



**HAL**  
open science

# Déséquilibre de liaison et cartographie de QTL en population sélectionnée

Florence Ytournel

► **To cite this version:**

Florence Ytournel. Déséquilibre de liaison et cartographie de QTL en population sélectionnée. Life Sciences [q-bio]. AgroParisTech, 2008. English. NNT : 2008AGPT0004 . pastel-00003789

**HAL Id: pastel-00003789**

**<https://pastel.hal.science/pastel-00003789>**

Submitted on 3 Jun 2008

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



UFR Génétique,  
Elevage et  
Reproduction



Agriculture, Alimentation,  
Biologie, Environnements, Santé



Station de Génétique  
Quantitative et  
Appliquée

Thèse

Pour obtenir le grade de

Docteur d'AgroParisTech

Discipline : Génétique animale

présentée et soutenue publiquement par

Florence YTOURNEL

le 28 janvier 2008

## **Déséquilibre de liaison et cartographie de QTL en population sélectionnée**

Directeurs de thèse : Hélène GILBERT et Didier BOICHARD

### Jury

Pascale LE ROY  
Frédéric FARNIR  
Laurence MOREAU  
Hélène GILBERT  
Etienne VERRIER

Directrice de Recherches, INRA, Rennes  
Professeur, Faculté de Médecine Vétérinaire, Liège  
Chargée de Recherches, INRA, Gif-S/Yvette  
Chargée de Recherches, INRA, Jouy-en-Josas  
Professeur, AgroParisTech

Rapporteur  
Rapporteur  
Examineur  
Examineur  
Président





UFR Génétique,  
Elevage et  
Reproduction



Agriculture, Alimentation,  
Biologie, Environnements, Santé



Station de Génétique  
Quantitative et  
Appliquée

Thèse

Pour obtenir le grade de

Docteur d'AgroParisTech

Discipline : Génétique animale

présentée et soutenue publiquement par

Florence YTOURNEL

le 28 janvier 2008

## **Déséquilibre de liaison et cartographie de QTL en population sélectionnée**

Directeurs de thèse : Hélène GILBERT et Didier BOICHARD

### Jury

Pascale LE ROY	Directrice de Recherches, INRA, Rennes	Rapporteur
Frédéric FARNIR	Professeur, Faculté de Médecine Vétérinaire, Liège	Rapporteur
Laurence MOREAU	Chargée de Recherches, INRA, Gif-S/Yvette	Examinateur
Hélène GILBERT	Chargée de Recherches, INRA, Jouy-en-Josas	Examinateur
Etienne VERRIER	Professeur, AgroParisTech	Président



THESE

**Déséquilibre de liaison  
et cartographie de QTL  
en population sélectionnée**

PHD THESIS

**Linkage disequilibrium  
and QTL fine mapping  
in a selected population**

Florence YTOURNEL, 2008

Thèse de doctorat  
UFR Génétique, Elevage et Reproduction  
AgroParisTech

Financement :

Bourse de la Direction Générale de l'Enseignement et de la Recherche  
Département de Génétique Animale

Encadrants principaux :

Hélène GILBERT, Chargée de Recherche à la Station de Génétique Quantitative et Appliquée – Institut National de la Recherche Agronomique, Domaine de Vilvert, 78352 Jouy-en-Josas, France

E-mail : [helene.gilbert@jouy.inra.fr](mailto:helene.gilbert@jouy.inra.fr)

Didier BOICHARD, Directeur du département de Génétique Animale – Institut National de la Recherche Agronomique, Domaine de Vilvert, 78352 Jouy-en-Josas, France

E-mail : [didier.boichard@jouy.inra.fr](mailto:didier.boichard@jouy.inra.fr)

Auteur:

Florence YTOURNEL, Station de Génétique Quantitative et Appliquée – Institut National de la Recherche Agronomique, Domaine de Vilvert, 78352 Jouy-en-Josas, France

E-mail : [florence.ytournel@jouy.inra.fr](mailto:florence.ytournel@jouy.inra.fr)

*A mes grands-pères*

*A Maxime*





## *Remerciements*

Je remercie les membres du jury : Etienne Verrier, qui a accepté la présidence du jury, Frédéric Farnir et Pascale Le Roy qui m'ont fait l'honneur d'accepter d'être rapporteurs, et Laurence Moreau qui m'honore d'être l'examinatrice de ce travail.

Je remercie tout particulièrement celui sans qui je n'aurais pas pu faire cette thèse. Il a réussi à concilier sa fonction de directeur du département de génétique animale et celui de directeur de thèse. Les mails reçus à minuit furent surprenants au début, mais le gage d'une présence précieuse, accompagnée de réponses et de conseils avisés. Merci Didier.

Je suis aussi reconnaissante à Jean-Pierre Bidanel de m'avoir accueillie à la SGQA, et d'avoir toujours été là quand j'ai eu des requêtes à lui soumettre, et ce dès mon premier jour à la SGQA (et même pour pouvoir y arriver !).

Et que dire pour Hélène Gilbert, qui a été mon encadrante principale. Tu as réussi à m'aiguiller dans ce projet tout en me laissant une certaine dose de liberté, et à gérer mon mauvais caractère dans les dernières semaines. Chapeau !

Je n'oublierai pas non plus Tom Druet qui a toujours répondu présent face à mes incompréhensions sur ses programmes. Merci aussi d'avoir su « mettre les pieds dans le plat » lorsque cela a été nécessaire. Sans toi, qui sait quelle serait ma situation actuelle...

Je dois aussi beaucoup à tous les membres de mon comité de thèse qui m'ont permis de dégager des besoins prioritaires, notamment lors de la création du simulateur. Merci donc à Christine Dillman, Mathieu Gautier, Etienne Verrier, et plus particulièrement à Brigitte Mangin et Hubert de Rochambeau.

Enfin, je tiens à faire part de toute ma gratitude à Andrés Legarra. Comment as-tu fait pour être toujours être aussi réactif à mes besoins et demandes ? Et le simulateur te remercie aussi de l'avoir déNAGué...

Et puis il y a tout le petit monde de la SGQA... Merci à Sylvie Nugier, Hervé Lagant et Fabien Dequine de m'avoir assistée face aux soucis informatiques, et à Thierry Coudert d'avoir été patient face à mes remplissages intempestifs de log... Je voudrais aussi remercier Manuëla Ferré, Christelle Gérardin et Sylvie Fillon pour m'avoir aidé dans les formalités administratives et de déplacement, et Serge Tignoux pour tous les éléments de reprographie et autres petits problèmes tôt le matin. J'ai également une pensée pour tous les animateurs du « café cochon » : Anne, Hélène, les Sophie, Armelle, Gilles, Florence, Eric, Thierry, Marie-

Pierre, Guillemette, Catherine, Mickaël, Marie-No, Pierre, Aurélie, Sébastien (vive les paris !), Denis,... Merci à Steph, Val, Delphine, Al, François et Hélène Leclerc (courage tous les deux, vous tenez le bon bout !) pour leur soutien dans les moments un peu plus difficiles, à Benj' pour ses discussions « palmipèdemment grasses », ainsi qu'à tous ceux qui ont partagé « mon » bureau, et qui m'ont supportée : Adalberto, Al, et dernièrement Aline. Je tiens aussi à remercier Marie-Yvonne, Vincent, François et Francis pour leurs discussions variées, mais toujours sources d'informations pertinentes, pendant notre petite marche quotidienne pour rejoindre la gare de Jouy en Josas.

Enfin, il y a tous ceux qui m'ont toujours soutenue : mes parents, Nico, et toute la famille en Auvergne. J'ai une pensée toute particulière pour mes grands-mères et mamie Dombey. Merci aussi à Pierre-Yves Moy, mon kiné, à qui mon genou et mon dos sont grandement redevables... Et je conclurai avec les amis de toujours : Isa, So-So, Antje, les deux Anne, Chris, Anne-So, Alex, Adeline, Ben, et tous les ENSARIENS... Merci d'être toujours là, même si je ne suis pas toujours des plus disponibles...

Et à tous ceux qui ne sont pas explicitement cités ici...

*Merci !!!*

## **Publications :**

### **Dans des revues scientifiques :**

Ytournel F., Gilbert H., Druet T., Boichard D. Location of maximum linkage disequilibrium in selected populations. *Soumis à Genetics Selection Evolution*. (Partie 2.2)

Ytournel F., Boichard D., Gilbert H., Legarra A. LDSO: A complete program for the simulation of pedigrees and molecular information under various evolutionary forces. *Soumission prévue à Journal of Heredity*. (Partie 2.1)

Ytournel F., Gilbert H., Boichard D. Comment affiner la localisation d'un QTL ? *Soumission prévue à INRA Productions Animales*. (Partie 1.2)

Ytournel F., Gilbert H., Boichard D. Comparison of estimated Identity-by-Descent probabilities and true Identity-by-Descent status of chromosomal regions in a population. *Soumission prévue à BMC Genetics*. (Partie 3.1)

### **En congrès internationaux :**

Ytournel F., Gilbert H., Druet T., Boichard D., 2006. Structure of linkage disequilibrium and length of IBD segments in simulated populations. 8<sup>th</sup> World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Brésil.

Ytournel F., Boichard D., Gilbert H., 2007. IBD probabilities: their discrimination ability and their co-evolution with linkage disequilibrium. 11<sup>th</sup> QTL MAS Workshop, Toulouse, France.

Ytournel F., Boichard D., Gilbert H., 2007. Concordance between IBD probabilities and linkage disequilibrium. 58<sup>th</sup> meeting of the European Association for Animal Production, Dublin, Ireland.



# Table des matières

INTRODUCTION GENERALE .....	15
PREMIERE PARTIE : REVUE BIBLIOGRAPHIQUE .....	19
Partie 1.1. Déséquilibre de liaison dans les populations animales .....	21
Introduction .....	23
I. Définition et propriétés .....	24
II. Mesures du déséquilibre de liaison entre locus .....	26
II.A. Mesures du LD entre deux locus .....	26
II.B. Mesures du LD avec plus de deux locus .....	29
III. Forces évolutives et déséquilibre de liaison .....	29
III.A. Mutation .....	29
III.B. Migration et mélange de populations .....	30
III.C. Dérive génétique .....	31
III.D. Sélection .....	32
IV. Déséquilibre de liaison dans des populations animales .....	33
Partie 1.2. Cartographie fine de QTL .....	39
Introduction .....	42
I. Principes et facteurs influençant la résolution de la cartographie .....	43
I.A. Principe général .....	43
I.B. Facteurs de précision de la cartographie .....	46
II. Dissection chromosomique .....	54
III. Méthodes statistiques de cartographie fine .....	59
III.A. Méthodes utilisant exclusivement le LD .....	59
III.B. Méthodes combinant LD et analyse de liaison .....	63
Conclusion .....	66
DEUXIEME PARTIE : EVOLUTION DE LA STRUCTURE DU DESEQUILIBRE DE LIAISON SOUS L'INFLUENCE DE LA SELECTION	73
Partie 2.1. Développement d'un simulateur : « Linkage Disequilibrium with several options » (LDSO) .....	77
Introduction .....	79
I. Article .....	80
II. Justification du choix de simulation des haplotypes .....	88
III. Dispositifs expérimentaux simulables .....	90
Conclusion .....	94

Partie 2.2. Structure du déséquilibre de liaison dans des populations sélectionnées .....	95
Introduction .....	97
I. Article.....	98
II. Résultats complémentaires sur les locus en LD maximum avec le QTL et discussion finale.....	124
III. Conclusion.....	128
<b>TROISIEME PARTIE : INFLUENCE DE LA SELECTION SUR LES METHODES DE CARTOGRAPHIE FINE .....</b>	<b>131</b>
Partie 3.1 : Influence de la sélection sur les probabilités d'IBD .....	133
Introduction .....	135
I. Article.....	136
II. Lien entre l'IBD et le LD .....	147
II.A. Critère de comparaison.....	147
II.B. Résultats .....	147
III. Conclusion.....	148
Partie 3.2 : Robustesse des méthodes de cartographie fine à la sélection .....	151
Introduction .....	153
I. Matériel et méthodes .....	154
I.A. Populations simulées .....	154
I.B. Cartes génétiques simulées.....	155
I.C. Méthodes de cartographie fine .....	155
I.D. Evaluation de la précision de cartographie des méthodes.....	160
I.E. Evaluation de la concordance entre LD maximum et position putative du QTL	161
II. Résultats .....	161
II.A. Comparaison des méthodes.....	161
II.B. Sensibilité des méthodes à la sélection .....	163
II.C. Concordance LD – cartographie du QTL.....	164
III. Discussion .....	168
IV. Conclusion.....	171
<b>QUATRIEME PARTIE : DISCUSSION GENERALE ET PERSPECTIVES..</b>	<b>173</b>
Introduction .....	175
I. Réalisme des simulations .....	176
I.A. Adéquation aux populations réelles .....	176
I.B. Adéquation aux données génétiques réelles.....	178
II. Effets de la sélection .....	179
II.A. LD et lien LD - IBD .....	179
II.B. Cartographie .....	182
Conclusion générale .....	185
<b>Références bibliographiques .....</b>	<b>189</b>

Glossaire .....	199
ANNEXES .....	201





# INTRODUCTION GENERALE

La disponibilité de la séquence des génomes complets d'un nombre croissant d'espèces, y compris chez les animaux domestiques, renouvelle considérablement les outils d'analyse de la diversité génétique. Ainsi, de très nombreux marqueurs génétiques sont disponibles, ainsi que des outils de génotypage à haut débit relativement peu coûteux. Cette situation nouvelle permet d'entrevoir des perspectives beaucoup plus favorables pour l'identification des gènes impliqués dans la variabilité génétique des caractères phénotypiques. Un important effort est consenti au niveau international dans le développement de méthodes de cartographie fine et de caractérisation de ces gènes, tant chez l'homme que les animaux et les végétaux. L'enjeu de cette recherche est double : d'une part, d'un point de vue fondamental, mieux comprendre la relation génotype-phénotype et les mécanismes mis en jeu ; d'autre part pour les espèces d'élevage, faciliter la mise en œuvre de la sélection pour les caractères économiquement importants, directement à partir de l'information moléculaire, en complément ou à la place de l'information phénotypique. Cette thèse consacrée au déséquilibre de liaison et son utilisation en cartographie fine de gènes s'inscrit dans cette dynamique internationale.

