, the power of our design was certainly more efficient as to detecting small and medium QTL effects (variance of QTL < 5%). A possible explanation for the differences in the QTLs identified in the various studies could be that OC at different anatomical sites is associated with different loci in different breeds. Other explanations could include sample size or the various phenotypic criteria applied when selecting cases.

In conclusion, a few QTLs associated with OC at different anatomical sites were revealed in this study. We show that OC, a multifactorial disease, is influenced by several genes and we **didn't find** that the genetic determinism is identical for OC developed on hock and on fetlock.

Further studies will now focus on the identification of candidate genes and screening for mutation in an attempt to clarify the molecular physiopathology of OC and develop efficient strategies for risk assessment. Meanwhile, markers could be used in a marker-assisted selection context to improve horse health and welfare.

LITERATURE CITED

- Balding, D. J. 2006. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* 7: 781-791.
- Ball, R. D. 2005. Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. *Genetics* 170: 859-873.
- Bonferroni, C. E. 1936. *Teoria statistica delle classi e calcolo delle probabilita*. Libreria internazionale Seeber.
- Burton, P. R., D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, M. I. McCarthy, W. H. Ouwehand, and N. J. Samani. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661-678.**
- Denoix, J. M., J. P. Valette, P. Heiles, X. Ribot, and L. Tavernier. 2000. Radiographic survey of juvenile osteoarticular lesions in 3 year old french breed horses : General results on 1180 horses. *Pratique vétérinaire équine* 32: 35-41.
- Dierks, C., K. Löhring, V. Lampe, C. Wittwer, C. Drögemüller, and O. Distl. 2007. Genome-wide search for markers associated with osteochondrosis in hanoverian warmblood horses. *Mammalian Genome* 18: 739-747.

- Druet, T., S. Fritz, M. Boussaha, S. Ben-Jemaa, F. Guillaume, D. Derbala, D. Zelenika, D. Lechner, C. Charon, and D. Boichard. 2008. Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on bta03 using a dense single-nucleotide polymorphism map. *Genetics* 178: 2227-2235.
- Druet, T., and M. Georges. 2010. A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics* 184: 789-798.
- Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson. 2006. *Asreml user guide release 2.0*. VSN International Ltd., Hemel Hempstead, UK.
- Grøndahl, A. M., and N. I. Dolvik. 1993. Heritability estimations of osteochondrosis in the tibiotarsal joint and of bony fragments in the palmar/plantar portion of the metacarpo-and metatarsophalangeal joints of horses. *Journal of the American Veterinary Medical Association* 203: 101-104.
- Guedj, M., D. Robelin, M. Hoebeke, M. Lamarine, J. Wojcik, and G. Nuel. 2006. Detecting local high-scoring segments: A first-stage approach for genome-wide association studies. *Statistical Applications in Genetics and Molecular Biology* 5:1: Art. 22.
- Hill, W. G. 1974. Estimation of linkage disequilibrium in randomly mating populations. *Heredity* 33: 229-239.
- Jeffcott, L. B. 1991. Osteochondrosis in the horse—searching for the key to pathogenesis. *Equine Veterinary Journal* 23: 331-338.
- Jeffcott, L. B., and F. M. D. Henson. 1998. Studies on growth cartilage in the horse and their application to aetiopathogenesis of dyschondroplasia (osteochondrosis). *The Veterinary Journal* 156: 177-192.
- Lampe, V., K. Komm, P. Lichtner, T. Meitinger, and O. Distl. 2010. Genome wide association analysis for osteochondrosis in hanoverian warmblood horses using a snp assay, University of Veterinary Medicine Hannover.
- Luo, Z. W. 1998. Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. *Heredity* 80: 198-208.
- Lykkjen, S., N. I. Dolvik, M. E. McCue, A. K. Rendahl, J. R. Mickelson, and K. H. Roed. 2010. Genome wide association analysis of osteochondrosis of the tibiotarsal joint in norwegian standardbred trotters. *Animal Genetics* 41: 111-120.
- Meuwissen, T. H. E., and M. E. Goddard. 2000. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* 155: 421-430.
- Meuwissen, T. H. E., and M. E. Goddard. 2001. Prediction of identity by descent probabilities from marker-haplotypes. *Genetics Selection Evolution* 33: 605-634.
- Meuwissen, T. H. E., A. Karlsen, S. Lien, I. Olsaker, and M. E. Goddard. 2002. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* 161: 373-379.
- Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, D. H. Lee, V. Ducrocq, J. M. Elsen, and F. Minvielle. 2002. *Blupf90* and related programs (bgf90).
- Philipsson, J., E. Andreasson, B. Sandgren, G. Dalin, and J. Carlsten. 1993. Osteochondrosis in the tarsocrural joint and osteochondral fragments in the fetlock joints in standardbred trotters. II. Heritability. *Equine Veterinary Journal* 25: 38-41.
- Pong-Wong, R., A. W. George, J. A. Woolliams, and C. S. Haley. 2001. A simple and rapid method for calculating identity-by-descent matrices using multiple markers. *Genetics Selection Evolution* 33: 453-471.
- Schober, M., M. Coenen, O. Distl, B. Hertsch, L. Christmann, and E. Bruns. 2003. Estimation of genetic parameters of osteochondrosis (oc) in hanoverian warmblood foals. In: 54th EAAP Meeting, Italy
- Self, S. G., and K. Y. Liang. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 82: 605-610.
- Stock, K. F., and O. Distl. 2006a. Genetic correlations between conformation traits and radiographic findings in the limbs of german warmblood riding horses. *Genetics, Selection, Evolution: GSE* 38: 657-671.
- Stock, K. F., and O. Distl. 2006b. Genetic correlations between osseous fragments in fetlock and hock joints, deforming arthropathy in hock joints and pathologic changes in the navicular bones of warmblood riding horses. *Livestock Science* 105: 35-43.

- Stock, K. F., H. Hamann, and O. Distl. 2005a. Estimation of genetic parameters for the prevalence of osseous fragments in limb joints of hanoverian warmblood horses. *Journal of Animal Breeding and Genetics* 122: 271-280.
- Stock, K. F., H. Hamann, and O. Distl. 2005b. Prevalence of osseous fragments in distal and proximal interphalangeal, metacarpo and metatarsophalangeal and tarsocrural joints of hanoverian warmblood horses. *Journal of Veterinary Medicine Series A* 52: 388-394.
- Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke, M. E. Goddard, and P. M. Visscher. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Research* 17: 520-526.
- Trotter, G. W., and C. W. McIlwraith. 1981. Osteochondrosis in horses: Pathogenesis and clinical syndromes. In: *Proc Ann Conv Am Ass Equine Pract.* p 141-160.
- van Grevenhof, E. M., A. Schurink, B. J. Ducro, P. R. van Weeren, J. M. F. M. van Tartwijk, P. Bijma, and J. A. M. van Arendonk. 2009. Genetic parameters of various manifestations of osteochondrosis and their correlations between and within joints in dutch warmblood horses. *Journal of animal science* 87: 1906-1912.
- van Weeren, P. R. 2006. Etiology, diagnosis, and treatment of oc (d). *Clinical Techniques in Equine Practice* 5: 248-258.
- Wittwer, C., C. Dierks, H. Hamann, and O. Distl. 2008. Associations between candidate gene markers at a quantitative trait locus on equine chromosome 4 responsible for osteochondrosis dissecans in fetlock joints of south german coldblood horses. *Journal of Heredity* 99: 125-129.
- Wittwer, C., H. Hamann, and O. Distl. 2009. The candidate gene *xirp2* at a quantitative gene locus on equine chromosome 18 associated with osteochondrosis in fetlock and hock joints of south german coldblood horses. *Journal of Heredity* 100: 481-486.
- Wittwer, C., K. Löhring, C. Drögemüller, H. Hamann, E. Rosenberger, and O. Distl. 2007. Mapping quantitative trait loci for osteochondrosis in fetlock and hock joints and palmar/plantar osseus fragments in fetlock joints of south german coldblood horses. *Animal Genetics* 38: 350-357.

TABLES AND FIGURES

Table 1: Frequency of osteochondrosis binary measurements

Traits	Code	No.	Cases	Controls
Fetlock	FM	525	0.32	0.68
Hock	HM	525	0.23	0.77
Others	OM	525	0.11	0.89

Table 2: Phenotypic correlation (Pearson) between traits¹

	GM	FM	HM	OM
GM	1	0.64	0.63	0.43
FM	-	1	0.08	0.12
HM	-	-	1	0.13
OM	-	-	-	1

¹**GM, global measurement of osteochondrosis; FM, fetlock measurement; HM, hock measurement; OM, osteochondrosis other than on the fetlock and the hock**

Table 3: Number of detected QTLs at different thresholds

Methods ¹	Traits ²	$P < 5 \times 10^{-4}$	$P < 5 \times 10^{-5}$	$P < 5 \times 10^{-6}$	Max ³
SNPMixed	GM	12	3	0	4.92
	FM	13	3	0	4.63
	HM	15	5	0	5.11
	OM	18	6	0	5.02
HaploIBD	GM	5	1	0	4.46
	FM	3	0	0	3.83
	HM	5	2	1	5.52
	OM	6	2	0	5.17
Combined	GM	4	1	-	-
	FM	2	-	-	-
	HM	4	2	-	-
	OM	4	1	-	-

¹**SNPMixed, SNP mixed-model analyses; HaploIBD, haplotype mixed-model analyses; Combined, QTLs detected with both SNPMixed and HaploIBD methods**

²**GM, global measurement of osteochondrosis; FM, fetlock measurement; HM, hock measurement; OM, osteochondrosis other than on the fetlock and the hock**

³**Max, maximum $-\log_{10}(P)$**

Table 4: QTLs between traits

QTL	Trait ¹	ECA ²	HaploIBD ³			SNPMixed ⁴		
			Bounds (Mb)	Pos (Mb)	Max	Bounds (Mb)	Pos (Mb)	Max
1	GM	3	104.94 - 110.12	105.95	3.64	105.13 - 110.01	106.05	3.48
	HM	3	100.39 - 107.92	105.05	5.52	102.03 - 107.37	105.88	4.94
2	GM	13	0.22 - 11.33	8.49	4.46	0.14 - 11.33	8.39	4.89
	FM	13	6.92 - 12.88	9.74	3.83	9.4 - 11.54	9.89	3.89
3	GM	14	68.00 - 78.91	73.63	3.90	68.87 - 79.44	74.08	4.75
	HM	14	67.97 - 77.90	73.87	4.47	67.94 - 76.02	73.76	5.09
4	GM	15	86.30 - 89.75	87.35	3.42	85.07 - 89.75	89.55	3.68
	FM	15	87.10 - 88.68	87.35	3.36	87.31 - 88.58	87.61	3.30

¹**GM, global measurement of osteochondrosis; FM, fetlock measurement; HM, hock measurement**

²**ECA, *Equus caballus* chromosome**

³**HaploIBD, haplotype mixed-model analyses; Pos, location of the QTL on the chromosome; Mb, Megabase; Max, maximum of -log₁₀(P)**

⁴**SNPMixed, SNP mixed-model analyses**

Figure 1. Distribution of the osteochondrosis Global Measurement (GM)

Figure 2. Extent of LD (r^2) in French Trotters (FT) assessed by using 525 samples from EquineSNP50 genotypes. The curve Ne150 (resp. Ne1000) represents the theoretical extent of LD in a population of effective size equal to 150 (resp. 1000)

Figure 3. Theoretical power of association analysis in function of LD between a bi-allelic marker and a QTL, explaining 3, 5 or 7% of the total phenotypic variance in a population size of 525 individuals

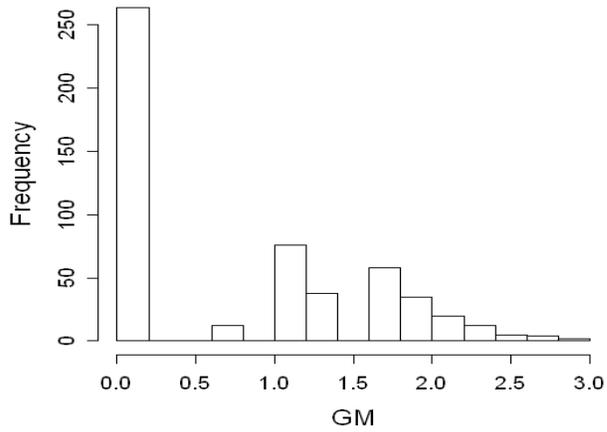
Figure 4. Quantile-quantile (QQ) plots of $-\log_{10}$ of P-values resulting from haplotype mixed-model analyses (first line) and SNP mixed-model analyses (second line). Each column represents a measurement of osteochondrosis (OC): all sites, fetlock, hock and other than fetlock and hock respectively. Deviations from the slope line correspond to loci that deviate from the null hypothesis of no association

Figure 5. Manhattan plot of $-\log_{10}$ of P-values for the global measurement (GM) of osteochondrosis (scores based on all radiographic findings). The plot displays the haplotype mixed-model test results for GM at each SNP position on each chromosome

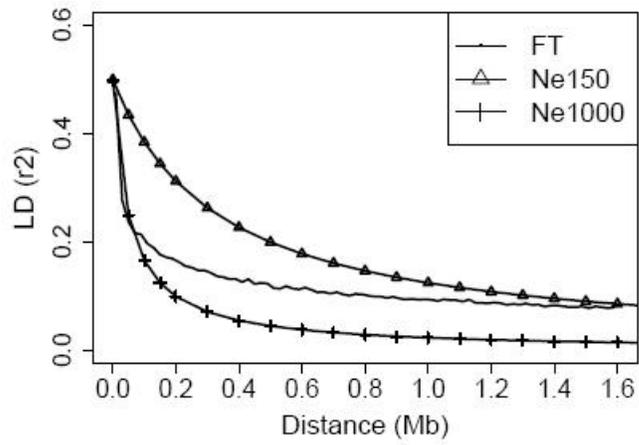
Figure 6. Manhattan plot of $-\log_{10}$ of P-values for the hock measurement (HM) of osteochondrosis (presence or absence of osteochondrosis on the hock). The plot displays the haplotype mixed-model test results for HM at each SNP position on each chromosome

Figure 7. Manhattan plot of $-\log_{10}$ of P-values for the fetlock measurement (FM) of osteochondrosis (presence or absence of osteochondrosis on the fetlock). The plot displays the haplotype mixed-model test results for FM at each SNP position on each chromosome

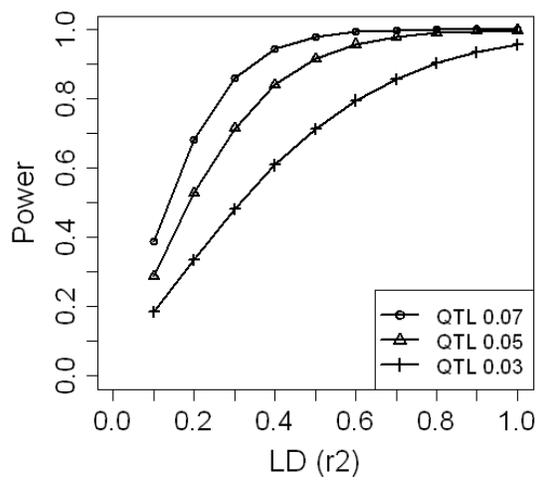
Figure 8. Manhattan plot of $-\log_{10}$ of P-values for the other measurement (OM) of osteochondrosis (presence or absence of osteochondrosis other than on the hock and the fetlock). The plot displays the haplotype mixed-model test results for FM at each SNP position on each chromosome



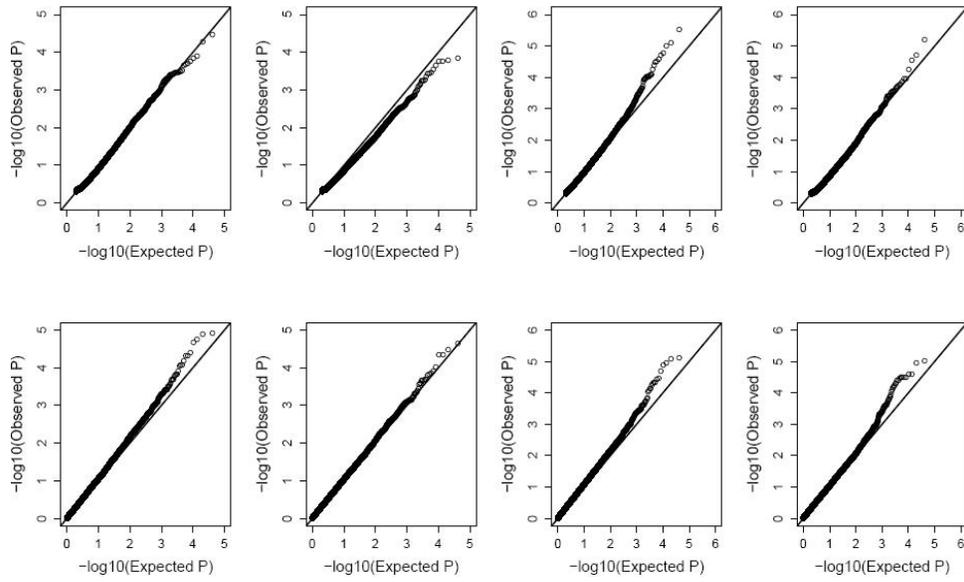
JAS-E-2011-4031, Figure 1

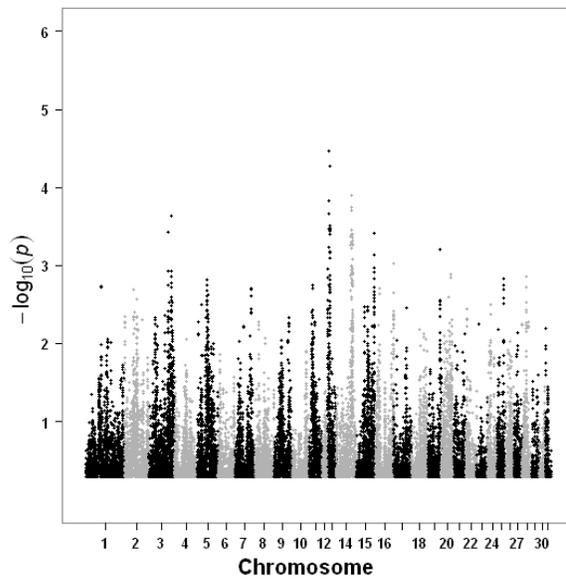


JAS-E-2011-4031, Figure 2

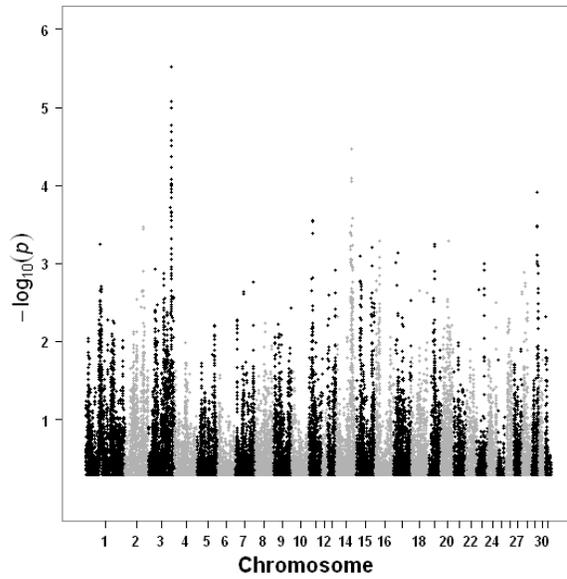


JAS-E-2011-4031, Figure 3

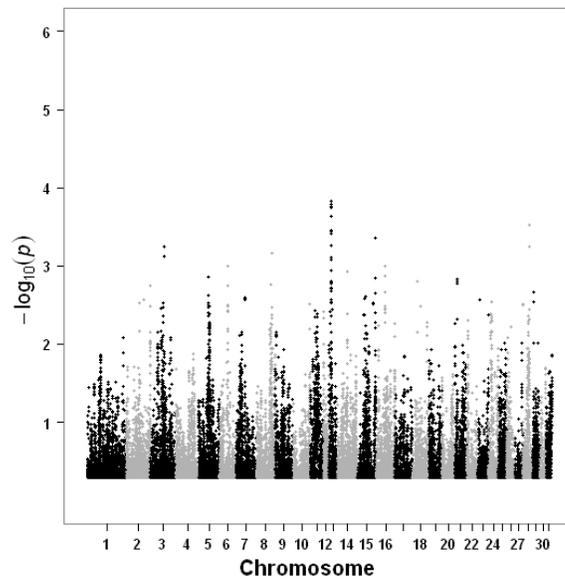
**JAS-E-2011-4031, Figure 4**



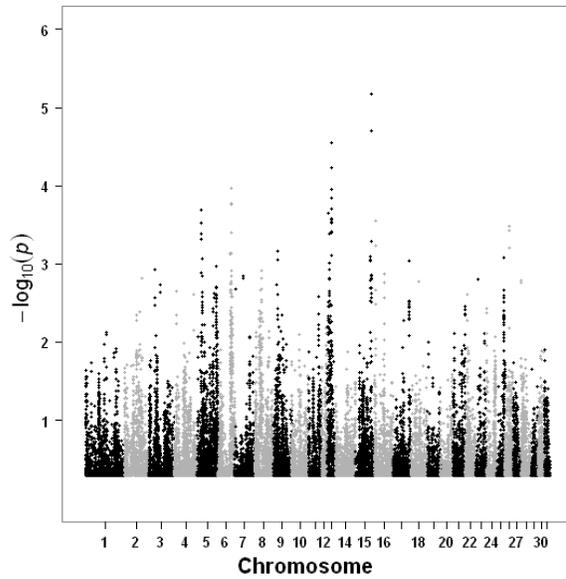
JAS-E-2011-4031, Figure 5



JAS-E-2011-4031, Figure 6



JAS-E-2011-4031, Figure 7



JAS-E-2011-4031, Figure 8

4.3 Compléments sur les phénotypes

L'article "Genome-wide association studies for osteochondrosis in French trotters" résume les principaux résultats obtenus au début de l'année 2011. Pour compléter ces résultats, nous nous proposons de détailler le choix des phénotypes retenus à partir des données radiographiques de base (689 chevaux), de discuter la qualité des données génotypiques issues du génotypage par la puce Illumina et de présenter des résultats complémentaires sur un effectif plus important (583 chevaux) que celui utilisé dans l'article (525 chevaux).

4.3.1 Description des données de base

Radiographies

L'ensemble des radiographies, prises dans diverses cliniques vétérinaires, ont été relues par deux vétérinaires spécialistes en orthopédie équine (Prof. Jean-Marie Denoix et Marie-Capucine Dupuis-Tricaud). Différents sites radiographiques ont été étudiés afin de visualiser l'ensemble des zones à risque des chevaux, à savoir les doigts antérieurs et postérieurs, le jarret, le carpe et le grasset (Figure 4.1).

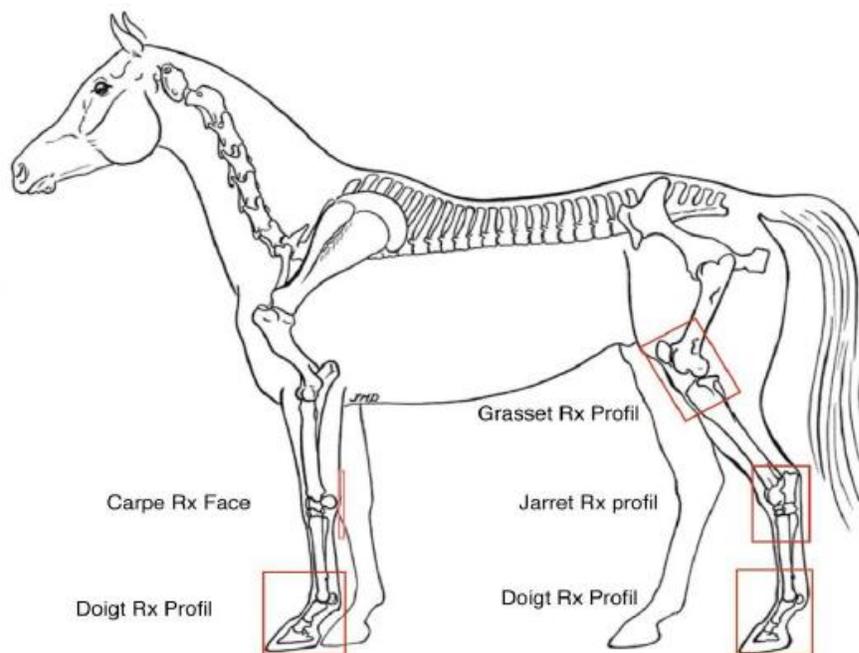


FIGURE 4.1 – Bilan radiographique GENEQUIN (d'après JM Denoix)

Les sites radiographiques étudiés pour le membre antérieur, au nombre de 6, sont l'articulation inter phalangienne proximale (IPP), l'articulation inter phalangienne distale (IPD), l'articulation métacarpo-phalangienne (AMcP), l'os sésamoïde distal (Sés dist), la troisième phalange (P dist) et d'autres parties du pied (autres). Pour les doigts postérieurs, on trouve 3 sites radiographiques qui sont le pied postérieur (Pied), l'articulation métatarso-phalangienne (Amt Ph) et l'articulation

inter phalangienne (IPP). Un seul cliché a été réalisé pour chacun des sites du jarret, du grasset et du carpe. Pour chacun des sites radiographiques étudiés, un cliché a été pris pour les membres de gauche et de droite.

Pour chacun des sites radiographiés, le praticien décrit l'anomalie observée et lui attribue un score en fonction de la gravité de la lésion. Le score est basé sur 4 type de graduation, le grade 1 correspondant à une petite lésion et le grade 4 correspondant à une lésion importante. La fonction $s = 2^{g-1}$, relie le grade g de chacune des lésions observées au score s associé. Une lésion a donc un score pouvant aller de 1 (petite lésion) à 8 (lésion importante). L'anomalie décrite est identifiée comme étant ou non de l'ostéochondrose, et provenant de 3 types de contraintes mécaniques sur les articulations : cisaillement, compression ou tension. Seule l'arthropathie juvénile ne relève pas d'une contrainte particulière et constituera donc une classe propre parmi les contraintes. Les données de base sont donc, pour chaque cheval, un tableau de deux fois 12 sites articulaires (gauche et droite) avec pour chaque site, le score de la lésion d'ostéochondrose éventuelle (0 si pas de lésion) et la cause mécanique de son origine. Il s'agit maintenant de synthétiser cette information en un ou plusieurs phénotypes par cheval qui à la fois résume l'information sans la dénaturer ou faire trop d'hypothèses sur les mécanismes biologiques et de choisir l'échantillon le plus intéressant à génotyper pour la recherche de QTL.

Description des données de base

Coordonné par le CIRALE (Centre d'Imagerie et de Recherche sur les Affections Locomotrices Equines), le recueil d'une cohorte de chevaux radiographiés sur place ou chez des vétérinaires praticiens s'est étalé sur deux années. L'objectif premier était d'avoir une population d'une même race (le Trotteur français) jeune, homogène en âge, issus d'un nombre restreint d'étalons pour favoriser les études familiales et aux phénotypes les plus extrêmes possibles. Finalement, c'est une population totale de 689 chevaux qui a été radiographiée.

Distribution des scores : Dans cette étude, on observe 725 lésions, soit 4.2% des radiographies. Parmi elles : 25% sont des lésions de grade 1 et correspondent à des lésions de très petite taille (score associé égal à 1) ; 55% sont des lésions de grade 2 et correspondent à des lésions de petite taille (score associé égal à 2) ; 19% sont des lésions de grade 3 et correspondent à des lésions de taille moyenne (score associé égal à 4) ; 1% sont des lésions de grade 4 et correspondent à des lésions de taille importante (score associé égal à 8).

Critères de regroupement : Pour synthétiser les 24 scores observés par cheval, il est possible de les regrouper soit par site articulaire soit par type de contrainte. D'autre part, lors de leur regroupement, les scores peuvent être soit sommés, soit permettre la création de 2 catégories : les sommes à 0 constituant le groupe des "sains" et toute somme > 0 constituant le groupe des "atteints". Trois considérations orientent ces choix :

- Le nombre de chevaux lésés par caractère doit être suffisant pour réaliser les analyses statistiques. C'est d'ailleurs aussi une des raisons de ne pas conserver les 24 variables élémentaires.
- Pris par paires, l'indépendance entre les valeurs élémentaires de 2 des 24 scores suggèrera l'existence de deux caractères différents, et donc ne militera pas pour le regroupement,

inversement une forte corrélation, même si elle n'est à ce stade que phénotypique, incitera au regroupement.

- Sommer les scores engendre de nombreuses classes de valeurs qui sont parfois trop peu fréquentes pour être analysées dans le détail. Dans ce cas, regrouper toutes les sommes > 0 est justifié même si il y a une perte potentielle d'information.

Nous discuterons d'abord des regroupements opérés par site articulaire puis par type de contrainte.

Regroupement par site articulaire

Le premier regroupement testé est celui des articulations semblables situées à droite et à gauche de l'animal. Une forte corrélation a été trouvée (par exemple, 0.5 sur le jarret) entre les lésions observées dans la même articulation à gauche et à droite, suggérant qu'un animal atteint d'un côté a plus de chance de l'être aussi de l'autre côté qu'un animal sain sur ce premier coté. Cette association n'a pu être testée sur toutes les articulations compte tenu des effectifs mais par homogénéité, toutes les articulations gauches et droites ont été regroupées.

Le second regroupement concerne l'articulation du boulet antérieure (AMcP) et postérieure (Amt Ph). Ces articulations sont assez similaires fonctionnellement et la corrélation (environ 0.2) montre qu'un regroupement est pertinent.

La faible fréquence d'apparition de lésions (moins de 0.5% par site) dans la troisième phalange (P dist), le paturon (IPP, IPD) et l'os sésamoïde dans les membres antérieurs et dans le pied et le paturon (IPP) postérieur oblige à les regrouper dans un ensemble (autres).

La fréquence de chevaux atteints d'au moins une lésion sur le grasset (7%) et le carpe (5%) est moins anecdotique mais il ne s'agit finalement que de 47 et respectivement 37 chevaux et nous avons donc préféré les regrouper avec l'ensemble des sites articulaires précédents.

Finalement nous avons donc regroupé les lésions d'un cheval en trois régions principales : le boulet, qui comprend donc les boulets gauche et droit et antérieur et postérieur, le jarret (gauche et droit) et tous les autres sites articulaires regroupés sous la dénomination "autres". Pour ces trois caractères, la mesure retenue est une variable binaire : 0 pour une somme des scores nulle et 1 pour une somme de score non nulle. Les fréquences des chevaux atteints sont dans le Tableau 4.1.

TABLE 4.1 – Tableau des effectifs et fréquences des caractères par site articulaire

Caractères	Effectifs chevaux		Fréquences chevaux	
	Atteints	Sains	Atteints	Sains
Boulet	246	443	0,36	0,64
Jarret	161	528	0,23	0,77
Autres	98	591	0,14	0,86

Regroupement par contrainte mécanique

Pour le regroupement par contrainte mécanique, 4 caractères ont été retenus : cisaillement, compression, tension et arthropathie juvénile. Pour ces caractères, on note surtout que la fréquence d'atteints pour l'arthropathie juvénile est très faible (2%) et sera abandonnée pour la recherche de QTL. Outre l'arthropathie juvénile, on ne fera qu'une partie des analyses pour les caractères tension et compression compte tenu de leurs faibles fréquences. Pour les 3 caractères étudiés (Cisaillement,

Tension et Compression), la mesure retenue est aussi une variable binaire : 0 pour une somme des scores nulle et 1 pour une somme des scores non nulle. Les fréquences des chevaux atteints sont présentées dans le Tableau 4.2.

TABLE 4.2 – Tableau des effectifs et fréquences des caractères par contrainte mécanique

Caractères	Effectifs chevaux		Fréquences chevaux	
	Atteints	Sains	Atteints	Sains
Cisaillement	311	378	0,45	0,55
Compréssion	123	566	0,18	0,82
Tension	54	635	0,08	0,92
ArthropJuv	15	674	0,02	0,98

Regroupement global

Une variable qui regroupe l'ensemble des 24 scores peut également être construite. Pour ce caractère deux mesures ont été retenues. Comme précédemment, une mesure binaire appelée "Total" avec 0 pour une somme des scores nulle (cheval "sain") et 1 pour une somme des scores non nulle (cheval "atteint"). Mais en regroupant tous les sites, on obtient une variabilité des sommes des scores des chevaux atteints suffisante pour être étudiée contrairement aux regroupements plus limités précédents. On appelle donc score global (SG) la somme des scores associés à chaque site radiographique. Les modèles statistiques utilisés pour la détection de QTL font souvent l'hypothèse de la normalité des données, ce qui n'est pas le cas de SG. De ce fait, nous avons préféré utiliser le caractère LSG qui est la transformation logarithmique de $1 + SG$. La Figure 4.2 donne les distributions de SG et LSG dans la population phénotypée.

Relations entre caractères

Nous avons finalement défini 9 caractères différents pour qualifier la présence d'ostéochondrose chez un cheval : 3 caractères binaires en fonction des sites articulaires, 4 caractères binaires en fonction des contraintes mécaniques appliquées à l'os, et 2 caractères globaux dont 1 binaire et 1 continu liés à la somme des scores. Pour tous les caractères binaires, tous les chevaux ayant au moins une lésion sont affectés du même phénotype, le cheval est "atteint". Pour le caractère continu on distingue des différences entre les chevaux atteints en fonction du nombre de site et de la gravité des lésions. Quelles sont les relations entre ces caractères et est-il pertinent de conserver autant de caractères différents ?

Liens entre les types de caractères : Intéressons nous tout d'abord aux tableaux croisés de la répartition des lésions d'OC observées pour les caractères par sites articulaires et pour les caractères basés sur les types de contrainte. Le Tableau 4.3 montre ainsi que le regroupement en type de contrainte cisaillement est majoritairement présent sur les sites du boulet et du jarret (99%). Pour les phénomènes de compression, on les retrouve un peu partout mais plus particulièrement ailleurs que sur les boulets et les jarrets (58%). Enfin, les phénomènes de tension et l'arthropathie juvénile sont principalement situés sur les sites articulaires autres que celui du jarret.

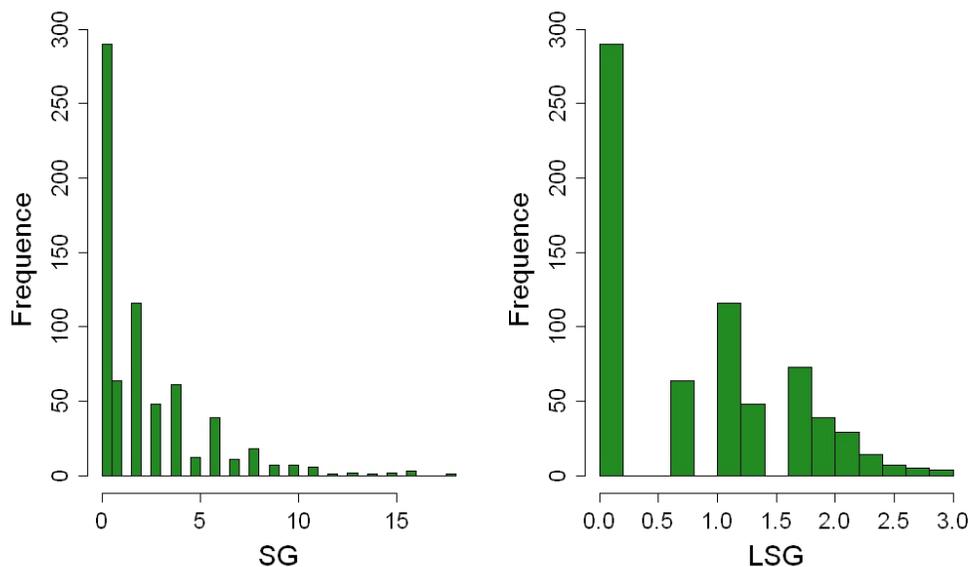


FIGURE 4.2 – Distribution des mesures SG et LSG du caractère global de l'OC

TABLE 4.3 – Tableau des fréquences des lésions par type de contrainte en fonction des sites articulaires étudiés

AOAJ	Nb Lésions	Fréquences		
		Boulet	Jarret	Autres
Cisaillement	469	0,54	0,45	0,01
Compression	163	0,28	0,16	0,56
Tension	62	0,48	0,04	0,48
ArthopJuv	31	0,61	0,07	0,32
Total	725	0,48	0,33	0,19

Le sens inverse, présenté dans le Tableau 4.4, montre que la plupart des lésions sur les articulations du boulet, du jarret et au niveau global sont de type cisaillement. Ailleurs que sur le boulet et le jarret, on retrouve principalement des lésions de type compression.

