Table des Matières

I Introduction générale.

I.1 Introduction.

Bien avant le développement de la génétique, l'homme était capable d'améliorer les plantes cultivées ou les animaux domestiques en sélectionnant pour la reproduction des individus semblables pour un caractère donné. Cette sélection s'appuie sur deux principes: (i) l'existence de disparités entre individus pour des caractères observables et (ii) la transmission de certains caractères des parents à leurs descendants. Cette transmission d'un caractère, qui a longtemps intrigué l'homme, a trouvé un début d'explication avec les premières théories génétiques proposées par Mendel. Au cours du 20e siècle, le développement des théories de transmission des caractères, combiné à nos connaissances des génomes, a donné naissance à une discipline appelée génomique statistique, qui s'intéresse aux causes génétiques impliquées dans les variations phénotypiques observées pour des caractères héréditaires. Bien qu'applicable à des différences phénotypiques inter-spécifiques (dans le cadre d'une perspective évolutive), nous nous limiterons ici à la génomique statistique portant sur les variations phénotypiques intra-spécifiques.

L'une des premières et principales approches de la génomique statistique est basée sur le concept suivant: en comparant le pattern d'hérédité d'un caractère au sein d'un pedigree avec celui des chromosomes, il est possible de découvrir les gènes impliqués sans connaissance à priori de la fonction de ces derniers. Cette approche complètement générique, n'est pas révolutionnaire et a été appliquée avec succès dès le début du 20e siècle sur des espèces expérimentales telles que la drosophile, le blé ou encore la levure[1]. Cette méthode appelée étude de liaison est la première étape du clonage positionnel. Elle est complétée par deux autres: (i) l'identification des mutations causales dans les régions mises en évidence lors de la première étape (c'est l'étape dite de cartographie fine), (ii) l'étude des fonctions cellulaires et moléculaires des gènes découverts.

Les études de liaison sont longtemps restées confinées à des espèces expérimentales et ont été très peu employées chez l'homme (ou d'autres espèces non expérimentales) pour les raisons suivantes: (i) l'impossibilité de mettre en œuvre des croisements expérimentaux et (ii) l'indigence en marqueurs génétiques. Il a fallu attendre le début des années 80 et le développement de marqueurs nucléotidiques qui de par leur nombre et leur polymorphisme élevé permettaient de suivre l'hérédité d'un segment chromosomique dans des pedigrees humains de manière analogue à des croisements expérimentaux[2]. Ces progrès, combinés à un engouement pour la cartographie génétique chez l'homme, ont conduit à la découverte de centaines de gènes impliqués dans des

maladies mendéliennes rares (OMIA). Plusieurs conclusions ont été tirées de ces résultats dont: (i) les allèles de susceptibilité de ces maladies sont le plus souvent rares et (ii) ces études peuvent échouer pour des maladies monogéniques quand il n'y a pas de correspondance simple entre un phénotype et un génotype, ce qui peut se produire en présence d'hétérogénéité génétique*, de pénétrance* incomplète ou d'expressivité variable*.

Par la suite les généticiens, désireux d'étendre le champ d'application de ces études de liaison, se sont tournés vers des maladies plus courantes, telles que le diabète, l'hypertension ou le cancer, présentant également un caractère familial mais nettement plus répandues que les maladies étudiés jusqu'alors. Malheureusement pour ce type de pathologies, les résultats obtenus via cette approche étaient la plupart du temps non convaincants, excepté pour des formes rares. Toutefois, les mutations génétiques découvertes pour ces sous-types mendéliens expliquent peu de cas dans la population[4,5]. Pour comprendre ces échecs, il est nécessaire d'introduire la notion de caractère complexe: la grande majorité des caractères médicaux (et également agronomique) ne sont pas influencés par un simple gène suivant les règles de Mendel, mais par de multiples facteurs d'ordre génétique et environnemental. Ces caractères sont dits polygéniques ou complexes. Ce concept « caractères discret courant = caractère polygénique » a été proposé par East (1910) et illustré par Altenburg[6] en découvrant plusieurs gènes impliqués dans les ailes tronquées chez la drosophile. Les modèles mathématiques reliant plusieurs *loci* à des caractères quantitatifs complexes ont été développés par Fisher (1918). Des études de liaison de caractères complexes ont été menées dès les années 80 avec succès chez des espèces expérimentales, dans lesquelles il est possible d'observer par des croisements des centaines de méioses d'un même parent[7]. Les résultats infructueux obtenus chez l'homme s'expliquent par le fait que le risque moyen relatif associé aux différents génotypes dans des maladies complexes, nécessiterait également de disposer de tels croisements ou encore d'un nombre très élevé de familles[8]. Ces problèmes de puissance statistique ont finalement poussé les généticiens à se tourner vers une autre stratégie pour étudier ce type de pathologie: plutôt que de suivre en parallèle le pattern d'hérédité d'un segment de chromosome avec celui d'un caractère au sein d'une famille, on regarde si pour un allèle particulier il existe des différences de fréquence entre des individus atteints et des individus contrôles dans la population. Cette approche, appelée étude d'association, n'est pas nouvelle: elle a d'abord été utilisée dés les années 50 dans des études de type gène candidat[9]. Dans ces dernières, on recherche des variations génétiques au sein de gènes dont on connaît ou suppose l'implication dans la maladie étudiée[10,11]. Il a été avancé qu'en étendant ce type d'approche à l'examen entier du génome, il est possible de détecter des facteurs génétiques influençant des maladies complexes avec des tailles de cohortes raisonnables. Toutefois, réaliser ce type d'étude implique d'examiner toutes les différences génétiques entre des cas et des contrôles et donc en principe de séquencer

* On parle d'hétérogénéité génétique quand différentes mutations génétiques conduisent à la même pathologie.
* La pénétrance est la portion d'individus possédant un génotype donné qui exprime le phénotype correspondant.
* De l'expressivité variable signifie que la probabilité d'être malade est une fonction de paramètre tel que l'age.

entièrement ces individus. Or au moment où ce type d'approche a été envisagé, les technologies de séquençage n'étaient pas aussi accessibles qu'actuellement.

Toutefois, en même temps, que l'idée de développer des études d'association génome entier (GWAS, pour *Genome Wide Association Studies*), une hypothèse concernant les variations génétiques influençant des maladies complexes est née: les généticiens ont supposé que ces pathologies génétiques complexes courantes étaient influencées en grande majorité par des variations génétiques courantes[12,13]. Cette hypothèse, appelée *Common Disease-Common Variants* (CDCV) s'appuient sur des arguments de génétique des populations, qui prédisent que la plupart des variations génétiques présentes dans les populations humaines sont des variations génétiques courantes. En effet suite à la brusque croissance des populations humaines, il n'existe qu'un polymorphisme toutes les 1000 paires de bases[14,15]. De plus, 90% des sites polymorphes chez un individu le sont également chez d'autres. Un autre argument en faveur de cette hypothèse de CDCV est: dans le cas de maladies mendéliennes rares la plupart des mutations impliquées sont rares, car ces dernières étant délétères, disparaissent au cours d'un processus de sélection naturelle. Cependant dans le cas de maladies courantes, le développement tardif de la pathologie et le faible impact sur la probabilité qu'un individu se reproduise (appelé valeur sélective ou *fitness*) laissent supposer un spectre de fréquences des mutations causales bien plus large que pour des pathologies mendéliennes. Enfin, cette hypothèse de CDCV est soutenue par quelques exemples comme l'allèle APOE4 et la maladie d'Azheimer ou encore l'allèle facteur V leiden et les risques de thrombose[16].

Sur base de cette hypothèse CDCV, il a été proposé de développer un catalogue des variations génétiques courantes et de caractériser la corrélation, appelée déséquilibre de liaison (DL) entre les allèles de ces différents *locus*. L'idée est de sélectionner par la suite des variations génétiques parmi ce catalogue représentant au mieux l'ensemble des variations génétiques courantes du génome et en cherchant à éviter toute redondance d'information. Ces variations génétiques sont ensuite testées pour une éventuelle association avec le caractère étudié. La plupart du temps, le *locus* de susceptibilité (supposé également comme étant une variation génétique courante) ne fera pas partie des variations génétiques testées. Toutefois, une association indirecte avec la pathologie étudiée pourra être mise en évidence grâce au DL entre ce dernier et les variations génétiques testées.

Au cours de 2007-2008, les efforts conjoints pour obtenir des cartes génétiques denses de variants communs, les progrès dans les technologies de génotypage à haut débit et l’obtention de grandes cohortes ont rendu possible les premiers progrès vers la compréhension de maladies complexes chez homme.

Cette introduction décrit d'abord les outils génétiques employés en cartographie génétique et se concentre ensuite sur l'évolution des différentes approches employées en cartographie aussi bien chez l'homme que chez les animaux domestiques.

I.2 Les Marqueurs génétiques.

Les marqueurs génétiques sont utilisés en cartographie génétique pour baliser le génome.

I.2.1 Historique des marqueurs génétiques.

I.2.1.1 Les Allozymes.

Les tout premiers marqueurs moléculaires étaient des protéines appelées allozymes (contraction des termes allèle et enzyme). Pour un même allozyme, il peut y avoir différentes versions dans la séquence en acides aminés. On peut distinguer les différentes copies d'un même allozyme par la taille et la charge lors d'une électrophorèse. On a observé un certain degré de polymorphisme de ce type de marqueur au sein de populations naturelles[17]. Ceci a permis de corroborer une théorie née ultérieurement et fondamentale en génétique des populations: la plupart des mutations sont neutres (Théorie de la sélection neutre[18,19]). À cause de leur faible coût, les allozymes ont été fréquemment employés dans des études de populations de grande taille. Cependant ils ont deux inconvénients majeurs: (i) ils sont trop peu nombreux pour être utilisés en cartographie et (ii) ils sont une représentation indirecte du polymorphisme d'une séquence nucléotidique.

I.2.1.2 Apparition des premiers marqueurs nucléotidiques.

L'avènement des techniques de manipulation de l'acide désoxyribonucléique (ADN), plus spécialement l'utilisation des endonucléases de restriction, a permis la mise au point des premiers marqueurs basés sur l'ADN: les RFLPs (*Restriction Fragment Lenght Polymorphism*). Dans une séquence d'ADN, un enzyme de restriction est capable de reconnaître un motif nucléotidique particulier, appelé encore site de restriction (le motif dépend de l'enzyme de restriction) et de cliver cette séquence en plusieurs fragments dont la taille est fonction de la position des sites de restriction. Une substitution au niveau d'un nucléotide peut apporter une modification dans la taille des fragments de restriction générés. Ces marqueurs ont permis d'établir la première carte génétique humaine[20] et de réaliser la première étude d'association[21].

Dans la même décennie, un autre type de marqueurs de nature nucléotidique hautement polymorphe a été découvert: les minisatellites[22]. Ce sont des séquences répétées en tandem dont la taille du motif unitaire varie

entre 10 et 100 paires de bases (pb). Le polymorphisme de ces séquences d'ADN vient des fluctuations du nombre de motifs. Les minisatellites sont les marqueurs de choix dans l'identification des individus (médecine légale et test de paternité) mais à cause de leur répartition génomique non uniforme (90% des séquences mini-satellites se trouvent dans des régions sub-télomériques), ils n'ont pas d'applications en cartographie.

I.2.1.3 Développement des marqueurs basés sur la PCR.

L'invention de la PCR (*Polymerase Chain Reaction*) a révolutionné les marqueurs génétiques basés sur l'ADN[23]. La PCR permet en effet d'amplifier et d'analyser une région génomique pour un grand nombre d'individus sans devoir cloner ou isoler de grandes quantités d'une séquence d'ADN donnée.

Les microsatellites ont été les premiers marqueurs à avoir tiré pleinement avantage de la PCR. Ils sont très similaires aux séquences minisatellites excepté que le motif répété est plus petit (entre 2-10pb)[24]. Ces marqueurs particulièrement polymorphes sont très abondants (plus de 100.000 microsatellites sont répertoriés dans le génome des mammifères) et sont répartis de façon aléatoire dans le génome. Grâce à leurs propriétés, ils sont devenus populaires pour la construction de cartes génétiques et dans les études d'association. Chez l'homme, le développement des microsatellites a été une étape clé pour le clonage positionnel d'un grand nombre de maladies monogéniques. Malgré le nombre important de microsatellites dans le génome, les artefacts nombreux pouvant se produire durant l'étape de PCR ont finalement empêché le développement de technologies permettant d'automatiser leur génotypage.

I.2.1.4 Arrivée des marqueurs de type SNP (Single Nucleotide Polymorphism).

L'utilisation de mutations ponctuelles dans le génome comme marqueurs existait déjà avec les RFLPs. Le développement de technologies à haut débit pour détecter ces polymorphismes n'est donc pas conceptuellement une avancée. Elles sont simplement une solution à grande échelle pour ce type de marqueurs.

Les SNPs (Single Nucleotide Polymorphism), marqueurs polymorphes (bialléliques) au niveau d'un nucléotide ont comme principaux intérêts de pouvoir être génotypés facilement et à un moindre coût et d'être extrêmement abondants (on compte un SNP pour 1000 paires de bases chez l'homme)[25]. Grâce à ces propriétés, ils sont devenus un outil précieux en cartographie fine (il faut 500,000 SNPs pour une étude de cartographie fine dans le génome humain). Par ailleurs, ils sont à l'origine du succès des études d'association génome-entier pour des maladies complexes humaines. Ils sont également très utilisés dans l'inférence de l'histoire génétique des populations. Ces marqueurs présentent cependant un certain nombre de désavantages comme d'être bialléliques

et donc peu informatifs quand ils sont pris individuellement (l'hétérozygotie maximale est de 50%). De plus, quelle que soit la stratégie mise en œuvre pour identifier des nouveaux SNPs, il existe un biais d'échantillonage systématique appelé *Ascertainment bias* dans les SNPs détectés. L'*Ascertainment bias* provient du fait que la taille de l'échantillon est souvent petite, et donc la probabilité de détecter un SNP est fonction de sa fréquence allélique. Ce biais a des conséquences importantes pour les paramètres corrélés aux spectres de fréquences alléliques (DL, Fst*, le test de Tajima*) et donc devra être pris en compte[26].

I.2.2 Développement en cours pour détecter de nouvelles sources de polymorphisme.

I.2.2.1 Copy Number Variant.

Même si les SNPs sont les marqueurs de choix dans les études d'association, il existe d'autres sources de polymorphisme dans le génome: ce sont les délétions, les insertions, les duplications ou, les translocations ou encore les réarrangements génomiques. Depuis des décennies, ce type de modifications structurales à une échelle microscopique est connu, grâce aux techniques d'hybridation fluorescente in-situ (*Fluorescent in situ hybridization* ou FISH). Avec les progrès de la biologie moléculaire et notamment avec l'avènement du séquençage, des petites altérations structurales comme des délétions et des insertions touchant quelques paires de bases ont pu être mises en évidence également. Cependant, il a fallu attendre les années 2000, pour découvrir l'existence de variations structurales à une échelle intermédiaire entre le caryotype et le nucléotide. Ces polymorphismes structuraux, qui incluent des *copy number variants* (CNV, Segment génomique de 1 kb ou plus grand, composé d'un nombre de copies variable par rapport à séquence de référence. CNVs incluent les délétions, les insertions et les duplications), ainsi que des translocations et des inversions, s'étendent sur des régions de 1kb à 3Mb.

Plusieurs exemples anciens montrent que des délétions ou des duplications de gènes peuvent affecter certains phénotypes. On sait également que bien que les CNVs soient moins abondants que les SNPs, ils représentent de par leur taille une fraction significative du génome. En effet, une étude ayant identifiée, 1447 CNVs montre qu'au moins 12% du génome humain contient des CNVs[27]. Il est donc fort probable que ces polymorphismes

* Le Fst (Indice de fixation ou *Fixation index*) est une mesure standardisée permettant d'évaluer la différenciation des populations à partir du polymorphisme génétique.

* Le test de Tajima est un test statistique permettant de déterminer si les mutations observées dans une population sont neutres ou non.

jouent un rôle important aussi bien dans des caractères mendéliens que dans des caractères complexes. Cependant, le rôle précis de ces CNVs dans chaque phénotype étudié reste encore largement incompris.

Pour des maladies monogéniques dominantes, une hypothèse avancée est que ces CNVs pourraient être l'origine d'une pénétrance incomplète. Par exemple, un individu porteur d'une mutation causale impliquant une perte de fonction dans le cadre d'une maladie dominante peut ne pas être atteint s'il a reçu de son autre parent l'allèle normal dupliqué, compensant l'expression de l'autre copie parentale défectueuse. Les CNVs peuvent donc être à l'origine de gains ou de pertes de copies normales ou défectueuses d'un gène dans une région donnée. Ces pertes ou gains pourraient également toucher des éléments régulant l'expression de gène et de cette façon affecter un phénotype. En modifiant le nombre de copies d'un gène présent chez un individu, les CNVs peuvent également affecter des caractères complexes. En effet, nos connaissances actuelles sur les caractères complexes montrent que ces derniers sont plus susceptibles à des formes de variations douces impliquant des fluctuations dans les concentrations en gène qu'à une détérioration complète de la fonction des gènes. Il existe déjà des exemples de CNVs montrant des associations avec des caractères complexes[28,29,30,31]. La Figure I.1 illustre les différentes manières dont un CNV peut affecter un phénotype.

Par ailleurs, des CNVs, qui peuvent sembler avoir des effets bénins sur un phénotype dans certaines populations, pourraient affecter ce même phénotype dans d'autres conditions par exemple en combinaison avec d'autres SNPs ou d'autres facteurs environnementaux. Utiliser des SNPs et CNVs conjointement dans des GWAS permettrait de comprendre la contribution de ces deux types de polymorphismes dans des phénotypes complexes.

À l'heure actuelle, la mise oeuvre de GWAS avec des CNVs courants reste un défi et pose davantage de difficultés que celles basées sur des SNPs courants. Différents problèmes et questions se posent à toutes les étapes conduisant à ce type d'analyse: (i) répertorier tous les CNVs courants et obtenir des informations précises sur chacun eux (position, nombre d'allèles et fréquences), (ii) obtenir des génotypes de qualité des CNVs catalogués, (iii) quels types de mesures utilisées dans une étude d'association: le nombre de copies en tenant compte de l'incertitude liée aux technologies actuelles génotypage pour ce type de variation ou transformer, avec le risque de perdre de l'information, le nombre de copies en une mesure de type gain ou perte. Pour contourner ces problèmes de mise en oeuvre de GWAS basés sur des CNVs courants, il a été proposé d'utiliser des tagSNPs pour capturer grâce au déséquilibre de liaison l'information génétique de CNVs courants. Toutefois, le problème est que DL entre SNPs et CNVs reste encore mal connu. Il existe deux raisons à cela: (i) les SNPs se trouvant dans les régions contenant des CNVs, sont davantage écartés des études d'association soit parce qu'ils ne remplissent pas les critères d'hérédité mendélienne ou encore parce qu'ils ne sont pas en équilibre Hardy-Weinberg (HWE), que des SNPs présents dans d'autres régions. (ii) On ne dispose pas jusqu'à récemment de beaucoup de CNVs courants génotypés. Le projet HapMap III, ainsi que le 1000 Genome Project devrait

permettre ces prochains mois d'avoir une idée plus précise du DL entre SNPs et CNVs.

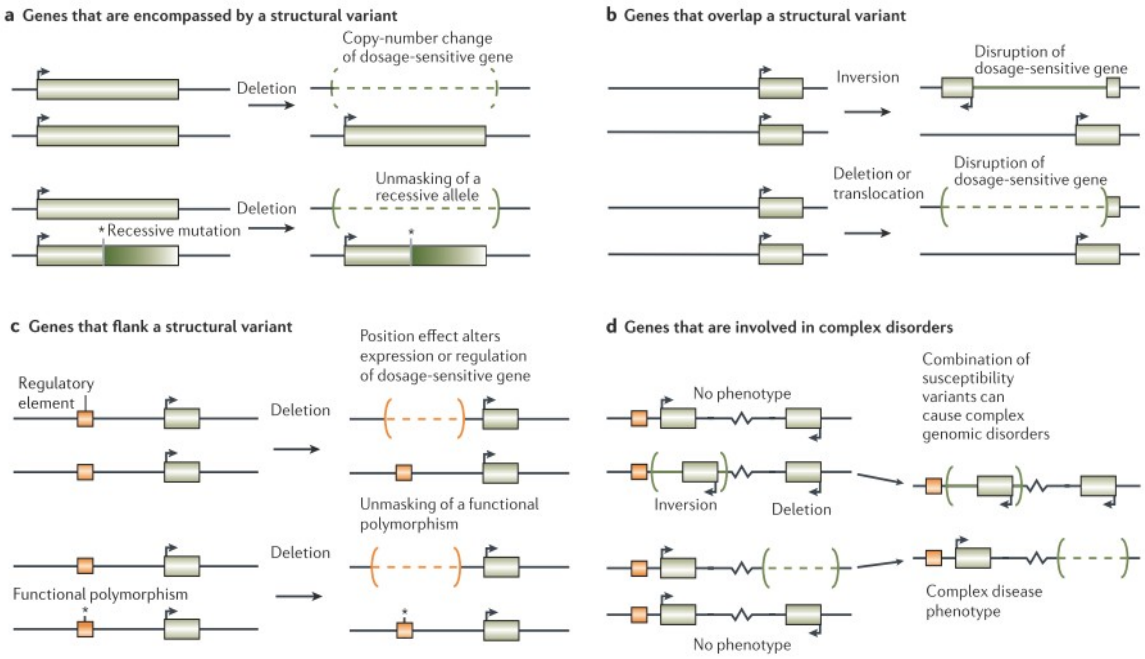


Figure I.1: *Différentes façons dont les CNV peuvent affecter un phénotype. Figure provenant de l'article de Feuk et al.*[206]

I.2.2.2 1000 Genome Project.

Actuellement, nos connaissances concernant l'effet des variations génétiques sur des phénotypes couvrent: (i) les variations génétiques rares ayant un effet majeur sur des caractères mendéliens et (ii) les variations génétiques courantes ayant un effet modéré sur des caractères complexes. Entre ces deux types de variations, on ne dispose de presque d'aucune information. L'objectif du 1000 Genome Project (http://www.1000genomes.org) est de combler ce vide en séquençant des milliers d'individus d'origine différentes, en tirant parti des améliorations survenues au cours de ces deux dernières années dans les technologies de séquençage qui ont diminué considérablement les couts. Le séquençage d'un génome consiste d'abord à découper en petits fragments d'ADN le génome d'un individu, que l'on séquence ensuite. Pour retrouver la séquence génomique, on aligne les fragments séquencés sur une séquence de référence. Si cette opération est effectuée une seule fois, il est fort probable que certaines régions ne soient couvertes par aucun fragment (ou que d'autres soient couvertes par plusieurs fragments), ce qui donne une séquence incomplète. Pour obtenir la séquence complète d'un individu, il

faut une profondeur de couverture de séquençage de 28. Les couts sont encore trop élevés pour réaliser le séquençage de milliers d'individus avec une telle couverture. Cependant, le 1000 Genome Project propose de séquencer avec une couverture de 4 des milliers d'individus d'origine différente, ce qui devrait permettre de détecter toutes les variations génétiques ayant un MAF (*Minor Allele Frequency*) > 1% (y compris CNV).

I.3 Construction de cartes de marqueurs.

Quelle que soit la méthode de cartographie employée, le généticien doit connaître la position génomique de ses marqueurs génétiques qui jouent un rôle de balise moléculaire.

I.3.1 Des cartes génétiques aux cartes physiques.

I.3.1.1 Les cartes génétiques.

Les premières cartes de marqueurs mises en place et longtemps utilisées ont été les cartes de liaison, encore appelées cartes génétiques. Elles sont nées de l'observation suivante: des allèles de *loci* proches ont tendance à être plus souvent transmis ensemble que des allèles de *loci* éloignés ou présents sur des chromosomes différents. Ainsi chez un individu double hétérozygote A1B1|A2B2, en cas de ségrégation indépendante des deux *loci* on s'attend à avoir des proportions identiques pour les quatre types de gamètes. Dans une situation où les 2 *loci* sont très proches l'un de l'autre, ils sont transmis ensemble sauf en cas de recombinaison. On peut alors constater une distorsion dans les fréquences haplotypiques attendues sous l'hypothèse d'indépendance, où deux types de gamètes correspondant aux recombinants peuvent être manquants. Dans des situations intermédiaires, la proportion de recombinants reflétera la distance entre les deux *loci*.

I.3.1.1.a Estimations des taux de recombinaisons.

Toutes les constructions de cartes génétiques commencent par des analyses de liaison deux points, dans lesquelles on cherche à calculer les taux de recombinaison θ entre toutes les paires de marqueurs. Pour estimer les taux de recombinaison entre deux marqueurs, on calcule pour chaque famille appartenant à un pedigree un lodscore $Z_i(\theta)$ dans lequel on compare la vraisemblance des données (L) pour une valeur θ donnée $0 \leqslant \theta < 0.5$ (Hypothèse H1 de liaison), par rapport à la vraisemblance des données pour une valeur $\theta = 0.5$ (Hypothèse H0 d'indépendance des deux marqueurs):

$$Z_i(\theta) = \log_{10} L*(\theta) = \log_{10}[L(\theta)/L(1/2)] \quad [32]$$

Pour calculer θ pour l'ensemble du pedigree, on somme les lodscores de chaque famille pour une valeur de θ donnée:

$$Z(\theta)=\sum_{i=1}^{F} Z_i(\theta)$$

On prend la valeur de θ donnant le lodscore le plus élevé. On considère qu'on a de la liaison génétique entre deux marqueurs lorsque $Z_{max}>3$, pas de liaison quand $Z_{max}<-2$ et on ne sait pas conclure dans un sens comme dans un autre quand $-2<Z_{max}<3$.

Pour calculer ces lodscores, il faut d'abord calculer la vraisemblance des données en fonction de θ. Cette expression dépend de la structure du pedigree (backcross, intercross, half-sib design...), des génotypes et des haplotypes connus. La vraisemblance des données peut s'écrire de la façon générale suivante:

$$L=P(x)=\sum_{g} P(x,g)=\sum_{g} P(x|g)\,P(g)$$ [32]

où g correspond au vecteur de génotypes de phase connus et x au vecteur des phénotypes. Les détails pour calculer la vraisemblance des données sont abordés plus loin dans la section concernant les études liaison.

I.3.1.1.b Ordonner les marqueurs[33].

Quand le nombre de marqueurs est supérieur à 2, il faut déterminer l'ordre des marqueurs sur la carte établie. Ceci nécessite de calculer la vraisemblance pour tous les ordres possibles, en choisissant les valeurs θ qui maximisent la vraisemblance. On garde l'ordre donnant la vraisemblance la plus élevée. Le problème est que quand le nombre de marqueurs augmente, il devient rapidement impossible d'explorer tous les ordres. En effet pour x marqueurs le nombre d'ordres est de $\frac{x!}{2}$ (10 marqueurs génèrent 1, 814, 400 ordres). Il est donc nécessaire de générer un nombre limité d'ordres approximatifs avant de procéder à des analyses de maximum de vraisemblance pour l'ordre définitif. Il existe un grand nombre de méthodes pour générer ces ordres approximatifs.

I.3.1.1.c Conversion des taux de recombinaison en distance génétique.

Les taux de recombinaison ne peuvent pas être utilisés directement comme distance génétique, car ils ne sont pas additifs. En effet si on a trois marqueurs dont les taux de recombinaison sont de 30% dans chaque intervalle

marqueur, le taux de recombinaison entre les marqueurs les plus éloignés n'est pas de 60%, car un taux de recombinaison de ne peut pas dépasser le seuil de 50%. On préfèrera utiliser le nombre de *crossing-over* (CO) entre deux *loci* comme mesure de distance *inter-loci*. Cependant le nombre de COs entre deux *loci* n'est pas directement estimable à partir des taux de recombinaison. En effet quand la distance *inter-loci* augmente, il est possible d'avoir plusieurs COs qui, selon que leur nombre soit pair ou impair, seront comptabilisés, soit comme recombinants, soit comme non recombinants. Toutefois quand la distance *inter-loci* devient petite, il est peu vraisemblable d'avoir de multiples COs, le pourcentage de recombinaison peut alors coïncider avec le pourcentage de CO. La mesure génétique couramment utilisée est le Morgan (x). Un Morgan correspond au nombre moyen de CO entre deux *loci* par gamète par chromosome. Pour avoir une mesure de la distance génétique entre deux *loci*, il faudra convertir, via des fonctions de cartographie, le pourcentage de recombinaison entre deux *loci* en distance génétique qui reflétera le nombre de CO moyen attendu. Il existe plusieurs fonctions de cartographie. Une des fonctions les plus simples et les plus utilisées est la fonction de Haldane. Elle suppose que les COs se produisent aléatoirement et indépendamment des uns des autres. À partir de cette hypothèse, la distribution du nombre de COs (n) entre deux *loci* peut se modéliser avec une distribution de Poisson de moyenne m:

$$f(n)=\frac{e^{-m}m^{n}}{n!}$$

Comme la distance d est définie comme le nombre moyen m de CO par chromatide dans un intervalle donné, le nombre moyen de COs par tétrade est donc 2d. On peut en déduire à partir de la formule ci-dessus que la probabilité d'avoir zéro CO est: $f(0)=e^{-2d}$. Comme toutes les méioses avec au moins un CO produisent 50% de gamètes recombinants et 50% de gamètes non recombinants, en prenant la probabilité d'avoir au moins un CO (l'événement complémentaire de zéro CO), on obtient:

$$\theta=\frac{1}{2}(1-e^{-2d})$$

ou

$$d=\frac{-1}{2}\ln(1-2\theta)$$

Beaucoup d'études ont montré que lorsqu'un CO se produisait dans une région donnée d'un chromosome, cela pouvait affecter le fait d'avoir un autre CO dans les régions adjacentes. Ce phénomène est appelé interférence. Il est possible de l'estimer en comparant la fréquence des doubles recombinants attendus sous l'hypothèse que les CO sont des événements indépendants avec la fréquence des doubles recombinants observés de la façon

suivante:

$$I = 1 - c = 1 - \frac{nb\ double\ recombinants\ observés}{nb\ double\ recombinants\ attendus}$$

Où c est le coefficient de coïncidence.

Si on a 3 *loci* dans cet ordre A, B et C le taux de recombinaison entre A et C s'écrit de la façon suivante en tenant compte de l'interférence: $r_{AC} = r_{AB}(1 - r_{BC}) + r_{BC}(1 - r_{AB}) = r_{AB} + r_{BC} - 2\mathrm{c}r_{AB}r_{BC}$

Si l'on cherche une fonction de cartographie telle que $f(d) = r$, en tenant compte de la relation précédente pour les taux de recombinaison et que l'on suppose que cette fonction peut satisfaire la relation suivante:

$$f(d+h) = f(d) + f(h) - 2\mathrm{c}f(d)f(h)$$

Si l'on travaille avec de petites distances, on peut poser: $r = f(d) = d$ ainsi que $\frac{f(h)}{h} = 1$. En divisant par h l'expression précédente, on obtient l'équation différentielle suivante:

$$f'(d) = \frac{f(d+h) - f(d)}{h} = 1 - c_0 r$$

c_0 est le coefficient dit marginal que l'on doit différencier de c car il correspond à une situation de deux intervalles marqueurs proches de zéro. Pour avoir une valeur de c_0 qui augmente avec r, qui est nulle quand r=0 et qui est égale à 1 quand r=1/2, Kosambi a proposé de poser $c_0 = 2\mathrm{r}$. En intégrant l'expression précédente, il a obtenu la fonction de cartographie suivante:

$$d = \frac{1}{4}\ln\frac{1+2\mathrm{r}}{1-2\mathrm{r}}$$

Un inconvénient avec la fonction de Kosambi comparée à celle d'Haldane est que les distances ne sont pas additives. Toutefois la fonction de Kosambi est la fonction de cartographie la plus utilisée chez les mammifères.

I.3.1.2 Les cartes physiques à basse résolution: les cartes d'hybrides irradiés.

L'ordre des marqueurs dans les cartes de liaison est bien souvent ambigu, quand les marqueurs sont très proches ou très éloignés. De plus les cartes génétiques sont très sensibles aux erreurs de génotypage. La détermination d'un ordre correct peut être primordiale dans certaines études de cartographie génétique en particulier celles exploitant le déséquilibre liaison. Le développement des hybrides irradiés a permis d'affiner l'ordre des *loci*.

I.3.1.2.a Construction d'hybrides irradiés.

Un hybride irradié est une cellule hôte (généralement une cellule de rongeur) contenant des fragments chromosomiques provenant d'un autre organisme (le donneur). Cette technologie est basée sur l'observation que des cellules humaines exposées à des rayons X subissent des cassures chromosomiques aléatoires. Ces fragments chromosomiques peuvent ensuite être propagés, si les cellules irradiées sont fusionnées avec des cellules de rongeur non irradiées. On teste ensuite les hybrides irradiés en les mettant dans un milieu sélectif, où seules les lignées cellulaires ayant intégré des fragments chromosomiques du donneur, porteurs d'un élément permettant la survie dans ces conditions peuvent continuer à se multiplier. Ces hybrides irradiés sont ensuite génotypés pour des marqueurs spécifiques du donneur. L'idée sous-jacente aux cartes d'hybrides irradiés est qu'il est moins vraisemblable que l'irradiation induise une cassure entre deux marqueurs relativement proches qu'entre deux marqueurs plus éloignés. Contrairement aux cartes génétiques, il n'est pas nécessaire de travailler avec des marqueurs polymorphes. On distingue selon le nombre de chromosomes fragmentés intégrés dans la cellule receveuse deux types d'hybrides irradiés: les hybrides irradiés haploïdes (un seul chromosome) et les hybrides irradiés diploïdes (ou polyploïdes, ou génome entier) (plusieurs chromosomes). L'avantage des hybrides irradiés diploïdes est qu'un jeu unique d'hybrides peut être utilisé pour cartographier tous les chromosomes du donneur. Cette approche est donc moins laborieuse que la première où un panel séparé d'hybrides doit être construit pour chaque chromosome. L'inconvénient est que quand un marqueur est présent, il n'est pas possible de savoir si il est présent en une ou deux copies, contrairement aux hybrides irradiés haploïdes, où un marqueur peut-être soit absent, soit présent, mais en une seule copie. On utilise le pattern de rétention des différents clones d'hybrides irradiés pour déterminer l'ordre et la distance entre *loci*.

I.3.1.2.b Estimation de la distance physique à partir de données d'hybrides irradiés.

En cartographie d'hybrides irradiés, la distance entre *loci* s'exprime en centirays. Cette mesure représente la probabilité de séparation par cassure pour une dose d'irradiation donnée. Celle-ci donne une meilleure estimation des distances physique que la distance génétique car la vulnérabilité aux cassures semble constante quelle que soit la position sur les chromosomes. Comme les cassures se produisent de manière aléatoire le long des chromosomes, ce processus peut être modélisé avec une distribution de Poisson. Pour n'importe quelle paire de *loci*, la probabilité d'au moins une cassure θ et la distance physique δ peuvent être reliées par:

$$1-\theta = e^{-\lambda \delta}$$

,où λ dépend de la dose d'irradiation et donc δ et λ ne peuvent être séparés lors de l'estimation. Cette fonction est similaire à la fonction de Haldane utilisée dans les cartes génétiques: δ peut-être interprété comme

le nombre de cassures chromosomiques attendues par hybrides entre deux *loci*. En dehors des cassures chromosomiques, il est nécessaire de prendre en compte que différents fragments chromosomiques peuvent avoir des probabilités de rétention différentes même s'il est souvent admis que les différents segments sont retenus de manière indépendante. La détermination de l'ordre des *loci* et des distances *inter-loci* peut devenir complexe si on considère que la rétention des fragments chromosomiques est non homogène. Par exemple pour un simple hybride irradié haploïde avec deux marqueurs, les probabilités des quatre patterns, (1,1) les deux marqueurs sont présents (1,0) le premier marqueur est présent, mais pas le second (0,1) le second marqueur est présent, mais pas le premier et (0,0) ni l'un ni l'autre des marqueurs n'est présent peuvent s'écrire de la façon suivante:

$$p_{11} = \theta P_A P_B + (1-\theta) P_{AB}$$
$$p_{10} = \theta P_A (1-P_B)$$
$$p_{01} = \theta P_B (1-P_A)$$
$$p_{00} = \theta (1-P_A)(1-P_B) + (1-\theta)(1-P_{AB})$$

où P_A, P_B et P_{AB} sont les probabilités qu'un hybride retienne un fragment avec le marqueur A uniquement, avec le marqueur B uniquement et avec à la fois avec les marqueurs A et B. Quand les probabilités de rétention varient d'un fragment à l'autre, le nombre de paramètres impliqués augmente de manière quadratique avec le nombre de marqueurs examinés. Bien qu'un certain nombre de méthodes de calcul peuvent être utilisées pour de tels problèmes[34], ceci peut poser de sérieux problèmes d'optimisation. Si le taux de rétention est supposé constant, les calculs peuvent être simplifiés. Si ceci n'affecte pas la capacité à retrouver l'ordre correct des *loci*, cela peut biaiser l'estimation des distances.

I.3.1.2.c Détermination de l'ordre des *loci* à partir de données d'hybrides irradiés.

Pour déterminer l'ordre des marqueurs en cartographie d'hybrides irradiés une manière simple et élégante est de rechercher l'ordre donnant le moins de cassures obligées. Pour un ordre spécifique des différents *loci* x = (x1, x2, ..., xm), ou xi est 1 si le marqueur est présent et 0 s'il est absent, on peut compter le nombre de changements de 1 à 0 et 0 à 1 dans chaque hybride. On peut calculer ensuite le nombre total de cassures impliquées pour un ordre spécifique en sommant sur l'ensemble des hybrides. On garde l'ordre qui minimise le nombre total de cassures obligées à travers tous les clones. Le problème de cette méthode est qu'elle ne donne ni d'estimation de distance entre *loci*, ni de comparaison de vraisemblance pour des ordres comparables. Une approche alternative est de construire un modèle pour les hybrides irradiés observés et d'estimer les paramètres de ce modèle par des méthodes de maximum de vraisemblance. De manière analogue aux cartes génétiques, le problème de cette approche est qu'il devient rapidement impossible d'évaluer la vraisemblance pour tous les ordres possibles. Il est alors indispensable de travailler avec des méthodes qui n'examinent qu'un nombre limité d'ordres[34].

L’intérêt des hybrides irradiés a rapidement diminué, car la séquence génomique d'un grand nombre d’espèces est devenue progressivement disponible.

I.3.1.3 Les premières cartes physiques à haute résolution: assemblage de contigs.

Les premières cartes physiques à haute résolution consistait à assembler des bouts de séquences d'ADN, appelé contig. La distance entre ces bouts de séquences d'ADN etait exprimée en kb de base. Ces cartes permettaient de positionner et d'établir un ordre pour des marqueurs polymorphes qui pouvaient être utilisés ultérieurement dans des études de cartographie génétiques. Ces cartes physiques ont également constitué la première étape vers l'assemblage complet de la séquence génomique des chromosomes.

Ces bouts de séquences d'ADN étaient obtenus par digestion partielle via des enzymes de restriction de la séquence génomique étudiée. Ces fragments d'ADN étaient partiellement chevauchants. Ces derniers étaient ensuite clonés par un système de BAC (*Bacterial artificial chromosome*) ou de YAC (*Yeast artificial chromosome*), permettant de mettre en place une banque génomique.

Il existait deux approches pour assembler un contig: (i) une approche dite ordonnée ou dite de marche sur le chromsome (*chromosome walking)*. Avec cette technique, on partait d'une région génomique donnée dans laquelle on identifiait tous les clones chevauchants de cette région génomique. On se servait ensuite de la séquence génomique de l'extrémité d'un des clones pour aller pécher les clones chevauchants de la région génomique adjacente et ainsi de suite. Pour rechercher les clones chevauchant une région génomique il existait deux techniques: (i) par hybridation. On fabriquait une petite sonde radioactive, si elle correspondait à une séquence non répétée, elle allait alors s'hybrider avec un nombre de clones limité. Une autre technique consistait à comparer les profils de restriction.

Quelle que soit la technique utilisée, il pouvait rester des trous après assemblage du contig. Une manière de procéder pour compléter ces trous etait de séquencer l'extrémité de certains clones ce qui permettait de définir des étiquettes, appelées STS (*sequence tagged site*), qui servaient soit à combler ces trous soit à identifier les clones chevauchant de la région génomique adjacente. (ii) Une approche dite désordonnée. Avec cette approche, les clones étaient pris au hasard et alignés soit en comparant les profils de restriction, soit en criblant une banque génomique avec des milliers de sondes réparties de manière aléatoire dans tout le génome, pour générer des ilots de clones chevauchants reliés par la suite.

I.3.1.4 Obtention de la séquence génomique entière: vers une cartographie physique complète.

I.3.1.4.a Le projet génome humain.

La séquence ADN complète d'un organisme constitue la carte physique ultime. Au début des années 90, un grand projet appelé *Human Genome Project* (HGP) a été lancé en vue d'obtenir la séquence génétique complète du génome humain et d'identifier et de cartographier les 25,000 gènes du génome humain. Le HGP, fruit d'une collaboration internationale s'est étalé sur un peu plus 10 ans et a aboutit à la mise en disposition dans des bases de données publiques de la séquence génomique complète de l'homme. Cette période a été marquée par une compétition âpre avec une société privée, Celera à partir de 1998, qui proposait de séquencer complètement le génome humain en trois ans avec une stratégie différente que celle qui était mis en oeuvre par le consortium: dans le HGP une approche de type séquençage ordonné a été adopté, tandis que Celera s'est basé sur une approche de type séquençage en vrac (ou *shotgun*). La distinction entre les deux types d'approches provient du fait que dans le séquençage ordonné il existe une étape dans laquelle on construit une carte physique de contigs du chromosome (on cherche des clones chevauchants (avec un minimum de recouvrement) ordonnés qui représentent la séquence entière du chromosome). On séquence ensuite ces contigs en les fragmentant en séquence de 500-800 paires bases. Dans l'approche en vrac, on fragmente directement la séquence d'un chromosome en petits fragments. La séquence entière est obtenue via des algorithmes utilisant la séquence de ces petits fragments. Finalement, la compétition entre les deux protagonistes s'est soldée par un match nul et la séquence brute du génome humain a été publiée en février 2001[35]. D'autres espèces, que l'homme ont été séquencées et annotées au niveau de la séquence: on dispose actuellement de la séquence de plus d'une cinquantaine d'espèces animales (http://www.ensembl.org/info/about/species.html)

I.3.1.4.b Les nouvelles technologies de séquençage à haut débit.

Depuis 2004, le séquençage des génomes s'est accéléré grâce à la commercialisation de nouvelle plateforme de séquençage à haut débit. Ces dernières ont permis de réduire considérablement les couts de séquençage par rapport aux plateformes de séquençage haut débit par capillaires. Ceci grâce à: (i) une simplification des étapes précédant le séquençage (notamment grâce à l'amplification des fragments amplifiés sélectivement par PCR) . (ii) et à la production extrêmement élevée de séquences courtes (appelé read en anglais) à la fin d'un cycle de séquençage: des centaines de milliers (454/FX) voir des des dizaines de millions sont séquencés de fragments

génomiques à la fin d'un cycle de séquençage (Illumina et SOLiD) alors que les anciennes technologies permettaient d'obtenir la séquence de seulement de 96 fragments simultanément.

Il existe 3 plateformes de séquençage, nouvelles générations très répandues actuellement:

(i) Roche/454 FLX qui allie différentes technologies: le pyroséquençage, les technologies des plaques en fibre optique picotitré, la PCR en émulsion (emPCR), ainsi que des technologies informatiques de pointe pour l'acquisition, le traitement et l'analyse des images.

(ii) Illumina Genome Analyzer est basé sur l'incorporation réversible de nucléotides fluorescents et par lecture optique de la fluorescence .

(iii) SOLiD (développé par ABI) emploie une technique de séquençage par ligation.

Ces nouvelles technologies de séquencage ouvrent de nouvelles perspectives dans une multitude de domaine: une des applications et déjà mis oeuvre à travers le 1000 Genome project est d'identifier et de cataloguer toutes les variations génétiques humaines en allant des SNPs jusqu'au CNVs.

I.3.2 Les cartes d'haplotypes: le projet HapMap.

Comme souligné plus haut, les généticiens avaient émis l'hypothèse dans les années 90, que la prédisposition à des maladies complexes était due à des polymorphismes courants (hypothèse CDCV, voir section 1 de l'introduction). Pour mettre en œuvre des GWAS, dans lesquelles des associations entre une pathologie complexe et des variations sont testées, il était nécessaire d'établir un catalogue le plus exhaustif possible des variations génétiques courantes. Le projet HapMap a été lancé pour répondre précisément à ce besoin.

I.3.2.1 Le déséquilibre de liaison.

Les études d'associations de ces variations génétiques courantes sont facilitées par le fait que les individus porteurs d'un allèle au niveau d'un site SNP donné, sont très souvent porteurs des mêmes allèles au niveau de sites SNPs voisins. Cette corrélation entre allèles de différents *loci* est appelée déséquilibre de liaison (DL ou *linkage disequilibrium* en anglais).

I.3.2.1.a Mesure du déséquilibre de liaison.

La manière la plus simple de définir du DL entre deux allèles i et j au niveau de deux *loci* différents est de poser:

$$D_{ij} = f_{ij} - f_i f_j$$

, où f_{ij} est la fréquence de l'haplotype*ij et f_i et f_j sont les fréquences des allèle i et j respectivement.

En cas d'équilibre de liaison: $D_{ij} = 0$

Toutefois quand on souhaite comparer le DL entre des paires de *loci*, cette mesure, appelée coefficient de déséquilibre de liaison souffre du fait que l'amplitude des valeurs prises dépende des fréquences alléliques.

C'est pour cette raison que Lewontin[36] a proposé de normaliser cette mesure de la façon suivante:

$$D'_{ij} = \frac{D_{ij}}{D_{min}} si\ D_{ij} < 0 \quad \text{ou} \quad D'_{ij} = \frac{D_{ij}}{D_{max}} si\ D_{ij} > 0$$

La valeur la plus faible que peut prendre D, Dmin correspond à la valeur la moins négative entre:

$$-f_i f_j \quad \text{et} \quad -(1-f_i)(1-f_j)$$

La valeur la plus élevée que peut prendre D, Dmin correspond à la plus grande valeur entre:

$$f_i(1-f_j) \quad \text{et} \quad (1-f_i) f_j$$

$D' = 1$ quand au moins un des 4 haplotypes possibles est absent.

On peut généraliser cette mesure à des *loci* ayant un nombre d'allèles > 2 de la façon suivante:

$$D' = \sum^{i} \sum^{j} f_{ij} |D'_{ij}|$$

,où f_{ij} est la fréquence de l'haplotype ij dans la population.

Une autre mesure de DL fréquemment utilisée est le r^2. Elle correspond au coefficient de corrélation entre deux allèles de deux *loci* bialléliques et s'écrit de la façon suivante:

$$r^2 = \frac{D^2}{f_i(1-fi) f_j(1-f_j)}$$

$r^2 = 1$ quand au moins deux des haplotypes possibles sont absent. Un des avantages de cette mesure de DL est que l'on sait comparer les valeurs observées à la distribution théorique sous une hypothèse d'équilibre liaison:

$$E(r^2) = \frac{1}{4N_e \theta + 1}$$

* Un haplotype est une combinaison d'allèles présents au niveau de *loci* différents d'un même chromosome.

, où Ne* est l'effectif efficace de la population et θ la distance génétique entre deux *loci*.

Cette mesure a également été généralisée à des *loci* ayant un nombre d'allèles > 2:

$$r^2 = \sum^{i} \sum^{j} f_{ij} r_{ij}^2$$

Ces mesures ne disent cependant pas si la corrélation entre deux allèles de *loci* différents est du à un simple effet du au hasard ou est causé par d'autre facteurs. Si les données sont phasées, on pourra tester l'indépendance des allèles entre deux *loci* avec test exact de Fisher (ou en utilisant une approximation du test exact de Fisher de type Monte-Carlo). Si les données ne sont pas phasées, on pourra alors employer des méthodes de type maximum de vraisemblances, dans lesquelles un des paramètres inconnus sera la fréquence des haplotypes.

I.3.2.1.b Origine du déséquilibre de liaison.

Dans une grande population panmictique, dans laquelle on suppose qu'il n'y a pas d'immigration, de mutation, de sélection et pas de dérive génique, les différents allèles au niveau de différents *loci* sont en équilibre de liaison. Si cette population s'écarte pour une raison ou une autre des ces caractéristiques du DL se créera. Si la cause à l'origine du DL disparait alors le DL décroisera à un taux qui dépendra de la distance génétique entre deux *loci*:

$$D_{AB}(t+1) = (1-\theta) D_{AB}(t)$$

, où $D_{AB}(t)$ et $D_{AB}(t+1)$ correspondent au niveau de DL entre deux *loci* AB à la génération t et t+1 respectivement et θ est la distance génétique entre A et B.

Même pour deux *loci* sont non liés, le DL ne diminuera que de moitié à la génération suivante.

Si le DL diminue à cause de la recombinaison quelles sont les raisons à l'origine de la création de DL. Les deux raisons principales à l'origine du DL touchant l'ensemble du génome sont l'introduction d'une manière ou d'une autre de dérive génique ou de mutations. Il est possible d'expliquer l'effet de la dérive génique sur le DL en reprenant la valeur attendue de r^2 qui sera inversement proportionnelle au taux mutation et à l'effectif efficace de la population. Le corollaire est que pour une distance génétique donnée on peut prédire l'effectif efficace d'une population. Sachant que des paires de *loci* proches donneront une information sur l'Ne remontant plus loin dans le passé qu'une paire de *loci* éloignés, il est possible d'inférer sur l'histoire de l'Ne d'une population sur base du DL. Par exemple chez l'homme des observations ont montré que du DL plus élevé qu'attendu sous un modèle EL (équilibre de liaison) pour des distances génétiques relativement longue. Ceci laisse supposé qu'un goulot

* L'effectif efficace d'une population observée est le nombre d'individus dans une population théorique panmictique ayant la même dérive génique que la population observée.

d'étranglement (*bottleneck* en anglais) s'est produit récemment dans les populations humaines.

Une autre façon de créer du DL est d'introduire des mutations dans une population: par exemple par immigration d'individus provenant d'une population où les fréquences alléliques sont différentes de la première population. Cet effet est simple à comprendre dans un cas extrême: si l'on suppose deux populations au niveau de deux *loci* différents et où les allèles sont fixés pour ces deux populations (population 1 AB et population 2 ab) et qu' un mélange se produit, on aura alors deux haplotypes: AB ab et ce qui amènera à un D'=1 alors que les deux populations de départ sont en équilibre de liaison.

Il existe d'autres sources à l'origine DL. L'une d'elles les plus citées est la sélection. Cependant ce facteur contrairement aux deux précédents affecte seulement les niveaux de DL de manière locale, c'est-à-dire les *loci* influencés par la sélection.

I.3.2.2 Définition du projet HapMap.

Le projet HapMap a démarré en 2002 avec comme but de développer un catalogue des variations génétiques communes dans différentes populations chez l'homme pour l'ensemble du génome et de caractériser le déséquilibre de liaison de la façon la plus précise possible entre ces polymorphismes.

L'objectif du HapMap I[37] était d'avoir un SNP tout les 5kb (pour le HapMap II 1 SNP par kb[38] et encore étendu dans le HapMap III[39] à des SNPs ayant un MAF < 5% et à des CNVs ayant MAF > 5%) pour chacun des 269 échantillons provenant de 4 populations différentes (dans le HapMap III l'étude a été étendue à 1184 échantillons provenant de 11 populations différentes). Dans le HapMap I seuls les SNP avec MAF > 0.05 (1 millions par populations) ont été ciblés tandis que le HapMap II et HapMap III ont intégré des SNP supplémentaires sans tenir compte de leur MAF (2.1 millions de SNPs supplémentaires par population dans le HapMap II). Pour évaluer ce catalogue de SNPs à travers le génome par rapport à une base de données exhaustive de polymorphismes, 10 régions de 500 kb ont été choisies dans le projet ENCODE*[40] (ENCyclopedia Of DNA Elements) et complément reséquencées.

I.3.2.3 Caractérisation du déséquilibre de liaison dans le génome humain.

Ces données ont permis d'étudier les propriétés du DL dans le génome humain. Il en est ressorti que (i) dans les régions sans recombinaison (D'=1, pour toutes les paires de marqueurs), seules les mutations sont responsables des variations du r^2 (r^2 est égal à 1 quand deux SNP se produisent sur la même branche de l'arbre généalogique,

* Le but du projet ENCODE était de trouver tous les éléments fonctionnels du génome humain.

et qu'il n'y a pas de recombinaison entre ces SNPs). Les fluctuations de r^2 observées dans ces régions proviennent non pas de la distance physique entre ces marqueurs, mais de l'ordre historique dans lequel ces SNP se sont produits. Malgré cela, il y a très peu d'haplotypes observés dans ces régions et il est tout à fait possible de prédire l'ensemble des haplotypes avec peu de SNPs. Par exemple dans une région comprenant 36 SNPs (liste exhaustive), on trouve seulement 7 haplotypes dont les 5 plus importants sont différentiables par seulement 4 SNPs. (ii) Dans les autres régions, les recombinaisons jouent un rôle déterminant dans les variations du DL et ont tendance à se concentrer dans des petites régions de quelques que kb appelées hotspot de recombinaison (80% des recombinaisons se produisent dans 15% de la séquence). (iii) Ces hotspots de recombinaison interrompent la structure en blocs haplotypiques du génome. Les blocs haplotypiques sont des segments chromosomiques où les niveaux de DL entre *loci*, indépendamment de la distance sont très élevés et dont les limites sont définies par des hotspots de recombinaisons. Beaucoup de SNPs sont dans des blocs haplotyques. (iv) Les blocs haplotypiques ne sont pas toujours clairement définis: la plupart des haplotypes s'arrêtent au niveau des hotspot de recombinaison identifiés alors que d'autres peuvent continuer au-delà. Par ailleurs, il existe des variations importantes dans la taille physique des blocs haplotypiques dans les génomes.

Cette architecture du DL dans le génome humain montre qu'il est possible de mettre en évidence un allèle de susceptibilité pour une maladie de manière indirecte, s'il est en déséquilibre de liaison avec un SNP ou un haplotype dont les effets sur la maladie ont été testés directement. À partir des données ENCODE, on estime que trois SNPs sur cinq sont en DL parfait ($r^2=1$) avec un SNP présent dans la même région qu'eux. Cependant pour trouver une association de manière indirecte, il n'est pas nécessaire d'avoir une corrélation parfaite. En effet si on double la taille des cohortes et que l'on teste des SNPs avec un $r^2=0.5$ avec la mutation causale, on garde la même puissance statistique que dans une situation où l'on a un $r^2 =1$. Le Hapmap II montre que pour capturer la grande majorité (> 95%) des variations génétiques courantes (SNPs > 5%) avec un $r^2 > 0.8$, il faut 500,000 et un million de SNPs respectivement dans les populations non Africaines et Africaines pour l'ensemble des haplotypes. Cette différence entre les deux populations vient du fait que le DL s'étend sur des distances plus longues dans les populations Caucasiennes et Asiatiques que dans les populations Africaines, conséquence du goulot d'étranglement qu'on subit leurs ancêtres ayant émigré d'Afrique.

En choisissant les SNPs qui représenteront l'ensemble des haplotypes présents dans le génome (appelé tag SNP) à génotyper parmi le catalogue proposé par HapMap dans le cadre d'études d'association, il est possible de capturer une grande majorité de la variation génétique courante des cohortes testées. Il existe de multiple approches pour choisir ces tag SNPs.

I.4 Des études de liaison aux études d'association.

I.4.1 Les études de liaison.

I.4.1.1 Introduction-Définition.

Une fois que le généticien dispose de marqueurs génétiques, ainsi que de leur position génomique, il peut regarder si, au sein d'un pedigree, des individus ayant hérité d'un phénotype similaire reçoivent également les mêmes haplotypes. S'il observe une telle coségrégation, il peut tenter de déterminer via les positions de ces marqueurs génétiques la localisation de la région génomique (voir d'un gène) impliquée dans le phénotype étudié. Cette approche est appelée étude de liaison (voir Figure I.2).

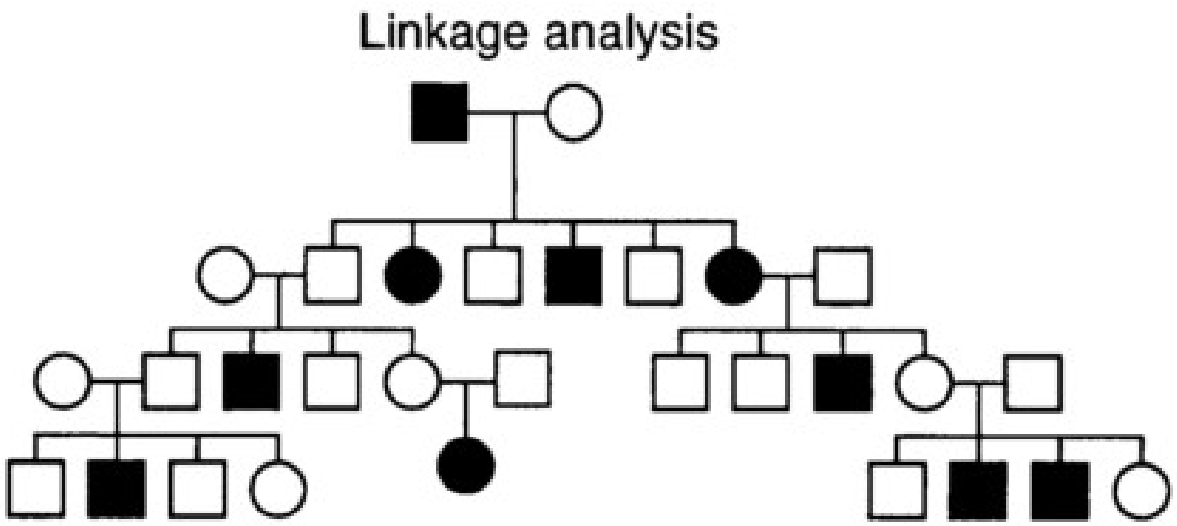


Figure I.2: *Une étude de liaison consiste à regarder dans un pedigree, si des individus ayant hérité d'un phénotype similaire (ici une pathologie), ont également reçu des allèles marqueurs similaires. Dans une analyse de liaison classique, on propose un modèle pouvant expliquer le mode d'hérédité d'un caractère. Figure provenant de l'article de Lander et Shork*[207]

Comme souligné plus haut, de par la petite taille des familles, les études de liaison chez l'homme ont été utilisées

quasi exclusivement pour des maladies mendéliennes (un seul gène impliqué), ou pour des formes mendéliennes de maladies courantes[5,4,41]. En effet, comme le nombre de descendants est souvent restreint, le nombre de méioses observables est faible, ce qui limite la résolution statistique de ce genre d'approche pour des caractères complexes, où généralement plusieurs *loci* représentant un faible pourcentage de la variation phénotypique sont impliqués. Toutefois en génétique animale, chez les espèces de productions, la structure des pedigrees est telle qu'il est fréquent qu'un seul individu engendre de multiples descendants. Il devient donc possible d'observer comme pour des croisements expérimentaux, un grand nombre de méioses issues d'un même individu. Ces caractéristiques ont permis d'employer dans ces espèces des études de liaison en cartographie de caractères complexes, en particulier pour des caractères quantitatifs.

I.4.1.2 Statistiques des études de *liaison*.

On peut distinguer deux types d'approches: les approches dites paramétriques et les approches dites non paramétriques.

I.4.1.2.a Les méthodes paramétriques.

I.4.1.2.a.1 LODSCORE.

L'approche la plus classique dans les études de liaison consiste à estimer la vraisemblance des données dans un modèle positionnant un *locus* dans une région donnée comme gène responsable du phénotype (hypothèse H1) et de comparer ensuite cette vraisemblance à celle d'un modèle où il n'existe aucune liaison entre le *locus* et la région testée (hypothèse H0). Cette comparaison de vraisemblances s'écrit sous la forme d'un LODscore, qui correspond au logarithme en base 10 du rapport entre le maximum de vraisemblance sous l'hypothèse H1 ($0 \leq \theta \leq 0.5$) et le maximum de vraisemblance sous l'hypothèse H0 ($\theta=0$):

$$LOD = \log_{10}[L(\theta)/L(\theta=0.5)] \quad 32$$

Classiquement, on considère que la liaison est significative quand LOD > 3, dans les cas simples de maladies monogéniques, où un seul modèle est testé. Cependant si l'on s'écarte de cette situation, on peut être amené à tester plusieurs modèles, ce qui nécessite de corriger ce seuil pour tests multiples, une tâche qui n'est pas toujours triviale. Un autre problème est d'estimer l'intervalle de confiance pour la position donnant le lodscore le plus élevé. On considère généralement que l'intervalle de confiance correspond à la région entourant le pic ayant un lodscore ne diminuant pas plus d'une unité par rapport au maximum (voire deux selon certains auteurs[42]).

I.4.1.2.a.2 Estimation de la vraisemblance.

I.4.1.2.a.2.1 Méthodes exactes.

Un des intérêts de ces méthodes basées sur la vraisemblance des données est qu'il est possible de tester un modèle dans lequel il existe ce que l'on appelle un certain nombre de variables latentes qui sont soient des données manquantes, soient des paramètres inconnus. On cherche les valeurs pour ces variables latentes qui maximiseront la vraisemblance des données observées. Dans une étude de liaison on exprime la vraisemblance d'un pedigree comme une fonction des paramètres du modèle, représenté collectivement par Θ.

Si on suppose que les probabilités Pr des phénotypes Y des n individus conditionnelles à leur génotype G sont indépendantes, on peut alors écrire:

$$Pr(Y|G) = Pr(Y_1 = y_1|G_1 = g_1) \times ... \times Pr(Y_n = y_n|G_n = g_n)$$

, où G et Y sont les vecteurs de phénotypes et génotypes respectivement et Pr(Yi=yi|Gi=gi) est la probabilité du phénotype l'individu i conditionnellement à son génotype.

Toutefois si on ne dispose pas des génotypes, les probabilités conjointes des phénotypes sont obtenues en sommant sur l'ensemble des combinaisons des génotypes possibles; la vraisemblance des données s'écrit alors de la façon suivante:

$$L(\Theta) = Pr(Y|\theta) = \sum_g Pr(Y, G = g) = \sum_g Pr(Y|G = g)\, Pr(G = g|q)$$

, où q est le vecteur donnant les fréquences génotypiques.

Le premier facteur du terme de droite dans cette équation correspond à l'expression précédente, tandis que le second facteur peut s'écrire, si l'on considère les lois de Mendel pour définir les génotypes des descendants selon les génotypes de leurs parents, de cette manière:

$$Pr(G) = \prod_{i=1}^{N} \begin{cases} Pr(G_i = g_i|G_{m_i} = g_{m_i}, G_{f_i} = g_{f_i}) \; non\ fondateurs \\ Pr(G_i = g_i|q) \; fondateurs \end{cases}$$

, où N correspond au nombre d'individus dans le pedigree, Gm_i et Gf_i correspondent aux génotypes des parents de l'individu i.