L'étape de cartographie fine est considérée comme indispensable dans la recherche des gènes impliqués dans l'expression d'un caractère. En effet, dans la plupart des cas, ces gènes impliqués dans le déterminisme d'un caractère sont inconnus et potentiellement nombreux. Souvent, n'utiliser qu'une approche de gène candidat, où on postule qu'un gène donné est impliqué dans les variations génétiques du caractère, paraît hasardeux. La stratégie de choix pour la caractérisation des gènes combine généralement une approche positionnelle et une approche fonctionnelle. L'approche positionnelle consiste à localiser la région chromosomique (Quantitative Trait Locus ou QTL) portant le ou les gènes à identifier. Ces méthodes, essentiellement statistiques, permettent de rechercher une association entre le niveau du caractère et la région chromosomique transmise.

Dans les populations animales, les QTL sont classiquement détectés et localisés grâce à des dispositifs expérimentaux familiaux sur au moins deux générations, sur lesquels sont appliquées des **analyses de liaison**. L'analyse de liaison consiste à rechercher au sein d'une famille une association statistique entre le phénotype du descendant et la région chromosomique transmise par le ou les parents, ces régions étant suivies du parent au

descendant grâce à des marqueurs moléculaires polymorphes. Elles exploitent en général une génération de crossing-over pour estimer la liaison entre le phénotype et le génotype. Cette méthode est très efficace et robuste pour détecter des gènes. Cependant, elle ne fournit qu'une résolution limitée de la localisation, résolution qui est fonction du nombre de recombinaisons dans la région du QTL. Plus cette région est petite, plus les recombinaisons sont rares. La taille du dispositif expérimental devient limitante et pas seulement la densité de marqueurs qui n'apportent plus d'information dès lors que les recombinaisons sont localisées précisément. Avec des dispositifs familiaux, l'intervalle de confiance est de l'ordre du centiMorgan (cM) pour les gènes majeurs (définis dans ce cas comme les gènes pour lesquels le génotype est déduit sans erreur du phénotype) mais de l'ordre de plusieurs dizaines de centiMorgans pour des QTL. De tels intervalles de confiance, qui contiennent plusieurs centaines de gènes, sont trop vastes pour permettre d'identifier un petit nombre de gènes candidats. Ils sont aussi trop vastes pour pratiquer une sélection assistée par marqueurs (SAM) simple et efficace.

Il est donc important de réduire l'intervalle de localisation des QTL. En 1986, Bodmer (repris en 1997 par Xiong et Guo) a proposé d'utiliser les recombinaisons qui s'accumulent au cours des générations pour préciser la **cartographie des locus quantitatifs** ou QTL. Le nombre de recombinaisons utiles ne dépend alors plus du nombre d'individus du dispositif familial mais du nombre de méioses dans l'ensemble du pedigree multigénérationnel de la population étudiée. Les méthodes reposant sur ce principe ont recours au **déséquilibre de liaison (Linkage Disequilibrium, LD)** entre les locus, qui correspond à une association non aléatoire entre allèles de différents locus dans une population. La première partie de cette thèse est consacrée à une synthèse sur le déséquilibre de liaison et les méthodes de cartographie fine de QTL.

On montre dans cette première partie qu'un grand nombre de ces méthodes de cartographie fine font l'hypothèse d'absence de sélection des populations, ce qui est rarement le cas en pratique, surtout dans les populations d'élevage qui sont fortement sélectionnées par l'homme. Un deuxième chapitre aborde l'impact de la sélection sur la structure du déséquilibre de liaison et son effet sur l'efficacité attendue de la cartographie de QTL. Cette étude repose sur le développement d'un simulateur de populations qui est préalablement présenté.

La sélection affecte non seulement la structure vraie du déséquilibre de liaison (chapitre 2) mais aussi l'efficacité des méthodes de cartographie qui supposent souvent l'équilibre de Hardy-Weinberg, et donc l'absence de sélection. Certaines méthodes reposent sur l'estimation de probabilités d'identité par descendance (IBD) entre QTL conditionnellement aux marqueurs qui les entourent. Le chapitre 3 montre dans quelle mesure la sélection impacte ces probabilités et donc l'efficacité de la cartographie. Il présente également une comparaison de différentes méthodes de cartographie fine dans des situations avec ou sans sélection.

Enfin, un dernier chapitre synthétise l'ensemble des résultats et les discute. En terme de valorisation, ce travail devrait donner lieu à une synthèse bibliographique sur les méthodes de cartographie fine, trois articles et une présentation d'un simulateur de populations.



**PREMIERE PARTIE :**  
**REVUE BIBLIOGRAPHIQUE**



**Partie 1.1.**

**Déséquilibre de liaison**

**dans les populations animales**





## Introduction





Le déséquilibre de liaison dans une population, ou déséquilibre de phase gamétique, se définit comme l'association non aléatoire d'allèles à des locus différents. Il est créé par différentes forces évolutives dont, entre autres, la sélection, les mutations, le mélange des populations, la dérive génétique et les goulets d'étranglement (cf. III). C'est un outil central pour les généticiens car il renseigne sur l'histoire de la population grâce aux outils de marquage moléculaire. Il constitue également le socle des méthodes de cartographie fine de gènes et de QTL.

Il a été étudié dès 1917 par Jennings, qui s'intéressait aux conséquences pour deux caractères, de la liaison génétique dans différents systèmes de reproduction (accouplement aléatoire, sélection sur l'un des caractères, auto-fécondation ou homogamie). Au cours des décennies suivantes, les études ont porté sur les méthodes de mesure du LD (cf. II) et sur ses facteurs de variation (cf. III). Bien que le concept de déséquilibre de liaison et les premières études soient déjà anciennes, les évaluations systématiques du LD à l'échelle de populations, avec un gradient de distance entre marqueurs n'ont eu lieu qu'après 2000, grâce à la disponibilité de données de marquage importantes. En effet, l'évaluation du LD nécessite des marqueurs génétiques répartis sur tout le génome, et des méthodes de génotypage afin d'estimer les fréquences alléliques et les fréquences de co-occurrence d'allèles à deux locus marqueurs différents. Or ce n'est que depuis peu d'années que l'on dispose pour les espèces d'élevage de nombreux marqueurs d'abord microsatellites (1995 environ) puis SNP, ouvrant la voie au génotypage à haut débit et à haute densité. Les marqueurs microsatellites correspondent à un motif de quelques nucléotides (généralement 2) répété un certain nombre de fois en tandem. Présentant un taux de mutation relativement élevé, le nombre de répétitions varie entre individus au sein d'une même espèce et constitue l'origine du polymorphisme. Le nombre d'allèles présents est souvent élevé, rendant le marqueur très informatif : les individus ont plus de chances d'être hétérozygotes, ce qui permet de tracer l'allèle transmis à leur(s) descendant(s). Dérivés des programmes de séquençage, très fréquents dans le génome (100 fois plus environ que les microsatellites), les marqueurs de type Single Nucleotide Polymorphisms (SNP) sont des marqueurs généralement bialléliques caractérisés par un changement d'une seule base. Bien que moins informatifs que les microsatellites, ils sont en train de les supplanter dans leur utilisation en génétique par leur nombre et par leurs méthodes de génotypage à très haut débit.

Ce chapitre introductif vise à présenter de façon synthétique successivement (i) le déséquilibre de liaison, (ii) les méthodes permettant de l'estimer, (iii) les effets sur le LD des forces évolutives présentes dans les populations de rente et (iv) un bilan des études réalisées sur l'étendue du LD dans ces populations.

## I. Définition et propriétés

Comme indiqué précédemment, le déséquilibre de liaison est défini comme l'association non aléatoire d'allèles à des locus différents. Dans le cas de locus bi-alléliques ayant respectivement pour allèles A/a et B/b, il se calcule comme  $D_{AB} = x_{AB} - p_A q_B$ , où  $x_{AB}$  est la fréquence de l'haplotype porteur des allèles A et B, et  $p_A$  ( $q_B$ ) la fréquence de l'allèle A (B) dans la population. Ce calcul est illustré par la figure 1.1.1, avec le cas 1 en équilibre de liaison ( $D = 0$ ) et le cas 2 en déséquilibre de liaison ( $D > 0$ ). Dans le cas de deux locus bi-alléliques, le LD est, en valeur absolue, le même pour toutes les combinaisons alléliques. La formule précédente se généralise pour les allèles de locus multi-alléliques de la manière suivante :  $D_{ij} = x_{ij} - p_i q_j$  avec  $x_{ij}$  la fréquence de l'haplotype porteur de l'allèle  $i$  au premier locus et de l'allèle  $j$  au second locus, et  $p_i$  ( $q_j$ ) les fréquences de ces allèles dans la population. La généralisation au déséquilibre global entre deux locus se fait selon les formules explicitées en II.A. Par la suite, on notera  $D$  le déséquilibre entre deux locus.

		Nombre d'haplotypes	
		Cas 1	Cas 2
		30	45
		30	15
		20	5
		20	35
Fréquences alléliques	A	0.6	0.6
	a	0.4	0.4
	B	0.5	0.5
	b	0.5	0.5
Valeur de D		0	0.15

**Figure 1.1.1 : Exemple de calcul du déséquilibre de liaison pour deux locus bi-alléliques**

La création d'un LD entre deux marqueurs n'implique pas de liaison physique entre eux. Cependant, l'évolution de ce déséquilibre est fonction de la distance génétique entre les locus et du nombre de générations écoulées entre la génération de création et la génération actuelle. Le taux de recombinaison entre deux locus (noté  $r$ ) est dépendant de leur distance génétique : pour des distances courtes, le taux de recombinaison (en %) est peu différent de la distance exprimée en centiMorgan (cM). Si  $D_0$  est le déséquilibre initial entre les deux locus dans une population de grande taille, en supposant qu'il n'y a pas de variation des fréquences alléliques entre deux générations, le LD entre les locus diminue d'un facteur  $(1-r)$  par génération dans le cas d'accouplements aléatoires. On obtient alors par récurrence la relation  $D_t = (1-r)^t D_0$ , avec  $D_t$  le LD à la génération  $t$ . De cette formule découlent les courbes présentées en Figure 1.1.2. Il apparaît nettement que le déséquilibre de liaison se maintient de manière prolongée entre des locus proches ( $r=0,01$ ), mais qu'il peut cependant rester un déséquilibre résiduel à plus forte distance si le déséquilibre initial était fort. Par exemple,  $D_{10}$  vaut environ 0,17 avec un taux de recombinaison de 10 cM si le déséquilibre initial valait 0.5.

Pour des locus indépendants ( $c=0.5$ ), le déséquilibre initial est divisé par deux à chaque génération, et redevient rapidement très faible (8 à 9 générations).

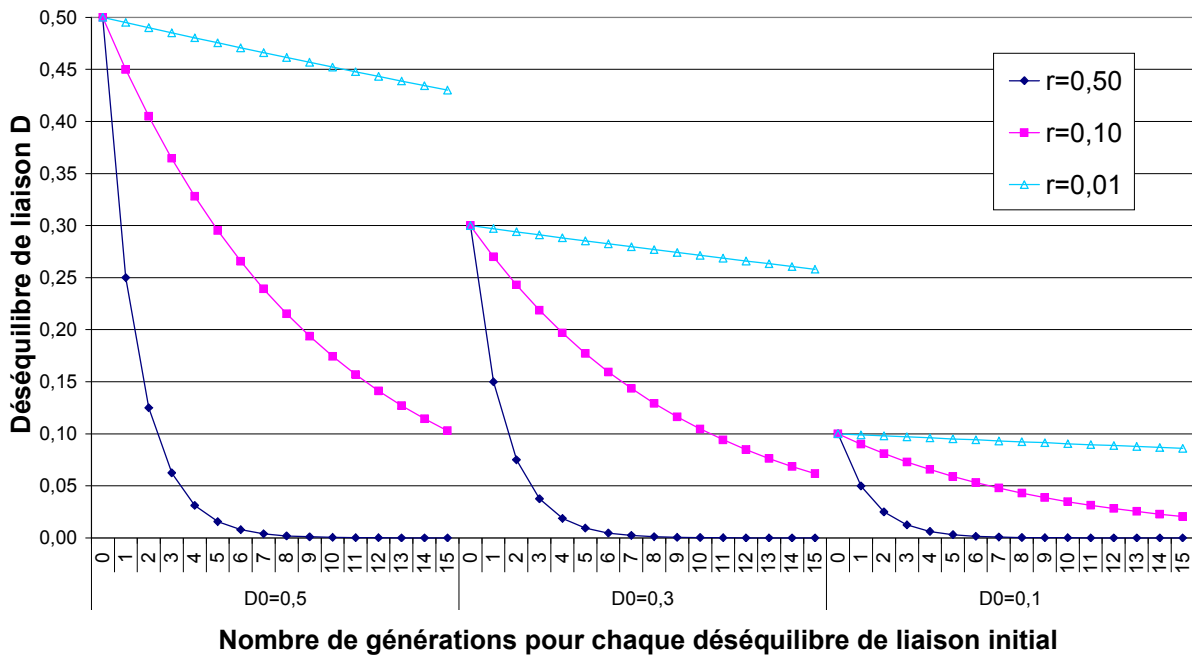


Figure 1.1.2 : Décroissance du déséquilibre de liaison en fonction du temps, du taux de recombinaison ( $r$ ) et du déséquilibre de liaison initial ( $D_0$ )

## II. Mesures du déséquilibre de liaison entre locus

### II.A. Mesures du LD entre deux locus

Les mesures de LD, calculé entre deux locus polymorphes, sont dépendantes de deux facteurs : les fréquences des haplotypes et les fréquences alléliques. La mesure idéale doit être indépendante des fréquences alléliques, et permettre de comparer les valeurs de LD obtenues pour différentes paires de locus. Par ailleurs, elle doit permettre de déterminer si des locus sont en déséquilibre de liaison, ce qui est facilité par la connaissance d'une loi de distribution sous l'hypothèse nulle d'équilibre de liaison. Sinon, on peut avoir recours à des méthodes empiriques (permutation ou bootstrapping par exemple) pour établir la distribution sous l'hypothèse nulle.