En regardant de près ces tableaux croisés, il semble difficile d'étudier les regroupements en type de contrainte un à un car la fréquence de certains d'entre eux est vraiment trop faible. L'ensemble des lésions est principalement de type cisaillement ou compression (87%). Etant donné les fréquences très élevées du cisaillement sur le boulet et le jarret, étudier le caractère boulet et jarret revient en fait à étudier le cisaillement sur le boulet et sur le jarret. De la même manière, étudier le caractère autres semble très fortement lié au fait d'étudier les phénomènes de compression. Le Tableau 4.5 montre les corrélations entre ces caractères et confirme ces liens. Enfin, étudier la variable globale "Total" se rapproche de la possible étude du caractère cisaillement. De ce fait, nous avons préféré étudier les caractères propres à chaque site articulaire, c'est à dire, au niveau global (Total ou LSG), le boulet, le jarret et ailleurs que sur le boulet et le jarret (autres).

TABLE 4.4 – Tableau des fréquences des lésions par sites articulaires étudiés en fonction des types de contrainte

Sites	Nb Lésions	Fréquences			
		Cisaillement	Compression	Tension	ArthropJuv
Boulet	347	0,73	0,13	0,09	0,05
Jarret	241	0,88	0,11	0,01	0,01
Autres	137	0,04	0,67	0,22	0,07
Total	725	0,65	0,22	0,09	0,04

TABLE 4.5 – Tableau des corrélations entre les caractères

	Boulet	Jarret	Autres	Cisaillement	Compression	Tension	ArthropJuv
Boulet	1						
Jarret	-0,02	1					
Autres	0	0,05	1				
Cisaillement	0,63	0,54	0,03	1			
Compression	0,1	0,18	0,69	0,04	1		
Tension	0,25	0,03	0,28	0,05	-0,03	1	
ArthropJuv	0,18	0,04	0	0,1	0,03	0	1

Corrélations avec les caractères binaires : Le Tableau 4.6 donne les corrélations obtenues entre les mesures globales, quantitative pour LSG et binaire pour Total, et les caractères des différents sites articulaires. Au vu des fortes corrélations, les deux mesures globales LSG et Total sont le reflet des différents autres caractères. Il sera certainement possible de détecter les QTLs spécifiques à un site articulaire avec ces mesures.

TABLE 4.6 – Tableau des corrélations entre mesures globales et sites articulaires

	Total	Boulet	Jarret	Autres
LSG	0.87	0.56	0.58	0.38
Total	1	0.64	0.47	0.35

Place et choix de la suppression de chevaux intermédiaires

Certains chevaux, ayant un SG égal à 1, ou ayant un SG égal à 0 ou 2 mais dont les lectures radiographiques sont douteuses, ont été considérés comme intermédiaires. Parmi notre population de 689 chevaux, 105 sont dans ce cas là. Pour éviter l'obtention de résultats erronés dus à une mauvaise classification de certains chevaux, seul certains cas intermédiaires ont par la suite été génotypés. Nous avons retenu 16 cas intermédiaires dont le génotype du père était disponible et dont les descendants de ce père étaient peu nombreux ou de même classe (i.e. SG égal à 0 ou SG > 2).

Ce choix ne devrait pas nous pénaliser pour la recherche de QTL car il a été montré que de génotyper les individus extrêmes apporte la plupart de la puissance (Darvasi and Soller, 1992;

Lander and Botstein, 1989). Néanmoins, en supprimant une partie des classes intermédiaires, l'effet du QTL est surestimé (Darvasi and Soller, 1992).

Conclusion : caractères retenus et effectifs analysés

Il y a finalement une assez forte confusion entre les caractères regroupés par site articulaire et par contrainte mécanique. C'est pourquoi, dans l'article présenté précédemment, nous nous sommes limités aux caractères par site articulaire afin d'éviter la multiplicité des tests qui est un des gros problèmes de la recherche de QTL. Cependant, dans un souci d'exhaustivité, nous rapportons ci-après l'ensemble des résultats obtenus sur tous les caractères.

L'analyse, effectuée uniquement sur les individus phénotypés et génotypés, a été réalisée à partir d'un échantillon de 525 chevaux (population P_1). Les distributions des caractères sur cet échantillon sont données dans le Tableau 1 et la Figure 1 de l'article précédent. Plus récemment, 58 autres phénotypes et génotypes ont été effectués et nous permettent d'analyser un échantillon de 583 TF (population P_2). Le Tableau 4.7 donne les fréquences des sains et atteints de ce nouvel échantillon. Pour la mesure LSG, dans cet échantillon, la moyenne était de 0.84 (écart-type de ± 0.87).

TABLE 4.7 – Tableau des effectifs et fréquences des chevaux en fonction des différents caractères

Caractères	Effectifs chevaux		Fréquences chevaux	
	Atteints	Sains	Atteints	Sains
Total	308	275	0,53	0,47
Boulet	195	388	0,33	0,67
Jarret	143	440	0,25	0,75
Autres	73	510	0,13	0,87
Cisaillement	259	324	0,44	0,56
Compression	94	489	0,16	0,84
Tension	38	545	0,07	0,93

4.3.2 Effets environnementaux

Liste des facteurs de variation environnementaux possibles

Nous avons vu dans le chapitre 3 que l'OC est une maladie multifactorielle et plusieurs effets environnementaux semblent jouer un rôle important dans sa pathogénie. On retrouve parmi eux en plus d'un effet génétique, des effets liés à l'entraînement, à l'âge, au sexe, au milieu (taille de l'aire, qualité des sols), à la nutrition du cheval etc... Certains d'entre eux sont facilement mesurables comme l'âge, le sexe, l'entraînement quand d'autres sont beaucoup moins, notamment la nutrition. Dans notre échantillon de la population TF, nous avons recueilli les informations de l'âge, du sexe et de l'entraînement des chevaux. Néanmoins, la collecte des données sur l'entraînement du cheval n'ayant pas été faite de façon homogène dans les diverses cliniques vétérinaires, nous avons choisi de ne pas inclure cet effet.

Distribution des facteurs environnementaux sélectionnés

Dans nos analyses de recherche de QTL, nous avons utilisé les échantillons P_1 de 525 et P_2 de 583 chevaux. Dans ces deux études, seuls les effets de l'âge au contrôle et du sexe n'ont pu être mesurés comme effet environnemental. On trouve plus de mâles que de femelles (Figure 4.3). Pour l'âge au contrôle, dont la distribution est présentée Figure 4.4, on note qu'environ 80% des chevaux avaient entre 1 an et 4 ans au moment du contrôle. L'idéal serait d'avoir des chevaux qui soient ni trop jeunes pour que les lésions d'OC puissent être définitives (autour de l'âge de 1 an) et ni trop vieux pour éviter qu'ils aient commencé leur entraînement (autour de l'âge de 3 ans), susceptible de créer également des lésions de type OC. Dans nos études, environ 6% des chevaux ont moins de 1 an et 28% ont plus de 3 ans. Beaucoup de classes d'âges sont peu fréquentes, ce qui peut poser un problème dans une analyse de variance. L'idée est alors de grouper certaines de ces classes d'âges. La Figure 4.5 présente la distribution des classes de 2 ans et moins et de plus de 2 ans qui sera par la suite utilisée pour tester l'effet fixe de l'âge au contrôle.

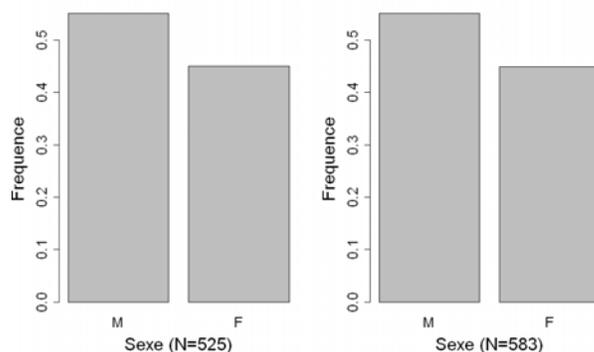


FIGURE 4.3 – Distribution des sexes des échantillons analysés (N=525 et N=583)

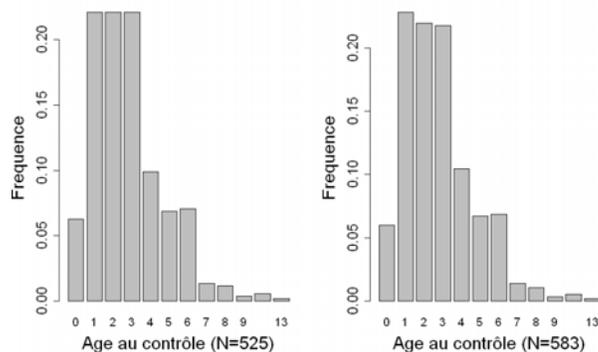


FIGURE 4.4 – Distribution des âges au moment du contrôle des échantillons analysés (N=525 et N=583)

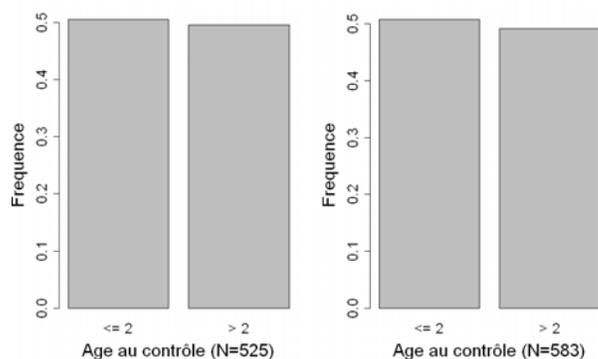


FIGURE 4.5 – Distribution des groupes d’âges au contrôle des échantillons analysés (N=525 et N=583)

Test des effets fixes dans notre population

Les effets fixes des groupes d’âges et du sexe ont été testés dans un modèle mixte incluant une moyenne, l’effet fixe testé en facteur, un effet aléatoire polygénique et une résiduelle. La matrice de variance-covariance A associé à l’effet polygénique a été créée à partir d’un pedigree de 2796 chevaux. Le Tableau 4.8 donne les résultats sur la significativité des tests pour les caractères étudiés. On note que l’effet des groupes d’âges sur les caractères jarret et LSG est significatif et nous conduit à prendre en compte ces effets dans les modèles pour la recherche de QTL. Pour ces deux caractères, on trouve que les individus âgés sont en moyenne moins lésés que les plus jeunes. Pour l’effet du sexe, aucun résultat n’est significatif.

TABLE 4.8 – Tableau des effets fixes sur les caractères étudiés

Caractères	Effets fixes	
	Sexe	Groupe d’âges
Total	NS	NS
Boulet	NS	NS
Jarret	NS	***
Autres	NS	NS
LSG	NS	*
Cisaillement	NS	NS
Compression	NS	NS
Tension	NS	NS

4.3.3 Description de la structure de parenté

Les données comprennent au total 689 TF phénotypés. Ces chevaux sont répartis dans 191 familles de pères dont la taille varie entre 1 et 23 descendants par famille. En moyenne, le nombre de descendants par famille est de 3.60 (± 3.98 e.t.) mais cette moyenne est due au fort nombre

d'étalons avec un seul descendant (79). En gardant uniquement les pères avec plusieurs descendants, la moyenne des descendants par familles est alors de 5.45.

La structure de parenté entre les individus de la population P_1 est décrites dans l'article. Dans la population P_2 , les 583 chevaux sont répartis dans 175 familles de pères dont la taille varie entre 1 et 20 descendants par famille. En moyenne, le nombre de descendants par famille est de 3.33 (± 3.28 e.t.) mais cette moyenne est due au fort nombre d'étalons avec un seul descendant (68). En gardant uniquement les pères avec plusieurs descendants, la moyenne des descendants par famille est alors de 4.81.

4.4 Compléments sur les marqueurs et les génotypes utilisés

4.4.1 Description de la puce Illumina BeadChip EquineSNP50

Les chevaux ont été génotypés à l'aide de la puce Illumina BeadChip EquineSNP50 qui contient 54602 marqueurs répartis sur l'ensemble du génome. La Figure 4.6 donne la distribution des SNPs par chromosome. En moyenne sur l'ensemble du génome, deux SNP sont séparés d'une distance de 0.043 ± 0.054 Mb. Cette moyenne est très peu variable entre les chromosomes puisque l'écart-type des moyennes des distances entre deux SNP adjacents entre les chromosomes est de ± 0.001 Mb.

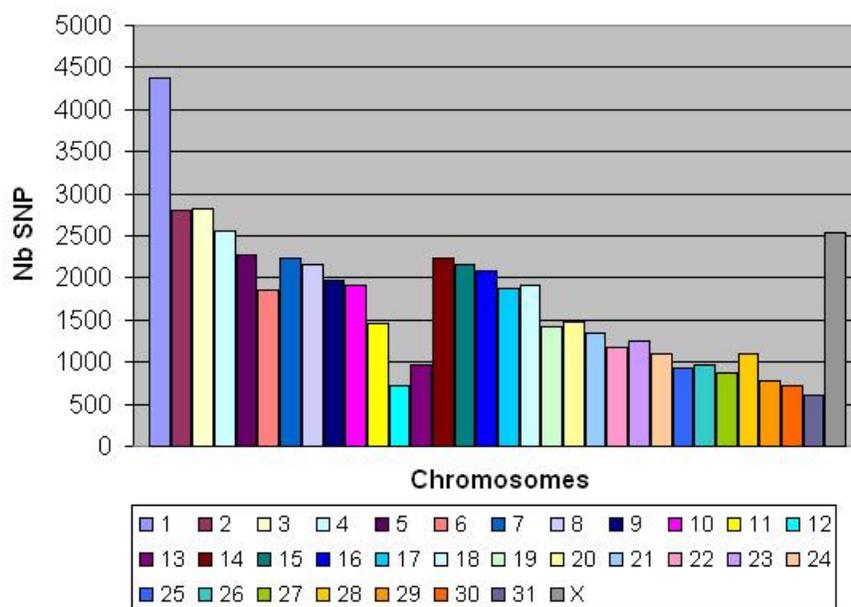


FIGURE 4.6 – Distribution des SNPs sur l'ensemble des chromosomes ECA

4.4.2 Qualité des typages

Taux d'échec des génotypages : *call freq* et *call rate*

Un premier filtrage dans le contrôle de la qualité de données des génotypes, qui permet de s'affranchir d'éventuelles erreurs qui risquent de ce produire, est la suppression des marqueurs et

des individus avec un taux d'échec au génotypage trop important (respectivement appelé *call freq* et *call rate*). En moyenne, dans nos données, le *call freq* est de 0.99 et nous avons choisi un seuil à 0.8. Seuls 675 SNP ne passaient pas ce seuil et ont donc été supprimés pour les analyses. En ce qui concerne le génotypage des individus, un total de 591 chevaux ont été typés mais 8 d'entre eux n'ont pas passé le seuil du *call rate* de 0.98 et ont été supprimés pour les analyses.

Filtrage sur la MAF et les tests HWE

La Figure 4.7 donne la distribution de la fréquence de l'allèle mineur (MAF) pour l'ensemble des marqueurs sur les chromosomes autosomes (soit 52063 SNP). Un total de 9801 SNP (dont 2850 SNP monomorphes) ont une $MAF < 0.05$ et ont été supprimés car de faibles MAF peuvent créer des problèmes d'estimations de l'effet du SNP ou d'haplotypes dans les analyses de recherche de QTL. La moyenne des MAF pour l'ensemble des marqueurs du génome est de 0.23 ± 0.15 et cette moyenne est très homogène entre les chromosomes puisque l'écart-type des moyennes entre chromosomes est de 0.009.

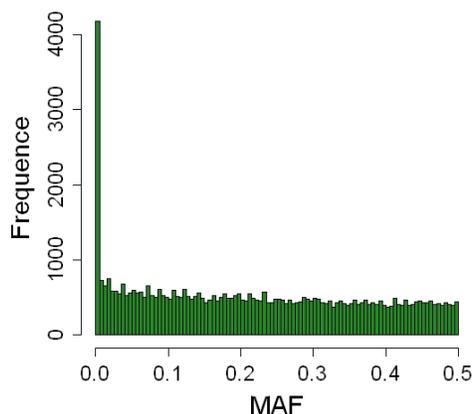


FIGURE 4.7 – Distribution des MAF dans la population GENEQUIN de TF

Une étape importante dans le contrôle de la qualité des données des génotypes est le test d'équilibre d'Hardy-Weinberg (HWE). A chaque SNP, un test de Pearson (ou test du χ^2) est utilisé afin de voir si le SNP testé est en HWE dans la population étudiée. Le but n'étant pas de supprimer tous les marqueurs qui ne sont pas en HWE car ceci peut arriver pour de multiples raisons (voir chapitre 1), mais de supprimer ceux qui dévient trop fortement car les principales causes des ces fortes déviations sont les erreurs de génotypages (par exemple à cause d'une mauvaise classification des hétérozygotes comme homozygotes). La Figure 4.8 montre le QQPlot de la distribution de ces tests d'Hardy-Weinberg (HWE) dans notre population. Même si l'impression visuelle ne donne pas ce sentiment, le nombre de marqueurs dont la valeur du test dépassait $\chi^2 = 36.84$ (soit une P-valeur égale à 10^{-8}) était égal à 216. L'ensemble de ces marqueurs ont été supprimés pour les analyses.

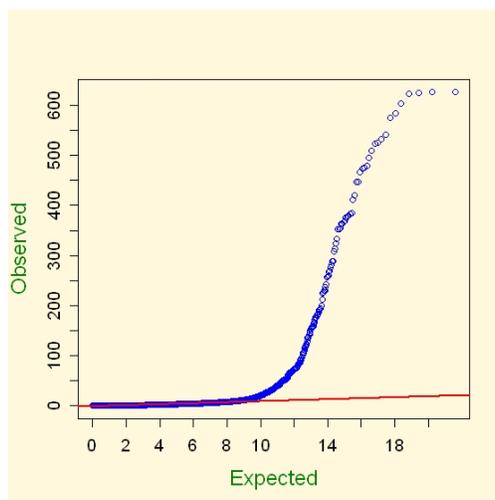


FIGURE 4.8 – QQPlot des tests HWE obtenus dans la population GENEQUIN de TF

Étude du déséquilibre de liaison

Dans notre population Dans la Figure 2 de l'article précédent, nous avons vu l'étendu du déséquilibre de liaison moyen obtenu dans la population de TF étudiée. Ici, nous présentons cette même figure avec les écart-types obtenus à chaque distance entre SNP (Figure 4.9). On s'aperçoit qu'ils sont importants et ce, même sur des grandes distances. La mesure du LD entre marqueurs SNP permet d'extrapoler le LD entre un SNP et un éventuel QTL. On suppose donc que la relation entre les LD et la distance est la même entre un allèle du QTL et un allèle du SNP qu'entre SNPs. Ainsi, l'observation de LD forts à une faible distance entre SNP laisse espérer un LD fort entre QTL et un marqueur de la puce situé à proximité. Plus ce LD est fort et plus l'effet "réel" du QTL sera visible dans l'effet "mesuré" du SNP sur la performance. Inversement un faible LD à grande distance évitera que les SNP situés à une grande distance du QTL émettent un signal important.

Dans notre étude, le LD moyen entre deux SNPs adjacents est de 0.35 ± 0.36 (mesure r^2) et on peut donc supposer que cette valeur est inférieure au LD moyen entre un marqueur et un possible QTL (au pire, le QTL est à mi distance entre 2 SNP). Notons que la valeur du LD moyen entre deux SNPs adjacents diffère de la valeur du LD obtenue à une distance moyenne entre deux SNPs (r^2 de 0.26 à une distance de 0.04Mb). Enfin, le LD entre SNP de chromosomes différents est en moyenne de 4.10^{-3} et peut donc être considéré comme inexistant.

La très grande variabilité du LD pour une même distance entre SNP (et par extrapolation entre SNP et QTL) implique qu'on peut très bien trouver un SNP en LD avec un QTL éloigné et inversement qu'un marqueur près du QTL ne soit pas en association avec lui. La décroissance du LD avec la distance entre marqueurs est un phénomène favorable "en moyenne" à la recherche de QTL mais cache une très grande hétérogénéité.

Comparaison avec d'autres espèces Le LD moyen à certaines distances entre marqueurs obtenu dans notre population de TF est relativement fort par rapport à d'autres espèces, ce qui nous laisse penser que l'utilisation de la puce Illumina BeadChip EquineSNP50 nous permet d'obtenir des puissances de détection élevées en analyse d'association.

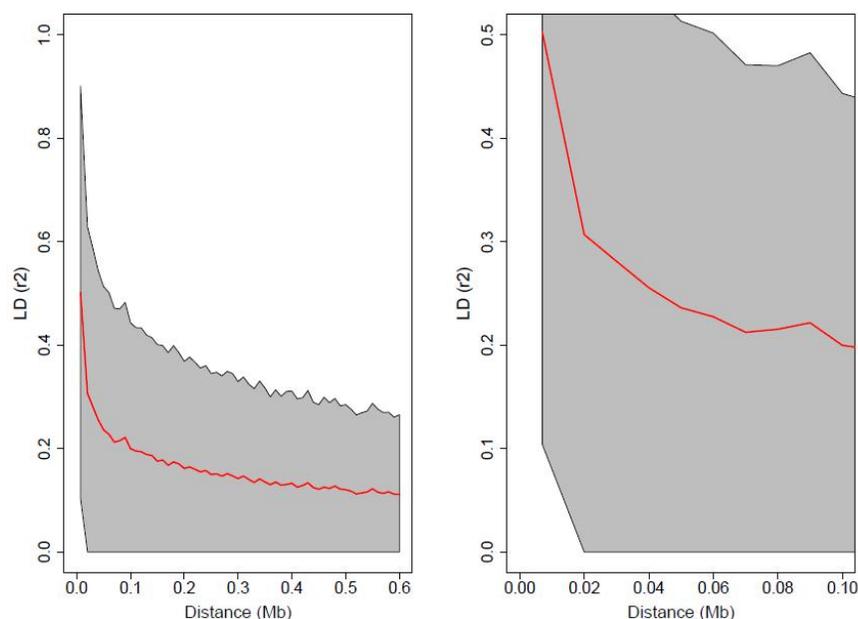


FIGURE 4.9 – LD moyen (et écart-type) en fonction des distances entre les SNP. La Figure de droite est un zoom sur la région 0-0.1Mb de celle de gauche.

McKay et al. (2007) ont étudié le LD moyen entre SNP en fonction des distances sur le génome dans 8 races bovines. Le LD qu'ils obtiennent dans la population Holstein, le plus fort de leur étude, semble être très proche de celui observé dans notre population puisqu'ils trouvent un r^2 de 0.5 à une distance entre SNP de 5Kb et un r^2 de 0.22 à une distance entre SNP de 0.1Mb (nous trouvons un LD de 0.5 à 7.5Kb et 0.2 à 0.1Mb). Ils concluent que le LD ne s'étend pas au delà de 0.5 Mb et qu'une puce contenant 30000 à 50000 SNP permet de réaliser une analyse d'association sur tout le génome avec un LD entre marqueurs adjacents proche de 0.2.

Chez les ovins et les humains, le LD moyen est beaucoup plus faible. Kemper et al. (2011) ont étudié 4 races ovines et montrent que le LD obtenu à une distance moyenne entre marqueurs adjacents (0.05Mb) varie de 0.12 à 0.19 selon les races et n'est plus que de 0.11 quand elles sont mélangées. Ces niveaux de LD sont beaucoup plus faibles que celui obtenu dans notre population (0.26). De plus, ce LD moyen décroît rapidement et n'est plus égal que de 0.05 à 0.12 selon les races pour une distance de 0.1 Mb. De ce fait, Kemper et al. (2011) suggèrent d'augmenter significativement les effectifs des populations ovines pour obtenir une puissance de détection de bonne qualité, ou d'augmenter le nombre de marqueur à 250000 (soit un SNP tous les 10kb pour avoir un LD moyen de 0.2 entre SNP adjacents). Chez les populations humaines, le LD est certainement l'un des plus faibles observés (Tenesa et al., 2007). Néanmoins, ce faible niveau de LD est compensé par des puces avec un grand nombre de SNP (de l'ordre du million).

Les plus fort LD sont observés chez les porcs et les canins (Uimari and Tapio, 2011; Wade et al., 2009). Uimari and Tapio (2011) ont étudié 2 races de porcs et trouvent que le LD s'étend sur une longue distance (parfois un LD de 0.16 sur 3 Mb) et que le LD moyen entre deux SNP adjacents est de l'ordre de 0.45.

Comparaison avec d'autres races équinnes La Figure 4.10 montre l'étendu du LD (mesure r^2) dans différentes populations équinnes. Il est intéressant de voir que le LD chez les TF est l'un des plus élevé.

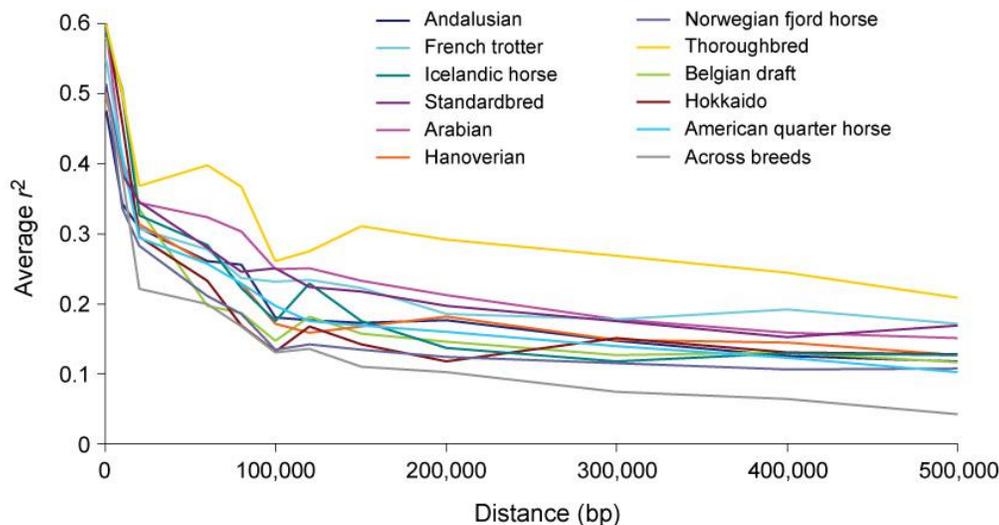


FIGURE 4.10 – Étendu du LD moyen dans différentes populations équinnes (Wade et al., 2009)

4.5 Compléments sur les QTL détectés

4.5.1 Nouveaux caractères analysés

Nous avons vu précédemment que notre choix de caractères à étudier portait sur les sites articulaires. Les caractères étudiés dans l'article étaient donc le LSG, le Boulet, le Jarret et le caractère Autres dans la population P_1 avec 525 chevaux phénotypés. Ici, nous proposons un complément à ces analyses en rajoutant la détection de QTL pour les caractères basés sur les différents type de contrainte, à savoir les phénomènes de cisaillement, de compression et de tension et en incluant les analyses de la populations P_2 avec 583 chevaux phénotypés. L'arthropathie juvénile, seule AOAJ qui ne peut être regroupée dans un type de contrainte, n'a pas été étudiée du fait de ces faibles fréquences dans notre population.

4.5.2 Compléments sur les méthodes d'analyses

Choix de nouvelles méthodes d'analyses Les caractères correspondant à des sites articulaires étudiés dans l'article précédent ont été analysés à partir de deux méthodes distinctes, SNPMixed et HaploIBD. La première était une méthode uni-marqueur dont le modèle comprenait un effet du SNP et un effet aléatoire polygénique avec une matrice de variance-covariance de cet effet calculée à partir du pedigree des individus phénotypés.

En complément à ces analyses, nous présentons ici les résultats obtenus avec un test Armitage et un modèle SNPMixed génomique, c'est à dire avec une matrice de variance-covariance de l'effet polygénique calculée à partir de l'information génomique de l'ensemble des individus génotypés. Par la suite, nous parlerons de SNPMixed en référence aux résultats de SNPMixed génomique et

non SNPMixed parenté. Le test Armitage, appliqué à tous les caractères binaires, ne prend pas en compte les effets fixes ni les effets liés à la structure de population. Par contre SNPMixed, appliqué pour les caractères par site articulaire et pour le Cisaillement, permet d'introduire les effets fixes et prend en compte la structure de la population. Les autres caractères, compression et tension, n'ont pas été analysés avec SNPMixed car les fréquences des chevaux lésés étaient trop faibles pour estimer correctement les variances.

Distribution des tests de nouvelles méthodes Les figures 4.11 et 4.12 montrent les QQPlots obtenus respectivement avec le test Armitage et le modèle SNPMixed (génomique). On peut voir que le test Armitage présente une inflation des valeurs des tests pour l'ensemble des caractères mais plus particulièrement pour le caractère Jarret et Cisaillement (respectivement $\lambda = 1.49$ et $\lambda = 1.43$ pour P_1 , le terme d'inflation λ étant décrit dans le chapitre 1 dans le paragraphe sur le contrôle génomique). Cette inflation résulte principalement de la non prise en compte de la structure de population dans le modèle. Le modèle SNPMixed donne quant à lui les bonnes distributions pour P_1 et P_2 et pour l'ensemble des caractères correspondant.

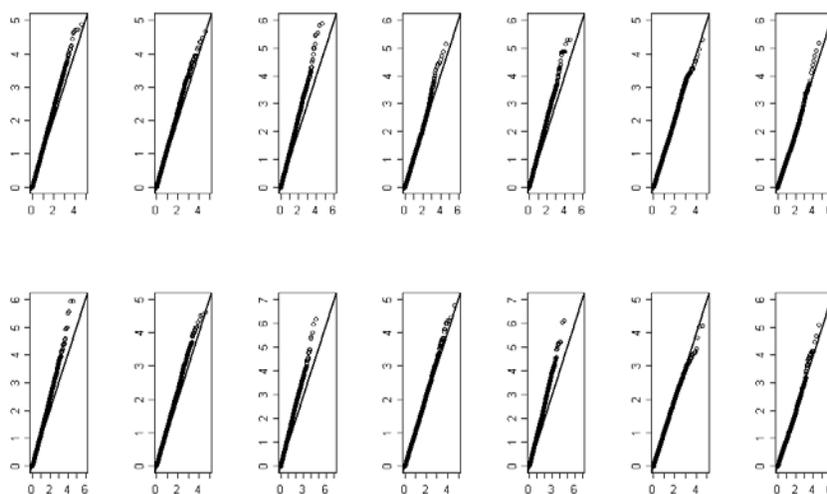


FIGURE 4.11 – QQPlot des $-\log_{10}(\text{Pvaleur Observées})$ (en ordonnées) par rapport aux $-\log_{10}(\text{Pvaleur Espérées})$ (en abscisses) obtenues avec le Test Armitage. Chaque colonne représente respectivement le QQplot d'un caractère étudié, dans l'ordre : Total, Boulet, Jarret, Autres, Cisaillement, Compression et Tension. Chaque ligne représente le QQplot dans une population étudiée, dans l'ordre : P_1 et P_2 .

Distribution du nombre de clusters pour HaploIBD La méthode HaploIBD n'a pas été réutilisée pour la population P_2 et pour d'autres caractères, principalement pour des raisons de temps. A chaque position testée du génome, cette méthode (décrite dans l'article) teste les effets des groupes (ou clusters) d'haplotypes sur le caractère correspondant. La figure 4.13 donne la distribution du nombre de clusters d'haplotypes sur l'ensemble des positions du génome. En moyenne, l'analyse a été réalisée sur 25.4 clusters avec un écart-type de ± 10.9 .

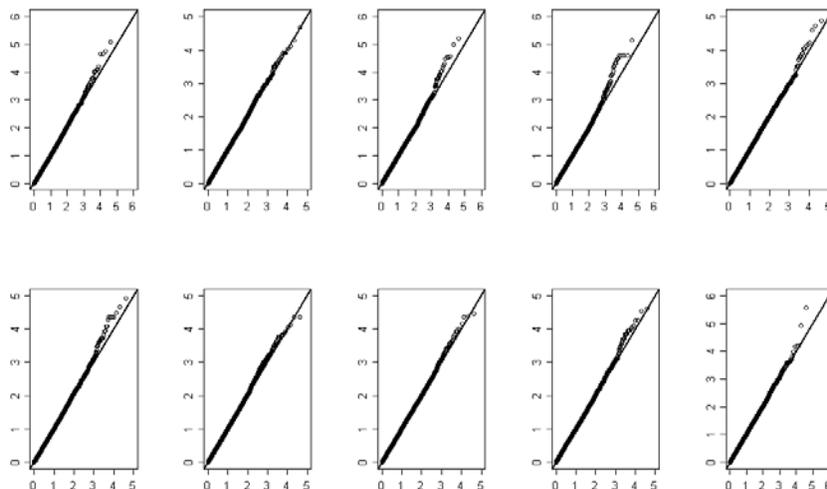


FIGURE 4.12 – QQPlot des $-\log_{10}(\text{Pvaleur Observées})$ (en ordonnées) par rapport aux $-\log_{10}(\text{Pvaleur Espérées})$ (en abscisses) obtenues avec la méthode SNPMixed génomique. Chaque colonne représente respectivement le QQplot d'un caractère étudié, dans l'ordre : LSG, Boulet, Jarret, Autres, Cisaillement. Chaque ligne représente le QQplot dans une population étudiée, dans l'ordre : P_1 et P_2 .

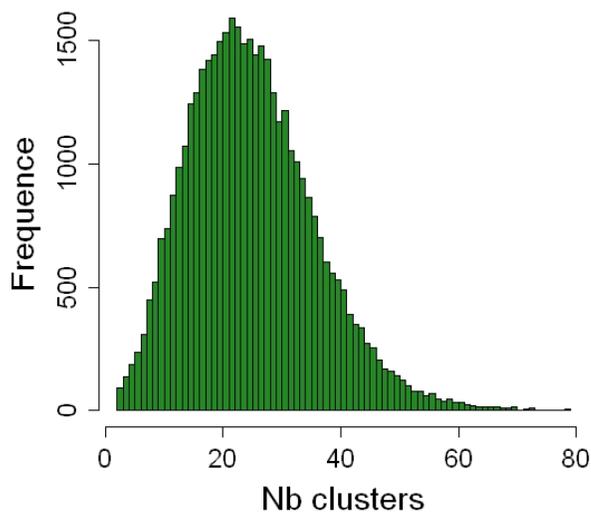


FIGURE 4.13 – Distribution des clusters d'haplotypes obtenus avec HaploIBD

4.5.3 Bilan des QTL détectés

Le Tableau 4.10, présenté en fin du chapitre, donne l'ensemble des QTLs détectés au seuil $P < 5.10^{-5}$ (correspondant à une valeur de $-\log_{10}(P) = 4.3$) pour les caractères et les méthodes

décrites précédemment.