En combinant les expressions précédentes, la vraisemblance peut s'écrire:

$$L(\Theta)=\sum_{g}\prod_{i}Pr(Y_i|g_i)\times\begin{cases}\prod Pr(g_i|g_{m_i},g_{f_i})\ non\ fondateurs\\ \prod Pr(g_i|q)\ fondateurs\end{cases}$$

Les différents termes dans cette expression correspondent soit à des valeurs numériques soit à une fonction avec des paramètres à estimer (θ, pénétrance...). Par exemple, dans une étude simple point Pr(G|Gf,Gm) est fonction des génotypes du marqueur M et de la distance génétique entre le marqueur M et le *locus* gène impliqué dans le caractère. La vraisemblance du pedigree en fonction de ces différents paramètres peut s'écrire aussi:

$$L(\theta,\omega)=\sum_{g}\prod_{i}Pr_f(Y_i|g_i)\times\begin{cases}\prod Pr_\theta(g_i|M_i,g_{m_i},g_{f_i},M_{f_i},M_{m_i})\ non\ fondateurs\\ \prod Pr(g_i|M_i)\ fondateurs\end{cases}$$

, où ω = (f,q) et f = (f_0,f_1,f_2) sont le vecteur des pénétrances pour les trois génotypes.

Le problème de cette formule est le fait d'avoir à faire une somme sur tous les génotypes possibles pour l'ensemble des membres du pedigree. Or quand le nombre d'individus augmente, le nombre de termes peut croire de manière exponentielle (3^Npour exprimé la vraisemblance avec une seul locus biallélique en utilisant pedigree contenant N individus).

I.4.1.2.a.2.1.1 Algorithme de Elston-Stewart.

L'énumération de tous les génotypes possibles est infaisable même quand le nombre d'individus est faible (avec 13 individus dans un pedigree, on a presque 1.6 millions de termes à évaluer). Une approche alternative pour évaluer la vraisemblance associée à un pedigree a été proposée par Elston et Stewart[43].

En exploitant l'indépendance des frères et sœurs conditionnellement aux génotypes des parents, il est possible de réécrire la vraisemblance pour une simple famille nucléaire de la façon suivante:

$$L(\Theta)=\sum_{g_p}\sum_{g_o}Pr(Y_p,G_p=g_p)Pr(Y_o,G_o=g_o|G_p=g_p)=\sum_{g_p}Pr(Y_p,G_p=g_p)(\prod_{o=1}^{s}Pr(Y_o,g_o|g_p))$$

, où le indices o et p référent aux descendants et aux parents respectivement.

Dans le cas d'un *locus* biallélique, il y a 9 génotypes possibles pour les parents, donc pour des parents avec un seul descendant, il y a en principe 27 configurations génotypiques; cependant il y a 12 situations impossibles (ex père: 11 mère: 11 et descendant: 22). Le nombre de configurations génotypiques possibles est donc de 15 par descendant. Si le nombre de descendants augmente, le nombre de configurations possibles à envisager croît de façon linéaire: pour s descendant, il y a 15s configurations possibles. Pour une famille avec 4 descendants, il y a donc 60 termes à évaluer dans le calcul de la vraisemblance, tandis qu'avec l'expression générale pour estimer la

vraisemblance d'un pedigree, il y a 3^6 termes = 768 à estimer.

Pour généraliser cette expression à un pedigree, l'algorithme de Elston-Stewart nécessite de subdiviser un pedigree en sous pedigrees indépendants. Cet algorithme exploite le fait que si l'on connaît le génotype d'un parent appartenant à une famille A et que cet individu est lui même un descendant dans une famille B, les phénotypes et génotypes des individus appartenant à la famille A sont indépendants des individus présents dans la famille B. Par exemple pour le pedigree montré (Figure I.3), on commence par subdiviser le pedigree en 3 sous-familles nucléaires A,B,C. En remontant dans le pedigree, on exprime la vraisemblance des données pour les individus 3 et 7 conditionnelle au génotype de l'individu 4: $A(g4) \equiv Pr(Y3, Y7|g4)$

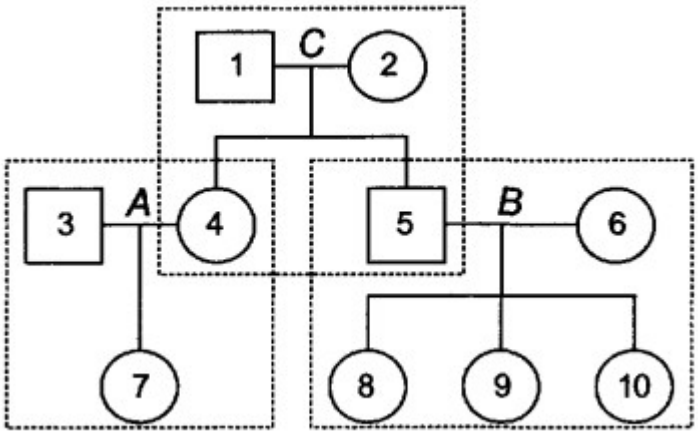


Figure I.3: *Pedigree simple pour illustrer l'algorithme de peeling de Elston-Stewart. Figure tirée de Statistical Methods in Genetic Epidemiology*[46].

L'information de la sous-famille A est contenue à travers l'individu 4: on dit qu'on a *pelé* les individus 3 et 7. On fait de même avec la famille B on exprime la vraisemblance des données pour les individus 6,8,9,10 conditionnelle au génotype de l'individu 5 $B(g5) \equiv Pr(Y6, Y8, Y9, Y10|g5)$. Dans la famille C on exprime ensuite la vraisemblance des phénotypes de l'individu 3 à 10 conditionnelle aux génotypes des parents 1 et 2 en utilisant les vraisemblances pelées précédentes:

$$C(g1, g2) \equiv Pr(Y3, .., Y10|g1, g2) = (\sum_{g_4} A(g_4) Pr(Y_4|g_4) Pr(g_4|g_1, g_2)) \times (\sum_{g_5} B(g_5) Pr(Y_5|g_5) Pr(g_5|g_1, g_2))$$

On calcule la vraisemblance du pedigree en sommant pour tous les génotypes possibles des individus 1 et 2:

$$Pr(Y1, ... Y10) = \sum_{g1} \sum_{g2} Pr(Y1|g1) Pr(Y2|g2) Pr(g1) Pr(g2) C(g1, g2)$$

Ici l'ordre de peeling des sous-familles était A,B et C. Cependant cet ordre peut être modifié et ne rien changer à l'expression finale de la vraisemblance. Par ailleurs, on n'est pas obligé de remonter dans le pedigree, par exemple si l'individu 6 a deux parents 11 et 12 (Figure I.4): on peut remplacer Pr(g6) en calculant la probabilité jointe des phénotypes des parents et du génotype de l'individu 6:

$$D(g6) \equiv Pr(Y11, Y12|g6) = \sum_{g_{11}} \sum_{g_{12}} Pr(Y_{11}|g_{11})\, Pr(Y_{12}|g_{12})\, Pr(g_6|g_{11}, g_{12})\, Pr(g_{11})\, Pr(g_{12})$$

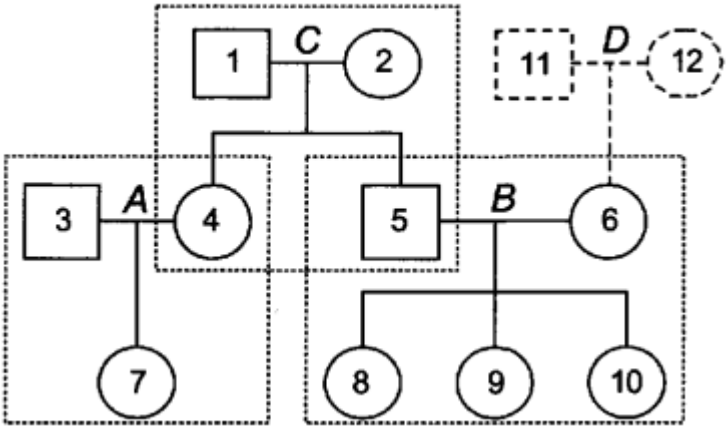


Figure I.4: *La famille D a été ajoutée afin de créer un pedigree plus complexe. Figure tirée de Statistical Methods in Genetic Epidemiology*[46].

Dans des pedigrees complexes, comme dans des pedigrees simples on a ce que l'on appelle des individus de liaison (ex l'individu 4) qui divise le pedigree en deux groupes: (i) *lower neighborhood*, les individus qui sont les descendants de l'individu de liaison, son époux (ou épouse) et les parents de son époux (ou épouse) (ii) u*pper neighborhood* qui sont les parents et tous les individus liés à ces parents. Ces deux groupes sont indépendants conditionnellement au génotype de l'individu de liaison.

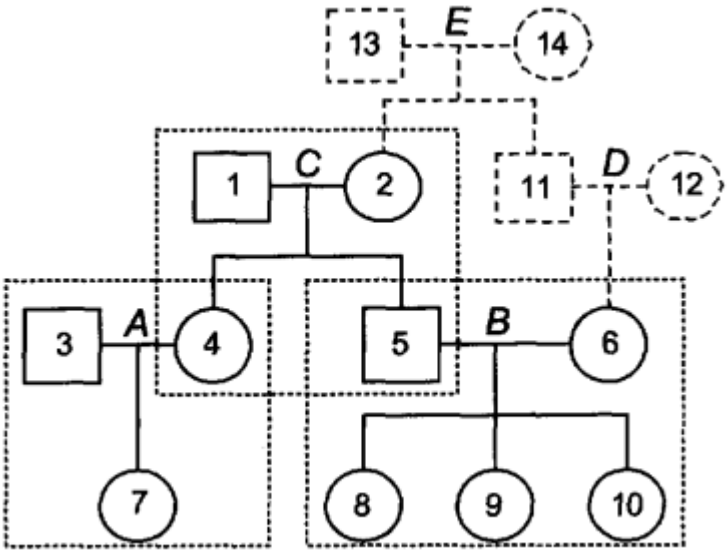


Figure I.5: *La famille E a été ajoutée, ce qui provoque une boucle de consanguinité dans le pedigree, formée par les individus 2, 11, 5 et 6. Figure tirée de Statistical Methods in Genetic Epidemiology*[46].

En cas de boucle de consanguinité le calcul peut se compliquer (Figure I.5) Par exemple si l'individu 2 et 11 sont frère et sœur (13 et 14 sont les parents) pour un individu de liaison donné (p.e. 5) des individus du groupe *lower neighborhood* peuvent être présents aussi dans le groupe *upper neighborhood* p.e. 13 et 14. La solution est de prendre non pas un individu mais un groupe d'individus appelé *cutset* pour diviser le pedigree en deux groupes *lower neighborhood* et *upper neighborhood*.

Cet algorithme de peeling d'Eslton et Stewart est au cœur d'un grand nombre de logiciels utilisés dans les études de liaison (p.e.: LINKAGE et VITESSE).

I.4.1.2.a.2.1.2 Algorithme de Lander et Green.

Au fur et à mesure du temps, le nombre de marqueurs disponibles a augmenté. Par ailleurs il est devenu de plus en plus simple de génotyper une grande majorité des individus présents dans un pedigree. Les études de liaison multipoint se sont développées car elles sont plus puissantes d'un point de vue statistique (ceci étant du à une augmentation de l'informativité des méioses).

Quand le nombre de marqueurs est faible (p.e. 2), il est possible d'utiliser l'algorithme de peeling de Elston-Stewart, et de considérer que dans l'équation générale de la vraisemblance d'un pedigree, le vecteur g représente un vecteur de génotype *multiloci.* Si l'on souhaite prendre en compte un nombre illimité de *loci* dans le calcul de la vraisemblance, une approche alternative a été proposée par Lander et Green[44]. Il faut souligner que cet algorithme est limité aux pedigrees simples avec un faible nombre d'individus. Contrairement à l'approche d'Elston et Stewart, où l'on examine les individus un à un pour tous les *loci* simultanément, l'algorithme de Lander et Green, est basé sur une chaîne de Markov cachée (HMM pour Hidden Markov Model), examinant *locus* par *locus* toute l'information du pedigree sous la forme d'un vecteur d'hérédité conditionnellement aux vecteurs d'hérédité au niveau des *loci* flanquants. Le vecteur d'hérédité correspond pour une paire d'individu, aux probabilités des différentes configurations génotypes: deux individus ont recus 0,1 ou 2 mêmes génotypes. Pour des marqueurs, il est possible de calculer directement le vecteur d'hérédité Q sur base des génotypes disponibles . Ainsi si on souhaite calculer le vecteur d'hérédité pour un *locus* gène B π_B voisin d'un marqueur flanquant A, il suffit de multiplier le vecteur d'hérédité π_A du locus A avec la matrice de probabilités de transition d'être identique par descente (IBD) pour le *locus* B sachant état IBD pour le *locus* A T_{AB}. Cette matrice de transition dépendra de la distance génétique θ entre les deux *loci*. On peut généraliser à deux marqueurs flanquants, on aura dans ce cas: $\pi_B = Q_A T_{AB} T_{BC} Q_C$. Dans une approche de type Lander et Green, la vraisemblance de la façon suivante:

$$L(x,\omega) = \sum_z Pr(Y|Z_x = z;\omega)\Pi(Z_x = z) \quad (1)$$

,où x correspond à position du *loci* impliqué dans le phénotype étudié et $\omega = (f,q)$ (voir au dessus)

$$Pr(Y|Z_x = z;\omega) = \sum_a \sum_{s|z} Pr(Y|G(s,a);f)\,Pr(a;q) \quad (2)$$

Le deuxième facteur de l'équation (1) consiste à calculer le vecteur d'hérédité d'un *locus* à une position x sachant que les individus non fondateurs sont dans une configuration z pour la matrice IBD. Dans l'équation 2 G(s,a) est un vecteur de génotypes correspondant à une configuration a des allèles chez les fondateurs et à une configuration s des allèles lors de la méiose chez ces mêmes fondateurs. Tous les paramètres sur lesquels on émet une hypothèse quant à leur valeur (f (fonction pénétrance) et q (fréquence allélique)) se trouvent dans l'équation 2 .

L'intérêt de cette méthode réside dans le fait que le temps de calcul augmentera linéairement avec le nombre de *loci* mais de manière exponentielle avec la taille du pedigree. Les logiciels GENHUNTER et MAPMARKER utilisent cette approche.

I.4.1.2.a.2.2 Méthodes d'échantillonnage.

Quand le nombre de marqueurs, la taille ou la complexité du pedigree grandissent, estimer la vraisemblance par des approches déterministes peut devenir rapidement fastidieux. Les méthodes d'échantillonnage de Monte-Carlo constituent une solution alternative pour estimer la vraisemblance: au lieu d'énumérer tous les génotypes possibles pour calculer la vraisemblance, on tire de manière répétée et aléatoire des configurations de génotypes possibles pour obtenir une solution approximative de la vraisemblance.

Une approche de type Monte-Carlo très répandue est le *Gibbs sampling*. Elle est fort utilisée quand le nombre de variables à estimer dans le modèle est élevé et qu'il devient impossible de calculer la distribution jointe de ces paramètres (ou variable latente). Cette approche se base: (i) sur le fait qu'il est facile de connaître la distribution marginale d'une variable sachant l'état de toutes les autres variables. (ii) Sur une mise à jour successive de l'état de chaque variable sur base de sa probabilité marginale calculée à partir des états des autres variables (iii) après n itérations (période de convergence) les états obtenus pour chaque variable à chaque itération successive correspondent à des échantillons provenant de la distribution conjointe de ces variables.

Appliqué à une étude de liaison cela donnera:

(i) On commence par assigner un génotype possible pour le *locus* gène à tous les membres du pedigree.

(ii) A chaque itération dans la boucle de *Gibbs sampling*: on considère tous les membres du pedigree un à un afin de lui attribuer un nouveau génotype pour le *locus* gène en utilisant la distribution marginale de son génotype sachant les génotypes de ses proches (parents, époux(ses), enfants) et son phénotype:

$$Pr(G_i = g_i | g_{-i}, Y_i) \propto Pr(Y_i | g_i) Pr(g_i | g_{m_i}, g_{f_i}) \prod_k Pr(g_{o_k} | g_{m_i}, g_{s_i})$$

, où g-i correspond au vecteur de génotype pour tous les individus excepté pour l'individu i.

(iii) On répète ce processus jusqu'à convergence.

À chaque itération on peut utiliser les génotypes échantillonnés pour estimer ensuite la vraisemblance des données (phénotype et génotype) conditionnellement aux valeurs données pour les paramètres du modèle et estimer un lodscore:

$$LR(\theta_1 : \theta_0) \approx \frac{1}{N} \sum_{G_n | Y, \theta} \frac{Pr(Y, G | \theta_1)}{Pr(Y, G | \theta_0)}$$

, où LR est une moyenne du rapport de vraisemblance et Gn corresponds au vecteur de génotype pour l'itération n.

I.4.1.2.a.3 Conclusions et limites de cette approche.

Ce type d'approche a permis d'élucider un grand nombre de maladies monogéniques. On l'a aussi utilisé pour des caractères hétérogènes. En effet, on peut parfois trouver de la liaison de manière significative (LOD> 3), bien qu'on ne trouve pas de liaison dans toutes les familles. Si on ne peut démontrer de la liaison directement, on a deux solutions: (i) on peut rajouter un paramètre d'hétérogénéité dans le modèle spécifiant la fraction attendue des familles avec de la liaison (ii) on peut redéfinir le phénotype et n'inclure que les familles correspondant à cette nouvelle définition. Par exemple dans une étude de liaison sur le cancer du sein, on a pu trouver de la liaison en incluant uniquement des familles avec des individus où la maladie s'était déclarée précocement[45].

Par ailleurs, redéfinir un caractère pour le rendre plus homogène est toujours intéressant, mais requiert des seuils de signification plus sévères pour compenser les tests multiples réalisés.

I.4.1.2.b Les méthodes non paramétriques: méthode d'allele sharing.

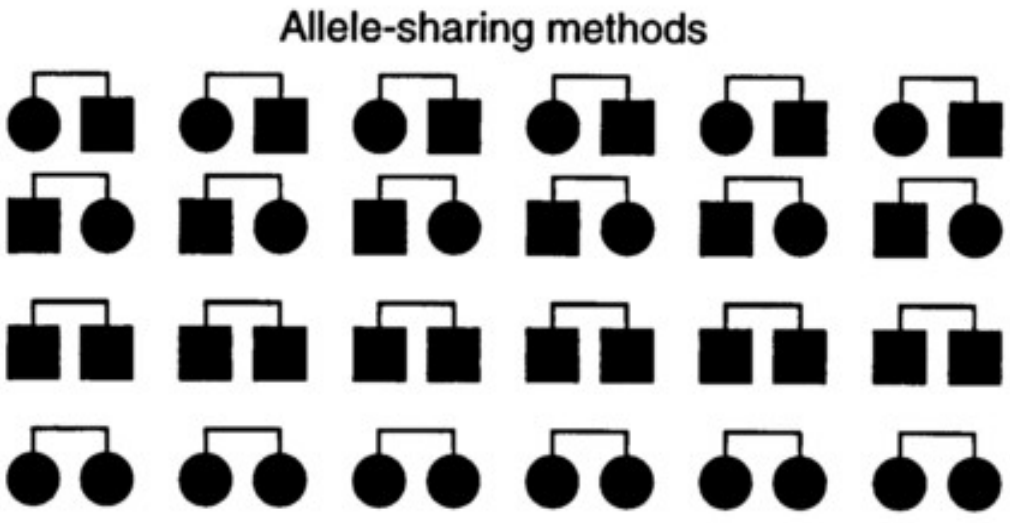


Figure I.6: *Le principe des méthodes de type « allele-sharing » est de tester si des paires d'individus partageant un phénotype commun héritent plus souvent qu'attendu des mêmes allèles. Dans une approche de type « Affected Sib Pair », on regarde pour des paires d'individus atteints frères et sœurs si ils ont reçu des génotypes marqueurs plus souvent qu'attendu sous l'hypothèse d'une ségrégation mendélienne aléatoire. Figure provenant de l'article de Lander et Shork*[207].

Une analyse de liaison basé sur un modèle est une approche très puissante pour découvrir les gènes impliqués dans des désordres de type mendélien. Cependant, même pour des maladies monogéniques, ce type d'approche peut échouer quand il n'existe pas de correspondance simple entre un génotype et un phénotype, ce qui est le cas en présence d'hétérogénéité génétique, ou de pénétrance incomplète ou encore d'expressivité variable. Une solution évidente est de trouver le modèle adéquat prenant en considération tous les facteurs pouvant influencer la relation phénotype-génotype. Toutefois, comme évoqué plus haut, essayer des modèles de plus en plus complexes pour détecter une éventuelle liaison implique de corriger les seuils de signification de manière plus sévère que lorsqu'on utilise un seul modèle. Une alternative consiste à rechercher des régions génomiques dont l'hérédité ne suit pas un modèle mendélien aléatoire: en cas de liaison de cette région avec le phénotype étudié, on montre que des individus partageant un phénotype commun héritent plus souvent qu'attendu des mêmes allèles (Figure I.6).

Cette approche appelée méthode d'*allele sharing* est basée sur le rejet d'un modèle (modèle: ségrégation mendélienne aléatoire), plutôt que sur l'acceptation d'un modèle (modèle dans lequel une hypothèse de liaison est explicitement défini): c'est une approche dite non paramétrique. L'avantage de ce type d'approche par rapport aux études de liaison génétique classiques est d'être plus robuste. En effet même en cas d'hétérogénéité génétique ou de pénétrance incomplète, le moindre écart par rapport à une ségrégation mendélienne aléatoire peut être détecté. Cependant, ces méthodes sont moins puissantes que les analyses de liaison classiques dans lesquelles un modèle est spécifié.

I.4.1.2.b.1 Méthodes de type ASP (Affected Sib Pair).

Les méthodes les plus simples d'*allèle sharing* sont les méthodes de type ASP (*Affected Sib Pair*) (voir Figure I.6). Ce type d'approches consiste à comparer au niveau d'un ou plusieurs marqueurs pour des paires d'individus atteints frères-soeurs la distribution des fréquences d'IBD avec la distribution des fréquences attendues pour un ou plusieurs marqueurs. Les fréquences attendues pour 0, 1 et 2 allèles IBD (marqueurs biallèliques) sont 25%, 50% et 25% respectivement. Si la distribution observée s'écarte de ce qui est attendu, une explication plausible est que ce ou ces marqueurs appartiennent à une région ayant un rôle dans le déterminisme de la maladie. Cette approche utilise des paires d'individus atteints car ce sont les paires les plus informatives. En effet si l'on part de l'hypothèse qu'une maladie est rare, alors un individu atteint est davantage susceptible de porter un ou plusieurs allèles alternatifs qu'un individu pris au hasard dans la population, tandis qu'un individu sain a légèrement moins de chance d'être porteur d'une mutation.

Il existe plusieurs tests statistiques pour comparer la distribution observée à celle attendue sous l'hypothèse d'une ségrégation mendélienne. La puissance statistique de ces tests dépendra du mode d'action du gène (récessif, dominant, additif) responsable.

Table I.1: *Distribution du nombre d'allèle partagé IBD (« identity by descent ») pour des paires d'individus atteints frères-soeurs. N est le nombre de paires.*

	Nombre d'allèles partagés IBD			
	0	**1**	**2**	**Total**
Observé	A	B	C	N
Attendu	N/4	N/2	N/4	N

Par exemple, une approche de type test t comparant la moyenne attendue 0.5 par rapport à la moyenne observée ($\hat{\pi} = \frac{(B+2C)}{2N}$, voir Table I.1) de la proportion des allèles partagés IBD (en utilisant comme déviation standard $SE(\hat{\pi}) = \sqrt{AB + 4AC + BC}/4N^3$ [46]), ou encore une approche de type chi2 à 1ddl regardant une éventuelle différence de répartition entre les classes zéro et deux allèles IBD sont des approches plus puissantes dans le cas d'un gène rare dominant ou additif qu'une approche de type chi2 à 2ddl comparant la répartition observée des trois classes (0,1,2 allèles partagés IBD) avec leur distribution attendue (IBD0=0.25, IBD1=0.5, IBD2=0.25). Toutefois en cas de faible liaison, avec un gène récessif, la proportion des allèles IBD1 s'écartant de 0.5, le test chi2 1ddl suivant s'avère être le test le plus avantageux en termes de puissance:

$$\frac{(3C-2B-A)^2}{(9C+4B+A)} \sim X_1^2$$

Il est possible d'étendre ce type de méthodes à n'importe quel type de paires d'individus atteints dans le pedigree. La Table I.2 donne les proportions attendues d'*allèle sharing* selon le type de relation de parenté. Cependant quand les individus sont éloignés, on préféra utiliser les méthodes à base de probabilités d'être identique par état (IBS ou identity by state) (voir paragraphe suivant). Parfois quand la proportion d'allèle sharing IBD ne peut être déterminée avec certitude, la valeur π_{obs} pour des paires d'individus atteints peut être biaisée. Une alternative est de comparer π_{obs} avec π_{est} à partir de paires d'individus atteints-sains, constituant ainsi un contrôle efficace à ce type d'artefact.

Table I.2: *Distribution des probabilités de partager des allèles IBD (« identity by descent ») pour des paires d'individus pour différents types de relation parenté. Table tirée de Statistical Methods in Genetic Epidemiology*[46].

Type of Relative Pair	Probability of Sharing IBD Alleles		
	π_0	π_1	π_2
Monozygotic twins	0	0	1
Full sibs	1/4	1/2	1/4
Parent-offspring	0	1	0
First cousins	3/4	1/4	0
Double first cousins	13/16	1/8	1/16
Grandparent-grandchild, half-sibs,avuncular	1/2	1/2	0

I.4.1.2.b.2 Méthodes de type APM (Affected Pedigree member).

Il n'est pas toujours évident de déterminer si deux individus partagent le même allèle en provenance d'un ancêtre commun (allèle IBD) ou de deux ancêtres disjoints (IBS): exemple deux frères ayant le même allèle, mais avec un des deux parents homozygotes pour cet allèle.

Il existe deux solutions à ce problème: la première consiste à tenter d'imputer l'état IBD des marqueurs sur base de leur état IBS et l'état IBD de marqueurs proches. Ce type d'approches est valable pour des cartes de marqueurs denses. La seconde solution est de développer une approche statistique directement basée sur les états IBS; exploitant le fait que des individus apparentés affectés et génotypés pour un des marqueurs, lié à un gène de susceptibilité à la maladie, auront des génotypes similaires pour ce marqueurs plus souvent que ceux attendus par chance.

Plusieurs améliorations ont été proposées pour ce type d'approches; comme par exemple intégrer les fréquences alléliques dans la population pour donner plus de poids à des allèles rares, ou encore utiliser l'information IBS sur plusieurs marqueurs. Ce type d'approche a été appliqué avec succès à des désordres donnant des résultats équivoques avec des analyses de liaison traditionnelles: on a pu ainsi découvrir un lien entre le gène de l'angiotensine et l'hypertension familiale[47] ou encore entre la maladie d'Alzheimer tardive et un *locus* sur chromosome 19[48].

I.4.1.2.b.3 Méthodes applicables à des caractères quantitatifs.

Haseman et Elston[49] ont étendu ce type d'approches à des caractères quantitatifs en se basant sur la notion suivante: la ressemblance phénotypique entre deux individus apparentés doit dépendre du nombre d'allèles IBD au niveau d'un *locus* impliqué. Via une régression on regarde l'effet du nombre d'allèles IBD en commun sur les différences phénotypiques au carré de paires d'individus apparentés:

$$Y_i = \alpha + \beta \pi_j$$

, où Yi est la différence phénotypique entre une paire d'individus élevée au carré, π_j est la fraction des marqueurs pour une paire j IBD0, IBD1 et IBD2, α l'intercept de y et β le coefficient de régression. Sous l'hypothèse nulle, pas liaison, β ne doit pas être significativement différent de zéro.

Ce type d'approches a permis de mettre en évidence chez l'homme le rôle du gène IL4 (*interleukin 4*) dans les différences de concentrations sériques en IgE (*immunoglobulins E*) entre individus[50] ou encore de relier la densité osseuse postménopause avec le gène codant pour le récepteur de la vitamine D[51].

I.4.1.3 Les études de liaison dans le cadre de caractères quantitatifs chez les espèces de production.

I.4.1.3.a Introduction.

Chez les animaux de production, où les enjeux agronomiques prévalent, les caractères étudiés sont le plus souvent des caractères quantitatifs et complexes (p.e. les différents paramètres de production laitière d'un bovin ou encore l'épaisseur du lard dorsal chez le porc). Dans ce type d'espèces, on dispose soit de croisements expérimentaux (de type F2* ou le plus souvent pour des raisons économiques de BC* (*Back Cross*), soit de pedigrees commerciaux. Ces derniers sont généralement le résultat d'une utilisation intensive de l'IA (insémination artificielle) dans le schéma de sélection, amenant certains parents à avoir de multiples descendants, de manière analogue à des croisements expérimentaux. Cette capacité d'observer un grand nombre de méioses par parent fait qu'il est possible de détecter de la liaison génétique même pour des *loci* expliquant une faible part de la variation génétique et donc plus que chez l'homme, les études de liaison génétique chez les espèces de production constituent une approche puissante en cartographie génétique pour des caractères

* Un croisement de type F2 est un croisement entre des individus hétérozygotes et génétiquement identiques.
* Un croisement backcross est un croisement entre un individu hétérozygote avec un de ses parents ou avec un individu génétiquement identique à ses parents.

complexes. Il faut cependant signaler que les croisements expérimentaux et les pedigrees commerciaux ont eu des utilisations divergentes. Les premiers s'intéressent davantage aux facteurs génétiques impliqués dans les variations phénotypiques entre races, ainsi par exemple, en croisant des lignées domestiques avec des lignées sauvages il est possible de découvrir les gènes ayant joué un rôle dans la domestication. Les seconds sont employés dans le but d'élucider les causes génétiques impliquées dans les variations phénotypiques entre individus d'une même race; le type de question auxquels ils peuvent contribuer à répondre est, par exemple: quels sont les gènes qui permettent de distinguer les meilleurs taureaux laitiers de leurs congénères moyens ?

Les études de liaison ont été appliquées avec succès pour différents caractères agronomiques chez différentes espèces de production. Cependant les propos développés par la suite traitent le cas particulier des populations de bovins laitiers. Deux arguments à cela: en premier lieu ces populations, pour des raisons économiques, ont fait l'objet d'une grande attention: estimations précises d'un grand nombre de caractères, pedigrees connus sur plusieurs générations, développement d'approches génétiques spécifiques ainsi que d'outils génétiques (cartes de marqueurs). Le second argument est qu'il est facile de transposer les approches développées pour cette population à d'autres espèces de production qui ont des structures de population analogues, de même qu'à des croisements expérimentaux, qui peuvent être vus comme des cas particulier des pedigrees de ces populations.

I.4.1.3.b Les daughter-granddaughter design.

Le développement du *progeny-test* comme processus de sélection de pères de taureaux laitiers, ainsi que de l'utilisation de l'IA a conduit à distinguer deux types de structure de pedigree dans les études de liaisons chez les bovins laitiers:

(i) les *Daughter Design* (DD) où on s'intéresse à la ségrégation d'un allèle QTL* (*Quantitative Trait Locus*) à travers un ensemble de familles de demi-sœurs paternelles et où les phénotypes sont des mesures réalisées sur chaque fille de chaque taureau et éventuellement corrigés pour l'environnement et la contribution maternelle (ii) des *Granddaughter-Design* (GDD) dans lesquels on regarde la transmission d'allèles QTL cette fois dans des familles de demi-frères paternels, le phénotype des fils étant une valeur d'élevage, c'est-à-dire une estimation de sa valeur génétique pour un caractère obtenue à partir de la moyenne des phénotypes de ses filles. Dans les GDD, seuls les mâles sont génotypés. À noter également que les GDD sont très bien adaptés aux populations laitières, car il n'existe pas de caractères laitiers observables pour les mâles dans ces populations (voir Figure I.7).

Le principe de base pour des études de liaison dans ce type de design, DD et GDD, consiste pour un parent

* Locus influençant un caractère quantitatif.

hétérozygote au niveau d'un marqueur à trier les demi-frères/demi-sœurs selon l'allèle marqueur reçu et de regarder ensuite s'il existe des différences significatives entre les deux groupes, pouvant refléter de la liaison génétique dans cette famille.

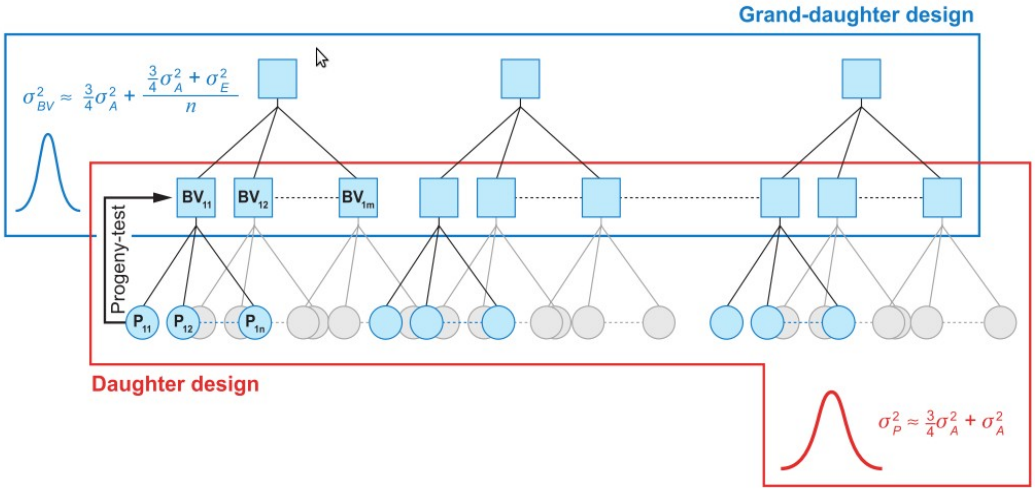


Figure I.7: *Pedigree de type DD (Daughter Design) et GDD (Grand-daughter design). Dans ce type de pedigree on regarde des familles soient de demi-sœurs soient de demi-frères paternels. Les DD sont constitués de familles de demi-soeurs maternelles. Un phénotype est estimé sur chaque fille. Les GDD sont constitués de familles de demi-frères paternelles. On estime la valeur génétique d'un taureau (Breeding Value BV) sur des mesures réalisées sur ses filles. Figure provenant de l'article de Georges*[93].

Bien que les effets estimés dans les GDD par rapport au DD soient réduits de moité, les études de liaison dans les GDD offrent une meilleure puissance que celles réalisées dans les DD. En effet, dans les GDD, la variance associée aux effets résiduels est réduite par le fait que l'estimation d'une BV d'un taureau laitier est réalisée sur un grand nombre de filles. Il faudra donc davantage d'individus pour détecter le QTL avec une étude de liaison réalisée dans un DD que dans un GDD.

I.4.1.3.c Les méthodes de cartographie par intervalle.

I.4.1.3.c.1 Avantage des méthodes de cartographie par intervalle sur les méthodes simple point.

Les études de liaison pour des caractères complexes (aussi bien quantitatifs que qualitatifs) ont un double objectif: (i) identifier les *loci* responsables, (ii) quantifier leurs effets sur le phénotype étudié. Le problème des approches simple point (ou marqueur par marqueur) est que l'on ne peut pas distinguer les situations où on a un QTL avec un petit effet mais fortement lié de celles où on a un QTL avec un effet important mais avec une liaison moindre. Par exemple supposons que l'on souhaite cartographier un QTL biallélique (Q,q) avec un effet substitution allélique α Ce QTL est à une distance génétique θ d'un marqueur. Si un père est double hétérozygote (hétérozygote à fois pour le marqueur (1,2) et le QTL (Q,q)) et que l'on connaît la phase (1Q/2q), on peut alors comparer la moyenne phénotypique du groupe ayant reçu l'allèle marqueur 1 $\bar{P}_1$ avec celle du groupe ayant reçu l'allèle marqueur 2 $\bar{P}_2$ de la façon suivante[52]:

$$\bar{P}_1 - \bar{P}_2 = \alpha(1-2\theta)$$

Le contraste entre les deux groupes dépendra à la fois de la position du marqueur par rapport au QTL (via θ) et de l'effet du QTL (via α). En d'autres termes dans les études de liaison simple point (c'est à dire utilisant un seul marqueur), il n'est pas possible de discerner la position du QTL de son effet.

Pour surmonter ce problème, la solution consiste à exploiter l'information de plusieurs marqueurs simultanément de part et d'autre de la position supposée du QTL. Ce type d'approche est appelée cartographie par intervalle ou *interval mapping*.

I.4.1.3.c.2 Méthodes d'interval-mapping paramétriques.

I.4.1.3.c.2.1 Approches par maximum de vraisemblance.

Comme précédemment, on peut écrire la vraisemblance des données observées pour un modèle donné en fonction d'une série de paramètres comme par exemple la distance génétique entre un QTL et un marqueur génétique θ ou encore un effet de substitution allélique α. On recherche ensuite les paramètres qui maximisent cette vraisemblance et on teste par la suite les hypothèses alternatives typiquement, l'absence de QTL dans le voisinage étudié par des rapports de maximum de vraisemblance.

Ainsi pour un descendant o la vraisemblance de son phénotype y_0 peut s'écrire, si le père est supposé Qq (à

gauche l'allèle d'origine paternel, à droite celui d'origine maternelle):

$$L(y_o|S(Qq)) = P(Q|M_s, M_o) \times N(y_p + \frac{\alpha}{2}; \sigma^2) + P(q|M_s, M_o) \times N(y_p - \frac{\alpha}{2}; \sigma^2)$$

Dans cette expression, le phénotype du descendant est supposé normalement distribué $y \sim N(\mu, \sigma^2)$: le paramètre μ dépend du génotype reçu par le descendant, tandis que σ^2 correspond à la variation résiduelle (estimée en maximisant la vraisemblance). Les fonctions de génotypes sont pondérées par les probabilités conditionnelles d'avoir reçu un allèle QTL x connaissant les génotypes marqueurs au niveau de l'intervalle marqueur supposé du QTL pour le père (Ms) et le descendant (Mo) considéré. On peut bien évidemment réécrire cette expression dans le cas où un père est homozygote qq ou QQ. Dans ce cas la distribution du phénotype du descendant suivra une distribution normale où on posera α=0:

$$L(y_o|S(QQ)) = L(y_o|S(qq)) = N(y_p; \sigma^2)$$

La vraisemblance d'un phénotype s'obtient donc en sommant sur toutes les combinaisons possibles d'haplotypes H au niveau QTL pour le père et en pondérant par la probabilité de leur contribution:

$$L(y_O) = \sum_H L(y_O|S(H))\, p(S=H)$$

Par ailleurs, si le père a plusieurs descendants, alors on multipliera les vraisemblances de ses descendants:

$$L(Ped_i) = \sum_H \prod_o L(y_O|S(H))\, p(S=H)$$

Si on a un pedigree constitué de N familles de type demi-frères paternels, on ne peut pas supposer que tous les pères fondateurs sont hétérozygotes. On émettra par contre l'hypothèse que le *locus* QTL est en équilibre Hardy-Weinberg. La vraisemblance des données peut alors s'écrire:

$$L_{NPeds} = \prod_{i=1}^{N} \begin{bmatrix} f_Q^2 P(Ped_i|Sire_i = QQ) \\ + f_q^2 P(Ped_i|Sire_i = qq) \\ + f_Q f_q P(Ped_i|Sire_i = Qq) \\ + f_q f_Q P(Ped_i|Sire_i = Qq) \end{bmatrix}$$

Les valeurs des paramètres qui maximisent la vraisemblance peuvent être obtenues par des méthodes de type "quasi-Newton" ou en utilisant un algorithme de type EM* (*espérance-maximisation*).

Pour tester la signification du modèle, on compare la vraisemblance des données sous des hypothèses

* L'algorithme espérance-maximisation (en anglais Expectation-maximisation algorithm) souvent abrégé EM est une classe d'algorithmes qui permettent de trouver le maximum de vraisemblance des paramètres de modèles probabilistes lorsque le modèle dépend de variables latentes non observables.

alternatives (pe H1 présence d'un QTL et H0 absence de QTL) en calculant un lodscore, qui est le ratio des maximums de vraisemblance des deux hypothèses:

$$LOD = \log_{10} \frac{L(NPeds|H1)}{L(NPeds|H0)}$$

ou encore:

$$2\ln \frac{L(NPeds|H1)}{L(NPeds|H0)} \sim X^2_{ddl\ H1 - ddl\ H0}$$

Le problème qui se pose maintenant est de choisir l'hypothèse H0. En effet, on émet comme hypothèse H0 soit θ=0.5 (pas de QTL lié) soit a=0 (pas effet QTL). Le choix de H0 est loin d'être anodin. En effet dans des positions où le contenu en information est moindre (p.e. extrémité des chromosomes), prendre a=0 comme H0 aura tendance à gonfler artificiellement le lodscore, cependant s'il existe réellement un QTL avec un gros effet mais non lié au chromosome testé alors cette hypothèse fera diminuer systématiquement le lodscore.

I.4.1.3.c.2.2 Approches de type moindres-carrés.

Il est possible également d'utiliser une méthode de type moindres-carrés pour estimer les différents paramètres. On peut écrire que le phénotype d'un individu appartenant à une famille demi-frères paternelles de la façon suivante:

$$y_{jl} = \mu + s_j + b_{kj} P_{kjl} + e_{jl}$$

y_{jl} est le phénotype de l'individu l (si GDD est une moyenne calculée sur les filles de l'individu l) ayant comme père j. μ est la moyenne générale. s_j est l'effet associé au père de la famille j, b_{kj} est l'effet du QTL à position k dans la famille j (α/2), P_{kjl} est la probabilité pour l'individu l d'avoir reçu l'allèle QTL Q (le QTL est supposé à une position k) et e_{jl} est le résidu associé à une variance $w_{jl}\sigma_e^2$ (w_{jl} est le poids associé au nombre de mesures pour l'individu l). À noter qu'il existe des différences notables en terme d'hypothèses par rapport à l'approche précédente: (i) l'hypothèse nulle est toujours de considérer que l'on n'a pas d'effet QTL (ii) dans l'approche précédente dans les familles avec une ségrégation possible des allèles QTL (Qq), on supposait un effet de substitution allélique non nul et identique à travers ces familles, tandis qu'on posait ce même effet à zéro dans les autres familles, ici α prend des valeurs différentes selon la famille (cet effet est niché dans celle-ci) et donc des familles peuvent ségréger pour des allèles QTL différents.

I.4.1.3.c.3 Méthodes de cartographie par intervalle non paramétriques.

Les méthodes précédentes s'appliquent très bien à des caractères distribués normalement. Cependant cette hypothèse de normalité des phénotypes n'est pas toujours la règle: on peut par exemple avoir des phénotypes tronqués, d'autres peuvent suivre une distribution de Poisson. Ces raisons ont motivé les généticiens à développer des méthodes de cartographie par intervalle non paramétriques basées sur les rangs pour des caractères quantitatifs. Les premières méthodes de cartographie par intervalle ont été proposées par Kruglyak [53] pour des croisements expérimentaux mais ont été étendues par la suite à des pedigrees commerciaux de bovins laitiers par Coppieters [54].

La première étape des méthodes de cartographie par intervalle consiste à attribuer un rang à chaque individu sur base de la valeur de son phénotype. Deux individus ayant deux phénotypes identiques auront également deux rangs identiques.

Dans une famille avec un père hétérozygote au niveau du QTL, on suppose que n1 individus ont reçu l'allèle Q et n2 individus ont reçu l'allèle q. On ne dispose pas des allèles QTL pour les membres de cette famille mais on peut toutefois, grâce à l'information marqueur, distinguer pour une position donnée l'homologue (A ou B) reçu par un descendant. En utilisant un test de Wilcoxon (test non paramétrique basé sur la somme de rang), et adapté par Coppieters[54] pour des pedigrees de type demi-frères paternels, on peut voir s'il existe une différence significative entre les valeurs phénotypiques transformées en rang entre le groupe ayant reçu le chromosome A avec le groupe ayant reçu le chromosome B. Pour cela on commence par calculer la déviation Y_w entre la somme des rangs pour les individus ayant reçu l'homologue A par rapport à la somme des rangs pour les individus ayant reçu l'homologue B.

$$Y_K = \frac{1}{2}\sum_{j=1}^{n}((n+1)-2\times rank_j)(P(A)_j - P(B)_j)$$

$P(A)_j$ (ou $P(B)_j$) correspond à la probabilité qu'un fils j ait reçu l'allèle A (ou l'allèle B) de son père à une position p conditionnellement à l'information des marqueurs flanquants. Une différence majeure par rapport un test de Wilcoxon classique est que P(A) et P(B) peuvent prendre des valeurs comprises entre zéro et un et pas uniquement zéro et un. Sous l'hypothèse nulle qu'il n'existe pas de différence significative dans la somme des rangs entre les groupes, Y_K suit asymptotiquement une distribution normale de moyenne $\mu=0$ et de variance

$$\langle \sigma^2_{Y_z} \rangle = \frac{n^3-n}{3}\langle \sigma^2_{P(A)-P(B)} \rangle \quad .$$

Alors $Z_K = \frac{Y_K}{\sqrt{\langle \sigma^2_{Y_K} \rangle}}$ est distribué comme une distribution normale standard. On peut donc aisément tester un effet de l'allèle transmis.

On peut généraliser cette expression à k familles de demi-frères paternels en élevant les scores Z_K de famille au carré et en les sommant, on obtient une statistique de type $X^2_{k\,ddl}$:

$$\sum_{i=1}^{k} Z^2_{K_i} \sim X^2_k$$

Les remarques concernant les caractéristiques de ce type d'approche sont les remarques usuelles: cette approche est plus robuste mais moins puissante qu'une approche paramétrique. À noter également que ce type de méthode ne permet ni d'estimer les effets QTL, ni d'estimer une position de QTL, elle aide uniquement à mettre en évidence une éventuelle liaison.

I.4.1.3.d Exploitation de toute l'information contenue dans un pedigree: modèles mixtes

I.4.1.3.d.1 Introduction: modèle animal.

Ces études de liaisons exploitant DD et GDD dans des pedigrees commerciaux peuvent ne pas être assez puissantes pour cartographier des QTL avec des effets trop modestes lors d'un examen du génome entier. La taille des échantillons restant limitée, des gains substantiels dans la puissance de détection d'un QTL ne pouvaient être obtenus qu'avec des méthodes exploitant l'information ignorée jusqu'à présent, c'est-à-dire toutes les relations de parenté hors DD ou GDD. Il arrive très souvent pour des individus que l'on considère comme des fondateurs dans un DD ou GDD de disposer de l'information de pedigree remontant sur des dizaines de générations et potentiellement reliant les familles entre elles.

Ces approches exploitant toute l'information pedigree sont nées bien avant les premières études de liaison chez les animaux domestiques. En effet elles découlent d'une longue tradition en génétique animale visant à estimer le mérite génétique d'un animal de production ou encore sa valeur d'élevage dans le but de sélectionner les meilleurs reproducteurs.

Les phénotypes des individus pour lesquels on dispose d'une mesure sont modélisés (en notation matricielle) selon une somme d'effets fixes et d'effets aléatoires de la façon suivante:

$$y = Xb + Zu + e$$

,où $y(n\times 1)$ est le vecteur des valeurs phénotypes pour n individus, $b(q\times 1)$ est le vecteur des effets fixes (p.e. saison, troupeau...), $X(n\times q)$ est la matrice d'incidence reliant le phénotype de chaque individu aux effets fixes correspondants, $u(m\times 1)$ est le vecteur des effets polygéniques de chaque animal (ou encore appelé mérite génétique, effet animal ou BV), $Z(n\times m)$ est la matrice d'incidence reliant le phénotype de chaque individu à son effet polygénique correspondant et $e(n\times 1)$ est le vecteur des résidus. Ce modèle très connu, appelé modèle animal, est un modèle mixte car il combine à la fois des effets fixes et des effets aléatoires. L'intérêt de ce type de modèles est de pouvoir estimer plus d'effets que l'on a d'observations. Pour cela on restreint l'espace des solutions en appliquant certaines contraintes:

(i) on suppose que les « effets animaux » sont tirés dans une distribution normale de moyenne μ=0 et de variance σ_A^2, (ii) par ailleurs on impose également que la covariance entre deux effets animaux soit égale à deux fois le coefficient de parenté (encore appelé apparentement) multiplié par σ_A^2 pour. Autrement dit, on suppose que la distribution jointe des effets animaux est une distribution normale multivariée de moyenne μ=0 avec une matrice de variance-covariance égale $A\times\sigma_A^2$. A, la matrice d'apparenté, contient les coefficients de parenté.

Les solutions pour les effets fixes encore appelées BLUE (*Best Linear Unbiased Estimates*), correspondent à une estimation par moindres carrés généralisées:

$$\hat{b}=(X'V^{-1}X)^{-1}X'V^{-1}y$$

Les solutions pour les effets animaux, appelées aussi BLUP (*Best Linear Unbiased Predictors*) sont obtenues en minimisant la prédiction de la variance résiduelle:

$$\hat{u}=\sigma_A^2 AZ'V^{-1}(y-X\hat{b})$$

Dans les deux expressions précédentes V correspond à la matrice de variance-covariance des observations y, laquelle peut s'écrire de la façon suivante:

$$V=\sigma_A^2 ZAZ'+R$$

, où R est la matrice variance-covariance des effets résiduels. Si on suppose que la covariance entre les effets résiduels est nulle, on peut écrire $R=I\sigma_E^2$, où I est la matrice d'identité.

Ces solutions imposent d'inverser la matrice V, ce qui peut poser des difficultés étant donné parfois la taille des données utilisés. Une alternative pour obtenir ces solutions est de résoudre les équations des modèles mixtes d'Henderson[55]:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z+\frac{1}{\sigma_A^2}A^{-1} \end{bmatrix}\begin{bmatrix}\hat{b}\\ \hat{u}\end{bmatrix}=\begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

Il faut souligner que ces équations très célèbres en génétique animale ont été proposées à une époque où l'on ne disposait pas des mêmes moyens informatiques qu'à l'heure actuelle. Quand on ajoute des effets aléatoires supplémentaires à ce modèle animal de base, les équations d'Henderson ne sont pas forcément la façon la plus simple pour obtenir les solutions de ces modèles mixtes (par exemple: il est possible de rencontrer des problèmes de singularité avec les matrices devant être inversées pour résoudre ces équations).

Il y a encore deux paramètres inconnus pour obtenir les solutions d'un modèle animal σ_A^2 et σ_E^2. Il existe différentes approches pour tenir compte de ces paramètres inconnus: (i) la plus simple est d'émettre une hypothèse sur le rapport $\alpha=\frac{\sigma_A^2}{\sigma_E^2}$, on peut alors récrire ces équations d'Henderson en fonction de ce rapport α:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z+A^{-1}\alpha \end{bmatrix}\begin{bmatrix}\hat{b}\\ \hat{u}\end{bmatrix}=\begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

(ii) La seconde est une approche particulière de type maximum de vraisemblance appelée REML (*Residual maximum likelihood estimation*). Le problème avec les estimateurs obtenus avec une approche type maximum de vraisemblance (ML, *maximum likelihood*) classique est illustré en regardant l'estimateur obtenu par ML pour la variance d'une distribution normale qui est: $\frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)^2$. Cet estimateur de la variance est une fonction du paramètre μ, qui est inconnu. Pour une estimation standard de la variance on remplace dans cette expression μ par la moyenne de l'échantillon et on divise par N-1 au lieu de N. On divise par N-1 pour tenir compte de l'incertitude sur la vraie valeur de, μ Dans une approche de type REML, ce problème des estimateurs des composantes de variance biaisés par les estimateurs des effets fixes est résolu par une transformation linéaire qui enlève les effets fixes du modèle. On définit une matrice K tel que KX=0. On a alors:

$$y^c=Ky=KZ'u+Ke$$

$$V(y^c)=KZAZ'K'\sigma_A^2+KRK'\sigma_E^2$$

I.4.1.3.d.2 Les études de liaison modélisant les effets gamétiques comme des effets aléatoires.

Le modèle ci-dessus a été étendu pour des analyses de liaison en ajoutant un effet aléatoire h lié aux gamètes parentaux reçus par un descendant ainsi que Zh sa matrice d'incidence:

$$y = Xb + Z_u u + Z_h h + e$$

Comme pour les effets polygéniques, on suppose que les effets gamétiques suivent une distribution normale multivariée de moyenne μ=0 et avec comme matrice-covariance $H \times \sigma^2_H$. La matrice Hp est la matrice des relations gamétiques, autrement dit, elle correspond à la matrice des probabilités IBD entre paires d'haplotypes à une position p donnée. Il existe différentes approches pour calculer cette matrice, dont certaines d'entre elles font appel à des techniques de type MCMC* (*Markov chain Monte Carlo*) pour exploiter les relations gamétiques d'individus n'ayant pas été génotypés.

Ces approches exploitant le modèle animal n'ont pas joué un rôle important jusqu'ici: tous les QTL découverts jusqu'à présent pouvaient l'être avec les approches plus simples d'analyse de liaison décrites précédemment. Cependant elles ont le mérite de pouvoir aisément intégrer n'importe quel type d'information comme par exemple du DL.

I.4.2 Les études d'association.

I.4.2.1 Introduction.

* Les méthodes MCMC sont une classe de méthodes d'échantillonnage à partir de distributions de probabilité. Ces méthodes se basent sur le parcours de chaînes de Markov qui ont pour lois stationnaires les distributions à échantillonner.

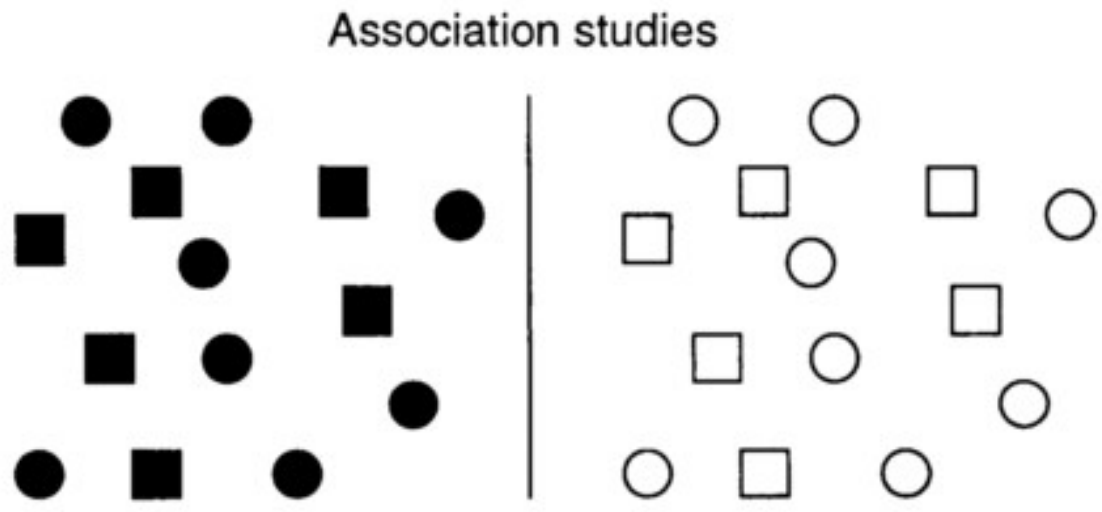


Figure I.8: *Le principe des études d'association est de regarder si un allèle est plus fréquent (ou moins fréquent) parmi les cas que parmi les contrôles. Figure provenant de l'article de Lander et Shork* [207]

Chez l'homme, les généticiens se sont rapidement heurtés aux limites des études de liaison en tentant de les appliquer à des maladies courantes telles que le diabète, les maladies cardio-vasculaires ou le cancer. Ces échecs, expliqués par le mode d'hérité complexe de ces maladies, ont poussé les généticiens vers une autre stratégie: les études d'association. Contrairement aux études de liaison, les études d'association ne s'intéressent pas aux patterns d'hérédité familiale d'un caractère, mais recherchent plutôt si un allèle particulier est plus fréquent (ou moins fréquent) dans un groupe d'individus atteints par rapport un groupe d'individus contrôles.

Les études d'association ont d'abord été employées dans une approche de type gène candidat[10,11]. Toutefois, il a fallu attendre plusieurs décennies pour voir le développement d'un certain nombre d'outils génétiques (pour rappel, mais déjà évoqués plus haut: développement d'un catalogue exhaustif des variations génétiques courantes et étude des patterns de DL chez l'homme avec le projet HapMap[51], développement de plate-forme de génotypage à haut débit et diminution des coûts de génotypage[25]) permettant de rechercher de manière systématique à travers tout le génome des associations entre des polymorphismes communs et des maladies communes. L'idée de base des GWAS peut paraître simple: rechercher à travers un catalogue de variations génétiques courantes (ayant un contenu en information suffisante pour couvrir l'ensemble du génome), celles pouvant être impliquées dans une maladie en comparant les fréquences alléliques entre un groupe d'individus atteints et un groupe d'individus contrôles. Cependant l'interprétation des résultats d'une GWAS peut

paradoxalement s'avérer être extrêmement complexe. En effet quand on détecte une association entre un SNP et une maladie, il peut y avoir 3 raisons à cela: (i) le SNP est une mutation causale (ii) le SNP est en déséquilibre de liaison avec la mutation causale (iii) une fausse association entre la maladie et un SNP.

Cette dernière résulte le plus souvent de problèmes de stratification dans la population dans laquelle les cas et les contrôles sont échantillonnés: une grande population non isolée est constituée de sous-populations d'origines ethniques différentes, si la maladie étudiée est présente à une fréquence plus élevée dans une de ces sous-populations, on trouvera systématiquement une association avec n'importe quel SNP présentant une distribution allélique différente entre cette sous-population et le reste de la population. Ces problèmes de stratification combinés aux problèmes de la multitude d'hypothèses testées dans les GWAS ont longtemps contribué à décrédibiliser les résultats obtenus avec ce type d'étude et ont été un challenge pour le généticien-statisticien.

Les études d'association ont un autre avantage majeur sur les études de liaison: les régions identifiées dans les études d'association sont beaucoup moins grandes (quelques kb) que dans les études de liaison (quelques Mb). En fait, les études d'association ont longtemps été considérées en cartographie fine comme une alternative bon marché au séquençage de cas et de contrôles visant à identifier les mutations causales dans les régions préalablement identifiées dans des études de liaison. Les généticiens ont tenté de tirer parti de cet avantage chez les espèces de production en développant des méthodes de cartographie fine de QTL exploitant simultanément la liaison et le déséquilibre de liaison.

I.4.2.2 Design d'une étude d'association génome-entier.

Si l'on veut éviter qu'une GWAS ne soit qu'un gaspillage de temps et d'énergie, il est indispensable de se poser un certain nombre de questions sur le design de l'étude avant d'entreprendre une quelconque action. En effet les choix préalables à une étude GWAS sont déterminants pour la puissance statistique de l'étude.

I.4.2.2.a Choix des cohortes.

Tout d'abord on doit s'intéresser aux critères de sélection des cas et des contrôles. Le choix des cas doit être guidé par cette idée d'enrichir les cas pour des allèles spécifiques prédisposant à la maladie. Pour cela on peut tenter de limiter l’hétérogénéité génétique en sélectionnant des cas extrêmes ou en se focalisant sur des cas familiaux. Pour les contrôles, il arrive très souvent que ces individus soient des donneurs de sang opportunistes n'ayant subi aucun examen médical préalable et servant de cohorte contrôle pour diverses études cas-contrôles pour différentes maladies. Ceci peut poser différents problèmes: Il faudra par exemple se prémunir d'éventuels

faux positifs dus à des problèmes de stratification en prenant garde que la cohorte contrôle ne divergent pas trop génétiquement des cohortes de cas (on testera un éventuel excès d'association positive, et l'on écartera les individus génétiquement trop divergents). Par ailleurs on veillera à ce que des individus n'aient pas été erronément considérés comme des contrôles, ce qui peut conduire à diminuer la puissance de l'étude. Ce problème survient essentiellement pour des désordres ayant une prévalence élevée, p.e. hypertension ou obésité. On peut remédier à cela en choisissant ses contrôles scrupuleusement, mais en tenant compte aussi du fait que choisir des contrôles extrêmes peut amener à sélectionner des individus atteints d'autres types de désordres (par exemple, choisir des individus maigres qui peuvent être atteints d'une maladie chronique dans une étude sur l'obésité).

I.4.2.2.b Taille des cohortes.

On doit par la suite s'interroger sur le nombre de cas et de contrôles que l'on veut génotyper sur une puce à haute densité en marqueurs. La réponse à cette question dépendra du type d'effet que l'on souhaite mettre en évidence. Les résultats des premières GWAS ont cependant montré que les variation génétiques détectés étaient la plupart du temps associées à des effets modestes, donc nécessitant des cohortes comprenant plusieurs milliers d'individus. Ainsi si l'on considère le seuil classique d'acceptation d'une association dans une GWAS (puce contenant 500,000 SNPs) c'est-à-dire une valeur $p = 5x10^{-8}$, (équivalent à une valeur p = 0.05 après une correction de Bonferonni), il faudra génotyper 6000 contrôles et 6000 cas pour détecter dans 80% des cas une association avec un allèle de susceptibilité ayant un MAF = 15% et un odds-ratio 1.25. Pour éviter d'avoir à génotyper autant d'individus, la plupart des GWAS adoptent une stratégie multiétapes: la première étape consiste à tester une association pour un nombre limité de cas et de contrôles (p.e. 1000 cas et 1000 contrôles) et de considérer toutes les associations au-dessus d'un seuil volontairement laxiste qui laissera passer à la fois des vraies associations, mais aussi majoritairement beaucoup de fausses associations (p.e. avec un seuil de 5%, on s'attend à détecter sous l'hypothèse H0 25,0000 associations (sur une puce de 500,000 SNP) dont seulement une faible fraction sont vraies). Dans l'étape suivante on teste des associations en génotypant des nouvelles cohortes ayant une taille égale voir supérieure aux cohortes utilisées dans la première étape, mais seulement pour des SNP ayant passé le filtre de la première étape. On adopte cette fois un seuil plus strict pour garder uniquement les vraies associations. Bien évidemment cette stratégie multiétapes soulève de nouvelles questions comme le choix des individus à génotyper lors de la seconde étape.

I.4.2.2.c Choix relatifs aux techniques de laboratoire.

On peut ensuite se poser des questions sur le choix de la plate-forme de génotypage et donc des SNPs à génotyper. Pour une couverture optimale du génome, les marqueurs génotypés doivent être en DL avec la mutation causale. Le choix des tags SNPs et de leur nombre dépendra de la méthode employée ainsi que de la population d'origine dont sont issues les cohortes. Par exemple du fait que le DL s'étend sur des distances moins longues dans les populations Africaines, on sait qu'il faudra pratiquement le double de SNP pour des cohortes issues de ces populations que pour des cohortes issues de populations européennes. Il faut toutefois souligner que le HapMap II[38] a montré que les deux principales plates-formes de génotypage (Illumina, Affymetrix) avaient une couverture génomique suffisante dans les populations caucasiennes. À l'heure actuelle, on voit arriver des plates-formes mixtes permettant de génotyper simultanément des SNPs et CNVs. Cependant, contrairement aux SNP, on ne dispose pas d'une liste exhaustive de CNVs pour le génome humain, il est donc difficile d'évaluer l'apport de ce type de plate-forme.

Autre point à souligner concernant les techniques de laboratoire: le traitement différent (technique d'extraction différente ou plusieurs plates-formes de génotypage) que peuvent subir des échantillons issus de cas et de contrôles pouvant être l'origine d'artefact. Par exemple si la cohorte des cas est génotypée sur une plate-forme où l'attribution des génotypes est biaisée vers le génotype hétérozygote, alors le statisticien peut être amené erronément à conclure à une association.