A ce titre, la mesure du LD précédemment définie ( $D_i$ ) présente deux inconvénients majeurs : elle dépend des fréquences alléliques, et elle n'est pas normalisée, ce qui ne permet pas de comparer des valeurs entre plusieurs paires de locus. C'est pourquoi d'autres mesures ont été proposées. Certaines sont plutôt utilisées dans les analyses de populations humaines ( $r^2$ ), alors que d'autres ont plus été employées dans les études sur les populations animales de production ( $D'$ ). Une revue de ces mesures a été réalisée par Cierco-Ayrolles *et al.* (2004).

Une première mesure est le coefficient de corrélation entre allèles (Hill et Robertson, 1968), qui est défini pour des locus bi-alléliques par la relation  $r^2 = \frac{D^2}{p_A p_a q_B q_b}$ . Cette mesure

est comprise entre 0 et 1, mais elle est très dépendante des fréquences alléliques. Elle ne prend la valeur de 1 que si chaque allèle du premier locus est associé à un allèle unique au second locus (déséquilibre de liaison complet) et que les fréquences de ces allèles sont identiques. Il s'agit de la mesure de référence dans les analyses de déséquilibre de liaison effectuées dans les populations humaines, et plus généralement dans les analyses sur des locus bi-alléliques. En effet, elle a l'avantage d'avoir une distribution connue sous l'hypothèse nulle d'équilibre de liaison puisqu'elle se comporte de façon équivalente à un test de  $\chi^2$  pour des tables de contingence (Pritchard et Przeworski, 2001), qui correspond à un test d'association entre les locus. Elle est également moins sensible à la taille de la population que d'autres mesures, notamment le  $D'$  (Weiss et Clark, 2002). Cette mesure est généralisée au cas multi-allélique

selon la formule  $r^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{D_{ij}^2}{p_i q_j}$ , avec  $I$  et  $J$  le nombre d'allèles respectifs des locus A et B.

Ce  $r^2$  a été adapté par Yamazaki en 1977 pour obtenir le  $\chi^{2'}$ , qui vaut  $\chi^{2'} = \frac{r^2}{l-1}$  où  $l$  est le minimum de ( $I, J$ ). Dans le cas où deux marqueurs encadrent un QTL, Zhao *et al.* (2005) ont démontré que cette mesure est celle pour laquelle la corrélation entre le LD mesuré entre les deux marqueurs et le LD mesuré entre le QTL et l'un des marqueurs est la plus forte. De plus, elle possède également une distribution de  $\chi^2$  sous  $H_0$ .

Le  $D'$ , appelé déséquilibre normalisé, est une autre mesure classique. Il fut proposé en 1964 par Lewontin puis généralisé pour les locus multi-alléliques par Hedrick en 1987. Il se

calcule de la manière suivante :  $D' = \sum_{i=1}^I \sum_{j=1}^J p_i q_j |D'_{ij}|$ , avec  $D'_{ij} = \frac{D_{ij}}{D_{\max}}$   
où  $D_{\max} = \begin{cases} \min [p_i q_j, (1-p_i) * (1-q_j)] & \text{si } D_{ij} < 0 \\ \min [p_i * (1-q_j), (1-p_i) * q_j] & \text{si } D_{ij} > 0. \end{cases}$

Première partie : revue bibliographique  
Partie 1.1. Déséquilibre de liaison dans les populations animales

Ce coefficient est normalisé afin d'être compris entre -1 et 1 quelles que soient les fréquences alléliques. Son indépendance par rapport aux fréquences alléliques est discutée (Hedrick, 1987 ; Lewontin, 1988). Cependant, il s'avère être surestimé dès lors que des haplotypes sont manquants, ce qui est classiquement le cas pour les haplotypes rares lors d'un échantillonnage, ou dans le cas de populations sélectionnées, du fait de la disparition des haplotypes défavorables (Zhao *et al.*, 2005). En effet, il prend la valeur de 1 dès lors que le déséquilibre de liaison est complet (Harmegnies *et al.*, 2006). Il a été très utilisé dans les études de déséquilibre de liaison dans les populations animales (cf. I.C.).

Il existe d'autres mesures ( $\Delta^2$ ,  $Q$ ,  $D^2$ ,  $\delta$ , cf. Tableau 1.1.1) qui sont utilisées dans les études épidémiologiques de type cas-contrôle en génétique humaine, mais n'ont historiquement pas été utilisées dans les populations animales.

Mesure	Référence	Expression en fonction de D
$D$		$D = p_{AB}p_{ab} - p_{Ab}p_{aB}$
$r^2$	Hill et Robertson (1968)	$\frac{D^2}{p_A p_B p_a p_b}$
$D'$	Lewontin (1964)	$\frac{D}{D_{\max}}$ où $D_{\max} = \begin{cases} \min(p_A p_b, p_a p_B) & \text{si } D > 0 \\ \min(p_A p_B, p_a p_b) & \text{si } D < 0 \end{cases}$
$\delta$	Bengtsson et Thompson (1981)	$\frac{D}{p_B p_{ab}}$
$d$	Nei (1980)	$\frac{D}{p_B p_b}$
$Q$	Yule (1990)	$\frac{D}{p_{AB} p_{ab} + p_{Ab} p_{aB}}$

**Tableau 1.1.1 : Mesures du LD entre deux locus bi-alléliques avec  $p_{ij}$  la probabilité d'avoir l'allèle  $i$  au premier locus et l'allèle  $j$  au second locus,  $p_i$  la probabilité d'avoir l'allèle  $i$**

## **II.B. Mesures du LD avec plus de deux locus**

Des méthodes ont été développées pour estimer le LD avec plus de deux marqueurs simultanément, dans l'objectif de capter plus d'information. Une méthode simple pour des marqueurs très liés consiste à considérer les haplotypes comme des allèles différents, avec l'inconvénient du nombre potentiellement très élevé d'allèles.

D'autres mesures sont intimement liées à la notion d'identité par descendance (Identity By Descent ou IBD), c'est-à-dire au fait que deux haplotypes soient identiques parce qu'ils ont été hérités du même ancêtre. C'est le cas du « Chromosome Segment Homozygosity » décrit en 2003 par Hayes *et al.*, qui est utilisé pour estimer la taille efficace de la population au cours des générations précédant la génération courante. Cette mesure, si elle est appliquée à de grands segments chromosomiques, est indicatrice de la taille efficace récente de la population, alors que si elle est utilisée sur de courtes distances, elle fournit des informations sur la taille efficace de la population sur un terme plus long.

Une dernière mesure est le « Decay of Haplotype Sharing » (DHS) proposé par McPeck et Strahs en 1999. Cette mesure, qui suppose une mutation unique ayant généré un LD complet entre les marqueurs de l'haplotype où elle est survenue et l'allèle muté, peut être interprétée comme la longueur attendue pour le segment hérité de l'haplotype fondateur sur lequel a eu lieu la mutation, ou comme l'inverse du nombre de générations écoulées depuis l'apparition de l'haplotype muté. Elle est généralement appliquée à des échantillons de population du type cas-contrôle, et est essentiellement utilisée dans les populations humaines.

## **III. Forces évolutives et déséquilibre de liaison**

Les forces évolutives abordées ci-dessous sont les plus actives dans les populations d'animaux de production ; les autres, telles que la croissance démographique, ne sont pas abordées.

### **III.A. Mutation**

Les mutations génèrent du déséquilibre de liaison. En effet, les taux de mutation sont généralement faibles, estimés entre  $10^{-5}$  et  $10^{-6}$  par individu et par génération (Falconer et Mackay, 1996). Une mutation apparaît ponctuellement dans un haplotype donné ; elle n'est donc associée qu'à cet haplotype et aux allèles qu'il contient. Le déséquilibre de liaison est



alors complet avec l'haplotype. La création de déséquilibre de liaison s'accompagne ici d'une augmentation du nombre d'haplotypes.

Le déséquilibre ainsi créé peut être ponctuel dans le temps (dans le cas d'une mutation létale ou perdue immédiatement par dérive génétique). Dans d'autres cas, la mutation peut se transmettre aux générations suivantes et voir sa fréquence augmenter, par dérive génétique ou par sélection. La dérive est fréquemment en cause dans les populations d'élevage, du fait de leur effectif génétique faible. L'utilisation massive d'un reproducteur porteur d'un allèle rare conduit à la diffusion de cet allèle et à l'augmentation de sa fréquence, même quand cet allèle n'est pas responsable de la valeur sélective du reproducteur. Dans le cas où une mutation confère un avantage sélectif, sa probabilité de propagation au cours des générations est supérieure, renforçant le déséquilibre de liaison dans un premier temps. Ces deux cas de propagation, sélection et dérive, sont a priori fréquents dans les populations animales.

### **III.B. Migration et mélange de populations**

Une autre voie de création du déséquilibre de liaison est le mélange de populations. En effet, comme le montre la Figure 1.1.3, le mélange de deux populations en équilibre de liaison constitue un ensemble globalement en déséquilibre dès lors que leurs fréquences alléliques sont différentes. Lors des générations suivantes, le déséquilibre décroît généralement, mais la décroissance dépend du régime de reproduction. En cas d'accouplements aléatoires, les locus indépendants deviennent rapidement en équilibre, tandis qu'ils restent en déséquilibre lorsque les accouplements restent partiellement dépendants de la population d'origine. Dans le cas de locus liés, le déséquilibre se maintient de toute façon plus longtemps. Un exemple est le gène rhésus dans les populations humaines. La fréquence de l'allèle *Rh-* est de 45% dans la population noire américaine et de seulement 3% dans la population blanche. Lorsqu'on considère la population nord-américaine dans son ensemble, on identifie un déséquilibre de liaison entre les gènes de la couleur de la peau et le locus rhésus. Ce phénomène appelé « stratification » est responsable de tests d'association faussement positifs.

	Population 1	Population 2	Réunion des populations 1 et 2																											
	<table border="1"> <tr> <td></td> <td>B</td> <td>b</td> </tr> <tr> <td>A</td> <td>81</td> <td>9</td> </tr> <tr> <td>a</td> <td>9</td> <td>1</td> </tr> </table>		B	b	A	81	9	a	9	1	<table border="1"> <tr> <td></td> <td>B</td> <td>b</td> </tr> <tr> <td>A</td> <td>2</td> <td>8</td> </tr> <tr> <td>a</td> <td>1</td> <td>89</td> </tr> </table>		B	b	A	2	8	a	1	89	<table border="1"> <tr> <td></td> <td>B</td> <td>b</td> </tr> <tr> <td>A</td> <td>83</td> <td>17</td> </tr> <tr> <td>a</td> <td>10</td> <td>90</td> </tr> </table>		B	b	A	83	17	a	10	90
	B	b																												
A	81	9																												
a	9	1																												
	B	b																												
A	2	8																												
a	1	89																												
	B	b																												
A	83	17																												
a	10	90																												
$p_A$	0,90	0,10	0,50																											
$p_a$	0,10	0,90	0,50																											
$p_B$	0,90	0,03	0,465																											
$p_b$	0,10	0,97	0,535																											
$D$	0	0	0,183																											

**Figure 1.1.3 : Exemple de création de déséquilibre de liaison par fusion de deux populations en équilibre de liaison et d'effectifs égaux**

Ce phénomène est courant dans les populations « hybrides » issues de croisements entre animaux de races distinctes. C'est le cas des croisements pratiqués entre races de bovins allaitants dans certains systèmes d'élevage (notamment en Irlande ou en Nouvelle-Zélande) afin de profiter de la vigueur hybride. Ce type de croisement est également pratiqué en vue de produire des lignées commerciales qui allient les qualités des deux (ou plus) races d'origine (lignées synthétiques porcines ou lignées commerciales chez la poule).