Bilan par méthodes Pour le test Armitage, 15 QTLs sont nouveaux (i.e. sont détectés dans P_2 et ne l'étaient pas dans P_1) et 63% des anciens QTLs (détectés dans P_1) sont confirmés dans P_2 . Si on s'intéresse uniquement aux QTLs détectés dans P_2 , on trouve le plus de QTL pour le caractère Jarret (13 QTL) et Cisaillement (10 QTL). A contrario, aucun QTL n'est détecté pour le caractère Compression et seulement 3 sont détectés pour les caractères Boulet et Tension. Le résultat le plus significatif ($-\log_{10}(P) = 6.18$) obtenu à partir du test Armitage est pour le caractère Jarret sur ECA 3 au SNP BIEC2-808879 en position 106.87 Mb. Il est intéressant de noter que parmi les 10 résultats les plus significatifs avec ce test, 9 sont détectés pour les caractères Jarret ou Cisaillement. Cependant, on a pu voir dans les QQplots que les tests sur ces caractères sont les plus inflatés et donc surestiment les valeurs des tests.

Pour le modèle SNPMixed, 7 QTLs sont nouveaux et 41% des anciens QTLs sont confirmés dans P_2 . Si on s'intéresse uniquement aux QTLs détectés dans P_2 , on trouve de 2 à 4 QTL détectés pour chacun des caractères étudiés. Le résultat le plus significatif pour ce modèle ($-\log_{10}(P) = 5.58$), est obtenu pour le caractère Cisaillement sur ECA 15 au SNP BIEC2-310861 en position 52.47 Mb. Il est intéressant de noter que le résultat le plus significatif obtenu avec la population P_1 sur ECA 3 du Jarret n'est plus significatif dans P_2 . Néanmoins, un nouvel SNP, très proche du précédent, devient significatif avec P_2 .

Comparaison des QTL entre méthodes Mis à part 3 QTL détectés avec SNPMixed, dont 2 pour LSG et 1 pour le Boulet, tous les QTL détectés avec SNPMixed l'ont également été avec le test Armitage. Néanmoins, principalement à cause de l'inflation des tests, le test Armitage a donné en moyenne plus de QTL. De même, il est rassurant de voir qu'à part un QTL détecté pour le caractère Autres, l'ensemble des QTL détectés avec HaploIBD l'ont été avec les deux autres méthodes dans la population P_1 .

Comparaison des QTL entre caractères Les caractères globaux LSG et Total mis à part, 1 seul QTL a été trouvé en commun entre les autres caractères. Il s'agit du QTL sur ECA 3 où un signal est significatif pour le Cisaillement et pour le Jarret entre les positions 105 et 107 Mb. Néanmoins, le signal pour le caractère Cisaillement n'est pas significatif pour le modèle SNPMixed. Par contre, et par définition des caractères globaux, on trouve des signaux significatifs partagés entre les caractères globaux et les autres caractères comme par exemple sur ECA 13 avec le Boulet ou sur ECA 15 avec le Cisaillement. Enfin, pour SNPMixed dans P_2 , mis à part un QTL détecté en début de ECA 3 avec LSG et en fin de ECA 16 avec Total, les caractères globaux ont donné les mêmes positions de QTL (3 QTL identiques).

Comparaison des QTL dans la bibliographie Le QTL présent sur ECA 5 pour le caractère Cisaillement est situé à la même position qu'un QTL détecté par Dierks et al. (2007) pour l'OC du jarret. Du fait que les contraintes de type cisaillement sont de loin les plus fréquentes dans les lésions sur les jarrets, ces QTL peuvent être comparés. De même, nos résultats semblent indiquer la position d'un même QTL sur ECA 15 pour l'OC du jarret autour des 24 Mb que les résultats de Wittwer et al. (2007). Notons également qu'un QTL détecté sur ECA 26 pour le caractère Cisaillement est à 3 Mb d'un QTL détecté par Lampe et al. (2009b) pour l'OC du boulet (les lésions de type cisaillement sont les plus fréquentes sur le boulet).

Ordre de grandeur des plus gros effets QTL obtenus Le Tableau 4.9 donne l'ordre de grandeur des effets des allèles obtenus avec SNPMixed sur la fréquence des chevaux atteints, pour les plus gros QTL détectés dans P_2 avec chacun des caractères. Prenons l'exemple du plus fort QTL détecté avec SNPMixed, qui est celui détecté pour le caractère Cisaillement au SNP BIEC2-310861 et dont la P-valeur vaut 5.58 (en $-\log_{10}(P)$). Pour ce SNP, la fréquence de l'allèle 2 (q) est de 0.71 et donc la valeur moyenne des génotypes (codés 0,1 ou 2) vaut 1.42 (car $0 \times p^2 + 1 \times 2pq + 2 \times q^2 = 2q$). Abstraction faite des effets polygénique de SNPMixed, l'intercept de la regression vaut donc 0.68 ($\hat{\mu} = \bar{y} - 2q$, où \bar{y} est la moyenne des phénotypes et représente la fréquence des atteints pour les variables binaires). L'estimation de β valant -0.17 dans ce cas, la valeur prédite des phénotypes des génotypes "11" est donc 0.68, celle des génotypes "12" est 0.51 et celle des génotypes "22" est 0.34. Ce qui signifie qu'on prédit que la proportion de chevaux atteints parmi les génotypes "11" est de 68% contre 34% parmi les génotypes "22", mais ces derniers sont plus fréquents (50%) que les premiers (8%).

TABLE 4.9 – Tableau des ordres de grandeur des plus forts QTL détectés dans P_2 avec SNPMixed

Trait	ECA	SNP	Pos	Pval	homo 11 ¹		hétéro 12 ²		homo 22 ³	
					freq ⁴	est ⁵	freq ⁴	est ⁵	freq ⁴	est ⁵
LSG	15	BIEC2-325537	87,91	4,92	0,61	0,96	0,34	0,69	0,05	0,42
Total	15	BIEC2-326013	89,62	5,02	0,66	0,59	0,31	0,42	0,04	0,25
Boulet	13	BIEC2-209568	9,89	4,36	0,05	0,55	0,35	0,41	0,59	0,27
Jarret	3	BIEC2-789537	68,81	4,46	0,77	0,21	0,21	0,39	0,01	0,57
Autres	6	BIEC2-1178497	60,44	4,60	0,01	0,38	0,16	0,24	0,83	0,10
Cisaillement	15	BIEC2-310861	52,47	5,58	0,08	0,68	0,41	0,51	0,50	0,34

1. homo 11 : Génotype homozygote 11

2. hétéro 12 : Génotype hétérozygote 12

3. homo 22 : Génotype homozygote 22

4. freq : fréquence estimée d'un génotype donné ($p^2, 2pq, q^2$)

5. est : valeur prédite des phénotypes pour un génotype donné

4.5.4 Discussion

Différences des résultats entre P_1 et P_2 L'apport des 58 individus supplémentaires dans P_2 ne contredit pas l'existence de 63% des QTL détectés dans P_1 avec le test Armitage et 41% des QTL avec SNPMixed. Ces pourcentages sont relativement faibles et semblent signifier la présence de plusieurs faux positifs. Une possible explication pourrait être que le seuil choisi n'est pas assez fort. En effet, le seuil $P < 5.10^{-5}$ est équivalent à un seuil de Bonferroni qui suppose 1000 tests indépendants sur l'ensemble du génome, soit environ 1 SNP sur 40 et donc une distance moyenne entre deux tests indépendants de 2Mb. Le DL chez les TF s'étend sur des longues distances mais on peut raisonnablement supposer qu'au delà de 0.5Mb (ou le r^2 vaut 0.11), les tests sont proches d'être indépendants. Le seuil correspondant à cette valeur aurait été de $-\log_{10}(P) = 4.90$. Néanmoins, les pourcentages de QTL confirmés pour ce seuil sont également faibles. Une autre explication réside

dans le fait qu'une grande majorité des QTL détectés dans P_1 l'ont été dans une tranche de valeurs comprises entre 4.30 et 4.90 (en $-\log_{10}(P)$). De ce fait, tout apport de nouveaux individus peut facilement faire passer ces valeurs en dessous du seuil (de même dans le sens inverse). Même si l'on observe des différences importantes entre le nombre de QTL détectés sous P_1 et sous P_2 , l'allure des courbes (non présentée ici) reste identique (la corrélation moyenne entre les Pvaleurs sous P_1 et sous P_2 pour une méthode et un modèle particulier est égale à 0.9).

Mesure globale de l'OC : quantitative ou binaire ? Les résultats des caractères globaux pour la méthode SNPMixed ont été obtenus pour la mesure quantitative LSG et la mesure binaire Total. En terme de détection de QTL, 4 QTL ont été détectés pour chacune des mesures dont 3 étaient localisés parfaitement au même SNP (sous P_2). Un QTL pour chacune des mesures n'est donc pas commun aux deux caractères. Si on regarde de près ces QTL, un QTL sur ECA 13 est détecté pour LSG avec une valeur de 4.48 (en $-\log_{10}(P)$) et un QTL sur ECA 16 est détecté pour Total avec une valeur de 4.49. La valeur associée au QTL sur ECA 13 pour le caractère Total était de 2.85 et celle associée au QTL sur ECA 16 pour le caractère LSG était de 4.08. Il semble donc que la tendance soit bien la même entre ces deux caractères. D'ailleurs, on peut noter que la corrélation entre les Pvaleurs de ces caractères est de 0.84 sous P_1 et P_2 (proche de la valeur des corrélations phénotypiques observée entre ces caractères qui était de 0.87). Il est donc difficile de conclure sur le meilleur choix à faire entre ces mesures.

Problèmes d'estimation avec HaploIBD L'estimation des variances des effets aléatoires des clusters d'haplotypes avec HaploIBD donne parfois des résultats aberrants. Ce problème est dû à la présence de clusters d'haplotypes de très faible effectif et est amplifié par le côté discret du caractère analysé. En effet, en considérant les effets comme aléatoires, chaque niveau (cluster) a un poids équivalent dans le calcul de la variance, même si il est très peu fréquent dans la population. Si l'estimation de cet effet a une valeur extrême, elle a un poids considérable dans le calcul de la variance. Avec un caractère continu, il y a peu de chance que cet effet ait une valeur extrême. Alors que ceci est possible avec un caractère binaire analysé comme une variable gaussienne comme c'est le cas avec HaploIBD. Prenons le cas du caractère "Autres" pour lequel on a observé ces variances aberrantes. La fréquence des chevaux atteints est 13%, donc le caractère a une variance phénotypique de 0.113 et un l'écart type phénotypique est 0.336. Ce qui veut dire que la différence de performance entre un cheval sain et atteint représente près de 3 écart types. Avec une variable continue normale, une performance a 3 écart type n'est observée que dans 0.3% des cas, ici il y en a 13%. Pour peu qu'un cheval atteint constitue un cluster à lui seul, l'estimation de l'effet de ce cluster va être extrême et gonfler artificiellement la variance. C'est ce qu'on a obtenu par deux fois pour la variable "Autres" (sur les chromosomes 13 et 15). Le problème, même moins visible, doit être récurrent dans toutes les analyses des variables binaires avec les clusters peu fréquents. Il faudrait donc améliorer la méthode d'analyse soit en ne permettant pas la présence de clusters rares soit en traitant correctement les valeurs binaires. Cependant, avec un modèle à seuil, l'EM classiquement utilisé pour le REML ne converge pas vers les vraies valeurs pour un modèle polygénique animal et qu'il faudrait sans doute recourir à un GIBBS sampling.

QTL à fort intérêt Certains des QTL détectés le sont par un signal fort et cohérent entre les méthodes. Ainsi, le QTL du chromosome ECA 3 entre 105 et 110 Mb paraît être la région la plus prometteuse pour une cartographie plus fine puisque toutes les méthodes le détectent, et de plus sur

plusieurs caractères : *Jarret*, *Cisaillement* et *Total*. La section suivante décrit de près cette région. D'autres régions sur ECA 13 pour l'OC du boulet, 14 pour l'OC du jarret, 15 pour l'OC ailleurs que sur le boulet et le jarret et 15 pour l'OC de type Cisaillement demanderaient également à être vues de plus près.

4.5.5 QTL de l'OC du jarret sur ECA 3 : description

La méthode HaploIBD appliquée à l'OC du jarret a donné en position 105.05 Mb de ECA 3 (SNP BIEC2-808442) une p -valeur en $-\log_{10}$ de 5.52 pour un QTL expliquant environ 7% de la variance phénotypique. Les méthodes SNPMixed ainsi qu'un test Armitage ont quant à elles donné un maximum en position 105.88 Mb (SNP BIEC2-808617). La Figure 4.14 montre les courbes des méthodes HaploIBD et SNPMixed (avec une matrice de variance-covariance construite à partir de la matrice de parenté comme dans l'article, et construite à partir des marqueurs comme dans les dernières études). Les causes de cette différence de localisation peuvent être recherchées.

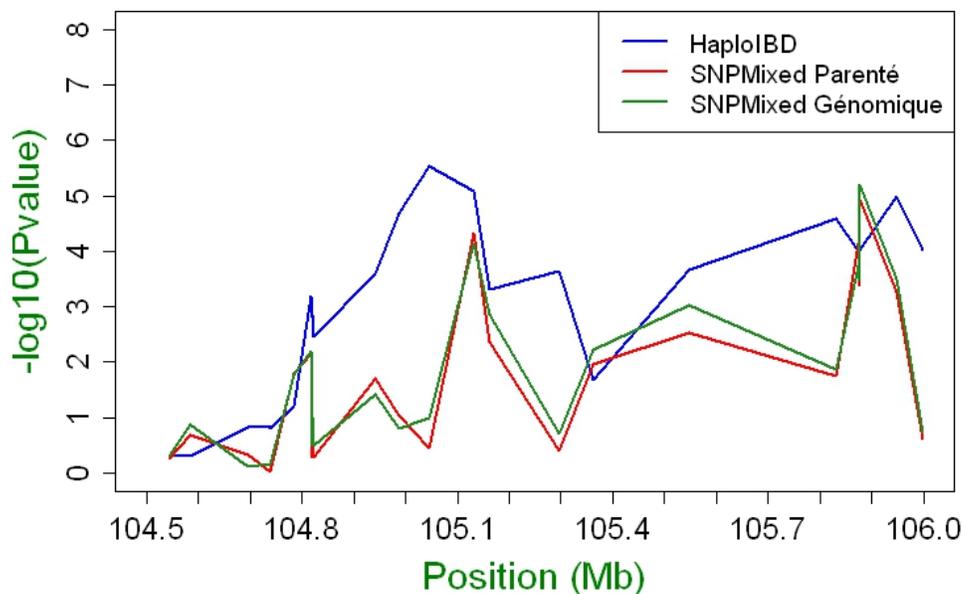


FIGURE 4.14 – Résultats de différentes méthodes sur la région QTL de ECA 3

Tout d'abord, la méthode HaploIBD utilise des haplotypes et un maximum en position 105.05 Mb signifie un maximum pour l'haplotype de 6 marqueurs compris entre la position 104.94 Mb et 105.3 Mb. Dans ce segment, 16 classes (ou clusters) d'haplotypes ont été identifiés. La Figure 4.15 donne la distribution des fréquences de ces classes au sein de l'échantillon analysé. Il apparaît assez nettement que 6 classes seulement sont à l'origine de l'effet du QTL, les autres étant en fréquence bien trop faible.

Les résultats des estimations des effets de ces classes d'haplotypes varient entre 0.16 pour le cluster 1 et -0.09 pour les clusters 3 et 7, le cluster 6 ayant une estimation de 0.09 et le reste

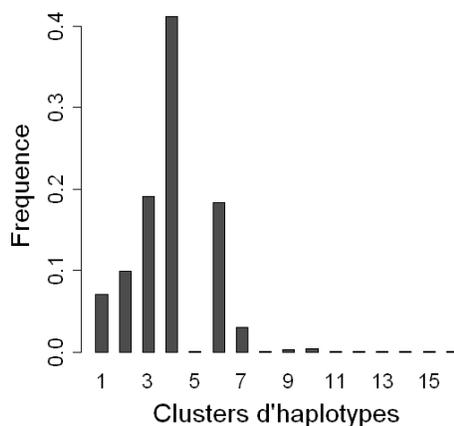


FIGURE 4.15 – Fréquence des classes d'haplotypes au QTL de ECA3

étant proche de 0. Le QTL est donc en très forte partie causé par les clusters 1 et 6 versus 3 et 7 (figure 4.16). Toutefois, le classe d'haplotype 7 possède une fréquence faible de 3%, ce qui fait partir l'essentiel du poids de ce QTL sur classes ne se voit qu'à la 4^{eme} position de l'haplotype, pour le SNP BIEC2-808456 en position 105.13 Mb. A cette position, l'allèle porté par les haplotypes présents dans les classes 1 et 6 est l'allèle 1 et l'allèle porté par les haplotypes présents dans la classe 3 est l'allèle 2. Il est intéressant de noter qu'à cette position, le test uni-marqueur réalisé avec SNPMixed avait donné une *p-valeur* élevée (4.13 en $-\log_{10}$).

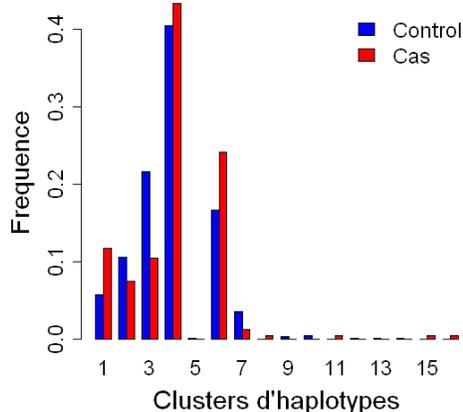


FIGURE 4.16 – Fréquence des clusters d'haplotypes au QTL de ECA3 en fonction des phénotypes

Les méthodes uni-marqueurs, Armitage ou SNPMixed, ne donnant pas de maximum à la position de BIEC2-808442, mais un peu plus loin sur le chromosome, au SNP BIEC2-808617 en position 105.88 Mb. Existe-t-il une relation entre les groupes d'haplotypes autour du SNP BIEC2-808442 et les génotypes au SNP BIEC2-808617 ?

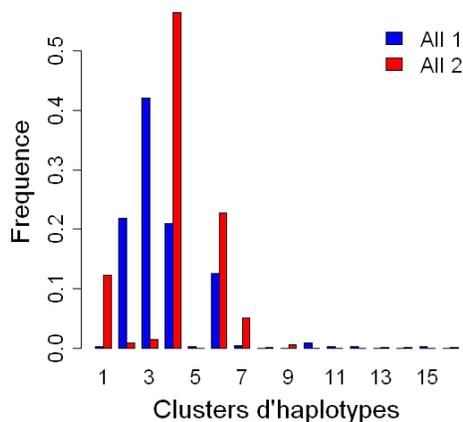


FIGURE 4.17 – Fréquence des clusters d'haplotypes au QTL de ECA3 en fonction de l'allèle au marqueur significatif avec SNPMixed

La Figure 4.17 montre la répartition avec les classes d'haplotypes précédents selon allèles associés au SNP BIEC2-808617 à la position 105.88 Mb. On observe que les classes 1 et 6 sont souvent associés avec l'allèle causal (2) du SNP BIEC2-808617 et que la classe 3 est plus fréquemment associé avec l'allèle sauvage (1) de ce même SNP. Il est intéressant de noter également que le SNP causal au sein des classes d'haplotypes (BIEC2-808456), est en déséquilibre de liaison modéré avec le SNP BIEC2-808617 (0.26 avec la mesure r^2).

Toutes ces informations semblent signifier qu'un QTL, localisé dans cette région de ECA 3, est en déséquilibre de liaison avec les SNP BIEC2-808456 et BIEC2-808617.

4.5.6 QTL de l'OC du jarret sur ECA 3 : test de validation

Le génotypage de 58 chevaux préalablement phénotypés a été réalisé afin de compléter la population P_1 de 525 chevaux déjà analysés et publiés (article précédent). Les résultats de l'analyse de la population P_2 de 583 chevaux ont été présentés plus haut. Ce nouvel échantillon peut aussi être traité comme un échantillon indépendant de validation du QTL de la région sur ECA 3 du jarret.

Ces 58 chevaux étaient composés de 23 atteints et 35 sains pour le caractère *Jarret*. Afin de valider la présence du QTL, un modèle mixte incluant l'effet du SNP, un effet age et un effet polygénique dont la variance est fixée a été mis en place pour tester l'effet du SNP causal BIEC2-808617. Ce test a donné un résultat non significatif au seuil de 5% ($-\log_{10}(P) = 0.56$) et ne validant pas la présence d'un QTL à cette position.

Néanmoins, ce résultat peut être un faux négatif. En effet, un échantillon de 58 individus est trop petit pour obtenir une puissance suffisante à la détection de ce QTL. A titre de démonstration, 100 tests ont été réalisés en tirant aléatoirement différentes tailles d'échantillons de la population P_1 . La Figure 4.18 montre la courbe des puissances obtenues à la suite de ces simulations. Dans notre cas avec 58 individus, la puissance n'est que de 0.36. Le QTL de l'OC du jarret n'avait que de faibles chances d'être trouvé à partir de cet échantillon. Un échantillon efficace eut été d'au moins

200 individus.

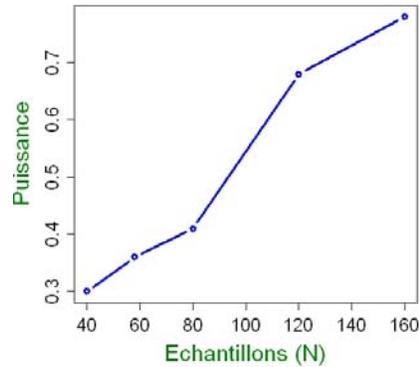


FIGURE 4.18 – Puissance empirique obtenues au SNP BIEC2-808617 après 100 simulations en fonction d’une taille N d’un échantillon tiré aléatoirement parmi 525 individus. Les points $N = 40, 80, 120, 160$ ont été obtenus avec la moitié de sains et d’atteints et le point $N = 58$ correspond à un tirage de 35 sains et 23 atteints

De plus, si on regarde les 100 simulations pour la taille $N = 58$ correspondant à la taille de notre échantillon de validation, on s’aperçoit que la valeur réelle obtenue de 0.56 (en $-\log(P)$) n’est qu’au niveau du quantile 0.25 de la distribution des $-\log(P)$ obtenues dans ces simulations (Figure 4.19). Ce résultat sous entend qu’en plus d’une faible puissance, notre échantillon n’était pas de bonne qualité pour valider ce QTL.

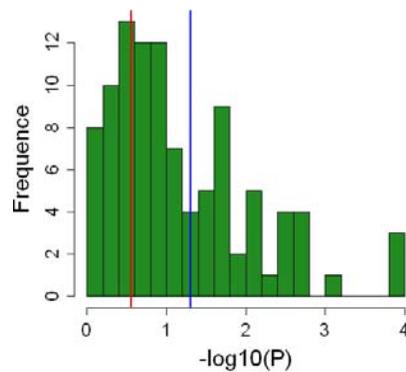


FIGURE 4.19 – Histogramme des P-valeurs obtenues au SNP BIEC2-808617 après 100 simulations de tirage aléatoire de 23 atteints et 35 sains à partir d’un échantillon de 525 individus. La ligne verticale rouge correspond à la valeur obtenue à partir du nouvel échantillon et la ligne bleue correspond au seuil significatif de 5%

4.6 Bilan et perspectives

Que pouvions-nous faire et trouver ? Un atout majeur de notre protocole et de la population étudiée est le niveau du LD entre les marqueurs. Ce niveau du LD, très fort puisque de 0.35 entre deux SNP adjacents, laissait espérer une forte association entre un SNP testé et un QTL proche (certainement supérieure à 0.4). Avec ce LD, la puissance de notre protocole (comprenant au maximum 583 chevaux) en analyse d'association est relativement élevée et nous permet de détecter avec quasi-certitude des QTL de moyens et forts effets (puissance supérieure à 90% pour des QTL qui expliquant > 5% de la variance phénotype). Par contre, cette puissance est bien moins élevée pour des QTL plus faibles (< 50% pour des QTL expliquant 2% de la variance phénotypique). En outre, le niveau de LD observé dans la population TF s'étend sur des distances relativement longues avec une forte variabilité, laissant attendre une localisation imprécise des QTL.

En ce qui concerne les phénotypes, les atouts majeurs de notre protocole sont l'homogénéité des lectures radiographiques par des mêmes vétérinaires (car il a été montré que la corrélation entre des lectures radiographiques de l'OC entre plusieurs vétérinaires était faible), et le nombre important de clichés radiographiques réalisés (un cheval avait au moins 10 clichés) qui nous ont permis de construire diverses mesures phénotypiques. Ces mesures, essentiellement binaires, ont été construites afin de rechercher les QTL. Pour certaines d'entre elles, la fréquence des chevaux atteints était trop faible pour permettre l'utilisation du modèle mixte pour l'analyse (variance génétique difficilement estimable avec des mesures binaires en fréquences < 15%).

Que trouvons-nous sur les mécanismes de l'OC ? Un premier résultat important, et qui confirme les précédents résultats sur la recherche de QTL de l'OC chez les chevaux, est que l'OC n'est pas gouverné par un gène majeur. Un tel gène eut été détecté compte tenu de la puissance de notre protocole. L'OC est donc un caractère multifactoriel avec un effet polygénique où plusieurs QTL à moyen et faible effets jouent un rôle sur le caractère. Nous trouvons d'ailleurs plusieurs de ces QTL.

Nous n'avons pas pu mettre en évidence de QTL ayant un effet à la fois sur l'OC du jarret et du boulet. Ce résultat, en accord avec les faibles corrélations génétiques trouvées précédemment entre ces sites (Grondahl and Dolvik, 1993; Stock and Distl, 2006), ouvre la porte à plusieurs questions sur la définition du caractère : le déterminisme génétique de l'OC sur le jarret et sur le boulet est-il totalement différent ? Les QTL détectés ne sont ils que des gènes qui permettent la révélation de la déficience du cartilage en des sites articulaires précis (comme des gènes de la morphologie) et donc indépendants de la susceptibilité générale des cartilages ? S'il n'est pas possible de répondre à ces questions ici faute d'informations supplémentaires, nos résultats ont au moins permis de les poser.

A-t-on réellement trouvé des QTL ? En ce référant aux méthodes dont les statistiques de test sont correctement déterminées (pas d'inflation des Pvaleurs), i.e. HaploIBD et/ou SNP Mixed, et en tenant compte d'un seuil qui prend en compte la multiplicité des test, nous détectons un nombre important de résultats significatifs sans qu'aucun ne soit hautement significatif (maximum $-\log_{10}(P) = 5.52$ dans P_1 et 5.58 dans P_2) et ils semblent de plus pas très bien localisés comme par exemple sur le chromosome 3.

Peu des QTL que nous avons détecté semblent communs à ceux détectés dans les précédentes études sur les QTL de l'OC. Ces résultats sont décevants et pourraient mettre en cause la validité de nos QTL. Plusieurs raisons peuvent expliquer ces divergences. D'abord, il est possible que les

QTL soient spécifiques aux différentes races. D'ailleurs peu de QTL étaient partagés entre les précédentes études dans la littérature. Ensuite, avec un effectif 2 à 3 fois supérieur aux études précédentes et donc une puissance bien supérieure, notre étude est probablement plus solide que les autres. Enfin, l'utilisation conjointe de plusieurs méthodes aux profils différents, l'une utilisant l'information provenant des haplotypes dans un modèle LDLA et l'autre les génotypes dans un modèle LDA, ajoute de la robustesse à notre approche : en combinant les résultats de ces analyses, nous espérons avoir diminué le nombre de faux positifs.

Que pouvons-nous conseiller ? En ce qui concerne le type de mesure à envisager, nous avons vu que les mesures binaires engendrent des problèmes liés à l'estimation des variances génétiques et haplotypiques lorsque les fréquences des chevaux atteints sont faibles. De ce fait, nous conseillons plutôt de rechercher des mesures biologiques de l'OC continues. Cette solution n'est pas toujours réaliste. Ainsi, pour des analyses par site articulaire où peu de clichés vont être réalisés, il ne sera pas facile d'imaginer une mesure biologique continue qui ait du sens. Dans ces situations, il serait bon d'utiliser des méthodes adaptées aux variables discrètes. De telles méthodes sont encore à concevoir et à tester.

Détailler un phénotype (l'OC) qui aurait pu être considéré a priori comme un caractère unique permet de montrer que le déterminisme génétique n'est pas toujours identique selon les définitions du caractère. Afin de ne pas multiplier artificiellement le nombre de tests statistiques, cela doit être réservé à des mesures susceptibles de ne pas être trop génétiquement ou phénotypiquement liées. C'est ce que nous montrons pour l'OC du jarret et du boulet. Il en est peut-être de même, et il serait intéressant de le montrer, pour d'autres caractères mais il nous a été impossible pour des raisons d'effectifs de le montrer. En revanche, il est important de noter que cela ne permet pas de démontrer un mécanisme différent. A l'inverse, étudier des caractères génotypiquement corrélés permettrait de comprendre les causes de ces corrélations.

En ce qui concerne les méthodes à utiliser, le choix d'une analyse haplotypique ou génotypique, LDLA ou LDA, point par point ou sur l'ensemble du génome pourrait longuement être discuté. Nous avons choisi une méthode haplotypique LDLA (HaploIBD) et une méthode génotypique LDA (SNPMixed). Les résultats ont montré un plus grand nombre de signaux significatifs avec SNP-Mixed, avec peut-être plus de faux positifs. Dans l'idéal, la méthode LDLA permet une meilleure localisation du QTL via son utilisation d'haplotypes mais il faut faire attention aux problèmes liés aux possibles erreurs dans reconstruction des phases, au choix du nombre de marqueurs à mettre dans l'haplotype et à l'estimation des effets des classes d'haplotypes en présence de variables binaires (surtout avec des fréquences faibles). La méthode LDA, moins contraignante, est puissante surtout en présence d'un niveau de LD entre marqueurs élevé. Cependant, elle est peu précise et sans doute moins que la méthode LDLA à cause de l'utilisation d'un seul SNP et du niveau élevé de LD entre SNP. De plus, elle est sujette à certains problèmes qui interviennent dans l'utilisation de variables binaires et dans l'estimation des variances. A ces méthodes, il serait judicieux d'adjoindre une méthode LDA qui utilise l'ensemble des marqueurs (type Bayes C π). Combiner ces approches aux profils différents, permettrait d'avoir une plus grande confiance en leurs résultats. Néanmoins, les défauts de ces méthodes en présence de variable binaires doivent être résolus au préalable.

Quelles sont les perspectives envisageables ? La suite de ce travail, partie intégrante du projet GENEQUIN, consiste à identifier les gènes candidats et les mutations causales. Pour ce faire, un séquençage complet (sur HiSeq2000) de 10 chevaux représentatifs de la variabilité allélique

des principaux QTL est en cours et devrait permettre l'identification et l'annotation fonctionnelle des SNP. Ainsi, le génotypage de nouveaux SNP au sein des régions QTL devrait améliorer leur localisation. Parallèlement, un autre projet ANR nommé BioCart, proposant une analyse comparée du protéome de l'os et du cartilage dont le but est de mettre en évidence le mécanisme moléculaire et les processus biologiques impliqués dans la physiopathologie de l'OC, se poursuit et les premiers résultats suggèrent que l'OC est une pathologie de l'os et du cartilage (dialogue os/cartilage). La coordination de ces deux projets, réalisée au sein de l'unité de Biologie Intégrative et Génétique Équine (BIGE) de l'INRA, devrait rapidement permettre de clarifier la physiopathologie moléculaire de l'OC et de développer des stratégies efficaces pour l'évaluation des risques.

Pendant ce temps, les marqueurs pourraient être utilisés dans un contexte de sélection assistée par marqueurs (SAM) afin d'améliorer la santé et le bien-être du cheval. Néanmoins, la construction d'une population de référence, nécessaire au fonctionnement de la SAM, ne semble pas envisageable à l'heure actuelle. Pour qu'une SAM soit efficace, il faudrait que cette population de référence, constituée de chevaux phénotypés et génotypés, soit de très grande taille pour que les effets des SNP soient le plus précis possible.