I.4.2.3 Analyses préliminaires.

I.4.2.3.a Des données brutes aux génotypes et les contrôles de qualité.

Les analyses statistiques dans les GWAS sont systématiquement précédées d'une batterie de tests permettant de passer des données brutes à des données ayant une qualité suffisante permettant ainsi d'éviter des artefacts pouvant conduire à des conclusions erronées.

Ces analyses préliminaires débutent par la transformation des données expérimentales en génotypes (Figure I.9). Étant donné la quantité monumentale de données à traiter, cette transformation a nécessité le développement de méthodes automatisées. La plupart des logiciels actuels n'attribuent pas directement à une mesure expérimentale un génotype discret, mais fournissent une probabilité pour chaque génotype possible en fonction des observations. Les critères de qualité de ce type de logiciel sont la précision liée à la capacité à bien distinguer les trois groupes de génotype pour un SNP et le *call-rate*. Ce dernier correspond pour un SNP au % d'individus pour

lesquels on peut attribuer un génotype.

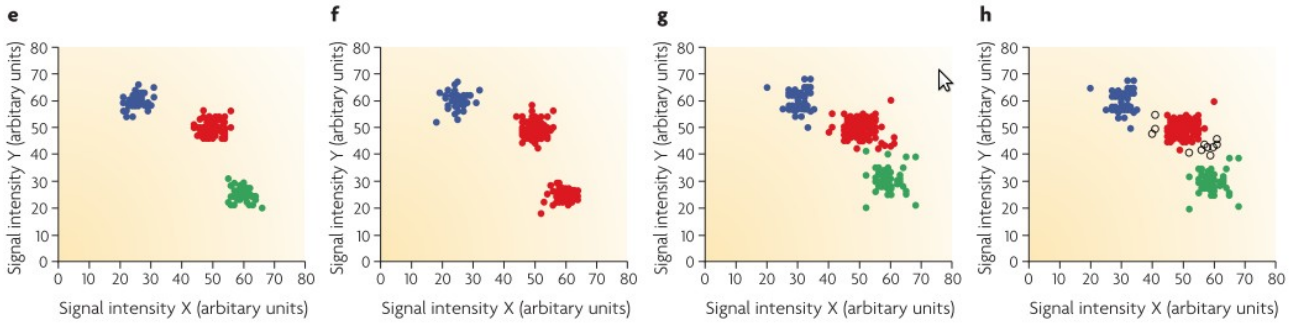


Figure I.9: *Procédure pour attribuer des génotypes pour un SNP donné. Les données brutes de 200 génotypes, niveaux d'intensité lumineuse sont représentées avec un graphe de type plot avec un axe des x pour le premier allèle et un axe des y pour le second allèle. Dans le premier graphique, les trois clusters sont correctement définis. Dans les autres graphiques différents problèmes dans l'attribution des génotypes surviennent. Figure provenant de l'article de McCarthy et al.* [205].

Le *call-rate* est un point capital comme critère de qualité des génotypes aussi bien des SNPs que des individus. En effet un *call-rate* trop faible peut amener de fausses associations du fait qu'un génotype est davantage systématiquement manquant que les deux autres. L'étape suivant l'attribution des génotypes sera de déterminer des seuils de *call-rate* et d'écarter les individus et les SNPs avec des *call-rate* trop bas.

Un autre test visant à améliorer la qualité des données est d'identifier et d'enlever des analyses ultérieures les SNPs en déséquilibre d'HW chez les contrôles. On suppose ici que des erreurs de génotypage sont responsables de ces déviations dans l'équilibre HW. Toutefois un déséquilibre d'HW peut se produire dans le cas d'une délétion ou d'une duplication lesquelles ayant peut-être elles-mêmes un rôle essentiel dans étiologie de la maladie. Actuellement on écarte systématiquement un SNP en déséquilibre HW sans se préoccuper de ces dernières considérations.

On teste généralement avec un test de Pearson (test chi²) l'équilibre HW en comparant les fréquences génotypiques observées avec les fréquences attendues en cas d'équilibre HW. Cependant quand les comptes génotypes sont trop faibles, il est recommandé de remplacer ce test par un test exact de Fisher.

Quand on dispose du sexe des individus, on peut confronter ces données aux génotypes sur le chromosome X. Ceci permettra d'une part de contrôler la qualité des SNPs sur le chromosome X et d'autre part de mettre évidence des erreurs dans la classification du statut d'un individu (malade ou contrôle).

Par ailleurs, il faudra systématiquement écarter des analyses statistiques les individus qui divergent trop génétiquement du reste du groupe. On pourra aussi mettre en évidence, en calculant les probabilités IBS, des relations de parenté qui étaient ignorées jusqu'à présent, pouvant également être une source de faux positifs.

I.4.2.3.b Imputation des données manquantes et phasage.

Il existe différentes raisons pour vouloir imputer des données manquantes: (i) nécessité de réaliser ultérieurement des analyses multipoints (ii) analyser des données provenant de plateforme de génotypage différente. Les logiciels d'imputation prédisent un génotype manquant en fonction des génotypes des SNPs voisins. La fiabilité de ce type de logiciel dépendra des niveaux de DL dans la région où se trouve le SNP avec les génotypes manquants. On distingue deux types de logiciels: ceux donnant un résultat unique d'imputation en utilisant par exemple des méthodes du type maximum de vraisemblance, de ceux sélectionnant un génotype sur base des probabilités associées à chacun des trois génotypes possibles. Cette dernière approche permet d'investiguer l'impact des imputations sur les analyses ultérieures. Par ailleurs la plupart des logiciels d'imputation supposent que le fait qu'un soit génotype soit manquant est événement indépendant du vrai génotype ainsi que du phénotype de l'individu. Ceci n'est pas toujours le cas: il arrive très souvent d'avoir davantage de données manquantes pour les hétérozygotes que pour les homozygotes ou encore on peut avoir une distorsion dans les données manquantes entre les cas et les contrôles si ceux-ci n'ont pas été génotypés sur la même plate-forme par exemple.

Quand on regarde les génotypes d'un individu au niveau de plusieurs SNPs dans une région, ils résultent de l'association de deux combinaisons d'allèles l'une et l'autre étant portées respectivement par les chromosomes d'origine maternelle et paternelle. Il est possible de déterminer ces combinaisons d'allèles appelées haplotype sur des individus non apparentés en se basant sur les caractéristiques du DL chez l'homme vu dans le paragraphe du projet HapMap, c'est-à-dire qu'il existe des régions avec des taux de recombinaison bas dans lesquelles il existe peu d'haplotypes dans la population. Les programmes exploitant ce principe fonctionnent bien quand d'une part le pourcentage de données manquantes est faible et quand d'autre part on dispose d'une carte dense en SNP. L'utilisation de ce type de logiciel est dictée par le besoin de réaliser des études d'association avec des haplotypes plutôt que sur des génotypes. À noter que beaucoup de logiciels d'imputation permettent également de phaser des données (ex.: PHASE[53]; FASTPHASE[54]). Le choix de phaser les données des cas et des contrôles ensemble ou séparément reste un sujet controversé. Phaser les données des cas et des contrôles ensemble peut induire un biais vers l'hypothèse de la non-existence d'une association entre les haplotypes et la maladie et donc de diminuer la puissance de l'étude. D'un autre côté, phaser les données des cas et des contrôles séparément peut gonfler le pourcentage d'erreur de type I.

I.4.2.3.c Évaluation des niveaux de déséquilibre de liaison et estimation des taux de recombinaison.

En principe le choix des SNPs génotypés a été préalablement réalisé en fonction des niveaux de DL dans les populations dont sont issues les cohortes, afin de refléter l'information de tous les SNPs courants présents dans le génome. Cependant ce type d'étude se réalise sur des populations proches, mais qui ne sont pas forcément les populations desquelles sont issues nos cohortes de cas et de contrôles. Afin d'estimer la couverture génomique dans le cas spécifique de nos cohortes et donc d'évaluer du même coup la puissance d'une GWAS, il est nécessaire de réaliser une étude de liaison en estimant soit un r^2 soit un D' entre des paires de marqueurs (voir plus haut pour les définitions et propriétés de ces mesures). On peut représenter les niveaux de DL dans un diagramme en couleurs représentant les variations dans les niveaux de DL entre paires de marqueurs.

I.4.2.4 Analyses statistiques des études d'association génome entier.

I.4.2.4.a Les études simple point.

I.4.2.4.a.1 Phénotype cas-contrôles.

Il existe différents tests pour mettre en évidence une association entre un SNP et la maladie. Les avantages et inconvénients de chacun de ces tests dépendent principalement du type d'effets des allèles d'un SNP sur la maladie: additifs, dominants ou récessifs. Le test le plus simple et le plus intuitif est le test de Pearson à 2df (ou un test exact de Fisher), où l'on compare les fréquences génotypiques observées chez les cas et les contrôles avec celles attendues sous l'hypothèse qu'il n'existe pas de différence entre les cas et les contrôles. Ce test a une puissance raisonnable pour différents types d'effets. Cependant il existe des tests avec une meilleure puissance dans le cas d'effets additifs comme par exemple un test de Pearson basé cette fois sur les fréquences alléliques. Toutefois ce test présente deux inconvénients majeurs: d'une part il suppose un équilibre HW combiné chez les cas et les contrôles et d'autre part il est difficile à interpréter du point de vue de l'estimation du risque. Une alternative à cette approche est le test de Cochran–Armitage, où on regarde dans une régression la proportion des cas pour chacun des 3 génotypes encodés en 0,1,2 (1 pour le génotype hétérozygote). S'il n'y pas d'association, on s'attend à une pente de régression $\beta=0$, tandis que dans le cas d'un SNP ayant des effets purement additifs, on pourra *fitter* parfaitement l'estimation des risques des 3 génotypes avec une droite de régression. L'avantage de

ce test par rapport au précédent est qu'il n'émet aucune hypothèse sur l'équilibre HW d'un SNP. Cependant il est déficient pour détecter des effets de sur-dominance. Il n'existe pas de test ayant une puissance constante quel que soit le type d'effet. Une approche consiste à tester plusieurs modèles et à pondérer les résultats de ces différents tests selon ce que l'on attend en termes d'effet (par exemple on donnera plus de poids à un test recherchant des effets additifs). Une autre approche est de réaliser un test d'Armitage quand le MAF est bas et un test exact de Fisher quand on a suffisamment d'individus pour chacun des 3 génotypes. Enfin on pourrait imaginer une approche Bayesienne dans laquelle le généticien exprime ses hypothèses sur la nature des risques de la maladie à l'intérieur de distributions prior. Ce type d'approche ne joue pour le moment aucun rôle dans les études d'association.

I.4.2.4.a.2 Phénotype continu.

Pour un phénotype continu (p.e. mesure de la tension artérielle), les outils statistiques les plus répandus pour tester une association avec les génotypes d'un SNP sont l'ANOVA 2 (analogue au test de Pearson avec 2df) et la régression linéaire (1df). Les deux tests supposent que pour chaque génotype le phénotype soit distribué normalement avec une variance similaire. Dans le cas d'un phénotype, où ces hypothèses de normalité et d'homosédasticité ne sont pas respectées, une solution est de réaliser la transformation logarithmique du phénotype. L'approche communément admise est généralement de comparer un modèle ANOVA avec une régression, la régression étant à son tour comparée au modèle nul où l'on n'a pas d'association. On garde le modèle le plus simple qui ne montre pas de différence significative avec un modèle plus général.

I.4.2.4.a.3 La régression logistique.

Il n'est pas possible d'appliquer directement une régression linéaire à un phénotype de cas-contrôles car ce dernier n'est pas distribué normalement et d'autre part les prédictions de probabilité d'être atteint pourraient tomber en dehors du range 0-1. Toutefois il existe une approche plus sophistiquée apparentée à la régression linéaire pour tester une association entre un SNP et un phénotype de type cas-contrôle: la régression logistique. Dans ce type de régression on opère une transformation en logit $logit(\pi)=\log(\pi/(1-\pi))$ du risque d'être atteint pour un individu i (pii). On peut ensuite proposer un modèle linéaire dans lequel le génotype d'un individu permet de prédire sa valeur de logit:

$$logit(\pi)=\beta_0\times x_0+\beta_1\times x_1+\beta_2\times x_2$$

Il est possible de comparer la vraisemblance de ce modèle avec la vraisemblance où on contraint tous les

coefficients à l'égalité ($\beta_0 = \beta_1 = \beta_2$). On peut tester également un modèle plus spécifique comme un modèle additif en contraignant cette fois le coefficient β_1 des hétérozygotes à être égal à la moitié d'un des coefficients correspondant à un génotype homozygote. Pour un modèle récessif et un modèle dominant, on posera respectivement $\beta_0 = \beta_1$ et $\beta_1 = \beta_2$.

La régression n'apporte rien de plus par rapport aux méthodes d'association simple point vues plus haut: pour chaque modèle de régression logistique il existe son homologue dans test score, plus simple et plus rapide à calculer. Cependant la régression logistique a le mérite d'être un outil flexible permettant d'ajouter des effets ou covariables (ex.: sexe) supplémentaires.

I.4.2.4.b Les études multipoints.

En examinant une éventuelle association marqueur par marqueur avec la maladie, on n'exploite pas l'information de leur distribution jointe. On pourrait pourtant tirer un bénéfice en terme de puissance statistique en exploitant cette information, sauf dans deux cas extrêmes: (i) une carte de trop faible densité pour avoir du DL entre SNP (ii) une carte très dense, où la mutation causale appartient peut-être aux SNPs génotypés. Dans les approches multipoints, on distingue celles basées sur les génotypes de celles basées sur les haplotypes.

I.4.2.4.b.1 Régression logistique multi-SNP.

Comme souligné plus haut il est facile d'étendre une régression logiste avec un SNP à plusieurs. Il est notamment possible de tester un modèle additif en posant $\beta_1=(\beta_0+\beta_1)/2$. Le "score test" correspondant est un test Armitage à n SNP appelé Hotelling T^2 test. Par ailleurs il est possible de tester dans une régression logistique multi-SNP des effets épistatiques. Cependant il faut souligner que le fait d'ajouter des effets supplémentaires peut avoir un coût en terme de puissance. C'est d'ailleurs pour cette raison que différentes stratégies pour limiter le nombre de SNP tester dans le modèle ont été développées.

I.4.2.4.b.2 Approches basées sur les haplotypes.

Les méthodes multi-SNPs souffrent d'avoir beaucoup de variables prédictives hautement corrélées. Une alternative, basée sur la structure en bloc haplotypiques des pattern de DL chez l'homme est d'exploiter des haplotypes dans des études d'association pour surmonter ce problème de SNP hautement corrélé. Cette approche permet de réaliser des analyses avec peu de degrée liberté. Toutefois l'intérêt de ce type de méthode est affaiblie par le fait que les SNPs génotypés sont préalablement sélectionnés par tagging-SNP.

Les études d'association basées sur des haplotypes posent plusieurs problèmes:

(i) tout d'abord ce type d'approche demande préalablement de phaser les données ce qui d'une part est lourd d'un point vu temps calcul et d'autre part pose le problème de prendre en compte l'incertitude associé à chaque haplotype dans des analyses ultérieure Une alternative est de déterminer les fréquences haplotypiques via des approches de types maximum de vraisemblance

(ii) Le second problème qui se pose est au niveau des analyses statistiques: si on travaille avec des haplotypes que faire des haplotypes rares ? quelles sont les limites des blocs haplotypiques ? certains haplotypes sont plus proches que d'autres d'un point évolutif: comment prendre en compte l'histoire de ces haplotypes dans des études d'association ? Une solution est de regrouper ces haplotypes dans des cluster par des approches retraçant l'histoire de ces haplotypes.

(iii) Par ailleurs toutes les études d'association basés sur des haplotypes en comparant la fréquences chez les cas par rapport aux contrôles font des suppositions sur l'équilibre HW ou sur le fait d'avoir des effets additifs. Une alternative aux analyses comparant les fréquences haplotypiques chez les cas et les contrôles est de regarder si il existe un excès d'*allèle sharing* parmi les haplotype des cas par rapport à ceux des contrôles. Un avantage de ce type d'approche est de pouvoir prendre en compte l'incertitude associée à chaque phase.

I.4.2.5 Approches permettant de régler les problèmes de stratification dans les études d'association.

Toutes les approches permettant de corriger d'éventuels effets de stratification supposent qu'on dispose d'au moins une centaine de SNP non liés qui ne jouent aucun rôle dans la maladie que ce soit d'une manière directe ou d'une manière indirecte en étant en DL avec un des SNP candidats.

I.4.2.5.a Approches exploitant les génotypes des parents.

Une approche simple qui permet de se prémunir des problèmes de stratification est d'utiliser les chromosomes des parents sains qui n'ont pas été transmis à leurs descendants atteints comme contrôles.

Une autre approche pour se prémunir des problèmes de stratification est de réaliser un test de TDT (transmission disequilibrium test). L'idée sous-jacente au test TDT, développé par Spielman et al[59] est de dire que si un allèle marqueurs est proche (en terme de distance génétique) d'un allèle susceptibilité, il aura alors tendance à être

surtransmis à la descendance atteinte comparé à l'autre allèle marqueur présent chez des parents hétérozygotes.

Table I.3: *Exemple de test TDT avec un marqueur bi-allélique.*

		Transmis	
		1	**2**
Non-Transmis	**1**	N_{11}	N_{21}
	2	N_{12}	N_{12}

On regarde si un allèle marqueur est davantage transmis que l'autre chez des parents hétérozygotes en utilisant le test suivant:

$$X_1^2 \sim \frac{N_{12}-N_{21}}{N_{12}+N_{21}}$$

I.4.2.5.b Approche de type contrôle génomique (GC = "Genomic Control").

Une manière simple de corriger des tests d'association pour d'éventuels effets de stratification est d'adopter une approche de type contrôle génomique (GC): approche dans laquelle on cherche à (i) évaluer le biais introduit par de la stratification dans la distribution statistique de tests d'association et (ii) à corriger ce biais.

Quand on regarde la distribution statistique de tests d'association pour des SNP, elle doit suivre une distribution de type X^2_{1ddl}. Or $E(X^2_{1ddl})=1$. En cas de problème de stratification on peut observer un déplacement de la courbe de distribution des tests d'association vers la droite et $E(X^2_{1ddl})>1$. Pour n SNP, on peut calculer Y^2 associé au test d'Armitage. Quand il n'y a pas de problème de stratification alors $Y^2 \sim X^2_{1ddl}$. En cas de stratification, on suppose que la statistique Y^2 est gonflée d'un facteur λ (appelé facteur d'inflation) et que donc $Y^2/\lambda \sim X^2_{1ddl}$. Il existe différentes méthodes pour estimer λ, mais une estimation très répandue est de prendre la médiane de la distribution empirique des valeurs d'un test d'Armitage pour des SNP nul et de diviser par sa valeur attendue en supposant que Y^2 suit une distribution de X^2_{1ddl}. On corrige la valeur Y^2 pour SNP candidat en divisant leur Y^2 par lambda.

Les inconvénients de cette approche sont: (i) d'être limitée à des analyses simple point (ii) de corriger de manière constante des effets des problèmes de stratification quel que soit le SNP (parfois la correction GC sera trop

conservative parfois pas suffisamment).

I.4.2.5.c Approche basée sur de l'inférence de la structure de la population.

L'idée de base de ce type d'approche est de relier le génome de chaque individu à des sous-populations et de tester une association conditionnellement à la répartition dans les sous-populations.

L'étape fastidieuse dans ce type d'approche est d'inférer la structure de la population. Ce type de méthode suppose 3 choses: (i) K populations fondatrices qui sont mélangées dans la population dans laquelle sont issues nos cohortes. (ii) Les n SNP utilisés dans ces K populations sont en équilibre HW et sont non liés. (iii) L'allèle au niveau d'un *locus* pour un marqueur provient d'une de ces populations fondatrices. On cherche à calculer pour chaque allèle de chaque individu la probabilité qu'il provienne de la population k, $q(i)_k$ (ces probabilités sont notées dans un vecteur Q). On note également $x(i,a)_l$ la copie allélique (a) pour l'individu a au niveau du *locus* l (l'ensemble des valeurs x est repris dans un vecteur X), $z(i,a)_l$ la population d'origine de la copie allélique $x(i,a)_l$ (vecteur Z) et enfin p_{klj} la fréquence de l'allèle j au niveau du *locus* l dans la population k. Le logiciel STRUCTURE, (la première méthode développée pour inférer la structure de la population) utilise une approche Bayésienne, où il cherche la distribution a posteriori des paramètres P,Q,Z. Il utilise des techniques de type MCMC pour estimer la distribution jointe de ces trois paramètres. Après cette étape d'inférence, on dispose pour chaque individu i de la proportion de son génome originaire de la population k à l'intérieure du vecteur Q. On peut alors comparer pour chaque SNP candidat, la vraisemblance des génotypes sous l'hypothèse nulle (pas d'association entre un gène candidat et un phénotype) avec la vraisemblance des génotypes sous l'hypothèse alternative.

Le problème de ce type d'approche est l'étape d'inférence de la structure de la population qui suppose de connaître le nombre K de sous-populations.

I.4.2.5.d Approche de type modèle mixte.

Une alternative développée par Yu *et al.* [60] est d'utiliser une approche déjà vue chez les espèces de production: les modèles mixtes

$$y = X\beta + S\alpha + Q\nu + Z\mu + e$$

On peut modéliser le phénotype y des individus comme une somme d'effets fixes tels le sexe β et les effets de sous-population γ et d'effets aléatoires comme les effets des SNP candidats α et les effets polygéniques μ.

I.4.2.5.e Approche basée sur une modélisation du génome en composante principale (PCA).

Cette approche nécessite de disposer d'un grand nombre de marqueurs répartis sur l'ensemble génome. À partir de cette information, on cherche à estimer des facteurs corrigeant pour la structure de la population. Ces facteurs seront utilisés ultérieurement pour ajuster les génotypes et les phénotypes dans des études d'association. Les facteur de correction sont calculés par une analyse en composante principale à partir de la matrice des génotypes (individus en colonne, les marqueurs dans les lignes et les génotypes sont annotés 0,1,2). L'analyse en composante principale consiste à transformer des variables liées entre elles en de nouvelles variables indépendantes les unes des autres, appelées composante principale. Cette approche de type PCA est appliquée à la matrice de covariance des génotypes, pour calculer les différents axes continus (vecteurs propres de matrice covariance) de la variation génétique, permettant ainsi de réduire les données à un petit nombre de dimensions.

I.4.2.6 Les études d'association dans le cadre de caractères quantitatifs chez les espèces de production.

Les études de liaison pour des caractères quantitatifs dans des pedigrees commerciaux ou dans des croisements expérimentaux chez les espèces de productions aboutissent le plus souvent à détecter un QTL dans un intervalle de confiance s'étendant sur des dizaines de centimorgans. Bien évidemment, ce niveau de résolution est insuffisant pour identifier le gène ou la mutation causale impliquée dans le caractère examiné dans une étude de cartographie QTL. Les raisons associées à ce niveau de résolution insuffisant sont: (i) une densité en marqueurs trop faible, (ii) une densité trop faible en CO dans la région d'intérêt, (iii) une pénétrance incomplète du phénotype étudié, (iv) ou encore plusieurs QTL influençant le caractère étudié dans l'intervalle marqueur considéré.

Différentes solutions ont été proposées pour résoudre ces problèmes de résolution: ces dernières années, notamment les cartes de densité moyenne en marqueurs microsatellites (comportant des centaines de marqueurs) ont été progressivement remplacées par des cartes à haute densité en marqueurs de type SNP (comportant des milliers de marqueurs). Pour augmenter la densité en CO dans la région d'intérêt, il a été proposé d'exploiter les recombinants historiques (événement de recombinaison chez les ancêtres des membres du pedigree disponible dans une étude de cartographie) en utilisant le signal de DL. Il est à noter que contrairement à l'homme, où le DL s'étend sur quelques kilobases, chez ces organismes en grande partie du à la faible taille de l'effectif efficace, il

est possible de détecter du DL sur plus 1 Mb. Le corollaire est qu'il possible de mener des études d'association dans ce type de population avec des cartes bien moins dense en marqueur que chez l'homme.

Pour exploiter le signal de déséquilibre de liaison, afin d'augmenter la résolution des études de cartographie QTL chez les espèces de production, il a été proposé de combiner des analyses de liaison avec des études d'association.

Ces approches exploitent le fait que le DL s'étend sur une distance inhabituellement longue chez les espèces de production, ce qui se traduit notamment par le fait que les individus partagent de long haplotypes provenant d'un ancêtre commun. Plusieurs études ont montré qu'il était possible d'une part de convertir cette information en une mesure appelé probabilité d'être identique par descente calculé entre chaque paire d'haplotype à une position p donnée et d'autre part d'utiliser cette matrice de probabilité IBD dans la covariance des effets haplotypes pour estimer d'éventuels effets (aléatoires) QTL à la position p via une approche de type REML [61]. Meuwissen et Goddard ont proposé une approche basée sur un modèle coalescent pour calculer la probabilité qu'une paire de segment chromosomique soient identiques par descente à une position A donnée connaissant l'état IBS des marqueurs dans la région étudié [62]. Cette approche pour déterminer si deux haplotypes ont coalescé plus ou moins vite tient compte de l'effectif efficace et des distances génétiques entre la position étudiée et les marqueurs flanquant.

Par la suite si l'on désire utilisé ces probabilités IBD pour améliorer la puissance et la précision de détection d'un QTL dans un pedigree de type GDD, on commencera par poser le modèle suivant:

$$y = Xb + Z_u u + Z_h h + e$$

On suppose que l'on dispose des génotypes pour les pères des taureaux (s pères) et les taureaux (n taureaux). On connaît également la valeur d'élevage de chaque taureau, laquelle est reprise dans le vecteur $y(n \times 1)$. En outre, les données de génotype sont phasées chez les pères et leurs fils. Les données sont donc constituées de(2s+n) haplotypes fondateurs pour lesquelles il est possible de calculer pour toutes les paires possibles, les pIBD. On peut également calculer pour une position donnée pour les haplotypes non fondateurs reçu par les fils d'origine paternelle la probabilité (par liaison génétique) que soit l'haplotype grand-mère paternelle (λ_p) (ou grand père paternelle ρ_p) qui était reçu . h correspond au vecteur des effets (aléatoires) haplotypiques de ((2s+n)x1) (2 s haplotypes paternelles et n haplotypes maternelles). Zh est la matrice d'incidence qui comprends n ligne et 3 élements par ligne: 1 (coeffecient d'incidence relatif à l'effet maternelle), λ_p et ρ_p).

La variance des effets haplotypiques est égale $H \times \sigma_H^2$ à Hp, cette matrice fournit les termes de covariance entre haplotypes pour une position donnée et correspond le plus souvent à la matrice de probabilités IBD.

Toutefois pour résoudre ces équations des modèle mixtes, il est nécessaire d'inverser cette matrice Hp, Ce qui peut être parfois compliquer d'un point vu numérique (problème de singularité). Pour contourner ce problème, il a été proposé de convertir ces probabilités en distance entre halotypes pour retracer un dendogramme sensé reflété l'histoire des différents haplotypes fondateurs de l'échantillon. Deux haplotypes sont d'autant plus proche dans l'arbre qu'ils ont une probabilité d'IBD élevé et inversement. En coupant l'arbre à une position donnée, il est possible de déterminer des groupes haplotypes, en considérant que tout les haplotype appartiennent au même groupe haplotypique si ils sont sur une branche issu du noeud le plus proche avant (dans le temps) la coupure. Les effet associés à ces différents groupes haplotype sont ensuite estimés par une approche REML, en posant que la covariance entre groupe haplotypique est égale à zéro. Ce qui amène à remplacer la matrice Hp en une matrice I dont la taille dépendra du nombre de groupe haplotype et donc de la position de la coupure dans le dendogramme.

I.4.3 Seuil de signification et intervalle de confiance.

I.4.3.1 Le problème: test multiple.

Dans la plupart des études de cartographie, l'analyse ne se limite pas à un unique segment chromosomique ou à un marqueur, mais concerne généralement l'entièreté d'un chromosome voir du génome. En outre, il arrive très fréquemment d'utiliser le même matériel génétique pour détecter des QTL affectant des phénotypes différents.
Ce qui conduit donc à tester plusieurs hypothèses nulles. Le seuil de signification dans ce type d'analyse ne correspond pas à la valeur nominale du seuil de signification pour un test unique.
Pour comprendre cela, on peut prendre l'exemple d'un dé jeté plusieurs fois. Pour savoir si un dé est pipé et qu'il sort plus souvent la face 1, on lance 100 fois ce dé. On note le nombre de fois où la face 1 sort. On ne peut pas considérer que le dé est pipé si le 1 ressort plus fois alors qu'on s'attend avec un dé normal à tirer en moyenne 17 un.

I.4.3.2 Correction pour multiple marqueurs testés.

I.4.3.2.a Carte de faible densité.

Si une carte de faible densité (ex dans le cas des études de liaison) est utilisée et que les marqueurs sont suffisamment distants pour considérer qu'ils ségrègent indépendamment. On peut alors considérer que le nombre de tests indépendants réalisés est égal au nombre de marqueurs. La valeur nominale des seuils de signification peut alors être ajustée en utilisant une correction de Bonferonni.

I.4.3.2.b Carte de haute densité.

Quand on utilise une carte de haute densité en marqueur, l'information apportée par certains marqueurs est corrélée à celle de marqueurs voisins. Appliquer une correction de Bonferonni dans ce cas peut amener à adopter un seuil de signification beaucoup trop strict. La meilleure alternative est de passer par des simulations (ou des permutations) pour calculer la distribution attendue sous l'hypothèse nulle des tests statistiques que l'on utilise . On peut générer l'hérédité des marqueurs génétiques indépendamment du phénotype étudié dans des simulations en suivant les lois Mendel et en recalculant à chaque fois pour chaque position les valeurs des tests statistiques. Pour chaque simulation, il est possible de déterminer la position du génome donnant la valeur la plus élevée pour le test statistique réalisé. On garde cette valeur pour calculer la distribution sous l'hypothèse nulle et déterminer le seuil de signification à adopter dans l'étude de cartographie mise en place. Garder la valeur la plus élevée du test statistique utilisé permet de diminuer le nombre de simulations en inférant seulement sur la queue de la distribution statistique sous l'hypothèse H0.
Il est également possible de procéder par des permutations des phénotypes des individus pour calculer la distribution des tests statistiques attendus sous l'hypothèse nulle. Que ce soit par des permutations ou des simulations, l'objectif est de déconnecter les phénotypes des génotypes.

I.4.3.3 Correction pour les multiples caractères étudiés.

Il arrive très souvent notamment chez les espèces de productions que l'on recherche des QTL ou des loci à risque pour plusieurs caractères simultanément. On peut procéder à une correction de Bonferonni en considérant que les phénotypes étudiés sont indépendant des uns des autres cependant comme pour les marqueurs génétiques ajuster les seuils de cette manière peut conduire à des seuils de signification trop stricts.
Il est possible également de procéder à des permutations (on permute tout les caractères d'un individu avec ceux

d'un autre individu) pour terminer le seuil de signification empirique à adopter pour tenir compte des tests multiples réalisés. Cependant, il a été montré que le seuil de signification déterminé avec cette stratégie peut différer de manière très importante d'un caractère à l'autre et en adoptant un seuil unique déterminer de manière empirique, on s'expose à définir un seuil parfois trop strict pour certains caractères.

Une alternative est de déterminer par des permutations le nombre de phénotypes indépendants n et d'utiliser cette valeur pour ajuster les seuils de signification de chaque phénotype de manière individuelle par une correction de Bonferonni.

II Linkage disequilibrium on the bovine X chromosome: characterization and use in Quantitative Trait Locus mapping.

Abstract:

We herein demonstrate that in the Holstein-Friesian dairy cattle population, microsatellites are as polymorphic on the X chromosome as on the autosomes but that the level of linkage disequilibrium between these markers is higher on the X chromosome than on the autosomes. The latter observation is not compatible with the small male to female ratio that prevails in this population and results in a higher gonosomal than autosomal effective population size. It suggests that the X chromosome undergoes distinct selective or mutational forces. We describe and characterize a novel Markovian approach to exploit this linkage disequilibrium in order to compute the probability that two chromosomes are identical-by-descent conditional on flanking marker data. We use the ensuing probabilities in a restricted maximum likelihood approach to search for QTL affecting 48 traits of importance to the dairy industry and provide evidence for the presence of QTL affecting five of these traits on the bovine X chromosome.

SANDOR, C.; FARNIR, F.; HANSOUL, S.; MEUWISSEN, T.; COPPIETERS, W.; GEORGES, M.

Genetics 173: 1777-1786 (2006)

II.1 Introduction.

Extensive use of artificial insemination (A.I.) in mammalian livestock species, especially cattle and pig, leads to the frequent occurrence of large paternal half-sibships. This pedigree structure has been abundantly exploited to map quantitative trait loci (QTL) influencing a variety of agronomically important phenotypes. Detection of significant differences between the phenotypic means of half-sibs sorted according to the homologue inherited from their sire point towards linked QTL[63]. However, as sires are hemizygous for the X chromosome, this strategy has precluded exploration of the sex chromosomes except for the pseudoautosomal region[64,65].

It has recently been shown that linkage disequilibrium (DL) extends over unusually long distances in livestock species due to reduced effective population size (N_e)[66]. As a result, significant DL can be detected using the medium density marker maps available in these species. This can be exploited to increase the power and resolution of QTL mapping[61,67]. It also opens possibilities to explore correlation between phenotype and marker genotype on the X chromosome, i.e. map QTL on that chromosome.

In this paper, we have (i) quantified the level of microsatellite polymorphism and DL on the X chromosome in Holstein-Friesian dairy cattle, (ii) used an approach based on Markov chains to compute identity-by-descent (IBD) probabilities of a pair of X chromosomes conditional on flanking marker data, and (iii) used to corresponding IBD probabilities to map QTL on the X chromosome.

II.2 Materials & Methods.

II.2.1 Pedigree material and phenotypes.

We used a previously described Holstein-Friesian grand-daughter design (GDD)[68] comprising 22 paternal half-brother families for a total of 929 bulls. The genealogies of these animals were obtained from Holland Genetics (Arnhem, The Netherlands). The average number of recorded ancestors per bull-dam was 40.4, whereas up to 11 generations separated the bulls from their most distant ancestor. Based on the available pedigree data and assuming that the founders were unrelated, we estimated the average inbreeding coefficient, *F*, of the bull-dams at 1.3% (range: 0% - 14%), and their average kinship coefficient, *f*, at 4% (range: 0%-57%)[66].

For each of these bulls we obtained estimated breeding values (BV) from Holland Genetics (Arnhem, The

Netherlands) for 48 phenotypes of interest to the dairy industry. The analyzed phenotypes can be grouped in (i) five milk yield and milk composition traits, (ii) 13 conformation traits, (iii) eight birth and fertility traits, (iv) ten udder health traits, (v) two workability traits, and (vi) ten productivity traits (Table 1). A detailed description of each of these traits can be found at http://www.nrs.nl/index-eng.htm.

II.2.2 Marker genotyping.

Using standard procedures[68], the entire GDD was genotyped for 22 X-specific and three pseudoautosomal microsatellite markers, which have all been previously described[69]. Marker order and recombination rates between adjacent markers were taken from[69] and are shown in Figure II.1. We also used phased genotypes available on the same GDD for 202 autosomal microsatellites[66]. Sex-averaged recombination rates between autosomal markers were obtained from Ihara et al.[70].

II.2.3 Measuring linkage disequilibrium.

Pair-wise linkage disequilibrium (LD) was measured using r^2 as previously described[71]. Briefly, r^2 was computed as

$$\sum_{i=1}^{u}\sum_{j=1}^{v} x_{ij}\,\frac{(x_{ij}-p_i q_j)^2}{p_i(1-p_i)q_j(1-q_j)} \quad (1)$$

where u and v are the respective number of alleles at the two marker loci A and B, p_i and q_j are the population frequencies of alleles A_i and B_j respectively, x_{ij} is the observed frequency of haplotype A_iB_j.

II.2.4 Computing identity-by-descent (IBD) probabilities conditional on marker genotype for pairs of X chromosomes.

The probability of IBD of two X chromosomes at a given map position conditional on multiple linked marker genotypes, Φ_p, was computed according to[72]. This requires the computation of the probability of coalescence without recombination, p_C^T, of a chromosome segment of size θ in a population with constant male and female population sizes of respectively N_m and N_f, within T generations where T is the number of generations separating the present from the unrelated base population. This probability is computed using hidden Markov theory,

knowing that:

$$\mathbf{P^{t+1}} = \mathbf{P^t M} \quad (2)$$

where $\mathbf{P^t} = (p^t_{MM}, p^t_{MF}, p^t_{FF}, p^t_C)$, p^t_{MM} is the probability that the two X chromosome segments are two different male chromosomes *t* generations before present, p^t_{MF} the probability that the two X chromosomes are one a male the other a female chromosome *t* generations before present, p^t_{FF} the probability that the two X chromosomes are two different female chromosomes *t* generations before present, p^t_C the probability that the two X chromosomes have coalesced between present and *t* generations back. **M** is a matrix of transition probabilities:

$$\mathbf{M} = \begin{pmatrix} p_{MM \to MM} & p_{MM \to MF} & p_{MM \to FF} & p_{MM \to C} \\ p_{MF \to MM} & p_{MF \to MF} & p_{MF \to FF} & p_{MF \to C} \\ p_{FF \to MM} & p_{FF \to MF} & p_{FF \to FF} & p_{FF \to C} \\ p_{C \to MM} & p_{C \to MF} & p_{C \to FF} & p_{C \to C} \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 0 & (1-\frac{1}{2N_f})(1-\theta_f)^2 & \frac{1}{2N_f}(1-\theta_f)^2 \\ 0 & \frac{1}{2}(1-\theta_f)(1-\theta_m) & \frac{1}{2}(1-\frac{1}{2N_f})(1-\theta_f)^2 & \frac{1}{4N_f}(1-\theta_f)^2 \\ \frac{1}{4}(1-\frac{1}{N_m})(1-\theta_m)^2 & \frac{1}{2}(1-\theta_m)(1-\theta_f) & \frac{1}{4}(1-\frac{1}{2Nf})(1-\theta_f)^2 & \frac{1}{4}(\frac{1}{N_m}(1-\theta_m)^2 + \frac{1}{2N_f}(1-\theta_f)^2 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (3)$$

$p_{MM \to MM}$ is the probability for the two X chromosomes to be different male chromosomes *t* generations ago and different male chromosomes *t+1* generations ago; $p_{MM \to MF}$ is the probability for the two X chromosomes to be different male chromosomes *t* generations ago and one a male chromosome and the other a female chromosome *t+1* generations ago; etc … θ_m and θ_f are male and female recombination rates, respectively. For all but the pseudoautosomal region of the sex chromosomes $\theta_m = 0$. As we are using sires in a GDD**, P⁰** was set at (1,0,0,0). *T* was set at 100, N_m at 25 and N_f at 1,000,000 yielding an effective population size N_e of ≈ 100.

Note that the same approach can also be used to compute coalescence probabilities for autosomal chromosome segments. **M** then becomes:

$$\mathbf{M}=\begin{pmatrix} \frac{1}{4}(1-\frac{1}{2N_m})(1-\theta_m)^2 & \frac{1}{2}(1-\theta_m)(1-\theta_f) & \frac{1}{4}(1-\frac{1}{2N_f})(1-\theta_f)^2 & \frac{1}{4}(\frac{1}{2N_m}(1-\theta_m)^2+\frac{1}{2N_f}(1-\theta_f)^2) \\ \frac{1}{4}(1-\frac{1}{2N_m})(1-\theta_m)^2 & \frac{1}{2}(1-\theta_m)(1-\theta_f) & \frac{1}{4}(1-\frac{1}{2N_f})(1-\theta_f)^2 & \frac{1}{4}(\frac{1}{2N_m}(1-\theta_m)^2+\frac{1}{2N_f}(1-\theta_f)^2) \\ \frac{1}{4}(1-\frac{1}{2N_m})(1-\theta_m)^2 & \frac{1}{2}(1-\theta_m)(1-\theta_f) & \frac{1}{4}(1-\frac{1}{2N_f})(1-\theta_f)^2 & \frac{1}{4}(\frac{1}{2N_m}(1-\theta_m)^2+\frac{1}{2N_f}(1-\theta_f)^2) \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

(4)

The probabilities of coalescence without recombination, p_C^T, computed as described above were then utilized to compute IBD probabilities for pairs of X chromosomes conditional on flanking marker genotypes, Φ_p, using the rules defined by Meuwissen & Goddard[72].

II.2.5 Mapping QTL on the X chromosome using a Restricted Maximum Likelihood Approach (REML).

QTL were mapped on the X chromosome using an "interval mapping approach", i.e. the position of the hypothetical QTL was slid across the BTAX marker map and lod scores were computed at regular intervals. In this study, a lod score was generated in the middle of each marker interval.

Lod scores were computed essentially as previously described[73,74]. To test map position *p*, we first clustered the *n* BTAX haplotypes in the data set in a rooted dendrogram. The *n(n-1)/2* pairwise (1 - Φ_p) values (computed as described above) were used as distance measures and UPGMA as a hierarchical clustering algorithm[75]. The tree was then used as a logical framework to group the haplotypes in functionally distinct clusters. To achieve this, the tree was scanned downwards from the top and branches cut such that the distance between haplotypes within resulting sub-trees would not exceed a threshold *C*.

We then modelled the breeding values of the sons using the following linear model:

$$\mathbf{y}=\mathbf{Xb}+\mathbf{Z_h h}+\mathbf{Z_u u}+\mathbf{e} \quad (5)$$

$\mathbf{y}$ $(n\times 1)$ is the vector of breeding values for all sons. $\mathbf{b}$ $(p\times 1)$ is a vector of fixed effects, which reduces here to the overall mean. $\mathbf{X}$ $(n\times p)$ is the incidence matrix relating fixed effects to individual sons, which here reduces to a vector of ones. $\mathbf{h}$ $(q\times 1)$ is the vector of random QTL effects corresponding to the functionally

distinct haplotype clusters defined as above. $\mathbf{Z_h}$ $(n \times q)$ is an incidence matrix that relates each son to his corresponding haplotype cluster. **u** $(r \times 1)$ is the vector of random individual autosomal polygenic effects ("animal model")[76]. $\mathbf{Z_u}$ $(n \times r)$ is a diagnonal incidence matrix that relates each son to its individual polygenic effect. **e** $(n \times 1)$ is the vector of individual error terms.

Haplotype cluster effects with corresponding variance, σ_H^2, individual polygenic effects with corresponding variance, σ_A^2, and individual error terms with corresponding variance, σ_E^2, were estimated using AIREML[77], by maximizing the restricted log likelihood function *L*:

$$L = -.5 \ln|\mathbf{V}| - .5 \ln\left|\mathbf{X^T V^{-1} X}\right| - .5(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) \quad (6)$$

In this, **V** equals:

$$\mathbf{V} = \sigma_H^2 \mathbf{Z_h H Z_h^T} + \sigma_A^2 \mathbf{Z_u A Z_u^T} + \sigma_E^2 \mathbf{I} \quad (7)$$

Because we assumed that the covariance between the QTL effects of the different haplotype clusters was zero, **H** reduces to an identity matrix. **A** is the additive genetic relationship matrix[76].

L was computed for all possible values of *C* (from 1 to 0), in order to identify a restricted maximum likelihood (REML) solution for map position *p* under the H_1 hypothesis.

The log likelihood of the data under H_1 was then compared with that under the null hypothesis, H_0, of no QTL at map position *p*. The latter was computed as described above but using the reduced model:

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{Z_u u} + \mathbf{e} \quad (8)$$

Evidence in favor of a QTL at map position, *p*, was expressed as a lod score:

$$z_p = \frac{(L_{H_1} - L_{H_0})}{\ln(10)} \quad (9)$$

The statistical significance of observed z_p values, accounting for the fact that we tested 24 marker intervals and multiple values of *C* in each interval, was estimated from the distribution of largest z_p values obtained along the X chromosome when analyzing 1,000 sets of simulated phenotypes. The latter were generated by "dropping" 30 unlinked, autosomal QTL through the known genealogy of the 929 bulls. All QTL were assumed to have four

alleles with random allelic frequency, and additive QTL effects sampled from $\frac{e^{-x}}{2}$ ($x \in [0;+\infty[$) with randomly assigned positive or negative sign. The polygenic effect of each bull was computed as the sum of its 60 QTL effects. In addition, each bull was assigned a non-genetic effect sampled from a normal distribution $N(0,\sigma_E^2)$ such that:

$$\sigma_E^2 = \frac{\sigma_P^2(1-h^2)}{h^2} \quad (10)$$

where σ_P^2 is the variance due to the polygenic effects and h^2 is the heritability which was set at 0.75.

II.3 Results.

276,048 phased autosomal and 22,181 X-linked microsatellite genotypes, produced on 929 bulls as described in Materials & Methods were available for analysis. Figure II.2 compares the number of alleles as well as the "effective" heterozygosity (computed as $1-\sum_{i=1}^{n} p_i$, where p_i is the frequency of allele i out of n) for the 202 autosomal microsatellites and the 25 gonosomal microsatellites. Only the alleles inherited from the dam were considered in this analysis. Average number of alleles and effective heterozygosity (± SE) were respectively 6.6 ± 0.4 and 0.59 ± 0.02 for the autosomal markers versus respectively 8.0 ± 1.1 and 0.61 ± 0.05 for the gonosomal markers. Microsatellite markers appeared thus slightly but not significantly more polymorphic on the X chromosome than on the autosomes.

Figure II.3A compares the level of pair-wise LD measured in the same sample for syntenic autosomal and X-specific markers as a function of genetic distance. To allow for proper comparison, the distance between markers was set at $\theta_A = \frac{1}{2}(\theta_f + \theta_m)$ for the autosomal markers, and at $\theta_X = [\theta_f(N_m + N_f)/(N_m + 2N_f)]$ for the gonosomal markers. In these, θ_m and θ_f correspond to the male and female recombination rates, while N_m and N_f correspond to the number of males and females respectively. For the autosomes, θ_A corresponded to the sex averaged θ values obtained from Ihara et al.[70]. For the X chromosome, θ_f values were obtained from Sonstegard et al.[69]. N_m and N_f were set at 25 and 1 million respectively, yielding an effective (autosomal) population size (N_{eA}) of ≈ 100. The latter value is close to recent, pedigree-based estimates of N_{eA} in this population[78]. It can be seen that r^2 values are systematically higher for the X chromosome than for the autosomes, being approximately

1.75 times as large and ranging from values of the order of 0.10 for recombination rates inferior to 5 % to values of the order of 0.01 for recombination rates between 20 and 25%. Note that these r^2 values, including between the more distant markers are highly significant[66]. Note also that choosing less extreme sex ratios accentuates the difference between r_X^2 and r_A^2.

Using the Markov model described in Materials and Methods, we computed the (929^2-929)/2 pair-wise IBD probabilities conditional on marker data, for the 24 BTAX marker intervals. In these calculations, we assumed that *T*, the number of generations to the base population, was 100, that N_m was 25 and N_f one million - as before. The overall IBD probability (averaged across all chromosome pairs and marker intervals) was 0.26. This is lower than the theoretical expectation of 0.36, corresponding to the probability of coalescence of a chromosome region of size 0 computed using the gonosomal-specific Markov model (see Materials & Methods), and hence to the coefficient of kinship of a chromosome region of size 0 (Meuwissen & Goddard)[72]. Adjusting these parameter values in order for the theoretical and observed values to match might be a way to fine-tune the model.

For a randomly selected interval (interval 12), we compared the IBD probabilities obtained with the X-specific transition probabilities, with those that would be obtained using autosomal transition probabilities, assuming (i) 25 males and 1 million females, but sex-specific recombination rates (θ_f and $\theta_m = 0$), (ii) a population size of 100 with equal numbers of males and females, as well as equal recombination rates in males and females ($= ½\ \theta_f$). The latter, yields in essence what would be obtained when directly applying the autosomal method proposed by Meuwissen and Goddard[72] to the X chromosome. The results obtained under scenario (i) are shown in Figure. II.4. It can be seen that - although the differences are quite modest - the correct gonosomal model yields more "conservative" IBD probabilities, in the sense that if marker information suggests that the considered pair of chromosomes have a higher probability of IBD than two randomly selected chromosomes this probability is inflated when using the erroneous autosomal model, while if the marker information suggests the opposite, the autosomal model underscores the IBD probabilities when compared to the correct gonosomal model. This tendency was even more pronounced when using an autosomal model assuming equal numbers of males and females and equal recombination rate in both sexes corresponding to $½\ \theta_f$ (scenario (ii); data not shown).

The pair-wise IBD probabilities described in the previous section were utilized to search for QTL influencing 48 dairying traits on the X chromosome using the variance component approach described in Materials and Methods. The proposed model includes a random individual animal effect which should properly correct for differences in autosomal co-ancestry that may be correlated with BTAX IBD probabilities, and thus protect against false-positive QTL effects as a result of stratification. The lod score threshold of 2.5 corresponding to a chromosome-wide type I error of ≈ 5 % was exceeded for five traits: fat yield, direct durability, durable

prestation, milking speed and rear leg set (Table II.1). Assuming that these are independent QTL (as suggested by the type of trait and the respective most likely QTL positions), and knowing that - because of their correlations - the 48 analyzed traits behave in essence as ≈ 20 independent traits (W Coppieters unpublished data), this is thus approximately five times more than expected by chance alone. The same five QTL emerged out of an analysis performed using less extreme parameter values N_m = 15,000, N_f = 1,000,000, and T = 100 (Table II.1). Altogether, this suggests that at least some of these QTL are likely to be genuine.

The proportion of the trait variance explained by the QTL ranged from 2 % to 8 %, and was - as expected - well correlated with the lod score value. The number of haplotype clusters associated with the REML solution ranged very widely from 2 to 745. Figure. II.5 shows representative examples of the distribution of haplotype-cluster effects with standard error of the estimates and corresponding cluster frequency in the analyzed sample. In all cases, we were able to identify common haplotype clusters with significantly different effects. These most likely explain the high lod scores that were obtained.

II.4 Discussion.

In this paper, we have compared the level of polymorphism and LD between X-linked and autosomal microsatellite markers in the Holstein-Friesian dairy cattle population. We find X-linked microsatellite markers to be at least as variable as autosomal ones, and - at comparable distance - LD to be considerable higher on the X chromosome than on the autosomes.

Typically, genetic polymorphism is expected to be lower and LD to be higher for markers on the X chromosome, as observed in the human[79,15,80,37]. The lower level of polymorphism on the X is thought to result from (i) higher genetic drift as $N_{eX} = \frac{3}{4} N_{eA}$, (ii) a reduced mutation rate as the male mutation rate (μ_M) is expected to be 1.7 to 4 times higher than the female mutation rate (μ_F) and $\mu_X = 2/3\,\mu_F + 1/3\,\mu_M$ while $\mu_A = 1/2\,\mu_F + 1/2\,\mu_M$ [81,15], and (iii) enhanced purifying selection due to male hemizygosity. Note that ascertainment bias (i.e. the fact that one selects on polymorphism when developing markers) is expected to considerably blur this picture. The higher level of LD on the X chromosome is primarily thought to result from higher genetic drift. Our results in the bovine may thus – at first glance - seem in reasonable agreement with theoretical predictions: the level of polymorphism on the X is comparable to that on autosomes as a result of ascertainment bias when developing microsatellite markers, but the higher level of LD on the X – which is not subject to the same ascertainment bias - is as predicted by basic population genetics.

As a matter of fact, $N_{eX} = \frac{3}{4} N_{eA}$ is only valid when the sex ratio ($SR = N_m / N_f$) is one. This is clearly not the case for domestic cattle and most probably wasn't for their wild-type ancestors. Assuming (Wright[82]) that

$$N_{eA} = \frac{4N_m N_f}{N_m + N_f} \quad (11)$$

and

$$N_{eX} = \frac{9N_m N_f}{4N_m + 2N_f} \quad (12)$$

one can easily show that N_{eX} becomes larger than N_{eA} when *SR* is smaller than 1/7 (Figure II.6). Equations 11 and 12 assume that all parents of a given sex have an equal chance of contributing offspring to the next generation, which is obviously not valid in livestock. Accounting for non-random parental contribution, however, does not alter these conclusions (Figure. II.6 and Appendix).

Thus in cattle N_{eX} is likely to have been higher than N_{eA} before and certainly after domestication. This may in part explain why - in the bovine - X-linked markers are as polymorphic as autosomal markers, contrary to what is observed in the human. However, it precludes stronger drift from underlying the higher level of LD between X-linked markers. This is at least in part corroborated by comparing the level of LD between autosomal and X-linked "in silico" markers whose segregation within the known genealogy of the 929 analyzed bulls is simulated by gene dropping as previously described by Farnir et al.[66]. While for autosomal markers in silico LD levels are of the same magnitude as actual LD levels, for X- linked markers the levels of in silico LD are of the same magnitude as for the autosomal markers and thus considerably lower than the actual LD levels (Figure II.3.B).

This strongly suggest that the higher level of LD observed on the X chromosome of cattle, and possibly of other species as well, is due to as of yet undetermined factors other than drift. The fact that the increased levels of LD are also observed for markers that are more than 10 cM apart suggests that these factors are still operating or have been operating until recently, certainly after domestication[83]

We have developed a novel approach based on a Markovian model, that allows computation of IBD probabilities conditional on flanking marker data accommodating both autosomes and gonosomes, varying *SR* and sex-specific recombination rates. At present, the model only accounts for recombination but it should be possible to extend it such that it includes mutation as well. This may be important when dealing with short chromosomal segments for which mutation and recombination rate are of the same order of magnitude.

It is worthwhile noting that it should be possible to use the variance of IBD probabilities (σ^2_{OBS}) as the basis for a measure of LD information content (IC). For a given marker interval, σ^2_{OBS} could be compared with the variance expected if the marker information allowed one to unambiguously distinguish IBD status from non-IBD status. The variance corresponding to such ideal map is:

$$\sigma^2_{MAX} = \bar{x}(1-\bar{x})^2 + (1-\bar{x})\bar{x}^2 = \bar{x}(1-\bar{x}) \quad (13)$$

where $\bar{x}$ corresponds to the average IBD probability in the considered interval. Indeed, a proportion $\bar{x}$ of chromosome pairs are expected to be IBD, thus have an IBD probability of 1 if the markers are fully informative, and hence contribute $(1-\bar{x})^2$ to the variance of IBD probabilities; a proportion $(1-\bar{x})$ of chromosome pairs are expected to be non-IBD, thus have an IBD probability of 0 if the markers are fully informative, and hence contribute $\bar{x}^2$ to the variance of IBD probabilities. The ratio $\sigma^2_{OBS} / \sigma^2_{MAX}$ could thus be viewed as a measure of the LD information content (IC) of the utilized map. Although, the detailed behaviour and suitability of the proposed IC measure still needs to be examined in more detail, the IC profile obtained with this method along the X chromosome map (data not shown) suggests - as expected - that one could gain considerable power by increasing the marker density. The recent availability of tens of thousands of bovine Single Nucleotide Polymorphisms (SNP) as well of array-based methods allowing for the effective genotyping now makes this possible for both autosomes and the X chromosome.

We present evidence in this work that the X chromosome harbours several QTL affecting traits of importance to the dairy breeding industry. This contradicts previous studies in which effects of the X chromosome on some of these traits were searched for by fitting models in which the expected phenotypic covariance between relatives included terms accounting for sex linkage (e.g. Lynch & Walsh[76]; Goddard personal communication). The discrepancy between these results might be due to substantial gains of detection power from considering marker information.

Contrary to Meuwissen & Goddard[72], but following Kim & Georges[74]and Blott and al.[73], we are not directly converting between haplotype IBD probabilities in covariances between haplotype effects, but are clustering haplotypes on the basis of the IBD probabilities and then testing the effect of haplotype clusters on phenotype. One of the advantages of this approach is that it allows one to identify the haplotypes that are the most likely to be functionally distinct (as illustrated in Figure II.5) and then to focus the molecular work on these haplotypes when searching for causal variants. In addition, the haplotype clustering method in essence models a discrete number of QTL alleles which may be closer to the actual biology than the infinite number of alleles that is modeled in the alternative approach.

Our results suggest that it might be advantageous to include X-linked markers when performing marker assisted selection or genomic selection[62] in dairy cattle. One could for instance select amongst full-brothers according to the X chromosome inherited from the dam.

II.5 Acknowledgments.

This work has been funded by grants from Holland Genetics (Arnhem, the Netherlands), Livestock Improvement Corporation (Hamilton, New Zealand), the Walloon Ministry of Agriculture, and the GAME ULg/ARC (Action de Recherche Concertée). Cynthia Sandor is a fellow of the F.R.I.A. (Fonds pour la formation à la Recherche dans l'Industrie et dans l'Agriculture). We are grateful to expert technical assistance of Paulette Berzi, Nadine Cambisano, Latifa Karim, Myriam Mni, Patricia Simon and Erica Davis in producing the microsatellite genotypes and to Alain Empain for expert administration of the computer grid.

II.6 Appendix.

In the case of non-random parental contribution and unequal numbers of males and females, effective population size for autosomes (N_{eA}) and the X chromosome (N_{eX}) can be estimated using respectively:

$$\frac{1}{N_{eA}} = \frac{1}{16N_m}\left[2 + \sigma^2_{mm} + 2SR\rho_{mm,mf} + SR^2\sigma^2_{mf}\right] + \frac{1}{16N_f}\left[2 + \sigma^2_{ff} + \frac{2}{SR}\rho_{fm,ff} + \frac{1}{SR^2}\sigma^2_{fm}\right] \quad (14)$$

and

$$\frac{1}{N_{eX}} = \frac{1}{9N_m}\left[1 + 2SR^2\sigma^2_{mf}\right] + \frac{1}{9N_f}\left[1 + \sigma^2_{ff} + \frac{2}{SR}\rho_{fm,ff} + \frac{1}{SR^2}\sigma^2_{fm}\right] \quad (15)$$

In these equations N_m, N_f and SR are as defined before, σ^2_{xy} is the variance of the number of offspring of sex "y" per parent of sex "x", and $\rho_{xy,xz}$ is the covariance between the number of offspring of sex "y" and the number of offspring of sex "z" per parent of sex "x"[84].

Unequal parental contribution can be modelled by assuming that the N_m males belong to K classes of size N_{mi} with probability of fatherhood p_{mi}, such that $\sum_{i=1}^{K} N_{mi}\,p_{mi}$, and likewise that the N_f females belong to L

classes of size N_{fi} with probability of motherhood p_{fi}, such that $\sum_{i=1}^{L} N_{fi} p_{fi}$. Using this model, one can show that:

$$\sigma^2_{mm} \approx N_m \sum_{i=1}^{K} N_{mi} p^2_{mi}$$

$$\sigma^2_{mf} \approx \frac{1}{SR^2}(\sigma^2_{mm} - 1) + \frac{1}{SR}$$

$$\sigma^2_{ff} \approx N_f \sum_{i=1}^{L} N_{fi} p^2_{fi}$$

$$\sigma^2_{fm} \approx SR^2(\sigma^2_{ff} - 1) + SR$$

$$\rho_{mm,mf} \approx \frac{1}{SR}(\sigma^2_{mm} - 1)$$

$$\rho_{fm,ff} \approx SR(\sigma^2_{ff} - 1) \quad \rho_{fm,ff} \approx SR(\sigma^2_{ff} - 1)$$

Figure II.6 compares estimates of N_{eA} and N_{eX} obtained with this model in a population of one million individuals and varying *SR* in the case of (i) equal probability of parenthood for all parents and (ii) unequal probability of parenthood between parents. In the latter case, it was assumed that (i) the males were subdivided in three categories representing respectively 0.001, 0.1 and 0.89 of the male population with relative probabilities of parenthood of 100, 10 and 1, and (ii) the females were subdivided in three categories representing respectively 0.001, 0.1 and 0.89 of the female population with relative probabilities of parenthood of 10, 2 and 1.

We used this model to estimate N_{eA} and N_{eX} in the Dutch Holstein-Friesian population. Using statistics for 2002, 2003 and 2004 obtained from the NRS (Arnhem, The Netherlands), we set the number of reproducing males at 15,650 and the number of reproducing females at 1,076,100. To match the known distributions of sons per bull-sire and bull-dam as well as daughters per son as closely as possible (data not shown), we subdivided (i) the reproducing males in five classes representing respectively 0.0005, 0.005, 0.2, 0.3 and 0.4945 of the male population with relative probabilities of parenthood of 15,000, 400, 10, 2 and 1 and (ii) the reproducing females in three classes representing respectively 0.05, 0.1 and 0.85 of the female population with relative probabilities of parenthood of 1,000, 100 and 1. This yielded estimates of N_{eA} and N_{eX} of 88 and 99 respectively, thus very similar to the corresponding values of 100 and 112 obtained when considering 25 reproducing males and

1,000,000 reproducing females with equal probability of parenthood (as utilized throughout the manuscript), as well as to pedigree-based estimates of N_{eA}[85].

Table II.1: *Analyzed phenotypes.*

Milk (5)	Conformation (13)	Birth / Fertility (8)	Udder health (10)	Workability (2)	Productivity (10)
Milk yield (Kg)	Angularity	Birth weight	Cell counts	Milking speed	Direct durability
Fat yield (Kg)	Body capacity	Calving ease	Fore udder attachment	Temperament	Direct viability cows
Protein yield (Kg)	Chest width	Calving ease of the daughters	Front teat placement		Durability
Fat %	Condition score	Fertility index	Rear teat placement		Durable prestation sum
Protein %	Feet and legs	Gestation length	Rear udder height		Index indirect viability
	Foot diagonal	Interval calving-1st insemination	Suspensory ligament		Index direct viability
	Frame	Non return rate at 56 days in the daughters	Teat length		Indirect viability cows
	Rear leg set	Stillbirth	Udder		Indirect viability heifers
	Rear leg rear view		Udder depth		Inet
	Rump angle		Udder health index		“Total”
	Rump width				
	Stature				
	Weight				

The phenotypes other than milk yield and composition are expressed using normalized scores with mean of 100 and variance of 16 (see http://www.nrs.nl/index-eng.htm)

Table II.2: *QTL mapped on BTAX.*

Trait	Marker bracket	Chromosome – wide P_value	N° clusters	r_H^2	r_A^2	r_E^2
Fat yield	[HAUT37-XBM16]	0.038	487	0.08	0.80	0.12
		0.042	499	0.08	0.80	0.12
Direct durability	[ILSTS017-BM4604]	0.050	745	0.08	0.90	0.02
		0.051	741	0.08	0.90	0.02
Durable	[TGLA325-MAF45]	0.051	2	0.07	0.86	0.07
prestation		0.053	2	0.06	0.85	0.08
Milking speed	[XBM7-BMS417]	0.051	64	0.03	0.95	0.02
		0.053	76	0.02	0.96	0.03
Rear leg set	[BMS1616-HUMM2.21]	0.050	23	0.03	0.96	0.01
		0.052	14	0.04	0.96	0

The first line in each cell correspond to values obtained when computing IBD probabilities using Nm = 25, Nf = 1,000,000 and T = 100, the second line to the corresponding values obtained using Nm = 25, Nf = 1,000,000 and T = 100., σ_A^2 and σ_E^2 correspond respectively to the estimated proportion of the trait variance explained by the BTAX QTL, the autosomal polygenes and the error term. “N° of clusters” correspond to the number of haplotype clusters yielding the model that maximizes the likelihood (REML) of the data. The high values of r_A^2 are explained by the fact that the analyzed traits are estimated breeding values (EBVs).