### III.C. Dérive génétique

Les populations animales de production sont généralement de taille efficace ( $N_e$ ) réduite, c'est-à-dire que seul un petit nombre d'individus indépendants contribue génétiquement à l'évolution de cette population. Plus ce nombre d'individus est réduit, plus la dérive génétique est forte (Hill et Robertson, 1968), et par conséquent la perte de variabilité génétique élevée. La dérive génétique correspond aux fluctuations aléatoires de fréquences alléliques liées à l'échantillonnage d'un nombre fini de reproducteurs, fluctuations pouvant aller jusqu'à la perte de certains allèles dans la population. Ainsi, pour la plupart des races de bovins laitiers français, cet effectif efficace a été estimé à moins de 50 individus (Boichard *et al.*, 1996). Des estimations du même ordre de grandeur ont été obtenues en 1998 par Maignel *et al.* pour des

rares porcines françaises (113 individus en race Large White, 78 en Piétrain et 55 en Landrace Français).

La dérive génétique est responsable de la disparition aléatoire de certains allèles et/ou de certains haplotypes. La réduction du nombre d'haplotypes conduit à une limitation de la décroissance du déséquilibre de liaison au cours des générations, certains allèles ne se trouvant plus liés qu'à un nombre réduit d'autres allèles aux locus proches. En effet, l'espérance de  $r^2$  est égale à :  $E(r^2) = \frac{1}{1 + 4N_e c}$ . Cette formule traduit l'existence d'un

déséquilibre asymptotique d'autant plus important que les locus sont plus liés et que la population est d'effectif génétique réduit (Figure 1.1.4). Deux raisons expliquent ce phénomène : (a) chaque reproducteur génère un déséquilibre intra famille et leur nombre limité ne permet pas de diluer ce phénomène sur un grand nombre ; (b) l'augmentation de consanguinité conduit à une augmentation d'homozygotie, rendant inefficaces les recombinaisons pour réduire le déséquilibre de liaison.

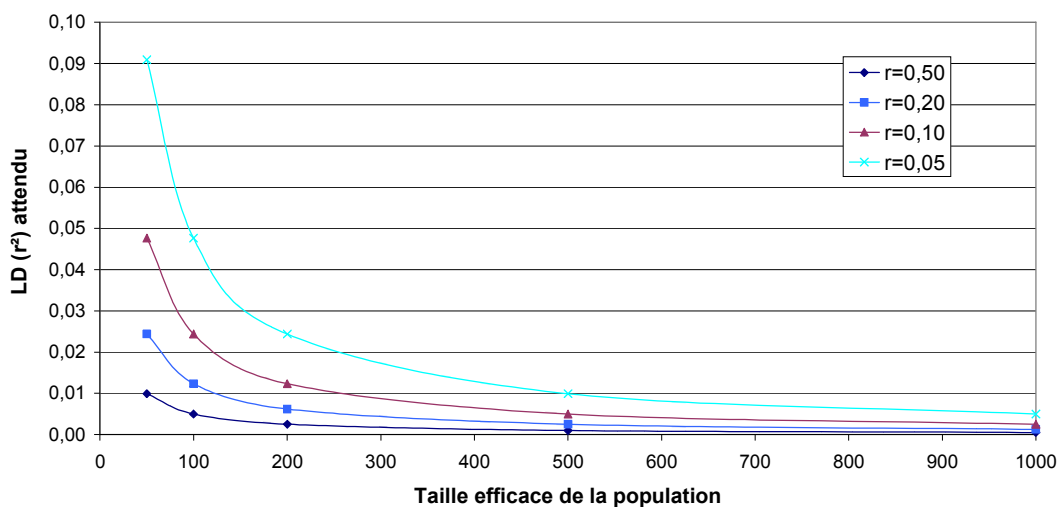


Figure 1.1.4 : Evolution du LD attendu (mesuré par le  $r^2$ ) en fonction de la taille efficace de la population et du taux de recombinaison ( $r$ )

### III.D. Sélection

Les relations entre la sélection et le LD ont été largement étudiées (Felsenstein, 1965 ; Hill et Robertson, 1966 ; Nei, 1967). La sélection induit plusieurs phénomènes : (a) comme la dérive génétique, elle induit une diminution du nombre de reproducteurs et donc une réduction du nombre d'haplotypes présents dans la population ; (b) elle augmente

l'apparementement entre les reproducteurs et donc la fréquence des allèles favorables ainsi que la consanguinité ; (c) elle induit un déséquilibre entre locus soumis à sélection (les combinaisons d'allèles défavorables étant sous-représentées). En conséquence, la sélection limite la décroissance du déséquilibre de liaison au cours des générations, voire peut en générer si on perd ou fixe des haplotypes.

La faible taille efficace des populations de production animale, par exemple des bovins laitiers en races Holstein et Normande, est très liée à la forte sélection à laquelle sont soumises ces races, notamment sur la voie mâle. En effet, l'utilisation massive de l'insémination artificielle a permis une plus grande diffusion du patrimoine génétique de certains individus, réduisant par là même le nombre de reproducteurs mâles utilisés dans l'ensemble de la population. Il est également important de noter que cette diminution du nombre de reproducteurs mâles a entraîné une hausse de la consanguinité, qui influence aussi le déséquilibre de liaison. Même si elle n'en engendre pas, elle est responsable d'une décroissance plus lente du LD à cause de la diminution de la taille efficace de la population et du déficit induit en hétérozygotes, qui réduit l'effet de la recombinaison sur le LD.

#### **IV. Déséquilibre de liaison dans des populations animales**

Compte tenu des forces précédemment évoquées, il convient de définir ce que serait un échantillon idéal pour l'étude du LD dans des populations animales. Tout d'abord, celui-ci devrait être composé d'individus aussi peu apparentés que possible, afin de représenter au mieux la variabilité haplotypique présente dans la population. L'étude d'un échantillon composé d'individus apparentés reviendrait à considérer une population de plus faible taille efficace, et donc probablement à surévaluer le LD. Ceci peut aussi conduire à la mise en évidence de locus en LD situés sur des chromosomes différents (locus non synténiques) du fait du hasard de la répartition des chromosomes dans les gamètes lors de la méiose. Un échantillon « idéal » ne correspond donc pas aux populations expérimentales classiques qui ont des structures familiales. Pour décrire le LD à l'échelle d'une population, il convient également de considérer un grand nombre de marqueurs permettant d'avoir une couverture dense du génome. Si l'objectif de l'étude est d'évaluer l'étendue du LD dans la population, il est nécessaire de sélectionner des marqueurs « neutres », c'est-à-dire situés dans des régions chromosomiques qui ne sont pas soumises à la sélection. Si les marqueurs retenus ne sont pas

Première partie : revue bibliographique  
Partie 1.1. Déséquilibre de liaison dans les populations animales

« neutres », des locus situés dans des régions chromosomiques co-sélectionnées seront en LD, ce qui introduit un biais dans l'inférence du LD à grande distance.

Les études sur données réelles dans des populations animales ont été initiées en l'an 2000 avec le travail de Farnir *et al.*. Il évalue le LD dans des populations de bovins laitiers avec le  $D'$ . Les études qui lui firent suite employèrent cette même mesure, qui fait donc référence dans la littérature, même si d'autres mesures sont également proposées. Aujourd'hui, il subsiste peu d'espèces de rente pour lesquelles aucune analyse n'a été réalisée. Il faut cependant noter que ces études sont menées sur des populations de structures très différentes, avec une densité et un nombre de marqueurs génétiques très variables, comme l'illustre le Tableau 1.1.2.

Première partie : revue bibliographique  
Partie 1.1. Déséquilibre de liaison dans les populations animales

Etude	Type de population	Mesure du LD	Effectif génotypé	Nombre de marqueurs	Principales conclusions
Farnir <i>et al.</i> (2000)	22 familles de demi-frères	D'	949 taureaux + 581 haplotypes maternels	284	* Décroissance rapide du LD avec la distance entre locus, mais LD à grande distance et entre locus non synténiques
	4 familles de pères		627 vaches + leurs parents	8 + 3 (BTA6) + 16 (BTA14)	
McRae <i>et al.</i> (2001)	Croisement Texel x Coopworth (ovins)	D' (LD à grande distance)	276 descendants + pères et grands-pères	90 (10 chromosomes)	* LD à grande distance et entre locus non synténiques
		D' (LD à faible distance)	482 descendants de 14 familles de pères	26 marqueurs (chromosome 6)	* Décroissance rapide du LD avec la distance entre locus * Valeurs du D' plus hétérogènes pour des locus proches
Lou <i>et al.</i> (2003)	Back-cross Greyhound x Labrador retriever	r <sup>2</sup>	147 individus issus des différentes générations	240 (39 chromosomes)	* Décroissance rapide du LD avec la distance entre locus, mais LD à grande distance et entre locus non synténiques
Vallejo <i>et al.</i> (2003)	Taureaux d'élite faiblement apparentés	Arlequin 2.0 (Schneider <i>et al.</i> , 2000)	23	54	idem Farnir <i>et al.</i>
Heifetz <i>et al.</i> (2005)	3 lignées commerciales de poulets	* D' * $\chi^2$	22 à 96, (apparemment entre individus limité)	24 (3 chromosomes) à 120 (3 chromosomes)	* Décroissance rapide du LD avec la distance entre locus * LD jusqu'à 20 cM, mais pas entre locus non synténiques
Harmegnies <i>et al.</i> (2006)	2 lignées commerciales porcines	* LRT (LE vs. LD) * r <sup>2</sup>	33 + 40	29 (SSC15) + 5 (SSC2p)	* Décroissance rapide du LD avec la distance entre locus * LD à grande distance, mais pas entre locus non synténiques
Khatkar <i>et al.</i> (2006)	Taureaux d'élite	D'	433	220 SNP sur le BTA6	idem Farnir <i>et al.</i>
Odani <i>et al.</i> (2006)	Familles de pères de 2 races (bovins)	D'	162 + 406	246 + 156	* LD à grande distance mais pas entre locus non synténiques
Gautier <i>et al.</i> (2007)	14 races bovines	r <sup>2</sup>	1800	696 SNP informatifs (surtout BTA3)	* Décroissance rapide du LD avec la distance entre locus

**Tableau 1.1.2 : Synthèse de résultats dans quelques études de LD menées sur des populations réelles d'animaux de rente**

Un grand nombre de ces études s'accordent sur :

- la présence d'un LD non nul à grande distance (jusqu'à 40 cM, voire plus) et même entre locus non synténiques, confirmant la première étude de Farnir et al (2000).
- une décroissance rapide du niveau de LD avec l'augmentation de la distance entre les marqueurs.

Deux d'entre elles (McRae *et al.*, 2001 ; Lou *et al.*, 2003) soulignent particulièrement la variabilité du LD entre des locus proches, ce qui signifie que, malgré la proximité de deux locus, ils peuvent ne pas paraître en LD, ce qui pourrait être un frein pour la cartographie fine.

L'étude de Heifetz *et al.* (2005) s'appuie sur 3 lignées commerciales de poules pondeuses. L'apparemment le plus fort entre individus était une relation de demi-frères et sœurs, et ce uniquement lorsque l'échantillon était de taille trop faible. Le nombre de marqueurs microsatellites variait entre 24 et 120 selon le nombre de chromosomes couverts (2 à 4) et la lignée considérée ; le nombre maximum d'individus pris en compte simultanément était de 96. Les auteurs ont utilisé le  $\chi^2$ ' et  $D'$  pour évaluer le LD, mettant en évidence une probable surévaluation de l'étendue du LD lorsque celui-ci est évalué avec le  $D'$ . En mesurant le LD avec le  $\chi^2$ ', les auteurs ont obtenu une proportion 100 fois plus faible de locus non synténiques ayant une très forte valeur de LD, comparée à celle enregistrée avec le  $D'$ , et un LD très fort principalement pour des marqueurs distants de moins de 5 cM. Ils ont également constaté que la réduction de la taille des échantillons d'individus augmentait le nombre de valeurs de LD très fortes, probablement en raison d'haplotypes manquants dans l'échantillon.

Les différences entre cette étude et les études antérieures peuvent être dues à différents facteurs :

- elle repose sur une population probablement moins apparentée que les autres, les lignées commerciales étant classiquement issues du croisement d'un grand nombre de lignées consanguines. Celles utilisées par Harmegnies *et al.* (2007) semblent hautement consanguines car elles sont maintenues en noyau fermé depuis au moins 20 générations, ce qui n'était peut-être pas le cas des lignées commerciales de Heifetz *et al.*,
- elle implique plus d'individus que l'étude de Vallejo *et al.* (2003),
- elle repose sur une densité de marqueurs assez forte, ces marqueurs étant *a priori* situés dans des zones assez neutres du génome,

- enfin, elle se fonde sur le  $\chi^2$  qui est moins sensible que le  $D'$  à l'absence de certains haplotypes qui devait être fréquente dans les autres études compte tenu de l'apparementement entre les individus.

La plupart de ces critères correspondent aux forces évolutives préalablement citées, et peuvent être responsables d'une éventuelle surévaluation de l'étendue du LD dans les populations animales. Cependant, compte tenu de la pression de sélection à laquelle ces populations sont soumises, il était attendu que le LD s'étende sur des distances assez longues.