TABLE 4.10: QTLs de l'OC détectés dans GENEQUIN

ECA ¹	SNP	Pos ²	Trait	Méthode	freq A_2 ³		P-valeur ⁴		β ⁵		V(QTL) ⁶	
					P_1 ⁷	P_2 ⁸	P_1 ⁷	P_2 ⁸	P_1 ⁷	P_2 ⁸	P_1 ⁷	P_2 ⁸
1	BIEC2-55545	128,91	Jarret	Armitage	0,47	0,47	4,31	<i>3,22</i>				
1	BIEC2-63220	147,61	Autres	Armitage	0,34	0,34	4,30	<i>3,08</i>				
1	BIEC2-63220	147,61	Autres	SNPMixed	0,34	0,34	4,49	<i>3,26</i>	0,26	0,20	3,02	1,80
1	BIEC2-83702	174,52	Cisaillage	Armitage	0,94	0,94	4,89	5,20				
2	BIEC2-508227	115,41	Jarret	Armitage	0,50	0,51	5,14	5,41				
2	BIEC2-508227	115,41	Total	Armitage	0,50	0,51	<i>3,70</i>	4,37				
3	BIEC2-789537	68,81	Jarret	Armitage	0,12	0,12	5,45	5,61				
3	BIEC2-789537	68,81	Jarret	SNPMixed	0,12	0,12	4,55	4,46	0,43	0,41	3,94	3,59
3	BIEC2-798300	84,89	Jarret	Armitage	0,86	0,85	4,50	5,37				
3	BIEC2-808442	105,05	Jarret	HaploIBD	0,90	0,89	5,52				6,96	
3	BIEC2-808606	105,87	Cisaillage	Armitage	0,67	0,67	4,86	5,14				
3	BIEC2-808617	105,88	Jarret	Armitage	0,56	0,56	5,89	4,42				
3	BIEC2-808617	105,88	Jarret	SNPMixed	0,56	0,56	5,20	<i>3,83</i>	0,29	0,24	4,14	2,84
3	BIEC2-808653	106,05	Cisaillage	Armitage	0,40	0,40	4,76	5,23				
3	BIEC2-808879	106,87	Jarret	Armitage	0,72	0,71	5,49	6,18				
3	BIEC2-808879	106,87	Jarret	SNPMixed	0,72	0,71	<i>4,04</i>	4,35	-0,28	-0,28	3,16	3,21
3	BIEC2-809387	110,01	Total	Armitage	0,59	0,59	4,45	4,40				
4	BIEC2-867526	58,66	Compression	Armitage	0,25	0,25	4,43	<i>4,19</i>				
5	BIEC2-898230	23,48	Autres	Armitage	0,71	0,71	4,71	<i>3,08</i>				
5	BIEC2-898230	23,48	Autres	SNPMixed	0,71	0,71	4,51	<i>2,82</i>	-0,28	-0,21	3,25	1,83
5	BIEC2-908686	50,02	Cisaillage	Armitage	0,39	0,39	5,31	4,92				
5	BIEC2-908686	50,02	Cisaillage	SNPMixed	0,39	0,39	4,60	<i>4,16</i>	0,28	0,25	3,71	2,97
6	BIEC2-1178497	60,44	Autres	Armitage	0,91	0,91	5,15	4,46				
6	BIEC2-1178497	60,44	Autres	SNPMixed	0,91	0,91	5,17	4,60	-0,47	-0,42	3,77	2,96
7	BIEC2-987750	25,86	Cisaillage	Armitage	0,71	0,70	4,31	<i>4,00</i>				
8	BIEC2-1050380	49,47	Jarret	Armitage	0,34	0,35	4,50	4,77				
8	BIEC2-1055892	59,98	Tension	Armitage	0,12	0,12	5,18	4,67				
8	BIEC2-1064290	82,88	Jarret	Armitage	0,40	0,42	<i>3,86</i>	4,89				
11	BIEC2-141902	17,78	Cisaillage	Armitage	0,73	0,73	4,85	<i>3,62</i>				
13	BIEC2-203529	2,16	Tension	Armitage	0,08	0,08	<i>3,85</i>	4,45				
13	BIEC2-207631	7,55	Total	Armitage	0,22	0,22	<i>3,58</i>	4,54				
13	BIEC2-208655	8,39	LSG	SNPMixed	0,49	0,49	4,75	<i>3,83</i>	0,27	0,23	3,64	2,64
13	BIEC2-208655	8,39	Total	Armitage	0,49	0,49	4,65	<i>4,18</i>				
13	BIEC2-208753	8,49	LSG	HaploIBD	0,79	0,78	4,46				1,88	
13	BIEC2-209278	9,56	LSG	SNPMixed	0,59	0,58	<i>3,65</i>	4,48	-0,25	-0,26	3,02	3,29
13	BIEC2-209568	9,89	Boulet	Armitage	0,76	0,77	4,37	4,62				
13	BIEC2-209568	9,89	Boulet	SNPMixed	0,76	0,77	<i>4,04</i>	4,36	-0,30	-0,30	3,31	3,23
13	BIEC2-215491	20,31	Autres	HaploIBD	0,63	0,63	4,54				25,97	
14	BIEC2-265953	73,76	Jarret	Armitage	0,23	0,23	5,54	4,49				
14	BIEC2-265953	73,76	Jarret	SNPMixed	0,23	0,23	4,99	<i>4,16</i>	0,35	0,30	4,35	3,15
14	BIEC2-265956	73,87	Jarret	HaploIBD	0,43	0,42	4,47				3,23	
14	BIEC2-265995	74,08	LSG	SNPMixed	0,24	0,24	4,67	<i>3,48</i>	0,39	0,31	5,60	3,50
14	BIEC2-265995	74,08	Total	Armitage	0,24	0,24	4,44	<i>3,25</i>				
14	BIEC2-265995	74,08	Total	SNPMixed	0,24	0,24	4,34	<i>3,19</i>	0,36	0,29	4,73	3,07
15	BIEC2-293832	20,22	Jarret	Armitage	0,80	0,81	5,83	5,95				
15	BIEC2-293832	20,22	Jarret	SNPMixed	0,80	0,81	4,57	4,39	-0,34	-0,32	3,66	3,20
15	BIEC2-296478	24,69	Jarret	SNPMixed	0,83	0,83	4,47	<i>3,94</i>	-0,35	-0,31	3,42	2,66
15	BIEC2-310861	52,47	Cisaillage	SNPMixed	0,72	0,71	4,87	5,58	-0,34	-0,35	4,71	5,02
15	BIEC2-310861	52,47	Cisaillage	Armitage	0,72	0,71	5,30	6,10				
15	BIEC2-320543	75,92	Autres	HaploIBD	0,76	0,77	5,17				40,63	
15	BIEC2-320636	76,22	Autres	Armitage	0,20	0,20	4,59	<i>3,67</i>				
15	BIEC2-320636	76,22	Autres	SNPMixed	0,20	0,20	4,45	<i>3,52</i>	0,34	0,29	3,69	2,67
15	BIEC2-325537	87,91	Cisaillage	Armitage	0,23	0,22	<i>3,96</i>	5,11				
15	BIEC2-325537	87,91	LSG	SNPMixed	0,23	0,22	<i>3,53</i>	4,92	-0,28	-0,31	2,74	3,27
15	BIEC2-325537	87,91	Total	SNPMixed	0,23	0,22	<i>3,60</i>	4,71	-0,28	-0,32	2,78	3,51
15	BIEC2-326013	89,62	LSG	SNPMixed	0,20	0,19	<i>3,54</i>	4,36	-0,30	-0,32	2,88	3,21
15	BIEC2-326013	89,62	Total	SNPMixed	0,20	0,19	<i>4,06</i>	5,02	-0,32	-0,35	3,28	3,77
15	BIEC2-326013	89,62	Total	Armitage	0,20	0,19	4,73	5,93				
16	BIEC2-328091	3,47	Jarret	Armitage	0,92	0,91	<i>3,23</i>	4,51				
16	BIEC2-329987	13,42	Total	Armitage	0,25	0,25	<i>4,24</i>	4,58				

Suite page suivante

ECA ¹	SNP	Pos ²	Trait	Méthode	freq A ₂ ³		P-valeur ⁴		β ⁵		V(QTL) ⁶	
					P ₁ ⁷	P ₂ ⁸	P ₁ ⁷	P ₂ ⁸	P ₁ ⁷	P ₂ ⁸	P ₁ ⁷	P ₂ ⁸
16	BIEC2-331093	18,56	Jarret	Armitage	0,10	0,10	3,93	4,42				
16	BIEC2-344288	42,93	Boulet	Armitage	0,84	0,83	4,66	3,56				
16	BIEC2-344288	42,93	Boulet	SNPMixed	0,84	0,83	4,30	3,10	-0,36	-0,28	3,55	2,18
16	BIEC2-365627	86,18	Boulet	SNPMixed	0,08	0,08	4,69	4,35	0,49	0,44	3,65	2,95
16	BIEC2-365627	86,18	Boulet	Armitage	0,08	0,08	4,47	4,19				
16	BIEC2-365627	86,18	Total	Armitage	0,08	0,08	4,89	4,91				
16	BIEC2-365627	86,18	Total	SNPMixed	0,08	0,08	4,52	4,49	0,48	0,45	3,39	2,98
17	BIEC2-384530	75,57	Autres	Armitage	0,31	0,31	4,47	4,35				
17	BIEC2-384530	75,57	Autres	SNPMixed	0,31	0,31	4,32	4,25	0,29	0,27	3,57	3,11
17	BIEC2-387172	79,68	Jarret	Armitage	0,37	0,37	3,31	4,73				
20	BIEC2-519527	13,04	Jarret	Armitage	0,70	0,69	2,87	4,90				
20	BIEC2-533812	45,00	Cisaillage	Armitage	0,07	0,07	5,14	4,46				
20	BIEC2-541255	56,24	Cisaillage	Armitage	0,19	0,19	4,42	6,02				
20	BIEC2-541255	56,24	Cisaillage	SNPMixed	0,19	0,19	3,87	4,93	-0,31	-0,34	3,02	3,55
20	BIEC2-541255	56,24	Total	Armitage	0,19	0,19	3,95	4,95				
22	BIEC2-577112	4,84	Boulet	Armitage	0,34	0,34	4,35	4,17				
22	BIEC2-579755	8,49	Autres	Armitage	0,86	0,86	4,88	4,80				
22	BIEC2-579755	8,49	Autres	SNPMixed	0,86	0,86	4,60	4,55	-0,40	-0,37	3,75	3,28
25	BIEC2-669447	31,72	LSG	SNPMixed	0,95	0,96	5,09	4,07	-0,62	-0,53	3,34	2,35
26	BIEC2-690165	22,45	Tension	Armitage	0,92	0,92	4,92	4,00				
26	BIEC2-690975	24,19	Cisaillage	Armitage	0,79	0,79	3,14	4,36				
27	BIEC2-720009	35,72	Boulet	Armitage	0,54	0,55	3,93	4,54				
28	BIEC2-729305	11,37	Tension	Armitage	0,79	0,78	3,84	5,09				
28	BIEC2-732594	17,95	Cisaillage	Armitage	0,63	0,62	3,71	4,56				
28	BIEC2-733463	19,53	LSG	SNPMixed	0,92	0,92	3,79	4,35	-0,46	-0,47	3,13	3,33
28	BIEC2-733463	19,53	Total	SNPMixed	0,92	0,92	3,76	4,66	-0,45	-0,49	2,98	3,53
28	BIEC2-733541	19,65	Cisaillage	Armitage	0,71	0,71	4,88	4,15				
28	BIEC2-733541	19,65	Cisaillage	SNPMixed	0,71	0,71	4,73	4,18	-0,31	-0,28	3,95	3,20
28	BIEC2-733541	19,65	Total	Armitage	0,71	0,71	4,70	4,13				
28	BIEC2-733541	19,65	Total	SNPMixed	0,71	0,71	4,35	3,72	-0,29	-0,26	3,46	2,78
28	BIEC2-735476	23,78	Boulet	Armitage	0,69	0,69	3,42	4,39				

1. ECA : *Equus Callabus* chromosome

2. Pos : Position en Mégabase (Mb)

3. freq A₂ : Fréquence de l'allèle 2

4. Pvalue : $-\log_{10}(P)$

5. β : Estimation de l'effet du SNP en écart-type phénotypique (possible uniquement pour SNP-Mixed)

6. V(QTL) : Estimation de la variance phénotypique expliquée par le QTL. Pour SNPMixed : $V(QTL) = 2pq\beta^2$; Pour HaploIBD : $V(QTL) = \sigma_{QTL}^2 / \sigma_y^2$

7. P₁ : Population avec 525 chevaux

8. P₂ : Population avec 583 chevaux

Conclusion générale

L'objectif de cette thèse consistait à trouver les QTL associés à l'ostéochondrose chez les Trotteurs Français. L'utilisation de marqueurs SNP cartographiés, dense, nous a permis d'utiliser des analyses d'association. Le principal problème identifié avec ces analyses est leur non robustesse lorsque une structure familiale entre les individus phénotypés et un fond polygénique sont présents. Dès lors, plusieurs travaux se sont développés pour prendre en compte cet apparentement entre individus. Un consensus autour de l'utilisation du modèle mixte, couplant l'effet d'un marqueur et un effet aléatoire polygénique, semble depuis peu voir le jour pour corriger ces problèmes.

Nous avons cherché à comprendre et évaluer ce phénomène en comparant algébriquement certains modèles et méthodes couramment utilisées (Régression, QTDT, et deux modèles simplifiés du modèle mixte). Ces dérivations algébriques des formules permettent une généralisation des résultats, alors que les simulations sont toujours réalisées dans un cadre précis. Elles permettent d'évaluer la robustesse et la puissance espérées de n'importe quel dispositif défini par un nombre d'individus, une matrice de parenté et une héritabilité. Elles permettent aussi de planifier le dispositif expérimental pour obtenir une puissance suffisante pour détecter un QTL d'effet donné sur la variance du caractère étudié. Ce travail était focalisé sur les problèmes liés à la structure familiale des données. Une extension à envisager serait l'étude de mélanges entre populations. De plus, d'autres modèles ou méthodes pourraient être testées de la même manière. Les méthodes présentées ici sont toutes uni-SNP et de ce fait uni-QTL. Il faudrait envisager l'extension vers des méthodes uni-QTL haplotypiques et multi-QTL.

La deuxième partie de la thèse était consacrée à l'étude des données de l'OC du programme GENEQUIN. L'objectif était de réaliser une cartographie fine des régions QTL de plusieurs caractères mesurant l'OC en différents sites articulaires et type de contrainte. Cette étude a permis de mettre en évidence plusieurs régions QTL d'effets moyens et faibles à un niveau significatif mais pas hautement significatif. Plusieurs régions QTL ont été détectés mais certaines, notamment sur les chromosomes ECA 3, 13, 15 et 28, semblent être d'un intérêt supérieur. Nous montrons que l'OC n'est pas gouvernée par un gène majeur car un tel gène eut été détecté avec la puissance de notre protocole (compte tenu du nombre d'individus et du niveau de déséquilibre de liaison). L'OC est donc, comme on le suspectait, un caractère multifactoriel et polygénique où plusieurs QTL de faibles et moyens effets sont responsables de la variabilité du caractère. De plus, nous n'avons pas pu mettre en évidence de QTL ayant un effet à la fois sur l'OC du jarret et l'OC du boulet, ce qui

infirmes l'hypothèse simple d'une cause génétique commune de la sensibilité à cette maladie sur les différents sites anatomiques. Des études plus approfondies sur ces régions QTL, notamment l'identification des gènes candidats et mutations causales devrait clarifier la physiopathologie moléculaire de l'OC.

Les avancées technologiques dans le monde de la génétique sont considérables depuis une dizaine d'années. On parlait au début des années 2000 de génotyper un animal avec un marqueur microsatellite pour 1 euro, vers 2008 de génotyper des dizaines de milliers de SNP autour des 300 euros et dans un avenir proche, de génotyper des centaines de milliers de SNP ou même séquencer des génomes entiers à des coûts très raisonnables. Ces avancées devraient amener un gain de puissance et de localisation important dans les études sur tout le génome sans que les modèles n'aient à être modifiés. Les véritables enjeux pour lutter contre l'OC seront dans la construction des phénotypes (regroupement des différentes formes de l'OC ? par site articulaire ? mesures ?), dans le développement de stratégies efficaces pour l'évaluation des risques individuels, et dans la création d'une population de référence de grande taille pour une possible mise en place d'une sélection assistée par marqueurs ou d'une sélection génomique.



Annexes A : Complément de l'article méthodologique 2.2.2

1 **Details about algebraic formulae.**

2

3

4 **MODEL 1, REGRESSION MODEL**

5
$$V(\hat{\beta}^{(1)}) = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'V(\mathbf{y})\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'(\mathbf{A}\sigma_u^2 + \mathbf{I}\sigma_e^2)\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1} = \sigma_e^2 \left[(\mathbf{x}'\mathbf{x})^{-1} + \frac{h^2}{1-h^2}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{A}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1} \right]$$

6
$$E(\hat{\mathbf{e}}^{(1)'}\hat{\mathbf{e}}^{(1)}) = tr((\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')V(\mathbf{y})) + E(\mathbf{y})'(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')E(\mathbf{y}),$$

7 Developing $E(\mathbf{y})'(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')E(\mathbf{y})$:

8
$$\begin{aligned} E(\mathbf{y})'(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')E(\mathbf{y}) &= (\mu\mathbf{1}' + \beta\mathbf{x}')(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')(\mu\mathbf{1} + \mathbf{x}\beta) \\ &= (\mu\mathbf{1}' + \beta\mathbf{x}')(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')(\mu\mathbf{1} + \mathbf{x}\beta - \mathbf{x}\beta - \mu\mathbf{1}) \\ &= 0 \end{aligned}$$

9 Because $\mathbf{x}'\mathbf{1} = (\mathbf{w} - \mathbf{1}\bar{w})'\mathbf{1} = 0$

10 Finally :

11
$$\begin{aligned} E(\hat{\mathbf{e}}^{(1)'}\hat{\mathbf{e}}^{(1)}) &= tr((\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')V(\mathbf{y})) \\ &= tr\left[(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')(\mathbf{A}\sigma_u^2 + \mathbf{I}\sigma_e^2)\right] \\ &= \sigma_e^2 \left[(n-2) + \frac{h^2}{(1-h^2)}(tr(\mathbf{A}) - (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{A}\mathbf{x} - (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{A}\mathbf{1}) \right] \end{aligned}$$

12

13 **MODEL 2 , GRAMMAR MODEL**

14 In order to simplify notation, in that entire section the reference (2a) was replaced by (2).

15 Following equalities were used:

16
$$\begin{bmatrix} \mathbf{C}_{11}^{(2)} & \mathbf{C}_{1u}^{(2)} \\ \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}' \\ \mathbf{1} & \mathbf{I} + \lambda^{(2)}\mathbf{A}^{-1} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

17
$$\begin{bmatrix} \mathbf{C}_{11}^{(2)} & \mathbf{C}_{1u}^{(2)} \\ \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}' \\ \mathbf{1} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{I} \end{bmatrix} - \begin{bmatrix} 0 & \lambda^{(2)}\mathbf{C}_{1u}^{(2)}\mathbf{A}^{-1} \\ \mathbf{0} & \lambda^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{A}^{-1} \end{bmatrix}$$

18 So :

$$19 \quad \mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)} = \mathbf{I} - \lambda^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{A}^{-1} \quad \text{ou} \quad \mathbf{1}\mathbf{C}_{1u}^{(2)} + \mathbf{C}_{uu}^{(2)} = \mathbf{I} - \lambda^{(2)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(2)}$$

$$20 \quad \mathbf{C}_{u1}^{(2)} = -\frac{1}{n}\mathbf{C}_{uu}^{(2)}\mathbf{1} \quad \text{or} \quad \mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)}\mathbf{1} = 0$$

$$21 \quad \mathbf{C}_{11}^{(2)}\mathbf{1}' + \mathbf{C}_{1u}^{(2)} = -\mathbf{C}_{1u}^{(2)}\lambda^{(2)}\mathbf{A}^{-1} \quad \text{or} \quad \mathbf{1}\mathbf{C}_{11}^{(2)} + \mathbf{C}_{u1}^{(2)} = -\mathbf{A}^{-1}\mathbf{C}_{u1}^{(2)}\lambda^{(2)}$$

$$22 \quad \mathbf{C}_{11}^{(2)}\mathbf{1}' + \mathbf{C}_{1u}^{(2)}\mathbf{1} = 1 \quad \text{or} \quad \mathbf{1}'\mathbf{1}\mathbf{C}_{11}^{(2)} + \mathbf{1}'\mathbf{C}_{u1}^{(2)} = 1$$

23 CALCULATION OF $V(\hat{\mathbf{u}}^{(2)})$

24 To express $V(\hat{\mathbf{u}}^{(2)})$, the objective was to separate the classical expression $(\mathbf{A}\sigma_u^2 - \mathbf{C}_{uu}^{(2)}\lambda^{(2)}\sigma_u^2)$
 25 from the part corresponding to the difference between variance components of the true model
 26 and variance components of the pure random model (2a): $(\sigma_e^2 - \lambda^{(2)}\sigma_u^2)\mathbf{C}_{uu}^{(2)}(\mathbf{I} - \lambda^{(2)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(2)})$

27 (if $\lambda^{(2)} = \frac{\sigma_e^2}{\sigma_u^2}$, instead of $\lambda^{(2)} = \frac{\sigma_e^{2(2a)}}{\sigma_u^{2(2a)}}$ this last term disappear)

$$28 \quad V(\hat{\mathbf{u}}^{(2)}) = (\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)})V(\mathbf{y})(\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)})' \\ = (\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)})(\mathbf{A}\sigma_u^2 + \mathbf{I}\sigma_e^2)(\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)})'$$

$$29 \quad = (\mathbf{I} - \mathbf{C}_{uu}^{(2)}\lambda^{(2)}\mathbf{A}^{-1})\mathbf{A}\sigma_u^2(\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)})' + \sigma_e^2(\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)})(\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)})'$$

30 (in left we used $\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)} = \mathbf{I} - \mathbf{C}_{uu}^{(2)}\lambda^{(2)}\mathbf{A}^{-1}$)

$$= \sigma_u^2\mathbf{A}(\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)})' - \sigma_u^2\mathbf{C}_{uu}^{(2)}\lambda^{(2)}(\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)})' + \sigma_e^2(\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)})(\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)})' \\ 31 \quad = \sigma_u^2\mathbf{A}(\mathbf{I} - \mathbf{A}^{-1}\mathbf{C}_{uu}^{(2)}\lambda^{(2)}) + (\sigma_e^2(\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)}) - \sigma_u^2\mathbf{C}_{uu}^{(2)}\lambda^{(2)})(\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)})' \\ = (\mathbf{A}\sigma_u^2 - \mathbf{C}_{uu}^{(2)}\lambda^{(2)}\sigma_u^2) + (\sigma_e^2\mathbf{C}_{uu}^{(2)} - \sigma_u^2\mathbf{C}_{uu}^{(2)}\lambda^{(2)})(\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)})'$$

32 (because $\mathbf{C}_{u1}^{(2)}\mathbf{1}'\mathbf{1}\mathbf{C}_{1u}^{(2)} + \mathbf{C}_{u1}^{(2)}\mathbf{1}'\mathbf{C}_{uu}^{(2)} = 0$)

33 So that :

$$34 \quad V(\hat{\mathbf{u}}^{(2)}) = \sigma_u^2(\mathbf{A} - \lambda^{(2)}\mathbf{C}_{uu}^{(2)}) + (\sigma_e^2 - \lambda^{(2)}\sigma_u^2)\mathbf{C}_{uu}^{(2)}(\mathbf{I} - \lambda^{(2)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(2)})$$

35 CALCULATION OF $E(\hat{\mathbf{u}}^{(2)})$

$$36 \quad E(\hat{\mathbf{u}}^{(2)}) = (\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)})E(\mathbf{y}) = (\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)})(\mathbf{1}\mu + \mathbf{x}\beta) = (\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)})\mathbf{x}\beta = \mathbf{C}_{uu}^{(2)}\mathbf{x}\beta$$

37 (because $\mathbf{C}_{u1}^{(2)}\mathbf{1}'\mathbf{1} + \mathbf{C}_{uu}^{(2)}\mathbf{1} = 0$ et $\mathbf{x}'\mathbf{1} = 0$)

38 CALCULATION OF $E(\hat{\beta}^{(2b)})$

$$39 \quad E(\hat{\beta}^{(2b)}) = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' E(\mathbf{y} - \mathbf{1}\hat{\mu}^{(2a)} - \hat{\mathbf{u}}^{(2a)}) \\ = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' E(\mathbf{y}) - (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' E(\hat{\mathbf{u}}^{(2a)})$$

40 (because $\mathbf{x}'\mathbf{1} = (\mathbf{w} - \bar{\mathbf{w}})' \mathbf{1} = 0$)

$$41 \quad E(\hat{\beta}^{(2b)}) = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' \mathbf{x} \beta - (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' \mathbf{C}_{uu} \mathbf{x} \beta \\ = \beta - (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' \mathbf{C}_{uu} \mathbf{x} \beta$$

42 CALCULATION OF $V(\hat{\beta}^{(2b)})$

43 Replacing $\hat{\mu}^{(2)}$ and $\hat{\mathbf{u}}^{(2)}$ by their expression in the mixed model equations:

$$44 \quad V(\hat{\beta}^{(2b)}) = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' V(\mathbf{y} - \mathbf{1}\hat{\mu}^{(2)} - \hat{\mathbf{u}}^{(2)}) \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1} \\ = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' V((\mathbf{I} - \mathbf{1} \begin{bmatrix} \mathbf{C}_{11}^{(2)} & \mathbf{C}_{1u}^{(2)} \\ \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix} - \begin{bmatrix} \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix}) \mathbf{y}) \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}$$

$$45 \quad V(\hat{\beta}^{(2b)}) = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' V((\mathbf{I} - \begin{bmatrix} \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix}) \mathbf{y}) \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}$$

46 Because $\mathbf{x}'\mathbf{1} = 0$

$$47 \quad V(\hat{\beta}^{(2b)}) = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' (\mathbf{I} - \begin{bmatrix} \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix}) \mathbf{A} \sigma_u^2 (\mathbf{I} - \begin{bmatrix} \mathbf{1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{1u}^{(2)} \\ \mathbf{C}_{uu}^{(2)} \end{bmatrix}) + \sigma_e^2 (\mathbf{I} - \begin{bmatrix} \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix}) (\mathbf{I} - \begin{bmatrix} \mathbf{1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{1u}^{(2)} \\ \mathbf{C}_{uu}^{(2)} \end{bmatrix}) \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1} \\ = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' ((\mathbf{C}_{uu}^{(2)} \boldsymbol{\lambda}^{(2)} \mathbf{A}^{-1}) \mathbf{A} \sigma_u^2 (\mathbf{A}^{-1} \mathbf{C}_{uu}^{(2)} \boldsymbol{\lambda}^{(2)}) \\ + \sigma_e^2 (\mathbf{I} - \mathbf{C}_{u1}^{(2)} \mathbf{1}' - \mathbf{C}_{uu}^{(2)} - \mathbf{1} \mathbf{C}_{1u}^{(2)} + \mathbf{C}_{u1}^{(2)} \mathbf{1}' \mathbf{1} \mathbf{C}_{1u}^{(2)} + \mathbf{C}_{uu}^{(2)} \mathbf{1} \mathbf{C}_{1u}^{(2)} - \mathbf{C}_{uu}^{(2)} + \mathbf{C}_{u1}^{(2)} \mathbf{1}' \mathbf{C}_{uu}^{(2)} + \mathbf{C}_{uu}^{(2)} \mathbf{C}_{uu}^{(2)})) \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}$$

48 because $\mathbf{C}_{uu}^{(2)} \boldsymbol{\lambda}^{(2)} \mathbf{A}^{-1} = \mathbf{I} - \mathbf{C}_{u1}^{(2)} \mathbf{1}' - \mathbf{C}_{uu}^{(2)}$

$$49 \quad = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' (\boldsymbol{\lambda}^{(2)} \boldsymbol{\lambda}^{(2)}) \sigma_u^2 (\mathbf{C}_{uu}^{(2)} \mathbf{A}^{-1} \mathbf{C}_{uu}^{(2)}) + \sigma_e^2 (\mathbf{I} - 2\mathbf{C}_{uu}^{(2)} + \mathbf{C}_{u1}^{(2)} \mathbf{1}' \mathbf{1} \mathbf{C}_{1u}^{(2)} + \mathbf{C}_{uu}^{(2)} \mathbf{1} \mathbf{C}_{1u}^{(2)} + \mathbf{C}_{u1}^{(2)} \mathbf{1}' \mathbf{C}_{uu}^{(2)} + \mathbf{C}_{uu}^{(2)} \mathbf{C}_{uu}^{(2)}) \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1} \\ 50 \quad \text{because } \mathbf{x}'\mathbf{1} = 0$$

$$51 \quad = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' (\boldsymbol{\lambda}^{(2)} \boldsymbol{\lambda}^{(2)}) \sigma_u^2 (\mathbf{C}_{uu}^{(2)} \mathbf{A}^{-1} \mathbf{C}_{uu}^{(2)}) + \sigma_e^2 (\mathbf{I} - 2\mathbf{C}_{uu}^{(2)} + \mathbf{C}_{u1}^{(2)} \mathbf{1}' \mathbf{C}_{uu}^{(2)} + \mathbf{C}_{uu}^{(2)} \mathbf{C}_{uu}^{(2)}) \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}$$

52 Because $\mathbf{C}_{u1}^{(2)} \mathbf{1}' \mathbf{1} + \mathbf{C}_{uu}^{(2)} \mathbf{1} = 0$

$$53 \quad = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' (\boldsymbol{\lambda}^{(2)} \boldsymbol{\lambda}^{(2)}) \sigma_u^2 (\mathbf{C}_{uu}^{(2)} \mathbf{A}^{-1} \mathbf{C}_{uu}^{(2)}) + \sigma_e^2 (\mathbf{I} - 2\mathbf{C}_{uu}^{(2)} + (\mathbf{I} - \boldsymbol{\lambda}^{(2)} \mathbf{C}_{uu}^{(2)} \mathbf{A}^{-1}) \mathbf{C}_{uu}^{(2)}) \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}$$

54 because $\mathbf{C}_{\text{ul}}^{(2)}\mathbf{1}' + \mathbf{C}_{\text{uu}}^{(2)} = \mathbf{I} - \mathbf{C}_{\text{uu}}^{(2)}\lambda^{(2)}\mathbf{A}^{-1}$

$$\begin{aligned} &= (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'(\lambda^{(2)}\lambda^{(2)}\sigma_u^2(\mathbf{C}_{\text{uu}}^{(2)}\mathbf{A}^{-1}\mathbf{C}_{\text{uu}}^{(2)}) + \sigma_e^2(\mathbf{I} - \mathbf{C}_{\text{uu}}^{(2)} - \lambda^{(2)}\mathbf{C}_{\text{uu}}^{(2)}\mathbf{A}^{-1}\mathbf{C}_{\text{uu}}^{(2)}))\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1} \\ 55 &= \sigma_e^2\left((\mathbf{x}'\mathbf{x})^{-1} - (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{C}_{\text{uu}}^{(2)}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\right) - (\sigma_e^2 - \lambda^{(2)}\sigma_u^2)\lambda^{(2)}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{C}_{\text{uu}}^{(2)}\mathbf{A}^{-1}\mathbf{C}_{\text{uu}}^{(2)}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1} \end{aligned}$$

56 To express this formula only with \mathbf{A} and $\mathbf{C}_{\text{uu}}^{(2)}$ and not \mathbf{A}^{-1} to understand directly the effect of
57 relationships, we know that:

$$58 \quad \mathbf{C}_{\text{ul}}^{(2)}\mathbf{1}' + \mathbf{C}_{\text{uu}}^{(2)} = \mathbf{I} - \mathbf{C}_{\text{uu}}^{(2)}\lambda^{(2)}\mathbf{A}^{-1} \text{ and } \mathbf{C}_{\text{ul}}^{(2)} = -\frac{1}{n}\mathbf{C}_{\text{uu}}^{(2)}\mathbf{1}, \text{ so } \mathbf{C}_{\text{uu}}^{(2)}\lambda^{(2)}\mathbf{A}^{-1} = \mathbf{I} - \mathbf{C}_{\text{uu}}^{(2)}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)$$

59 So :

$$60 \quad V(\hat{\beta}^{(2b)}) = \sigma_e^2\left((\mathbf{x}'\mathbf{x})^{-1} - (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{C}_{\text{uu}}^{(2)}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\right) - (\sigma_e^2 - \lambda^{(2)}\sigma_u^2)(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\left(\mathbf{C}_{\text{uu}}^{(2)} - \mathbf{C}_{\text{uu}}^{(2)}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{C}_{\text{uu}}^{(2)}\right)\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}$$

$$61 \quad = \sigma_e^2\left((\mathbf{x}'\mathbf{x})^{-1} - (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{C}_{\text{uu}}^{(2)}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\right) - (\sigma_e^2 - \lambda^{(2)}\sigma_u^2)(\mathbf{x}'\mathbf{x})^{-1}\left(\mathbf{x}'\mathbf{C}_{\text{uu}}^{(2)}\mathbf{x} - \mathbf{x}'\mathbf{C}_{\text{uu}}^{(2)}\mathbf{C}_{\text{uu}}^{(2)}\mathbf{x} + \frac{1}{n}\mathbf{x}'\mathbf{C}_{\text{uu}}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{C}_{\text{uu}}^{(2)}\mathbf{x}\right)(\mathbf{x}'\mathbf{x})^{-1}$$

62

63 The term $\mathbf{x}'\mathbf{C}_{\text{uu}}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{C}_{\text{uu}}^{(2)}\mathbf{x}$ was zero when family sizes are equal sizes because:

$$\begin{aligned} &E_x(\mathbf{x}'\mathbf{C}_{\text{uu}}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{C}_{\text{uu}}^{(2)}\mathbf{x}) = \mathbf{1}'\mathbf{A}\mathbf{1} - \lambda^{(2)}\mathbf{1}'\mathbf{C}_{\text{uu}}^{(2)}\mathbf{1} \\ &= \mathbf{1}'(\mathbf{A} - \lambda^{(2)}\mathbf{C}_{\text{uu}}^{(2)})\mathbf{1} \\ 64 &= \mathbf{1}'(\mathbf{C}_{\text{ul}}^{(2)}\mathbf{1}'\mathbf{A} + \mathbf{C}_{\text{uu}}^{(2)}\mathbf{A})\mathbf{1} \\ &= \mathbf{1}'\left(-\frac{1}{n}\mathbf{C}_{\text{uu}}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{A} + \mathbf{C}_{\text{uu}}^{(2)}\mathbf{A}\right)\mathbf{1} \\ &= -\frac{1}{n}\mathbf{1}'\mathbf{C}_{\text{uu}}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1} + \mathbf{1}'\mathbf{C}_{\text{uu}}^{(2)}\mathbf{A}\mathbf{1} \end{aligned}$$

65 Which is the covariance between the sum of column of \mathbf{A} and $\mathbf{C}_{\text{uu}}^{(2)}$, equal to zero when
66 equilibrium design because there was no variance between columns.