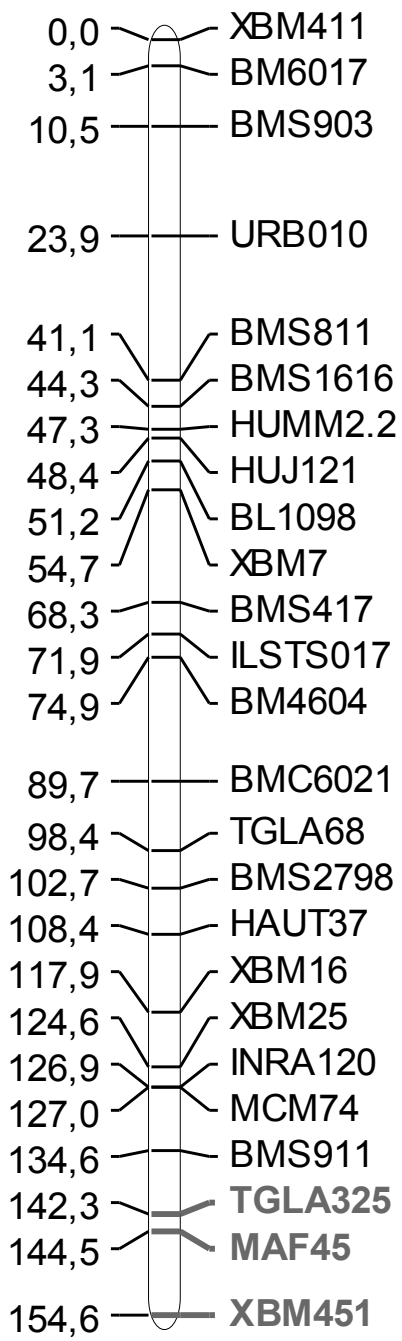


Figure II.1: *Order and distance in centimorgan (Kosambi) between the BTAX microsatellite markers used in the present study. Pseudo-autosomal markers are in bold.*

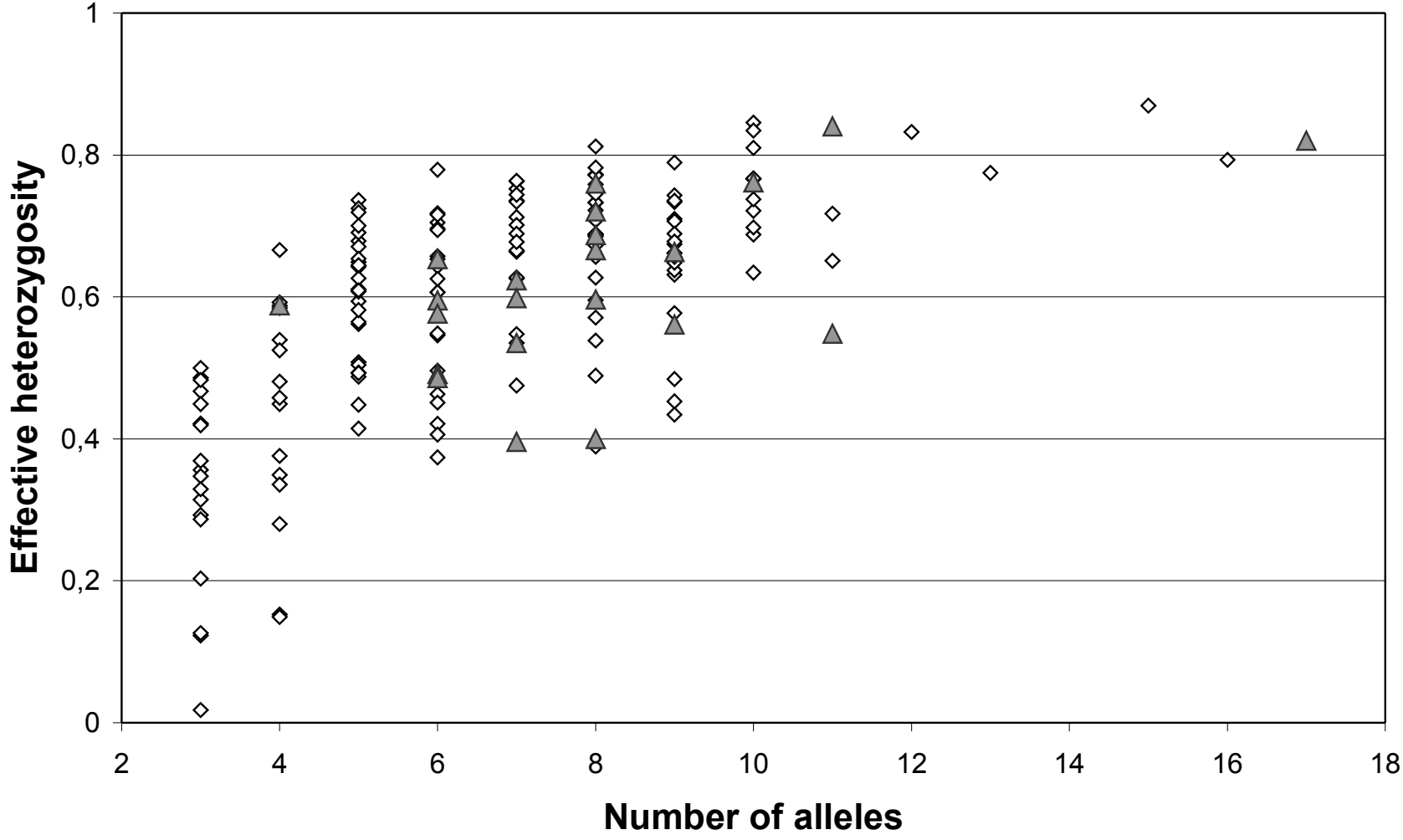


Figure II.2: *Number of alleles and effective heterozygosity* $1-\sum_{i=1}^{n} p_i$ *measured for 202 autosomal (white diamonds) and 25 X-specific (gray triangles) microsatellite markers in a Holstein-Friesian grand-daughter design.*

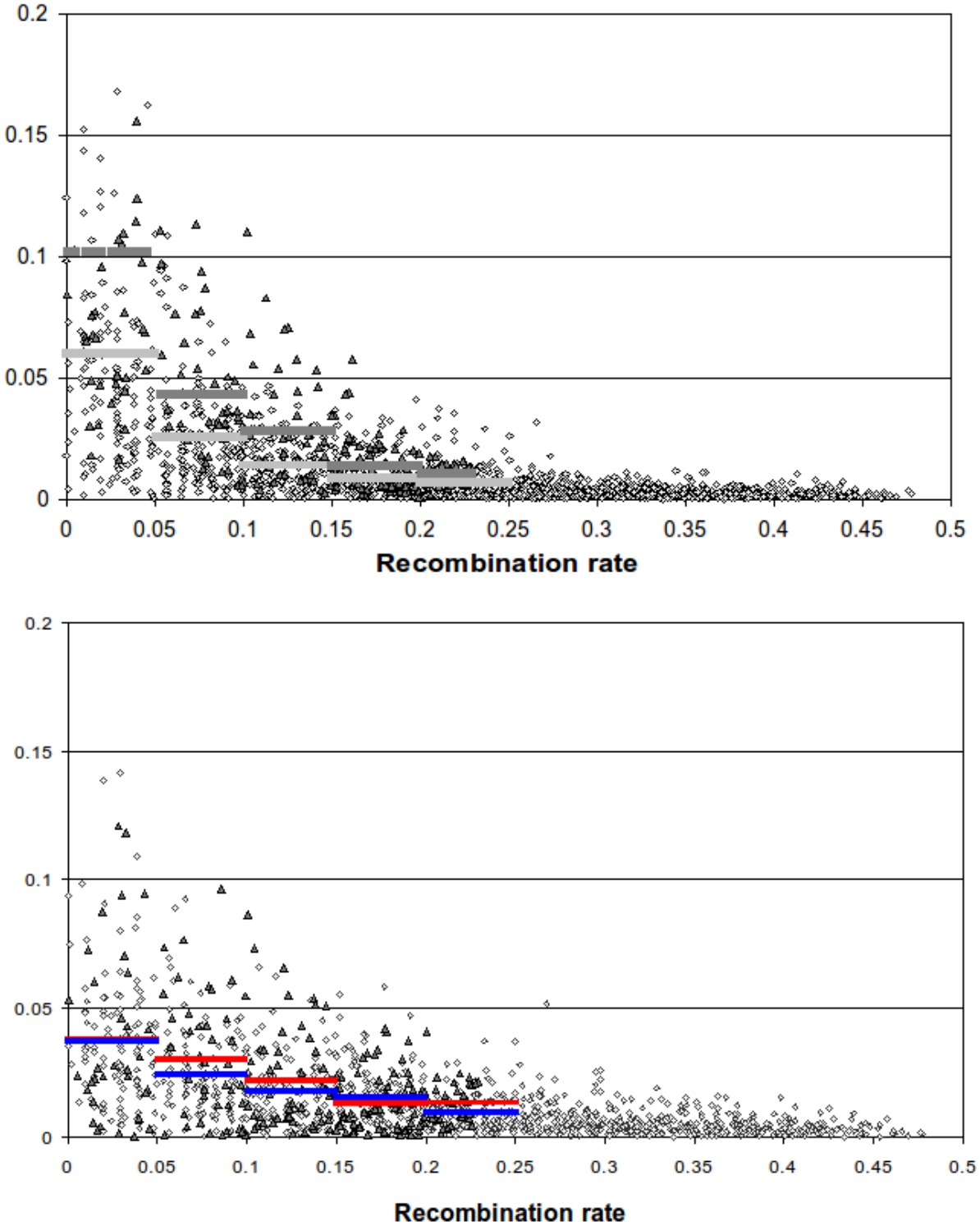


Figure II.3: *Linkage disequilibrium ((A) real data, (B) simulated data) measured using r² for syntenic autosomal (white diamonds) and X-specific (gray triangles) microsatellite markers. White and gray horizontal bars correspond to average autosomal and X-specific r2 values for windows of 5 recombination units (0 - 5 %, 5 - 10 %, 10 - 15 %, …).*

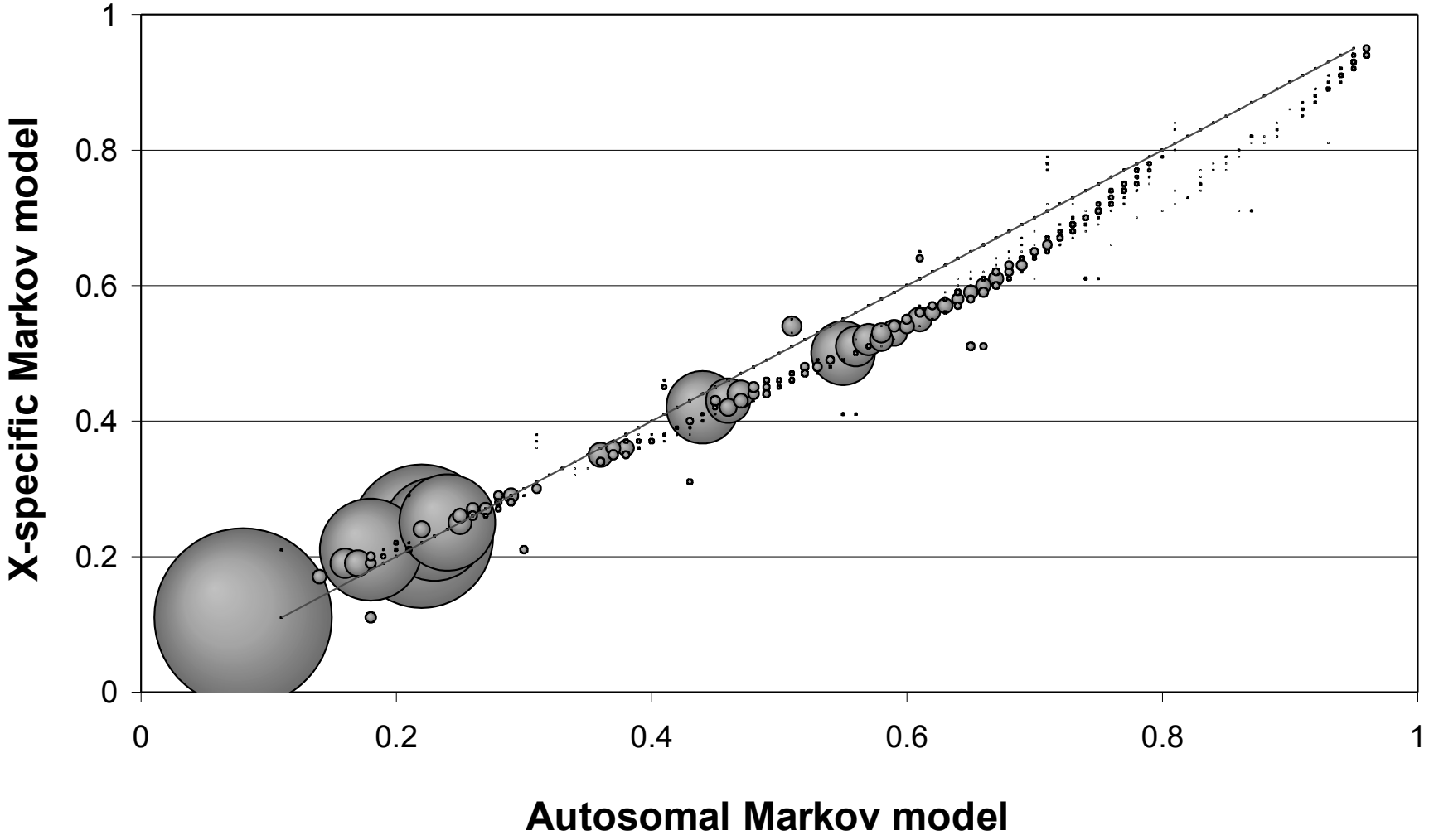


Figure II.4: *Linkage disequilibrium ((A) real data, (B) simulated data) measured using r2 for syntenic autosomal (white diamonds) and X-specific (gray triangles) microsatellite markers. White and gray horizontal bars correspond to average autosomal and X-specific r2 values for windows of 5 recombination units (0 - 5 %, 5 - 10 %, 10 - 15 %, …).*

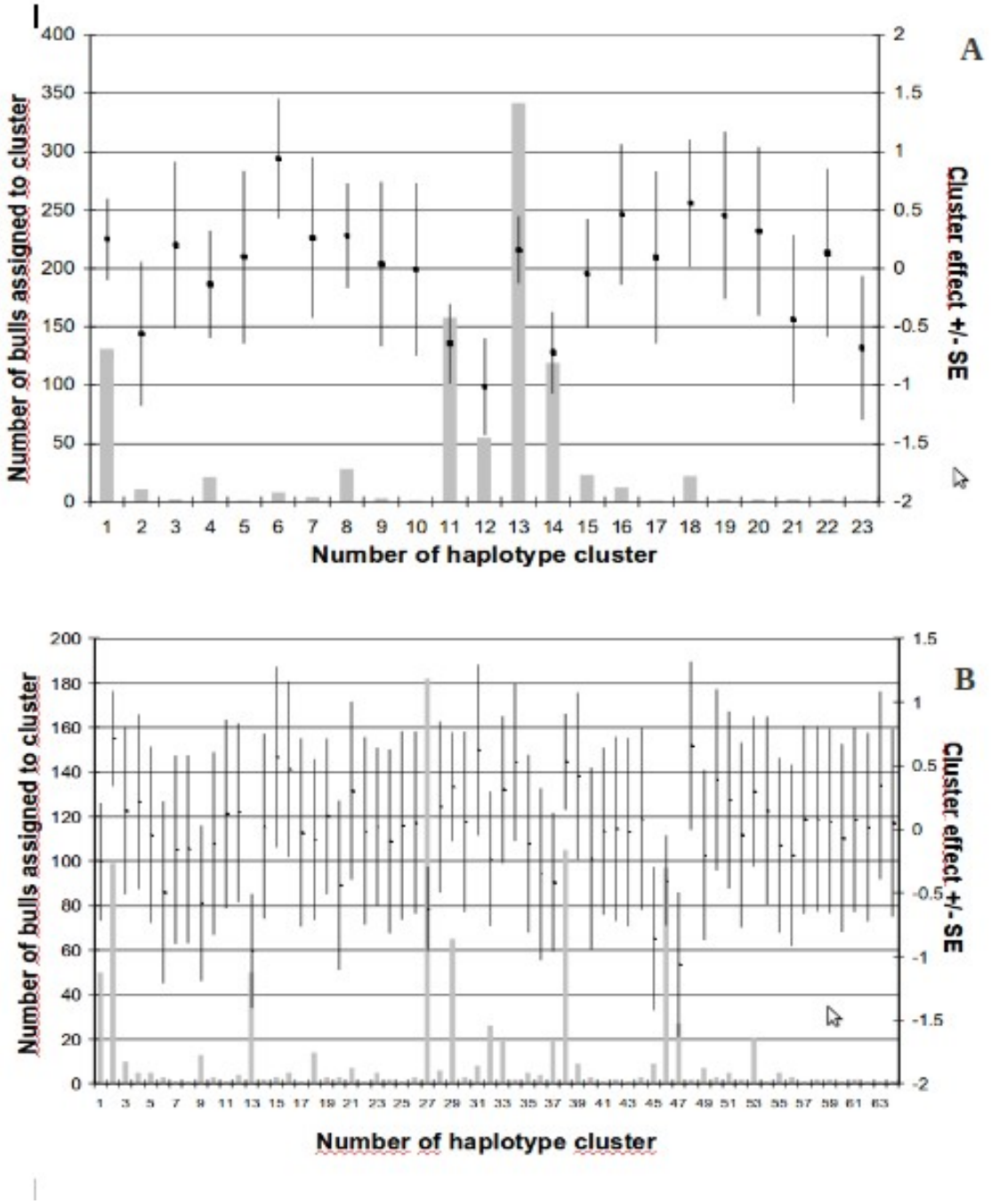


Figure II.5: *Representative examples of population frequency (gray bars) and effect ± standard error for the haplotype clusters yielding the highest lod scores for rear leg set (A) and milking speed (B).*

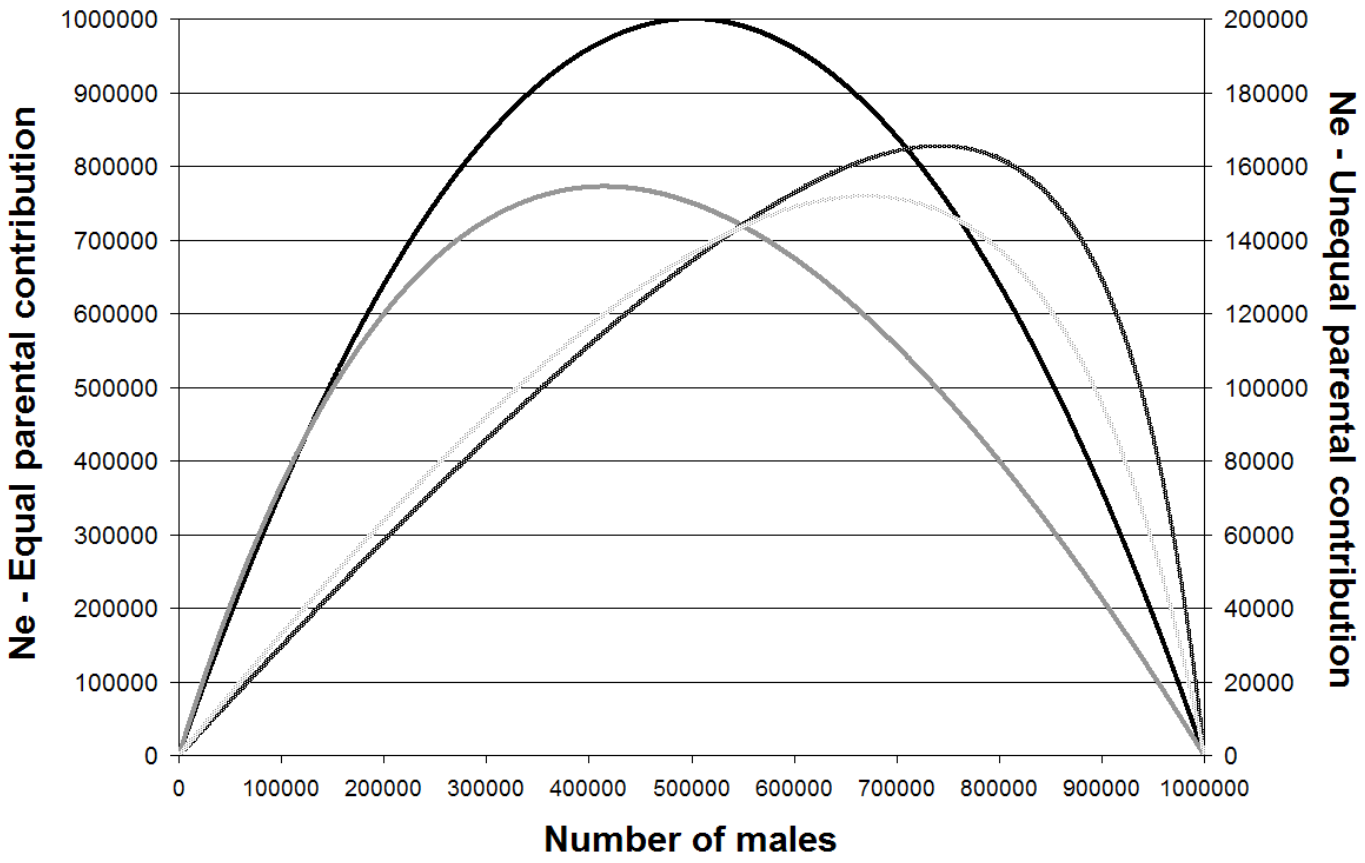


Figure II.6: *Estimates of autosomal (black) and X-specific (gray) effective population size for a population of constant size (1 million individuals) but with varying sex ratio assuming (i) equal probability of parenthood for all parents (continuous lines, left Y-axis), or (ii) unequal probability of parenthood amongst parents (dashed lines, right Y-axis). For the latter scenario we assumed that the males were subdivided in three categories representing respectively 0.001, 0.1 and 0.89 of the male population with relative probabilities of parenthood of 100, 10 and 1, while the females were subdivided in three categories representing respectively 0.001, 0.1 and 0.89 of the female population with relative probabilities of parenthood of 10, 2 and 1.*

III On the detection of imprinted QTL in line crosses: effect of linkage disequilibrium.

Abstract:

Imprinted QTL are commonly reported in studies using line-cross designs, especially in livestock species. It was previously shown that such parent-of-origin effects might result from the non-fixation of QTL alleles in one or both parental lines, rather than from genuine molecular parental imprinting. We herein demonstrate that if linkage disequilibrium exists between marker loci and non-fixed QTL, spurious detection of pseudo-imprinting is increased by an additional 40 to 80% in scenarios mimicking typical livestock situations. This is due to the fact that imprinting can only be tested in F2 offspring whose sire and dam have distinct marker genotypes. In the case of linkage disequilibrium between markers and QTL, such parents have a higher chance to have distinct QTL genotype as well thus resulting in distinct padumnal and madumnal allele substitution effect, i.e. QTL pseudo-imprinting.

SANDOR, C.; GEORGES, M.

Genetics 180: 1167-1175 (2008)

III.1 Introduction.

A minority of mammalian genes (~100) are subject to parental imprinting (http://www.mgu.har.mrc.ac.uk/research/imprinting/). Imprinted genes carry parent-of-origin specific epigenetic marks that cause tissue and developmental-stage specific silencing of either the padumnal (i.e. transmitted by the father) (~ 50% of imprinted genes) or madumnal (i.e. transmitted by the mother) (~ 50%) allele. Most imprinted genes cluster in imprinted domains, encompassing mono-allelically (both padumnal and madumnal) as well as bi-allelically expressed genes. Mono-allelically expressed genes at a given imprinted domain are co-ordinately regulated by imprinting control regions (ICR). In mammals, all known ICRs are characterized by allele-specific DNA methylation marks which in ~ 80% of the cases are imposed in the female germ-line[86]. Note that recent bioinformatic analyses suggest that the number of imprinted genes may be higher but this remains to be demonstrated experimentally[87,88].

In animals, parental imprinting has only been reported in placental mammals (marsupials and eutherians), but not in monotremes, birds or fish[89]. Parental imprinting is thought to have evolved as a result of unequal parental provision of resources to offspring in polygamous species (the “conflict hypothesis”)[90]. Padumnal alleles benefit from levying maternal resources even if at the expense of future offspring of the mother, while madumnal alleles benefit from sparing the reproductive potential of the mother. Concomitantly, imprinted genes that are preferentially expressed from the padumnal allele are enriched in genes that promote foetal and placental growth (e.g. *Igf2* and *Peg11*), while imprinted genes that are preferentially expressed from the madumnal allele are enriched in genes with opposite effects (e.g. *Igf2r*, *Phlda2, Cdkn1c*). There is evidence that some imprinted genes affect neonatal adaptation to suckling and metabolism (*Gnasxl, Peg3*)[91] suggesting that they may be imprinted in monotremes as well, an hypothesis that remains to be tested.

It is noteworthy that parental imprinting has evolved independently in flowering plants, affecting genes expressed in the endosperm, an organ which provides nutrients to the embryo developing inside the seed[92].

Growth and body composition are amongst the economically most important traits in livestock production, and identifying quantitative trait loci (QTL) and genes affecting these traits is one of the priorities of modern animal genetics[93]. Given the importance of imprinted genes in controlling prenatal growth it is not surprising that animal geneticists have searched for parent-of-origin effects on QTL and genes influencing these traits. As a matter of fact, parent-of-origin effects were observed early on for QTL increasing muscle mass in Callipyge sheep[94] and Piétrain pigs[95,96]. The corresponding QTL were subsequently shown to be caused by post-natal over-expression of the imprinted *DLK1*[97] and *IGF2* genes[98] in skeletal muscle.

These initial studies were followed by a flurry of papers reporting imprinted QTL in domestic animals. As an example, De Koning[99]detected four imprinted QTL (two paternally and two maternally expressed) out of five affecting body composition in a Meishan x European F2, and concluded that imprinting indeed plays an important role in the determinism of these traits. Reporting imprinted QTL has become standard in animal genetics, and the number of allegedly imprinted QTL abound across the genome. Imprinted QTL have been reported even in poultry despite the fact that there is no experimental evidence for parental imprinting in birds[100].

How can the present molecular knowledge of a handful of imprinted domains in mammals be reconciled with the abundance of imprinted QTL in domestic animals including birds ?

Studies reporting imprinted QTL share a "line-cross design", i.e. an F2 intercross between lines or breeds diverging for the traits of interest. In these studies, parental lines/breeds are assumed to be fixed for alternate QTL alleles (*Q* and *q*), all F1 individuals to be of "*Qq*" QTL genotype, and the three possible genotypes (*QQ, Qq, qq*) assumed to segregate in the F2 generation in approximate 1:2:1 Mendelian proportions. Testing imprinting consists in contrasting the madumnal versus padumnal allele substitution effect at the QTL map position. Assuming that the marker alleles originating from the parental lines are labelled "1" and "2" respectively, the madumnal allele substitution effect is estimated as

$$\alpha_M = \tfrac{1}{2}\left[(\overline{11} - \overline{21}) + (\overline{12} - \overline{22})\right] \quad (1),$$

and the padumnal allele substitution effect as:

$$\alpha_P = \tfrac{1}{2}\left[(\overline{11} - \overline{12}) + (\overline{21} - \overline{22})\right]. \quad (2).$$

In these $\overline{XX}$ is the phenotypic average of offspring with corresponding marker genotype; the first allele is madumnal, the second padumnal. The contrast that tests imprinting is

$$\Delta_{IMP} = \alpha_M - \alpha_P = (\overline{12} - \overline{21}) \quad (3).$$

As one cannot distinguish 12 from 21 offspring, imprinting cannot straightforwardly be tested in a typical F2 setting – at least not using genotypic means.

Note that Cui et al.[101] recently proposed to exploit known differences between male and female recombination rates to infer the most likely parental chromosomal origin in an F2 generation. This information can in theory be used to test for imprinted QTL in an F2 design. However, the effectiveness of this approach remains to be demonstrated, especially in species where differences between male and female recombination rates are mostly subtle, including several livestock species.

How are animal geneticists then testing the imprinting status of QTL in line-crosses ? The previous conclusion

assumes that alternate marker alleles (1 and 2) are fixed in the parental lines, as expected if these are inbred. However, lines or breeds of domestic animals are far from inbred, as demonstrated by the segregation of multiple marker alleles (e.g. 1^1, 1^2, 1^3, ...; 2^1, 2^2, 2^3 ...; note that there is a considerable degree of allele sharing between breed/lines, i.e. 1^m and 2^n may be identical-by-state for some *m*'s and *n*'s.). As a consequence F1 parents often have distinct heterozygous marker genotypes. In these cases the two classes of heterozygous F2 offspring (1^m2^n and $2^{n'}1^{m'}$) can be distinguished and imprinting can be tested.

The caveat resides in the fact that animal geneticist typically assume that alternate QTL alleles are still fixed in the parental lines/breeds (even if marker alleles obviously are not) and that all F1 individuals have *Qq* genotype[102]. If this is not the case, F1 animals may have different QTL genotypes, and the estimated QTL allele substitution effect will be the weighted average of allele substitution effects across F1 parents. As the number of F1 parents that is used to generate the F2 population is typically limited – especially on the male side – the average madumnal and padumnal substitution effects may differ as a result of varying QTL genotype frequencies between F1 males and females simply due to sampling. As recognized by De Koning[103], there is thus a much more trivial explanation for the frequent parent-of-origin effects observed for QTL in domestic animals than genuine parental imprinting.

It has recently become apparent that linkage disequilibrium (LD) extends over much longer chromosome segments in breeds of domestic animals than in human. Significant gametic associations between markers are readily detected using low-density microsatellites maps[66,104,105,106], and *a fortiori* using medium or high density density SNP panels[107,108,109]. We reasoned that if LD occurs between markers and QTL – as expected – the imprinting hypothesis will preferentially be testable in offspring from matings between F1 parents that differ for their QTL genotype as a result of their distinct marker genotype. This would irremediably result in a difference between the madumnal and padumnal allele substitution effect within families, i.e. a QTL with parent-of-origin effect. But would this hypothesis translate in an increase in "spurious" QTL imprinting across families ? To address this question, we herein examine the effect of LD on the incidence of parent-of-origin effects when mapping QTL in line-cross designs typical of studies in domestic animals.

III.2 Methods.

III.2.1 Simulations.

We first produced a "base" population by randomly mating 1,000 males and 1,000 females for 10,000

generations using a recombination-mutation-drift model. Twenty-one unlinked QTL were "dropped" in this population, with a mutation rate $\mu_{QTL}=10^{-5}$ under an ~ infinite-alleles model (maximum number of alleles = 1,000). Each QTL was positioned at position 22cM of a 50 cM chromosome (a different chromosome for each QTL) with six evenly spaced markers characterized by a mutation rate $\mu_M = 10^{-3}$ under the same ~ infinite-alleles model. QTL and marker were assumed to be monomorphic in the first generation.

At generation 10,000, we assigned an allelic effect to each QTL allele in the population. QTL effects were sampled in a double negative exponential distribution, $\frac{e^{-\delta x}}{2\delta}$, and randomly assigned a positive or negative sign[110]. The shape parameter, δ was set at one.

The "breeding value" (BV) of each individual was computed as the sum of effects across all its constituent – padumnal and madumnal - QTL alleles. The simulated QTL were thus considered to act in a Mendelian fashion, i.e. none of them were imprinted. In addition, each individual was assigned a non-genetic effect sampled from a normal distribution $N(0,\sigma_E^2)$.

The value of σ_E^2 was set such that $h^2 = \sigma_{BV}^2 / (\sigma_{BV}^2 + \sigma_E^2)$ was either 0.4 or 0.8 at generation 10,000. It remained unchanged in subsequent generations (see hereafter). The individual's non-genetic effect was added to the breeding value to yield its "phenotypic value", y_i.

From this "base" population we then generated two distinct populations (each with 200 males and 500 females) by divergent selection on the "phenotypic values". This was achieved by assigning differential probabilities of parenthood to individuals as a function of their phenotype. Probabilities of parenthood of individual *i* of sex *s* were computed as:

$$P_{is,H} = \frac{(y_{is} - Min + 1)^{\lambda}}{\sum_{i=1}^{n} (y_{is} - Min + 1)^{\lambda}} \quad (4)$$

and

$$P_{is,L} = \frac{(Max - y_{is} + 1)^{\lambda}}{\sum_{i=1}^{n} (Max - y_{is} + 1)^{\lambda}} \quad (5)$$

in the population selected for "high" ($P_{is,H}$) versus "low" ($P_{is,L}$) phenotypic value, respectively. *Min* and *Max*

are the lowest and highest phenotypic values in the respective populations, and λ determines the selection intensity. Selection was applied for up to 20 generations. Mutation and recombination operated during selection as before.

From these selected populations, we sampled 20 F_0 sires in one line, and 80 F_0 dams in the other. The choice of lines (high or low) from which to sample the F_0 sires or dams was random. Each F_0 sire was then mated to four F_0 dams producing five offspring each, for a total of 400 F_1. From these we randomly sampled 80 F_1 dams to generate five F_2 offspring each by equitably mating them with either (i) 2 F_1 sires, or (ii) 10 F_1 sires, or (iii) 20 F_1 sires, for a total of 400 F_2 offspring. This procedure was in essence identical to the one used by De Koning[99]

For each simulation, we selected the QTL contributing most to the genetic variance, σ^2_{BV} , for linkage disequilibrium analysis and QTL mapping (see hereafter). The selected QTL will be referred to as the "target QTL" in the remainder of the manuscript.

III.2.2 Linkage disequilibrium.

The simulation procedure generates LD between markers and between the markers and the QTL. Pair-wise LD was quantified using r^2 as described (Grisard et al.)[111]. To neutralize the effect of LD and hence being able to rigorously quantify its effect on the detection of "imprinted" QTL, we randomly permutated marker genotypes within F_0 sires and dams, respectively, prior to "dropping" the chromosomes down the F1 and F2 generations. The latter comprised exactly the same individuals as before, thus not altering the distribution of QTL genotypes in the pedigree.

III.2.3 QTL mapping.

QTL mapping of the target QTL was performed by least-square analysis using the line-cross model proposed Haley[102] and Knott[112]. Briefly, the presence of a Mendelian QTL at a given map position, *j*, was tested assuming that

$$y_i = \mu + a(p_{11}^{ij} - p_{22}^{ij}) + d(p_{12}^{ij} + p_{21}^{ij}) + \varepsilon_{ij} \quad (6)$$

while the presence of an "imprinted" QTL was tested assuming that

$$y_i = \mu + a(p_{11}^{ij} - p_{22}^{ij}) + d(p_{12}^{ij} + p_{21}^{ij}) + m(p_{12}^{ij} - p_{21}^{ij}) + \varepsilon_{ij} \quad (7)$$

In these y_i is the phenotypic value of individual *i*, μ is the mid-point between the phenotypic value of alternate

homozygotes, p_{xy}^{ij} is the probability that individual *i* has genotype "*xy*" at map position *j* (where *x* is the madumnal allele and *y* the padumnal allele) conditional on flanking marker genotypes, *a* is the additive effect, *d* the dominance deviation, *m* the "imprinting" deviation, and ε_{ij} the error term for individual *i* at map position j. The values of μ, *a*, *d* and *m* are determined in order to minimize the sum of squared error terms (*SSE*) over all offspring. Under the null hypothesis of no QTL

$$\frac{SSR/k}{SSE/(n-k-1)} \quad (8)$$

has an *F*-distribution with *k* and *n-k-1* degrees of freedom. *SSR* corresponds to the sum of squares due to the model, *n* to the number of offspring, and *k* to the number of parameters to estimate, i.e. two in the Mendelian scenario and three when testing "imprinting".

Nominal *p*-values of the *F*-statistics were corrected for multiple testing (51 positions and two models) by permutation testing[113]. The corrected *p*-value corresponds to their rank (divided by 1,000) with respect to the lowest *p*-values (across positions and models) obtained across 1,000 permutated data-sets. The threshold for significance was set at 0.05.

For data-sets yielding significant QTL, if the most significant *p*-value on the chromosome was obtained with the Mendelian model, the QTL was assumed to be Mendelian. If the most significant *p*-value was obtained with the "imprinting" model, we rejected the Mendelian model in favour of the "imprinting" model only if at the most significant position of the imprinted QTL.

$$\frac{SSR_{IMP} - SSR_{MEND}}{SSE_{MEND}/(n-4)} = F_{1,n-4} \quad (9)$$

had a *p*–value < 0.05

III.3 Results.

Figure III.1.A shows the progression of the level of polymorphism during the creation of the base population (N_e = 2,000). The proportion of polymorphic loci, defined as loci with major allele frequency ~ 0.95, increases to 0.90 for marker loci and 0.69 for QTL. Concomitantly the rate of heterozygosity plateaus at 0.44 and 0.28 for markers and QTL respectively, complying with theoretical expectation: $H_e = 4N_e\mu/(4N_e\mu+1)$

Figure III.1.B shows the evolution of the same metrics in the selection lines (N_e = 572). Without selection (λ=0),

proportion of polymorphic loci and rate of heterozygosity decrease slowly towards equilibrium values dictated by effective population size. With selection (λ=2), the reduction of polymorphism is accelerated at all QTL and particularly at the target QTL selected on the basis of its largest contribution to genetic variance. The proportion of populations for which the target QTL is polymorphic, decreases from ~ 0.70 to ~ 0.35 in 20 generations. Under selection, the decrease of polymorphism is also accelerated at marker loci, particularly those that undergo the strongest hitchhiking effects due to their proximity to the target QTL (data not shown). Figure III.2.A shows the evolution of LD between the target QTL and linked markers in the base population. As expected LD stabilizes at equilibrium values in accordance with theory: $r^2 = 1/(4N_e\theta + 1)$. On creation of the selection lines, r^2 values first undergo a slight increase due to reduction in population size (+ 1/*n*). In the absence of selection (λ=0), LD then increases towards new equilibrium values, while selection (λ=2) clearly augments LD build-up. Figure III.2.B also shows the effect of marker genotype permutations: LD becomes independent of genetic distance and is determined exclusively by population size.

Figure III.3.A shows the frequency distribution of allelic effects for target QTLs as well as for the remaining 20 "background" QTL. Target QTL are typically characterized by one allele with large substitution effect and an average of 3.4 alleles with modest effect (Figure III.6). This situation is reminiscent of what has been observed in reality at for instance the *DGAT1*, *GHR* and *ABCG2* loci in cattle[114;111;73;115], the *MSTN* locus in sheep[116] and the *IGF2* locus in the pig[98]. Many QTL are indeed likely to be poly-allelic *sensu stricto*, yet to behave largely as bi-allelic because of the occurrence of one allele with unusually large allelic effect. Figure III.3B shows the frequency distribution of the proportion of the genetic variance accounted for by the target QTL as well as of the remaining "background" QTL. The average proportion of genetic variance explained was 31% for the target QTL and 2% for a typical background QTL.

Figure. III.4.A-C shows the evolution of the power to detect the target QTL – using a line-cross model - as a function of (i) the number of generations of selection (1 to 20), (ii) the heritability of the trait (0.40 or 0.80), (iii) the F2 pedigree structure (2♂/80♀, 10♂/80♀ and 20♂/80♀). As expected, detection power is very low at onset of selection as both populations have not yet diverged and average substitution effect between QTL allele originating from the "low" and "high" line are concomitantly close to zero in the F2 population. Nevertheless, at selection onset power is slightly higher under the 2♂/80♀ scenario, supposedly reflecting situations where at least one sire is, or both are by chance "concordantly" heterozygous at the QTL. Detection power systematically increases with the number of generations of selection under all scenarios, as a result of the increasing divergence of the parental lines at the target QTL. It reaches slightly higher levels under the 10♂/80♀ and 20♂/80♀ scenarios, than under the 2♂/80♀ scenario. The reason for this remains uncertain but may be due at least in part to a higher variance in marker information content across simulations in the 2♂/80♀ than in the other scenarios.

As expected, decreased heritability decreases detection power under all scenarios as the additional non-genetic noise reduces the ratio of QTL variance to phenotypic variance and hence the ability to detect the QTL.

Strikingly, the proportion of cases in which one erroneously concludes that a detected QTL (i.e. exceeding the 0.05 significance threshold as defined in M&M) is imprinted is superior to 60% (average: 84%) for the ten first generations of selection under the 2♂/80♀ scenario higher than 20% (average: 34%) under the 10♂/80♀ scenario, and higher than 15% (average: 24%) under the 20♂/80♀ scenario (Figure III.4.A-C). The decrease in the proportion of erroneous conclusions of imprinting with increasing numbers of F1 sires agrees with the findings of De Koning[103], reflecting a decrease in the standard error of the estimate of the average padumnal QTL allele substitution effect with increasing number of F1 sires. Erroneous detection of imprinting decreases with increasing generations of selection, yet remains superior to 18%, 9% and 8% at generation 20 under the 2♂/80♀, 10♂/80♀ and 20♂/80♀ scenarios, respectively. This decrease is presumably due to the fact that alternate QTL alleles are getting to near fixation in the "high" and "low" line in an increasing proportion of simulations, thereby reducing the variation in QTL genotype amongst F1 parents - the source of the problem. Indeed, when restricting the analyses to crosses between selected lines for which the average QTL substitution effect *within* lines is larger than the average substitution effect *between* lines, the proportion of erroneous imprinting calls remains very high: 88% on average under the 2♂/80♀ scenario, 36% on average under the 10♂/80♀ scenario and 26% on average under the 20♂/80♀ scenario (Figure III.5.A-C).

Erroneous detection of imprinting increased slightly with decreasing heritability (Figure. III.3.A-C and data not shown). We believe that this reflects less effective selection and hence more prolonged segregation of distinct QTL alleles in the high and low lines. Indeed, this effect of heritability disappeared when restricting the analyses to crosses between selected lines for which the average QTL substitution effect *within* lines is larger than the average substitution effect between lines (Figure III.5.A-C).

The contribution of LD to erroneous detection of imprinting – the main hypothesis tested in this study - is clearly demonstrated by examining the effect of marker genotype permutation to neutralize LD between markers and QTL in the selection lines prior to mating (Figure III.5.A-C). While erroneous imprinting calls indeed occur in the absence of LD as well, corroborating the results of de Koning[103], the incidence of this artefact is increased by ~ 40% on average under the 2♂/80♀ scenario, by ~ 70% under the 10♂/80♀ scenario, and by ~ 80% under the 20♂/80♀ scenario. These results thus clearly show that - at least under the studied scenarios - LD considerably exacerbates the problem, as surmised. Note that the QTL detection power is unaffected by the presence or absence of LD (Figure III.4.A-C).

III.4 Discussion.

QTL mapping in domestic animals is routinely performed in F2 populations derived from parental lines that are divergent for at least some of the analyzed traits. The most commonly applied statistical models assume that the parental lines are fixed for alternate QTL alleles. Yet it is becoming increasingly apparent that in many cases this assumption is not valid: parental lines are often segregating for the QTL as well as sharing QTL alleles. With hindsight and having identified the causal mutation, this is now clearly demonstrated for the *IGF2*-intron-nt3072 QTN[98].

Applying a line-cross model to such data is suboptimal and may lead to loss of QTL detection power. Moreover, as previously noticed by De Koning[103] and extended in this work, it may lead to erroneous conclusions about the mode of action of the QTL, particularly with regards to parental imprinting De Koning[103] realized that if the parental lines are not fixed for alternate QTL alleles, the “average” QTL genotype of the utilized F1 sires and dams may differ as a result of sampling, thus resulting in a difference between the padumnal and madumnal QTL allele substitution effects even in the absence of true parental imprinting. We herein demonstrate that if LD exists between the markers and the QTL, “pseudo-imprinting” may become the rule rather than the exception in scenarios that are likely to properly mimic real livestock situations. We surmise that this is due to the fact that in the presence of LD the padumnal and madumnal QTL allele substitution effects are more effectively contrasted in matings between parents with distinct marker and hence QTL genotype. We believe that this accounts for the fact that imprinted QTL effects have been so commonly observed in line-crosses.

It may be argued that many commercial populations have been selected for more than 20 generations, the time horizon considered in our simulations. However, selection objectives often change over time, adjusting to evolving economic constraints. As an example, it is likely that the *DGAT1 K232* allele was initially selected for in dairy cattle because of its favourable effect on milk yield, until fat yield became the primary objective hence favouring the alternate *232A* allele, while this polymorphisms is now neutral with respect to the present-day selection Dutch index which weights both fat and protein yield thus explaining its continued segregation[114]. Moreover, QTL mapping experiments often target traits that are of major scientific interests without necessarily being the focus of ongoing breeding programs. Thus segregation of QTL alleles is likely to be a common occurrence in parental livestock populations used to generate line-crosses.

Such crosses are thus better analyzed using models that do not make the assumption that all alleles originating from a given parental line have the same QTL effect. This is sometimes achieved by analyzing line-cross data using a half-sib design, taking advantage of the common occurrence of large paternal half-sib pedigrees. This

approach, however, foregoes the information from the maternal chromosomes. More appropriate is the use of variance component (VC) approaches that either exploit linkage information alone[117], or simultaneously exploits linkage and LD signal to compute the probability of identity-by-descent (IBD) between chromosomes, used to constrain the covariances between the corresponding haplotype effects[118]. Parent-of-origin effects could be tested within a VC framework by estimating the variance associated with the madumnal and padumnal haplotypes separately[119] and comparing the likelihood of the data assuming that these are different versus identical. However, this approach would suffer from the same drawback as the line-cross model in that the identification of a statistically different madumnal versus padumnal VC does not imply that genuine parental imprinting is involved.

Demonstrating genuine imprinting thus requires the comparison of the QTL allele substitution effect of a proven IBD pair of alleles upon paternal versus maternal transmission. This is difficult to achieve in livestock as it is only exceptionally possible to have a sufficiently large number of F1 dams that have a genotype that is unambiguously IBD with that of one or more F1 sires.

The availability of hundreds of thousands to millions of SNPs for most livestock species, combined with cost-effective high throughput screening techniques, will soon allow the recognition of haplotypes that are known with virtual certainty to be IBD even in the absence of pedigree data. It should thus become possible to test the imprinting hypothesis by comparing the phenotypes of alternate heterozygotes (“ 12’s versus 21’s ”) at the population level. However, and as noted by others, this approach hides potential pitfalls as well, as maternal effects may render this contrast significant in the absence of genuine parental imprinting[120].

Imprinted genes certainly contribute to the genetic variation of quantitative traits[94,98], and forward genetics, including QTL mapping, may contribute to the identification of novel imprinted genes[89]. However, it is important to recognize that alternative “artefactual” explanations may account for the parent-of-origin effects that are often found for QTL detected in line-crosses. Demonstrating a role for genuine parental imprinting requires more stringent tests than what is usually being applied. For sure, available QTL evidence does not warrant re-evaluation of how common parental imprinting is both in terms of number of affected genes and species in which it occurs.

III.5 Acknowledgments.

Cynthia Sandor is a fellow of the “Fonds pour la Recherche dans l’Industrie et l’Agriculture” (FRIA). Support

for this work was provided by the Walloon Ministry of Agriculture, the Belgian Science Policy organization (SSTC Biomagnet PAI), and the Communauté Francaise de Belgique (Biomod ARC)

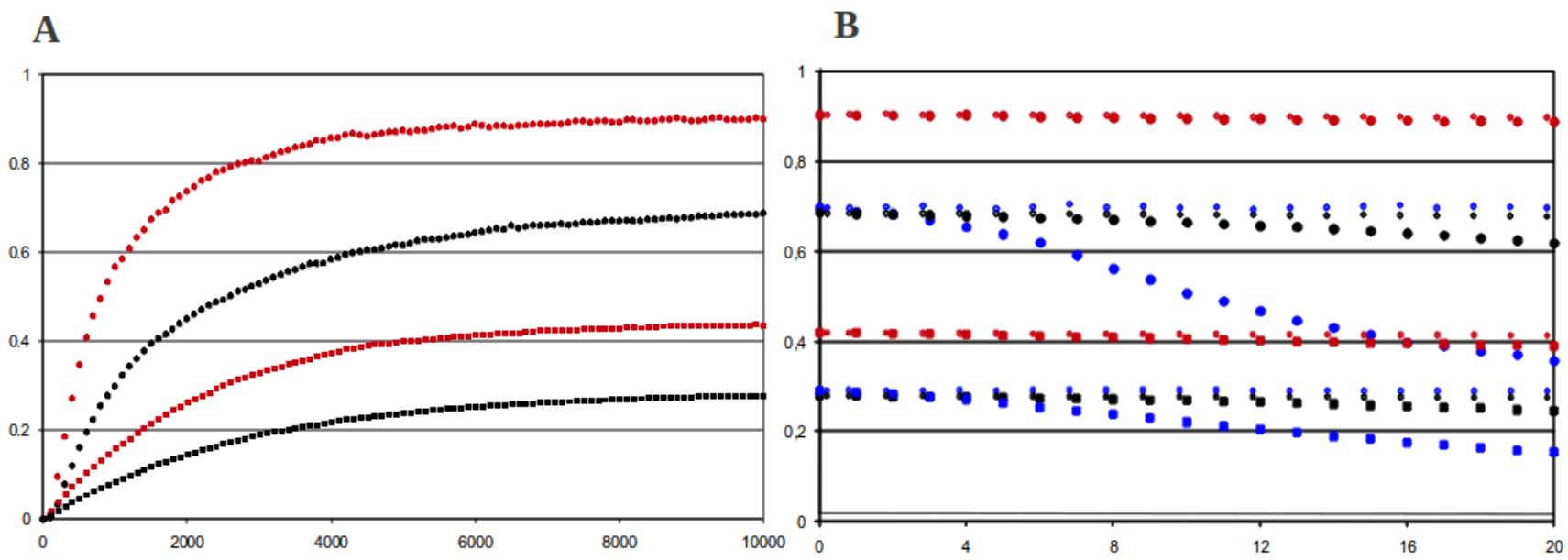


Figure III.1: *(A) Evolution of the proportion of polymorphic loci (major allele frequency < 0.95) (circles) and rate of heterozygosity (squares) for marker loci (red) and QTL (black) during the creation of the base population (10,000 generations). Data points are shown every 100 generation. Each data point corresponds to the average over six (markers), respectively 21 (QTL) loci and 10,000 simulations. (B) Evolution of the same metrics in the selection lines (20 generations) in the presence (λ=2) and absence (λ=0; miniatures) of selection. The data for the target QTL are shown in blue. Every data point corresponds to the average across six (markers), one (target QTL) and 20 (background QTL) loci times 10,000 simulations.*

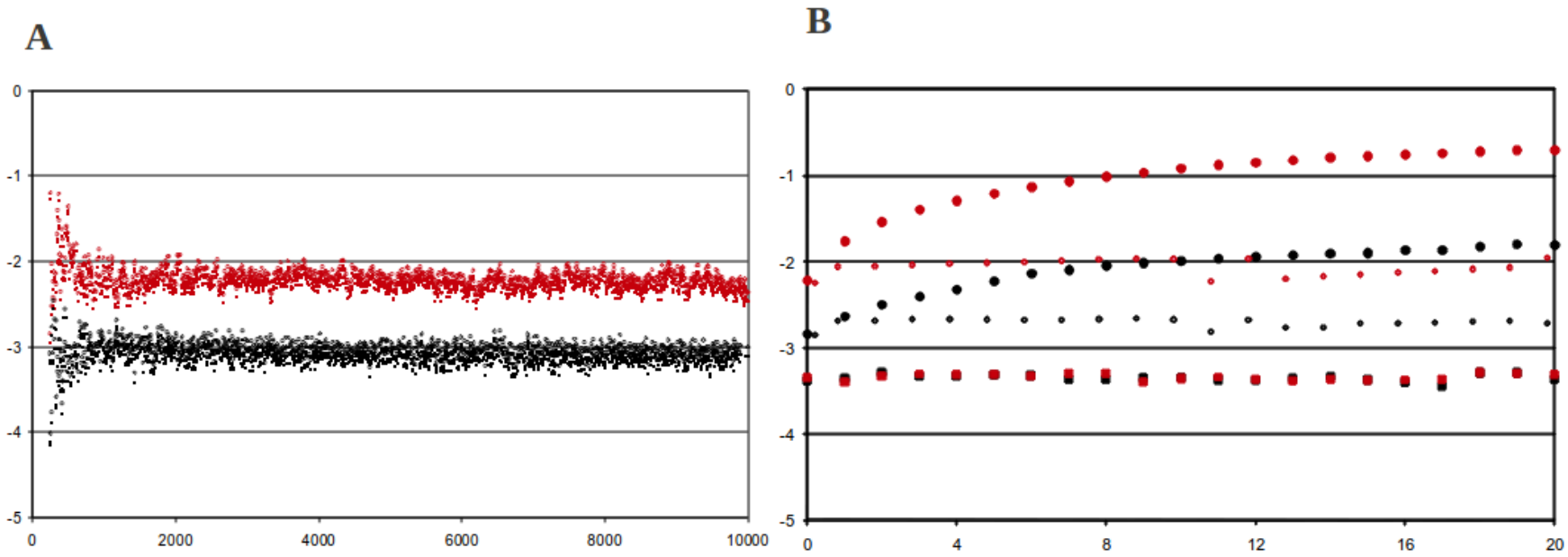


Figure III.2: *(A) Evolution of LD (log10(r2)) between markers and target QTL during creation of the base population. Data points correspond to the average of 1,000 simulations and are – for clarity - only shown for the marker at 2 cM (red circles) and the marker at 18 cM (black circles) from the QTL. (B) Same LD metrics in the selection lines (20 generations), with (λ=2) and without (λ=0; miniatures) selection. The effect of genotype permutations on LD is shown for the same two marker-QTL combinations (red and black squares). Every data point corresponds to the average across 1,000 simulations.*

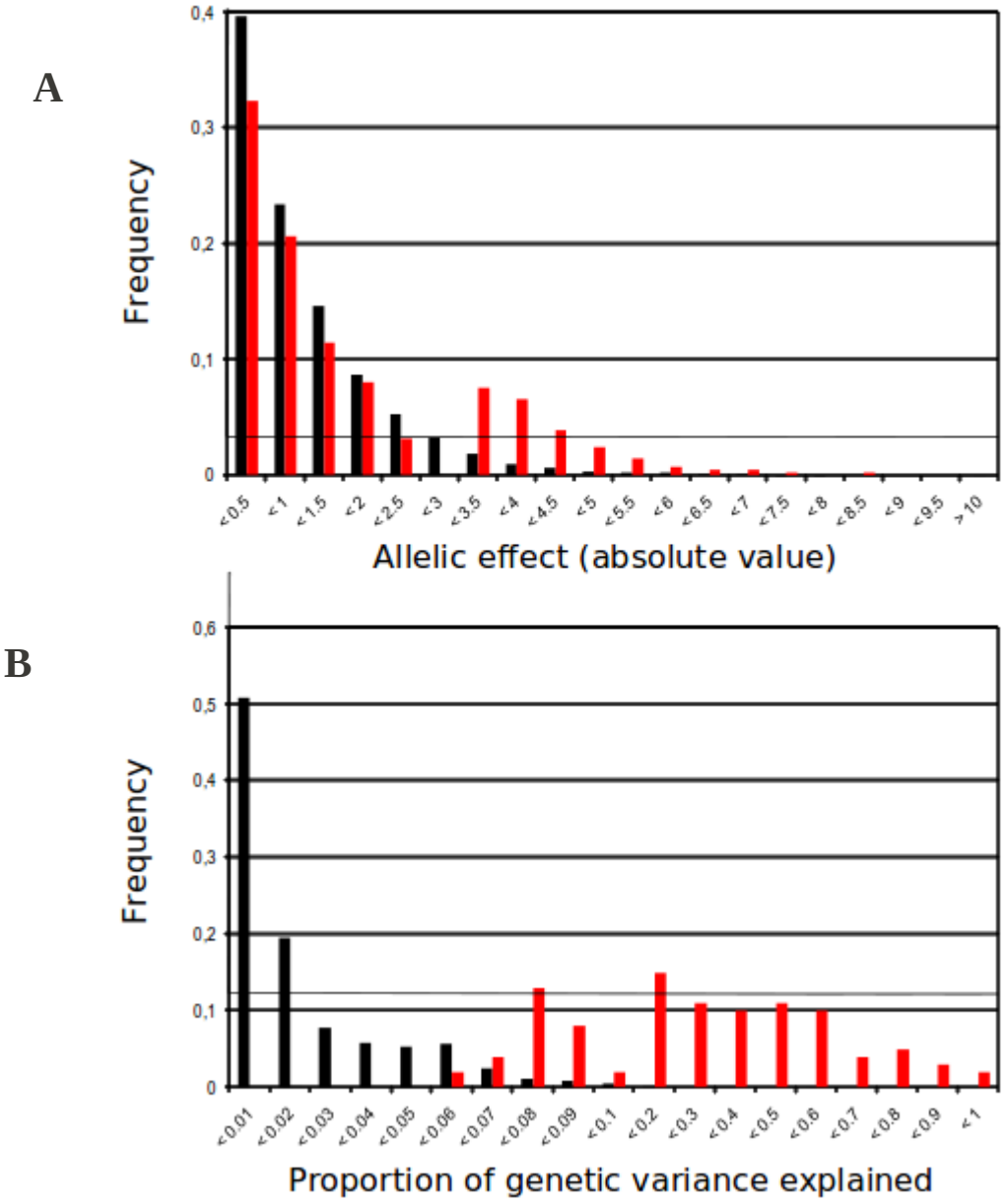


Figure III.3: *(A) Frequency distribution (over 100 simulations) of QTL allelic effects (absolute value) for the target QTL (red), as well as across the 20 remaining background QTL (black). (B) Frequency distribution of the proportion of genetic variance explained by the target QTL (red) as well as by individual background QTL.*

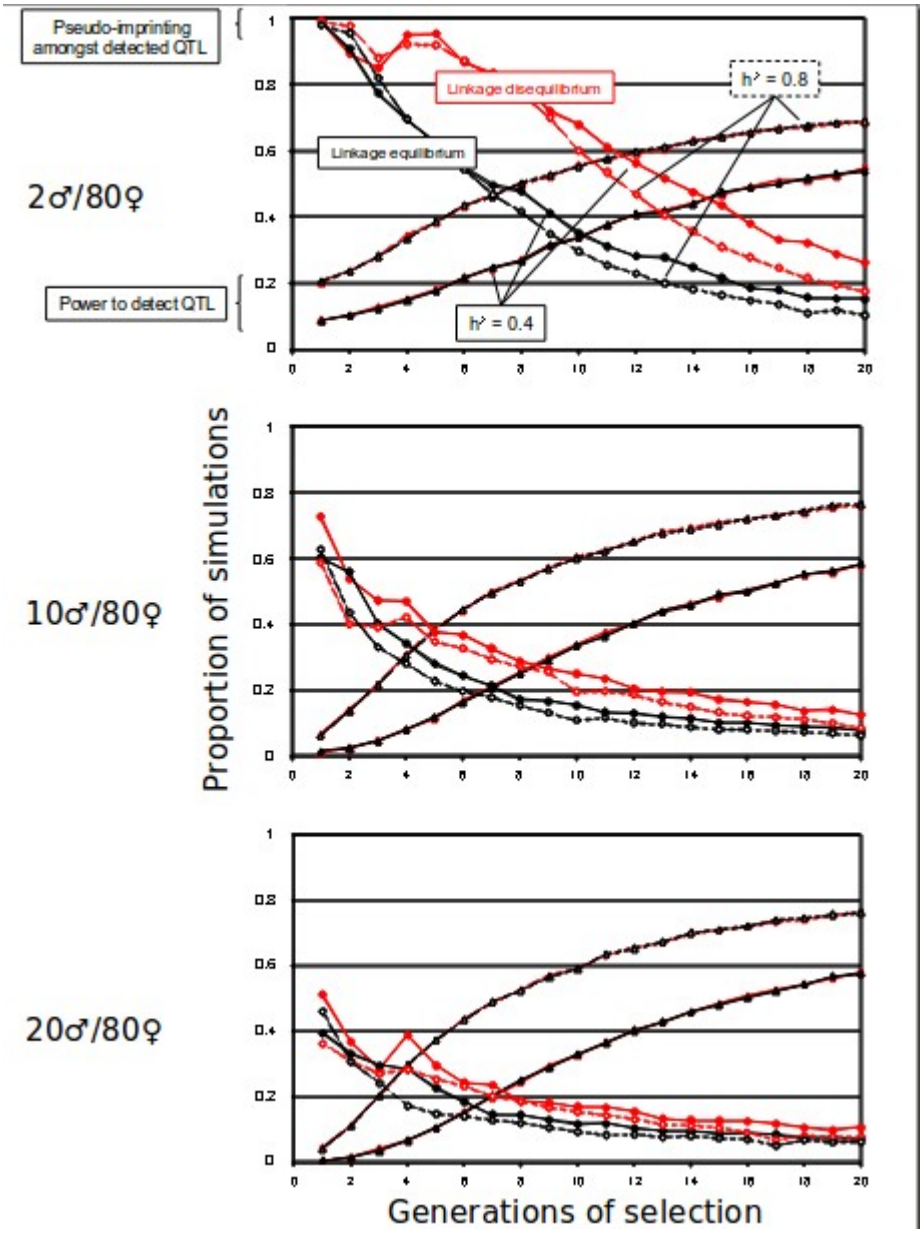


Figure III.4: *(A-C) Power to detect target QTL (triangles), and incidence of erroneous conclusions of imprinting amongst detected QTL (circles), as a function of the number of generations of selection (1-20), in the presence (red) and absence (black) of LD between markers and QTL, under the 2♂/80♀ (A), 10♂/80♀ (B), and 20♂/80♀ (C) F1 scenarios, for a trait with heritability of 40% (filled symbols and continuous lines) or 80% (empty symbols and dotted lines). Data points correspond to averages over 10,000 simulations.*

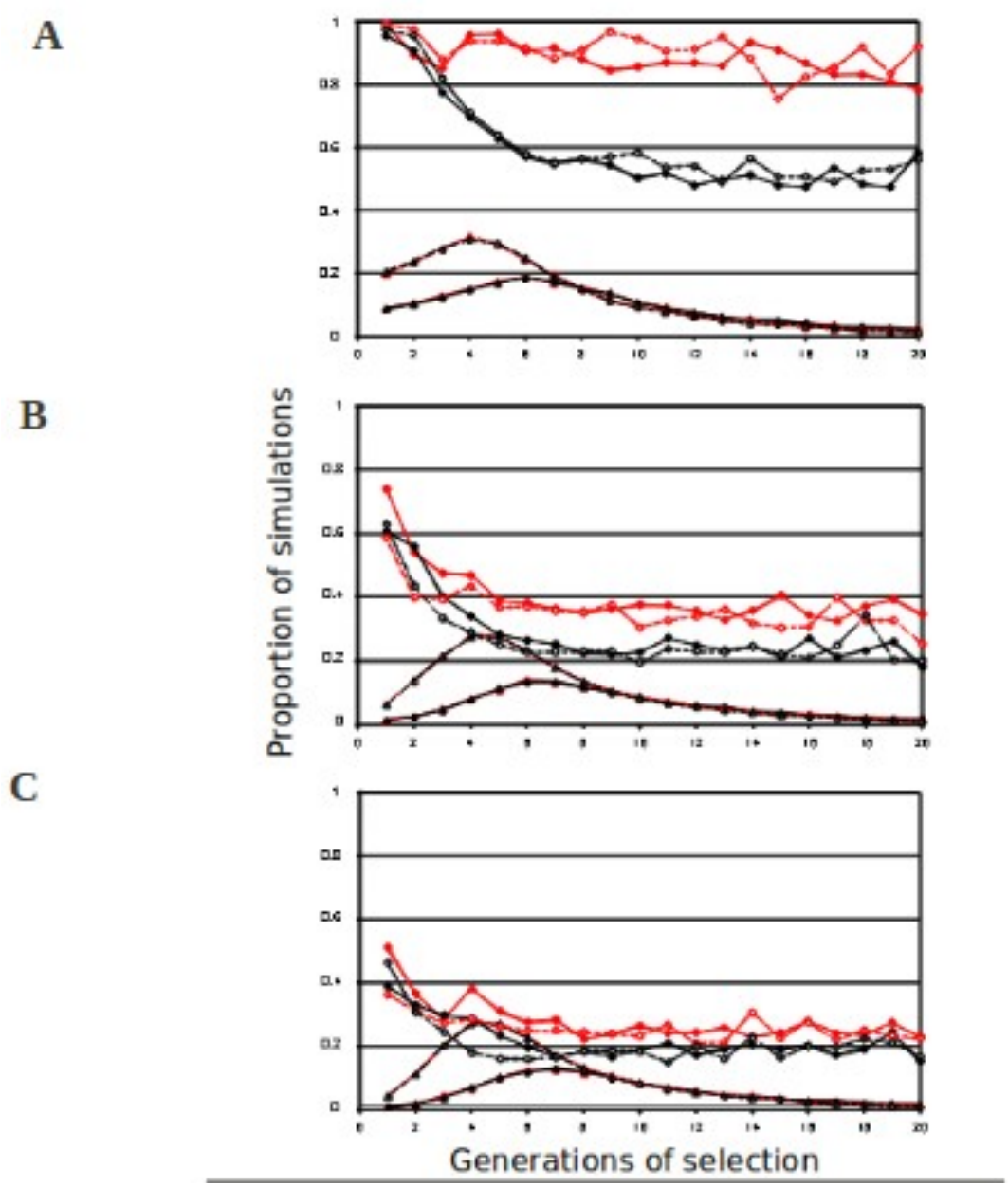


Figure III.5: *(A-C) Proportion of simulations for which the average QTL allele substitution effects within lines is superior to the average QTL allele substitution effects between lines (triangles). Incidence of erroneous conclusions of imprinting amongst detected QTL (circles) in the presence (red) and absence (black) of LD between markers and QTL, under the 2♂/80♀ (A), 10♂/80♀ (B), and 20♂/80♀ (C) F1 scenarios, for a trait with heritability of 40% (filled symbols and continuous lines) or 80% (empty symbols and dotted lines). Data points correspond to averages over 10,000 simulations.*

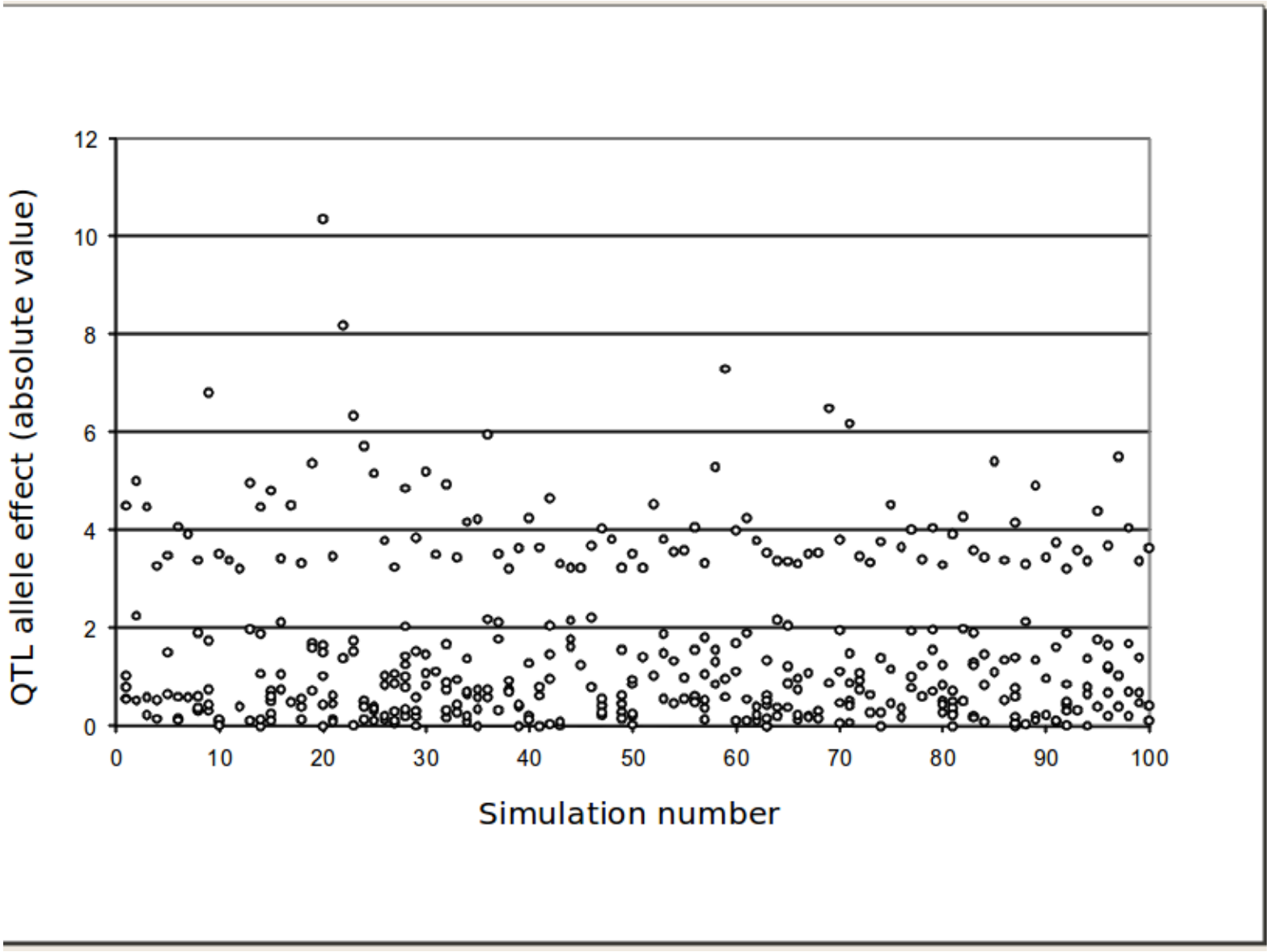


Figure III.6: *Hundred representative examples of QTL allele effects (absolute value) at the target QTL.*

IV Deux exemples d'utilisation des bases de données eQTL dans le cadre de l'élucidation des facteurs génétiques impliqués dans la maladie de Crohn.