Gautier et al (2007) présentent une étude à grande échelle sur l'étendue du LD intra et entre races bovines européennes et africaines, à l'aide de SNP principalement du chromosome 3. S'ils retrouvent le LD a grande distance, l'étude de la structure haplotypique suggère qu'au sein des races le génome s'organise en une mosaïque de blocs d'haplotypes d'une taille de quelques centaines de kilobases (kb), eux mêmes étant composés d'une mosaïque plus fine de blocs de moins de 10 kb conservés entre races. Ces petits blocs correspondent à des empreintes laissées au cours de la période précédant les débuts de la domestication. Cette structure haplotypique suggère qu'en considérant plusieurs races simultanément dans les projets de cartographie fine de gènes d'intérêt, il est possible de décupler la résolution attendue.





## **Partie 1.2.**

# **Cartographie fine de QTL**



**- Article 1 -**  
**Comment affiner la localisation d'un QTL ?**

YTOUNEL Florence, GILBERT Hélène, BOICHARD Didier

**Résumé :**

Au cours des trente dernières années, de nombreuses méthodes ont été proposées pour cartographier finement des locus impliqués dans la variabilité des caractères quantitatifs (Quantitative Trait Loci ou QTL). La variabilité de ces méthodes s'explique notamment par les différences qui existent entre les espèces, par exemple du fait des intervalles entre générations, des coûts de production, ou de la variabilité des coûts de phénotypage des caractères. Nous proposons ici une revue des méthodes permettant d'accéder à une cartographie fine de ces locus, décrivant au préalable les conditions d'une réduction de l'intervalle de localisation de ces QTL. Le paramètre clé permettant de réduire cet intervalle est le nombre de recombinaisons utiles. C'est pourquoi nous décrivons dans les parties suivantes les méthodes de cartographie fine en distinguant celles qui nécessitent la création de générations supplémentaires pour augmenter le nombre de recombinaisons de celles qui ont recours à des modèles statistiques pour inférer des recombinaisons ayant eu lieu au cours des générations qui ont précédé la génération actuelle.

- Article 1 -

**Introduction**

La variabilité des caractères quantitatifs mesurés dans les populations animales de rente résulte généralement de l'action combinée de gènes en nombre variable, mais souvent élevé, et d'effets de l'environnement. La variabilité d'origine génétique est due à des polymorphismes génétiques, c'est-à-dire des variations dans le code génétique, modifiant l'effet des gènes. Ce sont ces polymorphismes génétiques dits mutations causales, ou du moins les plus importants d'entre eux, que les généticiens cherchent à identifier à la fois pour comprendre le déterminisme des caractères et pour les sélectionner. Les locus impliqués dans la variation des caractères quantitatifs sont appelés Quantitative Trait Loci (QTL). Ils sont caractérisés par deux paramètres : leur effet sur le caractère cible et leur position sur le génome. L'identification et la caractérisation d'un QTL impliquent à la fois des approches de cartographie et des approches fonctionnelles. On considère généralement, sauf argument fonctionnel très fort, que l'approche cartographique est nécessaire dans une première approche pour réduire le nombre de gènes candidats fonctionnels. L'approche cartographique consiste à localiser le QTL dans une région chromosomique aussi réduite que possible. Si la cartographie des gènes majeurs est souvent assez précise, de l'ordre du centiMorgan (cM), celle des QTL (qui expliquent une fraction sensiblement inférieure de la variabilité génétique du caractère) l'est beaucoup moins, avec des intervalles de localisation de l'ordre de plusieurs dizaines de centiMorgan. Ces intervalles correspondent à plusieurs dizaines de millions de bases et à plusieurs centaines de gènes, rendant aléatoire l'identification de gènes candidats fonctionnels. La recherche du ou des gènes sous-jacents aux QTL doit donc être précédée d'une étape supplémentaire dite de cartographie fine.

**- Article 1 -**

La cartographie fine correspond à une étape de réduction de la taille de l'intervalle de localisation du QTL. Par le passé, cette étape consistait à augmenter à la fois le nombre de marqueurs et le nombre d'individus pris en compte dans le dispositif. De plus en plus aujourd'hui, des méthodes à haut débit, rendues possibles dans un nombre croissant d'espèces par la disponibilité de la séquence de leur génome complet, permettent de saturer le génome en marqueurs dès l'étape de primolocalisation. Il n'en demeure pas moins que l'étape de cartographie fine par extension du dispositif reste généralement nécessaire pour disposer d'une localisation précise du QTL. Dans le même temps, des méthodes statistiques particulières ont été développées pour exploiter au mieux ces données pour la cartographie fine.

Notre objectif est ici de présenter les principes généraux de la cartographie fine de QTL, puis de nous concentrer sur les facteurs influençant la résolution de la cartographie ainsi que sur les méthodes développées pour affiner la localisation du QTL.

**I. Principes et facteurs influençant la résolution de la cartographie**

**I.A. Principe général**

Le principe général de la cartographie de QTL est ancien (Niemann-Sorensen et Robertson, 1961) et on pourra se reporter à Boichard *et al.* (1998) pour une synthèse en français. Reposant sur les propriétés de la méiose, il utilise le fait que de larges segments chromosomiques sont transmis en bloc du parent au descendant avec conservation de l'association entre locus proches chez le parent. Rechercher des QTL consiste à rechercher des marqueurs dont le polymorphisme présente une association statistique avec le phénotype (Figure 1). S'il est relativement facile de détecter un QTL, il est plus complexe de le localiser

- Article 1 -

précisément. En effet, avec un seul marqueur par exemple, on ne peut pas faire la différence entre un QTL éloigné avec un effet fort, et un QTL proche avec un effet plus réduit, les deux conférant au marqueur le même effet apparent égal à  $a(1-2r)$ ,  $a$  étant l'effet du QTL (qui est l'écart phénotypique moyen entre les homozygotes QQ et qq) et  $r$  le taux de recombinaison entre marqueur et QTL.

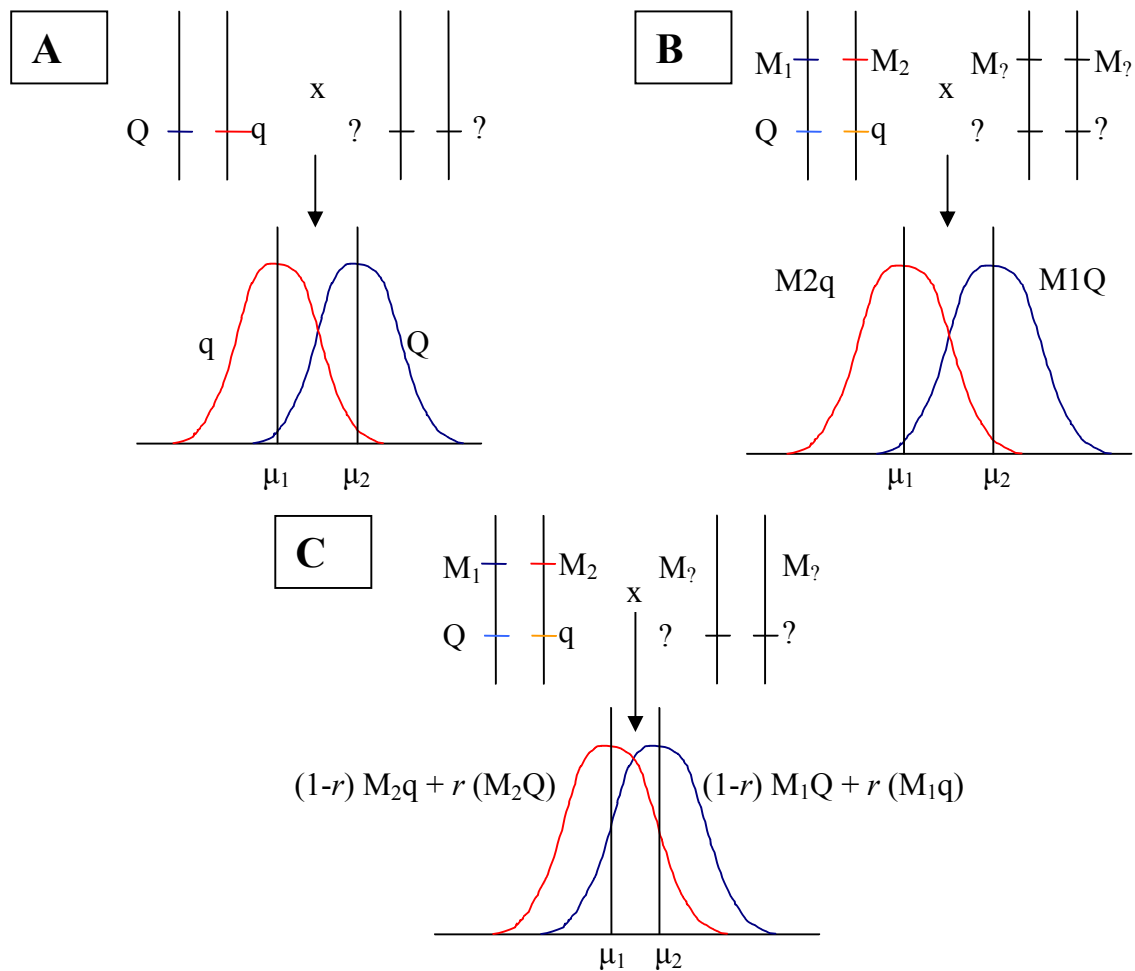


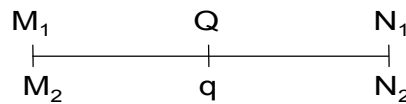
Figure 1 : Ségrégation des allèles d'un QTL (Q et q) dans la descendance d'un individu hétérozygote. A : sans marqueur ; B : avec un marqueur M (M<sub>1</sub> et M<sub>2</sub>) complètement lié au QTL, C : avec un marqueur M (M<sub>1</sub> et M<sub>2</sub>) incomplètement lié au QTL

**- Article 1 -**

Pour distinguer position et effet des QTL, Lander et Botstein ont proposé en 1989 d'utiliser l'information apportée par deux marqueurs informatifs encadrant la position testée: c'est la cartographie d'intervalle ou « interval mapping ». Dans l'intervalle considéré, la phase chez le parent (c'est-à-dire la succession des allèles des marqueurs ordonnés le long du chromosome) et les probabilités de recombinaison sont supposées connues. Connaissant les génotypes du parent et du descendant aux marqueurs flanquant la position testée, on estime la probabilité qu'un descendant ait reçu l'un ou l'autre des allèles parentaux à cette position. Cette probabilité est utilisée pour tester par régression ou maximum de vraisemblance la différence de moyennes de performances entre les deux groupes d'individus ayant reçus l'un ou l'autre allèle au QTL en probabilité. Chez des individus non recombinants, la probabilité de transmission est pratiquement constante pour toute localisation comprise entre les deux marqueurs assez proches. En effet, si  $r_1$  est le taux de recombinaison entre le 1<sup>er</sup> marqueur et la position testée et  $r_2$  le taux de recombinaison entre le 2<sup>ème</sup> marqueur et la position testée, la probabilité de transmission de l'allèle QTL parental est  $p=(1-r_1)(1-r_2)/(1-r)$ , proche de 1. On en déduit que seuls les recombinants permettent de localiser le QTL. Pour un gène majeur, la position estimée du gène par rapport à un des deux marqueurs est la proportion d'individus, parmi ceux ayant recombiné entre les deux marqueurs, qui n'ont pas recombiné entre le marqueur et le gène. Pour un QTL (dont le génotype ne peut généralement pas être déduit avec certitude à partir du phénotype), la position montrant l'association la plus significative dans le test de régression ou de maximum de vraisemblance est considérée comme la position la plus probable du QTL.



- Article 1 -



Taux de recombinaison (M-N) =  $2r$

Taux de recombinaison (M-Q) = taux de recombinaison (Q-N) =  $r$

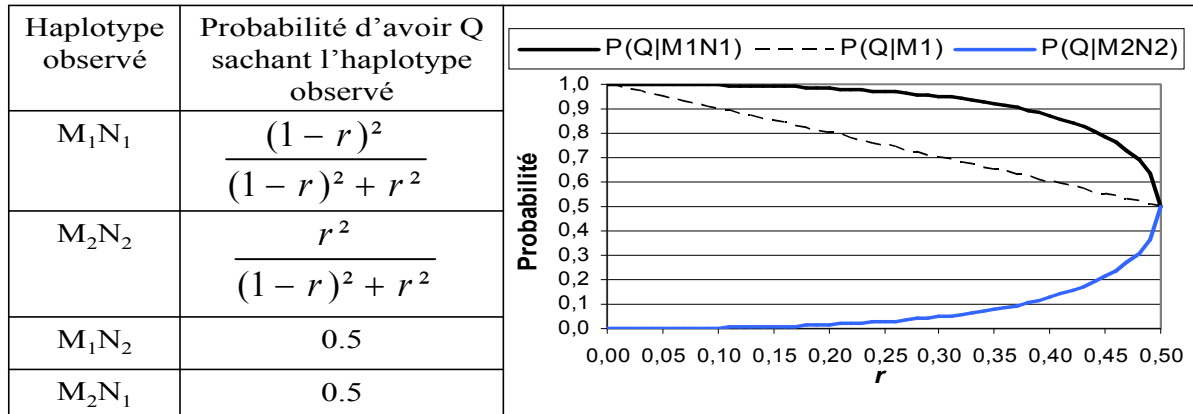


Figure 2 : Probabilités pour chaque haplotype défini par les locus marqueurs M et N de porter Q sachant l'haplotype observé aux marqueurs, et évolution graphique en fonction de  $r$

La résolution atteinte par les méthodes de cartographie relève donc de deux facteurs : l'information apportée par le phénotype et celle apportée par les marqueurs génétiques.