67 CALCULATION OF $E(\hat{\mathbf{e}}^{(2b)}\mathbf{1}'\hat{\mathbf{e}}^{(2b)})$

$$\begin{aligned} &\hat{\mathbf{e}}^{(2b)} = \hat{\mathbf{e}}^{(2)} - \mathbf{1}\hat{\mu}^{(2b)} - \mathbf{x}\hat{\beta}^{(2b)} \\ 68 &= \left(\mathbf{I} - [\mathbf{1} \quad \mathbf{x}]\begin{bmatrix} \mathbf{1}'\mathbf{1} & 0 \\ 0 & \mathbf{x}'\mathbf{x} \end{bmatrix}^{-1}\begin{bmatrix} \mathbf{1}' \\ \mathbf{x}' \end{bmatrix}\right)\hat{\mathbf{e}}^{(2)} \\ &= (\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')\hat{\mathbf{e}}^{(2)} \end{aligned}$$

69 because $\mathbf{x}'\mathbf{1} = 0$

$$70 \quad E(\hat{\boldsymbol{\epsilon}}^{(2b)})' \hat{\boldsymbol{\epsilon}}^{(2b)} = \text{tr}((\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')V(\hat{\boldsymbol{\epsilon}}^{(2)})) + E(\hat{\boldsymbol{\epsilon}}^{(2)})'(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')E(\hat{\boldsymbol{\epsilon}}^{(2)})$$

71 PART WITH $V(\hat{\boldsymbol{\epsilon}}^{(2)})$

72 We had:

$$73 \quad V(\hat{\boldsymbol{\epsilon}}^{(2)}) = V\left(\left(\mathbf{I} - \mathbf{1}\begin{bmatrix} \mathbf{C}_{11}^{(2)} & \mathbf{C}_{1u}^{(2)} \\ \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix} - \begin{bmatrix} \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \\ \mathbf{C}_{11}^{(2)} & \mathbf{C}_{1u}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix}\right)\mathbf{y})$$

$$= \left(\mathbf{I} - \mathbf{1}\begin{bmatrix} \mathbf{C}_{11}^{(2)} & \mathbf{C}_{1u}^{(2)} \\ \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix} - \begin{bmatrix} \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \\ \mathbf{C}_{11}^{(2)} & \mathbf{C}_{1u}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix}\right)V(\mathbf{y}) \left(\mathbf{I} - \begin{bmatrix} \mathbf{1} & \mathbf{I} \\ \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \end{bmatrix} \mathbf{1}' - \begin{bmatrix} \mathbf{1} & \mathbf{I} \\ \mathbf{C}_{11}^{(2)} & \mathbf{C}_{1u}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{u1}^{(2)} \\ \mathbf{C}_{uu}^{(2)} \end{bmatrix}\right)$$

74 So that the trace was :

$$75 \quad \begin{aligned} & \text{tr}((\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')V(\hat{\boldsymbol{\epsilon}}^{(2)})) \\ & = \text{tr}((\mathbf{I} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}') \left(\mathbf{I} - \mathbf{1}\begin{bmatrix} \mathbf{C}_{11}^{(2)} & \mathbf{C}_{1u}^{(2)} \\ \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix} - \begin{bmatrix} \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \\ \mathbf{C}_{11}^{(2)} & \mathbf{C}_{1u}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix}\right)V(\mathbf{y}) \left(\mathbf{I} - \begin{bmatrix} \mathbf{1} & \mathbf{I} \\ \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \end{bmatrix} \mathbf{1}' - \begin{bmatrix} \mathbf{1} & \mathbf{I} \\ \mathbf{C}_{11}^{(2)} & \mathbf{C}_{1u}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{u1}^{(2)} \\ \mathbf{C}_{uu}^{(2)} \end{bmatrix}\right)) \\ & \quad - \text{tr}((\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'V(\hat{\boldsymbol{\epsilon}}^{(2)})\mathbf{x}) \end{aligned}$$

76 In the first term, all terms which began by $\mathbf{1}$ or ended by $\mathbf{1}'$ were suppressed because

77 $\text{tr}((\mathbf{I} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')\mathbf{B}) = 0$ if $\mathbf{B} = \mathbf{1}\mathbf{G}\mathbf{1}'$ or $\mathbf{B} = \mathbf{1}\mathbf{G}$ or $\mathbf{B} = \mathbf{G}\mathbf{1}'$. The second term was used in the

78 calculation of $V(\hat{\boldsymbol{\beta}}^{(2b)})$ and we used the same developments:

$$79 \quad \begin{aligned} & = \text{tr}((\mathbf{I} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}') \left(\mathbf{I} - \begin{bmatrix} \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \\ \mathbf{C}_{11}^{(2)} & \mathbf{C}_{1u}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix}\right)V(\mathbf{y}) \left(\mathbf{I} - \begin{bmatrix} \mathbf{1} & \mathbf{I} \\ \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \end{bmatrix}\right)) \\ & \quad - \text{tr}(\sigma_e^2(1 - (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{x}) - (\sigma_e^2 - \lambda^{(2)}\sigma_u^2)\lambda^{(2)}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(2)}\mathbf{x}) \end{aligned}$$

80 In the first term, we used $V(\mathbf{y}) = \mathbf{A}\sigma_u^2 + \mathbf{I}\sigma_e^2$ and $\mathbf{C}_{u1}^{(2)}\mathbf{1}' + \mathbf{C}_{uu}^{(2)} = \mathbf{I} - \mathbf{C}_{uu}^{(2)}\lambda^{(2)}\mathbf{A}^{-1}$. For the second

81 term, the trace was useless because it was a scalar.

$$82 \quad \begin{aligned} & = \text{tr}((\mathbf{I} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')(\mathbf{C}_{uu}^{(2)}\lambda^{(2)}\mathbf{A}^{-1}\mathbf{A}\mathbf{A}^{-1}\lambda^{(2)}\mathbf{C}_{uu}^{(2)}\sigma_u^2 + \sigma_e^2(\mathbf{I} - 2\mathbf{C}_{uu}^{(2)} + (\mathbf{I} - \mathbf{C}_{uu}^{(2)}\lambda^{(2)}\mathbf{A}^{-1})\mathbf{C}_{uu}^{(2)})) \\ & \quad - \sigma_e^2(1 - (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{x}) + (\sigma_e^2 - \lambda^{(2)}\sigma_u^2)\lambda^{(2)}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(2)}\mathbf{x}) \end{aligned}$$

$$83 \quad \begin{aligned} & = \text{tr}((\mathbf{I} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')(\mathbf{C}_{uu}^{(2)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(2)}\lambda^{(2)}\lambda^{(2)}\sigma_u^2 + \sigma_e^2(\mathbf{I} - \mathbf{C}_{uu}^{(2)} - \mathbf{C}_{uu}^{(2)}\lambda^{(2)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(2)})) \\ & \quad - \sigma_e^2(1 - (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{x}) + (\sigma_e^2 - \lambda^{(2)}\sigma_u^2)\lambda^{(2)}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(2)}\mathbf{x}) \end{aligned}$$

$$84 \quad \begin{aligned} & = \text{tr}((\mathbf{I} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')(\sigma_e^2(\mathbf{I} - \mathbf{C}_{uu}^{(2)}) - (\sigma_e^2 - \lambda^{(2)}\sigma_u^2)(\mathbf{C}_{uu}^{(2)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(2)}\lambda^{(2)})) \\ & \quad - \sigma_e^2(1 - (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{x}) + (\sigma_e^2 - \lambda^{(2)}\sigma_u^2)\lambda^{(2)}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(2)}\mathbf{x}) \end{aligned}$$

85 Terms were then grouped in function of variances:

$$\begin{aligned}
86 & -\sigma_e^2 \left(1 - (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{x} \right) + (\sigma_e^2 - \lambda^{(2)}\sigma_u^2) \lambda^{(2)} (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(2)}\mathbf{x} \\
87 & = \sigma_e^2 (n-2 - \text{tr}(\mathbf{C}_{uu}^{(2)}) + (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{x} + (\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1}) \\
& - (\sigma_e^2 - \lambda^{(2)}\sigma_u^2) \lambda^{(2)} \left(\text{tr}(\mathbf{C}_{uu}^{(2)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(2)}) - (\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(2)}\mathbf{1} - (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(2)}\mathbf{x} \right)
\end{aligned}$$

88 PART WITH $E(\hat{\mathbf{e}}^{(2)})$

$$89 \quad E(\hat{\mathbf{e}}^{(2a)}) = (\mathbf{I} - \mathbf{1} \begin{bmatrix} \mathbf{C}_{11}^{(2)} & \mathbf{C}_{1u}^{(2)} \\ \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix} - \begin{bmatrix} \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \\ \mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix}) E(\mathbf{y})$$

$$= (\mathbf{I} - \mathbf{1} \begin{bmatrix} \mathbf{C}_{11}^{(2)} & \mathbf{C}_{1u}^{(2)} \\ \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix} - \begin{bmatrix} \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \\ \mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix}) (\mathbf{1}\mu + \mathbf{x}\beta)$$

$$90 \quad E(\hat{\mathbf{e}}^{(2)}) = \mathbf{1}\mu + \mathbf{x}\beta - \mathbf{1} \begin{bmatrix} \mathbf{C}_{11}^{(2)} & \mathbf{C}_{1u}^{(2)} \\ \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{1} & \mathbf{x} \\ \mu & \beta \end{bmatrix} - \begin{bmatrix} \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \\ \mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{1} & \mathbf{x} \\ \mu & \beta \end{bmatrix}$$

$$= \mathbf{1}\mu + \mathbf{x}\beta - \mathbf{1} \begin{bmatrix} \mathbf{C}_{11}^{(2)} & \mathbf{C}_{1u}^{(2)} \\ \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{1}'\mathbf{1} & 0 \\ \mathbf{1} & \mathbf{x} \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} - \begin{bmatrix} \mathbf{C}_{u1}^{(2)} & \mathbf{C}_{uu}^{(2)} \\ \mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{1}'\mathbf{1} & 0 \\ \mathbf{1} & \mathbf{x} \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix}$$

91 Using $\mathbf{x}'\mathbf{1} = 0$

$$= \mathbf{1}\mu + \mathbf{x}\beta - \mathbf{1} \begin{bmatrix} \mathbf{1} & \mathbf{C}_{1u}^{(2)}\mathbf{x} \\ 0 & \mathbf{C}_{uu}^{(2)}\mathbf{x} \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} - \begin{bmatrix} 0 & \mathbf{C}_{uu}^{(2)}\mathbf{x} \\ \mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix}$$

$$92 \quad = \mathbf{1}\mu + \mathbf{x}\beta - \mathbf{1}\mu - \mathbf{1}\mathbf{C}_{1u}^{(2)}\mathbf{x}\beta - \mathbf{C}_{uu}^{(2)}\mathbf{x}\beta \\ = (\mathbf{I} - \mathbf{1}\mathbf{C}_{1u}^{(2)} - \mathbf{C}_{uu}^{(2)})\mathbf{x}\beta$$

93 So

$$94 \quad E(\mathbf{e}^{(2)}) = (\mathbf{I} - \mathbf{1}\mathbf{C}_{1u}^{(2)} - \mathbf{C}_{uu}^{(2)})\mathbf{x}\beta$$

95 So that in the expression of $E(\hat{\mathbf{e}}^{(2b)'}\hat{\mathbf{e}}^{(2b)})$, the term with $E(\mathbf{e}^{(2)})$ in $E(\hat{\mathbf{e}}^{(2b)'}\hat{\mathbf{e}}^{(2b)})$ was:

$$96 \quad E(\hat{\mathbf{e}}^{(2)})'(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')E(\hat{\mathbf{e}}^{(2)})$$

$$= \mathbf{x}'(\mathbf{I} - \mathbf{C}_{u1}^{(2)}\mathbf{1}' - \mathbf{C}_{uu}^{(2)})(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')(\mathbf{I} - \mathbf{1}\mathbf{C}_{1u}^{(2)} - \mathbf{C}_{uu}^{(2)})\mathbf{x}\beta^2$$

$$97 \quad = \mathbf{x}'(\mathbf{I} - \mathbf{C}_{u1}^{(2)}\mathbf{1}' - \mathbf{C}_{uu}^{(2)})$$

$$(\mathbf{I} - \mathbf{1}\mathbf{C}_{1u}^{(2)} - \mathbf{C}_{uu}^{(2)} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' + \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{1}\mathbf{C}_{1u}^{(2)} + \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{C}_{uu}^{(2)} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' + \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{1}\mathbf{C}_{1u}^{(2)} + \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{C}_{uu}^{(2)})\mathbf{x}\beta^2$$

98 knowing $\mathbf{1}\mathbf{C}_{1u}^{(2)} = \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{1}\mathbf{C}_{1u}^{(2)}$ and $\mathbf{x}'\mathbf{1} = 0$ and $\mathbf{1}'\mathbf{x} = 0$

$$\begin{aligned}
& - \hat{\mathbf{x}}' \mathbf{C}_{uu}^{-1} \hat{\mathbf{x}} \\
99 \quad & (\mathbf{I} - \mathbf{C}_{uu}^{(2)} - (\mathbf{x}'\mathbf{x})' \mathbf{x}\mathbf{x}' + \mathbf{x}(\mathbf{x}'\mathbf{x})' \mathbf{x}' \mathbf{C}_{uu}^{(2)} + \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}' \mathbf{C}_{uu}^{(2)}) \mathbf{x} \beta^2 \\
& = \mathbf{x}' (\mathbf{I} - \mathbf{C}_{uu}^{(2)} - (\mathbf{x}'\mathbf{x})' \mathbf{x}\mathbf{x}' + \mathbf{x}(\mathbf{x}'\mathbf{x})' \mathbf{x}' \mathbf{C}_{uu}^{(2)} \\
100 \quad & + \mathbf{C}_{uu}^{(2)} \mathbf{1}' \mathbf{C}_{uu}^{(2)} - \mathbf{C}_{uu}^{(2)} \mathbf{1}' \mathbf{1} (\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}' \mathbf{C}_{uu}^{(2)} \\
& - \mathbf{C}_{uu}^{(2)} + \mathbf{C}_{uu}^{(2)} \mathbf{C}_{uu}^{(2)} + \mathbf{C}_{uu}^{(2)} (\mathbf{x}'\mathbf{x})' \mathbf{x}\mathbf{x}' - \mathbf{C}_{uu}^{(2)} \mathbf{x}(\mathbf{x}'\mathbf{x})' \mathbf{x}' \mathbf{C}_{uu}^{(2)} - \mathbf{C}_{uu}^{(2)} \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}' \mathbf{C}_{uu}^{(2)}) \mathbf{x} \beta^2 \\
101 \quad & = \mathbf{x}' (\mathbf{I} - 2\mathbf{C}_{uu}^{(2)} - (\mathbf{x}'\mathbf{x})' \mathbf{x}\mathbf{x}' + 2\mathbf{x}(\mathbf{x}'\mathbf{x})' \mathbf{x}' \mathbf{C}_{uu}^{(2)} \\
& + \mathbf{C}_{uu}^{(2)} \mathbf{C}_{uu}^{(2)} - \mathbf{C}_{uu}^{(2)} \mathbf{x}(\mathbf{x}'\mathbf{x})' \mathbf{x}' \mathbf{C}_{uu}^{(2)} - \mathbf{C}_{uu}^{(2)} \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}' \mathbf{C}_{uu}^{(2)}) \mathbf{x} \beta^2 \\
& = \mathbf{x}' \mathbf{x} - 2\mathbf{x}' \mathbf{C}_{uu}^{(2)} \mathbf{x} - \mathbf{x}' (\mathbf{x}'\mathbf{x})' \mathbf{x}\mathbf{x}' \mathbf{x} + 2\mathbf{x}' \mathbf{x}(\mathbf{x}'\mathbf{x})' \mathbf{x}' \mathbf{C}_{uu}^{(2)} \mathbf{x} \\
102 \quad & + \mathbf{x}' (\mathbf{C}_{uu}^{(2)} \mathbf{C}_{uu}^{(2)} - \mathbf{C}_{uu}^{(2)} \mathbf{x}(\mathbf{x}'\mathbf{x})' \mathbf{x}' \mathbf{C}_{uu}^{(2)} - \mathbf{C}_{uu}^{(2)} \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}' \mathbf{C}_{uu}^{(2)}) \mathbf{x} \beta^2 \\
& = \mathbf{x}' (\mathbf{C}_{uu}^{(2)} \mathbf{C}_{uu}^{(2)} - \mathbf{C}_{uu}^{(2)} \mathbf{x}(\mathbf{x}'\mathbf{x})' \mathbf{x}' \mathbf{C}_{uu}^{(2)} - \mathbf{C}_{uu}^{(2)} \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}' \mathbf{C}_{uu}^{(2)}) \mathbf{x} \beta^2 \\
103 \quad & = \mathbf{x}' (\mathbf{C}_{uu}^{(2)} \mathbf{C}_{uu}^{(2)} - \mathbf{C}_{uu}^{(2)} \mathbf{x}(\mathbf{x}'\mathbf{x})' \mathbf{x}' \mathbf{C}_{uu}^{(2)} - \mathbf{C}_{uu}^{(2)} \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}' \mathbf{C}_{uu}^{(2)}) \mathbf{x} \beta^2 \\
& = \beta^2 \mathbf{x}' \mathbf{C}_{uu}^{(2)} (\mathbf{I} - (\mathbf{x}'\mathbf{x})' \mathbf{x}\mathbf{x}' - (\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}\mathbf{1}') \mathbf{C}_{uu}^{(2)} \mathbf{x}
\end{aligned}$$

104 TOTAL OF $E(\hat{\mathbf{e}}^{(2b)}, \hat{\mathbf{e}}^{(2b)})$

$$\begin{aligned}
105 \quad & E(\hat{\mathbf{e}}^{(2b)}, \hat{\mathbf{e}}^{(2b)}) = \sigma_e^2 (n - 2 - \text{tr}(\mathbf{C}_{uu}^{(2)})) + (\mathbf{x}'\mathbf{x})' \mathbf{x}' \mathbf{C}_{uu}^{(2)} \mathbf{x} + (\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}' \mathbf{C}_{uu}^{(2)} \mathbf{1} \\
& - (\sigma_e^2 - \lambda^{(2)} \sigma_u^2) \lambda^{(2)} (\text{tr}(\mathbf{C}_{uu}^{(2)} \mathbf{A}^{-1} \mathbf{C}_{uu}^{(2)}) - (\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}' \mathbf{C}_{uu}^{(2)} \mathbf{A}^{-1} \mathbf{C}_{uu}^{(2)} \mathbf{1} - (\mathbf{x}'\mathbf{x})' \mathbf{x}' \mathbf{C}_{uu}^{(2)} \mathbf{A}^{-1} \mathbf{C}_{uu}^{(2)} \mathbf{x}) \\
& + \beta^2 \mathbf{x}' \mathbf{C}_{uu}^{(2)} (\mathbf{I} - (\mathbf{x}'\mathbf{x})' \mathbf{x}\mathbf{x}' - (\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}\mathbf{1}') \mathbf{C}_{uu}^{(2)} \mathbf{x}
\end{aligned}$$

106 EXPRESSION OF TERMS WITH \mathbf{A}^{-1} :

107 The traduction of:

$$108 \quad \lambda^{(2)} (\text{tr}(\mathbf{C}_{uu}^{(2)} \mathbf{A}^{-1} \mathbf{C}_{uu}^{(2)}) - (\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}' \mathbf{C}_{uu}^{(2)} \mathbf{A}^{-1} \mathbf{C}_{uu}^{(2)} \mathbf{1} - (\mathbf{x}'\mathbf{x})' \mathbf{x}' \mathbf{C}_{uu}^{(2)} \mathbf{A}^{-1} \mathbf{C}_{uu}^{(2)} \mathbf{x})$$

109 We know that

$$\begin{aligned}
& \mathbf{C}_{uu}^{(2)} \mathbf{1}' + \mathbf{C}_{uu}^{(2)} = \mathbf{I} - \mathbf{C}_{uu}^{(2)} \lambda^{(2)} \mathbf{A}^{-1} \\
& \mathbf{C}_{uu}^{(2)} \lambda^{(2)} \mathbf{A}^{-1} = \mathbf{I} - \mathbf{C}_{uu}^{(2)} \mathbf{1}' - \mathbf{C}_{uu}^{(2)} \\
110 \quad & \mathbf{C}_{uu}^{(2)} \lambda^{(2)} \mathbf{A}^{-1} = \mathbf{I} + \frac{1}{n} \mathbf{C}_{uu}^{(2)} \mathbf{1}\mathbf{1}' - \mathbf{C}_{uu}^{(2)} \\
& \mathbf{C}_{uu}^{(2)} \lambda^{(2)} \mathbf{A}^{-1} = \mathbf{I} - \mathbf{C}_{uu}^{(2)} \left(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}' \right)
\end{aligned}$$

111 So:

121 CALCULATION OF $V(\hat{\beta}^{(3)})$

$$\begin{aligned}
V(\hat{\beta}) &= V(\mathbf{C}_{\beta 1} \mathbf{1}' \mathbf{y} + \mathbf{C}_{\beta\beta} \mathbf{x}' \mathbf{y} + \mathbf{C}_{\beta u} \mathbf{y}) \\
&= (\mathbf{C}_{\beta 1} \mathbf{1}' + \mathbf{C}_{\beta\beta} \mathbf{x}' + \mathbf{C}_{\beta u}) (\mathbf{A} \sigma_u^2 + \mathbf{I} \sigma_e^2) (\mathbf{I} \mathbf{C}_{1\beta} + \mathbf{x} \mathbf{C}_{\beta\beta} + \mathbf{C}_{u\beta}) \\
122 \quad &= \sigma_u^2 \left(\begin{bmatrix} \mathbf{C}_{\beta 1} & \mathbf{C}_{\beta\beta} & \mathbf{C}_{\beta u} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{x}' \\ \mathbf{I} \end{bmatrix} \mathbf{A} \begin{bmatrix} \mathbf{1} & \mathbf{x} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{1\beta} \\ \mathbf{C}_{\beta\beta} \\ \mathbf{C}_{u\beta} \end{bmatrix} + \frac{\sigma_e^2}{\sigma_u^2} \begin{bmatrix} \mathbf{C}_{\beta 1} & \mathbf{C}_{\beta\beta} & \mathbf{C}_{\beta u} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{x}' \\ \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{1} & \mathbf{x} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{1\beta} \\ \mathbf{C}_{\beta\beta} \\ \mathbf{C}_{u\beta} \end{bmatrix} \right) \\
&= \sigma_u^2 \left(\begin{bmatrix} -\lambda^{(2)} \mathbf{C}_{\beta u} \mathbf{A}^{-1} \\ \mathbf{A} [-\lambda^{(2)} \mathbf{A}^{-1} \mathbf{C}_{u\beta}] \end{bmatrix} + \frac{\sigma_e^2}{\sigma_u^2} \begin{bmatrix} \mathbf{C}_{\beta 1} & \mathbf{C}_{\beta\beta} & \mathbf{C}_{\beta u} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \mathbf{1} & \mathbf{1}' \mathbf{x} & \mathbf{1}' \mathbf{I} \\ \mathbf{x}' \mathbf{1} & \mathbf{x}' \mathbf{x} & \mathbf{x}' \mathbf{I} \\ \mathbf{1} & \mathbf{x} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{1\beta} \\ \mathbf{C}_{\beta\beta} \\ \mathbf{C}_{u\beta} \end{bmatrix} \right)
\end{aligned}$$

$$123 \quad V(\hat{\beta}) = \sigma_u^2 \left(\lambda^{(2a)} \lambda^{(2a)} \mathbf{C}_{\beta u} \mathbf{A}^{-1} \mathbf{C}_{u\beta} + \frac{\sigma_e^2}{\sigma_u^2} (\mathbf{C}_{\beta\beta} - \lambda^{(2a)} \mathbf{C}_{\beta u} \mathbf{A}^{-1} \mathbf{C}_{u\beta}) \right)$$

$$V(\hat{\beta}^{(3)}) = \sigma_e^2 \mathbf{C}_{\beta\beta}^{(3)} - (\sigma_e^2 - \lambda^{(2a)} \sigma_u^2) \lambda^{(2a)} \mathbf{C}_{\beta u}^{(3)} \mathbf{A}^{-1} \mathbf{C}_{u\beta}^{(3)}$$

$$124 \quad \text{And with : } \mathbf{C}_{\beta u}^{(3)} = -\mathbf{C}_{\beta\beta}^{(3)} \mathbf{x}' \mathbf{C}_{uu}^{(2a)}, \quad \mathbf{C}_{\beta\beta}^{(3)} = [\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x}]^{-1} \quad \text{and} \quad \lambda^{(2a)} \mathbf{C}_{uu}^{(2a)} \mathbf{A}^{-1} = \mathbf{I} - \mathbf{C}_{u1}^{(2a)} \mathbf{1}' - \mathbf{C}_{uu}^{(2a)}$$

$$125 \quad V(\hat{\beta}^{(3)}) = \sigma_e^2 \mathbf{C}_{\beta\beta}^{(3)} - (\sigma_e^2 - \lambda^{(2a)} \sigma_u^2) \lambda^{(2a)} \mathbf{C}_{\beta\beta}^{(3)} \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{A}^{-1} \mathbf{C}_{uu}^{(2a)} \mathbf{x} \mathbf{C}_{\beta\beta}^{(3)}$$

$$126 \quad V(\hat{\beta}^{(3)}) = \sigma_e^2 \mathbf{C}_{\beta\beta}^{(3)} - (\sigma_e^2 - \lambda^{(2a)} \sigma_u^2) \lambda^{(2a)} [\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x}]^{-1} \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{A}^{-1} \mathbf{C}_{uu}^{(2a)} \mathbf{x} [\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x}]^{-1}$$

$$127 \quad V(\hat{\beta}^{(3)}) = \sigma_e^2 \mathbf{C}_{\beta\beta}^{(3)} - (\sigma_e^2 - \lambda^{(2a)} \sigma_u^2) [\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x}]^{-2} \mathbf{x}' (\mathbf{I} - \mathbf{C}_{u1}^{(2a)} \mathbf{1}' - \mathbf{C}_{uu}^{(2a)}) \mathbf{C}_{uu}^{(2a)} \mathbf{x}$$

$$128 \quad V(\hat{\beta}^{(3)}) = \sigma_e^2 [\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x}]^{-1} - (\sigma_e^2 - \lambda^{(2a)} \sigma_u^2) [\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x}]^{-2} (\mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{C}_{uu}^{(2a)} \mathbf{x})$$

129 CALCULATION OF $E(\hat{\epsilon}^{(3)} \mathbf{y})$

$$130 \quad \hat{\epsilon}^{(3)} = \left(\mathbf{I} - \begin{bmatrix} \mathbf{1} & \mathbf{x} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \mathbf{1} & \mathbf{1}' \mathbf{x} & \mathbf{1}' \\ \mathbf{x}' \mathbf{1} & \mathbf{x}' \mathbf{x} & \mathbf{x}' \\ \mathbf{1} & \mathbf{x} & \mathbf{I} + \lambda^{(2a)} \mathbf{A}^{-1} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{1}' \\ \mathbf{x}' \\ \mathbf{I} \end{bmatrix} \mathbf{y}$$

$$\begin{aligned}
131 \quad E(\hat{\boldsymbol{\epsilon}}^{(3)'} \mathbf{y}) &= \text{tr} \left(\mathbf{I} - \begin{bmatrix} \mathbf{1} & \mathbf{x} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{x} & \mathbf{1}' \\ \mathbf{x}'\mathbf{1} & \mathbf{x}'\mathbf{x} & \mathbf{x}' \\ \mathbf{1} & \mathbf{x} & \mathbf{I} + \lambda^{(2a)}\mathbf{A}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}' \\ \mathbf{x}' \\ \mathbf{I} \end{bmatrix} \right) V(\mathbf{y}) \\
&+ E(\mathbf{y})' \left(\mathbf{I} - \begin{bmatrix} \mathbf{1} & \mathbf{x} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{x} & \mathbf{1}' \\ \mathbf{x}'\mathbf{1} & \mathbf{x}'\mathbf{x} & \mathbf{x}' \\ \mathbf{1} & \mathbf{x} & \mathbf{I} + \lambda^{(2a)}\mathbf{A}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}' \\ \mathbf{x}' \\ \mathbf{I} \end{bmatrix} \right) E(\mathbf{y})
\end{aligned}$$

132 PART WITH THE TRACE:

$$\begin{aligned}
133 \quad &= \sigma_u^2 \text{tr} \left(\mathbf{I} - \begin{bmatrix} \mathbf{1} & \mathbf{x} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{x} & \mathbf{1}' \\ \mathbf{x}'\mathbf{1} & \mathbf{x}'\mathbf{x} & \mathbf{x}' \\ \mathbf{1} & \mathbf{x} & \mathbf{I} + \lambda^{(2a)}\mathbf{A}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}' \\ \mathbf{x}' \\ \mathbf{I} \end{bmatrix} \right) \mathbf{A} \\
&+ \sigma_e^2 \text{tr} \left(\mathbf{I} - \begin{bmatrix} \mathbf{1} & \mathbf{x} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{x} & \mathbf{1}' \\ \mathbf{x}'\mathbf{1} & \mathbf{x}'\mathbf{x} & \mathbf{x}' \\ \mathbf{1} & \mathbf{x} & \mathbf{I} + \lambda^{(2a)}\mathbf{A}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}' \\ \mathbf{x}' \\ \mathbf{I} \end{bmatrix} \right) \\
134 \quad &= \sigma_u^2 \text{tr} \left(\mathbf{I} - \begin{bmatrix} -\lambda^{(2a)}\mathbf{A}^{-1}\mathbf{C}_{u1}^{(3)} & -\lambda^{(2a)}\mathbf{A}^{-1}\mathbf{C}_{u\beta}^{(3)} & \mathbf{I} - \lambda^{(2a)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(3)} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{x}' \\ \mathbf{I} \end{bmatrix} \right) \mathbf{A} \\
&+ \sigma_e^2 \left(n - \text{tr} \left(\begin{bmatrix} 1 & 0 & \mathbf{0} \\ 0 & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} - \lambda^{(2a)}\mathbf{C}_{uu}^{(3)}\mathbf{A}^{-1} \end{bmatrix} \right) \right) \\
135 \quad &= \sigma_u^2 \text{tr} \left(\mathbf{A} + \lambda^{(2a)}\mathbf{C}_{u1}^{(3)}\mathbf{1}' + \lambda^{(2a)}\mathbf{C}_{u\beta}^{(3)}\mathbf{x}' - \mathbf{A} + \lambda^{(2a)}\mathbf{C}_{uu}^{(3)} \right) + \sigma_e^2 \left(\text{tr}(\lambda^{(2a)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(3)}) - 2 \right) \\
136 \quad &= \sigma_u^2 \text{tr} \left(\lambda^{(2a)}(\mathbf{I} - \lambda^{(2a)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(3)}) \right) + \sigma_e^2 \left(\text{tr}(\lambda^{(2a)}\mathbf{A}^{-1}\mathbf{C}_{uu}^{(3)}) - 2 \right) \\
137 \quad &= n\lambda^{(2a)}\sigma_u^2 - 2\sigma_e^2 + (\sigma_e^2 - \lambda^{(2a)}\sigma_u^2)\lambda^{(2a)}\text{tr}(\mathbf{A}^{-1}\mathbf{C}_{uu}^{(3)})
\end{aligned}$$

138 PART WITH EXPECTATION:

$$\begin{aligned}
 & E(\mathbf{y})' \left(\mathbf{I} - [\mathbf{1} \quad \mathbf{x} \quad \mathbf{I}] \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{x} & \mathbf{1}' \\ \mathbf{x}'\mathbf{1} & \mathbf{x}'\mathbf{x} & \mathbf{x}' \\ \mathbf{1} & \mathbf{x} & \mathbf{I} + \lambda^{(2a)}\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{x}' \\ \mathbf{I} \end{bmatrix} \right) E(\mathbf{y}) \\
 &= [\mu \quad \beta] \begin{bmatrix} \mathbf{1}' \\ \mathbf{x}' \end{bmatrix} \left(\mathbf{I} - [\mathbf{1} \quad \mathbf{x} \quad \mathbf{I}] \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{x} & \mathbf{1}' \\ \mathbf{x}'\mathbf{1} & \mathbf{x}'\mathbf{x} & \mathbf{x}' \\ \mathbf{1} & \mathbf{x} & \mathbf{I} + \lambda^{(2a)}\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{x}' \\ \mathbf{I} \end{bmatrix} \right) [\mathbf{1} \quad \mathbf{x}] \begin{bmatrix} \mu \\ \beta \end{bmatrix} \\
 &= [\mu \quad \beta] \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{x} \\ \mathbf{x}'\mathbf{1} & \mathbf{x}'\mathbf{x} \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} - [\mu \quad \beta] \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{x} & \mathbf{1}' \\ \mathbf{x}'\mathbf{1} & \mathbf{x}'\mathbf{x} & \mathbf{x}' \\ \mathbf{1} & \mathbf{x} & \mathbf{I} + \lambda^{(2a)}\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{x} \\ \mathbf{x}'\mathbf{1} & \mathbf{x}'\mathbf{x} \\ \mathbf{1} & \mathbf{x} \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} \\
 &= [\mu \quad \beta] \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{x} \\ \mathbf{x}'\mathbf{1} & \mathbf{x}'\mathbf{x} \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} - [\mu \quad \beta] \begin{bmatrix} 1 & 0 & \mathbf{0}' \\ 0 & 1 & \mathbf{0}' \\ \mathbf{1} & \mathbf{x} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{x} \\ \mathbf{x}'\mathbf{1} & \mathbf{x}'\mathbf{x} \\ \mathbf{1} & \mathbf{x} \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} \\
 &= [\mu \quad \beta] \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{x} \\ \mathbf{x}'\mathbf{1} & \mathbf{x}'\mathbf{x} \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} - [\mu \quad \beta] \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{x} \\ \mathbf{x}'\mathbf{1} & \mathbf{x}'\mathbf{x} \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} \\
 &= 0
 \end{aligned}$$

140 TOTAL OF $E(\hat{\mathbf{e}}^{(3)'}\mathbf{y})$

$$141 \quad E(\hat{\mathbf{e}}^{(3)'}\mathbf{y}) = n\lambda^{(2a)}\sigma_u^2 - 2\sigma_e^2 + (\sigma_e^2 - \lambda^{(2a)}\sigma_u^2)\lambda^{(2a)}tr(\mathbf{A}^{-1}\mathbf{C}_{uu}^{(3)})$$

142 Expression of $\lambda^{(2a)}tr(\mathbf{A}^{-1}\mathbf{C}_{uu}^{(3)})$ in terms with only \mathbf{A} and $\mathbf{C}_{uu}^{(2)}$:

143 We know that :

$$144 \quad \begin{bmatrix} \mathbf{C}_{\beta\beta}^{(3)} & \mathbf{C}_{\beta 1}^{(3)} & \mathbf{C}_{\beta u}^{(3)} \\ \mathbf{C}_{1\beta}^{(3)} & \mathbf{C}_{11}^{(3)} & \mathbf{C}_{1u}^{(3)} \\ \mathbf{C}_{u\beta}^{(3)} & \mathbf{C}_{u1}^{(3)} & \mathbf{C}_{uu}^{(3)} \end{bmatrix} \begin{bmatrix} \mathbf{x}'\mathbf{x} & \mathbf{x}'\mathbf{1} & \mathbf{x}' \\ \mathbf{1}'\mathbf{x} & \mathbf{1}'\mathbf{1} & \mathbf{1}' \\ \mathbf{x} & \mathbf{1} & \mathbf{I} + \lambda^{(2a)}\mathbf{A}^{-1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \mathbf{0}' \\ 0 & 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix} \text{ so :}$$

$$145 \quad \begin{bmatrix} \mathbf{C}_{11}^{(3)} & \mathbf{C}_{1u}^{(3)} \\ \mathbf{C}_{u1}^{(3)} & \mathbf{C}_{uu}^{(3)} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}' \\ \mathbf{1} & \mathbf{I} + \lambda^{(2a)}\mathbf{A}^{-1} \end{bmatrix}^{-1} + \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}' \\ \mathbf{1} & \mathbf{I} + \lambda^{(2a)}\mathbf{A}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}'\mathbf{x} \\ \mathbf{x} \end{bmatrix} \mathbf{C}_{\beta\beta}^{(3)} \begin{bmatrix} \mathbf{x}'\mathbf{1} & \mathbf{x}' \\ \mathbf{1} & \mathbf{I} + \lambda^{(2a)}\mathbf{A}^{-1} \end{bmatrix}^{-1}$$

146

147 And with the matrices used in GRAMMAR, the equalities were:

$$148 \quad \begin{bmatrix} \mathbf{C}_{11}^{(3)} & \mathbf{C}_{1u}^{(3)} \\ \mathbf{C}_{u1}^{(3)} & \mathbf{C}_{uu}^{(3)} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{11}^{(2a)} & \mathbf{C}_{1u}^{(2a)} \\ \mathbf{C}_{u1}^{(2a)} & \mathbf{C}_{uu}^{(2a)} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_{11}^{(2a)} & \mathbf{C}_{1u}^{(2a)} \\ \mathbf{C}_{u1}^{(2a)} & \mathbf{C}_{uu}^{(2a)} \end{bmatrix} \begin{bmatrix} \mathbf{1}'\mathbf{x} \\ \mathbf{x} \end{bmatrix} \mathbf{C}_{\beta\beta}^{(3)} \begin{bmatrix} \mathbf{x}'\mathbf{1} & \mathbf{x}' \\ \mathbf{1} & \mathbf{I} + \lambda^{(2a)}\mathbf{A}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{C}_{11}^{(2a)} & \mathbf{C}_{1u}^{(2a)} \\ \mathbf{C}_{u1}^{(2a)} & \mathbf{C}_{uu}^{(2a)} \end{bmatrix}$$