IV.1 Introduction.

Depuis 2007, les GWAS, utilisées à fin d'élucider les facteurs génétiques impliqués dans des maladies complexes chez l'homme, telles que les maladies inflammatoires intestinales, le diabète, l'obésité et le cancer se sont multipliées. Grâce à ces études, un grand nombre de nouveaux loci ont été découverts, dont certains associés à des valeurs p très significatives. Cependant il existe rarement de lien évident entre ces nouvelles variations génétiques et d'éventuels effets fonctionnels, comme peuvent montrer les trois mutations non synonymes dans le gène NOD2 (*nucleotide-binding oligomerization domain containing 2*) découvertes antérieurement aux GWAS dans le cadre de la maladie de Crohn (CD ou *Crohn Disease*)[121]. Une hypothèse avancée est que ces nouveaux polymorphismes de prédisposition à des maladies génétiques sont peut être impliqués dans des mécanismes qui affectent les niveaux d'expression de certains gènes, plutôt que dans des altérations moléculaires de gènes. En effet l'abondance des transcrits d'un gène peut être modulée par un polymorphisme présent dans un élément régulateur de ce gène[122]. Le corollaire est que le niveau d'expression d'un gène peut être considéré comme un phénotype quantitatif pour lequel il est possible de rechercher des régions génomiques expliquant les différences entre individus. Ces loci sont appelés des eQTL (*Expression Quantitative Trait Locus*). La réalisation sur un même groupe d'individus d'une GWAS et d'une étude d'expression sur un grand nombre de gènes permet grâce à des méthodes de cartographie classique d'identifier des eQTL. Plusieurs études de ce type ont été réalisées, elles répertorient les effets trans et cis de SNPs sur l'expression d'un grand nombre de gènes[123]. Ces bases de donnés d'eQTL (http://www.sph.umich.edu/csg/liang/asthma/) ont deux applications possibles: elle peuvent d'une part améliorer l'interprétation du rôle biologique des SNPs ressortant des GWAS en examinant si ces derniers sont associés également à l'expression d'un ou plusieurs gènes, d'autre part elles peuvent être utilisées dans l'identification d'un réseau de gènes impliqués dans une maladie. Nous présentons ici, deux exemples d'utilisation de ces bases d'eQTL dans le cadre de GWAS portant sur la CD. Dans les deux études nous avons

tenté de préciser le rôle biologique de ces nouveaux SNPs associés à CD, mais la première étude se focalise sur une région particulière du génome, un ensemble de SNPs présents dans un désert génétique sur le chromosome 5 dans une seule cohorte d'atteints (et de contrôles)[124] tandis que la deuxième étude s'intéresse à une trentaine de SNPs répartis dans tout le génome, ressortant d'une étude d'association combinant plusieurs cohortes d'atteints (et de contrôles) (d'origine caucasienne), de type méta-analyse[125].

IV.2 Matériels et Méthodes.

Dans les deux études nous avons employé la base de données d'eQTL développé par Dixon et al.[123]. Les données d'expression ont été générées à partir d'ARN (*acide ribonucléique*) provenant de lignées lymphoblastoïdes (lymphoblastes modifiées par le virus Epstein-Barr, rendant ces lignées cellulaires « immortelles ») de 378 descendants génotypés et appartenant à des familles nucléaires.

Grâce à l'utilisation de puces à haute densité en oligonucléotides (ici human GeneChip U133 Plus 2.0 d'Affymetrix) des mesures d'expression de milliers gènes peuvent être réalisées simultanément sur des centaines d'échantillons. Le principe de cette technologie repose sur l'existence de séquences consensus, qui servent de références et sont spécifiques d'un gène donné. A partir de ces séquences consensus on garde les oligonucléotides de 25 pb (sondes), les plus spécifiques d'un gène. Ces sondes fonctionnent par paire: une sonde référence, appelée PM (« perfect match ») s'appariant parfaitement avec la séquence de référence et une deuxième sonde, dite sonde MM (« mismatch ») qui diffère au centre d'un seul nucléotide. La différence entre le taux d'hybridation de la sonde PM et la sonde MM donne une idée de la spécificité d'hybridation de la sonde PM. Les échantillons d'ARN sont marqués et hybridés sur ces puces à hautes densité en oligonucléotides. Après un scan, on obtient une image de la lame représentant la fluorescence des différents spots qui après analyse donnera une mesure d'intensité pour chaque sonde.

Ces mesures d'intensité par sonde sont ensuite converties en mesure d'expression de gène en trois étapes. Les mesures d'intensité des sondes PM sont tout d'abord corrigées pour le bruit fond par une méthode appelée RMA (« Robust MultiChip Average »). Cette méthode utilise log(PM) comme mesure d'expression et l'ajuste pour le bruit fond en modélisant l'intensité des sondes PM comme une somme du vrai signal et du bruit fond causé par une hybridation non spécifique. Cette mesure a un double intérêt: (i) il ne s'agit pas d'une différence entre les taux d'hybridation (ou rapport) des sondes PM et MM, qui enlève du vrai signal et n'est pas toujours proportionnelle avec les concentrations en ARN (les sondes MM détectent aussi bien le bruit fond que du vrai signal); (ii) cette mesure permet de distinguer deux signaux de faible intensité mais qui différent l'un de l'autre

proportionnellement. Ces données sont ensuite normalisées par une méthode afin d'écarter au mieux la variation résiduelle introduite durant l'expérimentation. Les différences entre échantillons dans les mesures d'expression peuvent être dues à des facteur génétiques ou environnementaux mais aussi peuvent provenir d'erreurs ou de biais intervenant lors de l'expérimentation. Il a été montré que si les donnés ne sont pas normalisés, les comparaisons des mesure d'expression de différent échantillons peuvent mener à des faux positif. Les données ont donc été normalisées dans un deuxièmement temps avec une méthode de normalisation quantile. Le but de cette normalisation est de faire en sorte que la distribution des intensités des sondes pour chaque échantillon soit la même. L'idée vient d'une représentation graphique, appelée QQ plot permettant de comparer deux distributions. Si la distribution de deux vecteurs de données est la même alors le graphique donne une droite dans la diagonale et autre chose si ce n'est pas le cas. Pour normaliser les données de deux échantillons on prendra les valeurs projetées sur la diagonale. Ce concept peut s'étendre à n dimensions et donc peut s'appliquer à n échantillons. Cette méthode permet d'éviter de déterminer une référence pour normaliser les données. On réalise ensuite une normalisation inverse dans le but de détecter et d'écarter les valeurs aberrantes et extrêmes.

Une mesure d'association entre la mesure d'expression de chaque transcrit et chaque SNP a été calculée en utilisant le logiciel Merlin. On a estimé ensuite l'effet additif de chaque SNP sur l'expression de chaque gène et testé sa signification en tenant compte de la précision d'imputation des génotypes manquants.

IV.3 Résultats.

Dans la première étude, nous avons trouvé dans le cadre d'une GWA une nouvelle région associée à la CD[124]. Cette région était couverte par 111 SNPs répartis dans cinq blocs haplotypiques de tailles différentes. Cependant de part sa localisation, dans un désert génétique de 1.25 Mb, une interprétation biologique de son rôle dans la pathogénie ne pouvait être spontanément proposée. Grâce à la base d'eQTL de Dixon et al.[126], nous avons montré que plusieurs de ces SNPs agissaient comme des facteurs cis influençant l'expression du gène PTGER4 (*protaglandin E receptor 4*) présent à 270 kb en amont de la région d'intérêt. En effet, sur 26 marqueurs présents dans la base d'eQTL et dans la région d'intérêt huit montraient une association avec l'expression de PTGER4 avec une valeur $p < 0.001$ (voir Figure IV.2). Les deux marqueurs avec les valeurs p les plus faibles étaient, dans le bloc IV (rs4495224) et V (rs7720838), en déséquilibre de liaison (D'=0.84). Les allèles à risque de ces deux marqueurs étaient associés à une augmentation dans l'expression de PTGER4.

Dans la seconde étude d'association de type méta-analyse, une approche identique a été appliquée systématiquement à 39 SNPs répartis dans 32 régions génomiques différentes et ayant une association

significative avec la CD[125]. Nous avons recherché d'éventuels effets régulateurs cis de ces 39 SNPs sur l'expression de gènes voisins et distants de 250 kb maximum des régions génomiques d'intérêt dont la taille était fonction de l'étendue du déséquilibre de liaison. Nous avons détecté cinq eQTL associés à l'expression de gènes ayant un lodscore > 2. Nous avons ensuite voulu tester si ces co-localisations SNPs de susceptibilité à la CD avec des gènes dont les niveaux d'expression sont corrélés aux génotypes de ces SNPs étaient simplement dues au hasard ou si elles pouvaient indiquer un processus biologique sous-jacent. Pour cela nous avons réalisé des simulations dans lesquelles nous avons tiré 39 SNPs ayant des fréquences alléliques et des situations génomiques comparables aux vrais SNPs. Nous avons regardé ensuite la distribution du nombre de gènes avec un lodscore > 2, ainsi que la distribution de la somme des lodscores pour des gènes ayant un lodscore > 2 et comparer celles ci avec les valeurs correspondantes obtenues avec les 39 SNPs associés avec la CD. Les valeurs observées excédent largement les valeurs attendues par chance (P ~ 0.001) indiquant probablement un processus biologique sous-jacent (voir Figure IV.7 et IV.8)

IV.4 Discussion.

Ces deux exemples illustrent dans quelle mesure les eQTL offrent un nouvel éclairage sur les bases biologiques des loci ressortant dans les études d'association. Dans la première étude grâce à la base de données d'eQTL http://www.sph.umich.edu/csg/liang/asthma/, nous montrons que la nouvelle région génomique associée à la CD pourrait réguler l'expression du gène PTGER4. Or, le gène PTGER4 est un candidat sérieux dans la CD. En effet, des expériences sur des souris « knock-out » pour PTGER4 révèlent que ces dernières développent des colites très sévères sous traitement oral de Dextran Sulphate Sodium, contrairement à d'autres souris « knock-out » pour d'autres types de récepteur aux prostaglandines. Par ailleurs on a par ailleurs remarqué qu'en administrant chez des souris sauvages un antagoniste spécifique de PTGRE4, on augmentait leur susceptibilité aux colites [Kabashima2002]. Nous avons observé dans notre étude que les allèles à risque pour la CD des marqueurs rs4495224 et rs7720838 sont associés à une augmentation de l'expression du gène PTGER4 dans les LCL. Ces résultats semblent donc contredire ceux obtenus à partir du modèle murin.

Dans la seconde étude d'association de type méta-analyse [59], l'utilisation d'une base de donnée eQTL a permis d'identifier des effets régulateurs de type cis de SNPs associés de façon significative à la CD sur l'expression de gènes voisins. Ceci représente une avancée significative dans interprétation biologique des GWA pour la CD car la plupart de ces SNPs étaient dans des régions génomiques sans aucun gène candidat évident (aucun gène ou gène multiple).

Au niveau du *locus* IBD5, les allèles de susceptibilité à la CD de certains SNPs étaient corrélés avec une

diminution d'expression du gène SLC22A5. Par ailleurs l'eQTL le plus significatif (lodscore=20, gène ORMDL3) associé à la CD se trouve dans la même région que d'autre SNPs de suceptibilité à l'asthme. Ceci suggère que les mêmes polymorphismes en perturbant l'expression du gène ORMDL3, sont peut-être impliqués à la fois dans la CD et dans l'asthme.

Malgré la puissance de ces cartes d'eQTL dans l'élucidation des bases génétiques d'une maladie, il existe un grand nombre de limites dans la mise en place et l'utilisation de ces bases de données d'eQTL. La première et la plus importante est liée aux faiblesses des plateformes de type « microarray » employées pour la mesure d'expression de gènes. Les résultats obtenus sont très variables et dépendent de multiples facteurs durant les différentes étapes: (i) préparation de l'échantillon, (ii) l'hybridation et (iii) la mesure d'expression. Par ailleurs, dans le cas de SNPs inclus dans un transcrit, la réaction d'hybridation et donc la mesure d'expression peuvent être affectées par l'allèle présent, donnant ainsi un faux effet cis-eQTL. Cependant les « microarray » sont assez robustes par rapport à ce problème et de plus seulement 15% des sondes ont un SNP polymorphe dans une population donnée. Toutefois dans ce type situation, il sera préférable de confirmer un éventuel effet cis-eQTL par une technique alternative de mesure d'expression comme une PCR quantitative. La plateforme de « microarray » employée a aussi un impact non négligeable sur ces mesures d'expression: le chevauchement des résultats obtenus entre les différentes plateformes de « microarray » oscille entre 30 et 40%. Les nouvelles plateformes pour les mesures l'expression comme les « microarray » interrogeant tout les exons humains connus ou encore des systèmes basés sur le séquençage à haut débit d'ARN devraient rendre à l'avenir les mesures d'expression de gènes plus robustes.

Une autre limite avec ces bases de données eQTL est liée à l'utilisation pratiquement exclusive de lignées lymphoblatoïdes, qui sont des lymphoblastes B infectés par le virus d'Epstein-Barr provocant des divisions cellulaires incontrôlées et les rendant ainsi « immortels ». Des études ont montré que 60 % des gènes exprimés dans d'autres types cellulaires le sont également dans les LCL. Par ailleurs en comparant les eQTL détectés à partir de données d'expression mesurées sur des différents tissus, on obtient un nombre d'eQTL comparables et un chevauchement de 50 % entre les eQTL mis en évidence. Toutefois certains gènes sont vraisemblablement exprimés uniquement dans des types cellulaires spécialisés ou dans le cas de pathologies: c'est l'expression spatiale du transcriptome. De plus les LCL cultivés dans des conditions in vitro, ne sont pas soumis aux stimuli habituels. Les gènes connus comme étant exprimés dans les lymphocytes B des voies aériennes d'individus asthmatiques, ne le sont pas dans les LCL provenant de patients également asthmatiques.

Un autre problème actuel des cartes d'eQTL est d'être établies à partir de cohortes de petite taille comprenant quelques centaines d'individus permettant seulement de détecter les eQTL ayant des effets importants. Le projet Gtex devrait permettre d'améliorer cela anisi que le problème précédent car il vise la création d'une base de

données d'eQTL obtenue à partir d'un millier d'échantillons récoltés à partir d'une trentaine de tissus différents.

Comme dans toutes les études GWA portant sur des caractères et des maladies complexes, les eQTL découverts expliquent une fraction mineure de l'héritabilité et de la composante génétique des niveaux d'expression des gènes. Les causes probables sont: (i) un grand nombre de loci à faibles effets, non détectés à cause de la taille de l'échantillon. (ii) une mauvaise couverture du génome en SNPs testés ou autre types de polymorphismes comme des copy number variants (CNV) (iii) des effets de dominance et d'interaction entre eQTL.

Par ailleurs on sait par ailleurs que des modifications épigénétiques comme la méthylation dans les lignées germinales des ilots CpG ou des altérations d'histones peuvent jouer un rôle très important dans l'expression des gènes. Certains gènes subissent des modifications épigénétiques rendant leur expression monoallélique et dépendante de l'origine parentale: ces gènes sont soumis à de l'empreinte parentale.

IV.5 Paper I: A novel susceptibility locus for Crohn's disease identified by whole genome association maps to a gene desert on chromosome 5p13.1 and modulates the level of expression of the prostaglandin receptor EP4.

Abstract:

To identify novel susceptibility loci for Crohn's disease (CD), we undertook a genome-wide association study with more than 300,000 SNPs characterised in 547 patients and 928 controls. We found three chromosomes regions that provided evidence of disease association with p-values between 10^{-6} and 10^{-9}. Two of these (*IL23R* on chromosome 1 and *CARD15* on chromosome 16) correspond to genes that have been previously reported in CD. In addition, a 250 Kb region of chromosome 5p13.1 was found to contain multiple markers with strongly suggestive evidence of disease association (including four markers with $p<10^{-7}$). We replicated the results for 5p13.1 by studying 1,266 additional CD patients, 559 additional controls and 428 trios. Significant evidence of association ($p < 4x10^{-4}$) was found in case/control comparisons with the replication data, while associated alleles were over-transmitted to affected offspring ($p < 0.05$), thus confirming that the 5p13.1 locus contributes to CD susceptibility. The CD-associated 250 Kb region was saturated with 111 SNP markers. Haplotype analysis supports a complex locus architecture with multiple variants contributing to disease susceptibility. The novel 5p13.1 CD locus is contained within a 1.25 Mb gene desert. We present evidence that disease-associated alleles correlate with quantitative expression levels of the prostaglandin receptor EP4, *PTGER4*, the gene that resides closest to the associated region. Our results identify a major new susceptibility locus for CD, and suggest that genetic variants associated with disease risk at this locus could modulate cis-acting regulatory elements of *PTGER4*.

LIBIOULLE, C.; LOUIS, E.; HANSOUL, S.; SANDOR, C.; FARNIR, F.; FRANCHIMONT, D.; DE WIT, O.; DE VOS, M.; VERMEIRE, S.; DEMARCHE, B.; GUT, I.; HEATH, S.; MNI, M.. ZELENIKA, D.; BELAICHE, J.; RUTGEERTS, P.; LATHROP, M.; GEORGES, M.

Plos Genetics 4e158 (2007)

IV.5.1 Introduction.

Crohn's disease (CD) is a chronic relapsing inflammatory disorder of the intestinal tract, described for the first time in the 1920s[127]. Lifetime prevalence has increased to current estimates of ~ 0.15% in Caucasians. The precise environmental causes underlying this rise remain essentially unknown, but familial clustering and twin-studies clearly identify an inherited component to predisposition. More than ten susceptibility loci have been identified by linkage and/or association studies and convincing causative mutations have been reported, particularly in *CARD15*[128,129]. As known loci don't fully account for the genetic risk for CD we performed a genome-wide association scan (WGA) to contribute to the identification of additional susceptibility loci.

IV.5.2 Results/Discussion.

Genotype data from the Illumina HumanHap300 Genotyping Beadchip[130] were obtained on 547 Caucasian CD patients from Belgium and compared to genotypes for 928 healthy controls from Belgium and France. Genotype call rates were > 93% for all individuals included in the study. Of the total 317,497 SNPs available, 5,615 with genotyping success rate of less than 91% or deviating from Hardy-Weinberg proportions in controls (Fisher's exact test $p \leq 10^{-3}$) were eliminated from further analysis as it is known that less reliable markers generate spurious associations. For the remaining 311,882 SNPs, we compared allele frequencies between cases and controls as outlined in Methods.

Figure VI.1 shows the 10,000 most significant p-values obtained across the human genome. Regions on chromosomes 1, 5 and 16 harboured clusters of markers with suggestive evidence of association at significance levels between 10^{-6} and 10^{-10}. The significance of tests of association with these markers remained within this range after controlling for possible effects of population structure using a backwards stepwise regression[131]. The strongest association was found with markers of the *IL23R* gene on chromosome 1 which has recently been identified as a novel CD susceptibility locus in a case-control and family-based association study of Caucasian and Jewish cohorts[128]. In our data, two markers of the *IL23R* gene, rs11209026 and rs11465804, gave the most significant association signals ($p < 10^{-9}$). Rs11209026 corresponds to an Arg381Gln substitution in IL23R while rs11465804 is intronic and in strong LD with the former marker. A marker within the *CARD15* gene on chromosome 16, which is the first susceptibility gene to have been identified in CD[125], also showed suggestive evidence of association (rs5743289; $p < 10^{-6}$). We also examined the results of the WGA with respect to other previously reported susceptibility loci, including *OCTN*[133], *DLG5*[134], *TNFSF15*[135] and *ATG16L1*[136]. None of these

obtained a similar level of significance for association in our study. Genotyping our cohorts for other SNPs at these loci that are reported in the literature to be associated with CD did not improve the signals, with the exception of rs224188 corresponding to a Thr to Ala substitution within *ATGL16L1* ($p < 2x10^{-4}$), thus providing confirmation of this novel susceptibility locus for the first time[136].

On chromosome 5p13.1, we identified a region of approximately 250 Kb that contained six markers with $p<10^{-6}$ in the association test. This region has not previously been reported as a CD susceptibility locus. We selected 10 markers from the regions of *IL23R* and 5p13.1 for confirmation genotyping in up to 1,266 additional Caucasian CD patients and 559 additional controls. The *IL23R* locus was included in the confirmation genotyping as it had not yet been reported at the time of our study[132]. The associations at these two loci were clearly replicated with p-values as low as $4.2x10^{-7}$ at the *IL23R* and $3.7x10^{-4}$ at 5p13.1 (Table IV.1). In the combined data from the WGA and replication studies, we obtained p-values as low as $2.2x10^{-18}$ at *IL23R* and $2.1x10^{-12}$ at the 5p13.1 locus. In addition, we genotyped trios with non-affected parents for the same SNPs to perform a transmission disequilibrium test (TDT). The 10 SNPs were typed on 137 trios with affected offspring included in the case-control study, while two of the 5p13.1 SNPs were typed on an additional 291 independent trios originating also from Belgium.. Significant over-transmission of the associated alleles were found at both loci, thus providing additional confirmatory evidence in support of the *IL23R1* and 5p13.1 susceptibility loci (Table IV.1).

To further characterize the novel 5p13.1 locus, we genotyped a subset of 1,092 CD patients and 374 Belgian controls for 111 markers (average interval: 2.3 Kb) spanning the 250 Kb segment. We determined the most likely linkage phase for each individual using PHASE[49], and used the corresponding haplotype frequencies to quantify the level of linkage disequilibrium (LD) between all marker pairs. The 250 Kb encompass five clearly delineated LD blocks, the central one (block III) being the largest and spanning 122 Kb (Figure IV.2.A). We first performed single-marker association analyses. The strongest effects were observed within the 122 Kb block III with several SNPs yielding p-values $< 10^{-5}$. P-values $< 10^{-3}$ and 10^{-4} were observed in flanking blocks II and IV, respectively (Figure IV.2.B). We then performed haplotype analysis of the region spanned by blocks II to IV. For block III, 20 haplotypes accounted for 93% of the observed chromosomes. These could be grouped in three clades comprising respectively six (IIIA), six (IIIB) and two (IIIC) haplotypes, plus a group of six haplotypes that apparently originated from various recombination events. Likewise, evaluation of block II revealed three clades (with respectively two (IIA), three (IIB) and two (IIC) haplotypes) and two recombinant haplotypes, while block IV was characterized by two clades with two (IVA) and one (IVB) haplotype respectively. We compared the clade frequencies in cases and controls at intervals bounded by ancestral recombination events (Figure IV.2.C). In agreement with the results of the single-marker analysis, the most significant associations were found in block III followed by IV and II. To verify whether the entire 5p13.1 effect could be attributed to block III (i.e. the

effects observed for blocks II and IV would be mere echos of the block III effect), we performed a multi-variate analysis as described in Methods. The clade effects of blocks II and IV conditional on the effect of block III and *vice versa*, remained significant ($p_{(II|III)}$=0.023; $p_{(III|II)}$=0.0004; $p_{(IV|III)}$=0.003; $p_{(III|IV)}$=0.026), suggesting that multiple variants in the region may jointly account for the observed effect on CD. Commonly occurring recombinant haplotypes in blocks II and III caused local drops in significance thus suggesting that causal variants lie outside the corresponding sub-segments (Figure.IV.2.C).

No known genes or CpG islands were found within the region of association on 5p13.1 after examination with the Ensembl and UCSC genome browsers. The region has an average G+C content of 38%, and an excess of interspersed repeats given GC content (58.36% vs 42.3%), which is mainly due to an excess of LINE1's (33.05% vs 19.6%) and LTR elements (15.36% vs 7.70%)[137]. It contains 98 Phastcons conserved elements[138]. It is part of a 1.25Mb gene desert between *DAB2* (850Kb distally from the block) and *PTGER4* (270Kb proximally from the block). Interestingly several of the genes flanking the region have been implicated in pathogenesis of CD, or are related to genes that have been implicated in the disease. These include a member of the caspase recruitment domain family (*CARD6*), three complement factors (*C6*, *C7* and *C9*), and - most notably - the prostaglandin receptor EP4 (*PTGER4*), which resides closest to the group of disease associated markers.

One hypothesis is that the disease-associated region contains *cis*-acting regulatory elements that control the expression levels of the causal gene(s) located in the vicinity, and that the causal variants modulate the activity of these elements. As a first step to test this, we studied the effect of SNPs in the disease-associated region on the expression levels of neighbouring genes. To that end we exploited a database of genome-wide gene expression (Affymetrix HG-U133 Plus 2.0 chips) measured in EBV-transformed lymphoblastoid cell lines from 378 individuals genotyped with the Illumina HumanHap300 Genotyping Beadchip (W. Cookson, unpublished data). Remarkably, seven of the 26 Illumina markers spanning 264 Kb coinciding precisely with the CD-associated region yielded p-values between $6.7x10^{-5}$ and $1x10^{-3}$ for *PTGER4* (Figure IV.2.B). Three of the markers influencing *PTGER4* expression are located in block III (rs16869977, rs10512739 and rs6880934). The first two are tagging the IIIBa sub-clade (r^2=1) (Figure IV.2.C), while the third one is in complete LD with it (D'=1). The corresponding SNPs and IIIBa haplotypes did not show evidence for association with CD. Two strongly associated SNPs (D'=0.84) located respectively in block IV (rs4495224) and V (rs7720838) were showing the most significant effect on *PTGER4* expression and were also associated with CD (Table IV.1). The rs4495224 A and rs7720838 T risk alleles were associated with increased *PTGER4* expression. Although these results must be treated as preliminary, they tend to support the hypothesis that the disease-associated polymorphisms may be related to the expression levels of one or more genes in the region.

CD is the most common form of inflammatory bowel disease (IBD), the other being ulcerative colitis (UC). We

genotyped a cohort of 246 Belgian UC patients (Caucasians) for *IL23R* (rs11209026), *ATG16L1* (rs2241880) and the novel 5p13.1 locus (rs4613763). Consistent with published results[132,136] we found a significant association for *IL23R* ($p = 1.2x10^{-3}$; OR: 2.51) but not for *ATG16L1* ($p = 0.78$). There was no effect of the novel 5p13.1 locus on UC ($p = 0.54$). While additional studies will be needed to exclude completely a role in UC, these results suggests that the principal susceptibility effects of the 5p13.1 locus are for CD. The restriction to CD risk observed for *ATG16L1* and the 5p13.1 locus is similar to that found for *CARD15* [129].

We herein describe the localisation of a novel major susceptibility locus for CD on 5p13.1 by WGA. The region of strongest association coincides with a gene desert devoid of known protein-coding genes. The observed effect may be mediated by as of yet unknown transcripts mapping within the region. As a matter of fact limited numbers of spliced and unspliced ESTs originating from the HT1080 fibrosarcoma cell line or medulla (e.g. BG182136, BG184600) map to the region. An alternative explanation, however, is that the disease-associated region contains cis-acting elements controlling the expression of more distant genes. We provide evidence in support of this hypothesis by demonstrating that genetic variants in the CD-associated region differentially regulate the expression levels of *PTGER4*, the closest known gene located at 270 Kb proximally. *PTGER4* is a strong candidate gene for CD as it is known that knock-out (KO) mice develop severe colitis upon dextran sodium sulphate treatment contrary to mice deficient in either of the seven other types of prostanoid receptors. Increased susceptibility to colitis is also observed in wild-type mice administered an EP4-selective antagonist, while EP4-selective agonist are protective[139]. We observe in particular that the CD susceptibility allele at marker rs4495224 is associated with increased *PTGER4* transcript levels in lymphoblastoid cell lines. This finding establishes a direct link between disease susceptibility and *PTGER4* expression, although the direction of the effect apparently contradicts the results in KO mice. Detailed studies of the effect of genetic variants in the disease-associated region on *PTGER4* expression in different tissues and of a possible connection between *PTGER4* levels and CD susceptibility are certainly needed and work towards that goal is in progress. The hypothesis that the 5p13.1 CD-susceptibility locus operates by modulating *PTGER4* expression levels could – at least in theory – be tested by replacing the corresponding murine sequences with the human orthologous variants and quantitatively complement the murine KO allele[140]. Our results suggest that the 5p13.1 effect on CD could result from the combined action of multiple susceptibility variants. Extensive sequencing of the most common haplotypes in the region of association is being conducted towards their identification.

IV.5.3 Methods.

IV.5.3.1 Genotyping.

Genotyping for the whole genome scan was performed on a Illumina HumanHap300 Genotyping Beadchip[130]. Genotyping of individual SNPs was performed on an ABI7900HT Sequence Detection System using TaqMan MGB probes from “Pre-designed SNP Genotyping” or “Custom TaqMan SNP Genotyping” Assays (Applied Biosystems, Foster City, CA).

IV.5.3.2 Association analyses.

Association analyses were conducted using Fisher’s exact test (whole genome scan) or chi-squared tests of independence (confirmation analysis). We applied the logistic regression method of Setakis et al.[131] to test for the possible effect of population structure on the most significant association results. The 110 control markers included in the logistic regression had 100% genotype success rate with minor allele frequency >30%, and no two markers were within 20Mb. To test for an effect of block I conditional on the effect of an adjacent block II, we compared the proportion of I haplotype clades nested within a given II clade (f.i. proportion of IA, IB and IC within IIA) between cases and controls by chi-squared. Chi-squared values (and d.f.) were summed across II clades to yield an overall (I|II) test statistic.

IV.5.3.3 Expression database.

The database genome-wide expression analysis data was provided by W. Cookson (Imperial College, London). Briefly, expression data were generated from RNA extracted from EBV-transformed cells from 378 genotyped offspring in nuclear families. Annotations for individual transcripts on the Affymetrix arrays were extracted from the Affyemtrix NetAffx database (www.affymetrix.com). Data from the gene expression experiment was normalized together using the RMA (Robust Multi-Array Average) package[141,142] to remove any technical or spurious background variation. An inverse normalization transformation step was also applied to each trait to avoid any outliers. A variance components method was used to estimate heritability of each trait using the Merlin-regress (RandomSample option)[143,144]. For *PTGER4*, we obtained a mean quantitative expression value of -0.017 and a variance of 0.722 while the heritability estimate for *PTGER4* estimated using the sibship data was 0.844. Association analysis was applied with Merlin (FASTASSOC option). We estimated an additive effect for

SNPs and tested its significance using a score test that adjusts for familiality and takes into account uncertainty in the inference of missing genotypes.

IV.5.4 Acknowledgements.

This work was supported by grants from (i) the DGTRE from the Walloon Region, (ii) from the Communauté Française de Belgique (Game and Biomod ARC), (ii) the Belgian Science Policy organisation (SSTC Genefunc and Biomagnet PAI), and (iv) the University of Liège. Edouard Louis, Sarah Hansoul, Denis Franchimont and Severine Vermeire are fellows of the Belgian FNRS and NFWO. Cynthia Sandor is a fellow of the FRIA. We are grateful to Véronique Dhennin et Stéphanie Glineur for their assistance with the collection of intestinal biopsies, to Dimitri Pirotin for his assistance with the Q-RTPCR, and to all the clinicians that have taking part to the recruitment of patients: Jean-Marc Maisin*, Vinciane Muls*, Jean Van Cauter*, Marc Van Gossum*, Philippe Closset*, Pierre Hayard* et Jean Michel Ghilain*, Paul Mainguet°, Faddy Mokaddem°, Fernand Fontaine°, Jacques Deflandre°, Hubert Demolin° (* Erasme-BBIH-IBD; ° Ulg Collaborators). Sincere thanks to W. Cookson for providing us access to the genome-wide expression data prior to publication.

Table IV.1: *Results of primary and confirmatory association analysis for the IL23R and 5p13.1 loci, as well as of TDT for 5p13.1.*

Locus	SNP	Primary data		Confirmatory data		Combined		TDT
		Controls	Cases	Controls	Cases	Controls	Cases	
IL23R	rs11465804 67475114 [§]	0.915[#] 923[&] 0.98[$]	0.971[€] 553[£] 3.2E-8[%] 3.00[*]	0.934 555 0.96	0.970 928 1.7E-5 2.30	0.922 1,478 0.99	0.970 1,481 3.5E-15 2.74	16:4° 137[@] 0.04[ϕ]
	rs11209026 (67478546) Arg381Gln	0.918 906 0.93	0.972 550 1.5E-8 3.20	0.934 550 0.64	0.972 1,255 4.2E-7 2.48	0.924 1,456 0.99	0.972 1,807 2.2E-18 2.92	17:5 135 0.045
	rs1343151 (67491717)	0.641 928 0.88	0.712 554 3.0E-4 1.38	0.655 556 0.32	0.722 1,266 2.9E-4 1.36	0.646 1,484 0.87	0.719 1,820 2.3E-9 1.40	76:39 137 0.0003
	rs10889677 (67497708)	0.291 927 0.91	0.354 550 0.002 1.33	0.31 559 0.75	0.36 1,263 0.015 1.25	0.30 1,486 0.73	0.36 1,813 2.4E-6 1.31	69:44 135 0.009
5p13.1	rs348601 (40355763)	0.589 928 0.24	0.686 552 5.1E-7 1.54	0.629 545 0.53	0.668 1,261 0.067 1.19	0.604 1,473 0.82	0.673 1,813 6.6E-7 1.36	72:64 138 0.05
	rs1002922 (40422312)	0.665 903 0.46	0.762 550 9.1E-8 1.63	697 441 0.45	0.741 1,212 0.04 1.25	0.675 1,344 0.95	0.747 1,762 1.7E-9 1.43	62:44 134 0.040
	rs4613763 (40428485)	0.120 929 0.99	0.191 553 6.1E-7 1.74	0.139 545 0.13	0.183 1,247 6.2E-3 1.38	0.127 1,474 0.37	0.185 1,800 1.2E-9 1.56	139:113 428 0.050
	rs10512734 (40429362)	0.666 929 0.30	0.762 553 9.7E-8 1.63	0.685 543 0.91	0.742 1,236 1.8E-3 1.33	0.673 1,472 0.62	0.748 1,789 9.2E-11 1.45	61:46 136 0.073
	rs1373692 (40466940)	0.585 929 0.13	0.690 554 4.1E-8 1.59	0.607 552 0.89	0.674 1,235 3.7E-4 1.35	0.593 1,481 0.43	0.679 1,789 2.1E-12 1.46	214:177 428 0.030
	rs4495224 (40513272)	0.651 926 0.60	0.746 552 2.2E-7 1.59	0.675 544 0.99	0.708 1,237 0.134 1.17	0.659 1,470 0.71	0.720 1,789 6.6E-7 1.33	66:43 137 0.013

[§]Chromosomal position on march 2006 assembly.

Controls: [#]allelic frequency of risk allele; [&]number of individuals with genotype; [$]p-value of Hardy-Weinberg proportions (Fisher's exact test).

Cases: [€] allelic frequency of risk allele; [£]number of individuals with genotype; [%]p-value of allelic association (chi-squared test); [*]Odds Ratio

Results in "Primary data" were obtained after re-genotyping of the initial samples using the Taqman assay conducted to verify the Illumina genotypes.

TDT: ° times transmitted:times non-transmitted; [@]number of genotyped trios; [ϕ]p-value of segregation distortion (one-sided chi-squared test)

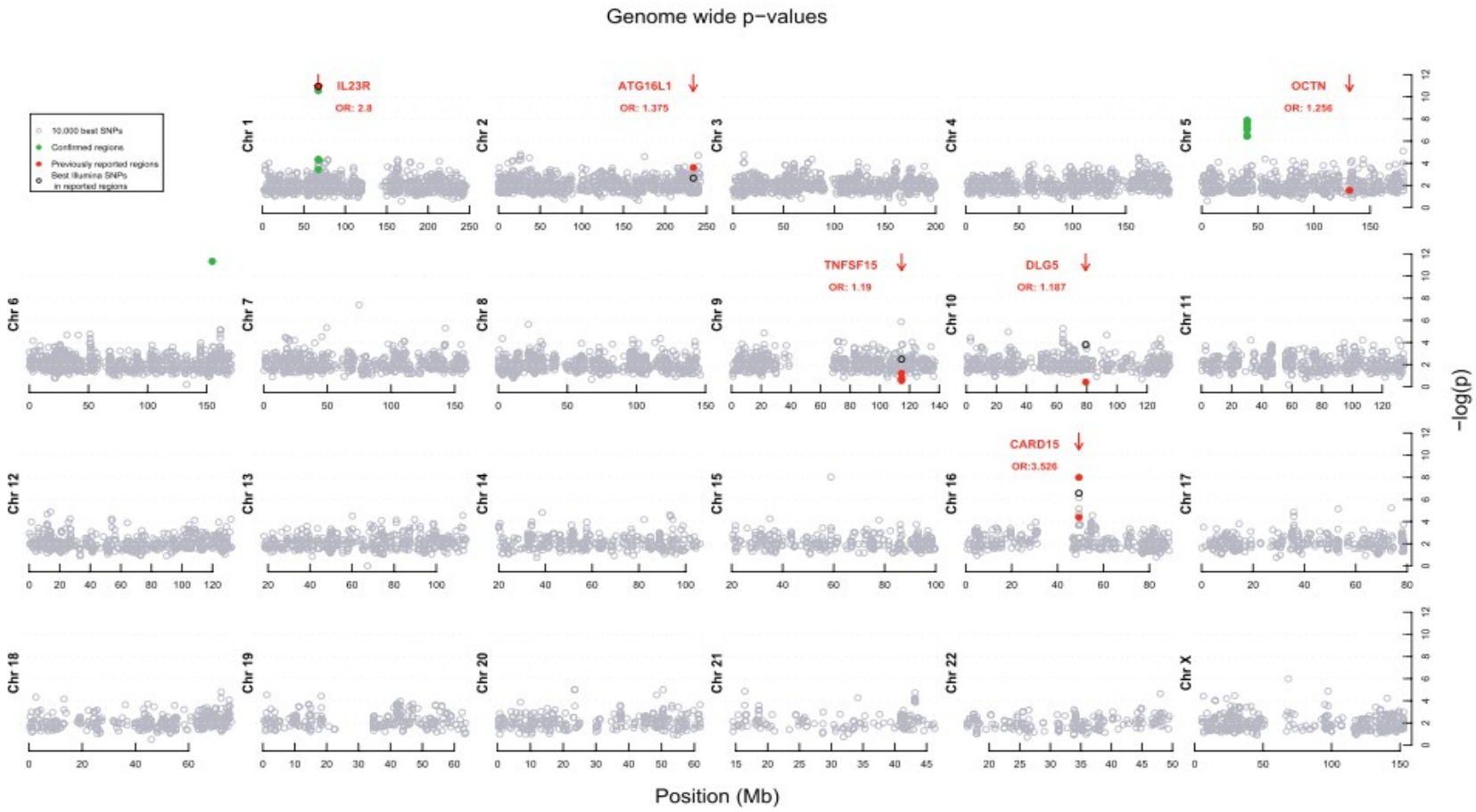


Figure IV.1: *Results of the whole genome association for CD. P-values (-log(p)) for the 10,000 best SNPs out of 311,882 are shown (gray circles). The position of previously described susceptibility loci are marked by red arrows. The p-values obtained in our cohorts with the reportedly associated SNPs/mutations are shown by the red dots, and the corresponding odds ratios (OR) indicated. The p-values obtained with SNPs included in the Illumina panel at ≤ 50 Kb from these SNPs/mutations are marked by black circles. SNPs genotyped in the confirmation cohort are shown as green dots. Two singleton SNPs, located respectively on chromosome 3 (rs11128423) and 6 (rs10485060), yielding p-values < 10^{-10} in the WGA experiment were genotyped in the replication samples but did not provide confirmatory evidence of association (data not shown).*

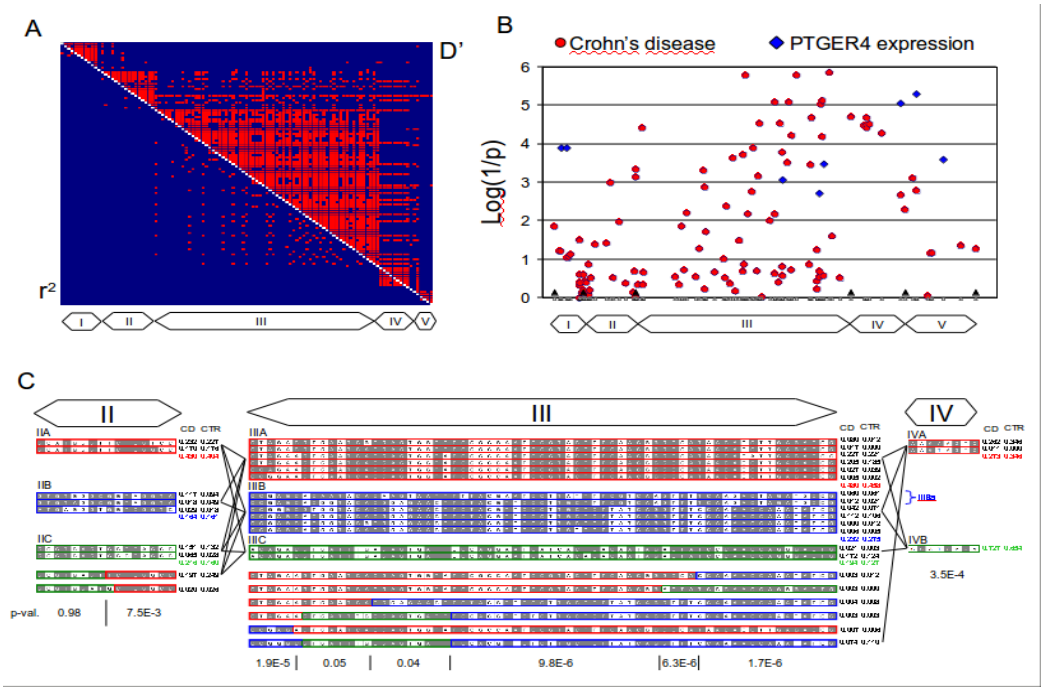


Figure IV.2: *(A) Pair-wise LD analysis between the 111 SNPs in the 250 Kb window. r2 (lower left) and D' (upper right) values were computed using standard procedures from the genotypes phased with PHASE[57]. Values > 0.93 are marked in red, values ≤ 0.93 in blue. The five LD blocks are easily identified and marked by corresponding boxes I to V. (B) Red dots: results of single-marker association analyses for CD using 111 SNPs located in a 250 Kb window spanning the positions of the most significant 5p13.1 markers in the WGA. The results are expressed as log(1/p) where p corresponds to the p-value of the association determined by chi-squared analysis. The positions of the 111 markers are indicated by the small triangles. The limits between the LD blocks (I-V) are indicated. Blue diamonds: log(1/p) values of the effect of marker genotype on PTGER4 expression levels for the 28 HumanHap300 Genotyping Beadchip SNPs mapping to the 250 Kb window. Values are only shown when exceeding 2. (C) Haplotype analysis of LD blocks II, III and IV. Haplotypes accounting jointly for > 93% of studied chromosomes are shown. The ancestral allele is in grey when known. Within each block, similar haplotypes are grouped in "clades" (e.g. IIA, IIB and IIC) marked by different colours (red, blue or green). For blocks II and III, supposedly recombinant haplotypes are represented under the major clades and coloured accordingly. The frequency of the corresponding haplotypes (black & white) and clades (coloured) in CD patients (CD) and controls (CTR) are given. p-values (chi-squared test) of the clade-based association tests for CD are given underneath for intervals bounded by recombination events. The approximate positions of within-block recombinations are marked by vertical lines between p-values. The two haplotypes forming the IIIBa sub-clade are indicated.*

IV.6 Paper II: Genome-wide association defines more than thirty distinct susceptibility loci for Crohn's disease.

Abtract:

Several new risk factors for Crohn's disease have been identified in recent genome-wide association studies. To advance gene discovery further we have combined the data from three studies (a total of 3,230 cases and 4,829 controls) and performed replication in 3,664 independent cases with a mixture of population-based and family-based controls. The results strongly confirm 11 previously reported loci and provide genome-wide significant evidence for 21 new loci, including the regions containing *STAT3*, *JAK2*, *ICOSLG*, *CDKAL1*, *and ITLN1*. The expanded molecular understanding of the basis of disease offers great promise for informed therapeutic development.

BARRET, J.C.; HANSOUL, S.. CHO, J.H.; NICOLAE, D.L.; BARMADA, M.M.; BITTON, A.; BRANT, S.R.; DASSOPOULOIS, T.; WU DATTA, L.; DUERR, R.H.; GREEN, T.; GRIFFITHS, A.M.; KISTNER, E.O.; MURTHA, M.T.; REGUEIRO, M.D.; RIOUX, J.D.; ROTTER, J.I.; SCHUMM, L.P.; SILVERBERG, M.S.. STEINHART, A.H.; TARGAN, S.R.; TAYLOR, K.D.; XAVIER, R.; THE NIDDK IBD GENETICS CONSORTIUM; LIBIOULLE, C.; SANDOR, C.; LATHROP, M.; BELAICHE, J.; DEWIT, O.; GUT, I.; HEATH, S.; LAUKENS, D.; MNI, M.; RUTGEERTS, P.; VAN GOSSUM, A.; ZELENIKA, D.; FRANCHIMONT, D.; HUGOT, J.P.; DE VOS, M.; VERMEIRE, S.; LOUIS, E.; THE BELGIAN-FRENCH IBD CONSORTIUM, THE WELLCOME TRUST CASE CONTROL CONOSRTIUM, CARDON, L.; ANDERSON, C.; DRUMMOND, H.; NIMMO, E.; AHMAD, T.; PRESCOTT, N.J.; ONNIE, C.M.; FSIHER, S.A.; GHORI, J.; BUMPSTEAD, S.; GWILLAM, R.; TREMELLING, M.; DELOUKAS, P.; MANSFIELD, J.; JEWELL, D.; SATSANGI, J.; MATHEW, C.G.; PARKES, M.; GEORGES, M.; DALY, M.J.

Nature Genetics 40:955-962 (2008)

IV.6.1 Introduction.

The first genome-wide association studies (GWAS) have identified many common variants associated with complex diseases, and have rapidly expanded our knowledge of the genetic architecture of these traits. Progress in Crohn's disease (CD), a common idiopathic inflammatory bowel disease (IBD) with high heritability (λ_s ~ 20-35), has been especially striking, with recent GWAS publications increasing the number of confirmed associated loci from two to more than ten[145]. The results have identified new pathogenic mechanisms of IBD and promise to advance fundamentally our understanding of CD biology. These recent discoveries highlight, for instance, the key importance of autophagy and innate immunity[136,146,147,148] as determinants of the dysregulated host-bacterial interactions implicated in disease pathogenesis. Furthermore, genetic associations have been shown to be shared between CD and other auto-inflammatory conditions – for example, *IL23R* variants[132] are also associated with psoriasis[149] and ankylosing spondylitis[150], and *PTPN2* variants with type 1 diabetes[146,148]). As in other complex diseases, restricted sample sizes have resulted in early CD studies focusing on only the strongest effects, which turn out to explain only a fraction of the heritability of disease.

We recently published three separate GWA scans for CD in European-derived populations – the details of which are shown in Table IV.2[147,148,124]. Motivated by the need for larger datasets to improve power to detect loci of modest effect [151], we carried out a genome-wide meta-analysis from our three CD scans. These analyses, together with a replication study in an equivalently sized, independent panel, have enabled us to identify at genome-wide levels of significance 21 novel Crohn's disease susceptibility genes and loci. This brings the total number of independent loci conclusively associated with Crohn's disease to more than 30 and provides unprecedented insight into both CD pathogenesis as well as the general genetic architecture of a multifactorial disease.

IV.6.2 Results.

IV.6.2.1 Meta-analysis of three genome-wide association scans.

The combined GWAS study samples (Table IV.2) consisted of 3,230 cases and 4,829 controls, all of European descent. While the individual scans did identify new risk factors, they were only well-powered to discover common alleles with odds-ratios (ORs) above 1.3 (in the case of the WTCCC) or 1.5 (the smaller two scans, Figure IV.3). By contrast, the combined sample has 74% power at an OR of 1.2, allowing evaluation of the role of alleles with smaller effect sizes for the first time. As two different genotyping technologies were used in the

constituent scans, we utilized recently developed imputation[152,153] methods to assess association across all three studies at 635,547 SNPs contained on one or both platforms. A quantile-quantile (Q-Q) plot of the primary meta-statistic (single SNP Z-scores, Figure IV.4) shows a striking excess of significant associations, well beyond what would be attributable to the modest overall distributional inflation (genomic control $\lambda < 1.16$). Despite the large sample size, the overall inflation is modest because (1) each group had separately tested for evidence of population stratification, and the meta-analysis used a test that combined the results from each study (rather than mixing the raw data and compromising the case-control matching of each study), and (2) imputation was done on all samples ignoring case status and thus would not introduce artifactual differences between cases and controls[129]

We focus our attention in this study specifically on the 526 SNPs from 74 distinct genomic loci which were associated with $p < 5x10^{-5}$ – more than 7 times the number of SNPs expected by chance even after correction for the modest overall inflation detected. This threshold for follow-up is not meant to imply that there are no genuine associations among SNPs with less significant association in the meta-analysis, but rather reflects a practical desire to prioritize as many true positives as possible for immediate replication. Eleven associations previously replicated and established at genome-wide significance levels (Methods, Table IV.3), including both "historical" associations at *NOD2*[154,155] and 5q31 (IBD5)[135] as well as recent replicated findings from individual GWA scans such as *IL23R*, *ATG16L1*, *IRGM*, *TNFSF15* and *PTPN2*[136,146,147,148,132,156] were among the 74 regions represented in this tail of the distribution of association statistics. Even after removing all SNPs in LD with these eleven loci, however, there continued to be a substantial excess of associated alleles beyond that which would be expected by chance (Figure IV.4).

IV.6.2.2 Replication of 21 new loci.

As these 74 regions included the 11 already reported as independently replicated and meeting genome-wide significance thresholds, this replication experiment effectively explored 63 putative associations in novel regions with 11 positive controls. To identify the true risk factors from these 63 regions, we undertook a replication study involving a total of 2,325 additional Crohn's disease cases and 1,809 controls alongside an independent family-based dataset of 1,339 parent-parent-affected offspring trios.

Results (significance levels and odds ratios) for strongly replicating loci, including all positive controls, are presented in Table IV.3. The distribution of Z-scores from the 63 putative regions shows a dramatic departure from the null distribution (Figure IV.5) with 19 novel regions showing significant replication ($p < 0.0008$ – a value of 0.05/63 representing a conservative threshold expected to be exceeded only once by chance in 20 such

replication experiments). SNPs on chromosome 19p13 (replication p = 0.00347, combined p = $1.06x10^{-9}$) and in the MHC (replication p = 0.006, combined p = $2.6x10^{-9}$ - suspected but not previously conclusively established in Crohn's disease) did not reach this conservative threshold, but so convincingly satisfy proposed thresholds for genome-wide significance ($p < 5x10^{-8}$, Methods) that we propose these as the 20th and 21st additional Crohn's disease associated loci defined here. A further 8 of the 42 remaining loci showed nominal replication (Table IV.3).

It is possible that extreme population substructure in the replication sample could give rise to such a striking excess of hits. While unlikely, this was directly evaluated by the large family-based component of the replication study. Odds ratio estimates from the TDT analysis of the North American, French and Belgian families alone are consistent with those from the UK and Belgian case/control samples (Tables IV.3 & IV.4), with all 21 newly defined loci showing odds ratios in the same direction of association with the original scan in the family-based component (and nearly half showing greater OR than in the case-control arm). Importantly, none of the significantly or nominally replicating loci show significant evidence for heterogeneity (across studies or between family-based and population-based arms) when corrected for the number of tests performed. This independent family based evidence confirms these alleles constitute true Crohn's disease loci.

For this newly expanded set of 32 unequivocally associated loci, we assessed whether there was evidence of significant pairwise interactions which could add further to the overall variance in liability explained by this set of loci. We performed a case-only analysis of the 3,664 cases in the replication study and observed no interactions that withstood a correction for the number of tests performed.

IV.6.2.3 Deciphering the genetic architecture of CD.

The contributions of the 32 loci to disease risk were computed using a standard liability threshold model[133] and are displayed as a histogram of individual variances (Figure IV.6). The observations from this variance analysis that many loci were detected for which the current study had low power, and that only a minority of the variance in risk is explained by these 32 loci, suggest that many additional loci are yet to be identified. This is reinforced by the additional 8 nominal replications (Table IV.4) where only 2 or 3 would be expected by chance, and by the continued excess of small p values when these 40 total regions are removed (Figure IV.3).

While recognizing that fine-mapping is required to identify specific causal variants, we performed a series of analyses to gain some general insight into the CD associations. We first queried HapMap to discover any instances where a non-synonymous SNP (nsSNP) was correlated ($r^2 > 0.5$) to the most associated variant discovered in this study. Accepting that HapMap is not a complete catalogue of nsSNPs, but including four loci

where fine-mapping has identified coding variants, just 9 of the 32 genomewide significant associations were correlated with a known nsSNP. To explore whether any of the associations reflect a cis-acting regulatory effect on a nearby gene, we evaluated genotype-expression correlation using the panel of 400 lymphoblastoid cell lines described by Dixon et al.[154]. From all genes within 250 kb of the LD-based intervals defined in Table IV.3 and IV.4, five correlations between expression of a nearby gene and a CD-associated variant were identified (LOD > 2). This was far in excess of chance ($p \sim 0.001$) and suggests that regulatory variation also contributes to the genetic architecture identified.

IV.6.3 Discussion.

Genome-wide association studies provide a systematic assessment of the contribution of common variation to disease pathogenesis. A limiting factor is often the size of the case-control dataset, and hence the power to detect any but the most strongly associated loci. Meta-analysis of existing data provides an obvious potential solution. As Figure IV.3 demonstrates, our expectation was that the additional power of the combined dataset would result in the identification of a substantially larger number of readily replicating associations than were derived from any of the smaller, constituent datasets. However, the paradigm of exploring common genetic variation with similar effects across studies (in this case all of European descent) needs testing before its results can be accepted as valid.

On the validity of the method our results are substantially reassuring. All 11 previously confirmed CD susceptibility loci were strongly replicated both in the meta-analysis and follow-up experiment. These include the two widely replicated findings from studies published in 2001[151,152,132] as well as all of the compelling findings from individual GWAS (Table IV.3). Significantly, we have also identified and replicated 21 new CD susceptibility loci. Using a conservative threshold for significance (only 1 such region would be expected by chance in 20 such experiments), the loci with clear evidence for association in the replication panel include a very high proportion of those showing strongest signals in the meta-analysis – 9 of 9 previously unreported regions with $p < 5x10^{-7}$ in the combined scan were replicated convincingly - emphasizing the validity of the meta-analysis results. Further emphasizing the robustness of these results, all 21 of these loci exceed a conservative genome-wide level of significance ($p < 5x10^{-8}$) by a significant margin (all have $p < 5x10^{-9}$) - and equivalent strength of association was observed in the family-based subset of our replication sample.

In keeping with other regions recently identified as associated with CD, the 21 new loci do not conform to any obvious pattern in terms of gene content. Thus, as shown in Table IV.3, some loci (defined by HapMap recombination hotspots flanking the set of correlated, associated variants) contain just a single gene, some

contain many genes and others none. Clearly the first category provides the most immediate clues regarding pathogenic mechanisms. These genes are discussed briefly in Box 1 (see paper), together with a number of genes which constitute striking candidates from regions with only a handful of transcripts. Included among these are compelling functional candidates such as *STAT3*, *JAK2* and *IL12B* while others, such as *CDKAL1* and *PTPN22*, highlight potentially intriguing contrasts between genetic susceptibility to Crohn's disease and some other complex disorders. It is noteworthy – and consistent with previous findings from CD and other complex diseases – that we did not find any strong evidence of deviation from the model of multiplicative (random) effects when we tested for gene-gene interactions among the 32 confirmed associations. This is in spite of the fact that some of these genes seem to affect the same or overlapping pathways.

For loci containing multiple genes or no genes the picture is less well defined. The identified paucity of correlation between associated SNPs and coding variation suggests that these loci may, in particular, benefit from eQTL (expression quantitative trait locus) analysis. This seeks correlation between genotype and expression patterns – bearing in mind that such functional relationships need not respect the specific boundaries of LD around the association. One of our groups previously reported an eQTL effect incriminating *PTGER4* at the 5p13 locus[122]. A striking outcome from our present analysis was at the established IBD5 locus[132], where CD-associated SNPs were associated with decreased *SLC22A5* mRNA expression levels. While a SNP had previously been proposed as regulating *SLC22A5* transcriptional activity[155], these data suggest for the first time that the most disease-associated variants in the IBD5 region, including a coding variant in neighboring *SLC22A4*, are the same variants most associated with *SLC22A5* expression. Equally striking, the most significant Crohn's disease associated eQTL reported here affects *ORMDL3* (LOD = 20) on chromosome 17 and SNPs in precisely the same region were recently shown to be strongly associated with childhood asthma[156] This suggests that the same polymorphisms might underlie susceptibility to both CD and asthma, possibly by perturbing *ORMDL3* expression.