## I.B. Facteurs de précision de la cartographie

### I.B.1. Connaissance du phénotype

Un premier facteur jouant sur la résolution est la capacité à prédire le génotype d'après le phénotype observé. Lorsque le génotype est déduit du phénotype sans incertitude, la précision est maximale. Au contraire, lorsque les distributions des phénotypes conditionnellement aux génotypes se superposent, le génotype n'est déduit qu'en probabilité, ce qui induit une importante perte de précision dans la localisation du QTL. Idéalement, les distributions des phénotypes en fonction des génotypes doivent donc être peu chevauchantes : pour des phénotypes à distribution gaussienne, le cas idéal correspond à une différence entre les moyennes phénotypiques de chaque groupe d'au moins quatre, voire six écart-types résiduels

- Article 1 -

(Figure 3). Cette situation favorable correspond davantage à des gènes majeurs plus qu'à des QTL. Chez le porc, c'est par exemple le cas du gène RN pour lequel le phénotype RTN (rendement à la cuisson) est un excellent prédicteur (Milan *et al.* 1996), ou du gène Hal, pour lequel le test de réaction au gaz Halothane prédit le génotype (Ollivier 1975). Il est à noter que la variabilité intra-génotype au QTL peut dépendre à la fois de la variabilité génétique du caractère due à d'autres locus, de la variabilité environnementale et de la précision de la mesure.

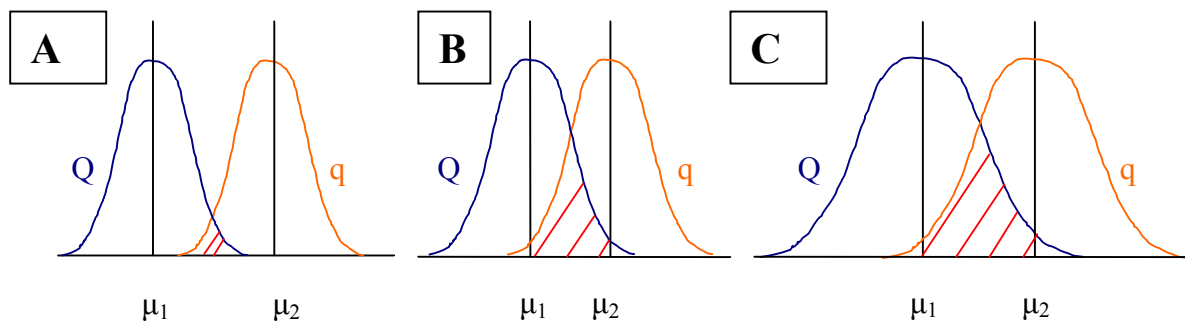


Figure 3 : Distribution des phénotypes pour deux génotypes au QTL selon l'effet du QTL (A vs. B), ou la variabilité du phénotype (A vs. C)

Pour optimiser ces critères, plusieurs types de stratégies ont été proposés : ne retenir que des individus de phénotypes extrêmes pour inférer leur génotype avec certitude, diminuer la variabilité résiduelle relativement à l'effet du QTL, ou prendre simultanément en compte plusieurs phénotypes de caractères corrélés.

I.B.1.a. Génotypage sélectif

Cette méthode, proposée en 1987 par Lebowitz *et al.*, et adaptée en 1989 par Lander et Botstein aux caractères quantitatifs, exploite le fait que l'essentiel de l'information sur le lien entre le QTL et le marqueur est apporté par les individus de phénotypes extrêmes : on estime que 81% de l'information quant à la liaison phénotype/génotype est apportée par les 33% des

- Article 1 -

individus s'écartant de plus de un écart-type de la moyenne de la distribution. Cette propriété vient du fait que la connaissance du génotype est généralement bonne pour les extrêmes et beaucoup plus incertaine sur la zone de chevauchement des distributions. Pour conserver la même puissance de détection que si tous les individus avaient été génotypés, il convient, si on ne génotype que les individus de phénotype extrême, d'augmenter le nombre d'individus phénotypés (Lebowitz *et al.* 1987) : augmenter d'un quart le nombre de phénotypes enregistrés permet de maintenir le niveau d'information en divisant par 2,5 le nombre de génotypes à mesurer. Par ailleurs, l'intervalle de confiance de la localisation du QTL reste inchangé si 40 à 50% des individus sont génotypés par rapport au génotypage de l'ensemble de la population. Cette solution s'applique donc principalement à des populations où le nombre d'individus à produire et le nombre de phénotypes à mesurer n'est pas limitant. Dans le cas de cartographie très fine, la localisation est très dépendante de chaque recombinant pris individuellement. En conséquence, aucune erreur sur le phénotype n'est acceptable et cette stratégie est généralement nécessaire pour inférer le plus justement possible les allèles au QTL.

I.B.1.b.Réduction de la variabilité résiduelle

Pour un effet de QTL donné, une meilleure discrimination entre distributions peut être obtenue en maîtrisant mieux leurs variances résiduelles. La répétition des mesures phénotypiques pour un individu permet de limiter la variabilité d'origine environnementale et les erreurs de mesure dans l'estimation du phénotype. Ceci est possible pour un grand nombre de caractères qui peuvent être mesurés plusieurs fois au cours de la vie d'un animal : poids, taille et conformation ; d'autres, mesurables plusieurs fois dans la vie d'un animal, peuvent aussi être considérés comme différents en fonction du numéro de la répétition : production laitière au cours des différentes lactations, taille de portée,... Pour nombre de caractères, la

- Article 1 -

mesure est cependant généralement unique (relative à un moment de la vie – naissance, croissance, abattage) ou son coût trop élevé pour être répétée.

Une autre façon de réduire la variabilité est d'utiliser comme phénotype la moyenne de nombreux descendants. Ainsi, dans le cas du testage sur descendance utilisé dans les dispositifs de détection de QTL dits « petites filles » ou « granddaughter designs » (Weller *et al.* 1990) (Figure 4), le phénotype du fils de première génération est la moyenne des performances de ses propres filles (de seconde génération). Par rapport à un dispositif « filles » (Figure 4) où le phénotype est mesuré directement chez les descendants de première génération (Soller et Genizi 1978), le phénotype présente une variance résiduelle divisée par le nombre de petites-filles. Cette méthode permet de réduire les effets de l'environnement, des polygènes et autres sources de variabilité qui rendent les mesures directes moins précises. Ainsi, dans un dispositif de type « petite-filles », bien que les effets associés au marqueur du grand-père soient divisés par 2 par rapport aux effets estimés chez des filles (l'allèle grand-paternel du fils n'est transmis qu'à une petite-fille sur deux), cette réduction de l'effet est généralement plus que compensée par la réduction de la variance résiduelle. Les dispositifs sur trois générations sont donc généralement plus puissants que ceux à deux générations (Weller *et al.* 1990, Van der Beek *et al.* 1995), pour des coûts de génotypage réduits. Au-delà du gain de puissance qui est connu, on obtient également un gain de précision de localisation du QTL. On peut également combiner performances propres et performances de la descendance (Tribout *et al.*, 2008)

- Article 1 -

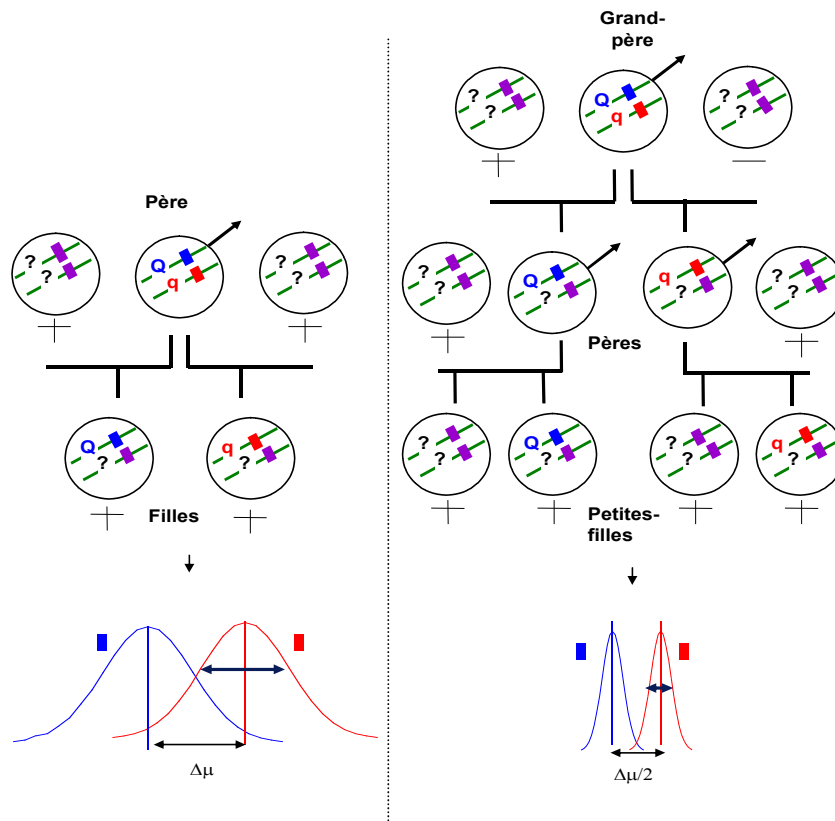


Figure 4 : Dispositifs « filles » (à gauche) et « petites-filles » (à droite) et distributions des phénotypes selon l'allèle hérité au QTL tracé par le marqueur A/a

I.B.1.c. Utilisation de caractères corrélés

La différence de phénotype entre génotypes peut parfois être accentuée en considérant des caractères corrélés au caractère d'intérêt. Citons l'exemple du gène RN mis en évidence par le phénotype RTN pour lequel une différence de 3 écarts types est mise en évidence, mais qui induit de surcroît une différence de potentiel glycolytique dans le muscle de 7 écarts types. Ce nouveau phénotype permet une discrimination parfaite entre les génotypes, ce qui est un avantage considérable dans l'étape de cartographie fine.

L'utilisation des distributions conjointes entre caractères pour affiner la connaissance d'un phénotype donné a été proposée par Weller *et al.* (1996). Elle permet d'affiner la localisation du QTL, mais aussi de préciser ses effets grâce à la corrélation entre les caractères. Le gain

- Article 1 -

d'information obtenu par ces stratégies s'avère particulièrement important lorsque les corrélations dues au QTL ont des directions différentes des corrélations phénotypiques entre les caractères (Korol *et al.* 1995). L'utilisation de la seule corrélation résiduelle entre les caractères permet également d'augmenter l'information et la puissance du test (Korol *et al.* 1995, Gilbert et Le Roy 2003).

Les données d'expression, qui peuvent par exemple refléter dans un tissu l'expression différentielle de gènes entre deux états (animal sain / malade, carencé ou non,...), peuvent elles aussi être utilisées comme des phénotypes. On parlera alors de détection de QTL d'expression ou de eQTL. Bien que des développements méthodologiques soient encore nécessaires, on peut imaginer que l'accès à l'expression de certains gènes ou certaines combinaisons de gènes impliqués dans la différence entre phénotypes montre des différences très supérieures à la différence entre génotypes. De façon analogue au lien entre le gène RN et le potentiel glycolytique du muscle cité plus haut, ce phénotype nouveau permet alors une cartographie beaucoup plus fine.

Un problème fréquemment rencontré en cartographie de QTL est l'hétérogénéité génétique de phénotypes semblables. Les données d'expression peuvent permettre de préciser le phénotype et de concentrer l'étude sur des groupes homogènes, dont la différence est liée uniquement au QTL recherché (Schadt et Lum, 2006).

I.B.2. Informativité des marqueurs

Dans un dispositif donné, la résolution est maximale lorsque chaque événement de recombinaison est localisé exactement. En pratique, ceci nécessite pour chaque recombinaison de disposer de deux marqueurs très proches, informatifs et flanquants. Ceci peut être obtenu

- Article 1 -

en saturant le génome, ou tout au moins la région considérée, par des marqueurs génétiques informatifs nombreux.

Jusqu'à une date récente, les dispositifs usuels de cartographie de QTL étaient réalisés avec un maillage de marqueurs peu dense, soit un marqueur tous les 10 à 20 cM. Une région d'intérêt n'était donc couverte que par 2 à 3 marqueurs, pas toujours informatifs. La phase de cartographie fine impliquait donc de rajouter de nombreux marqueurs dans cette région. Avant le séquençage des génomes complets, ces marqueurs n'étaient pas toujours disponibles et étaient souvent développés spécifiquement.

Avec les nouvelles technologies de génotypage, les phases de primolocalisation et de génotypage dense sont simultanées, grâce à l'utilisation de puces permettant le génotypage simultané de plusieurs dizaines de milliers de marqueurs couvrant tout le génome. Même si cette densité n'est pas encore totalement optimale (on considère qu'un marqueur tous les 10 kilobases serait en théorie nécessaire pour obtenir la résolution maximale dans les espèces d'élevage, soit une puce de 300 000 SNP), ces puces constituent déjà un progrès considérable en termes de temps, de coût et de travail. Elles ne sont par ailleurs pas encore disponibles pour toutes les espèces d'animaux de rente.