149 Soit

$$150 \quad \begin{bmatrix} \mathbf{C}_{11}^{(3)} & \mathbf{C}_{1u}^{(3)} \\ \mathbf{C}_{u1}^{(3)} & \mathbf{C}_{uu}^{(3)} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{11}^{(2a)} & \mathbf{C}_{1u}^{(2a)} \\ \mathbf{C}_{u1}^{(2a)} & \mathbf{C}_{uu}^{(2a)} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_{11}^{(2a)} & \mathbf{C}_{1u}^{(2a)} \\ \mathbf{C}_{u1}^{(2a)} & \mathbf{C}_{uu}^{(2a)} \end{bmatrix} \begin{bmatrix} \mathbf{1} & \mathbf{x} \mathbf{C}_{\beta\beta}^{(3)} \mathbf{x}' & \mathbf{1} & \mathbf{1} \\ \mathbf{x} \mathbf{C}_{\beta\beta}^{(3)} \mathbf{x}' & \mathbf{1} & \mathbf{x} \mathbf{C}_{\beta\beta}^{(3)} \mathbf{x}' & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{11}^{(2a)} & \mathbf{C}_{1u}^{(2a)} \\ \mathbf{C}_{u1}^{(2a)} & \mathbf{C}_{uu}^{(2a)} \end{bmatrix}$$

$$151 \quad \begin{bmatrix} \mathbf{C}_{11}^{(3)} & \mathbf{C}_{1u}^{(3)} \\ \mathbf{C}_{u1}^{(3)} & \mathbf{C}_{uu}^{(3)} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{11}^{(2a)} & \mathbf{C}_{1u}^{(2a)} \\ \mathbf{C}_{u1}^{(2a)} & \mathbf{C}_{uu}^{(2a)} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_{11}^{(2a)} & \mathbf{C}_{1u}^{(2a)} \\ \mathbf{C}_{u1}^{(2a)} & \mathbf{C}_{uu}^{(2a)} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{x} \mathbf{C}_{\beta\beta}^{(3)} \mathbf{x}' \end{bmatrix} \begin{bmatrix} \mathbf{C}_{11}^{(2a)} & \mathbf{C}_{1u}^{(2a)} \\ \mathbf{C}_{u1}^{(2a)} & \mathbf{C}_{uu}^{(2a)} \end{bmatrix}$$

$$152 \quad \begin{bmatrix} \mathbf{C}_{11}^{(3)} & \mathbf{C}_{1u}^{(3)} \\ \mathbf{C}_{u1}^{(3)} & \mathbf{C}_{uu}^{(3)} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{11}^{(2a)} & \mathbf{C}_{1u}^{(2a)} \\ \mathbf{C}_{u1}^{(2a)} & \mathbf{C}_{uu}^{(2a)} \end{bmatrix} + \begin{bmatrix} 0 & \mathbf{C}_{1u}^{(2a)} \mathbf{x} \mathbf{C}_{\beta\beta}^{(3)} \mathbf{x}' \\ 0 & \mathbf{C}_{uu}^{(2a)} \mathbf{x} \mathbf{C}_{\beta\beta}^{(3)} \mathbf{x}' \end{bmatrix} \begin{bmatrix} \mathbf{C}_{11}^{(2a)} & \mathbf{C}_{1u}^{(2a)} \\ \mathbf{C}_{u1}^{(2a)} & \mathbf{C}_{uu}^{(2a)} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{C}_{11}^{(3)} & \mathbf{C}_{1u}^{(3)} \\ \mathbf{C}_{u1}^{(3)} & \mathbf{C}_{uu}^{(3)} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{11}^{(2a)} & \mathbf{C}_{1u}^{(2a)} \\ \mathbf{C}_{u1}^{(2a)} & \mathbf{C}_{uu}^{(2a)} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_{1u}^{(2a)} \mathbf{x} \mathbf{C}_{\beta\beta}^{(3)} \mathbf{x}' \mathbf{C}_{u1}^{(2a)} & \mathbf{C}_{1u}^{(2a)} \mathbf{x} \mathbf{C}_{\beta\beta}^{(3)} \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \\ \mathbf{C}_{uu}^{(2a)} \mathbf{x} \mathbf{C}_{\beta\beta}^{(3)} \mathbf{x}' \mathbf{C}_{u1}^{(2a)} & \mathbf{C}_{uu}^{(2a)} \mathbf{x} \mathbf{C}_{\beta\beta}^{(3)} \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \end{bmatrix}$$

153 So

$$154 \quad \mathbf{C}_{uu}^{(3)} = \mathbf{C}_{uu}^{(2a)} + \mathbf{C}_{uu}^{(2a)} \mathbf{x} \mathbf{C}_{\beta\beta}^{(3)} \mathbf{x}' \mathbf{C}_{uu}^{(2a)}$$

155 And knowing:

$$156 \quad \begin{aligned} \mathbf{C}_{\beta\beta}^{(3)} &= \left(\mathbf{x}' \mathbf{x} - \begin{bmatrix} \mathbf{x}' \mathbf{1} & \mathbf{x}' \end{bmatrix} \begin{bmatrix} \mathbf{C}_{11}^{(2a)} & \mathbf{C}_{1u}^{(2a)} \\ \mathbf{C}_{u1}^{(2a)} & \mathbf{C}_{uu}^{(2a)} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \mathbf{x} \\ \mathbf{x} \end{bmatrix} \right)^{-1} \\ &= \left(\mathbf{x}' \mathbf{x} - \begin{bmatrix} 0 & \mathbf{x}' \end{bmatrix} \begin{bmatrix} \mathbf{C}_{11}^{(2a)} & \mathbf{C}_{1u}^{(2a)} \\ \mathbf{C}_{u1}^{(2a)} & \mathbf{C}_{uu}^{(2a)} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{x} \end{bmatrix} \right)^{-1} \\ &= \left(\mathbf{x}' \mathbf{x} - \begin{bmatrix} \mathbf{x}' \mathbf{C}_{u1}^{(2a)} & \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{x} \end{bmatrix} \right)^{-1} \\ &= \left(\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} \right)^{-1} \end{aligned}$$

157 Replacing $\mathbf{C}_{\beta\beta}^{(3)}$:

$$158 \quad \begin{aligned} \mathbf{C}_{uu}^{(3)} &= \mathbf{C}_{uu}^{(2a)} + \mathbf{C}_{uu}^{(2a)} \mathbf{x} \mathbf{C}_{\beta\beta}^{(3)} \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \\ \mathbf{C}_{uu}^{(3)} &= \mathbf{C}_{uu}^{(2a)} + \mathbf{C}_{uu}^{(2a)} \mathbf{x} \left[\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} \right]^{-1} \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \\ \mathbf{C}_{uu}^{(3)} &= \mathbf{C}_{uu}^{(2a)} + \left[\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} \right]^{-1} \mathbf{C}_{uu}^{(2a)} \mathbf{x} \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \end{aligned}$$

159 So that, for the function of the trace: $\lambda^{(2a)} tr(\mathbf{A}^{-1} \mathbf{C}_{uu}^{(3)})$

$$\begin{aligned}
& \lambda^{(2a)} \text{tr}(\mathbf{A}^{-1} \mathbf{C}_{uu}^{(3)}) = \lambda^{(2a)} \text{tr}(\mathbf{A}^{-1} (\mathbf{C}_{uu}^{(2a)} + [\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x}]^{-1} \mathbf{C}_{uu}^{(2a)} \mathbf{x} \mathbf{x}' \mathbf{C}_{uu}^{(2a)})) \\
& = \lambda^{(2a)} \text{tr}(\mathbf{A}^{-1} \mathbf{C}_{uu}^{(2a)}) + \lambda^{(2a)} \text{tr}(\mathbf{A}^{-1} \mathbf{C}_{uu}^{(2a)} \mathbf{x} \mathbf{x}' \mathbf{C}_{uu}^{(2a)}) [\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x}]^{-1} \\
160 \quad & = \text{tr}(\mathbf{I} - \mathbf{C}_{uu}^{(2a)} \left(\mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}' \right)) + \mathbf{x}' \left(\mathbf{I} - \mathbf{C}_{uu}^{(2a)} \left(\mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}' \right) \right) \mathbf{C}_{uu}^{(2a)} \mathbf{x} [\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x}]^{-1} \\
& = n - \text{tr}(\mathbf{C}_{uu}^{(2a)}) + \frac{1}{n} \mathbf{1}' \mathbf{C}_{uu}^{(2a)} \mathbf{1} + \frac{\mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{C}_{uu}^{(2a)} \mathbf{x} + \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{1} \mathbf{1}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} / n}{\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x}}
\end{aligned}$$

161 And the expression of $E(\hat{\boldsymbol{\epsilon}}^{(3)'} \mathbf{y})$ was

$$\begin{aligned}
& E(\hat{\boldsymbol{\epsilon}}^{(3)'} \mathbf{y}) = n \sigma_u^2 \lambda^{(2a)} - 2 \sigma_e^2 + \\
162 \quad & (\sigma_e^2 - \sigma_u^2 \lambda^{(2a)}) \left(n + \frac{1}{n} \mathbf{1}' \mathbf{C}_{uu}^{(2a)} \mathbf{1} - \text{tr}(\mathbf{C}_{uu}^{(2a)}) + \frac{\left(\mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} + \frac{1}{n} \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{1} \mathbf{1}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{C}_{uu}^{(2a)} \mathbf{x} \right)}{\left(\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} \right)} \right)
\end{aligned}$$

163 which was equal to

$$\begin{aligned}
& E(\hat{\boldsymbol{\epsilon}}^{(3)'} \mathbf{y}) = (n-2) \sigma_e^2 + \\
164 \quad & (\sigma_e^2 - \sigma_u^2 \lambda^{(2a)}) \left(\frac{1}{n} \mathbf{1}' \mathbf{C}_{uu}^{(2a)} \mathbf{1} - \text{tr}(\mathbf{C}_{uu}^{(2a)}) + \frac{\left(\mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} + \frac{1}{n} \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{1} \mathbf{1}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{C}_{uu}^{(2a)} \mathbf{x} \right)}{\left(\mathbf{x}' \mathbf{x} - \mathbf{x}' \mathbf{C}_{uu}^{(2a)} \mathbf{x} \right)} \right)
\end{aligned}$$

165

166 MODEL 4, QTDT

167 For the QTDT method, lot of simplifications happened when replacing $\mathbf{Q}'\mathbf{Q}$ by its expectation
168 which was a diagonal matrix, easy to invert. So that

$$\begin{aligned}
169 \quad & E(\hat{\beta}_w^{(4)}) = \frac{1}{\text{tr}(\mathbf{D})} (\text{tr}(\mathbf{D}) \frac{n-1}{n}) \beta = \frac{n-1}{n} \beta \\
& V(\hat{\beta}_w^{(4)}) = [(\mathbf{Q}'\mathbf{Q})^{-1} \mathbf{Q}' \mathbf{A} \mathbf{Q} (\mathbf{Q}'\mathbf{Q})^{-1}]_{3,3} \sigma_u^2 + [(\mathbf{Q}'\mathbf{Q})^{-1}]_{3,3} \sigma_e^2 \\
170 \quad & = \frac{1}{\text{tr}(\mathbf{D})} \text{tr}(\mathbf{A} \mathbf{D}) \frac{1}{\text{tr}(\mathbf{D})} \sigma_u^2 + \frac{1}{\text{tr}(\mathbf{D})} \sigma_e^2
\end{aligned}$$

171 And without inbreeding, $\text{tr}(\mathbf{A} \mathbf{D}) = \text{tr}(\mathbf{D}) = \frac{n}{2}$ so that:

$$172 \quad V(\hat{\beta}_w^{(4)}) = \frac{2}{n} (\sigma_u^2 + \sigma_e^2)$$

173 When replacing $\mathbf{Q}'\mathbf{Q}$ by its expectation, the sum of square of residuals were:

$$174 \quad E(\hat{\mathbf{e}}^{(4)'}\hat{\mathbf{e}}^{(4)}) = (n-3)\sigma_e^2 + \left(tr(\mathbf{A}) - \frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{1} - \frac{tr(\mathbf{A}'\mathbf{A}) - \frac{2}{n}\mathbf{1}'\mathbf{A}'\mathbf{A}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1} - \frac{n}{2} + \frac{1}{2n}\mathbf{1}'\mathbf{A}\mathbf{1}}{tr(\mathbf{A}) - \frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{1} - \frac{n-1}{2}} - 1 \right) \sigma_u^2 + \beta^2 \frac{(n-1)}{2n}$$

$$175 \quad E(\tau^{(4)}) = \sqrt{tr(\mathbf{D})} \frac{n-1}{n} \frac{\sqrt{(n-3)}}{\sqrt{E(\hat{\mathbf{e}}^{(4)'}\hat{\mathbf{e}}^{(4)})}} \beta$$

$$176 \quad V(\tau^{(4)}) = \left(\frac{tr(\mathbf{AD})}{tr(\mathbf{D})} \sigma_u^2 + \sigma_e^2 \right) \frac{(n-3)}{E(\hat{\mathbf{e}}^{(4)'}\hat{\mathbf{e}}^{(4)})}$$

177 **MARGINAL EXPECTATION OF QUADRATIC FORMS ACCORDING TO**
178 **DISTRIBUTION OF GENOTYPES**

179 We will need $V(\mathbf{x})$

180 The coefficients of the matrix $V(\mathbf{x})$ were:

$$181 \quad E((w_i - \bar{w})(w_j - \bar{w})) = E(w_i w_j) - \frac{1}{n} \sum_{k=1}^n E(w_i w_k) - \frac{1}{n} \sum_{k=1}^n E(w_i w_k) + \frac{1}{n^2} \mathbf{1}'\mathbf{A}\mathbf{1}$$

$$182 \quad = a_{ij} - \frac{1}{n} \sum_{k=1}^n a_{ik} - \frac{1}{n} \sum_{k=1}^n a_{jk} + \frac{1}{n^2} \mathbf{1}'\mathbf{A}\mathbf{1}$$

183 So that:

$$184 \quad V(\mathbf{x}) = \mathbf{A} - \frac{1}{n} \mathbf{1}\mathbf{1}'\mathbf{A} - \frac{1}{n} \mathbf{A}\mathbf{1}\mathbf{1}' + \frac{1}{n^2} \mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'$$

185 CALCULATION OF $E_x(\mathbf{x}'\mathbf{x})$

$$E_x(\mathbf{x}'\mathbf{x}) = tr(V(\mathbf{x})) + E(\mathbf{x}')E(\mathbf{x}') \\ = tr\left(\mathbf{A} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{A} - \frac{1}{n}\mathbf{A}\mathbf{1}\mathbf{1}' + \frac{1}{n^2}\mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\right)$$

$$186 \quad = tr(\mathbf{A}) - \frac{2}{n}\mathbf{1}'\mathbf{A}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{1} \\ = tr(\mathbf{A}) - \frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{1}$$

187 CALCULATION OF $E_x(\mathbf{x}'\mathbf{A}\mathbf{x})$

$$E_x(\mathbf{x}'\mathbf{A}\mathbf{x}) = tr(\mathbf{A}V(\mathbf{x})) + E(\mathbf{x}')\mathbf{A}E(\mathbf{x})$$

$$188 \quad = tr\left(\mathbf{A}\left(\mathbf{A} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{A} - \frac{1}{n}\mathbf{A}\mathbf{1}\mathbf{1}' + \frac{1}{n^2}\mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\right)\right) \\ = tr(\mathbf{A}\mathbf{A}) - \frac{2}{n}\mathbf{1}'\mathbf{A}\mathbf{A}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1}$$

189 Or, for an easier interpretation:

$$190 \quad E_x(\mathbf{x}'\mathbf{A}\mathbf{x}) = tr(\mathbf{A}\mathbf{A}) - \frac{2}{n}\mathbf{1}'\mathbf{A}\mathbf{A}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1} \\ = tr(\mathbf{A}\mathbf{A}) - \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1} - 2\left(\frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{A}\mathbf{1} - \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1}\right)$$

191 Because $tr(\mathbf{A}\mathbf{A}) - \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1}$ was equal to the variance of the coefficients of the matrix \mathbf{A}

192 multiplied by n^2 and $\left(\frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{A}\mathbf{1} - \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1}\right)$ was the variance of the sum of columns of \mathbf{A}

193 CALCULATION OF $E_x(\mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{x})$

$$\begin{aligned}
 E_x(\mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{x}) &= tr(\mathbf{C}_{uu}^{(2)}V(\mathbf{x})) = \\
 &tr(\mathbf{C}_{uu}^{(2)}(\mathbf{A} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{A} - \frac{1}{n}\mathbf{A}\mathbf{1}\mathbf{1}' + \frac{1}{n^2}\mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}')) \\
 &= tr(\mathbf{C}_{uu}^{(2)}\mathbf{A}) - \frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2)}\mathbf{1} - \frac{1}{n}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{A}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1} \\
 194 &= tr(\mathbf{A}) - \lambda^{(2)}tr(\mathbf{C}_{uu}^{(2)}) - tr(\mathbf{C}_{u1}^{(2)}\mathbf{1}'\mathbf{A}) - \frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2)}\mathbf{1} - \frac{1}{n}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{A}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1} \\
 &= tr(\mathbf{A}) - \lambda^{(2)}tr(\mathbf{C}_{uu}^{(2)}) + \frac{1}{n}tr(\mathbf{C}_{uu}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{A}) - \frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2)}\mathbf{1} - \frac{1}{n}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{A}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1} \\
 &= tr(\mathbf{A}) - \lambda^{(2)}tr(\mathbf{C}_{uu}^{(2)}) - \frac{1}{n}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{A}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1} \\
 &= tr(\mathbf{A}) - \lambda^{(2)}tr(\mathbf{C}_{uu}^{(2)}) - (\frac{1}{n}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{A}\mathbf{1} - \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1})
 \end{aligned}$$

195 Using the permutation property of the trace in order to obtain a scalar

196 CALCULATION OF $E_x(\mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{x})$

$$\begin{aligned}
 E_x(\mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{x}) &= tr(\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}V(\mathbf{x})) = \\
 &tr(\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}(\mathbf{A} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{A} - \frac{1}{n}\mathbf{A}\mathbf{1}\mathbf{1}' + \frac{1}{n^2}\mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}')) \\
 &= tr(\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{A}) - \frac{2}{n}\mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{1} \\
 197 &= tr(\mathbf{C}_{uu}^{(2)}(\mathbf{A} - \lambda^{(2)}\mathbf{C}_{uu}^{(2)} - \mathbf{C}_{u1}^{(2)}\mathbf{1}'\mathbf{A})) - \frac{2}{n}\mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{1} \\
 &= tr(\mathbf{A}) - \lambda^{(2)}tr(\mathbf{C}_{uu}^{(2)}) - tr(\mathbf{C}_{u1}^{(2)}\mathbf{1}'\mathbf{A}) - \lambda^{(2)}tr(\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}) - tr(\mathbf{C}_{uu}^{(2)}\mathbf{C}_{u1}^{(2)}\mathbf{1}'\mathbf{A}) - \frac{2}{n}\mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{1} \\
 &= tr(\mathbf{A}) - \lambda^{(2)}tr(\mathbf{C}_{uu}^{(2)}) - \lambda^{(2)}tr(\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}) - \mathbf{1}'\mathbf{A}\mathbf{C}_{u1}^{(2)} - \mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2)}\mathbf{C}_{u1}^{(2)} - \frac{2}{n}\mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{1} \\
 &= tr(\mathbf{A}) - \lambda^{(2)}tr(\mathbf{C}_{uu}^{(2)}) - \lambda^{(2)}tr(\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}) + \frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2)}\mathbf{1} - \frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{1}
 \end{aligned}$$

198 CALCULATION OF $E_x(\mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{x})$

$$\begin{aligned}
E_x(\mathbf{x}'\mathbf{C}_{uu}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{x}) &= \text{tr}(\mathbf{C}_{uu}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{V}(\mathbf{x})) = \\
&= \text{tr}(\mathbf{C}_{uu}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}(\mathbf{A} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{A} - \frac{1}{n}\mathbf{A}\mathbf{1}\mathbf{1}' + \frac{1}{n^2}\mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}')) \\
199 \quad &= \text{tr}(\mathbf{C}_{uu}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{A}) - \frac{2}{n}\mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1} \\
&= \text{tr}(\mathbf{C}_{uu}^{(2)}\mathbf{1}\mathbf{1}'(\mathbf{A} - \lambda^{(2)}\mathbf{C}_{uu}^{(2)} - \mathbf{C}_{u1}^{(2)}\mathbf{1}'\mathbf{A})) - \frac{2}{n}\mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1} \\
&= \mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2)}\mathbf{1} - \lambda^{(2)}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{1} + \frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1} - \frac{2}{n}\mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1} \\
&= \mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2)}\mathbf{1} - \lambda^{(2)}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{C}_{uu}^{(2)}\mathbf{1} - \frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{C}_{uu}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1}\mathbf{1}'\mathbf{C}_{uu}^{(2)}\mathbf{1}
\end{aligned}$$

200 CALCULATION OF $E_x(\mathbf{Q}'\mathbf{Q})$

201 With the notations:

202
$$a_{ii} = \frac{1}{4}a_{s_i s_i} + \frac{1}{4}a_{d_i d_i} + \frac{1}{2}a_{s_i d_i} + d_{ii} = 1 + f_i$$

203 where a_{ij} the relationship coefficient between individual i and j , s_i the sire of the individual i ,
204 d_i the dam of individual i , d_{ii} the Mendelian sampling effect and f_i the inbreeding coefficient
205 of i . We had:

206
$$d_{ii} = \frac{1}{2} - \frac{1}{4}(f_{s_i} + f_{d_i})$$

207 Let note

208
$$\zeta = \frac{\mathbf{w}_s + \mathbf{w}_d}{2}, \text{ so that } \mathbf{z} = \zeta - \mathbf{1}\bar{w}$$

209 We had

210 $\mathbf{x} - \mathbf{z} = \mathbf{w} - \zeta$ and $\mathbf{z} - \mathbf{1}\bar{z} = \zeta - \mathbf{1}\bar{\zeta}$ which were replaced in each coefficient of \mathbf{Q}

211 The following results were useful:

212
$$E(w_i w_i) = a_{ii}$$

213
$$E(w_i w_j) = a_{ij}$$

214
$$E(w_i \zeta_i) = E(w_i \left(\frac{w_{s_i} + w_{d_i}}{2} \right)) = \frac{1}{2}a_{s_i} + \frac{1}{2}a_{d_i} = \frac{1}{4}a_{s_i s_i} + \frac{1}{4}a_{d_i d_i} + \frac{1}{2}a_{s_i d_i} = a_{ii} - d_{ii}$$

$$215 \quad E(w_i \zeta_j) = a_{ij}$$

$$216 \quad E(\zeta_i \zeta_i) = a_{ii} - d_{ii}$$

$$217 \quad E(\zeta_i \zeta_j) = a_{ij}$$

$$E(\bar{\zeta}^2) = \frac{1}{n^2} E\left(\left(\sum_{i=1}^n \zeta_i\right)\left(\sum_{i=1}^n \zeta_i\right)\right) = \frac{1}{n^2} \left[\sum_{i=1}^n E(\zeta_i^2) + \sum_{i=1}^n \sum_{j=1, j \neq i}^n E(\zeta_i \zeta_j) \right] = \frac{1}{n^2} \sum_{i=1}^n (a_{ii} - d_{ii}) + \sum_{i=1}^n \sum_{j=1, j \neq i}^n a_{ij}$$

$$218 \quad = \frac{1}{n^2} [tr(\mathbf{A}) - tr(\mathbf{D}) + \mathbf{1}'\mathbf{A}\mathbf{1} - tr(\mathbf{A})]$$

$$= \frac{1}{n^2} [\mathbf{1}'\mathbf{A}\mathbf{1} - tr(\mathbf{D})]$$

219 with \mathbf{D} the diagonal matrix of Mendelian sampling coefficients d_{ii}

$$220 \quad E(\bar{\zeta} \bar{w}) = \frac{1}{n^2} E\left(\left(\sum_{i=1}^n \zeta_i\right)\left(\sum_{i=1}^n w_i\right)\right) = \frac{1}{n^2} \left[\sum_{i=1}^n E(\zeta_i w_i) + \sum_{i=1}^n \sum_{j=1, j \neq i}^n E(\zeta_i w_j) \right] = \frac{1}{n^2} \sum_{i=1}^n (a_{ii} - d_{ii}) + \sum_{i=1}^n \sum_{j=1, j \neq i}^n a_{ij}$$

$$= \frac{1}{n^2} [\mathbf{1}'\mathbf{A}\mathbf{1} - tr(\mathbf{D})]$$

221 So, the coefficients of the matrix $\mathbf{Q}'\mathbf{Q}$ and $\mathbf{Q}'\mathbf{x}$ were:

$$222 \quad E(\mathbf{1}'(\mathbf{z} - \mathbf{1}\bar{z})) = E(\mathbf{1}'(\zeta - \mathbf{1}\bar{\zeta})) = 0$$

$$223 \quad E(\mathbf{1}'(\mathbf{x} - \mathbf{z})) = E(\mathbf{1}'(\mathbf{w} - \zeta)) = 0$$

$$E((\mathbf{z} - \mathbf{1}\bar{z})'(\mathbf{z} - \mathbf{1}\bar{z})) = E((\zeta - \mathbf{1}\bar{\zeta})'(\zeta - \mathbf{1}\bar{\zeta}))$$

$$= E(\zeta'\zeta) - nE(\bar{\zeta}^2)$$

$$224 \quad = tr(\mathbf{A}) - tr(\mathbf{D}) - \frac{1}{n} [\mathbf{1}'\mathbf{A}\mathbf{1} - tr(\mathbf{D})]$$

$$= tr(\mathbf{A}) - \frac{1}{n} \mathbf{1}'\mathbf{A}\mathbf{1} - tr(\mathbf{D}) \frac{n-1}{n}$$

$$E((\mathbf{z} - \mathbf{1}\bar{z})'(\mathbf{x} - \mathbf{z})) = E((\zeta - \mathbf{1}\bar{\zeta})'(\mathbf{w} - \zeta)) = E(\zeta'\mathbf{w}) - E(\zeta'\zeta) - E(\bar{\zeta}\mathbf{1}'\mathbf{w}) + E(\bar{\zeta}\mathbf{1}'\zeta)$$

$$225 \quad = tr(\mathbf{A}) - tr(\mathbf{D}) - (tr(\mathbf{A}) - tr(\mathbf{D}))$$

$$= 0$$

$$E((\mathbf{x} - \mathbf{z})'(\mathbf{x} - \mathbf{z})) = E((\mathbf{w} - \zeta)'(\mathbf{w} - \zeta)) = E(\mathbf{w}'\mathbf{w}) - E(\mathbf{w}'\zeta) - E(\zeta'\mathbf{w}) + E(\zeta'\zeta)$$

$$226 \quad = tr(\mathbf{A}) - 2(tr(\mathbf{A}) - tr(\mathbf{D})) + (tr(\mathbf{A}) - tr(\mathbf{D}))$$

$$= tr(\mathbf{D})$$

227 CALCULATION OF $E_x(\mathbf{Q}'\mathbf{x})$

228 Using the previous results, the expectations involved were:

$$229 \quad E(\mathbf{1}'(\mathbf{w} - \bar{\mathbf{w}})) = 0$$

$$230 \quad \begin{aligned} E((\mathbf{z} - \bar{\mathbf{z}})'(\mathbf{w} - \bar{\mathbf{w}})) &= E((\boldsymbol{\zeta} - \mathbf{1}\bar{\boldsymbol{\zeta}})'(\mathbf{w} - \bar{\mathbf{w}})) = E(\boldsymbol{\zeta}'\mathbf{w}) - E(\boldsymbol{\zeta}'\mathbf{1}\bar{\mathbf{w}}) - E(\bar{\boldsymbol{\zeta}}'\mathbf{1}'\mathbf{w}) + E(\bar{\boldsymbol{\zeta}}'\bar{\mathbf{w}}\mathbf{1}'\mathbf{1}) \\ &= \text{tr}(\mathbf{A}) - \frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{1} - \text{tr}(\mathbf{D})\frac{n-1}{n} \end{aligned}$$

$$231 \quad \begin{aligned} E((\mathbf{x} - \bar{\mathbf{x}})'(\mathbf{w} - \bar{\mathbf{w}})) &= E((\mathbf{w} - \boldsymbol{\zeta})'(\mathbf{w} - \bar{\mathbf{w}})) = E(\mathbf{w}'\mathbf{w}) - E(\mathbf{w}'\mathbf{1}\bar{\mathbf{w}}) - E(\boldsymbol{\zeta}'\mathbf{w}) + E(\boldsymbol{\zeta}'\mathbf{1}\bar{\mathbf{w}}) \\ &= \text{tr}(\mathbf{A}) - \frac{1}{n}\mathbf{1}'\mathbf{A}\mathbf{1} - (\text{tr}(\mathbf{A}) - \text{tr}(\mathbf{D})) + \frac{1}{n}[\mathbf{1}'\mathbf{A}\mathbf{1} - \text{tr}(\mathbf{D})] \\ &= \text{tr}(\mathbf{D})\frac{n-1}{n} \end{aligned}$$

232 CALCULATION OF $E_x(\mathbf{Q}\mathbf{A}\mathbf{Q}')$

233 Using the previous results, the expectations involved were:

$$234 \quad \begin{aligned} E((\mathbf{z} - \bar{\mathbf{z}})' \mathbf{A}(\mathbf{z} - \bar{\mathbf{z}})) &= E((\boldsymbol{\zeta} - \mathbf{1}\bar{\boldsymbol{\zeta}})' \mathbf{A}(\boldsymbol{\zeta} - \mathbf{1}\bar{\boldsymbol{\zeta}})) = E(\boldsymbol{\zeta}'\mathbf{A}\boldsymbol{\zeta}) - E(\boldsymbol{\zeta}'\mathbf{A}\mathbf{1}\bar{\boldsymbol{\zeta}}) - E(\bar{\boldsymbol{\zeta}}'\mathbf{1}'\mathbf{A}\boldsymbol{\zeta}) + E(\bar{\boldsymbol{\zeta}}'\mathbf{1}'\mathbf{A}\mathbf{1}) \\ &= E\left(\sum_{i=1}^n \sum_{j=1}^n z_i a_{ij} \zeta_j\right) - \frac{2}{n} E\left(\sum_{i=1}^n \zeta_i \left(\sum_{k=1}^n a_{ik}\right)\right) \left(\sum_{i=1}^n \zeta_i\right) + \mathbf{1}'\mathbf{A}\mathbf{1}E(\bar{\boldsymbol{\zeta}}^2) \\ &= \sum_{i=1}^n a_{ii}(a_{ii} - d_{ii}) + \sum_{i=1}^n \sum_{j=1, j \neq i}^n a_{ij}^2 - \frac{2}{n} \left[\sum_{i=1}^n \left(\sum_{k=1}^n a_{ik}\right)(a_{ii} - d_{ii}) + \sum_{i=1}^n \sum_{j=1, j \neq i}^n \left(\sum_{k=1}^n a_{ik}\right)a_{ij} \right] + \mathbf{1}'\mathbf{A}\mathbf{1}\frac{1}{n^2}[\mathbf{1}'\mathbf{A}\mathbf{1} - \text{tr}(\mathbf{D})] \\ &= \text{tr}(\mathbf{A}'\mathbf{A}) - \text{tr}(\mathbf{A}\mathbf{D}) - \frac{2}{n}(\mathbf{1}'\mathbf{A}'\mathbf{A}\mathbf{1} - \mathbf{1}'\mathbf{A}\mathbf{D}\mathbf{1}) + \frac{1}{n^2}(\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1} - \mathbf{1}'\mathbf{A}\mathbf{1}\text{tr}(\mathbf{D})) \\ &= \text{tr}(\mathbf{A}'\mathbf{A}) - \frac{2}{n}\mathbf{1}'\mathbf{A}'\mathbf{A}\mathbf{1} + \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{A}\mathbf{1} - \text{tr}(\mathbf{A}\mathbf{D}) + \frac{2}{n}\mathbf{1}'\mathbf{A}\mathbf{D}\mathbf{1} - \frac{1}{n^2}\mathbf{1}'\mathbf{A}\mathbf{1}\text{tr}(\mathbf{D}) \end{aligned}$$

235 And:

$$236 \quad \begin{aligned} E((\mathbf{z} - \bar{\mathbf{z}})' \mathbf{A}(\mathbf{x} - \bar{\mathbf{x}})) &= E((\boldsymbol{\zeta} - \mathbf{1}\bar{\boldsymbol{\zeta}})' \mathbf{A}(\mathbf{w} - \boldsymbol{\zeta})) = E(\boldsymbol{\zeta}'\mathbf{A}\mathbf{w}) - E(\boldsymbol{\zeta}'\mathbf{A}\boldsymbol{\zeta}) - E(\bar{\boldsymbol{\zeta}}'\mathbf{1}'\mathbf{A}\mathbf{w}) + E(\bar{\boldsymbol{\zeta}}'\mathbf{1}'\mathbf{A}\boldsymbol{\zeta}) \\ &= 0 \end{aligned}$$

$$237 \quad \begin{aligned} E((\mathbf{x} - \bar{\mathbf{x}})' \mathbf{A}(\mathbf{x} - \bar{\mathbf{x}})) &= E((\mathbf{w} - \boldsymbol{\zeta})\mathbf{A}(\mathbf{w} - \boldsymbol{\zeta})) = E(\mathbf{w}'\mathbf{A}\mathbf{w}) - E(\mathbf{w}'\mathbf{A}\boldsymbol{\zeta}) - E(\boldsymbol{\zeta}'\mathbf{A}\mathbf{w}) + E(\boldsymbol{\zeta}'\mathbf{A}\boldsymbol{\zeta}) \\ &= \text{tr}(\mathbf{A}'\mathbf{A}) - (\text{tr}(\mathbf{A}'\mathbf{A}) - \text{tr}(\mathbf{A}\mathbf{D})) \\ &= \text{tr}(\mathbf{A}\mathbf{D}) \end{aligned}$$

1 **SIMULATION PARAMETERS**

2 A certain number of simulations were performed in order to validate the algebraic formulae
3 described in our paper. All the methods were tested, that is the REGRESSION, QTDT,
4 GRAMMAR and FASTA methods. The present validation was restricted to the family
5 structures and heritability values used in the Application section of the paper. The population
6 used for the simulations therefore comprised 600 genotyped individuals, offspring of 120, 20
7 and 10 sires that respectively produced 5, 30 and 60 offspring. To do this, the genotypes for a
8 SNP were simulated for sires and dams with a MAF of 0.5, and the genotypes of the offspring
9 were extrapolated from their parents' genotypes. Next, the polygenic values of the sires and
10 offspring and the phenotypes of the offspring were computed with and without the effect of a
11 corresponding QTL with an allele substitution effect of 0.20 (equivalent to a regression
12 coefficient of 0.141 for phenotypic standard deviation or even a QTL effect explaining 2% of
13 the phenotypic variance). The robustness and power of each method were then evaluated
14 using these two phenotypes (with or without a QTL) with a significance threshold of 5%
15 (which is different from the 1% threshold used in the paper). The simulations were performed
16 with heritability values ranging from 0 to 1 by 0.1 steps. 10,000 simulations were carried out
17 for each scenario. In all, 1,320,000 simulations were performed.

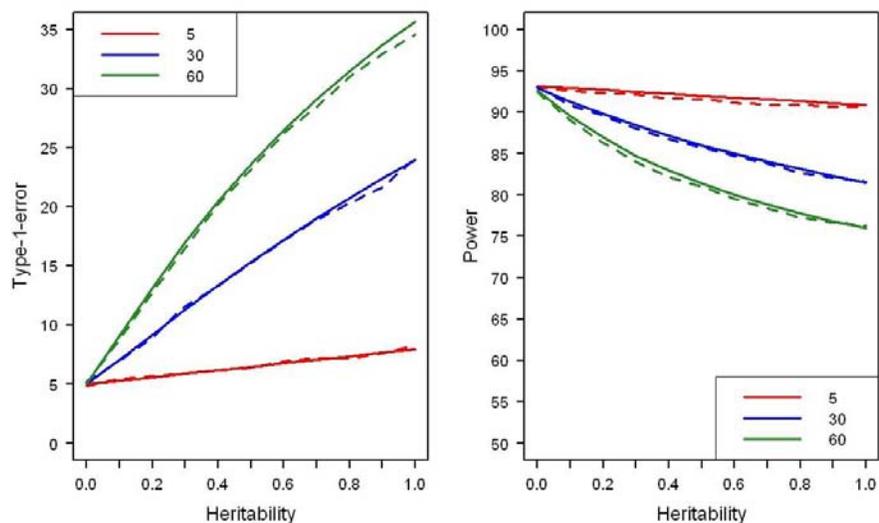
18 For the GRAMMAR and FASTA methods, the ASREML software (Gilmour et al., 2006) was
19 used for estimating variance components. It should also be noted that the relationship matrix
20 used for these two methods is derived from pedigree data and not genomic data.