The new loci that we have identified are of modest effect size, which is unsurprising given all loci with larger impact on disease risk were – as might be expected – discovered in the original scans. The small sizes of these effects explains the lack of overlap between linkage results in CD and these newly discovered loci, with the possible exceptions of combined effects of multiple high ranking associations on chromosomes 5q and 6p. Indeed, the linkage evidence that led to the discovery of the IBD5 locus was very likely boosted by the nearby effects at *IL12B* and *IRGM*. As expected, the only gene conclusively discovered via linkage (*NOD2*) is one of two loci which stand well out from the remainder of the distribution of effect sizes (Figure IV.6). The other outlier, *IL23R*, illustrates an interesting characteristic of linkage – because (unlike *NOD2*) the most penetrant risk allele has very high frequency (93%), it is nearly invisible to linkage analysis despite the high OR; highly

protective rare alleles are simply not present in multiplex affected families and thus do not influence allele sharing substantially.

Using a liability-threshold model, we estimate that the 32 loci identified to date explain about 10% of the overall variance in disease risk, which may be as much as a fifth of the genetic risk, given previous estimates of CD heritability of approximately 50%[157] This observation is consistent with the fact that these loci collectively contribute only a factor of two to sibling relative risk (λ_s), and even this figure is dominated by the substantial contribution of NOD2 variants. However, it should be emphasized that the full impact of the new loci cannot be determined until causal variants have been identified by directed sequencing and fine-mapping experiments. Until then the proportion of the variance in Crohn's disease risk explained must be measured from the confirmed SNPs, where association is due to LD with causal variants. Since multiple causal variants might exist at each locus (ranging in frequency from rare to common) our estimates of variance explained provide only a lower bound for the true contribution of each locus.

In conjunction with results from very similar gene discovery efforts in cholesterol levels[158] and type 2 diabetes[25], common lessons are beginning to emerge with respect to the genetic architecture of complex traits. In each example, substantial increase in sample size achieved through meta-analysis has led to dramatic success in gene discovery. In all cases, this progress has revealed an underlying architecture consistent with many individually modest effects which conventional genetic linkage analysis, and even the largest individual genome-wide association studies, are not well powered to detect. Common variants explaining more than 1% of the genetic variance are rare, whereas well-powered studies have found dozens of variants contributing 0.1% of overall variance in liability. Perhaps surprisingly, neither we nor others have yet to document a substantial role for epistasis among these loci and a number of associated loci are conclusively mapped to regions with no currently annotated protein coding genes. Despite the considerable concordant success, a distinct minority of the overall heritability has been explained by these documented associations.

Since our study is well-powered to identify loci that explain > 0.2% of the overall variance, but the sum of such loci explains a relatively small fraction of the total, it seems likely that many loci with even more modest effect sizes remain undiscovered. Of particular note is the continued excess of associations outside of the regions studied here, as well as the nominal replication of an additional 8 loci, notably greater than expected by chance. Overall, the distribution of Z scores in the replication experiment is clearly skewed towards replication – only 11 of the 63 Z-scores in this replication experiment generate $Z < 0$. If only the 21 strongly confirmed loci were genuinely associated, half of the 42 remaining should end up with $Z < 0$. Indeed, observing 8 of the 42 remaining tests with $Z > 1.5$ is itself a highly significant observation ($p < 0.0001$). Although modest in terms of effect size, identification of such loci is likely to still provide important insights into pathogenic mechanisms, as biological

importance need not be proportional to the statistical evidence for genetic association. Closer inspection of regions showing nominal association in the replication experiment reveals that a number of transcripts in these loci are of considerable interest, including *CCL2*/*CCL7*, *IL18RAP* and *GCKR*.

It is important to note that the generation of GWAS arrays used in the scans here did not offer complete genome coverage of common variation (additional loci may reside in poorly covered intervals) and did not address either rare SNPs or copy number variation effectively. Thus in spite of the wealth of new susceptibility genes and loci identified by the current study, it seems implausible that there are not more to be found – albeit very large datasets are likely to be required to achieve robust statistical support for them. With respect to the present findings, there is much work to be done in resequencing and fine mapping to identify causal variants. While we do not yet have a complete understanding of the genetic architecture of Crohn's disease, dramatic progress has now been made towards this goal - and with it the prospect of directed functional exploration of the pathways identified, insight into how risk alleles interact with environmental modifiers, and the hope of new avenues for treatment.

IV.6.4 Methods.

IV.6.4.1 Crohn's disease patients, controls, and GWAS.

The meta-analysis was based on data from the 3 genome-wide scans of the NIDDK[147], WTCCC[148] and Belgian/French[124] studies. Details of the numbers of cases and controls genotyped in the respective scans and of the genotyping platforms used are shown in Table IV.2, as are case/control and family cohorts genotyped in the replication study of the meta-analysis. Details of the ascertainment and characterization of these cohorts, as well as quality control procedures applied to the GWA datasets, were provided in the original scan and replication publications[146,147,148,132,124]. Recruitment of study subjects was approved by local and national institutional review boards, and informed consent was obtained from all participants.

IV.6.4.2 Imputation.

Briefly, these methods rely on observed haplotype patterns in a set of reference data (the HapMap) and the actual genotype data from each project to make predictions (along with a measure of statistical certainty) at un-genotyped SNPs. We used the program MACH[151] with the NIDDK and Belgian/French data, and IMPUTE[152] with the WTCCC data. Comparisons between the two algorithms yielded very similar results (data

not shown). We imputed the superset of polymorphic markers which passed QC in the original scans. This set was comprised of SNPs on either the Affymetrix 500K only (n = 350,507), Illumina HumanHap300 version 1 only (n = 238,935), or both panels (n = 46,105) such that all association tests performed were at least partially based on observed genotype data.

IV.6.4.3 Test for association, effect size estimation and interactions.

Using the genotype probabilities (rather than best-guess genotypes) and empirical variances for imputed markers in the case and control tallies, we summarized the standard 1 d.f. allele-based test of association as a Z-score within each scan and combined scores across studies to produce a single meta-statistic for each SNP across all three datasets. Odds ratios were estimated separately in TDT samples and each case/control replication collection, and then combined and tested for heterogeneity[162]. Interaction tests were performed using the case-only epistasis test implemented in PLINK[163].

IV.6.4.4 Critical regions.

Given that most associations contain many correlated SNPs showing signal, we demarcated independent loci by first defining the set of HapMap SNPs with $r^2 > 0.5$ to the most significantly associated SNP. We then bounded the “critical region” by the flanking HapMap recombination hotspots which contained this set. These windows very likely contain the causal polymorphisms explaining the associations.

IV.6.4.5 Replication

We defined loci to have been previously confirmed if an earlier study had both detected and replicated the association in independent samples and the association achieved $p < 5 \times 10^{-8}$ (recently proposed as an appropriate genome-wide significance level for GWAS[164]). For replication genotyping, we selected the most significantly associated SNP from each region along with a second, correlated SNP with $p < 0.0001$ or a second assay on the opposite strand in order to have a technical backup should the first fail genotyping .

Replication genotyping for the putatively associated loci was performed using primer extension chemistry and mass spectrometric analysis (iPLEX, Sequenom) using Sequenom Genetics Services (N. American panel) and Genome Research Limited, Wellcome Trust Sanger Institute (UK panel), and using a custom-made Golden Gate assay on a Beadstation500 (Illumina), following the manufacturer's recommendations (Belgian/French panel). The more completely genotyped SNP of the two from each region was chosen to represent that regional

association in analysis (if both were completely typed, the SNP that was more strongly associated in the scan was used). Samples with >10% missing data (n = 267 for Belgian/French data, 111 for the UK data and 8 for the N. American data; these samples are not included in the tallies for Table IV.2), as well as SNPs with >10% missing data or Hardy-Weinberg p value < 0.001 were excluded from this analysis.

IV.6.4.6 Regional Annotation: eQTL analysis.

Effects of SNPs in Tables IV.3 & IV.4 on expression levels of neighbouring genes was studied using transcriptome data from the ~400 lymphoblastoid cell lines described by Dixon et al.[156]. SNPs that were not genotyped on this panel (n=14) were replaced with a proxy with $r^2 > 0.95$ when possible (n=12). LOD scores > 2 for genes (probe average) located within 250 Kb of the corresponding LD windows were retrieved from http://www.sph.umich.edu/csg/liang/asthma/. To evaluate the significance of the findings with the CD associated SNPs, we compared the observed (i) number of genes yielding LOD scores > 2, and (ii) sum of these LOD scores, with the corresponding frequency distributions for 1,000 randomly selected sets of 31 SNPs, matched for allele frequency (± 0.02) and gene context. Window sizes determined for associated SNPs were used for the matched simulated SNPs.

IV.6.5 Acknowlegments.

We acknowledge use of DNA from the 1958 British Birth Cohort collection (R.Jones, S. Ring, W. McArdle and M. Pembrey), funded by the Medical Research Council (grant G0000934) and The Wellcome Trust (grant 068545/Z/02) and the UK Blood Services Collection of Common Controls (W. Ouwehand) funded by the Wellcome Trust. We also acknowledge the National Association for Colitis and Crohn's disease and the Wellcome Trust for supporting the case DNA collections, and support from UCB Pharma (unrestricted educational grant) and the NIHR Cambridge Biomedical Research Centre. The National Institute of Diabetes and Digestive and Kidney Disease (NIDDK) IBD Genetics Consortium is funded by the following grants: DK62431 (S.R.B.), DK62422 (J.H.C.), DK62420 (R.H.D.), DK62432 and DK064869 (J.D.R.), DK62423 (M.S.S.), DK62413 (K.D.T.), and DK62429 (J.H.C.). Additional support was provided by the Burroughs Wellcome Foundation (J.H.C.), the Crohn's and Colitis Foundation of America (S.R.B., J.H.C.). We thank Peter Gregersen and Annette Lee (Feinstein Medical Research Institute) for their efforts and the use of control samples. This work was supported by grants from (i) the DGTRE from the Walloon Region (n°315422 and CIBLES), (ii) from the Communauté Française de Belgique (Biomod ARC), and (iii) the Belgian Science Policy organisation (SSTC

Genefunc and Biomagnet PAI). Edouard Louis, Sarah Hansoul, Denis Franchimont and Severine Vermeire are fellows of the Belgian FNRS and NFWO. Cynthia Sandor is a fellow of the FRIA. We are grateful to all the clinicians, consultants and nursing staff who recruited patients, including: Jean-Marc Maisin*, Vinciane Muls*, Jean Van Cauter*, Marc Van Gossum*, Philippe Closset*, Pierre Hayard* and Jean Michel Ghilain*; Paul Mainguet°, Faddy Mokaddem°, Fernand Fontaine°, Jacques Deflandre°, and Hubert Demolin°; Jean-Frédéric Colombel[#], Marc Lemann[#], Sven Almer[#], Curt Tysk[#], Yigael Finkel[#], Miquel Gassul[#], Colm O'Morain[#], Vibeke Binder[#] and Jean-Pierre Cézard[#] (*Erasme-BBIH-IBD; ° Ulg Collaborators; [#]INSERM collaborators). Sincere thanks to L. Liang for his assistance in accessing the eQTL database, and to Françoise Merlin for expert technical assistance. Finally, we thank all subjects who contributed samples.

Table IV.2: *Samples used (post QC) in this study*

	NIDDK	BEL/FR	UKIBDGC	Total
Scan cases	946	536	1,748	3,230
Scan controls	977	914	2,938	4,829
Replication cases	0	1,082	1,243	2,325
Replication controls	0	787	1,022	1,809
Replication Trios	720	619	0	1,339
Nationality	USA/Canadian	Belgian/French	British	
Scan Platform	Illumina HumanHap300	Illumina HumanHap300	Affymetrix GeneChip 500K	
Replication Platform	Sequenom	Illumina GoldenGate	Sequenom	

Table IV.3: *Convincingly (Bonferroni p < 0.05) replicated CD risk loci*

			p values							Odds ratios	
SNP	**Chr**	**Critical region**	**Scan**	**Replication**	**Combined**	**Num. genes**	**Gene of interest**	**RAF**	**Risk allele**	**Case Ctrl**	**TDT**
(a) Previously published loci											
rs11465804	1p31	67.4*	1.01×10^{-35}	3.1×10^{-29}	3.33×10^{-63}	NA	*IL23R*	0.933	T	2.50	2.77
rs3828309	2q37	230.9*	1.13×10^{-20}	7.67×10^{-14}	1.18×10^{-32}	NA	*ATG16L1*	0.533	G	1.28	1.30
rs3197999	3p21	48.73 - 49.87	2.16×10^{-7}	5.64×10^{-7}	5.76×10^{-13}	35	*MST1*	0.271	A	1.20	1.20
rs4613763	5p13	40.32 - 40.48	4.52×10^{-22}	2.79×10^{-8}	3.41×10^{-27}	0	*PTGER4***	0.125	C	1.32	1.28
rs2188962	5q31	131.44 - 131.90	4.58×10^{-9}	3.52×10^{-11}	1.16×10^{-18}	7		0.425	T	1.25	1.26
rs11747270	5q33	150.15 - 150.32	6.36×10^{-11}	2.57×10^{-7}	1.70×10^{-16}	3	*IRGM*	0.090	G	1.33	1.31
rs4263839	9q32	114.61 - 114.78	3.92×10^{-7}	6.58×10^{-5}	1.30×10^{-10}	2	*TNFSF15*	0.677	G	1.22	1.07
rs10995271	10q21	64.05 - 64.12	1.90×10^{-11}	1.61×10^{-10}	2.23×10^{-20}	1	*ZNF365*	0.387	C	1.25	1.53
rs11190140	10q24	101.26 - 101.32	1.71×10^{-10}	1.69×10^{-7}	1.53×10^{-16}	1	*NKX2-3*	0.478	T	1.20	1.28
rs2066847	16q12	49.3*	NA	1.49×10^{-24}	1.49×10^{-24}	NA	*NOD2*	0.018	C	3.99	2.57
rs2542151	18p11	12.73 - 12.88	1.19×10^{-11}	2.41×10^{-7}	2.55×10^{-17}	1	*PTPN2*	0.152	G	1.35	1.14
(b) Novel loci											
rs2476601	1p13	113.79 - 114.17	1.81×10^{-5}	0.000101	7.30×10^{-9}	7	*PTPN22*	0.899	G	1.31	1.17
rs2274910	1q23	157.65 - 157.72	3.50×10^{-7}	0.000481	7.30×10^{-10}	2	*ITLN1*	0.682	C	1.14	1.62
rs9286879	1q24	169.54 - 169.67	4.02×10^{-7}	0.000321	7.66×10^{-10}	0		0.243	G	1.19	1.08
rs11584383	1q32	197.60 - 197.77	6.82×10^{-7}	2.34×10^{-6}	7.17×10^{-12}	3		0.697	T	1.18	1.20
rs10045431	5q33	158.69 - 158.76	8.80×10^{-9}	3.66×10^{-6}	1.93×10^{-13}	1	*IL12B*	0.708	C	1.11	1.36
rs6908425	6p22	20.63 - 20.84	2.52×10^{-7}	0.000278	4.48×10^{-10}	1	*CDKAL1*	0.780	C	1.21	1.09
rs7746082	6q21	106.52 - 106.62	3.70×10^{-6}	7.7×10^{-6}	1.22×10^{-10}	0		0.289	C	1.17	1.19
rs2301436	6q27	167.32 - 167.52	3.30×10^{-7}	3.26×10^{-7}	5.22×10^{-13}	3	*CCR6*	0.463	T	1.21	1.16
rs1456893	7p12	50.03 - 50.11	4.92×10^{-5}	1.1×10^{-5}	2.30×10^{-9}	0		0.678	A	1.20	1.14
rs1551398	8q24	126.60 - 126.62	4.90×10^{-6}	0.000109	2.25×10^{-9}	0		0.619	A	1.08	1.25
rs10758669	9p24	4.94 - 5.26	6.80×10^{-7}	0.00043	1.73×10^{-9}	3	*JAK2*	0.348	C	1.12	1.21
rs17582416	10p11	35.30 - 35.60	8.48×10^{-6}	2.53×10^{-5}	8.93×10^{-10}	3		0.345	G	1.16	1.26
rs7927894	11q13	75.80 - 76.02	1.43×10^{-7}	0.000732	6.60×10^{-10}	1	*C11orf30*	0.386	T	1.16	1.07
rs11175593	12q12	38.61 - 39.31	1.33×10^{-7}	0.000165	1.54×10^{-10}	3	*LRRK2,MUC19*	0.017	T	1.54	1.44
rs3764147	13q14	43.13 - 43.54	1.61×10^{-7}	1.33×10^{-7}	1.04×10^{-13}	3		0.221	G	1.25	1.19
rs2872507	17q21	34.63 - 35.34	2.12×10^{-6}	0.000292	2.50×10^{-9}	17	*ORMDL3*	0.473	A	1.12	1.24
rs744166	17q21	37.74 - 37.95	5.94×10^{-6}	9.15×10^{-8}	3.41×10^{-12}	4	*STAT3*	0.565	A	1.18	1.25
rs1736135	21q21	15.73 - 15.76	2.06×10^{-5}	4.58×10^{-5}	3.70×10^{-9}	0		0.565	T	1.18	1.10
rs762421	21q22	44.43 - 44.48	1.08×10^{-5}	1.59×10^{-5}	7.04×10^{-10}	1	*ICOSLG*	0.389	G	1.13	01/01/21

RAF is risk allele frequency in control samples. Critical region is in NCBI B35 coordinates, with definition as described in Methods. Risk alleles are defined relative to the + strand of the reference. * regions where causal variants have been convincingly mapped, rendering the LD window uninformative. **PTGER4 is outside the critical region, but was implicated via eQTL analysis.

Table IV.4: *Nominally (p < 0.05) replicated CD risk loci*

SNP	Chr	Critical region	*p* values			Num. genes	Gene of interest	RAF	Risk allele	Odds ratios	
			Scan	Replication	Combined					CaseCtrl	TDT
rs4807569	19p13	1.05 -1.15	$1.16x10^{-8}$	0.00347	$1.06x10^{-9}$	0		0.217	C	1.02	1.26
rs780094	2p23	27.30 - 27.77	$3.82x10^{-6}$	0.00381	$1.57x10^{-7}$	22	*GCKR*	0.397	T	1.08	1.13
rs3763313	6p21	32.44-32.79 *	$1.45x10^{-8}$	0.00602	$2.60x10^{-9}$	7	*BTNL2, DRA, DRB, DQA*	0.188	C	1.19	1.01
rs13003464	2p16	61.09 - 61.14	$3.44x10^{-5}$	0.00565	$2.30x10^{-6}$	1	*CCDC139*	0.376	G	1.16	1.08
rs991804	17q12	29.57 - 29.70	$4.02x10^{-6}$	0.0135	$5.34x10^{-7}$	4	*CCL2, CCL7*	0.726	C	1.1	1.08
rs12529198	6p25	5.04 - 5.11	$7.08x10^{-7}$	0.0192	$3.48x10^{-7}$	1	*LYRM4*	0.062	G	1.12	1.19
rs17309827	6p25	3.36 - 3.42	$2.08x10^{-6}$	0.0391	$1.37x10^{-6}$	1	*SLC22A23*	0.639	T	1.1	1.02
rs7758080	6q25	149.54 - 149.65	$7.28x10^{-6}$	0.044	$4.39x10^{-6}$	0		0.274	G	1.12	0.99
rs8098673	18q11	17.74 - 17.93	$3.18x10^{-5}$	0.0443	$1.44x10^{-5}$	0		0.329	C	1.05	1.09
rs917997	2q11	102.31 - 102.64	$2.16x10^{-5}$	0.0493	$1.11x10^{-5}$	5	*IL18RAP*	0.222	T	1.05	1.11

RAF is risk allele frequency in control samples. Critical region is in NCBI B35 coordinates, with definition as described in Methods. Risk alleles are defined relative to the + strand of the reference. * SNPs with p<0.0001 were observed throughout the MHC from 30.2 – 32.9 Mb but only this largest signal from the region was followed up. More detailed study of the MHC will be required to identify and localize potentially independent signals from this region.

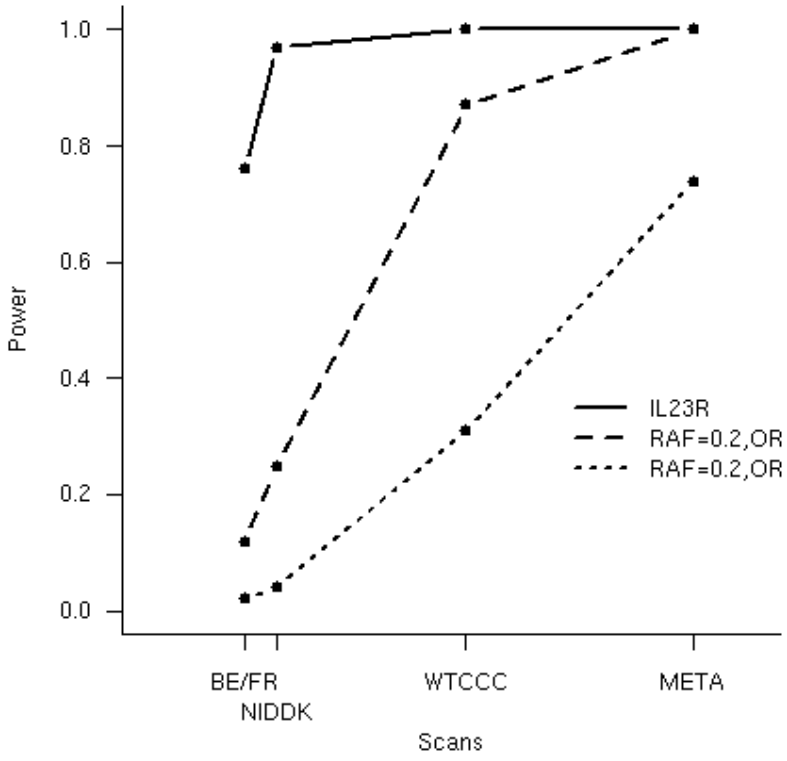


Figure IV.3: *Power to detect a genetic effect of various sizes (odds ratio 1.2, 1.3, 1.5) versus study sample size. Power is reported here as the probability (given a multiplicative model and risk allele frequency of 20%) of p < 5x10-5 in a scan – the value used to define regions for attempting replication in a larger sample set. Vertical dotted lines show the sample sizes for the three constituent scans and the meta-analysis. Relatively large effects are likely to be detected by any of these scans, whereas only the combined analysis is well powered to detect more*

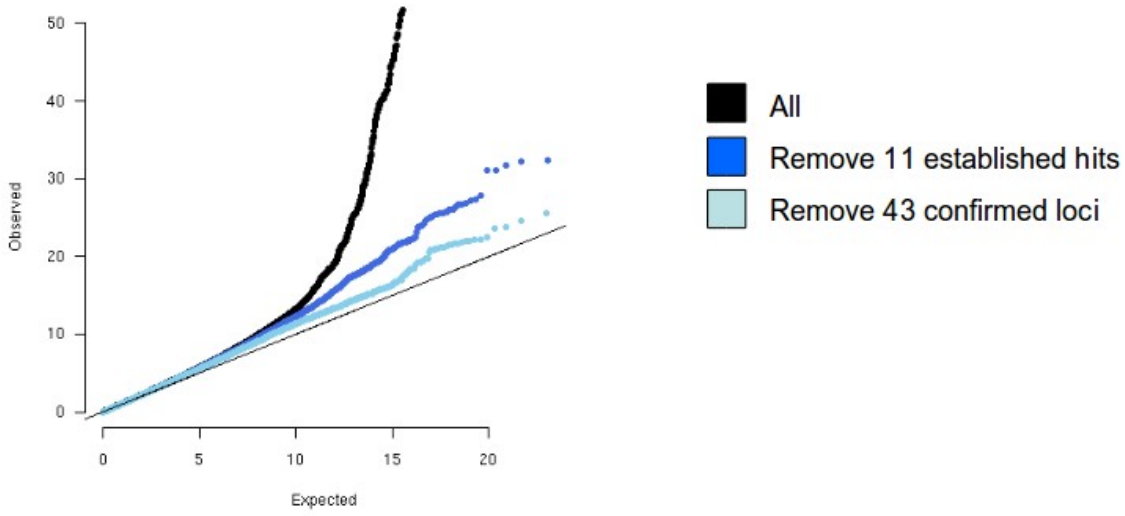


Figure IV.4: *A quantile-quantile plot of observed -log10 p values versus the expectation under the null. Black points represent the complete meta-analysis, with a substantial departure from the null at the tail (values > 8 are represented along the top of the plot as triangles). Dark blue points show the distribution after removing 11 previously published loci, demonstrating a still notable excess. Light blue points show the distribution after removing all 40 loci which replicate at least nominally. In all the cases the overall distribution is marginally inflated (λGC < 1.16).*

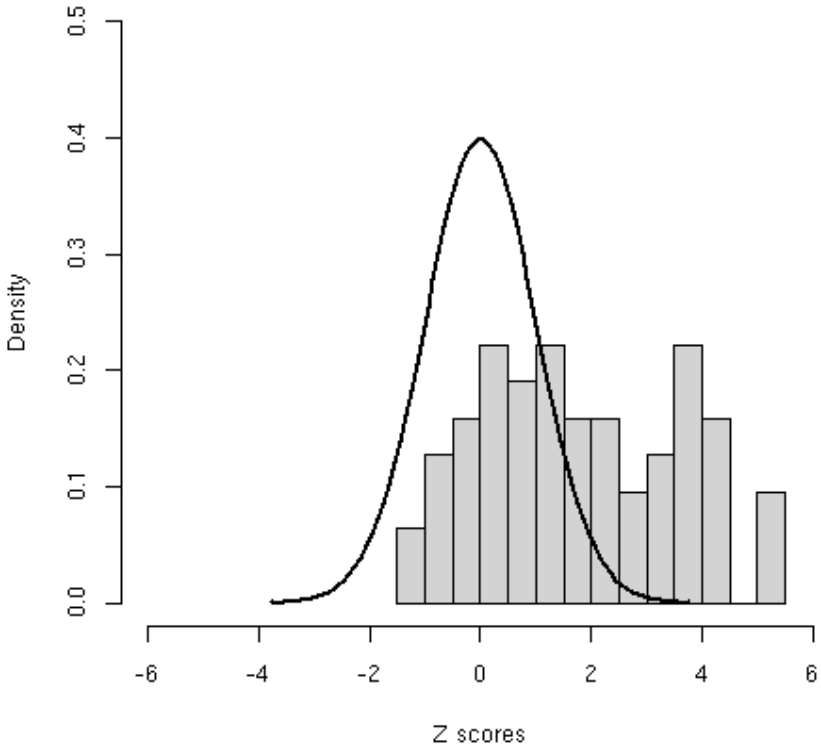


Figure IV.5: *Distribution of observed Z scores from the 63 novel regions explored, along with the expected distribution under the null (a standard normal with mean 0 and variance 1). Even setting aside the 21 regions reaching genome-wide significance, the distribution is highly skewed – 4 more results exceed a Z of 2 (1 would be expected by chance under the null) whilst none showed a Z of less than -2 (same expectation under the null) suggesting that even more of the regions investigated here are likely to constitute true positive associations when additional data become available.*

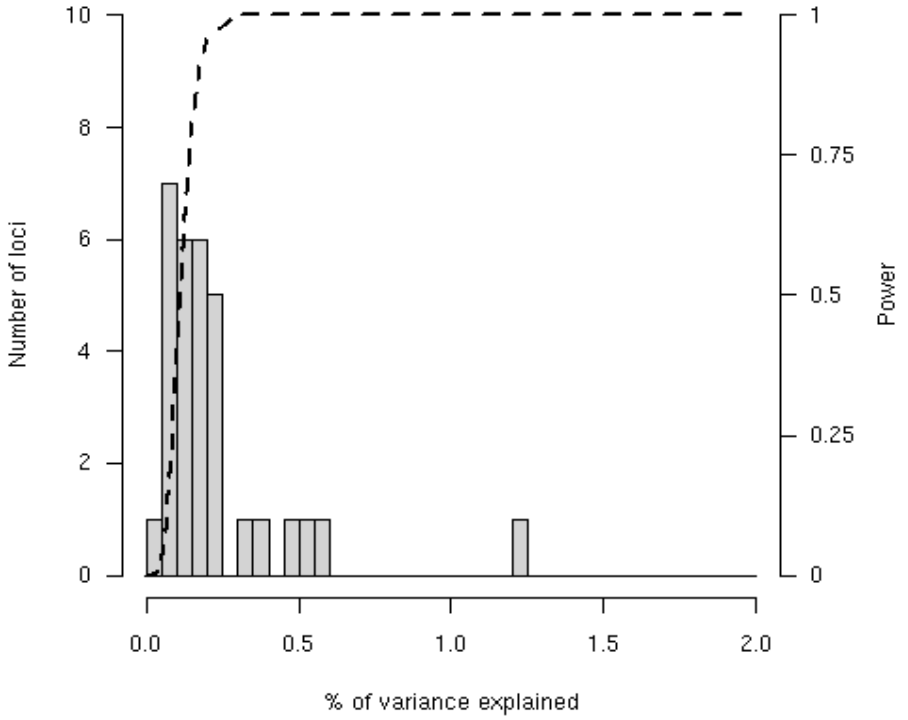


Figure IV.6: *Histogram of percent variance explained by each of the 32 established CD risk loci. The distribution resembles the long postulated exponential distribution of effect sizes. Dashed line shows the joint power for our meta-analysis to detect ($p < 5x10-5$), and for our replication sample to replicate (at Bonferroni corrected p values), a 20% variant explaining a given fraction of variance. Note how quickly this curve moves from nearly zero power to detect tiny effects (less than one tenth of one percent) to nearly full power to detect larger effects (presuming they are well covered by the current generation of GWAS chips). Complete power near the origin would likely reveal a more complete exponential distribution, with many very small effects. These are likely to increase somewhat once the causal variant or variants are identified in each locus. Indeed, NOD2 and IL23R are distant outliers, each explaining 1-2% of total variance, partially because multiple causal variants have already been discovered at these loci6,14.*

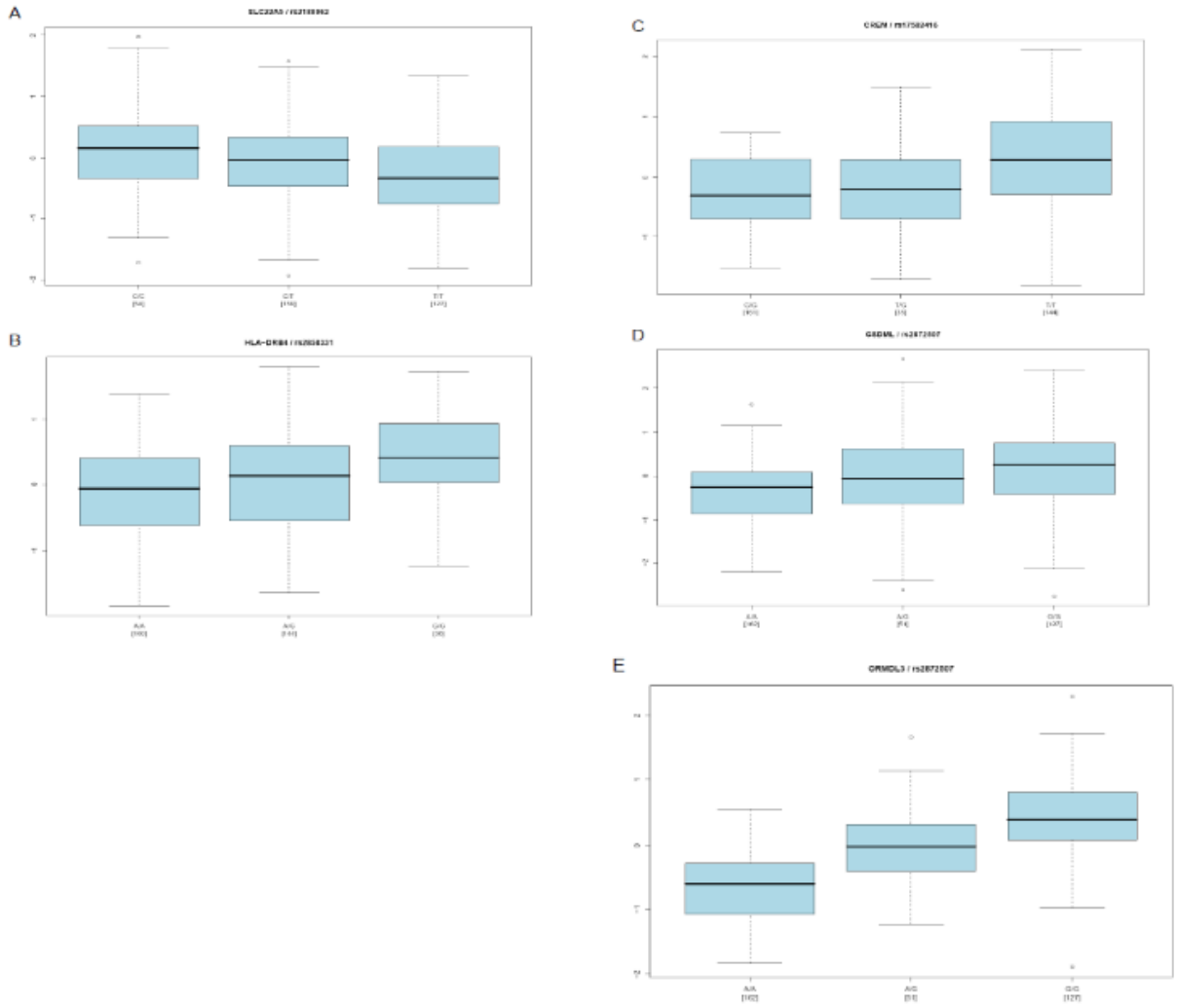


Figure IV.7: *(A-E) Boxplots of gene (probe average) expression levels bygenotype for the eQTL effects reported in Table 4. Respective gene and SNP identifiersare given for each graph. Numbers of individuals of a given genotype are given inparentheses. Data are from Dixon et al. (2007).*

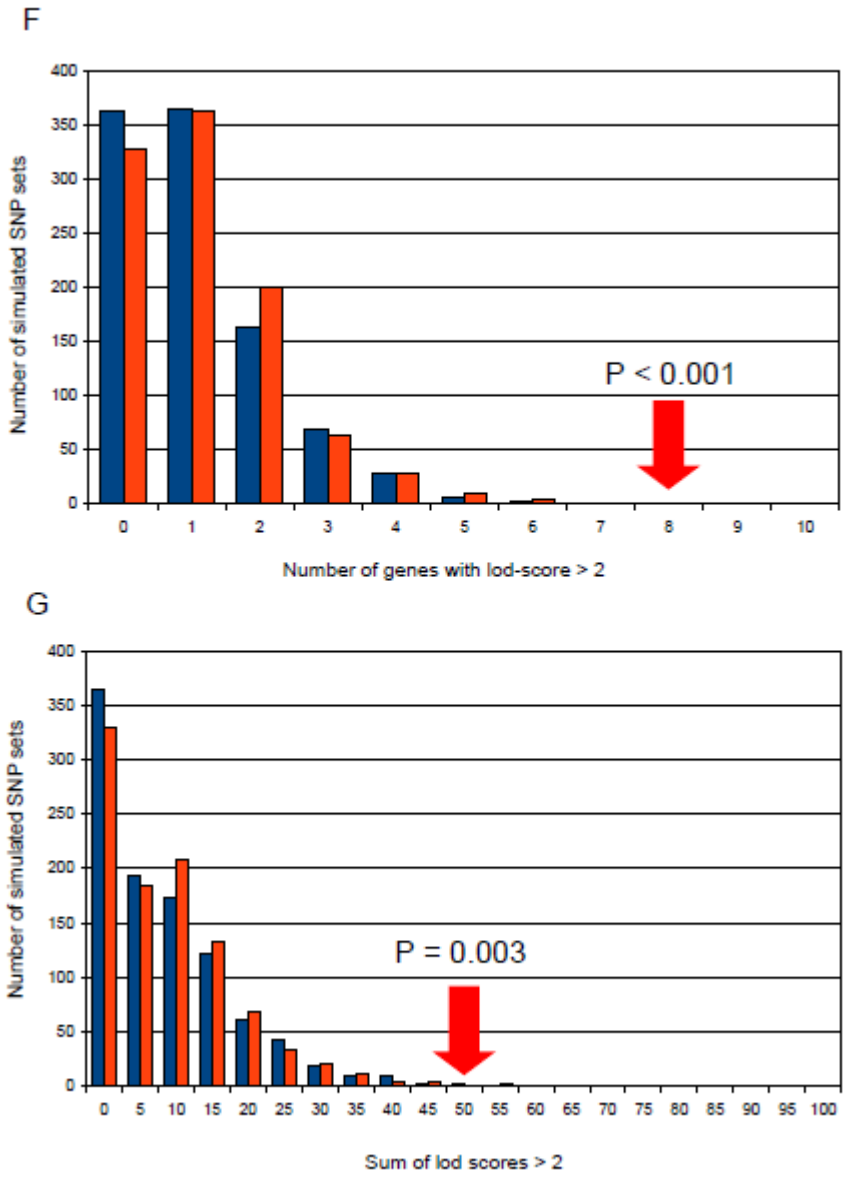


Figure IV.8: *(F-G) Blue bar graphs: Frequency distribution of number of genes with lod scores > 2 (F),and sum of lod scores > 2 (G) for 1,000 sets of 39 SNPs matching the true SNP set interms of allelic frequency and gene context. Red bar graphs: cfr. blue bar graph, exceptthat associated intergenic SNPs were replaced by intronic SNPs in the simulations toincrease their proximity to genes. Red arrows: corresponding values obtained with the 39CD associated SNPs.*

V Characterization and genetic analysis of male recombination in cattle.

Abstract:

In this work, we take advantage of three-generational paternal half-sib pedigrees genotyped for ~ 50K genome-wide SNP panels to map 267,801 CO events in 10,218 sperm cells produced by 796 bulls, amounting to a total map length of 26.21 M. Genome-wide recombination rate (GRR) was shown to differ significantly between bulls and to be characterized by a heritability of 23%. We identified one putative QTL (z=5.3) for GRR on BTA 10, encompassing three genes expressed in testis including *REC8* known to play an important role in mammalian recombination. A lodscore peak encompassing *RNF212* reached nominal significance (z=2.9), suggesting that this gene influences GRR in cattle as it does in human. The frequency distribution of number of CO events per chromosome supported the requirement for at least one chiasma per arm, with ~ 85% of bivalents for larger chromosomes (≤ 16) exhibiting two or more chiasmata. CO events tended to cluster in recombination “jungles” (mapping preferentially to proximal sub-terminal and distal terminal regions) and avoid recombination “deserts” (mapping preferentially to proximal terminal and central regions). Genome-wide jungle usage differed significantly between bulls, and was characterized by a heritability of 24%. We were not able to identify QTL with significant effect on jungle usage. Inflated distances between pairs of CO events supported positive interference on all autosomes. Normalized inter-CO distance measuring global interference differed significantly between bulls, yet was characterized by a low heritability (4.5%) and lack of evidence for genome-wide significant QTL in the studied population.

SANDOR, C.; COPPIETERS, W.; DRUET, T.; CHARLIER, C.; GEORGES, M.

V.1 Introduction.

Reciprocal recombination between homologues fulfills an essential mechanistic role during meiosis[165,166]. It is required for proper bivalent alignment on the metaphase I plate preceding disjunction and segregation at anaphase I. Correct segregation of the full chromosome complement demands tight, sex-specific control of the number of CO per arm, as well as of their position relative to chromosomal landmarks (centromeres and telomeres) and other CO (in the case of multichiasmatic meioses)[167,168]. Failures in this process underlie aneuploidies affecting more than 10% of human oocytes[169].

At the population level, recombination affects the rate of creation and loss of haplotypes with cis-configured favourable alleles, placing second order selection pressure on modifiers of global and/or local recombination including inversions[167].

Components of the recombination apparatus are well described in yeast, but remain largely undefined in most other organisms including mammals[168,170]. One strategy to identify such components is to positionally clone the genes and variants that underlie inherited variation in recombination phenotypes. Global recombination is characterized by considerable inter-individual variation which is in part inherited[171,172,173,174]. GWAS have identified several loci influencing global recombination rate (GRR) in human[175,174,176]. These include the 17q21.31 inversion[175], as well as the *RNF212* gene harboring common variants with antagonistic effects on GRR in males and females[174]. Of note, women's recombination rate correlates positively with reproductive succes[177]. In human, 80% of CO events map to 25,000 1-2Kb recombination hotspots[178]. Hotspot usage was shown to considerably differ between individuals[179] and this was shown to involve variation in cis-acting hotspot-triggering sequences[180], as well as in the trans-acting *PRDM9* H3K4 trimethyltransferase and hotspot regulator[181,182,183,184]. Of note, recombination hotspots and their *PRDM9* regulator undergo accelerated evolution (explained in part by the self-destructive drive of hotspot motifs due to biased gene conversion)[185,186,183], and *PRDM9* has been identified as a hybrid sterility gene in the mouse[187]. Genome-wide levels of cross-over interference were also suggested to differ between women[171], but corresponding genetic variants – if existing - have not been identified thus far.

We herein describe our attempts to take advantage of (i) the large multigenerational half-sib pedigrees typifying dairy cattle population, and (ii) the systematization of genome-wide SNP genotyping with ~ 50K medium density arrays for "genomic selection" purposes[62], to quantify inter-individual variation in recombination phenotypes as well as to map contributing genetic loci. The bovine haploïd genome is estimated at 2.87Gbp distributed over 29 acrocentric chromosomes, and a pair of metacentric sex chromosomes (The Bovine Genome

Sequencing and Analysis Consortium[188]). Total map length was previously estimated at ~ 31M and shown (contrary to most other mammals) not to differ between sexe[70]. The potential corrrelation between recombination rate and fertility, as well as the hypothesized effect of domestication on recombination rates[189] adds to the interest of a detailed characterization of recombination phenotypes in livestock.

V.2 Results and Discussion.

V.2.1 Identifying cross-over (CO) events.

We took advantage of a dataset of 10,192 bulls from Holland (H) and 3,942 bulls from New-Zealand (NZ) that were genotyped for marker panels comprising respectively 50,876[190] and 51,456[191] genome-wide SNPs of which 19,487 in common. The 14,134 bulls assort in 432 three-generational paternal half-sib pedigrees of the structure shown in Figure V.1. All Dutch bulls were from the Holstein-Friesian breed, while in NZ 59% of the bulls were Holstein-Friesian, 37% Jerseys, and 4% of mixed-breed origin. Marker genotypes were phased following Druet & Georges[192], i.e. by using Mendelian rules to phase heterozygous genotypes of sons (generation II (GII) and GIII) and linkage information to phase heterozygous genotypes of sires (GI and GII). CO events having occured in the germ line of the 796 GII sires were then identified as marker intervals in which the paternal homologue of GIII sons switched phase. Double-CO occuring in intervals that were separated by less than three informative intervals were ignored. The distribution of CO-events was surveyed using a graphical interface to identify as many other artifacts as possible. The size-distribution of the marker intervals with assigned CO is shown in Figure_Sup V.1. Median size was 899kb and 721kb in the Dutch and New-Zealand populations, respectively.

V.2.2 Genome-wide recombination rate (GRR).

We identified a total of 267,801 CO events in 10,218 gametes, corresponding to an average genome size of 26.21 Morgan. This value is lower than previous estimates of total bull map length (f.i. 31.58 M in Ihara[70]), but in better agreement with the general relationship between number of chromosome arms and total map length in mammals[167].

We noted a small but significant effect of the number of half-sibs on GRR (Figure_Sup V.2) which is thought to reflect errors in determining the sire's phase in smaller families. The effect of the number of half-sibs was estimated by simulation as detailed in Supplemental Methods and GRR corrected accordingly.

Figure V.2 shows the distribution of corrected GRR in the paternal genome of the 10,218 GIII sons sorted by GII sire. Average GRR of the GII sires ranged from 20.2 to 27.1. The fixed effect of GII sire on GRR was highly significant ($p < 0.0001$).

To evaluate the repeatability of the differences in average GRR, we took advantage of 72 GII sires shared by the Dutch and New-Zealand populations. Figure V.3 shows the correlation between the average GRR estimated for the same GII sires using information from non-overlapping sets of GIII sons from H and NZ, respectively. Spearman's rank correlation was 0.61 ($p < 4 \times 10^{-8}$).

We estimated the h^2 of GRR using an individual animal model[76] accounting for all known genealogical relationships between the GII sires (Methods). h^2 was estimated at 23% and 22% in the Dutch and New-Zealand populations respectively, i.e. comparable to the 30% h^2 of global female recombination in the Islandic population[177].

To map QTL influencing GRR we used a HMM-based approach that simultaneously exploits linkage and LD information[192]. Individual chromosomes were assigned to 20 hidden states (or ancestral haplotype clusters) using linkage and LD information, and "random" ancestral haplotype effect on GRR estimated by REML using a mixed model including an individual animal effect to correct for stratification. Significance thresholds were determined by permutation (Methods). Figure V.4 shows the location scores obtained across the genome in the Dutch population. We obtained a genome-wide significant QTL on BTA10 (z= 5.3; Figure_Sup V.3.A), and near genome-wide significant QTL on BTA19 (z=4.4; Figure_Sup V.3.B). The lod-2 drop-off QTL confidence interval of the BTA10 QTL spans 1.4Mb encompassing 47 genes. Three of these are worth mentioning as they are strongly expressed in testis: *TBC1D21* (TBC1 domain family, member 21), *TSSK4* (Testis-specific serine kinase 4), and *REC8* (Rec8 homolog (yeast)). *REC8* codes for a member of the kleisin family of SMC (structural maintenance of chromosome) proteins, that localizes to the axial elements of chromosomes during meiosis in both oocytes and spermatocytes. The mouse homologue is a key component of the meiotic cohesion complex, which regulates sister chromatid cohesion and recombination between homologous chromosomes[193,194]. The lod-2 drop-off QTL confidence interval of the BTA19 QTL spans ~ 0.6Mb encompassing two genes: *KCNJ2* and *KCNJ16* (Potassium inwardly-rectifying channel, subfamily J, members 2 and 16) which are not known to be expressed in gonads or involved in recombination. On BTA6 we obtained a lodscore peak of 2.9 that spans the position of the *RNF212* gene. This suggests that variation in *RNF212* affects GRR in cattle as it is in human[174].

Of note, we have so far failed to confirm either of these three QTL in the NZ population. The reasons for this discrepancy are being examined.

V.2.3 Chromosome-specific recombination rates (CRR).

Figure V.5 shows the relationship between chromosome size (in bp according to UMD3.0 build) and average number of observed CO-events. As in human, average number of CO-events is remarkably well (r^2=0.96) predicted by the requirement for at least one chiasma (Y-intercept β_0 of 0.5 CO at size 0 bp) and chromosome size. The slope of the regression (β_1=0.07CO/10Mb) appeared intermediate between the slopes characterizing male and female recombination in human[195].

Also in agreement with the obligate chiasma theory, the proportion of gametes with zero CO was less than expected for all 29 chromosomes assuming a Poisson distribution of CO events (data not shown). We used the frequency distribution of gametes with 0, 1, 2, ... CO-events to estimate the proportion of meioses with 0, 1, 2, ... chiasmata assuming absence of chromatid interference. Figure V.6 shows the corresponding estimates for the 29 autosomes. The data are best explained assuming a near absence of nullichiasmatic meioses for autosomes 1 to 16, and frequencies < 5% for the smaller chromosomes. For the same 16 chromosomes, the most likely (ML) frequency of meioses with at least 2 chiasmata is considerably higher than expected under a truncated Poisson model (forcing the proportion of nullichiasmatic meioses at zero)[196], supporting the obligate occurrence of a second chiasma for larger chromosomes.

CRR were shown to differ very significantly between GII-sires for all chromosomes (Table V.1). The correlation (Spearman's rank correlation) between CRR measured independently in the Dutch and New-Zealand populations for the 72 common sires was positive for 26/29 autosomes but only significant (Bonferroni-corrected) for one (BTA7). Heritabilities for CRR, estimated using an individual animal model (cfr. above), were low (< 7%). We conclude that genetic variation in recombination rates are more likely controlled at the level of the genome rather than chromosome.

V.2.4 Locus-specific recombination rates.

Figure V.7 and Figure_Sup V.4 show maps of male recombination intensity in non-overlapping windows of 60Kb. Recombination rate per 60Kb averaged 0.00062 in the Dutch (range: 0 to 0.0042) and 0.00063 in the NZ material (range: 0 to 0.0065). The correletion between window-specific recombination rate in the Dutch and NZ population was high (r^2 =0.76; $p < 0.0001$), despite the use of distinct SNP panels. Window-specific RR were normalized for local variation in marker density and informativeness as described in Methods. The distribution of normalized recombination rates strongly departed from that expected assuming a uniform distribution of CO events, with a large excess of both "hot" and "cold" windows (Figure V.8). Figure V.7 and Figure_Sup V.4 show

the location of windows in which the observed RR deviates by more than 2.5 (local) SD from the mean. Paraphrasing Chowdhury et al.[176], "hot" and "cold" windows clustered in recombination "jungles" and "deserts", respectively. Jungles tend to concentrate in subterminal (proximal chromosome end) and terminal regions (distal chromosome end), while deserts concentrate in the middle of the chromosome arms as well as in terminal regions (proximal chromosome end) coinciding with the centromeres (Figure V.9). Of note, because of the difference in map resolution, "jungles" and "deserts" as defined in this work (≥ 60Kb) cannot be compared with recombination hot and cold-spots defined in human genetics (~ 5Kb[178,197]).

We examined whether recombination "jungles" and "deserts" differed from each other with regards to sequence composition and content. Remarkably, the abundance of 29 of the 51 considered repeat families (as defined by Repbase) differed significantly (Bonferroni corrected) between the two compartments (Table V.2). Twenty-three repeat types were overrepresented in "jungles" (including SINE/MIR, LINE/L2, and DNA/MER1 which are also very significantly overrepresented in human hotspots[178], and six underrepresented (including LINE/L1 found to be underrepresented in human hostspots[178]).

In mice and human, genome-wide hotspot usage (GHU) has been shown to be genetically controlled with variation at the *PRDM9* gene having a major effect[181,182,183,184]. We computed the proportion of CO events falling in recombination "jungles" for the 10,218 paternal GIII genomes. On average, 34% (range: 8% - 53%) of Dutch CO events and 34% (range: 0% - 62%) of NZ CO events could be assigned to "hot" windows representing respectively 14% (H) and 12% (NZ) of the genome. Average genome-wide jungle usage (GJU) differed significantly between GII sires ($p < 0.002$)(Figure V.10), but was not affected by family size ($p=0.86$) or GRR ($p=0.86$). For the 72 shared sires, differences in average GJU in the Dutch and NZ population were significantly correlated ($\rho=0.31$; $p < 0.01$)(Figure V.11). The heritability of GJU was estimated at 23% in the Dutch and 21% in the NZ population. We scanned the genome for QTL affecting GJU using the same mixed model as before. A suggestive QTL ($z = 3.6$) was obtained on BTA3 (Figure V.12). The lod-2 drop-off QTL confidence interval spans ~ 1.16 Mb and encompasses three genes *LOC781798* (similar to Integrin β1 binding protein 1), *LOC522984* (similar to Eucaryotic translation elongation factor 1 α1) and *OLFM3* (olfactomedin 3) not obvioulsy related to recombination. There was no evidence for lod score peaks spanning the two adjacent autosomal PRDM9 paralogues.

The genome is being scanned for possible cis-acting haplotype effects on local recombination rates (cfr. Methods).

V.2.5 Genome-wide and chromosome-specific CO interference.

Figure_Sup V.5 shows the CO positions for paternal GIII homologues with two recombination events (Dutch population; n=36,933) across the 29 autosomes. As expected, CO pairs tend to be further apart than expected assuming independent CO positioning (i.e. positive interference). Following Lian et al.[198], we quantified the degree of interference for each chromosome using the shape parameter (ν) of a gamma distribution[199] maximizing the likelihood of the observed inter-CO distances (paternal homologues with two recombination events). Chromosome-specific ν parameters estimated from non-overlapping Dutch and New-Zealand GIII sons were highly repeatable (ρ=0.76, $p < 0.00001$)(Figure_Sup V.6). Interference tended to increase with increasing size for chromosomes 29 to 16, yet to decrease with increasing size from chromosomes 15 to 1 (Figure_Sup V.7). Chromosomes-specific ν-values were shown to correlate (Spearman's rank correlation) positively with DNA/Mariner content and negatively with RC/Helitron content (Bonferroni-corrected $p < 0.01$).

We then evaluated inter-individual variation in genome-wide interference levels. To that end, we expressed inter-CO distances (homologues with two recombination events) in standardized (in SD) deviations from the chromosome mean. GII sire proved to have a significant fixed effect on this standardized distance ($p < 0.001$) (Figure V.13). GII sire-specific effects were repeatable between the Dutch and NZ samples (ρ=0.46; $p< 0.0002$) (Figure V.14). There was no convincing evidence, however, for a strong heritable component underlying the observed differences (h^2 = 4.5%). A genome-scan for QTL influencing genome-wide interference levels did not reveal signals exceeding z=4.7 corresponding to the approximate threshold for genome-wide significance (Figure V.15).

V.3 Methods.

V.3.1 Marker phasing.

Marker phasing was conducted with the Phasebook software package[192]. To identify CO events that occured in the germ line of the GII sires, we exploited Mendelian rules to phase SNP genotypes in sons (GII and GIII), and linkage information to phase SNP genotypes in sires (GI and GII). CO events were then identified as phase switches in the gametes transmitted by the GII sires to their GIII sons. To map QTL, we additionally exploited LD information as described in Druet & Georges[192]. As result, and at each SNP position, all 2x10,192 homologues in the data set are assigned to one of 20 hidden states corresponding to "ancestral haplotype

clusters” .

V.3.2 Estimating h^2.

Narrow sens heritabilities (h^2) of recombination phenotypes (measured in the GIII sons) were estimated using two mixed models. The first modelled average phenotypes of GII sires, and included an overall mean, a random individual animal effect (with variance-covariance structure proportionate to twice the coefficient of kinship between corresponding GII sires), and a random error proportionate to the inverse of the number GIII sons per GII sire. The second modelled the individual phenotypes of the gametes transmitted to GIII sons. It included an overall mean, a random individual animal effect (with variance-covariance structure proportionate to twice the coefficient of kinship between corresponding GII sires), a random permanent non-additif GII sire effect, and a random error. Variance components were estimated by REML analysis[77].

V.3.3 QTL mapping.

QTL were mapped using a previously described mixed model approach that simultaneously exploits linkage and LD information[192]. The utilized mixed models are the same as those used to estimate h^2 with addition of a random “ancestral haplotype cluster” effect. The covariance between ethe effects of the 20 possible “ancestral haplotype clusters” is assumed to be zero. Significance thresholds are empirically determined by phenotype permutation[200] conducted within GII brother families.

V.3.4 Measuring and normalizing 60 Kb window-specific recombination rates.

The recombination rate in a defined 60Kb window was computed as $\frac{1}{T}\sum_{i=1}^{n} o_i / x_i$ where n is the total number of CO events identified on the corresponding chromosome in the analyzed population, x_i is the size (in bp) of the marker interval to which CO i has been mapped, o_i the overlap (in bp) between the 60 Kb window and CO interval i, and T is the total number of analyzed gametes. To normalize window-specific recombination rates for local marker density and informativeness, we simulated (1,000 times) genotypes for the GIII sons by randomly “dropping” CO events on the phased GII chromosomes assuming a uniform distribution of CO events following

a Poisson process (with mean corresponding to the real data), randomly sampling one of the two paternal chromosomes, while keeping the original maternal chromosome intact. The entire phasing and CO mapping process was then reinitiated with these *in silico* generated SNP genotypes.

V.3.5 Identifying distinctive features of recombination "jungles" and "deserts"

We used a permutation test to evaluate the statistical significance of the difference in repeat content (number of events identified with Repeatmasker: http://www.repeatmasker.org/) of recombination "jungles" and "deserts". Obtained p-values were Bonferroni-corrected for the number of families/elements evaluated.

V.3.6 Scanning the genome for cis-acting haplotype effects on local recombination rate.

To identify cis-acting haplotype effects on local recombination rate, we defined 800 Kb windows centered around the interrogated marker position. At that marker position, we selected the GIII sons of the GII sires that were heterozygous for "ancestral haplotype clusters"[192] and tested the additive effect of "ancestral haplotype cluster" of the GII sires on the recombination phenotype of their GIII sons by ANOVA. The recombination phenotype of GIII sons was defined as the probability that a paternal CO event would have occurred in the interrogated window measured as the degree of overlap between CO encompassing marker intervals and interrogated window.

V.4 Acknowledgments.

We are grateful to CRV and LIC for providing the SNP genotype and pedigree information. Cynthia Sandor has benefitted from financial support of the Fonds National de la Recherche Scientifique (FNRS; FRIA fellowship), and of the Communauté Française de Belgique (BIOMOD Action de Recherche Concertée).

V.5 Supplemental method: correcting GRR for family size.

We noted that estimates of GRR decreased with increasing family size (Figure_Sup V.2) and attributed this to errors in determining the sire's phase. To correct GRR for this factor we used 10 paternal half-sib families with > 100 GIII sons. From these families we randomly sampled (1,000 times) from 1 to 10 sons with corresponding SNP genotypes. Phasing of the GI, GII and GIII bulls was conducted with Phasebook on these purposely limited data-sets, including determination of CO events in the paternal gametes transmitted to GIII sons. For each of the 10 families we then compared average GRR estimated with 1, 2, ... 10 sons (over the 1,000 simulations) with GRR estimated with all sons (> 100), yielding a set of $\bar{\Delta}_{ij}$ values where *i* corresponds to the number of used sons (1 to 10) for family *j*. These values were averaged across families to generate $\bar{\Delta}_{i}$, i.e. an overall effect on GRR of family size *i*, used to correct the actual GRR estimates obtained from families with < 10 half-sibs.

Table V.1: *Statistical significance of "fixed" GII-sire effect on chromosome-specific recombination rate (CRR), assessed by ANOVA (Pval: nominal p-values; Pval*: p-values obtained by permutation). Spearman's rank correlation and significance between the average CRR of 72 shared GII sires, estimated separately from non-overlapping sets of Dutch and NZ GIII sons. Heritability of CRR estimated using an animal model*[76].