Cependant, une forte densité en marqueurs informatifs est une condition nécessaire mais pas suffisante pour obtenir une localisation précise des QTL. En effet, Darvasi *et al.* (1993) ont démontré que, pour un effet de QTL et une taille de population donnés, on peut définir une résolution maximale du dispositif expérimental, limitée par le nombre de recombinaisons informatives présentes dans le dispositif. Par définition, entre deux marqueurs séparés de 1 cM (soit 1 million de paires de bases en moyenne chez les mammifères), il y a une recombinaison toutes les cent méioses. Les dispositifs expérimentaux classiques pour la cartographie de QTL, composés en général de quelques centaines d'individus répartis en

- Article 1 -

familles de 50 à 100 descendants, ne permettent d'espérer que quelques recombinaisons par cM, ce qui limite la précision de localisation à l'échelle du dispositif de détection à des segments de plusieurs dizaines de cM. Il est donc généralement admis qu'une forte augmentation de la densité des marqueurs dans ce type de dispositif n'augmente que faiblement la résolution offerte par les dispositifs sur deux ou trois générations, le facteur limitant étant le nombre de recombinaisons. Darvasi et Soller (1997) obtiennent deux estimations de l'intervalle de confiance  $I$  de localisation (en cM) dans des dispositifs expérimentaux F2 ou Back Cross complètement informatifs (les parents F1 sont hétérozygotes au QTL) et la carte est saturée en marqueurs :

$$I = 3000 / (N d^2)$$

$$I = 530 / (N p)$$

où  $N$  est le nombre de méioses informatives,  $d$  l'effet de substitution et  $p$  la proportion de la variance phénotypique expliquée par le QTL. Il est à noter que dans certains dispositifs animaux intra race usuels, seules les méioses paternelles sont utilisées et seulement une fraction des pères est hétérozygote au QTL, ce qui limite la valeur de  $N$ .

La disponibilité de nombreuses recombinaisons et la capacité à les identifier sont donc cruciales pour la cartographie fine. La recombinaison est un phénomène aléatoire : pour en obtenir plus, il faut augmenter le nombre de méioses informatives disponibles. Une première solution est de maximiser la proportion de parents hétérozygotes au QTL, d'où la stratégie fréquente de croisement entre animaux divergents pour procréer des parents F1. Dans les populations commerciales, le choix des descendances les plus variables peut être intéressant. Dans la mesure du possible, l'utilisation d'un dispositif avec les deux méioses parentales utilisables est à privilégier, ce qui incite à constituer de larges familles non seulement de père mais aussi de mère. L'utilisation d'un très grand dispositif expérimental sur une génération est



- Article 1 -

une option possible mais coûteuse : Darvasi et Soller (1995) ont démontré que, pour cartographier un QTL ayant un effet de substitution  $\alpha = 0,25$  avec un intervalle de confiance à 95% de 5 cM, il faut 4500 individus F2.

Deux autres stratégies peuvent être privilégiées : la procréation expérimentale de nouveaux individus, et l'utilisation de méiotes qui ont eu lieu au cours des générations passées (recombinaisons « historiques »). Des méthodologies différentes sont alors utilisées, basées sur des stratégies dites de dissection chromosomique dans le premier cas, et de nouvelles méthodes statistiques dédiées à la cartographie fine dans le deuxième cas.

## II. Dissection chromosomique

La dissection chromosomique résulte de la production de nouvelles méiotes sur plusieurs générations afin d'identifier les recombinaisons produites et de réduire progressivement la longueur de la région chromosomique d'intérêt. Les *lignées d'intercroisement avancé* (ou AIL pour Advanced Intercross Lines) sont créées à partir de croisements entre individus de la génération  $F_i$  précédente, en évitant les accouplements frère-sœur (Figure 5). Les recombinaisons ainsi obtenues sont accumulées dans une seule population et réduisent la taille des segments parentaux initiaux. Ces croisements sont optimaux en terme de nombre de méiotes utiles. Ainsi Darvasi et Soller (1995) ont démontré que, sans génotypage supplémentaire et après 10 générations, l'intervalle de confiance de la localisation d'un QTL peut idéalement être jusqu'à cinq fois plus petit qu'avec un dispositif BC ou F2. La ségrégation des allèles au QTL génère ici trois génotypes au QTL, comme dans les populations F2. L'existence d'un effet est donc testée de la même manière. L'avantage des AIL est particulièrement important lorsque les populations parentales sont consanguines :

- Article 1 -

dans ce cas, seuls deux allèles par marqueur, d'origine connue, sont en ségrégation dans la population d'AIL, simplifiant grandement l'analyse. Lorsque les populations initiales ne sont pas fixées, l'analyse nécessite un suivi des segments chromosomiques dans la population, ou se résume à une analyse d'association marqueur par marqueur. Que les populations parentales soient consanguines ou non, la méthode repose sur une forte densité de marqueurs.

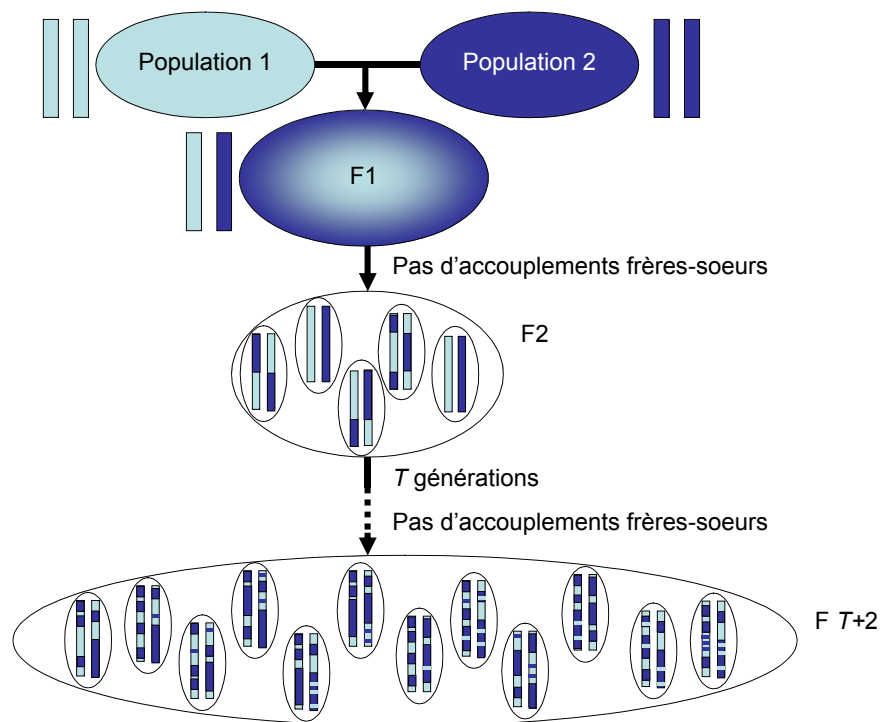


Figure 5 : Schéma d'obtention de *lignées d'intercroisement avancé (AIL)*

Les *lignées recombinantes consanguines* (ou RIL pour *Recombinant Inbred Lines*) sont rares chez les animaux domestiques car très coûteuses à obtenir. Elles sont par contre fréquentes chez les végétaux ou la souris. Issues d'un certain nombre de générations d'accouplements frère-sœur (Figure 6A) ou d'autofécondations après la F1, ces lignées sont génétiquement fixées pour la plus grande partie de leur génome. Elles peuvent être utilisées en cartographie (fine) de deux façons. D'une part, comme pour une F2, la présence d'un QTL peut être testée par une différence phénotypique significativement associée à une région différenciellement

- Article 1 -

fixée entre lignées. On observe un gain par rapport à la F2 dans la mesure où le nombre de générations réalisées correspond à une accumulation de recombinaisons et donc une augmentation du taux de recombinaison équivalent sur une génération, dans un rapport qui tend vers  $4r(1+6r)$  pour les croisements frères-sœurs et  $2r(1+2r)$  pour les autofécondations,  $r$  étant le taux de recombinaison observé dans la F2 (Lynch et Walsh 1998). Les segments conservés sont donc de plus petite taille, ce qui améliore la résolution. On gagne également en résolution par une diminution de la variance intra-lignée et par l'information du dispositif, tous les individus étant homozygotes et totalement informatifs (AA vs aa à chaque locus). La puissance des RIL est généralement très élevée, surtout si le caractère est gouverné par peu de QTL et qu'il a une faible héritabilité, du fait de la réduction de la variabilité environnementale et polygénique qu'elles permettent. Les RIL sont aussi un outil facilitant grandement la recherche des épistasies, qui sont les interactions existant entre deux ou plusieurs gènes, en individualisant des combinaisons de gènes. Autre atout, comme toute lignée fixée, les RIL, si elles sont conservées, peuvent être phénotypées pour tout nouveau caractère et ne nécessitent pas de génotypage nouveau.

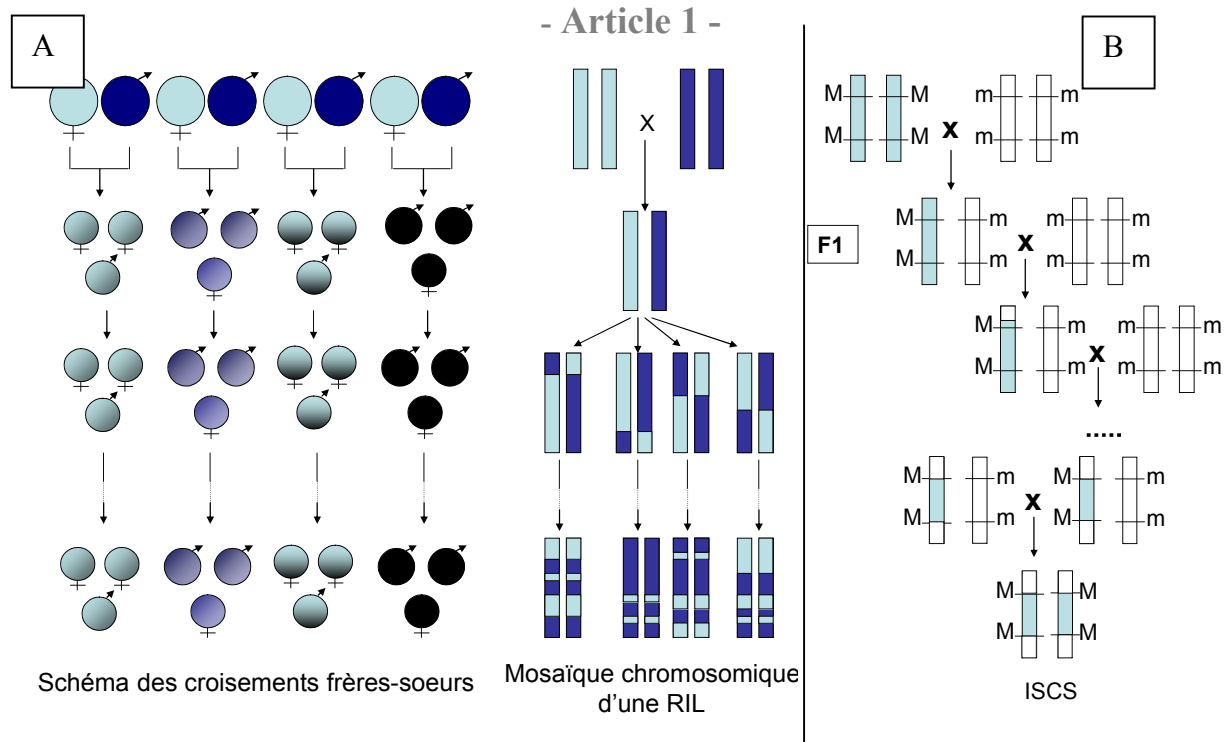


Figure 6 : Lignées congéniques. A : dispositif pour l'obtention de *lignées recombinantes consanguines* et chromosomes obtenus. B : Obtention d'une *lignée congénique spécifique d'un intervalle*

Les RIL présentent un autre avantage moins connu en cartographie fine. Elles gardent généralement une petite fraction du génome non fixée. Utilisée comme parent de dispositif de détection de QTL, un RIL peut exprimer une ségrégation du caractère, ce qui permet de localiser finement le QTL à l'intérieur de la région encore hétérozygote, région qui peut ensuite être réduite par des recombinaisons dans une génération suivante. Cette approche a été utilisée avec succès par Loudet *et al.* (2007) chez *Arabidopsis*.

Une approche comparable est parfois utilisée chez les espèces d'élevage, le porc par exemple, dans le cadre de back-cross successifs. On choisit comme parents des individus recombinants à différents points de la région étudiée (Figure 7), donc homozygotes ou hétérozygotes au QTL recherché. Ces parents sont ensuite accouplés à des conjoints homozygotes. On déduit le statut du parent en fonction de la présence ou l'absence de ségrégation du QTL chez ses

- Article 1 -

descendants. Ceci permet de localiser le QTL dans une région donnée (le rectangle rouge sur la figure 8). Cette méthode est très efficace pour des régions de taille moyenne, mais elle nécessite des effectifs très élevés lorsque la région devient petite et n'est réaliste que lorsque la prolificité est élevée et le coût d'élevage et de phénotypage réduit.