21 **RESULTS**

22 In this section, all the differences between the theoretical and simulated values are given as a
23 function of the type-1 error and type-2 error (robustness and power), which are both expressed
24 in percentage. Theoretical values were computed using a R program named RobPower.

25 **Regression model**

26 Figure 1 shows the simulation results (dashed lines) and the theoretical results (solid lines) for
 27 the regression model. As regards to robustness, the average absolute value of the difference
 28 between the curves was of 0.26 % (± 0.25) and was maximal (1.09%) for families with 60
 29 offspring and a heritability value of 1. The difference was, on average, greater for larger
 30 family structures (0.46% for families with 60 offspring compared to 0.11% for families with 5
 31 offspring). As regards to power, the average absolute value of the difference between the
 32 curves was of 0.37% (± 0.17) and was maximal (0.76%) for families with 60 offspring and a
 33 heritability value of 0.3. The difference was, on average, greater for family structures with 60
 34 offspring per family (0.46%).
 35 On the whole the differences were weak, hence validating the theoretical results for the
 36 regression model.



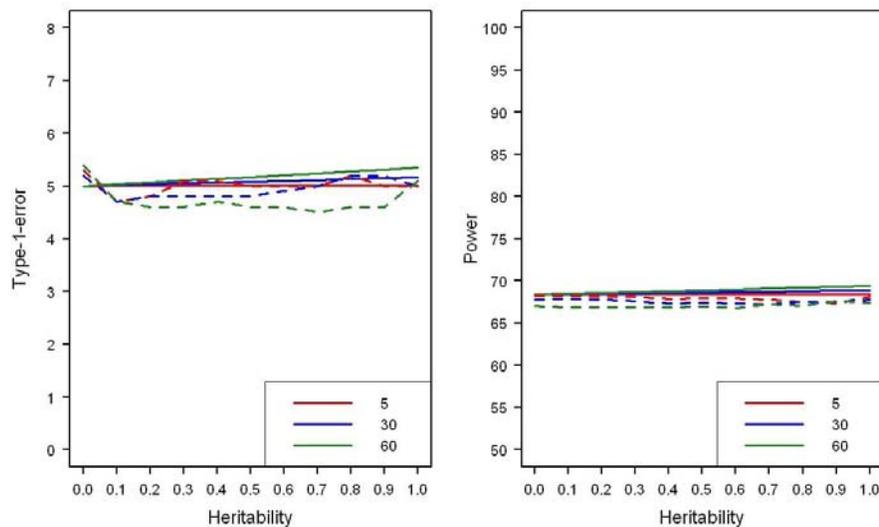
37

38 *Figure 1: Robustness and power for the regression model*

39

40 *QTD model*

41 Figure 2 shows the simulation results (dashed lines) and the theoretical results (solid lines) for
 42 the QTD model. As regards to robustness, the average absolute value of the difference
 43 between the curves was of 0.28% (± 0.22) and was maximal (0.74%) for families with 60
 44 offspring and a heritability value of 0.7. The difference was, on average, greater for larger
 45 family structures (0.52% for families with 60 offspring compared to 0.11% for families with 5
 46 offspring). As regards to power, the average absolute value of the difference between the
 47 curves was of 1.12% (± 0.63) and was maximal (0.76%) for families with 60 offspring and a
 48 heritability value of 0.3. The difference was, on average, greater for larger family structures
 49 (1.81% for families with 60 offspring compared to 0.47% for families with 5 offspring).
 50 On the whole the differences were relatively small, hence validating the theoretical results for
 51 the QTD model.



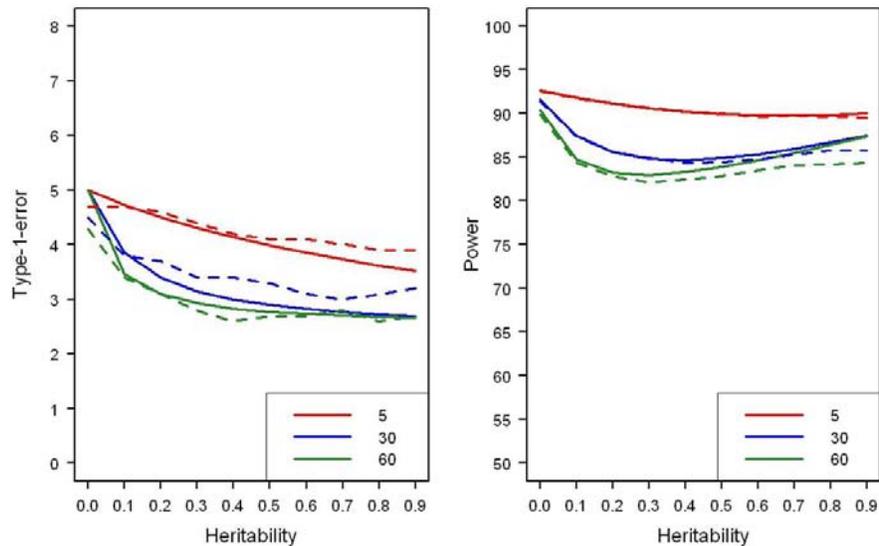
52

53 *Figure 2: Robustness and power for the QTD model*

54

55 **GRAMMAR model**

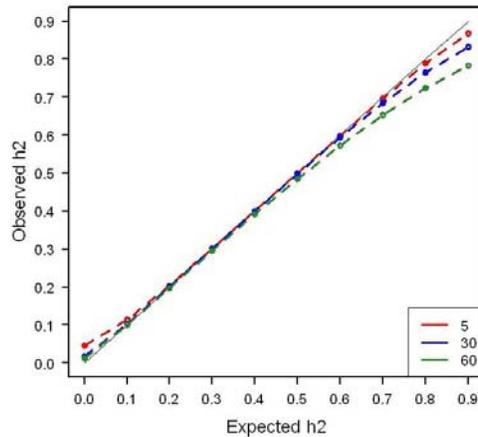
56 Figure 3 shows the simulation results (dashed lines) and the theoretical results (solid lines) for
 57 the GRAMMAR method. As regards to robustness, the average absolute value of the
 58 difference between the curves was of 0.22 % (± 0.17) and was maximal (0.7%) for families
 59 with 60 offspring and a null heritability value. As regards to power, the average absolute
 60 value of the difference between the curves was of 0.58% (± 0.71) and was maximal (2.97%)
 61 for families with 60 offspring and a heritability value of 0.9. The difference was, on average,
 62 greater for larger family structures (1.16% for families with 60 offspring compared to 0.11%
 63 for families with 5 offspring). It should be noted that differences seem to increase with the
 64 heritability value. A possible explanation for this is that the estimation of heritability obtained
 65 with ASREML was biased (underestimated) for the higher simulated heritability values.
 66 Figure 4 shows the distribution between expected (theoretical) and observed (simulated)
 67 heritability values.



68

69 *Figure 3: Robustness and power for the GRAMMAR model*

70



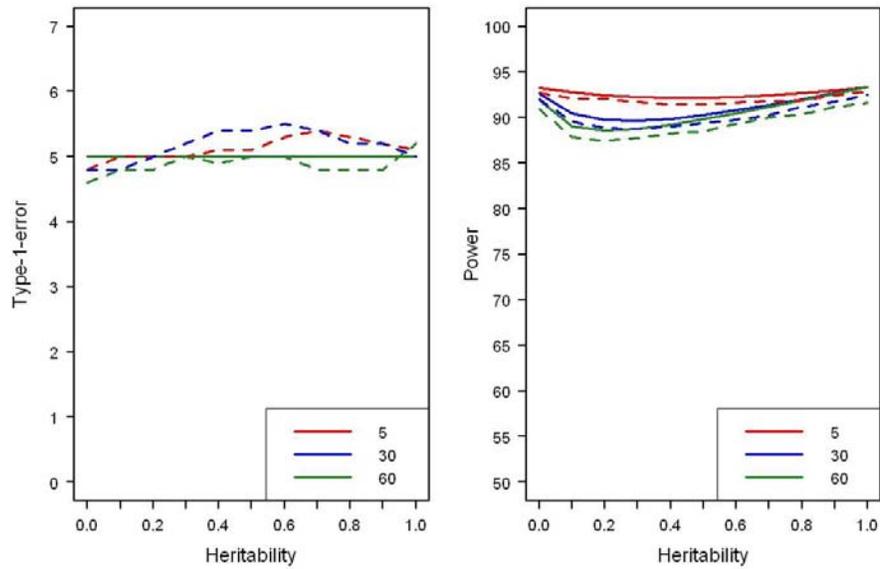
71

72 *Figure 4: Bias between expected and observed heritability values for the GRAMMAR method*

73

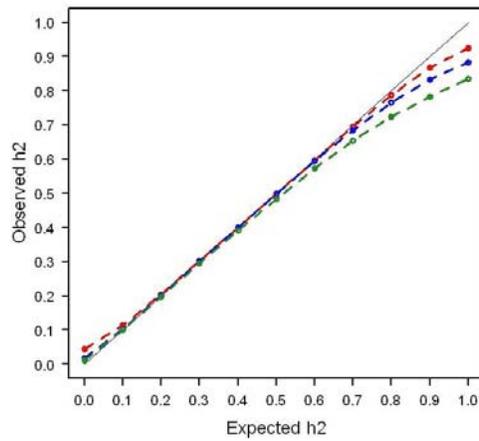
74 ***FASTA model***

75 Figure 5 shows the simulation results (dashed lines) and the theoretical results (solid lines) for
 76 the FASTA model. As regards to robustness, the average absolute value of the difference
 77 between the curves was of 0.18% (± 0.14) and was maximal (0.5%) for families with 30
 78 offspring and a heritability value of 0.6. As regards to power, the average absolute value of
 79 the difference between the curves was of 0.90% (± 0.32) and was maximal (1.74%) for
 80 families with 60 offspring and a heritability value of 1. The difference was, on average,
 81 greater for larger family structures (1.25% for families with 60 offspring compared to 0.59%
 82 for families with 5 offspring). As with the GRAMMAR method, it should be noted that
 83 differences seem to increase with the heritability value and are potentially caused by a bias
 84 between the expected and observed heritability values. Figure 6 shows the distribution
 85 between the heritability values.



86

87 *Figure 5: Robustness and power for the FASTA model*



88

89 *Figure 6: Bias between expected and observed heritability values for the FASTA method*

90 Gilmour, A. R., Gogel, B. J., Cullis, B. R., et Thompson, R. (2006). ASREML user guide
 91 release 2.0. VSN International Ltd, Hemel Hempstead, UK, page 320.



Annexes B : Détails des lésions d'ostéochondrose

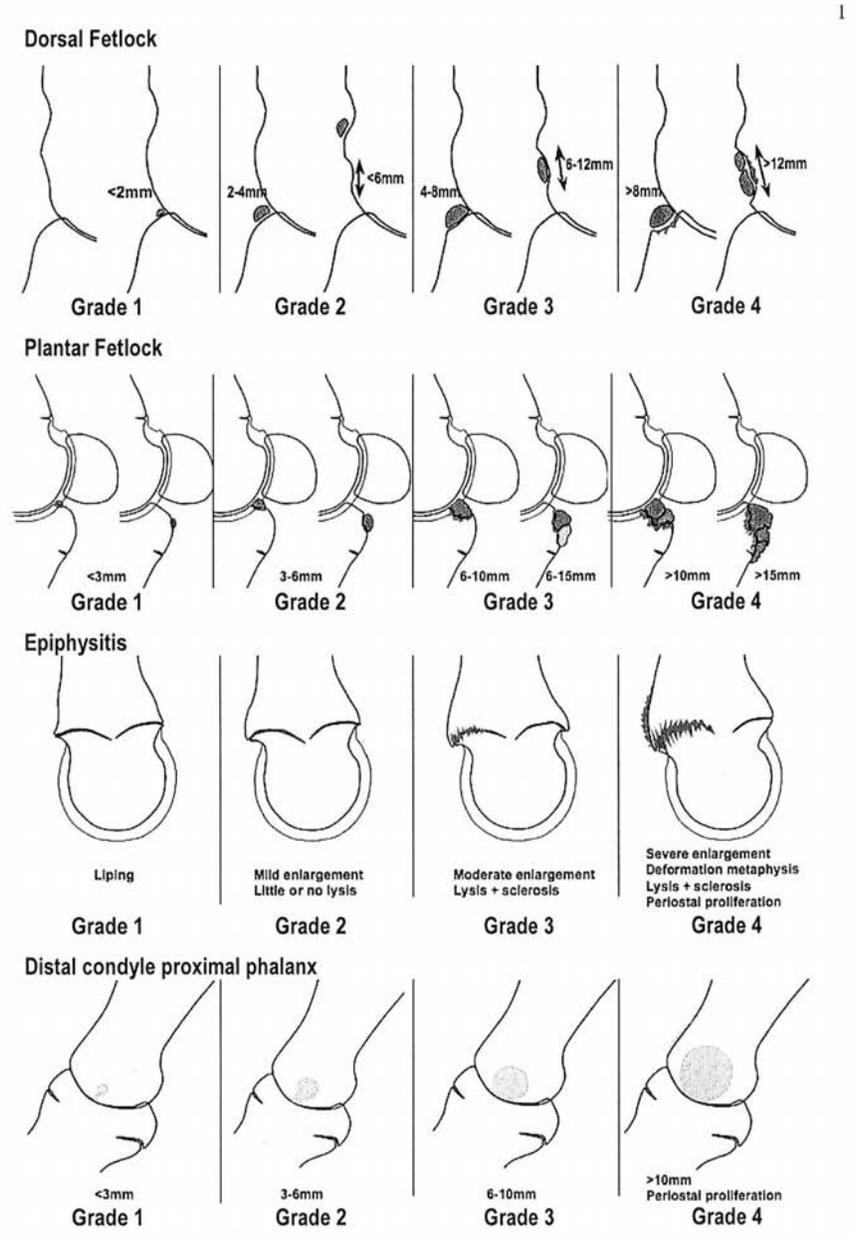


FIGURE 4.20 – Bilan radiographique GENEQUIN (d’après JM Denoix)

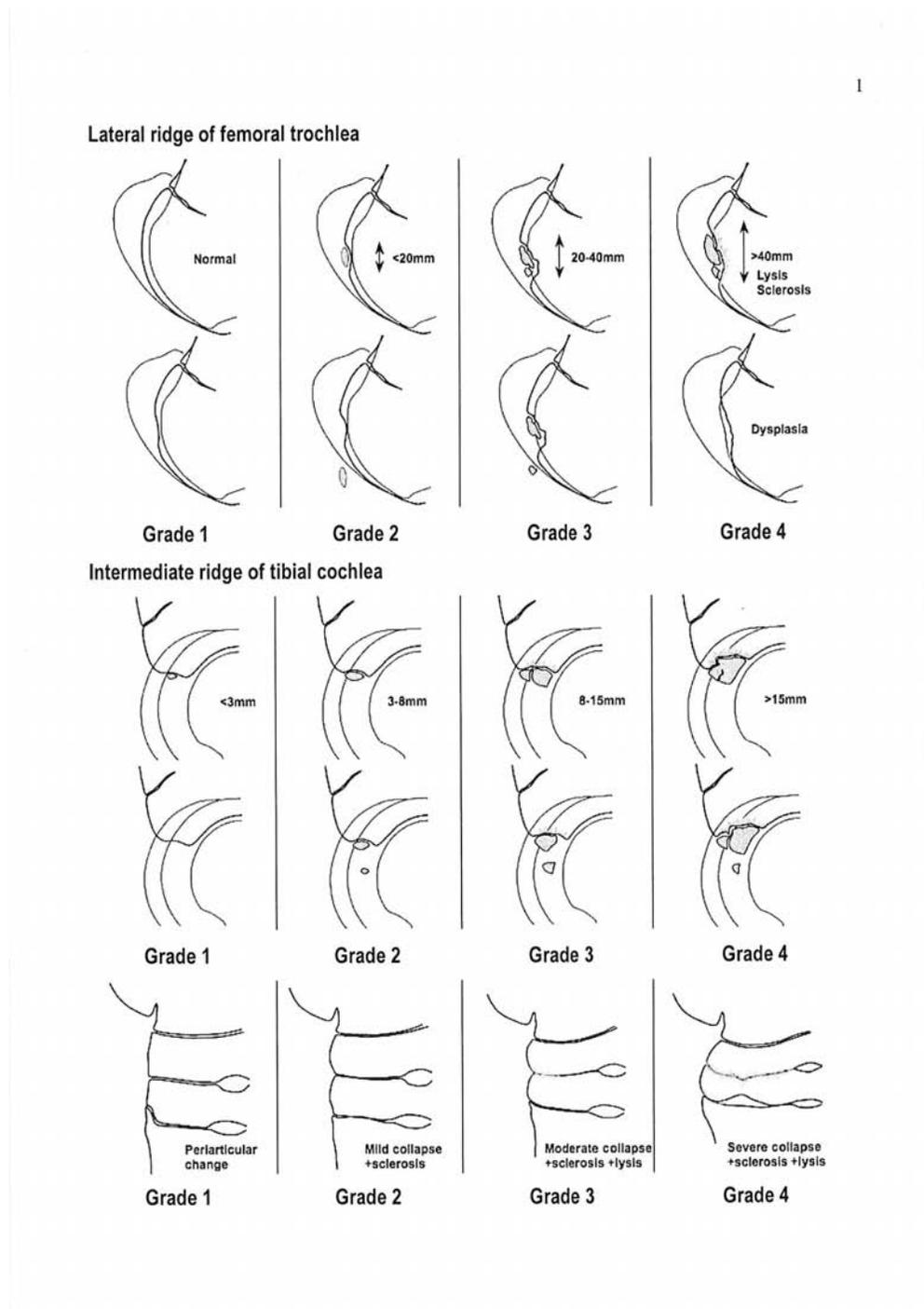


FIGURE 4.21 – Bilan radiographique GENEQUIN (d'après JM Denoix)



Annexes C : Article LDSO

Cet article a été soumis à *Journal of Animal Breeding and Genetics* en 2010

1 INTRODUCTION

2 Linkage Disequilibrium (LD), the non random association of alleles at different loci, is a tool of
3 increasing importance in animal genetics. LD is required for fine mapping of Quantitative Trait
4 Loci (QTL) (Meuwissen and Goddard 2000), for the estimation of the effective population size
5 (Hayes et al. 2003) or for genomic selection procedures (Meuwissen et al. 2001). In animal
6 breeding, only one program simulating LD is publically available to evaluate and validate the
7 methods: QMSim (Sargolzaei and Schenkel 2009). According to its manual, QMSim accounts for
8 a single way of selecting in the recent generations although the way of selecting has changed over
9 the time, and it accounts only for additive effects. But most of all, it does not allow including a
10 known pedigree in the simulation although this inclusion allows better mimicking the LD
11 structure in existing populations. As the optimal design for a fine mapping experiment depends
12 among other on the existing LD and the relative contribution of LD and family linkage to the
13 information, the use of a real pedigree may improve the ability of designing optimized designs by
14 varying the quantity of information used from the pedigree.

15 We herein present a simulation program, LDSO, that overcomes these drawbacks, and we present
16 some outcomes.

17 METHODS

18 Program

19 LDSO (Linkage Disequilibrium with Several Options) is a Fortran90, completely self-contained
20 program for simulations of whole diploid population histories under various historical scenarios
21 based on the gene-dropping method (MacCluer et al. 1986). The random number generator from
22 L'Ecuyer (1996) is used. The genetic history of one or two populations can be simulated; the

1 output files can deliver various statistics (inbreeding rates, allele frequencies, linkage
2 disequilibrium) on these populations. Evolutionary forces that are classically found in livestock
3 populations, such as mutation, selection, changes in the population size or random drift, can be
4 taken into account, allowing the simulation of a wide-range of situations. The parameters have to
5 be provided by the user using simple text files.

6 **Simulation of populations**

7 The history of the populations is subdivided into two parts, “historical” and “recent”. In each
8 generation, a set of individuals is chosen at random, with replacement, as parents of the next
9 generation (e.g., the Wright-Fisher model, Wright 1931). If two populations are simulated, their
10 genetic structure and histories are independent, but the genetic parameters (mutation rates, initial
11 LD and QTL effects) are the same.

12 In the “recent” part of the simulation, five population types (Table 1) can be derived from the last
13 historical generation:

- 14 – Type 1: no particular design. If there was a single historical population, a random mating
15 population will be simulated . If two populations were simulated, the two populations may:
16 (a) mate at random, (b) produce an hybrid F1 population (c) be used to introgress one
17 population into the other.
- 18 – Type 2: a grand-daughter design, following one unique historical population.
- 19 – Type 3: Back-Crosses (BC) after a simulation with two historical populations.
- 20 – Type 4: from an existing pedigree. Founders sampled from the historical part are used as
21 parents of the founders of the pedigree. Once the founders have received their genotypes and
22 phenotypes, a classical gene-drop transmission is operated throughout the pedigree. If non-

1 founder animals happen to have an unknown parent, the missing haplotype is chosen from the
2 base generation. The pedigree is assumed to be correct.

3 – Type 5: to determine the optimal experimental size. Half-sibs (paternal and maternal) and
4 full-sibs designs with one or two litter/dam can be modelled. Individuals produced in the
5 historical part and/or that exists in an actual pedigree can be used as parents of the sibs.

6 **Simulation of genetic architecture**

7 The number of marker loci, QTL and chromosomes, and the number of alleles at each locus are
8 unlimited, providing the possibility of simulating the evolution of a whole genome. The
9 chromosomes may have different length to mimic a real genome.

10 Fine mapping methods may differ in their assumptions relative to the initial LD in the population.

11 For that reason, two initial situations of LD have been considered:

12 – 1st situation: no initial LD assumed (Meuwissen and Goddard 2000). Alleles are drawn at
13 random.

14 – 2nd situation: a unique mutation is introduced into the population in a haplotype that may not
15 be unique. This classically corresponds to a mutation in a gene involved in a quantitative trait
16 (corresponding to the QTL) in a population that was formerly in Linkage Equilibrium (LE).

17 Some fine-mapping methods suppose this initial situation (Terwilliger 1995, Farnir 2002).

18 **Simulation of evolutionary forces**

19 Four evolutionary forces are considered: genetic drift, selection, changes of population sizes and
20 mutation. Genetic drift occurs because of limited populations sizes. To create a new generation,

21 only the s percents (s ranging from 0 to 1) of the individuals with the best performances will
22 become parents. In the recent part of the population history, selection can be operated on different

23 evaluations of the breeding values: phenotypic selection, selection on breeding values with a

1 given accuracy or BLUP selection. In the historical part, the change of population sizes may be
 2 punctual (the number of individuals changes in a single generation) or there may be a constant
 3 expansion or contraction rate (i.e. a constant slope between the two extreme generations).
 4 Mutations may occur with three different rates corresponding to biallelic markers (generally
 5 Single Nucleotide Polymorphisms, SNP), multi-allelic markers (generally microsatellites) and
 6 QTL. Mutations on microsatellites follow the Stepwise Mutation Model (Kimura and Ohta 1978),
 7 while there is a switch from one allele to the other for SNP markers. For both kinds of genetic
 8 markers, the number of mutations is sampled from a binomial distribution. When only mutation
 9 and random drift are assumed, a random-drift equilibrium will be established.

10 **Simulation of phenotypes**

11 Phenotypes are simulated as the sum of the QTL additive effects, possible dominant, epistatic and
 12 imprinting QTL effects, an infinitesimal additive polygenic effect and an environmental effect.

13 The polygenic Mendelian sampling effect of the individual i is sampled from a normal

14 distribution with mean 0 and variance $\sigma_{Poly_{t+1},i}^2 = \sigma_{Poly}^2 * \left(\frac{2 - F_s + F_d}{4} \right)$ where σ_{Poly}^2 is the

15 polygenic variance in the founder generation and F_s (resp. F_d) is the inbreeding coefficients of

16 the sire (resp. dam) of the offspring i . Residual effects are normally distributed with mean 0 and a

17 variance constant over time. Two kinds of epistatic QTL effects were considered:

- 18 - 1st kind: multiplicative effects. The effects of the genotypes at both loci are multiplied
 19 (Cordell 2002),
- 20 - 2nd kind: "compositional epistasis" (Phillips 2008). The effect of the genotype at the second
 21 locus is expressed only if the individual has at least one favourable allele at the first locus.

1 The loci showing epistatic effects are associated at random. A QTL can only be involved in one
2 pair of epistatic QTL. QTL effects may be defined by the user or be drawn at random from a
3 Gamma distribution (Hayes and Goddard 2001) with parameters determined by the user.
4 Phenotypes are computed in the historical part only if selection is applied. Otherwise, they are
5 computed in the last generation of the historical population.

6 **General algorithm**

7 In each generation, the individuals are first subjected to selection if $s < 1$. Once the potential
8 parents have been selected, the actual ones are chosen and mated at random. The offspring are
9 created according to gene dropping method (MacCluer et al. 1986), which is a forward process.
10 Each generation is created by transmitting the parental alleles to its offspring following the
11 Mendelian rules. Recombinations are implemented as a Poisson process, assuming no crossover
12 interference. The number of recombinations is drawn from a Poisson distribution with parameter
13 the length of the chromosome. The recombinations to obtain the haplotype transmitted to the
14 offspring are then positioned at random on the chromosome. The principle of gene dropping
15 makes it possible to apply selection in each generation, contrary to a coalescent process.

16 When the population is created, each allele can be given a unique founder number (independently
17 from the QTL allelic state) corresponding to the founder haplotype, resulting in $2N$ founder
18 numbers, where N is the founder population size. These numbers can be traced on all along the
19 simulations, so that the IBD status at each locus and thus the average inbreeding rate for each
20 locus in the whole population can be computed. A mutation leads to the appearance of a new
21 founder number, indicating that this locus is no longer IBD with that of any ancestor. These
22 founder numbers are stored additionally to the allele numbers. Both the haplotypes (relying on
23 the alleles) and the founder origins can thus be obtained in a single replicate, while two runs (one

1 with the initial allelic state situation and one with unique alleles for each founder) would be
2 required with QMSim.

3 If a real pedigree is used in the “recent” part, the paternal and maternal haplotypes of each of the
4 pedigree founders are sampled randomly with replacement among the haplotypes of the sires and
5 dams of the last historical generation. Once the founders have received their genotypes and
6 phenotypes, a forward process is initiated along the pedigree. If one parent is unknown, the
7 haplotype for this parent is randomly chosen in the last historical generation.

8 **OUTPUTS**

9 Different files are created according to the user’s directions. They can contain information of
10 historical generations, such as allelic frequencies, founder number frequencies, average
11 inbreeding at each locus, Polymorphism Information Content (Botstein *et al.* 1980) of the QTL,
12 or LD information. Two LD measures are computed, the D' (Lewontin 1964, Hedrick 1987) and
13 the standardized χ^2 (Yamazaki 1977). The latter corresponds to the r^2 (Hill and Robertson 1968)
14 when applied to biallelic markers. Contrary to the output of QMSim, LDSO can also provide the
15 LD values between all loci, making the study of the LD pattern in the neighbourhood of the QTL
16 in contrast to that in neutral regions possible.

17 The output files for the “recent” part are: the pedigree, the molecular information of the
18 individuals in the two generations, and phenotypes. The haplotypes provided in the output files
19 can be obtained with typing errors and missing data.

20 The high number of outputs may be a limiting factor as it slows down the program and that these
21 outputs may require a large storage place.

22 **General considerations**

1 The time required for one simulation depends mainly on the number of requested output files, but
2 also on the population size and on the number of simulated loci. Two populations ending with
3 five generations of back-cross with 10 males at each generation mated each to 5 females with 20
4 offspring per female were simulated on an AIX5.2.0 web server with 62,720 megabytes of
5 internal memory and a 64 bits-processor. For an initial population size of 500 individuals and 101
6 loci simulated, 58 seconds were needed. With a real pedigree, 1,000 individuals in the initial
7 population and 1,000 SNPs simulated over 1,000 historical generations, 729 seconds were
8 required per simulation. For that simulation, less than 1Gb of memory was required.

9 LDSO provides a very large range of options on population history and population structures,
10 covering a great number of situations observed in livestock populations. However, it does not
11 account for overlapping generations. It is different from other simulation programs mainly
12 through the opportunity of applying several types of selection and the variety of non-additive
13 effects at the QTL loci. It is particularly suited to investigate the effects of selection on mapping
14 methods or LD, or to test fine mapping strategies in existing pedigrees.

15 The program LDSO can be downloaded on the quantitative genetics platform of the INRA
16 (<https://qgp.jouy.inra.fr/>). It is provided as a source code with a manual and examples of
17 simulations and outputs.

18 **References**

- 19 Botstein, D., White, R.L., Skolnick, M., Davis, R.W. (1980) Construction of a genetic linkage
20 map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.*, 32, 314-331.
21 Cordell, H.J. (2002) Epistatis: what it means, what it doesn't mean, and statistical methods to
22 detect it in humans. *Human Molecular Genetics*, 11: 2463-2468.

- 1 Farnir, F., Grisart, B., Coppeters, W., Riquet, J., Berzi, P., Cambisano, N., Karim, L., Mni, M.,
2 Moisisio, S., Simon, P., Wagenaar, D., Vilkkki, J., Georges, M. (2002) Simultaneous mining of
3 linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib
4 pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production
5 on bovine chromosome 14. *Genetics*, 161, 275-287.
- 6 Haldane, J.B.S. (1919) The combination of linkage values, and the calculation of distances
7 between loci of linked factors. *J. Genet.*, 8, 299-309.
- 8 Hayes, B.J., Visscher, P.M., McPartlan, H.C., Goddard, M.E. (2003) Novel Multilocus Measure
9 of Linkage Disequilibrium to Estimate Past Effective Population Size. *Genome Res.*, 13, 635-
10 643.
- 11 Hayes, B.J., Goddard, M.E. (2001) The distribution of the effects of genes affecting quantitative
12 traits in livestock. *Genet. Sel. Evol.*, 33, 209-229.
- 13 Hedrick, P.W. (1987) Gametic disequilibrium measures: proceed with caution. *Genetics*, 117,
14 331-341.
- 15 Hill, W.G., Robertson, A. (1968) Linkage disequilibrium in finite populations. *Theoretical and*
16 *Applied Genetics*, 38, 226-231.
- 17 Kimura, M., Ohta, T. (1978) Stepwise mutation model and distribution of allelic frequencies in a
18 finite population. *Proc. Natl. Acad. Sci. U.S.A.*, 75, 2868-2872.
- 19 L'Ecuyer, P. (1996) Maximally equidistributed combined Tausworthe generators, *Math. of*
20 *Comput.*, 65, 203-213, available at <http://jblevins.org/mirror/amiller/taus88.f90>
- 21 Lewontin, R.C. (1964) On measures of gametic disequilibrium. *Genetics*, 49, 49-67.
- 22 MacCluer, J.W., VandeBerg, J.L., Read, B., Ryder, O.A. (1986) Pedigree analysis by computer
23 simulation. *Zoo. Biol.*, 5, 147-160.

- 1 Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E. (2001) Prediction of total genetic value using
- 2 genome-wide dense marker maps. *Genetics*, 157, 1819–1829.
- 3 Meuwissen, T.H.E., Goddard, M.E. (2000) Fine mapping of quantitative trait loci using linkage
- 4 disequilibria with closely linked marker loci. *Genetics*, 155, 421-430.
- 5 Phillips P.C. (2008) Epistasis – the essential role of gene interactions in the structure and
- 6 evolution of genetic systems. *Nature Reviews*, 9, 855-867.
- 7 Sargolzaei, M., Schenkel, F.S. (2009) QMSim: a large-scale genome simulator for livestock.
- 8 *Bioinformatics*. 25, 680-681.
- 9 Terwilliger, J.D. (1995) A powerful likelihood method for the analysis of LD between trait loci
- 10 and one or more polymorphic loci. *Am. J. Hum. Genet.*, 56, 777-787.
- 11 Tranulis, M.A. (2002) Influence of the prion protein gene, Prnp, on scrapie susceptibility in
- 12 sheep. *APMIS*, 110, 33-43.
- 13 Wright, S. (1931) Evolution in Mendelian populations. *Genetics*, 16, 97–159.
- 14 Yamazaki, T. (1977) The effects of overdominance on linkage in a multilocus system. *Genetics*,
- 15 86, 227-236.

Table 1: Summary of the options for the “recent” part of the population history

Option	Number of populations in the historical part	Type of population in the recent part	Number of generations genotyped
1	1	Random mating	n
	2	- Random mating	n
		- F1 and Fn generations	n
		- introgression	n
2	1	Grand-daughter design	2
3	2	F2 or Backcross	n
4	1 or 2	Real pedigree	n
5	1	Half-sibs and Full-sibs	2

n: as many generations as wished can be produced.

Bibliographie

- Abdallah, J. M., Mangin, B., Goffinet, B., Cierco-Ayrolles, C., and Perez-Enciso, M. (2004). A comparison between methods for linkage disequilibrium fine mapping of quantitative trait loci. Genet. Res., 83 :41–47.
- Abecasis, G. R., Cardon, L. R., and Cookson, W. O. C. (2000). A general test of association for quantitative traits in nuclear families. Am. J. Hum. Genet., 66 :279–292.
- Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet., 30 :97–101.
- Abecasis, G. R. and Cookson, W. O. C. (2000). Gold - graphical overview of linkage disequilibrium. Bioinformatics, 16 :182–183.
- Akey, J., Jin, L., and Xiong, M. (2001). Haplotypes vs single marker linkage disequilibrium tests : what do we gain ? European Journal of Human Genetics, 9 :291–300.
- Allison, D. B. (1997). Transmission-disequilibrium tests for quantitative traits. Am J Hum Genet, 60 :676–690.
- Alvarado, A. F., Marcoux, M., and Breton, L. (1990). The incidence of osteochondrosis in a standardbred breeding farm in quebec. In Proceedings of the Annual Convention of the American Association of Equine Practitioners, volume 35, pages 293–307.
- Ambrosius, W. T., Lange, E. M., and Langefeld, C. D. (2004). Power for genetic association studies with random allele frequencies and genotype distributions. Am J Hum Genet, 74 :683–693.
- Amin, N., Duijn, C. M. V., and Aulchenko, Y. S. (2007). A genomic background based method for association analysis in related individuals. PLoS ONE, 12 :e1274.
- Andersson-Eklund, L., Uhlhorn, H., Lundeheim, N., Dalin, G., and Andersson, L. (2000). Mapping quantitative trait loci for principal components of bone measurements and osteochondrosis scores in a wild boar x large white intercross. Genetical Research, 75 :223–230.
- Ardlie, K. G., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics, 3 :299–309.
- Armitage, P. (1955). Tests for linear trends in proportions and frequencies. Biometrics, 11 :375–386.
- Aulchenko, Y. S., de Koning, D.-J., and Haley, C. (2007a). Genomewide rapid association using mixed model and regression : A fast and simple method for genomewide pedigree/based quantitative trait loci association analysis. Genetics, 177 :577–585.
- Aulchenko, Y. S., Ripke, S., Isaacs, A., and Duijn, C. M. V. (2007b). GenABEL : an R library for genome-wide association analysis. Bioinformatics, 23 :1294–1296.
- Bacanu, S.-A., Devlin, B., and Roeder, K. (2002). Association studies for quantitative traits in structured populations. Genetic Epidemiology, 22 :78–93.
- Balding, D. J. (2006). A tutorial on statistical methods for population association studies. Nature Review Genetics, 7 :781–791.
- Ball, R. D. (2005). Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. Genetics, 170 :859–873.