Chr	Effect of GII sire on CRR		"Repeatability"		Heritability
	Pval	Pval*	Spearman's Rho	Pval	h^2(%)
1	**5,07E-060**	**<1/1000**	0,14	0,23	4,01
2	3,85E-009	<1/1000	0,18	0,13	**5,52**
3	1,96E-004	<1/1000	0,06	0,60	4,52
4	1,85E-006	<1/1000	0,05	0,66	**5,23**
5	**9,07E-002**	**8,80E-002**	-0,02	0,86	4,58
6	2,66E-006	<1/1000	0,05	0,68	**6,69**
7	7,38E-004	2,00E-003	**0,40**	**0,00**	**5,32**
8	4,52E-009	<1/1000	**0,28**	**0,02**	3,90
9	1,90E-012	<1/1000	0,17	0,14	6,33
10	2,89E-011	<1/1000	**0,30**	**0,01**	**5,61**
11	6,58E-005	<1/1000	0,00	0,99	4,56
12	8,28E-005	<1/1000	**0,29**	**0,01**	**5,39**
13	8,96E-008	<1/1000	0,04	0,72	4,26
14	5,04E-007	<1/1000	0,00	0,98	**7,16**
15	1,46E-006	<1/1000	0,11	0,35	**5,08**
16	3,19E-003	4,00E-003	0,18	0,13	4,90
17	1,33E-017	<1/1000	0,04	0,76	**7,51**
18	3,39E-008	<1/1000	**0,26**	**0,03**	**5,60**
19	2,97E-011	<1/1000	-0,07	0,54	**5,15**
20	1,93E-015	<1/1000	**0,26**	**0,02**	**5,81**
21	7,83E-018	<1/1000	0,13	0,26	4,21
22	1,06E-005	<1/1000	0,10	0,41	**5,26**
23	2,73E-006	<1/1000	0,14	0,24	4,87
24	1,09E-004	<1/1000	0,21	0,08	**6,62**
25	3,51E-014	<1/1000	0,17	0,16	4,30
26	9,35E-008	<1/1000	0,20	0,09	4,75
27	1,57E-032	<1/1000	0,11	0,38	**5,34**
28	8,10E-010	<1/1000	0,09	0,45	4,06
29	**5,21E-002**	**6,40E-002**	-0,19	0,12	4,81

Table V.2: *Statistical significance of the difference in repeat content of recombination "jungles" and "deserts"; Pval: nominal p-values obtained by permutation test to evaluate the statistical significance of the difference in repeat content; Pval*: pval were Bonferroni-corrected for the number of familie/elements evaluated (51). Ratio between the "jungle" repeat content and the "desert" repeat content (after normalization for the size).*

Name	Ratio	Pval	Pval*
LTR/ERVK	0,61	<1/10000	0,0051
SINE/MIR	1,42	<1/10000	0,0051
SINE/BovA	1,21	<1/10000	0,0051
LINE/RTE-BovB	0,66	<1/10000	0,0051
LINE/L1	0,88	<1/10000	0,0051
LTR/ERVL-MaLR	1,32	<1/10000	0,0051
LTR/ERV1	0,75	<1/10000	0,0051
LINE/L2	1,19	<1/10000	0,0051
SINE/tRNA-Glu	1,18	<1/10000	0,0051
DNA/hAT-Charlie	1,26	<1/10000	0,0051
tRNA	0,39	<1/10000	0,0051
LINE/CR1	1,27	<1/10000	0,0051
DNA/MER1_type	1,27	<1/10000	0,0051
SINE/RTE-BovB	0,94	<1/10000	0,0051
DNA/AcHobo	1,77	<1/10000	0,0051
DNA/hAT	1,42	<1/10000	0,0051
DNA/hAT-Blackjack	1,32	<1/10000	0,0051
LTR/Gypsy	1,4	<1/10000	0,0051
DNA	1,84	<1/10000	0,0051
DNA/hAT?	2,11	0,0002	0,0101
DNA/hAT-Tip100	1,17	0,0005	0,0252
LTR/Gypsy?	1,41	0,0004	0,0202
DNA?	1,76	0,0001	0,0051
SINE?	4,51	0,0001	0,0051
Unknown	1,33	0,0007	0,0351
DNA/TcMar-Tigger	1,11	0,0004	0,0202
LINE/RTE	1,16	0,0039	0,1807
SINE	1,45	0,0009	0,0449
SINE/tRNA	1,91	0,0011	0,0546
DNA/Tip100	1,18	0,0264	0,7445
DNA/Mariner	0,66	0,0212	0,6647
DNA/TcMar-Mariner	1,24	0,0069	0,2975
DNA/Tc2	1,3	0,0373	0,8561
RC/Helitron	1,39	0,0287	0,7735
DNA/TcMar?	1,29	0,0124	0,4708
rRNA	0,73	0,0092	0,3759
DNA/PiggyBac	0,63	0,0890	0,9914
snRNA	0,82	0,0409	0,8811
DNA/TcMar-Tc2	1,19	0,0270	0,7524
DNA/MER2_type	1,07	0,1170	0,9982
LINE/Dong-R4	1,62	0,0845	0,9889
Satellite/centr	0,91	0,3917	1,0000
SINE/Deu	1,13	0,1152	0,9981
DNA/hAT-Tip100?	0,78	0,0927	0,9930
LTR/ERVL	1,02	0,2458	1,0000
LTR/ERV	0,79	0,1919	1,0000
DNA/Charlie	1,16	0,3423	1,0000
LTR/ERVL?	0,91	0,2837	1,0000
RNA	1,19	0,3346	1,0000
LTR	1,08	0,3411	1,0000
DNA/Tigger	0,98	0,4449	1,0000

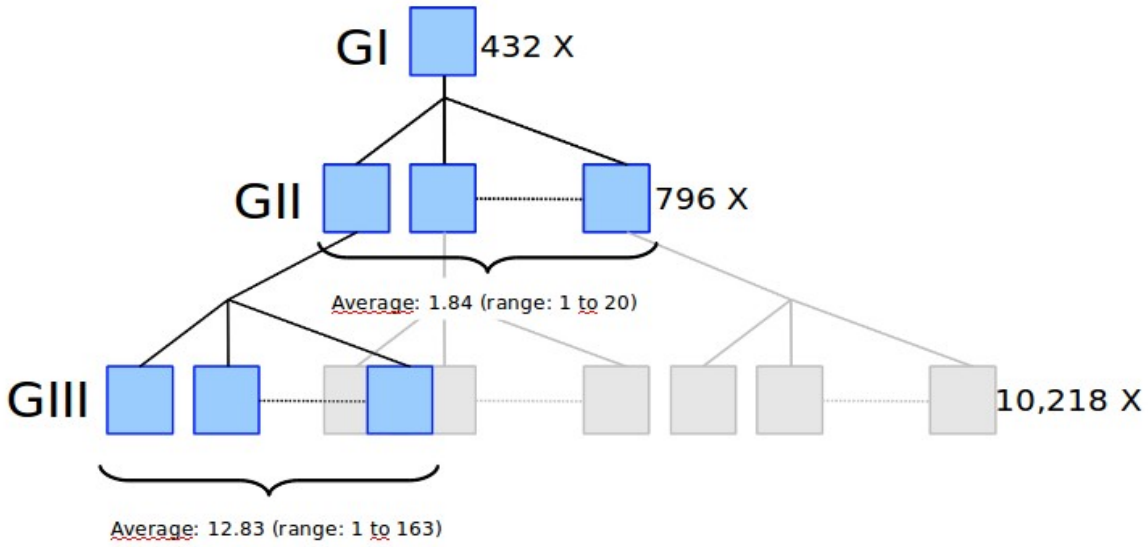


Figure V.1: *Three-generational pedigrees used to map genetic determinants of variation in male recombination rate in cattle. 10,192 (Dutch population) and 3,942 (NZ population) bulls, genotyped for 60K SNP panels assort in 432 three-generational pedigrees of the kind illustrated. Each pedigree is composed of one grand-sire with 1.84 GII sons on average (range: 1 to 20). Each GII sire has 12.84 GIII sons on average (range: 1 to 163). We have used the available SNP genotypes to identify 267,801 CO events that occured in the sperm cells transmitted by the GII sires to their 10,218 GIII sons. QTL affecting variation in recombination rates were mapped by exploiting linkage information (effect of the homologues transmitted by the GI grand-sires to their GII sons) and LD information (effect of haplotypes transmitted by the GI-grand-sires and GI-grand-dams (not genotyped) to their GII sons).*

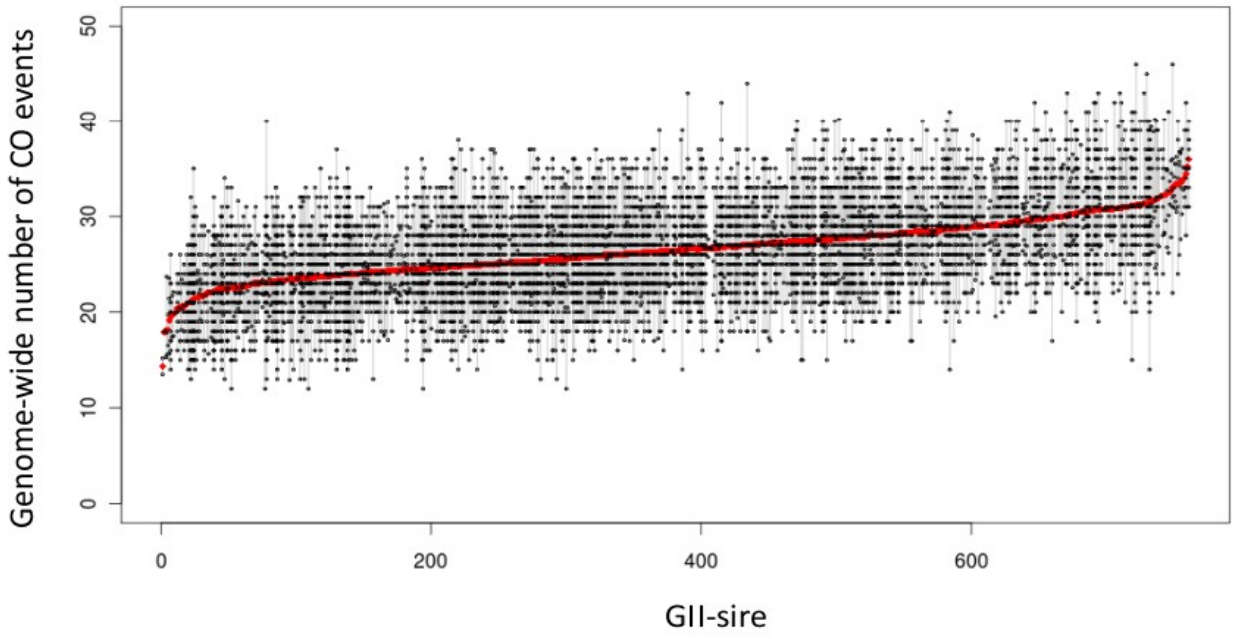


Figure V.2: *Black dots correspond to the total number of CO events identified in the paternal genome of 10,218 GIII sons sorted by GII sire. The red dots mark the average GRR for each GII sire. (B) Correlation between the GRR estimated for 72 GII sires separately from the number of CO events transmitted to Dutch versus NZ GIII sons.*

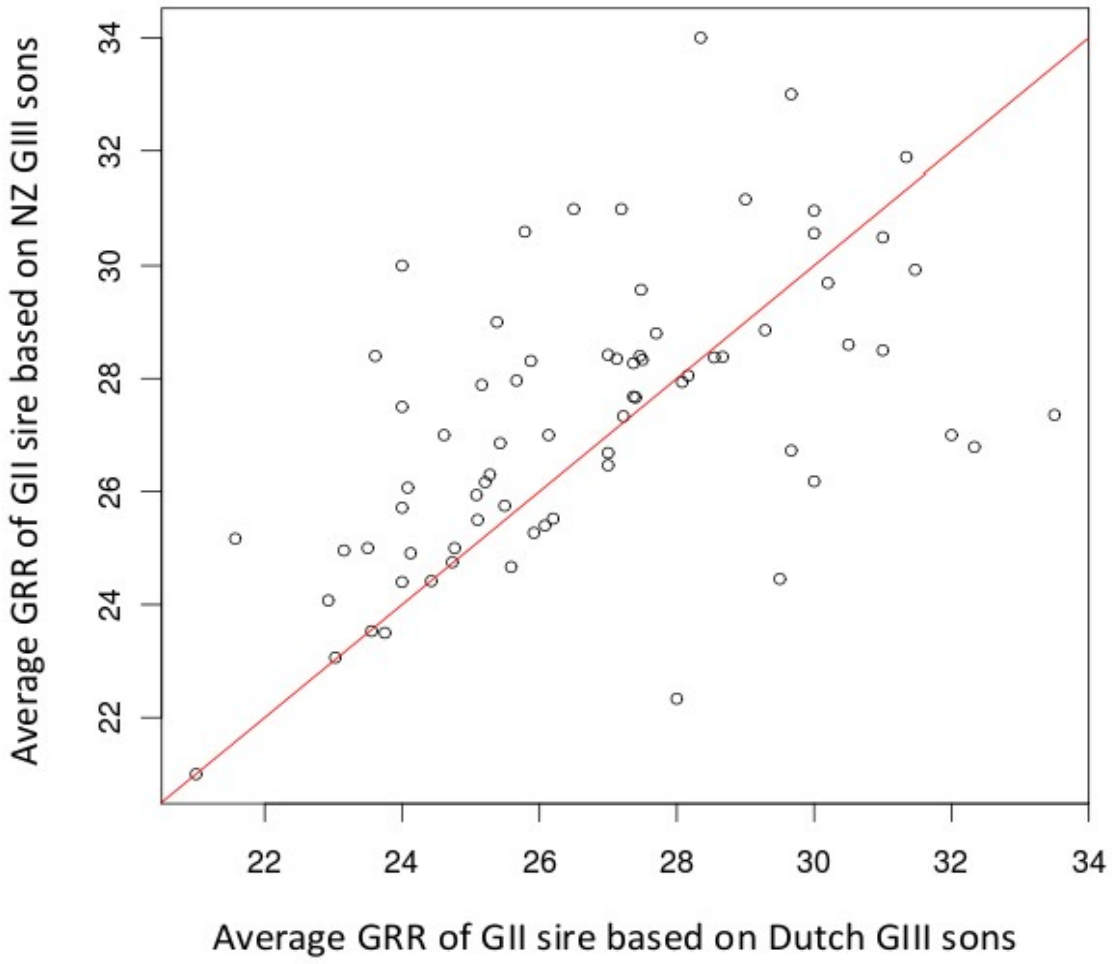


Figure V.3: *Correlation between the GRR estimated for 72 GII sires separately from the number of CO events transmitted to Dutch versus NZ GIII sons.*

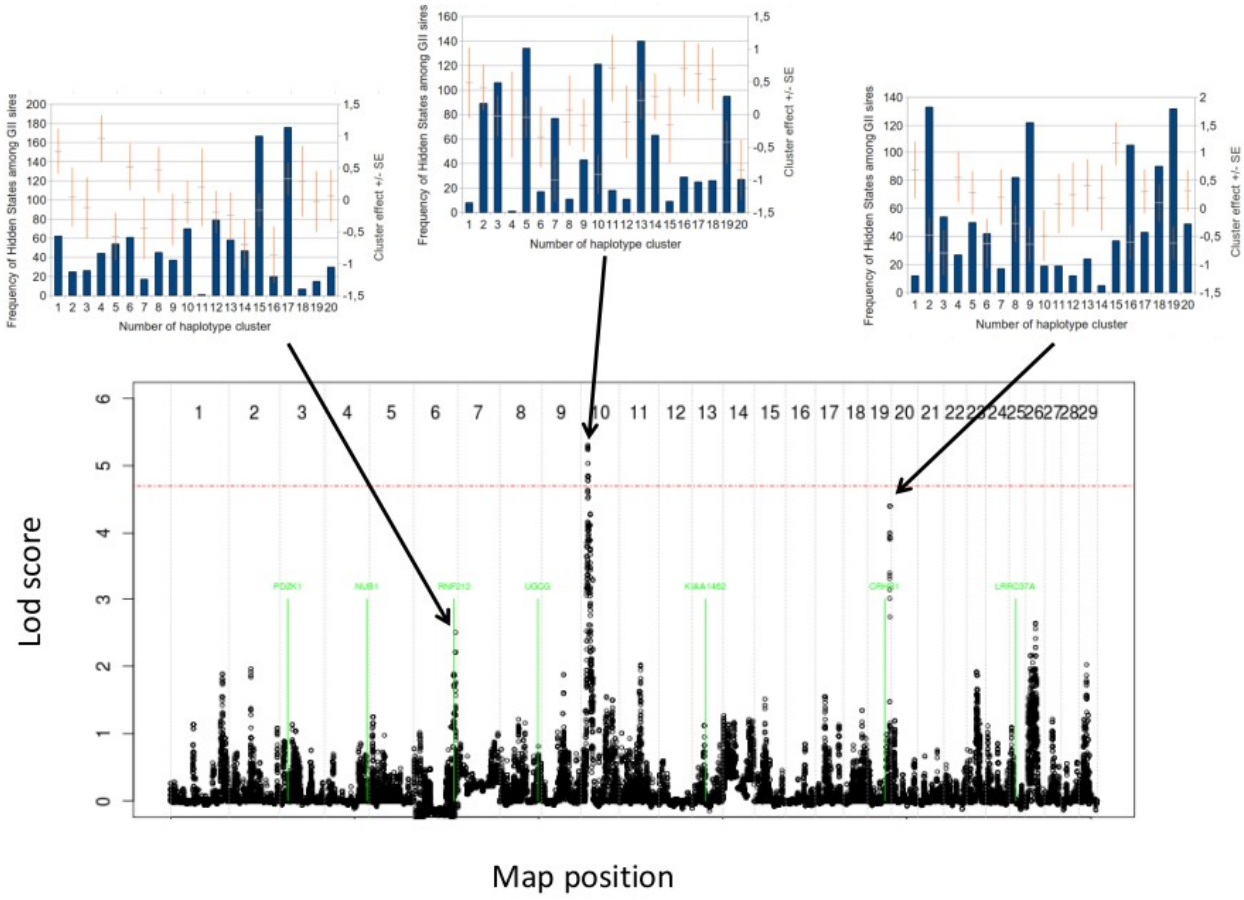


Figure V.4: *Genome-wide lod score profiles obtained for GRR in the Dutch sample set. The red dotted horizontal line corresponds to the empirical genome-wide 5% significance threshold determined by permutation. The position of seven loci that have been previously implicated as determinants of variation in GRR*[174,176] *are shown in green. The population frequency (blue bars) and phenotypic effects (± SE; orange bars) of the 20 ancestral haplotype clusters are shown as insets for three putative QTL on BTA10, BTA19 and BTA6 (at most likely position).*

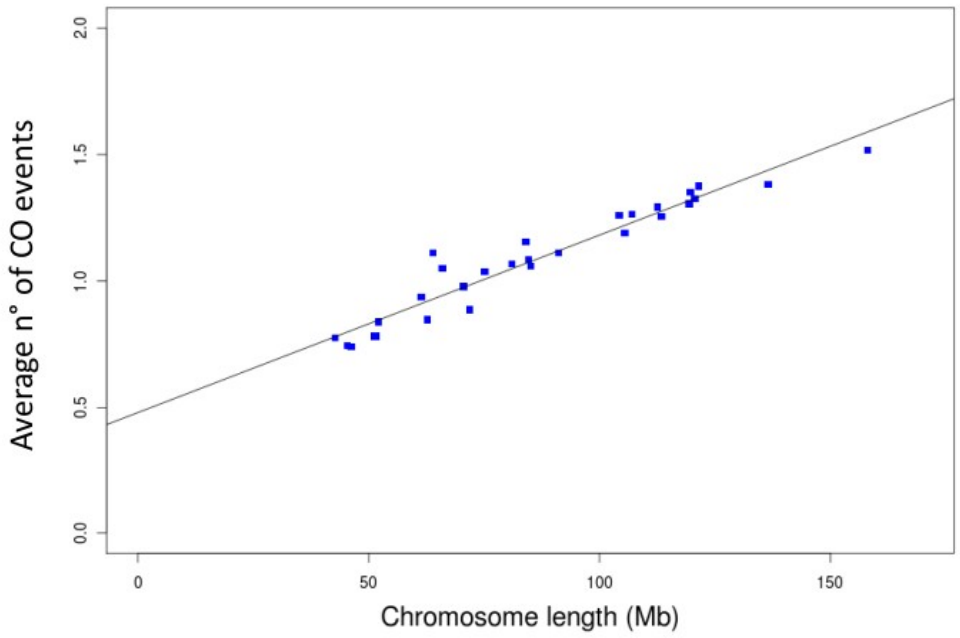


Figure V.5: *Linear relationship between chromosome length in Mb and average number of CO-events for the 29 bovine autosomes. The least square regression is characterized by a Y-intercept β0 = 0.48 and a slope β1 =0.07CO/10Mb.*

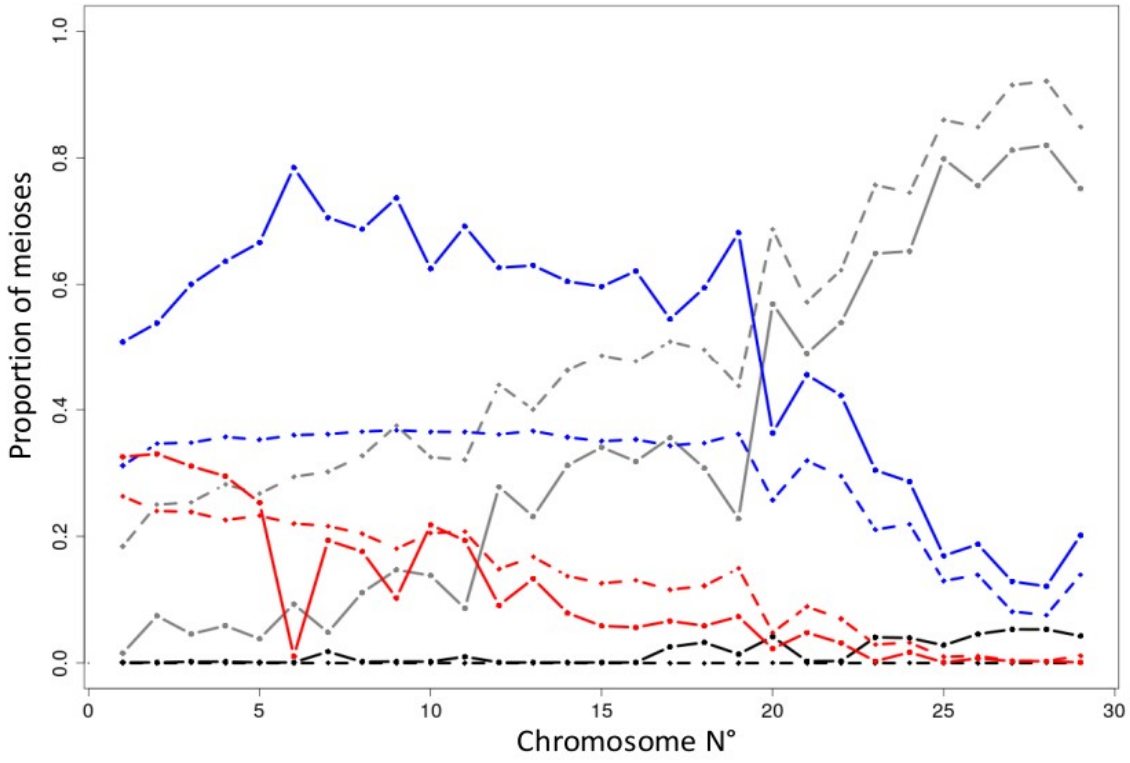


Figure V.6: *Proportion of meioses with zero (black), one (gray), two (blue) and three (red) chiasmata for the 29 bovine autosomes. Plain lines: proportions maximizing the likelihood of the data (assuming no chromatid interference). Dotted lines: expected proportions assuming a truncated Poisson distribution of number of chiasmata (proportion of meioses with zero chiasmata forced at zero).*

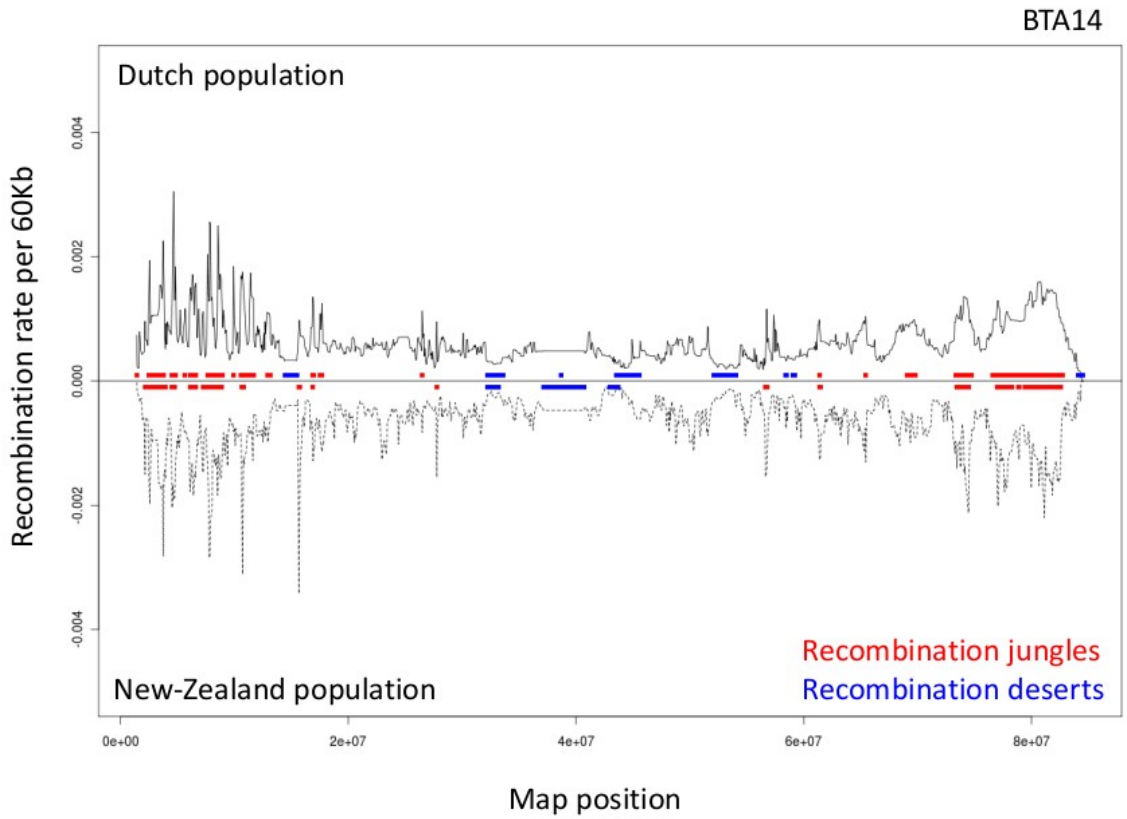


Figure V.7: *Representative example of the variation in male recombination in 60Kb windows across a bovine autosome (BTA14). The plain black line (upper halve) corresponds to recombination rate estimated in the Dutch population, while the dotted black line (lower halve) corresponds to the recombination rate estimated in the NZ population. The red and blue horizontal lines correspond to recombination "jungles" and "deserts", respectively, i.e. segments in which the observed recombination rate deviates by more than 2.5 standard deviations from the local recombination rate expected under a model of uniform distribution of CO events.*

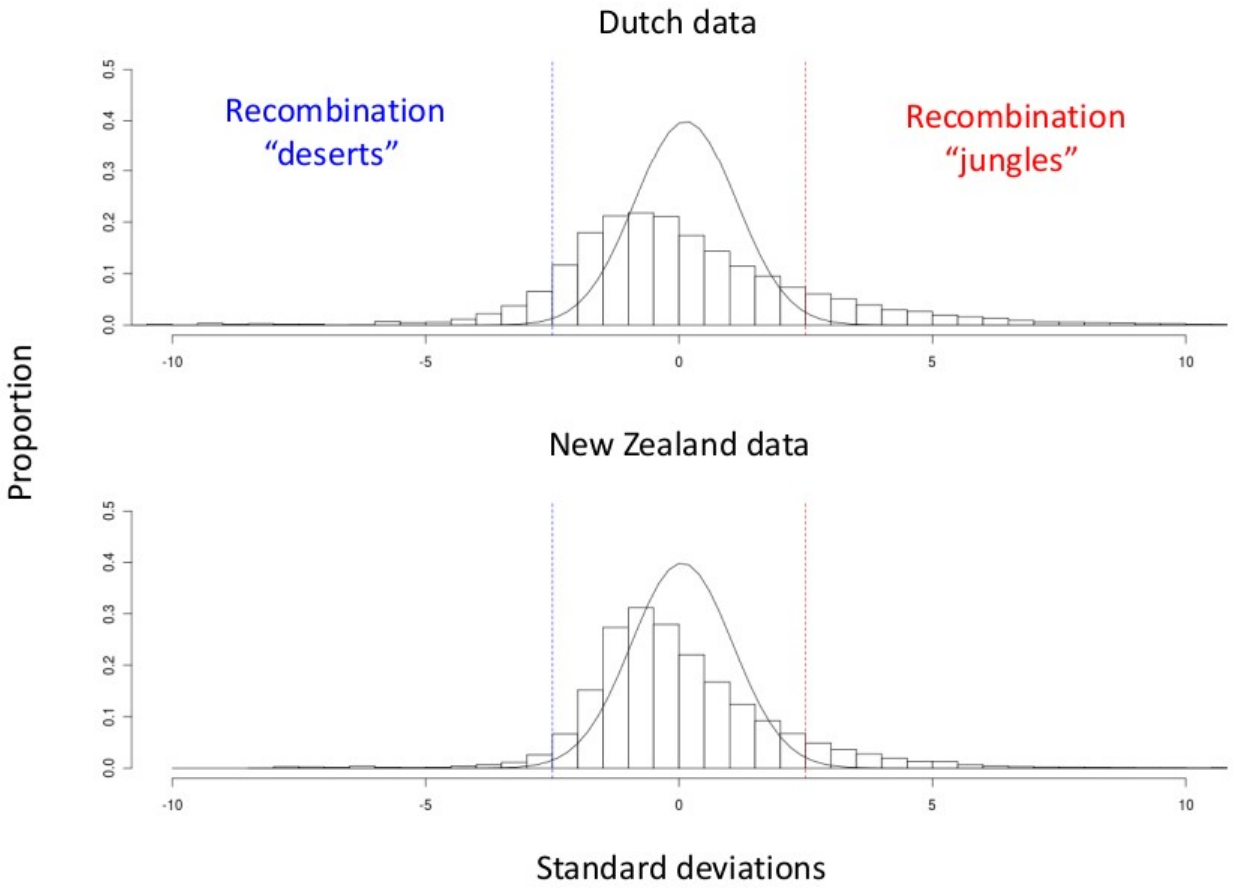


Figure V.8: *Bar graphs: Frequency distribution of local (60Kb window) recombination rate normalized for local marker density and informativeness as described in Methods. Curve: Standard normal distribution with same mean as actual distribution, and variance of one. Red and Blue vertical lines mark the thresholds defining recombination jungles (mean + 2.5 SD) and deserts (mean – 2.5 SD), respectively.*

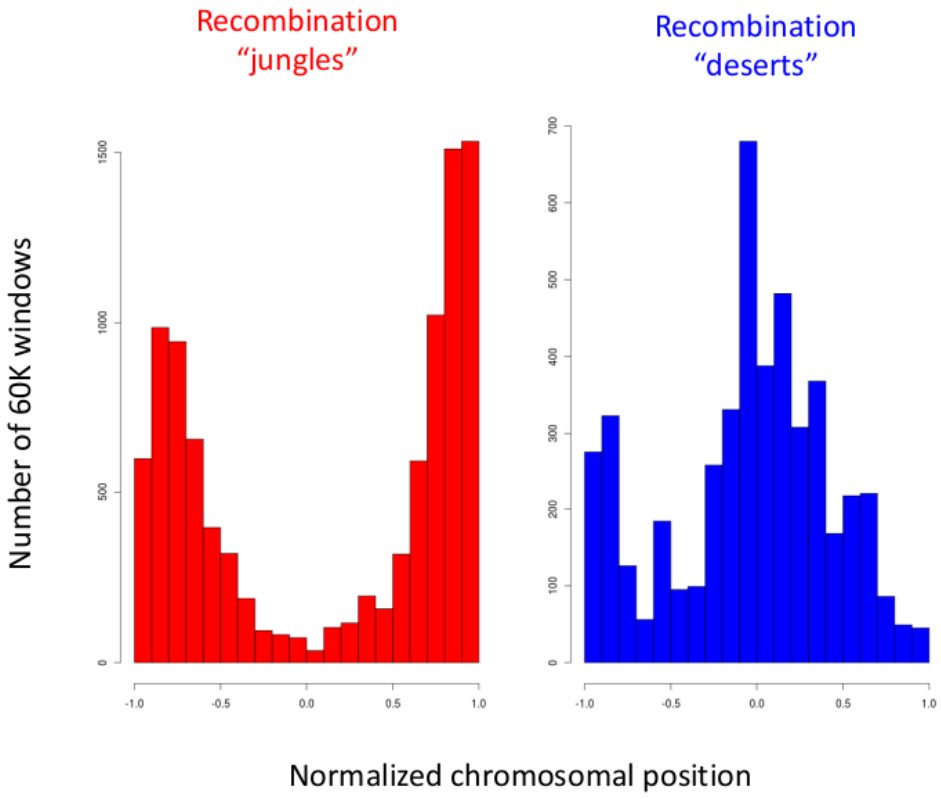


Figure V.9: *Location of recombination “jungles” (red) and “deserts”(blue) relative to normalized chromosome length. All 29 achrocentric autosomes were aligned with their centromere towards the left of the graphs. Jungles tend to concentrate in subterminal (proximal chromosome end) and terminal regions (distal chromosome end), while deserts concentrate in the middle of the chromosome arms as well as in terminal regions (proximal chromosome end) coinciding with the centromeres.*

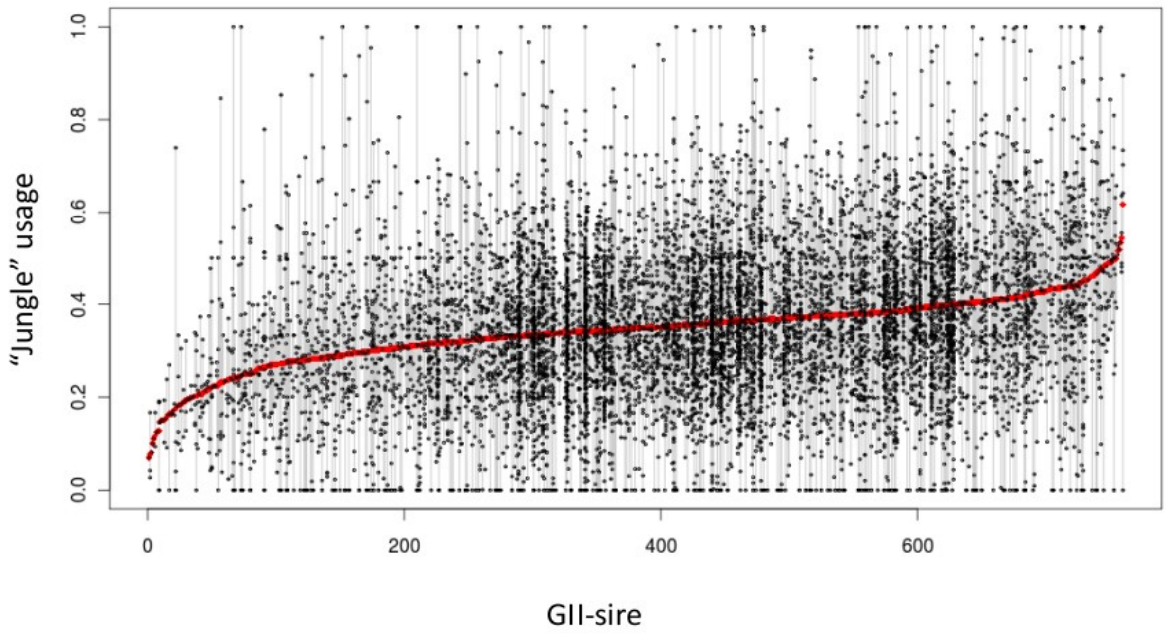


Figure V.10: *Black dots: Average overlap (0 to 1) between marker intervals (< 800 Kb) with assigned CO events and "hot" 60K windows for GIII-sons sorted by GII-sire. Red dots: Average overlap for all CO events transmitted by corresponding GII-sire.*

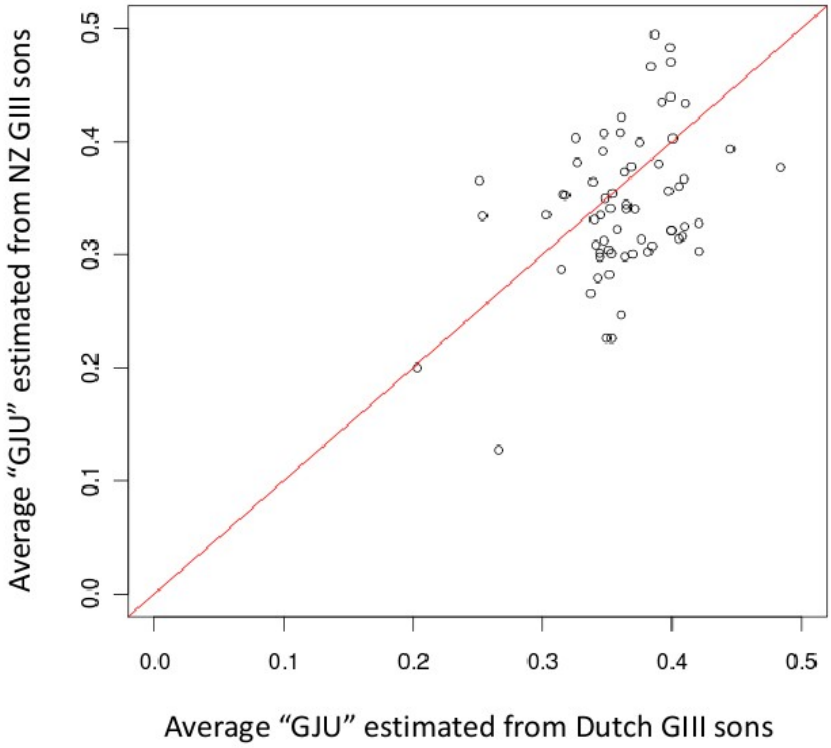


Figure V.11: *Correlation between average jungle usage estimated for the 72 shared GII-sires respectively from gametes transmitted to Dutch versus New-Zealand GIII sons.*

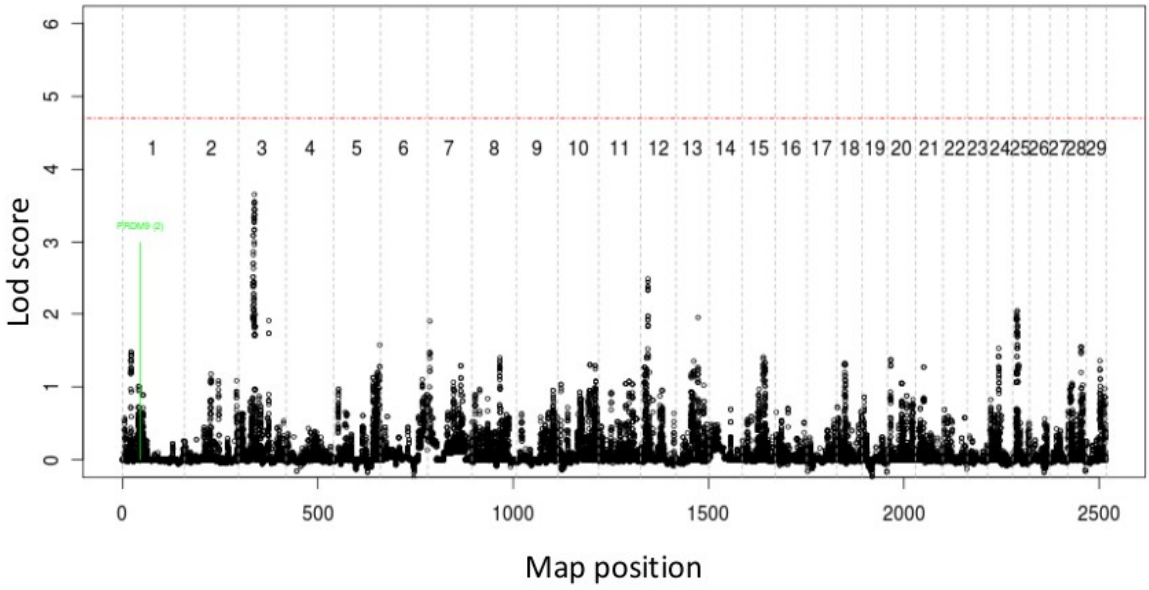


Figure V.12: *Results of genome-scan for QTL affecting genome-wide jungle usage (GJU) using a method that simultaneously extracts linkage and LD signal*[192]*. The red horizontal lines marks the genome-wide threshold for significance (5%) determined by permutation testing.*

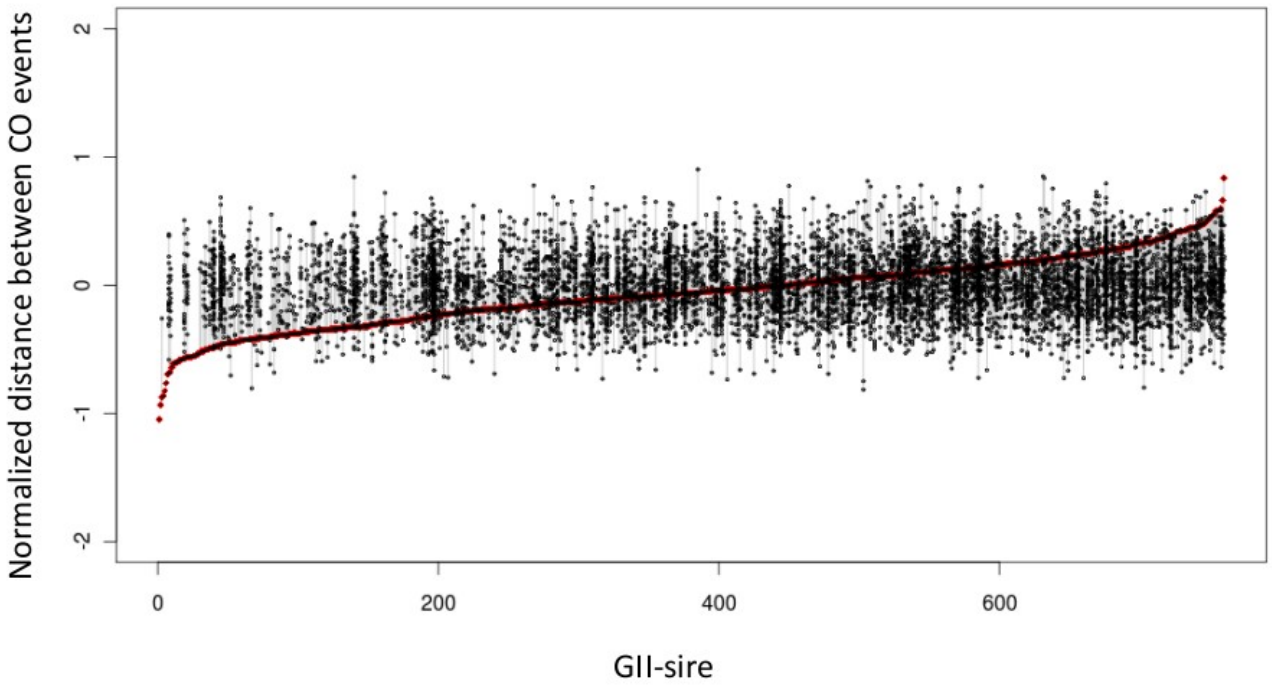


Figure V.13: *Black dots: Average normalized distance between CO events for all paternal homologues with two recombination events for corresponding GIII-sons sorted by GII sire. Red dots: Average normalized distance between CO events for all homologues with two recombination events transmitted by GII-sire to its GIII-sons.*

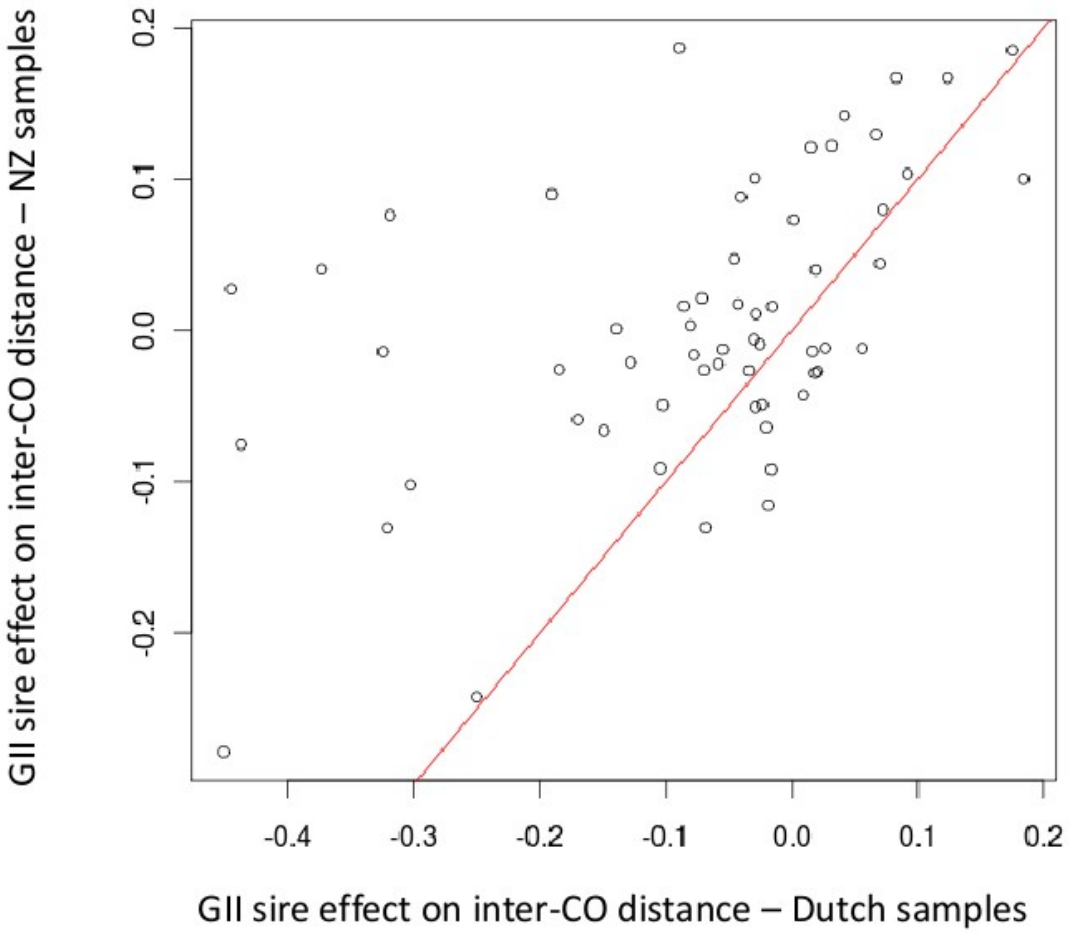


Figure V.14: *Correlation between average normalized distance between CO events for all homologues with two recombination events transmitted by 72 shared GII-sire to (i) their Dutch GIII-sons, and (ii) their NZ GIII-sons.*

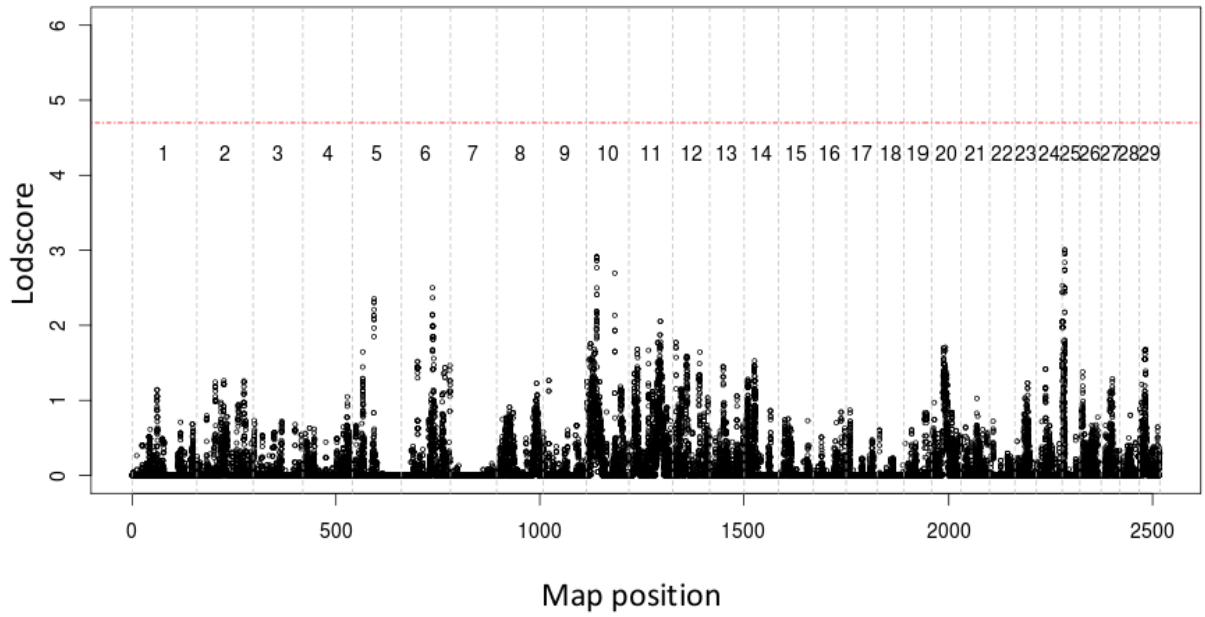


Figure V.15: *Results of genome-scan for QTL affecting the normalized distance between pairs of CO events, using a method that simultaneously extracts linkage and LD signal*[192] *The red horizontal lines marks the genome-wide threshold for significance (5%) determined by permutation testing.*

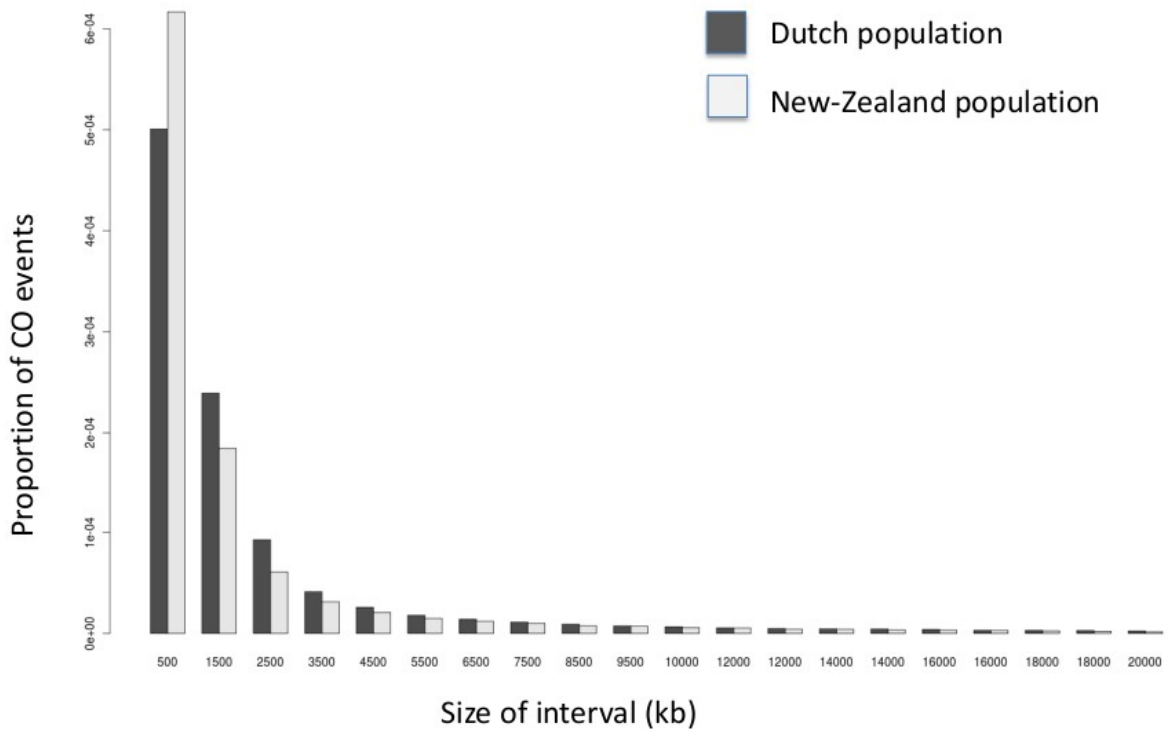
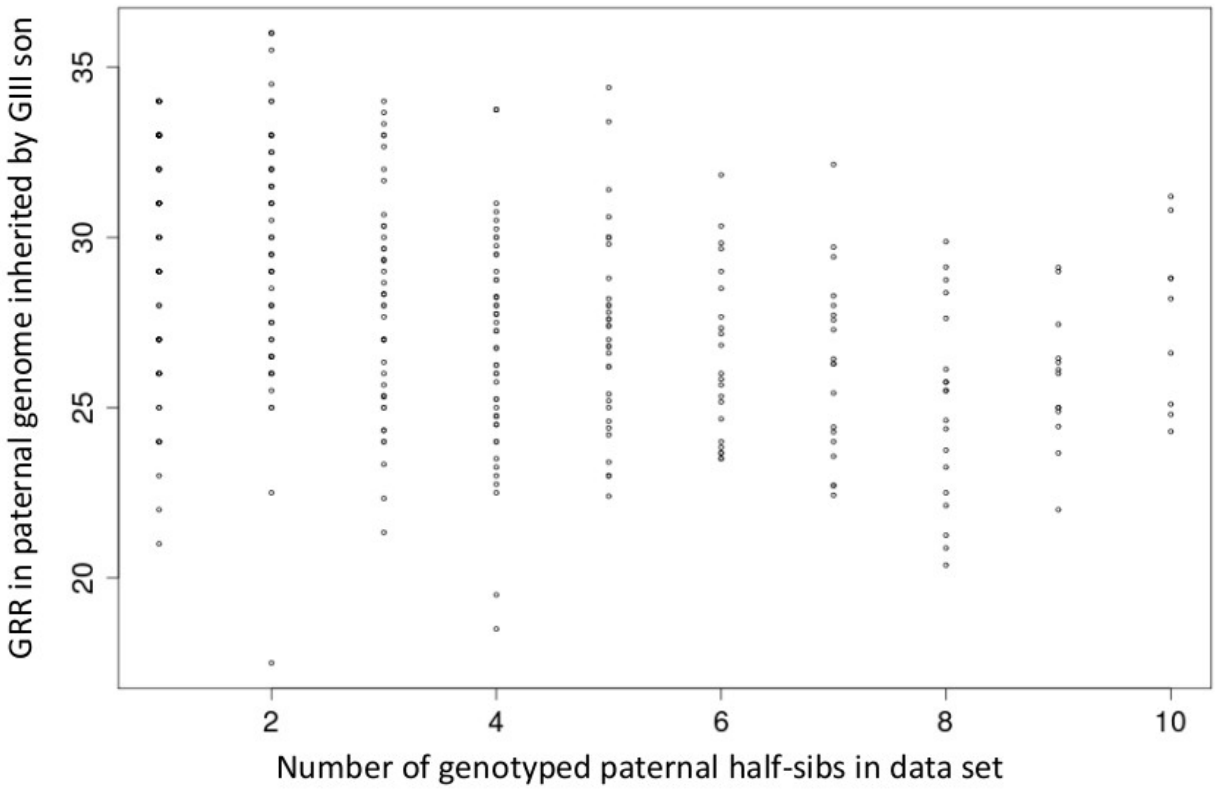
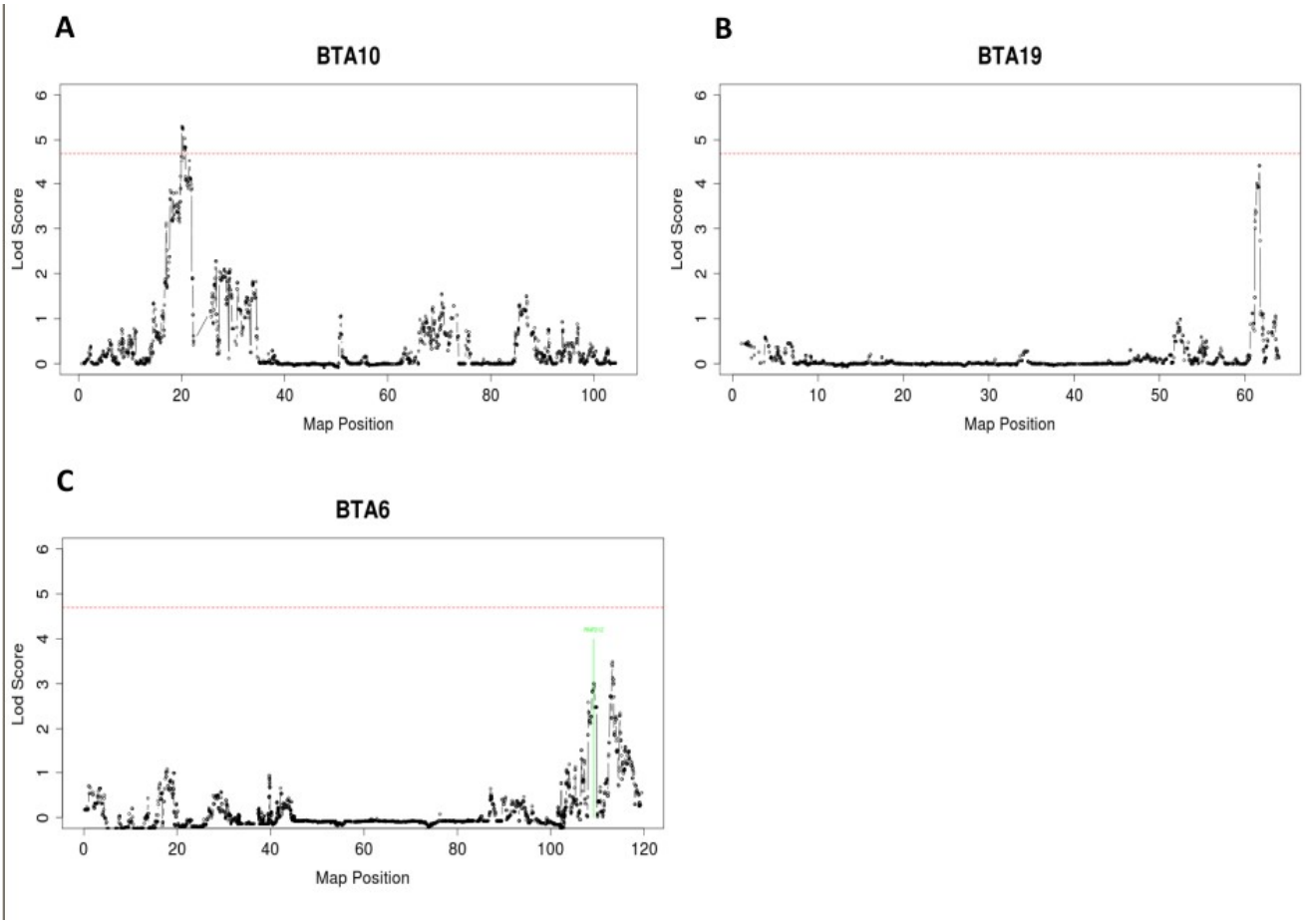


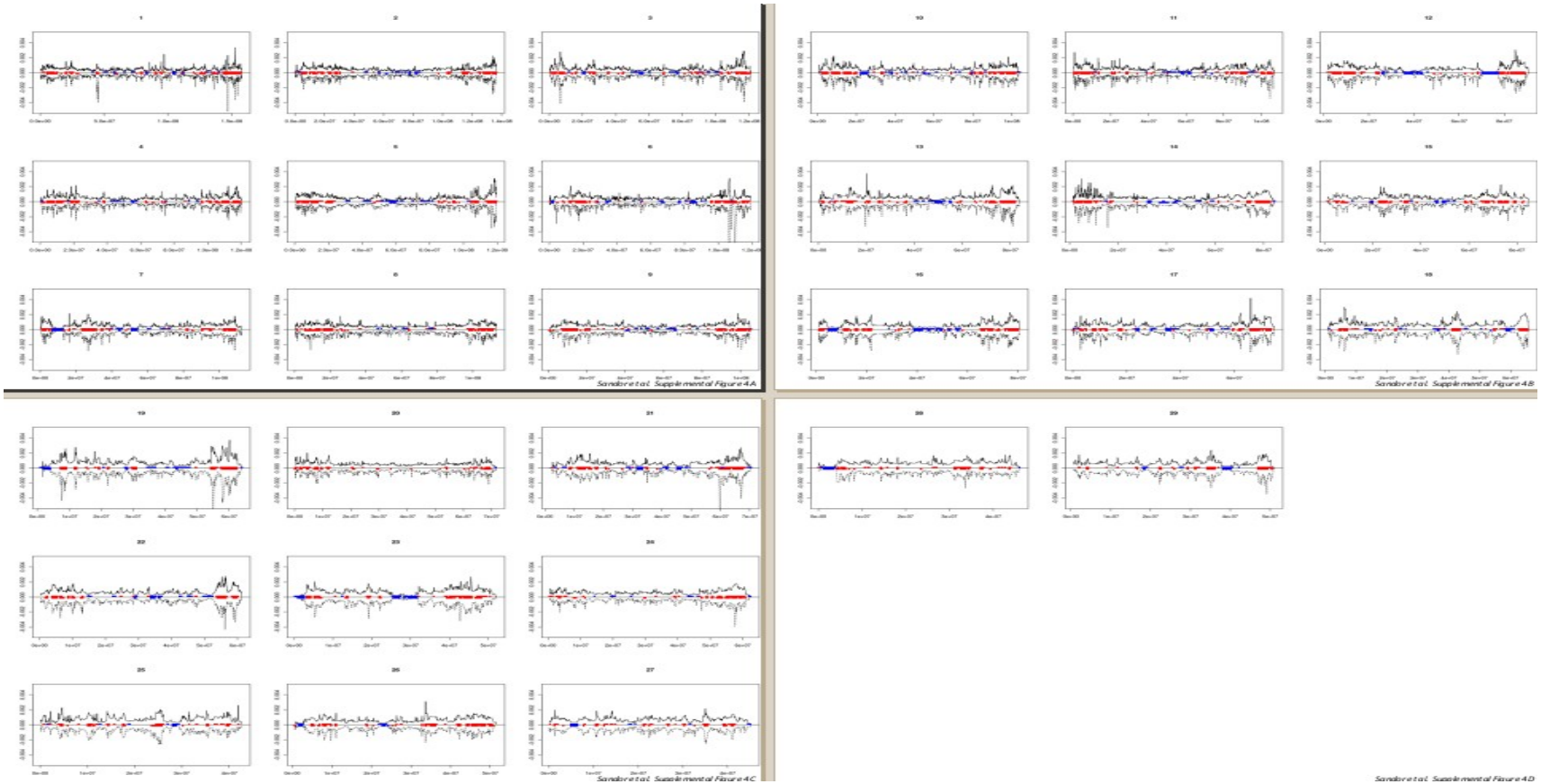
Figure Sup V.1: *Frequency distribution of the size (in Kb) of the marker intervals to which CO events were mapped (i.e. distance separating the closest fully informative markers flanking the obligated CO event), in the Dutch and New-Zealand samples, respectively.*



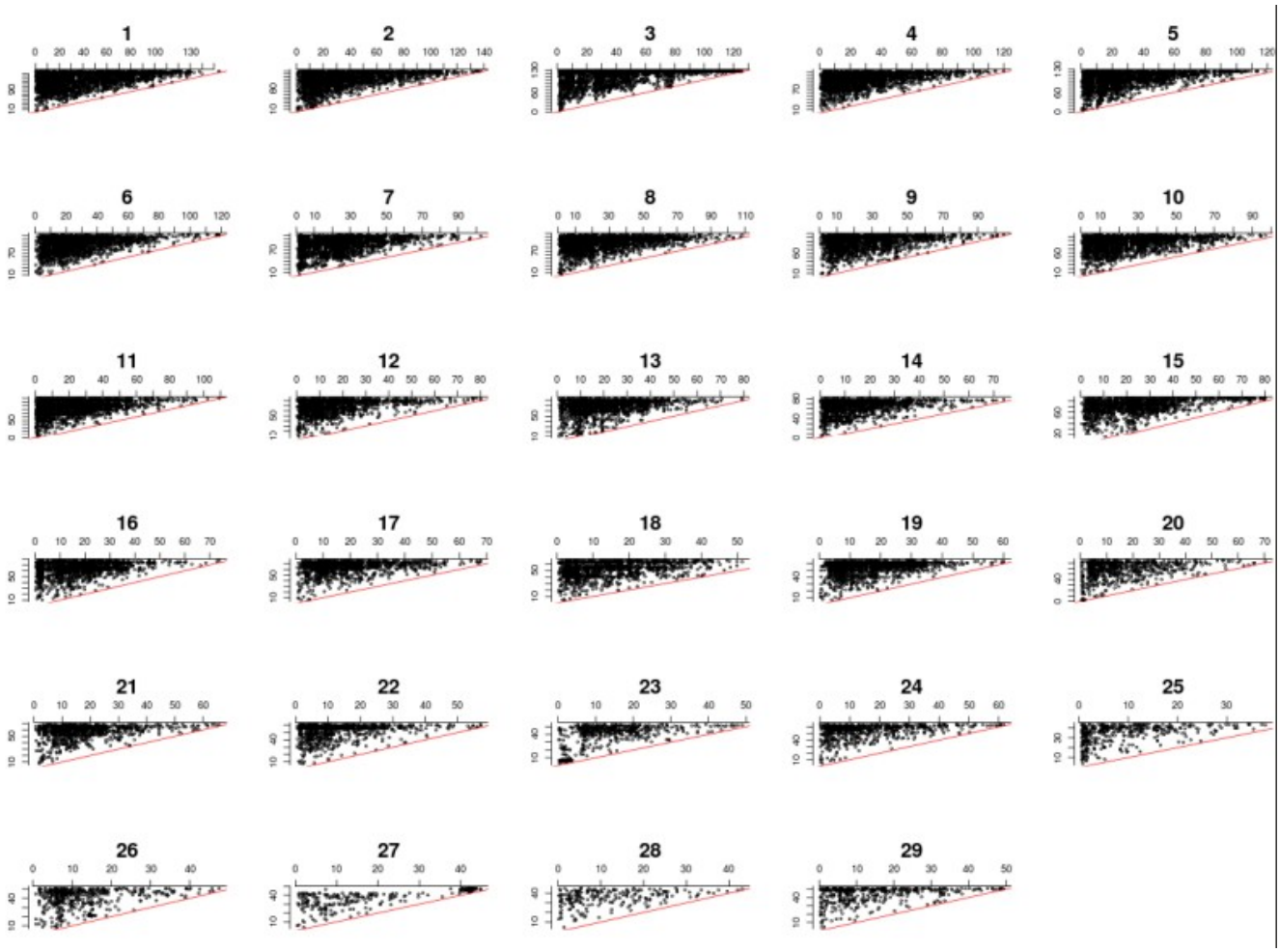
Figure_Sup V.2: *Total number of CO events (GRR) in the genome transmitted by GII sires to their GIII sons. GIII sons are sorted according to the number of half-brothers in the data set. The increase of GRR with decreasing family size is clearly visible.*



Figure_Sup V.3: *Lod score profiles for three chromosomes with significant (BTA9) or suggestive (BTA19, BTA6) QTL for GRR. The red horizontal lines marks the threshold for genome-wide significance (5%).*



Figure_Sup V.4: *Variation in male recombination in 60Kb windows across the bovine genome. The plain black line (upper halve) corresponds to recombination rate estimated in the Dutch population, while the dotted black line (lower halve) corresponds to the recombination rate estimated in the NZ population. The red and blue horizontal lines correspond to recombination "jungles" and "deserts", respectively, i.e. segments in which the observed recombination rate deviates by more than 2.5 standard deviations from the local recombination rate expected under a model of uniform distribution of CO events.*



Figure_Sup V.5: *"Folded" positions of CO events for paternal homologues with two recombinations. The depletion of datapoints along the diagonal is indicative of positive interference and is observed for all 29 autosomes.*

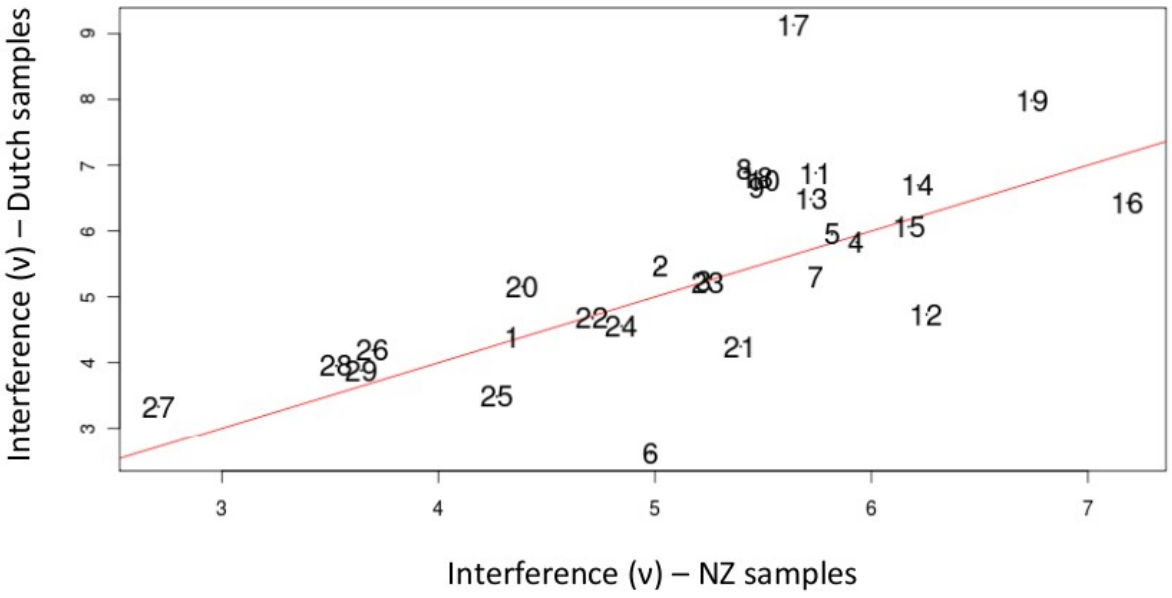
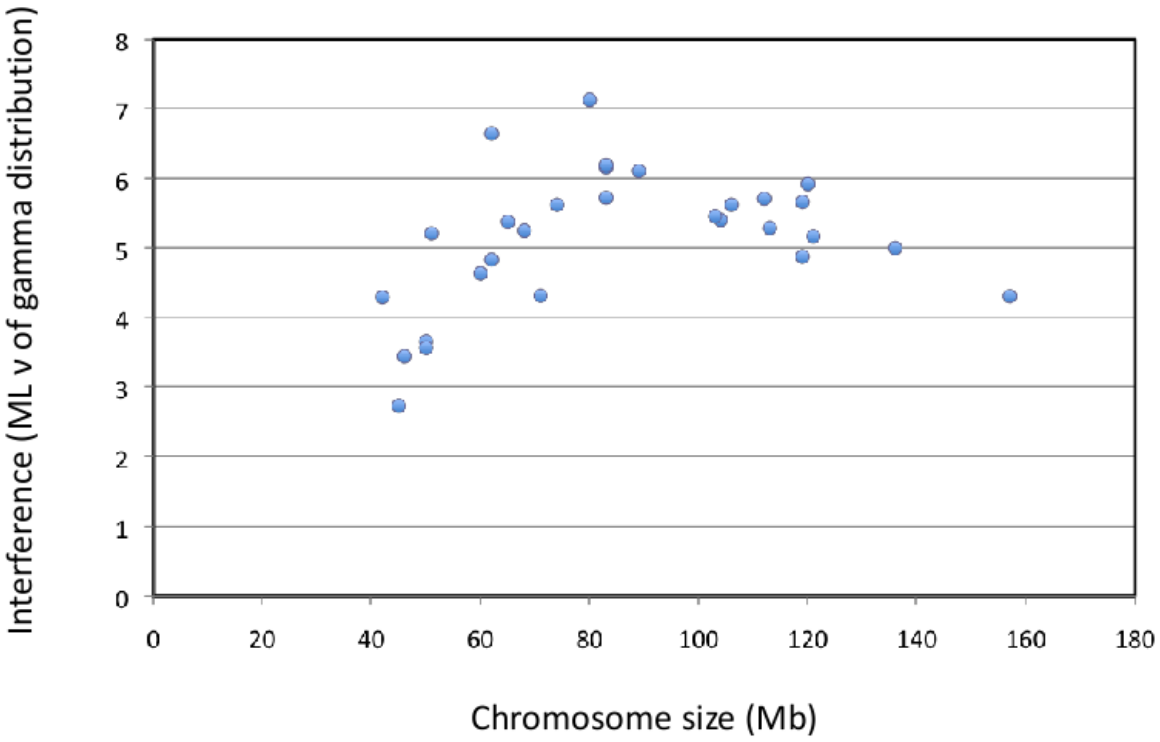


Figure Sup V.6: *Correlation between chromosome-specific estimates of v (gamma model) computed respectively in the Dutch and NZ samples.*



Figure_Sup V.7: *Relationship between estimates of v (gamma model) and chromosomze size in Mb.*

VI Discussion.