- Barneveld, A. and VanWeeren, P. R. (1999). Conclusions regarding the influence of exercise on the development of the equine musculoskeletal system with special reference to osteochondrosis. Equine Veterinary Journal, 31 (S31) :112-119.
- Barret, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haplowiew : analysis and visualization of ld and haplotype maps. Bioinformatics, 21 :263-265.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. J. R. Stat. Soc. B, 57 :289-300.
- Bertone, A. L., Bramlage, L. R., McIlwraith, C. W., and Malemud, C. L. (2005). Comparison of proteoglycan and collagen in articular cartilage of horses with naturally developing osteochondrosis and healing osteochondral fragments of experimentally induced fractures. American journal of veterinary research, 66 (11) :1881-1890.
- Blott, S., Kim, J.-J., Moiso, S., Schmidt-Küntzel, A., Cornet, A., Berzi, P., Cambisano, N., Ford, C., Grisart, B., Johnson, D., Karim, L., Simon, P., Snell, R., Spelman, R., Wong, J., Vilkki, J., Georges, M., Farnir, F., and Coppeters, W. (2003). Molecular dissection of a quantitative trait locus : A phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. Genetics, 163 :253-266.
- Boitard, S., Abdallah, J., de Rochambeau, H., Cierco-Ayrolles, C., and Mangin, B. (2006). Linkage disequilibrium interval mapping of quantitative trait loci. BMC Genomics, 7 :54-68.
- Boitard, S., Mangin, B., and Azaïs, J. M. (2010). Asymptotic distribution of the orthogonal quantitative transmission disequilibrium test in a structured population : Exact formula. Statistical Applications in Genetics and Molecular Biology, 9(1) :Art 11.
- Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8 :3-62.
- Bridges, C. H. and Harris, E. D. (1988). Experimentally induced cartilaginous fractures (osteochondritis dissecans) in foals fed low-copper diets. Journal of the American Veterinary Medical Association, 193 (2) :215-221.
- Browning, S. R. (2008). Missing data imputation and haplotype phase inference for genome-wide association studies. Hum. Genet., 124 :439-450.
- Browning, S. R. and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. Am. J. Hum. Genet., 81 :1084-1097.
- Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., and Samani, N. J. (2007). Genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls. Nature, 447 :661-678.
- Calus, M., Meuwissen, T. H. E., de Roos, A. P. W., and Veerkamp, R. (2008). Accuracy of genomic selection using different methods to define haplotypes. Genetics, 178(1) :553-561.
- Calus, M., Mulder, H., and Veerkamp, R. (2011). Estimating genomic breeding values and detecting qtl using univariate and bivariate models. BMC proceedings, 5(S3) :S5.
- Cardon, L. R. and Abecasis, G. R. (2003). Using haplotype blocks to map human complex trait loci. Nature Genet., 19(3) :135-140.
- Cardon, L. R. and Palmer, L. J. (2003). Population stratification and spurious allelic association. Lancet, 361 :598-604.
- Carlson, C. S., Cullins, L. D., and Meuten, D. J. (1995). Osteochondrosis of the articular-epiphyseal cartilage complex in young horses : evidence for a defect in cartilage canal blood supply. Veterinary Pathology Online, 32 (6) :641.
- Carlsten, J., Sandgren, B., and Dalin, G. (1993). Development of osteochondrosis in the tarsocrural joint and osteochondral fragments in the fetlock joints of standardbred trotters. i. a radiological survey. Equine Veterinary Journal, 25 (S16) :42-47.
- Caure, S., Tourtoulou, G., Valette, J. P., Cosnier, A., and Lebreton, P. (1998). Prévention de l'ostéochondrose chez le trotteur au sevrage : étude expérimentale. Pratique Vétérinaire Équine, 30 :49-59.
- Chapman, J. M., Cooper, J. D., Todd, J. A., and Clayton, D. G. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags : A class of tests and the determinants of statistical power. Hum. Hered., 56 :18-31.
- Christensen, O. F., Busch, M. E., Gregersen, V. R., Lund, M. S., Nielsen, N., Vingborg, R. K. K., and Bendixen, C. (2010). Quantitative trait loci analysis of osteochondrosis traits in the elbow joint of pigs. Animal, 4(3) :417-424.
- Churchill, G. A. and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. Genetics, 138 :963-971.
- Clark, A. G. (1990). Inference of haplotypes from pcr-amplified samples of diploid populations. Mol. Biol. Evol., 7 :111-122.

- Clark, A. G. (2004). The role of haplotypes in candidate gene studies. *Genetic Epidemiology*, 27 :321–333.
- Clayton, D., Chapman, J., and Cooper, J. (2004). Use of unphased multilocus genotype data in indirect association studies. *Genetic Epidemiology*, 27 :415–428.
- Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., Smink, L. J., Lam, A. C., Ovington, N. R., Stevens, H. E., Nutland, S., Howson, J. M. M., Faham, M., Moorhead, M., Jones, H. B., Falkowski, M., Hardenbol, P., Willis, T. D., and Todd, J. A. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics*, 37(11) :1243–1246.
- Cordell, H. J. and Clayton, D. G. (2002). A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data : Application to hla in type 1 diabetes. *Am. J. Hum. Genet.*, 70 :124–141.
- Crenshaw, T. D. (2006). Arthritis or ocd-identification and prevention. *Advances in Pork Production*, 17 :199–208.
- Darvasi, A. and Soller, M. (1992). Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theor Appl Genet*, 85 :353–359.
- Denoix, J. M., Audigié, F., Tapprest, J., Jacquet, S., and Coudry, V. (2002). Les affections ostéo-articulaires juvéniles (aoaj) : Nature des lésions et diagnostic. In *Compte-rendu de l'AVEF. Le Touquet, France*, pages 217–220.
- Denoix, J. M., Valette, J. P., Heilès, P., Ribot, X., and Tavernier, L. (2000). Etude radiographique des affections ostéo-articulaires juvéniles (aoaj) chez des chevaux de races françaises, âgés de 3 ans : présentation globale des résultats sur 1180. *Pratique Vétérinaire Equine*, 126 :35–41.
- Denoix, J. M., Valette, J. P., Robert, C., Houliez, D., and Heiles, P. (1996). Prévalence des images radiographiques anormales dans les membres de 575 chevaux de races françaises, âgés de 3 ans. *Pratique Vétérinaire Equine*, 28 :97–104.
- Der Kinderen, L. (2005). *Heritability of osteochondrosis in Dutch warmblood stallions from the second stallion inspection*. PhD thesis, MSc Thesis. Wageningen University, The Netherlands.
- Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, 55 :997–1004.
- Dierks, C., Komm, K., Lampe, V., and Distl, O. (2010). Fine mapping of a quantitative trait locus for osteochondrosis on horse chromosome 2. *Stichting International Foundation for Animal Genetics*, 41(S2) :87–90.
- Dierks, C., Löhring, K., Lampe, V., Wittwer, C., DrÖgemü, C., and Distl, O. (2007). Genome-wide search for markers associated with osteochondrosis in hanoverian warmblood horses. *Mamm Genome*, 18 :739–747.
- Dik, K. J., Enzerink, E. E., and VanWeeren, P. R. (1999). Radiographic development of osteochondral abnormalities, in the hock and stifle of dutch warmblood foals, from age 1 to 11 months. *Equine Veterinary Journal*, 31 (S31) :9–15.
- Donabedian, M., Fleurance, G., Perona, G., Robert, C., Lepage, O., Trillaud-Geyl, C., Leger, S., Ricard, A., Bergero, D., and Martin-Rosset, W. (2006). Effect of fast vs. moderate growth rate related to nutrient intake on developmental orthopaedic disease in the horse. *Animal Research*, 55 (5) :471–486.
- Druet, T., Fritz, S., Boussaha, M., Ben-Jemaa, S., Guillaume, F., Derbala, D., Zelenika, D., Lechner, D., Charon, C., Boichard, D., Gut, I. G., Eggen, A., and Gautier, M. (2008). Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on bta03 using a dense single-nucleotide polymorphism map. *Genetics*, 178 :2227–2235.
- Druet, T. and Georges, M. (2010). A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics*, 184 :789–798.
- Durrant, C., Zondervan, K., Cardon, L., Hunt, S., Deloukas, P., and Morris, A. (2004). Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *Am. J. Hum. Genet.*, 75 :35–43.
- Epstein, M. P., Allen, A. S., and Satten, G. A. (2007). A simple and improved correction for population stratification in case-control studies. *Am. J. Hum. Genet.*, 80 :921–930.
- Erbe, M., Ytournal, F., Pimentel, E. C. G., Shari, A. R., and Simianer, H. (2011). Power and robustness of three whole genome association mapping approaches in selected populations. *J. Anim. Breed. Genet.*, 128 :3–14.
- Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc KDD'96*, pages 226–231.
- Ewens, W. J., Li, M., and Spielman, R. S. (2008). A review of family-based tests for linkage disequilibrium between a quantitative trait and a genetic marker. *PLoS Genetics*, 4(9) :1–6.

- Fan, R. and Xiong, M. (2002). High resolution mapping of quantitative trait loci by linkage disequilibrium analysis. European Journal of Human Genetics, 10 (10) :607–615.
- Farnir, F., Grisart, B., Coppieters, W., Riquet, J., Berzi, P., Cambisano, N., Karim, L., Mni, M., Moisisio, S., Simon, P., Wagenaar, D., Vilkki, J., and Georges, M. (2002). Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees : revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. Genetics, 161 :275–287.
- Fernando, R. L. and Grossman, M. (1989). Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol., 21 :467–477.
- Freidlin, B., Zheng, G., Li, Z. H., and Gastwirth, J. L. (2002). Trend tests for case-control studies of genetic markers : Power, sample size and robustness. Hum. Hered., 53 :146–152.
- Fulker, D. W., Cherny, S. S., Sham, P. C., and Hewitt, J. K. (1999). Combined linkage and association sib-pair analysis for quantitative traits. Am. J. Hum. Genet., 64 :259–267.
- Geffroy, O., Couroucé, A., Valette, J. P., and Kraft, E. (1997). Pathologie ostéo-articulaire juvénile chez les cheval trotteur français : étude préliminaire. Pratique Vétérinaire Équine, 29 :191–199.
- George, A. W., Visscher, P. M., and Haley, C. S. (2000). Mapping quantitative trait loci in complex pedigrees : a two-step variance component approach. Genetics, 156 :2081–2092.
- Gilmour, A. R., Gogel, B. J., Cullis, B. R., and Thompson, R. (2006). Asreml user guide release 2.0. VSN International Ltd, Hemel Hempstead, UK, page 320.
- Grapes, L., Dekkers, J. C. M., Rothschild, M. F., and Fernando, R. L. (2004). Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. Genetics, 166 :1561–1570.
- Grapes, L., Firat, M., Dekkers, J. C. M., Rothschild, M. F., and Fernando, R. L. (2006). Optimal haplotype structure for linkage disequilibrium-based fine mapping of quantitative trait loci using identity by descent. Genetics, 172 :1955–1965.
- Gron Dahl, A. M. (1991). The incidence of osteochondrosis in the tibiotarsal joint of norwegian standardbred trotters : a radiographic study. J. Equine Vet. Sci., 11 :272–274.
- Gron Dahl, A. M. and Dolvik, N. I. (1993). Heritability estimations of osteochondrosis in the tibiotarsal joint and of bony fragments in the palmar/plantar portion of the metacarpal- and metatarsophalangeal joints of horses. J. Am. Vet. Med. Assoc., 203 :101–104.
- Gron dalen, T. (1974). Osteochondrosis and arthrosis in pigs 1. incidence in animals up to 120 kg live weight. Acta Vet. Scand., 15 :1–25.
- Guedj, M., Robelin, D., Hoebeke, M., Lamarine, M., Wojcik, J., and Nuel, G. (2006). Detecting local high-scoring segments : A first-stage approach for genome-wide association studies. Statistical Applications in Genetics and Molecular Biology, 5(1) :Art. 22.
- Guo, S. W. and Thompson, E. A. (1992). Performing the exact test of hardy-weinberg proportion for multiple alleles. Biometrics, 48 :361–372.
- Hardy, G. (1908). Mendelian proportions in a mixed population. Science, 28 :49–50.
- Hayes, B., Visscher, P., McPartlan, H., and Goddard, M. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res., 13 :635–643.
- Hayes, B. J., Chamberlain, A. J., McPartlan, H., MacLeod, I., Sethurman, L., and Goddard, M. E. (2007). Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. Genet. Res., 89 :215–220.
- Henderson, C. (1973). Sire evaluation and genetic trends. Journal of Animal Science, 1973(Symposium) :10–41.
- Hill, W. G. (1975). Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theoretical Population Biology, 8(2) :117–126.
- Hill, W. G. and Robertson, A. (1968). The effects of inbreeding at loci with heterozygote advantage. Genetics, 60 :615–628.
- Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. Nature Review Genetics, 6 :95–108.
- Hoggart, C. J., Parra, E. J., Shriver, M. D., Bonilla, C., Kittles, R. A., Clayton, D. G., and McKeigue, P. M. (2003). Control of confounding of genetic associations in stratified populations. Am. J. Hum. Genet., 72 :1492–1504.
- Hurtig, M. B. and Pool, R. R. (1996). Pathogenesis of equine osteochondrosis. Joint disease in the horse, pages 335–358.

- Jeffcott, L. B. (1991). Osteochondrosis in the horse-searching for the key to pathogenesis. *Equine Veterinary Journal*, 23 (5) :331–338.
- Jeffcott, L. B. (1993). Problems and pointers in equine osteochondrosis. *Equine Veterinary Journal*, 25 (S16) :1–3.
- Jeffcott, L. B. (1997). Osteochondrosis in horses. *In Practice*, 19 (2) :64.
- Jelan, Z. A., Jeffcott, L. B., Lundeheim, N., and Osborne, M. (1996). Growth rates in thoroughbred foals. *Pferdeheilkunde*, 12 :291–295.
- Kang, H., Sul, J., Service, S. K., Zaitlen, N., Kong, S., Freimer, N., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42 :348–354.
- Kauppi, L., Jasin, M., and Keeney, S. (2007). Meiotic crossover hotspots contained in haplotype block boundaries of the mouse genome. *PNAS*, 104(33) :13396–13401.
- Kemper, K. E., Emery, D. L., Bishop, S. C., Oddy, H., Hayes, B. J., Dominik, S., Henshall, J. M., and Goddard, M. E. (2011). The distribution of snp marker effects for faecal worm egg count in sheep, and the feasibility of using these markers to predict genetic merit for resistance to worm infections. *Genet. Res. Camb.*, 93 :203–219.
- Kizilkaya, K., Fernando, R. L., and Garrick, D. J. (2010). Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *J. Anim. Sci.*, 88 :544–551.
- Knight, D. A., Weisbrode, S. E., Schmall, L. M., Reed, S. M., Gabel, A. A., Bramlage, L. R., and Tyznik, W. I. (1990). The effects of copper supplementation on the prevalence of cartilage lesions in foals. *Equine Veterinary Journal*, 22 (6) :426–432.
- Kozlitina, J., Xing, C., Pertsemliadis, A., and Schucany, W. R. (2010). Power of genetic association studies with fixed and random genotype frequencies. *Annals of human genetics*, 74 (5) :429–438.
- Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, 22 :139–144.
- Laird, N. M., Horvath, S., and Xu, X. (2000). Implementing a unified approach to family-based tests of association. *Genet. Epidemiol.*, 19 (Suppl. 1) :S36–S42.
- Laird, N. M. and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. *NATURE REVIEWS*, 7 :385–394.
- Laird, N. M. and Lange, C. (2008). Family-based methods for linkage and association analysis. *Advances in Genetics*, 60 :219–252.
- Lampe, V., Dierks, C., Komm, K., and Distl, O. (2009a). Identification of a new quantitative trait locus on equine chromosome 18 responsible for osteochondrosis in hanoverian warmblood horses. *Journal of animal science*, 87 (11) :3477–3496.
- Lampe, V., Komm, K., Lichtner, P., Meitinger, T., and Distl, O. (2009b). *Fine mapping of quantitative trait loci (QTL) for osteochondrosis in Hanoverian warmblood horses*. PhD thesis, Hannover, Chap 8, 149-166.
- Lander, E. S. and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121 :185–199.
- Lange, C., DeMeo, D. L., and Laird, N. M. (2002). Power and design considerations for a general class of family-based association tests : Quantitative traits. *Am J Hum Genet*, 71 :1330–1341.
- Lee, G. J., Archibald, A. L., Garth, G. B., Law, A. S., Nichol森, D., Barr, A., and Haley, C. S. (2003). Detection for quantitative trait loci for locomotion and osteochondrosis-related traits in large white x meishan pigs. *Animal Science*, 76 :155–165.
- Lepeule, J., Bareille, N., Robert, C., Ezanno, P., Valette, J. P., Jacquet, S., Blanchard, G., Denoix, J. M., and Seegers, H. (2009). Association of growth, feeding practices and exercise conditions with the prevalence of developmental orthopaedic disease in limbs of french foals at weaning. *Preventive veterinary medicine*, 89 :167–177.
- Lepeule, J., Seegers, H., Rondeau, V., Robert, C., Denoix, J. M., and Bareille, N. (2011). Risk factors for the presence and extent of developmental orthopaedic disease in the limbs of young horses : Insights from a count model. *Preventive veterinary medicine*, 101 :96–106.
- Leupeule, J. (2007). *Epidémiologie Descriptive et Analytique des Affections Ostéo-Articulaires Juvéniles chez les Cheval*. PhD thesis, Thèse doctorale. Université de Rennes 1, France.
- Lewontin, R. (1964). The interaction of selection and linkage. i. general considerations ; heterotic models. *Genetics*, 49 :49–67.
- Li, J. and Jiang, T. (2005). Haplotype-based linkage disequilibrium mapping via direct data mining. *Bioinformatics*, 21 :4384–4393.

- Li, J., Zhou, Y., and Elston, R. C. (2006). Haplotype-based quantitative trait mapping using a clustering algorithm. *BMC Bioinformatics*, 7 :258.
- Li, N. and Stephens, M. (2003). Modelling ld and identifying recombination hotspots from snp data. *Genetics*, 165 :2213–2233.
- Lienasson, D. (2005). *Contribution à l'étude du traitement arthroscopique de l'ostéochondrose disséquante du relief intermédiaire du tubia distal chez le cheval : étude rétrospective sur 110 Trotteurs Français opérés en basse-normandie (1992-2002)*. PhD thesis, Université Paul-Sabatier, Toulouse.
- Luo, Z. W. (1998). Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. *Heredity*, 80 :198–208.
- Lykkjen, S., Dolvik, N. I., McCue, M. E., Rendahl, A. K., Mickelson, J. R., and Roed, K. H. (2010). Genome-wide association analysis of osteochondrosis of the tibiotarsal joint in norwegian standardbred trotters. *Animal Genetics*, 41(S2) :111–120.
- MacLeod, I. M., Hayes, B. J., Savin, K., Chamberlain, A. J., McPartlan, H., and Goddard, M. E. (2008). Power of dense bovine single nucleotide polymorphisms (snps) for genome scans to detect and position quantitative trait loci (qtl). *Genetics*, (inpress).
- Mailund, T., Besenbacher, S., and Schierup, M. (2006). Whole genome association mapping by incompatibilities and local perfect phylogenies. *BMC Bioinformatics*, 7 :454.
- Maliepaard, C., Bastiaansen, J. W. M., Calus, M. P. L., Coster, A., and Bink, M. C. A. M. (2010). Comparison of analyses of the qtlmas xiii common dataset. ii : Qtl analysis. *BMC Proceedings*, 4(Suppl 1) :S2.
- Manly, K. F., Nettleton, D., and Hwang, J. T. G. (2004). Genomics, prior probability, and statistical tests of multiple hypotheses. *Genome Res.*, 14 :997–1001.
- Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics*, 36(5) :512–517.
- Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z. S., Munro, H. M., Abecasis, G. R., and Donnelly, P. (2006). A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.*, 78 :437–450.
- Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39 :906–912.
- Martin-Rosset, W. (2001). Croissance osseuse chez le cheval. In *Actes de la 27ème Journée de la Recherche Equine*. Paris, Les Haras Nationaux, pages 73–100.
- McIlwraith, C. W. (1986). Incidence of developmental joint problems. In *Proceeding of AQHA Developmental orthopedic disease symposium*, Amarillo, USA, pages 15–20.
- McIlwraith, C. W. (2004). Developmental orthopedic disease : problems of limbs in young horses. *Journal of Equine Veterinary Science*, 24 (11) :475–479.
- McKay, S., Schnabel, R., Murdoch, B., Matukamalli, L., Aerts, J., Coppieters, W., Crews, D., Neto, E., Gill, C., Gao, C., et al. (2007). Whole genome linkage disequilibrium maps in cattle. *BMC genetics*, 8(1) :74–86.
- Meuwissen, T. H. E. and Goddard, M. E. (2000). Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics*, 155 :421–430.
- Meuwissen, T. H. E. and Goddard, M. E. (2001). Prediction of identity by descent probabilities from marker-haplotypes. *Genet. Sel. Evol.*, 33 :605–634.
- Meuwissen, T. H. E. and Goddard, M. E. (2004). Mapping multiple qtl using linkage disequilibrium and linkage analysis information and multitrait data. *Genet. Sel. Evol.*, 36 :261–279.
- Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157 :1819–1829.
- Meuwissen, T. H. E., Karlsen, A., Lien, S., Olsakerand, I., and Goddard, M. E. (2002). Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics*, 161 :373–379.
- Misztal, I., Tsuruta, S., Auvray, B., Druet, T., Lee, D. H., Ducrocq, V., Elsen, J. M., and Minvielle, F. (2002). Blupf90 and related programs. (*bgf90*).
- Mucha, S., Pszczola, M., Strabel, T., Wolc, A., Paczynska, P., and Szydlowski, M. (2011). Comparison of analyses of the qtlmas xiv common dataset. ii : Qtl analysis. *BMC Proceedings*, 5(Suppl 3) :S2.

- Onteru, S. K., Fan, B., Nikilä, M. T., Garrick, D. J., Stalder, K. J., and Rothschild, M. F. (2011). Whole-genome association analyses for lifetime reproductive traits in the pig. American Society of Animal Science, 89 (4) :988–995.
- Pagan, J. D. and Jackson, S. G. (1996). The incidence of developmental orthopedic disease on a kentucky thoroughbred farm. Pferdeheilkunde, 12 :351–354.
- Park, T. and Casella, G. (2008). The bayesian lasso. J. of the Am. Stat. Assoc., 103(482) :681–686.
- Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genetics, 2(12) :2074–2093.
- Pearce, S. G., Grace, N. D., Wichtel, J. J., Firth, E. C., and Fennessy, P. F. (1998). Effect of copper supplementation on copper status of pregnant mares and foals. Equine veterinary journal, 30(3) :200–203.
- Philipsson, J., Andreasson, E., Sandgren, B., Dalin, G., and Carlsten, J. (1993). Osteochondrosis in the tarsocrural joint and osteochondral fragments in the fetlock joints in standardbred trotters. ii. heritability. Equine Veterinary Journal, 25(S16) :38–41.
- Pieramati, C., pepe, M., Silvestrelli, M., and Bolla, A. (2003). Heritability estimation of osteochondrosis dissecans in maremmano horses. Livest. Prod. Sci., 79 :249–255.
- Pong-Wong, R., George, A. W., Woolliams, J. A., and Halay, C. S. (2001). A simple and rapid method for calculating identity-by-descent matrices using multiple markers. Genet. Sel. Evol., 33 :453–471.
- Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics, 38(8) :904–909.
- Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics, 11 :459–463.
- Pritchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans : models and data. Am. J. Hum. Genet., 69 :1–14.
- Pritchard, J. K. and Rosenberg, N. A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. Am. J. Hum. Genet., 65 :220–228.
- Pritchard, J. K., Stephens, M., and Donnelly, P. (2000a). Inference of population structure using multilocus genotype data. Genetics, 155 :945–959.
- Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000b). Association mapping in structured populations. Am. J. Hum. Genet., 67 :170–181.
- Recht, M. P., Goodwin, D. W., Winalski, C. S., and White, L. M. (2005). Mri of articular cartilage : revisiting current status and future directions. American Journal of Roentgenology, 185 (4) :899–914.
- Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., et al. (2001). Linkage disequilibrium in the human genome. Nature, 411 (6834) :199–204.
- Rejnö, S. and Strömberg, B. (1978). Osteochondrosis in the horse. ii. pathology. Acta Radiologica, Supplement 358 :153–178.
- Ricard, A. (2002). Genetic background of osteochondrosis. 55th Annu. Meet. Eur. Assoc. Anim. Prod., Bled, Slovenia, Sep, pages 5–8.
- Ricard, A. (2007). Hérité des affections ostéo-articulaires juvéniles. Pratique Vétérinaire Équine, 39 :103–110.
- Ricard, A., Perrocheau, M., Couroucé-Malblanc, A., Valette, J. P., Tourtoulou, G., Dufosset, J. M., Robert, C., Chaffaux, S., Denoix, J. M., and Guérin, G. (2010). Genetic parameters of juvenile osteoarticular conditions (joac) in french trotter. Equine Veterinary Journal, In :Press.
- Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science, 273(5281) :1516.
- Ritland, K. (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. Genet. Res. Camb., 67 :175–185.
- Sandgren, B., Dalin, G., and Carlsten, J. (1993a). Osteochondrosis in the tarsocrural joint and osteochondral fragments in the fetlock joints in standardbred trotters. i. epidemiology. Equine Veterinary Journal, S16 :31–37.
- Sandgren, B., Dalin, G., Carlsten, J., and Lundeheim, N. (1993b). Osteochondrosis in the tarsocrural joint and osteochondral fragments in the fetlock joints in standardbred trotters. ii. body measurements and clinical findings. Equine Veterinary Journal, S16 :48–53.
- Sasieni, P. D. (1997). From genotypes to genes : doubling the sample size. Biometrics, 53 :1253–1261.
- Satten, G., Flanders, W. D., and Yang, Q. (2001). Accounting for unmeasured population structure in case-control studies of genetic association using a novel latent-class model. Am. J. Hum. Genet., 68 :466–477.

- Savage, C. J., McCarthy, R. N., and Jeffcott, L. B. (1993a). Effects of dietary energy and protein on induction of dyschondroplasia in foals. Equine Veterinary Journal, 25 (S16) :74–79.
- Savage, C. J., McCarthy, R. N., and Jeffcott, L. B. (1993b). Effects of dietary phosphorus and calcium on induction of dyschondroplasia in foals. Equine Veterinary Journal, 25 (S16) :80–83.
- Schaid, D. J. (2004a). Evaluating associations of haplotypes with traits. Genetic Epidemiology, 27 :348–364.
- Schaid, D. J. (2004b). Linkage disequilibrium testing when linkage phase is unknown. Genetics, 166 :505–512.
- Schaid, D. J., Rowland, C. M., Tines, D. E., and Jacobson, R. M. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am. J. Hum. Genet., 70 :425–434.
- Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data : applications to inferring missing genotypes and haplotype phase. Am. J. Hum. Genet., 78 :629–644.
- Schober, M., Coenen, M., Distl, O., Hertsch, B., Christmann, L., and Bruns, E. (2003). Estimation of genetic parameters of osteochondrosis (oc) in hanoverian warmblood foals. In 54th Annual Meeting European Association Animal Production.
- Schougaard, H., Falk-Ronne, J., and Phillipson, J. (1990). A radiographic survey of tibiotarsal osteochondrosis in a selected population of trotting horses in denmark and its possible genetic significance. Equine Vet. J., 22 :288–289.
- Schulze, T. G. and McMahon, F. J. (2002). Genetic association mapping at the crossroads : Which test and why ? overview and practical guidelines. American Journal of Medical Genetics (Neuropsychiatric Genetics), 114 :1–11.
- Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association, 82 :605–610.
- Setakis, E., Stirnadel, H., and Balding, D. J. (2006). Logistic regression protects against population structure in genetic association studies. Genome Res., 16 :290–296.
- Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. J. Am. Stat. Assoc., 62 :626–633.
- Soller, M., Brody, T., and Genizi, A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theoretical and applied genetics, 47 :35–39.
- Souverein, O. W., Zwinderman, A. H., and Tanck, M. W. T. (2006). Multiple imputation of missing genotype data for unrelated individuals. Ann. Hum. Genet., 70 :372–381.
- Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993). Transmission test for linkage disequilibrium : The insulin gene region and insulin-dependent diabetes mellitus (iddm). The American Society of Human Genetics, 52 :506–516.
- Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet., 68 :978–989.
- Stock, K. F. and Distl, O. (2006). Genetic correlations between osseous fragments in fetlock and hock joints, deforming arthropathy in hock joints and pathologic changes in the navicular bones of warmblood riding horses. Livest. Sci., 105 :35–43.
- Stock, K. F., Hamann, H., and Distl, O. (2005). Estimation of genetic parameters for the prevalence of osseous fragments in limb joints of hanoverian warmblood horses. J. Anim. Breed. Genet., 122 :271–280.
- Stock, K. F., Hamann, H., and Distl, O. (2006). Factors associated with the prevalence of osseous fragments in the limb joints of hanoverian warmblood horses. The veterinary journal, 171 :147–156.
- Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. PNAS, 100(16) :9440–9445.
- Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theoretical population biology, 2 :125–141.
- Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E., and Visscher, P. M. (2007). Recent human effective population size estimated from linkage disequilibrium. Genome Res., 17 :520–526.
- Terwilliger, J. D. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am. J. Hum. Genet., 56 :777–787.
- Thompson, K. N., Jackson, S. G., and Rooney, J. R. (1988). The effect of above average weight gains on the incidence of radiographic bone aberrations and epiphysitis in growing horses. Journal of Equine Veterinary Science, 8 (5) :383–385.

- Thornton, T. and McPeck, M. S. (2010). Roadtrips : case-control association testing with partially or completely unknown population and pedigree structure. The American Journal of Human Genetics, 86 (2) :172–184.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, 58 :267–288.
- Trotter, G. W. and McIlwraith, C. W. (1981). Osteochondrosis in horses : pathogenesis and clinical syndromes. In American Association of Equine Practitioners, volume 27, pages 141–160.
- Uimari, P. and Tapio, M. (2011). Extent of linkage disequilibrium and effective population size in finnish landrace and finnish yorkshire pig breeds. J.Anim. Sci., 89 :609–614.
- VanGrevenhof, E. M., Schurink, A., Ducro, B. J., VanWeeren, P. R., Tartwijk, J. M. F. M., Bijma, P., and VanArendonk, J. A. M. (2009). Genetic parameters of various manifestations of osteochondrosis and their correlations between and within joints in dutch warmblood horses. J.Anim. Sci., 87 :1906–1912.
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. J. Dairy Sci., 91 :4414–4423.
- VanWeeren, P. R. (2006a). Etiology, diagnosis, and treatment of oc (d). Clinical Techniques in Equine Practice, 5 (4) :248–258.
- VanWeeren, P. R. (2006b). Osteochondrosis. Equine surgery, pages 1166–1178.
- VanWeeren, P. R. and Barneveld, A. (1999). The effect of exercise on the distribution and manifestation of osteochondrotic lesions in the warmblood foal. Equine Veterinary Journal, 31 :16–25.
- VanWeeren, P. R., Oldruitenborgh-oosterbaan, M. M. S., and Barneveld, A. (1999). The influence of birth weight, rate of weight gain and final achieved height and sex on the development of osteochondrotic lesions in a population of genetically predisposed warmblood foals. Equine Veterinary Journal, S31 :26–30.
- Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T. L., Adelson, D. L., Bailey, E., Bellone, R., et al. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. Science, 326 :865–867.
- Waldron, E. R. B., Whittaker, J. C., and Balding, D. J. (2006). Fine mapping of disease genes via haplotype clustering. Genetic Epidemiology, 30 :170–179.
- Wall, J. D. and Pritchard, J. K. (2003). Assessing the performance of the haplotype block model of linkage disequilibrium. Am. J. Hum. Genet., 73 :502–515.
- Watkins, J. P., Auer, J. P., and Stick, J. A. (1999). Osteochondrosis. In Auer, J.A. (Ed.), Equine Surgery. Saunders, Philadelphia, pages 765–778.
- Weinberg, W. (1908). Uber den Nachweis der Vererbung beim Menschen. Jahreshefte Verein f. vaterl. Naturk, in Wurttemberg, 64 :368–82.
- Weir, B. S., Hill, W. G., and Cardon, L. R. (2004). Allelic association patterns for a dense snp map. Genetic Epidemiology, 27 :442–450.
- Wigginton, J. E., Cutler, D. J., and Abecasis, G. R. (2005). A note on exact tests of hardy-weinberg equilibrium. Am. J. Hum. Genet., 76 :887–883.
- Wilke, A., Coenen, M., Distl, O., Hertsch, B., Christmann, L., and Bruns, E. (2003). The incidence of osteochondrosis in a standardbred breeding farm in quebec. In Proceedings of the 54th Annual Meeting of Europe Association for Animal Production, Rome, Italy.
- Wittke-Thompson, J. K., Pluzhnikov, A., and Cox, N. J. (2005). Rational inferences about departures from hardy-weinberg equilibrium. Am. J. Hum. Genet., 76 :967–986.
- Wittwer, C., Dierks, C., Hamann, H., and Distl, O. (2008). Associations between candidate gene markers at a quantitative trait locus on equine chromosome 4 responsible for osteochondrosis dissecans in fetlock joints of south german coldblood horses. Journal of Heredity, 99 (2) :125–129.
- Wittwer, C., Hamann, H., and Distl, O. (2009). The candidate gene xirp2 at a quantitative gene locus on equine chromosome 18 associated with osteochondrosis in fetlock and hock joints of south german coldblood horses. Journal of Heredity, 100(4) :481–486.
- Wittwer, C., Hamann, H., Rosenberger, E., and Distl, O. (2006). Prevalence of osteochondrosis in the limb joints of south german coldblood horses. J. Vet. Med. A., 53 :531–539.
- Wittwer, C., Löhring, K., Drögemüller, C., Hamann, H., Rosenberger, E., and Distl, O. (2007). Mapping quantitative trait loci for osteochondrosis in fetlock and hock joints and palmar/plantar osseus fragments in fetlock joints of south german coldblood horses. Animal Genetics, 38 :350–357.

- Wolter, R. (1996). Ostéochondrose et alimentation chez le cheval. *Prat Vet Equine*, 28 :85–96.
- Wu, C., DeWan, A., Hoh, J., and Wang, Z. (2011). A comparison of association methods correcting for population stratification in case-control studies. *Annals of Human Genetics*, 75 :418–427.
- Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. (2010). Common snps explain a large proportion of the heritability for human height. *Nature Genetics*, 42 :565–569.
- Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38 :203–208.
- Zaykin, D. V., Westfall, P. H., Young, S., Karnoub, M. A., Wagner, M. J., and Ehm, M. G. (2002). Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human Heredity*, 53 :79–91.
- Zhang, F., Wang, Y., and Deng, H.-W. (2008). Comparison of population-based association study methods correcting for population stratification. *PLoS ONE*, 3(10) :e3392.
- Zhao, H. H., Fernando, R. L., and Dekkers, J. C. M. (2007a). Power and precision of alternate methods for linkage disequilibrium mapping of quantitative trait loci. *Genetics*, 175 :1975–1986.
- Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P., and Nordborg, M. (2007b). An arabidopsis example of association mapping in structured samples. *PLoS Genetics*, 3(1) :e4.
- Zheng, G., Freidlin, B., and Gastwirth, J. L. (2006). Robust genomic control for association studies. *Am. J. Hum. Genet.*, 78 :350–356.
- Zöllner, S. and Pritchard, J. K. (2005). Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*, 169 :1071–1092.