VI.1 Connaissances actuelles des caractères complexes.

Les différentes études de cartographie QTL menées chez les animaux de productions, ainsi que les multiples études d'association génome entier portant sur des maladies complexes humaines ont contribué à fournir une première ébauche des facteurs génétiques sous-tendant ces caractères complexes. Que ce soit des QTL affectant des caractères agronomiques ou des nouveaux *loci* à risque pour des maladies complexes humaines, plusieurs conclusions peuvent être d'or et déjà tirées.

VI.1.1 Conclusions d'un point de vue statistique.

Les facteurs génétiques impliqués dans des caractères complexes expliquent une faible part de la variation génétique. Par exemple, dans le cas de la maladie de Crohn, on estime que pris ensemble les nouveaux *loci* à risque n'expliquent pas plus de 10% de la variation génétique totale. La question qui se pose c'est d'où viennent les 90 % restant. Plusieurs raisons sont avancées.

(i) L'une d'elles est que la plupart des études d'association GWA menées jusqu'à présent se sont concentrées sur un certain type de polymorphismes: des variations généiques courantes (MAF > 5%). Les effets associés à ces polymorphismes sont relativement modestes, dans la majorité des cas le risque augmente de 1.1 à 1.5 avec l'allèle de susceptibilité. Or il est tout à fait probable que des SNPs avec un MAF plus faible, mais avec des effets plus importants sur le phénotype étudié expliquent tout autant de la variation génétique totale que ces variations génétiques courantes avec MAF plus élevée. Par ailleurs, un caractère complexe pourrait être affecté de la même façon qu'un caractère mendélien par une mutation génétique rare. Toutefois étant donné leur nombre et leur rareté, il n'est pas possible de les énumérer dans les populations comme les variations génétiques courantes l'avaient été par le passé avec le projet Hapmap [38]. La seule manière de les identifier est de séquencer les cas et les contrôles.

(ii) Une autre hypothèse avancée est que d'autres types de polymorphismes comme des modifications structurales sont peut-être impliquées dans des caractères complexes. Les raisons pour lesquelles ces polymorphismes structuraux peuvent affecter des caractères complexes ont déjà été évoquées plus haut dans

l'introduction dans la section concernant les CNV.

(iii) Une autre raison invoquée pour expliquer la faible contribution des polymorphismes identifiés dans les premières études GWA est la faible puissance statistique de ces analyses. Cette faiblesse est illustrée, par le fait qu'en réalisant des méta-analyses, c'est à dire en combinant des jeux de donnés de GWAS individuelles, il est possible d'identifier davantage de *loci* à risque que plusieurs GWAS, où les jeux de données sont considérés séparément. Par exemple, trois études de type GWA portant sur la maladie de Crohn avaient identifié une dizaine de *loci à* risque[63,64,58], en combinant les données de trois précédentes études dans une méta-analyse, il a été possible de mettre en évidence une trentaine de *loci* à risque pour la maladie[59].

(iv) Une autre cause possible, c'est que les effets associés à de ces nouveaux *loci* à risque sont peut-être sous estimés du fait que les SNP utilisés dans une GWAS sont en déséquilibre de liaison imparfait (r^2!=1) avec la mutation causale. De plus, une nouvelle région génomique à risque et couverte par un seul SNP peut contenir plusieurs mutations ayant un impact sur la maladie.

(v) Un autre reproche qui peut être fait aux GWAS actuelles est le fait d'avoir été pratiquement exclusivement réalisées sur des cohortes de même origine ethnique, essentiellement caucasienne. Un SNP à risque peut par exemple ne pas être détecté dans des cohortes issues de populations d'une certaine ethnie du fait de sa faible fréquence, mais pourra l'être dans une cohorte issue d'une population d'une autre ethnie ou d'une population isolée, du fait d'une fréquence plus élevée ou encore d'un effet plus élevé dû à un environnement différent de la précédente population.

(vi) Les estimations du % de la variation génétique expliquée par différents facteurs sont basées sur un certain nombre de suppositions pouvant être erronées. Ces estimations émettent l'hypothèse que les *loci* sont additifs. Elles ne tiennent pas compte des interactions gènes-environnement ou gènes-gènes. Or ces interactions pourraient avoir un impact dans le risque associé à une maladie. Il existe plusieurs exemples d'épistasie chez les organismes modèles (ex.: drosophile), toutefois il n'existe pas encore d'études chez l'homme ou chez les espèces de production montrant que l'épistasie joue un rôle dans les caractères complexes. Ceci est probablement dû au fait que les études actuelles ne sont pas assez puissantes pour évaluer toutes les interactions possibles.

VI.1.2 Conclusions d'un point de vue biologique.

Si les facteurs génétiques identifiés représentent une faible proportion de la variation génétique, des enseignements peuvent être néanmoins tirés d'un point vu biologique.

Tout d'abord, il y a peu de situations où les nouveaux *loci* génétiques identifiés se trouvent dans des gènes

connus comme étant des acteurs dans le caractère étudié. L'exception est une étude d'association génome-entier chez l'homme montrant que parmi les 19 *loci* influençant de manière significative les niveaux en lipoprotéines et triglycérides, 12 sont dans des gènes dont on connaissait déjà le rôle dans la biologie des lipides[65,66].

Certains des nouveaux *loci* génétiques identifiés comme influençant un caractère complexe se trouvent dans des gènes dont on ne soupçonnait pas qu'ils pouvaient avoir un lien avec le caractère étudié.

Cependant la plupart du temps les associations pointent vers des régions non-codantes. Une partie de ces associations peut être due au LD: un SNP de susceptibilité dans une région non-codante est peut-être en LD avec une mutation causale. Toutefois, le plus souvent les régions pointées ou leurs voisines dans des études de cartographie génétique de caractère complexe sont dépourvues de gènes. Cette observation n'est pas tellement surprenante étant donné qu'il existe seulement 5% de régions conservées et donc fonctionnelles dans le génome et que parmi ces 5%, un tiers seulement est représenté par des gènes. Plusieurs hypothèses sont avancées pour expliquer le rôle de ces régions non-codantes associées à un caractère complexe. L'une d'elles est que le niveau d'expression de certains gènes est peut-être modulé par ces polymorphismes montrant une association avec un phénotype complexe (voir chapitre traitant des eQTL).

Par ailleurs, certaines régions exhibent des associations avec plusieurs caractères complexes. Pour certains d'entre eux, on connaissait le lien d'apparenté notamment dans le cas de maladies sur base de la pathogénie (ex.: Crohn et spondylarthrite ankylosante), dans d'autres cas ces associations d'une même région avec différents caractères s'est avéré plus surprenantes (ex.: diabète de type II avec la Maladie de Crohn et le Psoriasis [67]).

VI.2 Voies à suivre pour améliorer nos connaissances sur les caractères complexes.

Pour identifier de nouveaux facteurs génétiques impliqués dans des caractères complexes, des efforts seront nécessaires dans deux domaines: (i) dans le design des études de cartographie génétique et (ii) dans les outils génétiques.

VI.2.1 Efforts dans le design des études génétiques.

Les études de cartographie actuelles ne sont pas assez puissantes pour détecter des *loci* même associés à des effets moyens. Il sera donc nécessaire d'étendre les cohortes afin d'améliorer la puissance de détection. On sait par exemple qu'une étude d'association menée avec 1000 cas et 1000 contrôles aura une puissance de seulement 1% pour détecter un allèle de susceptibilité associé à facteur de risque de 1.3 et dont la fréquence dans la population est de 0.2. En étendant une telle étude à 5000 cas et 5000 contrôles, la puissance de détection passerait à 98%. En outre, des études GWA avec des polymorphismes moins fréquents dans la population contraindront également à augmenter la taille des cohortes utilisées.

On pourra également étendre les études de cartographie génétique du point vue de l'origine des individus employés dans ces études. Toutes les études GWA, ces dernières années, se sont focalisées sur des individus issus de populations de type européenne. Or par exemple, un polymorphisme ayant le même effet sur un caractère dans les deux populations pourrait être plus facilement détectable dans la population où il est le plus fréquent.

Par ailleurs, beaucoup de caractères restent à étudier notamment des sous-phénotypes qui pourraient offrir un éclairage nouveau sur les facteurs génétiques impliqués dans le phénotype principal. Par exemple, les premières études d'association sur les maladies inflammatoires intestinales chroniques se sont concentrées sur la maladie de Crohn. Or il existe d'autres formes de maladies inflammatoires intestinales apparentées à la maladie de Crohn, par exemple la colite-ulcéro hémorragique. En outre, même au sein de patients atteints de la maladie de Crohn, il existe différentes formes de maladies: on classifie les malades par exemple selon la partie de l'intestin touché ou encore selon le traitement reçu (chirurgical ou non).

La détection de vraies associations est également affectée par la précision avec laquelle on affecte à la fois un

phénotype et un génotype à un individu. Toutefois, en général, les génotypes sont déterminés avec une plus grande certitude que les phénotypes. Toutes les approches qui permettront de définir un phénotype plus précisément contribueront à améliorer la puissance de détection des études d'association.

Les facteurs environnementaux jouent un rôle très important dans les variations phénotypiques, néanmoins il est très difficile de les identifier et d'estimer les effets qui leur sont associés, d'autant plus qu'ils peuvent fluctuer au cours du temps. Cependant des améliorations dans les approches permettant de les évaluer devraient augmenter la puissance de détection des études de cartographie génétique en diminuant la variance résiduelle.

VI.2.2 Développement de nouveaux outils génétiques.

Un des principaux reproches qui peut être émis à l'égard des études de cartographie génétique actuelles est d'utiliser des outils génétiques qui ciblent un seul type de polymorphisme: des variations génétiques fréquentes dans la population. Pour identifier de nouveaux facteurs génétiques affectant des caractères complexes, les outils génétiques devront être étendus à d'autres types de polymorphismes, notamment à des variations génétiques peu fréquentes dans la population ou encore à des polymorphismes structuraux. Si l'on veut étendre des études d'association génome-entier à ce type de polymorphismes, il faudra créer des catalogues de ces variations génomiques et caractériser les niveaux de LD entre eux. Cela devrait être réalisable avec l'essor des technologies de séquençage à haut de débit, lesquelles permettront d'identifier des polymorphismes ayant une fréquence dans la population supérieure à 1%. Le projet « 1000 Genome Projet » (http://www.1000genomes.org/) a été lancé chez l'homme à cette fin.

Toutefois, il est fort probable que ces nouveaux outils manqueront certains facteurs génétiques tels que des mutations génétiques rares. La seule solution pour identifier ces *loci* est de séquencer les cohortes étudiées. Actuellement, des outils sont en cours de développement pour capturer des mutations génétiques rares en se concentrant sur les exons. Il est fort probable qu'il sera possible d'obtenir dans le futur ce type d'information en dehors des régions codantes en séquençant complètement les cohortes. Cependant les *loci* identifiés dans des régions non codantes continueront à poser des problèmes pour la détermination de leur rôle biologique.

La possibilité d'obtenir la séquence complète pour l'ensemble des individus des cohortes devrait bouleverser les méthodes de cartographie actuelle: les approches futures devront exploiter toute l'information disponible simultanément, c'est-à-dire combiner l'information concernant toutes les variations génétiques possibles ainsi que les phénotypes et d'éventuels effets environnementaux.

VI.3 Approches de type sélection génomique pour des caractères complexes humains ou agronomiques.

Comme indiqué plus haut un des problèmes principaux des études GWA actuelles est leur faiblesse en terme de puissance de détection: une proportion importante des *loci* à risque ont vraisemblablement des effets (« risque relatif ») trop faibles pour dépasser les seuils de significations très sévères requis par le grand nombre de tests statistiques effectués lors de GWA. Sous ces seuils, les vraies hypothèses alternatives et zéros « cohabitent » sans qu'on puisse les distinguer.

Ce même problème est rencontré en production animale où il a été contourné par une approche dite de « sélection génomique ». Celle-ci a pour but de prendre en compte les signaux d'association sur l'entièreté du génome, indépendamment des seuils de signification associés, et de les intégrer en une prédiction la plus précise possible de la valeur d'élevage d'un individu[68]. L'objectif passe donc de l'identification la plus précise possible de *loci* (« QTL ») individuels, à la prédiction la plus précise possible d'une valeur d'élevage individuelle globale sans nécessairement savoir exactement quels sont les *loci* qui y contribuent mais en intégrant plutôt de façon pondérée sur l' ensemble de possibilités.

Or il serait tout à fait imaginable d'adapter ce type d'approche à des maladies complexes humaines et de déterminer à partir d'une GWAS, un « risque relatif génome entier » (GWRR). Les principales modifications à apporter par rapport à l'approche initiale seraient: (i) d'adapter ce type d'approche développé pour des phénotype continus à des phénotypes méristiques (ii) En génétique, on dispose d'une information d'ascendance très importante sur chaque individu, ce qui permet de mettre des effets polygéniques et de se prémunir d'éventuels effets de stratification. Il faudra dans le cas de population humain trouver un moyen d'inclure dans les modèles des effets permettant d'éviter les problèmes de stratification. (iii) Enfin, il faudra extrapoler l'impact des résultats à une population non-biaisée (« ascertainment biais »). Les GWAS portant sur des maladies complexes humaines sont des études de types rétrospectives, dans lesquelles on a généralement 50% de cas et 50% d'individus contrôles. Ceci ne correspond pas forcément à la fréquence de la maladie dans un échantillon aléatoire. Une réflexion s'impose donc quand à la spécificité et la sensibilité de l'approche diagnostique proposée dans des conditions plus proches de la réalité.

VII Résumé:

VII.1 Description du sujet de recherche abordé.

La plupart des caractères ayant un intérêt médical ou agronomique sont dits complexes. Ce qui signifie qu'ils sont influencés par plusieurs gènes, des facteurs environnementaux et des interactions gènes-environnement. L'identification de gènes affectant de tels caractères est un des sujets les plus importants de la génétique moderne: d'un point vu médical, cela offrirait de nouvelles perspectives, tant d'ordre diagnostic que thérapeutique, tandis que sur un plan agronomique, cela ouvrirait de nouvelles voies dans la sélection et la manipulation des animaux domestiques.

L'approche la plus efficace pour identifier des gènes impliqués dans des caractères complexes est le clonage positionnel. Cette approche comporte trois étapes: (i) identifier les régions génomiques contenant les facteurs génétiques impliqués dans les caractères étudiés, (ii) identifier les mutations causales (iii) et enfin étudier le fonctionnement cellulaire et moléculaire des gènes responsables.

La première étape du clonage positionnel, appelé cartographie génétique consiste à regarder au sein d'un groupe d'individus, s'il existe une corrélation entre l'histoire des différents chromosomes et celle du caractère étudié. Les outils génétiques permettant de suivre l'état d'un chromosome chez un individu, ont connu ces dernières années des progrès remarquables, notamment grâce à l'apparition des plates-formes de génotypage à haut débit et au développement de nouveaux marqueurs génétiques. Actuellement, la principale difficulté en cartographie génétique réside dans le choix de la meilleure stratégie pour détecter des variations génétiques affectant des caractères complexes, qui le plus souvent sont associés à des effets relativement modestes. Le choix d'une approche dépendra: (i) du caractère étudié – caractère discret ou continu (ii) du dataset – les individus sont apparentés ou non et (iii) des outils génétiques disponibles.

Les travaux réalisés au cours de cette thèse s'inscrivent précisément dans cette problématique, c'est à dire développer des approches statistiques afin d'identifier des gènes affectant des caractères complexes ayant un intérêt médical ou agronomique.

VII.2 Résultats.

La plupart des études de cartographie QTL (*Quantitative Trait Loci*, loci influençant un caractère continu), dans des populations de bovins laitiers, exploitent la structure en GDD (*Grand Daughter Design*) des pedigrees et procèdent en regardant au sein de familles de demi-frères paternelles, s'il existe des différences phénotypiques entre les taureaux en fonction de l'homologue reçu. Étant donné que ce type d'approche exploite la transmission de chromosome de taureau à taureau et que le chromosome X est d'origine maternelle chez les mâles, la cartographie de QTL sur le chromosome X dans ce type d'espèce a longtemps été exclue. Pour cartographier des QTL sur le X, nous avons proposé une approche retraçant l'histoire des différents segments chromosomiques dans notre échantillon, sur base de son mode de transmission et de la structure de la population. Cette approche suppose que s'il existe un QTL au niveau d'une région donnée, deux individus ayant reçu le même segment chromosomique se ressembleront davantage que deux individus ayant reçu deux segments chromosomiques différents. Cette approche suppose également qu'il existe une corrélation, appelée déséquilibre de liaison (DL) entre les allèles QTL et allèles marqueurs. Afin d'évaluer l'intérêt d'une telle approche, nous avons caractérisé les niveaux de DL sur le X. Nous avons montré que le X exhibait dans ce type de population des niveaux de DL particulièrement élevé et inattendu. Parmi les 48 caractères laitiers étudiés, nous avons trouvé en utilisant une méthode de type maximum de vraisemblance restreint (REML, *Residual maximum likelihood estimation*) 5 QTL significatifs sur le X.

Au cours de ces dernières années, le nombre de publications chez les espèces de productions mettant en évidence de l'empreinte parentale (y compris chez les oiseaux) comme étant associé aux QTL découverts (*imprinted QTL*, l'effet d'un allèle QTL dépendra de son origine parentale) n'a cessé de croitre. Ces résultats contredisent ceux de la biologie moléculaire, qui montrent que l'empreinte parentale est un phénomène rare et uniquement observé chez les mammifères placentaires. Une précédente étude, pointe le problème d'ordre statistique soulevé par toutes ces études détectant de l'empreinte parentale de façon quasi systématique. Ces études emploient un design de type *line-cross* qui suppose, pour cartographier des QTL et tester une hypothèse d'empreinte parentale, d'un coté que les lignées parentales sont fixées pour les allèles QTL et de l'autre qu'elles peuvent ségréger pour différents allèles marqueurs. Si cette hypothèse est fausse et que les lignées parentales ne sont pas fixées pour les allèles QTL, tous les individus en F1 ne sont pas hétérozygotes ou pas hétérozygotes pour les mêmes allèles QTL. Si le nombre de parents en F1 est restreint (cas typique du coté paternel), il sera possible d'avoir un effet de substitution allélique qui dépendra de l'origine parentale en F2 et de conclure erronément à de l'empreinte parentale. Dans notre étude, nous avons montré que ce problème pouvait être exacerbé de 40 à 80% en cas de DL. Pour tester une hypothèse d'empreinte parentale, il faut que les parents en F1 soient hétérozygotes pour des

allèles marqueurs différents. En cas de DL la probabilité que des allèles marqueurs différents soient associés à des allèles QTL augmente et la détection de fausse empreinte parentale également.

Depuis 2007, le nombre de loci à risque associés à des maladies complexes humaines et découverts dans des études d'association génome-entier (GWAS = *Genome Wide Association Study* n'a cessé de croire. Beaucoup de ces loci tombent dans des régions non-codantes et une hypothèse avancée pour expliquer leur rôle biologique est qu'ils moduleraient le niveau d'expression de certains gènes à travers des éléments cis. Des investigations combinant des études d'expression sur un grand nombre de gènes avec des études GWA ont été mises en œuvre afin de répertorier dans des bases de données les effets trans et cis de polymorphismes (appelés eQTL pour *Expression Quantitative Trait Locus*) sur le niveau d'expression de ces gènes. En utilisant ces bases de données d'eQTL, nous avons pu découvrir: (i) des SNP à risque pour la maladie de Crohn et présent dans une région dépourvue de gène, régulaient probablement le niveau d'expression du gène PTGER4 (*protaglandin E receptor 4*), codant pour un récepteur à une prostaglandine et candidat sérieux à ce type désordre. (ii) Parmi les 39 SNPs à risque dans la maladie, on observe 5 effets de type cis eQTL, non dus au hasard. Ces effets eQTL ouvrent de nouvelles perspectives dans l'architecture génétique d'une maladie complexe: la maladie de Crohn.

L'intensité des recombinaisons ainsi que leur position dans le génome sont dictées par le rôle fondamental que joue la recombinaison dans la ségrégation correcte des chromosomes lors de la première division méiotique. Néanmoins, on observe, des différences entre individus de même sexe et de même âge aussi bien dans l'intensité que dans la position des recombinaisons. Une idée pour explorer les causes génétiques sous-tendant ces variations est d'appliquer des méthodes de cartographie QTL classique à la recombinaison elle-même, qui sera traitée comme un phénotype quantitatif. Nous avons réalisé ce type d'étude à différente échelle, sur une population de taureaux laitiers génotypés pour des milliers de marqueurs de type SNP, en exploitant le fait: (i) qu'un grand nombre d'entre eux disposent d'un nombre suffisant de descendants pour estimer précisément leur taux de recombinaison, (ii) qu'ils appartiennent à des familles de demi-frères paternelles, pour employer des méthodes de cartographie QTL exploitant le DL et la liaison génétique. Dans cette étude, nous caractérisons préalablement les niveaux de recombinaison sur différentes échelles: (i) distribution du phénotype, (ii) répétabilité du caractère (iii) suivi d'une étude d'héritabilité. Nous montrons que plusieurs QTL affectent de manière significative les taux de recombinaisons et cela à différentes échelles.

VII.3 Conclusions et Perspectives.

Plusieurs conclusions peuvent être tirées de ces études de cartographie génétiques tentant d'identifier des gènes influençant des caractères complexes d'intérêt agronomique ou médical.

D'un point vu statistique, on montre que dans la plupart des cas les loci identifiés représentent une faible part de la variation génétique totale (10-15%). Plusieurs hypothèses sont avancées pour expliquer la variation génétique restante. (i) La première est que les études actuelles ne sont pas suffisamment puissantes pour détecter des loci même associés à des effets moyens. Cette faiblesse est illustrée par le fait qu'en réalisant une GWAS, dans laquelle on regroupe les données de GWAS individuelles, on augmente considérablement la puissance de détection de loci associés au caractère étudié. (ii) Une autre hypothèse est que les études de cartographie actuelles utilisent des outils génétiques qui ciblent un seul type de polymorphismes: des variations génétiques fréquentes dans la population. Or un caractère complexe peut très bien être influencé par des variations génétiques peu fréquentes dans la population, voire des mutations rares ou encore des polymorphismes structuraux (p.e: CNV: *Copy Number Variant*). (iii) La plupart des études actuelles ne recherchent que des effets de type additif. Il est fort probable qu'il existe des effets de type gène-gène, appelés épistasie, qui affectent des caractères complexes. Cependant, détecter de tels effets nécessite de mettre en œuvre des études beaucoup plus puissantes que celles existantes actuellement.

D'un point vu biologique, si certains loci détectés se trouvent dans des gènes dont on connait le rôle dans le caractère étudié, beaucoup d'entre eux se trouvent dans des régions non-codantes. Ce résultat n'a rien de surprenant quand on sait que seulement 5 % du génome est conservé et donc fonctionnel et que parmi ces 5%, un tiers correspond à des gènes. Toutefois, comprendre le rôle biologique de ces loci dans des caractères complexes est un challenge. Une hypothèse avancée est que le niveau d'expression de certains gènes est peut-être régulé par ces polymorphismes influençant des caractères complexes.

Pour améliorer nos connaissances sur les caractères complexes, il sera nécessaire, au cours des prochaines années:

(i) d'étendre le design des études génétiques afin d'en augmenter leur puissance de détection. Ceci passera notamment par: (1) une augmentation de la taille des échantillons, (2) étudier le même phénotype dans des populations ayant une origine différente (un polymorphisme, ayant le même effet dans deux populations différentes peut être plus facilement détecté dans la population où il est le plus fréquent), (3) améliorer la précision de l'estimation des phénotypes (4) étudier des phénotypes apparentés ou des sous-phénotypes (p.e. Crohn et les colites ulcéro-hémorragiques) (5) s'intéresser davantage aux facteurs environnementaux.

(ii) Il faudra également étendre la palette des outils génétiques disponibles, pour rechercher des polymorphismes peu fréquents ou des polymorphismes structuraux pouvant affecter des caractères complexes. Ceci devrait être prochainement réalisable grâce à l'essor des technologies de séquençage à haut débit qui devrait permettre de cataloguer des polymorphismes avec une fréquence > 1% dans une population (1000 Genome Project).

On peut penser également qu'il sera possible dans quelques années de séquencer complètement tous les individus d'une étude de cartographie et d'identifier ainsi des mutations génétiques rares. Néanmoins, ces nouveaux outils génétiques bouleverseront les méthodes de cartographie génétique actuelles. Les approches futures devront exploiter toute l'information disponible simultanément, c'est-à-dire combiner l'information concernant toutes les variations génétiques possibles ainsi que les phénotypes et d'éventuels effets environnementaux.

En génétique animale, le problème des loci influençant un caractère complexe et écartés par manque de puissance statistique a été contourné par une approche dite de sélection génomique. Celle-ci a pour but de prendre en compte les signaux d'association sur l'entièreté du génome, indépendamment des seuils de signification associés, et de les intégrer en une prédiction la plus précise possible de la valeur d'élevage d'un individu. L'objectif passe donc de l'identification la plus précise possible de loci individuels à la prédiction la plus précise possible d'une valeur d'élevage individuelle globale, sans nécessairement savoir exactement quels sont les loci qui y contribuent, mais en intégrant plutôt de façon pondérée sur l'ensemble des possibilités.

Or il serait tout à fait imaginable d'adapter ce type d'approche à des maladies complexes humaines et de déterminer à partir d'une GWAS un « risque relatif génome entier » (GWRR) pour chaque individu.

VIII Summary.

VIII.1 Research topic description.

The most important medical or agronomical phenotypes are “complex traits”. This means that they are influenced by multiple genes, environmental factors and genes-environment interactions. The identification of genes affecting « complex traits » is a major topic of modern genetic: in the medical field this could provide new diagnostic and therapeutic perspectives, while from an agronomical point-view, this could open new ways in the selection and the manipulation of domestic animals.

The most efficient approach to identify the genes involved in “complex traits” is positional cloning. This approach includes three steps: (i) identifying the genomic regions containing the genetics factors involved in the phenotypes studied, (ii) identifying the causal mutations, and finally, (iii) studying the cellular and molecular function of genes involved.

The first step of positional cloning, genetic mapping, consist of looking in a group of individuals to find out if there is a correlation between the history of different chromosomes and the history of the studied phenotype. The genetic tools allowing us to apprehend the state of a chromosome in an individual have made remarkable progress in recent years, thanks especially to the emergence of high-throughput genotyping platforms and the development of new genetic markers. Currently, the main difficulty encountered in genetic mapping is the selection of a strategy best suited for detecting genetic variant often associated with weak effects in the case of complex traits. The choice of an approach will depend on: (i) the phenotype studied – be it continuous or discreet, (ii) the dataset – related or unrelated individuals and (iii) the genetic tools available. To develop new approaches suited to the multiple context in genetics mapping has become an essential need in recent years.

The works carried out in this thesis are taking part precisely in this problematic, which is to develop statistical approaches to identify genes affecting complex trait presenting either a medical or an agronomical interest.

VIII.2 Results.

Most of the QTL (« Quantitative Trait Loci », loci influencing a continous phenotype) mapping studies exploit the GDD (Grand Daughter Design) structure of pedigree, and seek if there are significant differences between

the phenotypic means of half-sibs sorted according to the homolog inherited . Since this type of approach exploits the transmission of chromosome from bull to bull and the X chromosome is of maternal origin in males, the QTL mapping on the X chromosome in this type of species had long been excluded. To map QTL on the X in this context, we proposed an approach retracing the history of chromosomal segments in our sample using the effective population size and the specific transmission mode of sexual chromosomes. This approach assume that if there is a QTL at a genomic position, two individuals, who received the same chromosomal segment, would look more alike than two individuals who received two different chromosomal segments. This approach also assume that there is a correlation, called linkage disequilibrium (LD), between the QTL alleles and the markers alleles. To evaluate the interest of such an approach, we characterized the LD on the X chromosome. We showed that, in these populations, the X chromosome exhibited unexpected high levels of LD. Among the 48 dairy characters studied, we found, using a Residual maximum likelihood estimation (REML) approach, 5 significant QTL on the X chromosome.

In recent years, the number of publications in livestock species reporting imprinted QTL (imprinted QTL, the effect of an allele QTL depends on its parental origin), including birds, hasn't stopped growing. These results contradict those of molecular biology, showing that imprinting is a rare phenomenon and is observed only in the placental mammals. A previous study highlight the statistical problem raised by all these academic work detecting systematically imprinted QTL: they all use a “line-cross” design that assumes that, for mapping imprinted QTL and testing for imprinting, (a) the parental lines are fixed for QTL alleles and (b) they can segregate for different marker alleles. If this hypothesis is false and the parental lines are not fixed for the QTL alleles, then all F1 individuals are not heterozygous or not heterozygous for the same QTL alleles. If the number of F1 parents is small, (which is typically the case for males), it would be possible to have an allelic substitution effect depending on the parental origin in F2 and to conclude erroneously to imprinted QTL. In our study, we demonstrated that if linkage disequilibrium exists between marker loci and nonfixed QTL, spurious detection of pseudo-imprinting is increased by an additional 40–80% in scenarios mimicking typical livestock situations. This is due to fact to test for an imprinting hypothesis in a F2 pedigree, it is necessary that the F1 parents be heterozygous for different markers alleles. In case of LD, the probability that different markers alleles are associated with different QTL alleles increases and the detection of false imprinting too.

Since 2007, the number of risk loci associated with complex human diseases and discovered by Genome Wide Association studies (GWAS) hasn't stopped growning. The great majority of these loci fall in no-coding regions. An hypothesis for explaining their biological role is that they could modulate the level of expression of some genes by cis elements. Investigations combining expression studies on a large number of genes with GWAS have been performed to identify and catalog (in databases) the cis and trans effects of polymorphisms (named eQTL,

Expression Quantitative Trait Locus) on the expression level of these genes. Using such databases, we discovered: (i) a some number of risk SNPs for the Crohn disease (CD), present in a desert genetic region, and likely regulating the expression level of a gene PTGR4 (protaglandin E receptor 4) coding for a prostaglandin receptor and serious candidate for this type of disorder. (ii) Among the 39 risk SNPs discovered in a second GWAS on the (CD), 5 non-random cis eQTL effects were observed. These eQTL open new perspectives in the genetic architecture of a complex disease: the Crohn's disease.

The intensity of the recombinations as well as their position is governed by the fundamental role played by the recombination in the correct segregation of chromosomes during the first meiotic division. However, differences were observed between individuals of the same sex and age both in the intensity and the positions of recombinations. To explore the genetics causes underlying these variations, it is possible to apply classical QTL mapping methods to recombination itself, which is then considered as a quantitative phenotype. We performed this type of study on different scales of recombination, using a population of dairy cattle bulls genotyped for thousands of SNP markers, exploiting the fact that: (i) a large number of bulls had many descendants, allowing us to estimate the recombination rates accurately. (ii) They belonged to half-sibs families, this allowed us to exploit both the genetic linkage and the linkage disequilibrium. In this study, we first characterized the levels of recombination on different scales: distribution, repeatability and heritability of phenotype. We then performed the QTL analysis and we found several QTL significantly affecting the recombination rates at different scales.

VIII.3 Conclusions and Perspectives.

Several conclusions can be drawn from the current genetic mapping studies attempting to identify genes influencing complex traits.

From a statistical point of view, it was shown that most of the identified loci represent a small proportion of the genetic variance (max 10-15%). Several hypothesis have been advanced to explain the remaining genetic variation. (I) The first hypothesis is that the current studies are not powerful enough to detect loci even associated with average effects. This weakness is illustrated by the fact that performing a GWA study combining dataset provided by individual GWAs can increase the statistical power considerably. (ii) Another hypothesis is that the current mapping studies use genetic tools targeting a single type of polymorphisms: common genetic variants. A complex trait could very well be influenced by less frequent genetic variations in the population or even by rare mutations or structural polymorphism such as a Copy Number Variant (CNV). (iii) Most of the current studies look only for additive genetic effects. It is likely that gene-gene effect, called epistasis, could affect the complex traits. However, detecting such effects would require us to work with more powerful studies

than with the ones we are currently using.

From a biological point of view, even if some risk loci are detected in known genes, the vast majority of them are detected in non-coding regions. This is not surprising since only 5% of the genome is conserved and thus functional. Among these 5 % only a third corresponds to genes. However to be able to understand the role of these new risk polymorphism remains a challenge. An hypothesis is that some of these polymorphism could modulate the expression level of some genes.

Improving our understanding of complex traits will necessitate that over the coming years:

(i) we extend the design of genetic studies to in order to increase their statistical power. This will involve (1) an increase in the size of our samples, (2) the study of the same phenotypes in populations having different origins (a polymorphism with the same effect in two populations can be more easily detected in the population where it is more frequent. (3) an improvement in the accuracy of phenotypes estimations (4) the study of related phenotypes (ie Crohn and Ulcerative colitis) or sub-phenotypes (5) an increased interest on environmental factors.

(ii) we broaden the range of genetics tools available to look for uncommon or structural polymorphism. This should soon be feasible thanks to the advances in high throughput sequencing technologies that will allow for the identification of polymorphisms with a frequency > 1% in a population (1000 genome project). Presumably it will be also possible in a few years to completely sequence all the individuals in a mapping study and thus identify rare genetic mutations. However, these new genetics tools will revolutionize the current methods of genetic mapping and future approaches will need to exploit all available information simultaneously, in extenso, to combine information from all possible genetic variations as well as phenotypes and possible environmental effects.

In animal genetics, the problem of loci with weak effect influencing a complex trait and removed due to a lack of statistical power has been circumvented by a genomic selection approach. The goal of this approach is to account for all of the association signals on the whole genome, independently of associated significant thresholds, and to integrate these signal in the most accurate prediction possible of the breeding value of an individual. The objective thus shift from the most accurate prediction possible of individual loci to the most accurate prediction possible of the global individual breeding value, without necessarily knowing exactly which loci are contributing, but rather by integrating on all the possibilities. It would be possible to think of a similar approach tailored to complex human diseases and to determine from a GWA study the genome wide relative risk (GWRR) for each individual.

IX Bibliographie.

1. Of, R., By, H. & Of, O.D.E. Electronic Scholarly Publishing http://www.esp.org. *Scholarly Publishing* 43-59(1913).

2. Petes, T.D. & Botstein, D. Simple Mendelian inheritance of the reiterated ribosomal DNA of yeast. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5091-5(1977).

3. OMIA (Online Mendelian Inheritance in Man). at <http://www.ncbi.nlm.nih.gov/omim>

4. Welcsh, P.L. & King, M.-claire Brca1 brca2. **10**, 705-714(2001).

5. Bell, G.I. & Polonsky, K.S. Diabetes mellitus and genetically programmed defects in beta-cell function. *Nature* **414**, 788-91(2001).

6. Altenburg, E. & Muller, H.J. The Genetic Basis of Truncate Wing,-an Inconstant and Modifiable Character in Drosophila. *Genetics* **5**, 1-59(1920).

7. Paterson, A.H. et al. Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* **335**, 721-726(1988).

8. Risch, N. & Merikangas, K. pq(y+ 1)2. *Science* **273**, 0-1(1996).

9. AIRD, I. et al. The blood groups in relation to peptic ulceration and carcinoma of colon, rectum, breast, and bronchus; an association between the ABO groups and peptic ulceration. *Br Med J* **2**, 315-321(1954).

10. Klein, J. & Sato, A. The HLA system. First of two parts. *N Engl J Med* **343**, 702-709(2000).

11. Strittmatter, W.J. & Roses, A.D. Apolipoprotein E and Alzheimer's disease. *Annu Rev Neurosci* **19**, 53-77(1996).

12. Lander, E.S. The new genomics: global views of biology. *Science* **397**, (1996).

13. Chakravarti, a Population genetics--making sense out of sequence. *Nature genetics* **21**, 56-60(1999).

14. Li, W.H. & Sadler, L.A. Low nucleotide diversity in man. *Genetics* **129**, 513-523(1991).

15. Sachidanandam, R. et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-933(2001).

16. Reich, D.E. & Lander, E.S. On the allelic spectrum of human disease. *Trends Genet* **17**, 502-510(2001).

17. Harris, H. Genes and enzymes in man. *Cancer Res* **26**, 2054-2062(1966).

18. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624-626(1968).

19. King, J.L. & Jukes, T.H. Non-darwinian evolution. *Science* **164**, 788–798(1969).

20. Botstein, D. et al. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**, 314-331(1980).

21. Kerem, B. et al. Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073-1080(1989).

22. Jeffreys, A.J., Wilson, V. & Thein, S.L. Hypervariable “minisatellite” regions in human DNA. *Nature* **314**, 67-73(1985).

23. Saiki, R.K. et al. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350-1354(1985).

24. Litt, M. & Luty, J.A. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am J Hum Genet* **44**, 397-401(1989).

25. Chen, X. & Sullivan, P.F. Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput. *Pharmacogenomics J* **3**, 77-96(2003).

26. Nielsen, R. & Signorovitch, J. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theor Popul Biol* **63**, 245-255(2003).

27. Beckmann, J.S., Estivill, X. & Antonarakis, S.E. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nature reviews. Genetics* **8**, 639-46(2007).

28. Aitman, T.J. et al. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851-855(2006).

29. Fanciulli, M. et al. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat Genet* **39**, 721-723(2007).

30. Yang, Y. et al. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European America. *Am J Hum Genet* **80**, 1037-1054(2007).

31. Gonzalez, E. et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434-1440(2005).

32. Jurg Ott *Analysis of Human Genetic Linkage*. (1991).

33. Wu, R., Ma, C.-X. & Casella, G. *Statistical Genetics of Quantitative Traits*. (2007).

34. Boehnke, M., Lange, K. & Cox, D.R. Statistical methods for multipoint radiation hybrid mapping. *American journal of human genetics* **49**, 1174-88(1991).

35. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921(2001).

36. Lewontin, R.C. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics* **49**, 49-67(1964).

37. International, T. & Consortium, H. A haplotype map of the human genome. *October* **437**, 1299-1320(2005).

38. Frazer, K. a et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-61(2007).

39. Altshuler, D.M. et al. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58(2010).

40. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636-640(2004).

41. Lifton, R.P. Genetic dissection of human blood pressure variation: common pathways from rare phenotypes. *Harvey Lect* **100**, 71-101

42. Visscher, P.M., Thompson, R. & Haley, C.S. Confidence intervals in QTL mapping by

bootstrapping. *Genetics* **143**, 1013-1020(1996).

43. Elston, R.C. & Stewart, J. A general model for the genetic analysis of pedigree data. *Hum Hered* **21**, 523-542(1971).

44. Lander, E.S. & Green, P. Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S A* **84**, 2363-2367(1987).

45. Hall, J.M. et al. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**, 1684-1689(1990).

46. Duncan *Statistical Methods in Genetic Epidemiology*. (2004).

47. Jeunemaitre, X. et al., et al. Molecular basis of human hypertension: role of angiotensinogen. *Cell* **71**, 169-180(1992).

48. Pericak-Vance, M.A. et al., et al. Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. *Am J Hum Genet* **48**, 1034-1050(1991).

49. Haseman, J.K. & Elston, R.C. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* **2**, 3-19(1972).

50. Marsh, D.G. et al. Linkage analysis of IL4 and other chromosome 5q31.1 markers and total serum immunoglobulin E concentrations. *Science* **264**, 1152-1156(1994).

51. Morrison, N.A. et al. Prediction of bone density from vitamin D receptor alleles. *Nature* **367**, 284-287(1994).

52. Georges, M. *Positional identification of Quantitative Trait Loci in livestock populations*. (2005).

53. Kruglyak, L. & Lander, E.S. A nonparametric approach for mapping quantitative trait loci. *Genetics* **139**, 1421-1428(1995).

54. Coppieters, W. et al. A rank-based nonparametric method for mapping quantitative trait loci in outbred half-sib pedigrees: application to milk production in a granddaughter design. *Genetics* **149**, 1547-55(1998).

55. Henderson Animal Breeding and Genetics Symposium in Honor of Dr Jay Lush. (1973).

56. Thorisson, G. a et al. The International HapMap Project Web site. *Genome research* **15**, 1592-

3(2005).

57. Stephens, M., Smith, N.J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**, 978-989(2001).

58. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**, 629-644(2006).

59. Mellitus, D. Transmission Test for Linkage Disequilibrium: *Test* 506-516(1993).

60. Yu, J. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics* **38**, 203-8(2006).

61. Meuwissen, T.H. & Goddard, M.E. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* **155**, 421-430(2000).

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

63. Georges, M. et al., et al. Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* **139**, 907-920(1995).

64. Kühn, C. et al. Quantitative trait loci mapping of functional traits in the German Holstein cattle population. *J Dairy Sci* **86**, 360-368(2003).

65. Hiendleder, S. et al. Mapping of QTL for Body Conformation and Behavior in Cattle. *J Hered* **94**, 496-506(2003).

66. Farnir, F. et al. Extensive genome-wide linkage disequilibrium in cattle. *Genome Res* **10**, 220-227(2000).

67. Farnir, F. et al. 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(2002).

68. Coppieters, W. et al. A QTL with major effect on milk yield and composition maps to bovine chromosome 14. *Mammalian genome : official journal of the International Mammalian Genome Society* **9**, 540-4(1998).

69. Sonstegard, T.S. et al. Consensus and comprehensive linkage maps of the bovine sex chromosomes. *Anim Genet* **32**, 115-117(2001).

70. Ihara, N. et al. A comprehensive genetic map of the cattle genome based on 3802 microsatellites. *Genome Res* **14**, 1987-1998(2004).

71. Grisart, B. et al. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 2398-403(2004).

72. Meuwissen, T.H. & Goddard, M.E. Prediction of identity by descent probabilities from marker-haplotypes. *Genet. Sel. Evol.* **33**, 605-634(2001).

73. Blott, S. et al. 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(2003).

74. Kim, J. & Georges, M. Evaluation of a New Fine-mapping Method Exploiting Linkage Disequilibrium: a Case Study Analysing a QTL with Major Effect on Milk Composition on Bovine Chromosome 14. 1250-1265(2001).

75. Mount, D.W. *Bioinformatics: Sequence and Genome Analysis.* (NY, 2001).

76. Walsh, L.M.; A.B. *Genetics and Analysis of Quantitative Traits.* (Sunderland, 1998).

77. Johnson, D.L. & Thompson, R. Restricted Maximum Likelihood Estimation of Variance Components for Univariate Animal Models Using Sparse Matrix Techniques and Average Information. *Journal of Dairy Science* **78**, 449-456(1995).

78. Boichard, D., Maignel, L. & Verrier, E. No Title. *Evolution* **9**, 323-335(1996).

79. Dib, C. et al. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152-154(1996).

80. Schaffner, S.F. The X chromosome in population genetics. *Nat Rev Genet* **5**, 43-51(2004).

81. Ellegren, H. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat Genet* **24**, 400-402(2000).

82. Wright, S. Evolution and the Genetics of Populations. *The theory of gene frequencies* 213(1969).

83. Hayes, B.J. et al. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res* **13**, 635-643(2003).

84. Caballero, A. Review article Developments in the prediction of effective population size. *Population (English Edition)* **73**, 657-679(1994).

85. Boichard, D. Analyse genealogique des races bovines laitieres francaises. *INRA Prod. Anim.* **9**, 323-325(1996).

86. Edwards, C.A. & Ferguson-Smith, A.C. Mechanisms regulating imprinted genes in clusters. *Curr Opin Cell Biol* **19**, 281-289(2007).

87. Luedi, P.P., Hartemink, A.J. & Jirtle, R.L. Genome-wide prediction of imprinted murine genes. *Genome Res* **15**, 875-884(2005).

88. Luedi, P.P. et al. Computational and experimental identification of novel human imprinted genes. *Genome Res* **17**, 1723-1730(2007).

89. Reik, W. & Lewis, A. Co-evolution of X-chromosome inactivation and imprinting in mammals. *Nat Rev Genet* **6**, 403-410(2005).

90. Wilkins, J.F. & Haig, D. What good is genomic imprinting: the function of parent-specific gene expression. *Nat Rev Genet* **4**, 359-368(2003).

91. Constancia, M., Kelsey, G. & Reik, W. Resourceful imprinting. *Nature* **432**, 53-57(2004).

92. Feil, R. & Berger, F. Convergent evolution of genomic imprinting in plants and mammals. *Trends Genet* **23**, 192-199(2007).

93. Georges, M. Mapping, fine mapping, and molecular dissection of quantitative trait Loci in domestic animals. *Annual review of genomics and human genetics* **8**, 131-62(2007).

94. Cockett, N.E. et al. Polar overdominance at the ovine callipyge locus. *Science (New York, N.Y.)* **273**, 236-8(1996).

95. Nezer, C. et al. An imprinted QTL with major effect on muscle mass and fat deposition maps to the IGF2 locus in pigs. *Nature genetics* **21**, 155-6(1999).

96. Jeon, J.T. et al. A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the IGF2 locus. *Nat Genet* **21**, 157-158(1999).

97. Davis, E. et al. Ectopic Expression of DLK1 Protein in Skeletal Muscle of Padumnal Heterozygotes Causes the Callipyge Phenotype. *October* **14**, 1858-1862(2004).

98. Van Laere, A.-S. et al. A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* **425**, 832-6(2003).

99. Koning, D.J. de et al. Genome-wide scan for body composition in pigs reveals important role of imprinting. *Proc Natl Acad Sci U S A* **97**, 7947-7950(2000).

100. Tuiskula-Haavisto, M. et al. Quantitative trait loci with parent-of-origin effects in chicken. *Genet Res* **84**, 57-66(2004).

101. Cui, Y., Cheverud, J.M. & Wu, R. A statistical model for dissecting genomic imprinting through genetic mapping. *Genetica* **130**, 227-239(2007).

102. Haley, C.S., Knott, S.A. & Elsen, J.M. Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* **136**, 1195-1207(1994).

103. Koning, D.-J. de, Bovenhuis, H. & Arendonk, J.A.M. van On the detection of imprinted quantitative trait loci in experimental crosses of outbred species. *Genetics* **161**, 931-938(2002).

104. McRae, A.F. et al. Linkage disequilibrium in domestic sheep. *Genetics* **160**, 1113-1122(2002).

105. Nsengimana, J. et al. Linkage disequilibrium in the domesticated pig. *Genetics* **166**, 1395-1404(2004).

106. Harmegnies, N. et al. Results of a whole-genome quantitative trait locus scan for growth, carcass composition and meat quality in a porcine four-way cross. *Animal genetics* **37**, 543-53(2006).

107. Sutter, N.B. et al. Extensive and breed-specific linkage disequilibrium in Canis familiaris. *Genome Res* **14**, 2388-2396(2004).

108. Jungerius, B.J. et al. Estimation of the extent of linkage disequilibrium in seven regions of the porcine genome. *Anim Biotechnol* **16**, 41-54(2005).

109. Lindblad-Toh, K. et al. Genome sequence, comparative analysis and haplotype structure of the

domestic dog. *Nature* **438**, 803-819(2005).

110. Sandor, C. et al. Linkage disequilibrium on the bovine X chromosome: characterization and use in quantitative trait locus mapping. *Genetics* **173**, 1777-86(2006).

111. Grisart, B. et al. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc Natl Acad Sci U S A* **101**, 2398-2403(2004).

112. Knott, S.A. et al. Multiple marker mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. *Genetics* **149**, 1069-1080(1998).

113. Churchill, G.A. & Doerge, R.W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963-971(1994).

114. Grisart, B. et al. Positional Candidate Cloning of a QTL in Dairy Cattle: Identification of a Missense Mutation in the Bovine. *Genome Research* 222-231(2001).doi:10.1101/gr.224202.1

115. Cohen-Zinder, M. et al. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Res* **15**, 936-944(2005).

116. Clop, A. et al. A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nature genetics* **38**, 813-8(2006).

117. George, A.W., Visscher, P.M. & Haley, C.S. Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics* **156**, 2081-2092(2000).

118. Uleberg, E. et al. Fine mapping of a QTL for intramuscular fat on porcine chromosome 6 using combined linkage and linkage disequilibrium mapping. *J Anim Breed Genet* **122**, 1-6(2005).

119. Heuven, H.C.M. et al. Efficiency of population structures for mapping of Mendelian and imprinted quantitative trait loci in outbred pigs using variance component methods. *Genet Sel Evol* **37**, 635-655(2005).

120. Hager, R., Cheverud, J.M. & Wolf, J.B. Maternal effects as the cause of parent-of-origin effects that mimic genomic imprinting. *Genetics* **178**, 1755-62(2008).

121. Abraham, C. & Cho, J.H. Functional consequences of NOD2 (CARD15) mutations. *Inflamm*

Bowel Dis **12**, 641-650(2006).

122. Cookson, W. et al. Mapping complex disease traits with global gene expression. *Nature reviews. Genetics* **10**, 184-94(2009).

123. Dixon, A.L. et al. A genome-wide association study of global gene expression. *Nature genetics* **39**, 1202-7(2007).

124. Libioulle, C. et al. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet* **3**, e58(2007).

125. Barrett, J.C. et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature genetics* **40**, 955-62(2008).

126. Dixon, A.L. et al. A genome-wide association study of global gene expression. *Nat Genet* **39**, 1202-1207(2007).

127. Crohn, B.B., Ginzburg, L. & Oppenheimer, G.D. Landmark article Oct 15, 1932. Regional ileitis. A pathological and clinical entity. By Burril B. Crohn, Leon Ginzburg, and Gordon D. Oppenheimer. *JAMA* **251**, 73-79(1984).

128. Schreiber, S. et al. Genetics of Crohn disease, an archetypal inflammatory barrier disease. *Nat Rev Genet* **6**, 376-388(2005).

129. Hugot, J.P. et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599-603(2001).

130. Gunderson, K.L. et al. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet* **37**, 549-554(2005).

131. Setakis, E., Stirnadel, H. & Balding, D.J. Logistic regression protects against population structure in genetic association studies. *Genome Res* **16**, 290-296(2006).

132. Duerr, R.H. et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**, 1461-1463(2006).

133. Peltekova, V.D. et al. Functional variants of OCTN cation transporter genes are associated with Crohn disease. *Nat Genet* **36**, 471-475(2004).

134. Stoll, M. et al. Genetic variation in DLG5 is associated with inflammatory bowel disease. *Nat Genet* **36**, 476-480(2004).

135. Yamazaki, K. et al. Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. *Hum Mol Genet* **14**, 3499-3506(2005).

136. Hampe, J. et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nat Genet* **39**, 207-211(2007).

137. Smit, A.F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9**, 657-663(1999).

138. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050(2005).

139. Kabashima, K. et al. The prostaglandin receptor EP4 suppresses colitis, mucosal damage and CD4 cell activation in the gut. *J Clin Invest* **109**, 883-893(2002).

140. Yalcin, B. et al. Genetic dissection of a behavioral quantitative trait locus shows that Rgs2 modulates anxiety in mice. *Nat Genet* **36**, 1197-1202(2004).

141. Irizarry, R.A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264(2003).

142. Bolstad, B.M. et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193(2003).

143. Abecasis, G.R. et al. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**, 97-101(2002).

144. Sham, P.C. et al. Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *American journal of human genetics* **71**, 238-53(2002).

145. Mathew, C.G. New links to the pathogenesis of Crohn disease provided by genome-wide association scans. *Nat Rev Genet* **9**, 9-14(2008).

146. Parkes, M. et al. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat Genet* **39**, 830-832(2007).

147. Rioux, J.D. et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet* **39**, 596-604(2007).

148. WTCCC Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78(2007).

149. Cargill, M. et al. A large-scale genetic association study confirms IL12B and leads to the identification of IL23R as psoriasis-risk genes. *Am J Hum Genet* **80**, 273-290(2007).

150. Burton, P.R. et al. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat Genet* **39**, 1329-1337(2007).

151. Li, Y.A.A. Rapid haplotype reconstruction and missing genotype inference. *Am. J. Hum. Genet.* **S79**, 2290(2006).

152. Marchini, J. et al. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-913(2007).

153. Clayton, D.G. et al. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* **37**, 1243-1246(2005).

154. Ogura, Y. et al. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**, 603-606(2001).

155. Rioux, J.D. et al. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat Genet* **29**, 223-228(2001).

156. Dixon, A.L. et al. A genome-wide association study of global gene expression. *Nat Genet* **39**, 1202-1207(2007).

157. Moffatt, M.F. et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448**, 470-473(2007).

158. Tysk, C. et al. Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. *Gut* **29**, 990-996(1988).

159. Zeggini, E. et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* **40**, 638-645(2008).

160. Wedemeyer, J. et al. Enhanced production of monocyte chemotactic protein 3 in inflammatory bowel disease mucosa. *Gut* **44**, 629-635(1999).

161. Dinarello, C.A. Interleukin-18 and the pathogenesis of inflammatory diseases. *Semin Nephrol* **27**, 98-114(2007).

162. Kazeem, G.R. & Farrall, M. Integrating case-control and TDT studies. *Ann Hum Genet* **69**, 329-335(2005).

163. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575(2007).

164. Pe'er, I. et al. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* **32**, 381-385(2008).

165. Roeder, G.S. Meiotic chromosomes: it takes two to tango. *Genes & Development* **11**, 2600-2621(1997).

166. Page, S.L. & Hawley, R.S. Chromosome choreography: the meiotic ballet. *Science (New York, N.Y.)* **301**, 785-9(2003).

167. Coop, G. & Przeworski, M. An evolutionary view of human recombination. *Nature reviews. Genetics* **8**, 23-34(2007).

168. Martinez-Perez, E. & Colaiácovo, M.P. Distribution of meiotic recombination events: talking to your neighbors. *Current opinion in genetics & development* **19**, 105-12(2009).

169. Hassold, T. & Hunt, P. To err (meiotically) is human: the genesis of human aneuploidy. *Nature reviews. Genetics* **2**, 280-91(2001).

170. Handel, M.A. & Schimenti, J.C. Genetics of mammalian meiosis: regulation, dynamics and impact on fertility. *Nature reviews. Genetics* **11**, 124-36(2010).

171. Broman, K.W. & Weber, J.L. Characterization of human crossover interference. *American journal of human genetics* **66**, 1911-26(2000).

172. Cheung, V.G. et al. Polymorphic variation in human meiotic recombination. *American journal of human genetics* **80**, 526-30(2007).

173. Kong, A. et al. A high-resolution recombination map of the human genome. *Nature genetics* **31**, 241-7(2002).

174. Kong, A. et al. Sequence variants in the RNF212 gene associate with genome-wide recombination rate. *Science (New York, N.Y.)* **319**, 1398-401(2008).

175. Stefansson, H. et al. A common inversion under selection in Europeans. *Nature genetics* **37**, 129-37(2005).

176. Chowdhury, R. et al. Genetic analysis of variation in human meiotic recombination. *PLoS genetics* **5**, e1000648(2009).

177. Kong, A. et al. Recombination rate and reproductive success in humans. *Nature genetics* **36**, 1203-6(2004).

178. Myers, S. et al. A fine-scale map of recombination rates and hotspots across the human genome. *Science (New York, N.Y.)* **310**, 321-4(2005).

179. Coop, G. et al. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science (New York, N.Y.)* **319**, 1395-8(2008).

180. Jeffreys, A.J. & Neumann, R. The rise and fall of a human recombination hot spot. *Nature genetics* **41**, 625-9(2009).

181. Parvanov, E.D., Petkov, P.M. & Paigen, K. Prdm9 controls activation of mammalian recombination hotspots. *Science (New York, N.Y.)* **327**, 835(2010).

182. Baudat, F. et al. *PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science (New York, N.Y.)* **327**, 836-40(2010).

183. Myers, S. et al. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science (New York, N.Y.)* **327**, 876-9(2010).

184. Berg, I.L. et al. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nature Genetics* 1-6(2010).doi:10.1038/ng.658

185. Ptak, S.E. et al. Fine-scale recombination patterns differ between chimpanzees and humans. *Nature genetics* **37**, 429-34(2005).

186. Winckler, W. et al. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science (New York, N.Y.)* **308**, 107-11(2005).

187. Mihola, O. et al. A mouse speciation gene encodes a meiotic histone H3 methyltransferase. *Science (New York, N.Y.)* **323**, 373-5(2009).

188. Mehta, J. et al. The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. *Science* 522-528(2009).

189. Ross-Ibarra, J. The evolution of recombination under domestication: a test of two hypotheses. *The American naturalist* **163**, 105-12(2004).

190. Charlier, C. et al. Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nature genetics* **40**, 449-54(2008).

191. Matukumalli, L.K. et al. Development and characterization of a high density SNP genotyping assay for cattle. *PloS one* **4**, e5350(2009).

192. Druet, T. & Georges, A.M. A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and QTL fine mapping. *Genetics* **54**, 258(2010).

193. Bannister, L.A. et al. Positional cloning and characterization of mouse mei8, a disrupted allelle of the meiotic cohesin Rec8. *Genesis* **40**, 184-194(2004).

194. Xu, H. et al. Absence of mouse REC8 cohesin promotes synapsis of sister chromatids in meiosis. *Dev Cell* **8**, 949-961(2005).

195. Fledel-Alon, A. et al. Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS genetics* **5**, e1000658(2009).

196. Sturt, E. & Smith, C.A. The relationship between chromatid interference and the mapping function. *Cytogenet Cell Genet* **17**, 212-220(1976).

197. Paigen, K. & Petkov, P. Mammalian recombination hot spots: properties, control and evolution. *Nature reviews. Genetics* **11**, 221-33(2010).

198. Lian, J. et al. Variation in crossover interference levels on individual chromosomes from human males. *Human molecular genetics* **17**, 2583-94(2008).

199. McPeek, M.S. & Speed, T.P. Modeling interference in genetic recombination. *Genetics* **139**, 1031-44(1995).

200. Doerge, R.W. & Churchill, G.A. Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285-94(1996).

201. Barrett, J.C. et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40**, 955-962(2008).

202. Kathiresan, S. et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet* **40**, 189-197(2008).

203. Willer, C.J. et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* **40**, 161-169(2008).

204. Wolf, N. et al. Psoriasis is associated with pleiotropic susceptibility loci identified in type II diabetes and Crohn disease. *J Med Genet* **45**, 114-116(2008).

205. McCarthy, M.I. et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics* **9**, 356-69(2008).

206. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nature reviews. Genetics* **7**, 85-97(2006).

207. Lander, E.S. & Schork, N.J. Genetic dissection of complex traits. *Science (New York, N.Y.)* **265**, 2037-48(1994).