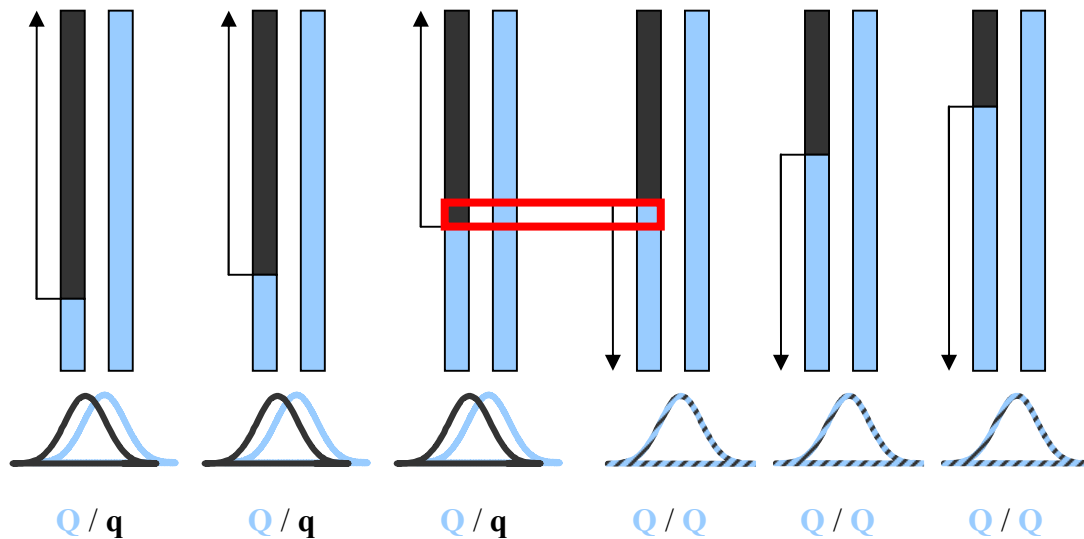


Figure 7. Méthode de dissection chromosomique permettant de localiser le QTL (dans le rectangle rouge). D'après Riquet, communication personnelle

Les *lignées congéniques* (CL) ou *lignées quasi-congéniques* (NIL) sont issues de BC récurrents sur l'une des lignées parentales. Ces lignées sont plus intéressantes si on adjoint au back cross récurrent une introgression assistée par marqueurs. On obtient alors des *lignées congéniques spécifiques d'un intervalle* (*Interval-Specific congenic strains, ISCS*) (Darvasi 1996) qui ne contiennent qu'une fraction de chromosome provenant d'une autre lignée (lignée donneuse) (Figure 6B). Toutes les lignées ainsi créées peuvent posséder un fragment introgressé différent. On peut alors tester la présence d'un QTL par comparaison des moyennes phénotypiques des différentes lignées (L'Hote *et al.*, 2007)

- Article 1 -

**III. Méthodes statistiques de cartographie fine**

Les méthodes précédemment citées reposent sur des populations expérimentales. De telles populations ne peuvent pas être créées dans toutes les espèces d'animaux de rente ou de végétaux, par exemple à cause d'intervalles de génération trop longs et / ou de coût de production des animaux ou des mesures élevé. Elles ne sont bien sûr pas envisageables chez l'homme. Une alternative pour la cartographie fine de QTL consiste à utiliser les recombinaisons historiques, grâce aux approches de déséquilibre de liaison. Entre deux générations consécutives, et entre deux locus séparés par une distance  $c$ , le LD diminue d'un facteur  $(1-c)$  à chaque génération selon la formule:  $D_t = (1-c)^t D_0$ , ce qui permet d'estimer le LD à la génération  $t$  en fonction du LD à la génération 0.

On peut distinguer les méthodes statistiques de cartographie fine selon deux grands critères :

- d'une part, en fonction de l'hypothèse faite sur l'origine et la nature du polymorphisme au QTL : apparition d'un second allèle par mutation unique, ou absence d'hypothèse sur le nombre d'allèles au QTL (Terwilliger 1995, Boitard 2006, Farnir *et al.* 2002)
- d'autre part en fonction de la nature de l'information utilisée, LD exclusivement (Spielman 1993, Terwilliger 1995, Boitard 2006), ou combinant LD et l'information de liaison apportée par le pedigree (Meuwissen et Goddard 2000, Farnir *et al.* 2002).

**III.A Méthodes utilisant exclusivement le LD**

**III.A.1 Etude cas-témoin sur une mutation unique dans une population isolée**

Cette méthode, proposée en 1992 par Hästbacka *et al.*, suppose qu'une mutation unique est apparue dans un haplotype (haplotype fondateur, cf. Figure 8)  $t$  générations avant la génération actuelle. Elle utilise directement la modélisation de la décroissance du LD avec le temps.

- Article 1 -

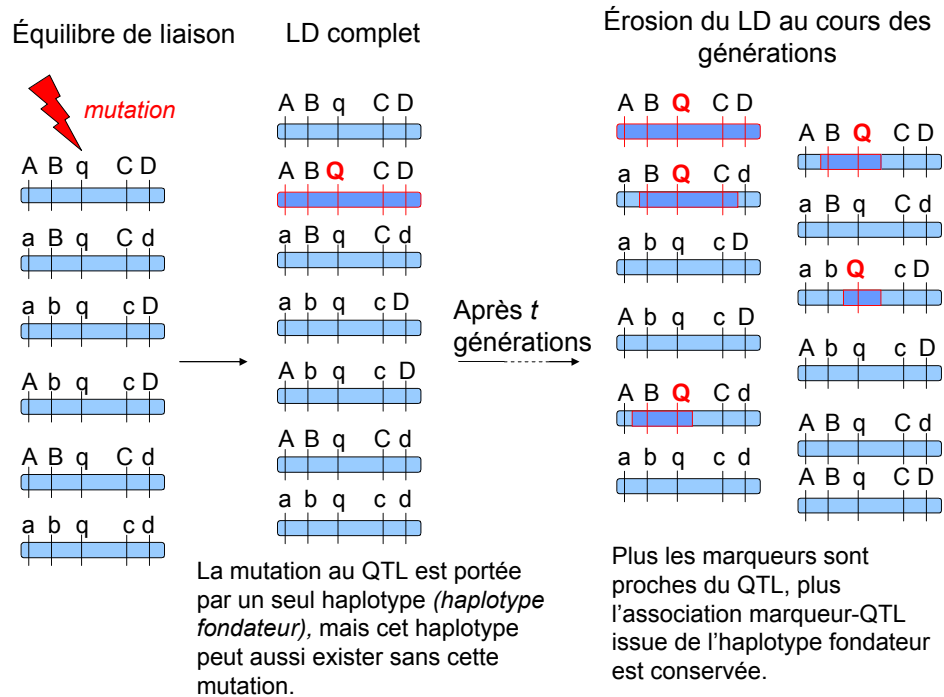


Figure 8 : Haplotype fondateur et érosion du LD au cours des générations

Le principe est d'identifier un haplotype très majoritaire dans la population malade. Cet haplotype est alors considéré comme la copie exacte de l'haplotype initialement porteur de la mutation, transmis identiquement au cours des générations (haplotype IBD – identique par descendance - avec l'haplotype fondateur). Les autres haplotypes porteurs de l'allèle muté et présents dans la population malade sont alors supposés être issus de recombinaisons. La proportion d'haplotypes IBD dans la population malade est utilisée comme une estimation du rapport du LD restant dans la population par rapport au LD initial, égale à  $(1-c)^t$ , ce qui, lorsque  $c$  tend vers 0 (c'est-à-dire que le marqueur est très proche de la mutation responsable de la maladie), équivaut à  $exp(-ct)$ . En faisant une hypothèse sur l'âge de la mutation, la position du locus peut alors être estimée par rapport à celle du marqueur. L'estimation de la distance entre le locus de la maladie et le marqueur est cependant très sensible à cette hypothèse, alors que cet âge n'est que très rarement connu avec précision. Cette méthode est donc à privilégier uniquement dans le cas de populations isolées d'âge connu ce qui était le

- Article 1 -

cas de la population étudiée dans l'article original de Hästbacka *et al.*. Elle a été appliquée à la cartographie du gène de la dysplasie diastrophique dans la population finnoise avec une résolution de 0,06 cM.

III.A.2 Etude cas – témoin en population ouverte avec des structures familiales

Le Transmission Disequilibrium Test (TDT) (Spielman *et al.* 1993) repose initialement, en génétique humaine, sur l'étude de la (non-)transmission des allèles d'un marqueur bi-allélique entre des parents hétérozygotes et des enfants malades. Il utilise donc des données issues de structures familiales dans des populations de type cas (individus malades) – contrôle (individus sains). Si la ségrégation des locus marqueurs est indépendante de celle du gène de la maladie, alors les individus malades doivent hériter des deux allèles parentaux dans les mêmes fréquences que les individus sains. Le test de différence de fréquences est alors un  $\chi^2$  à un degré de liberté. Cette statistique dépend à la fois de la liaison physique et du LD entre les deux locus.

Les trois inconvénients de cette méthode sont : (a) elle est peu puissante du fait de l'utilisation exclusive des parents hétérozygotes, (b) elle reste sensible aux effets de structuration de la population, même si cette sensibilité est beaucoup plus faible que celle des méthodes de type cas-contrôle ne tenant pas compte des structures familiales et (c) elle permet seulement de détecter une association entre un marqueur unique et le gène, et non de localiser le gène par rapport au marqueur. Il s'agit cependant de la méthode la plus fréquemment utilisée pour les recherches de locus de maladie dans les populations humaines.



- Article 1 -

III.A.3 Etude cas – témoin pour une mutation unique en population ouverte  
(Terwilliger 1995)

De même que dans la méthode de Hästbacka *et al.* (1992), une mutation causale unique est postulée, mais cette fois-ci, l'haplotype fondateur n'est pas pré-identifié. Après  $t$  générations, du fait du LD créé entre les marqueurs de l'haplotype et la mutation, certains allèles marqueurs devraient être surreprésentés chez les individus malades par rapport aux individus sains. La probabilité de présence de chaque allèle conditionnellement à la présence de l'allèle muté causal s'écrit en fonction (*a*) des fréquences des allèles au marqueur dans la population avant l'apparition de la mutation causale, (*b*) de la fréquence actuelle de la maladie dans la population, et (*c*) de l'excès d'association ( $\lambda$ ) entre l'allèle marqueur et la mutation causale. La vraisemblance que chaque allèle marqueur soit lié à l'allèle de la maladie est calculée puis ces vraisemblances sont sommées sur tous les allèles d'un marqueur pour tester l'association du marqueur avec le locus de la maladie. L'hypothèse d'une liaison entre le marqueur et le gène de la maladie est testée par test de rapport de vraisemblance contre l'hypothèse qu'il n'y a pas d'excès d'association entre l'allèle et le locus de la maladie ( $\lambda=0$ ). Ce rapport de vraisemblance suit un demi  $\chi^2$  à un degré de liberté, et le test est unilatéral.

L'information sur plusieurs marqueurs peut être intégrée dans une approche multipoint, obtenue en multipliant les vraisemblances des différents marqueurs en supposant qu'ils sont indépendants. Cette hypothèse est évidemment fautive pour les marqueurs de la région d'intérêt, rendant la distribution de la statistique de test inconnue sous l'hypothèse nulle. Une alternative est alors de construire la distribution sous  $H_0$  par simulation. Cette extension de la méthode permet d'obtenir une estimation de la distance entre les marqueurs et le gène, en supposant que la décroissance du LD est fonction du temps et du taux de recombinaison entre le QTL et le marqueur sous l'hypothèse d'équilibre de Hardy-Weinberg. Cet équilibre

- Article 1 -

correspond à une constance des fréquences des haplotypes au cours des générations en l'absence de dérive génétique, de mutation, de migration et de sélection, et avec des accouplements aléatoires.

Cette méthode a été étendue par Abdallah *et al.* (2004) à la détection de QTL. Les probabilités de porter un allèle marqueur donné conditionnellement au génotype au QTL sont alors écrites en fonction des distributions des phénotypes selon le génotype.

Boitard *et al.* (2006) ont développé la méthode HapIM, qui est une extension de la précédente à des haplotypes de 2 marqueurs. Cette méthode utilise l'espérance des fréquences de la mutation causale au QTL conditionnellement à l'haplotype à la place des fréquences de la mutation causale au QTL conditionnellement à un allèle au marqueur.

### **III.B Méthodes combinant LD et analyse de liaison**

#### **III.B.1 Pas d'hypothèse sur le LD initial dans des populations à pedigree (Meuwissen et Goddard 2000, 2001)**

La méthode utilise les ingrédients bien connus des généticiens quantitatifs, le modèle linéaire mixte, le BLUP et l'estimation des composantes de variance des effets aléatoires par REML. Pour chaque QTL (supposé additif), le modèle suppose deux effets gamétiques par individu, la corrélation entre effets étant supposée égale à leur probabilité IBD. Ce cadre général étant connu, il convient d'estimer les probabilités IBD, d'une part à l'intérieur du pedigree connu, d'autre part entre les fondateurs (individus sans parents connus). A l'intérieur du pedigree, les probabilités sont déduites à la fois des probabilités de transmission conditionnellement à l'information marqueur et de l'apparentement entre QTL des fondateurs. L'apport essentiel de Meuwissen et Goddard (2000, 2001) est le développement de deux approches, l'une numérique, l'autre déterministe, pour prédire la probabilité IBD entre allèles au QTL des

- Article 1 -

fondateurs. Cette approche repose sur l'hypothèse que des allèles au QTL portés par deux haplotypes identiques par état (*Identical By state, IBS*), c'est-à-dire dont tous les allèles aux marqueurs sont identiques, ont plus de chances d'être IBD que des haplotypes porteurs d'allèles marqueurs différents. En fonction de leur degré de similarité, la probabilité que ces deux haplotypes soient porteurs du même allèle au QTL est alors directement reflétée par la probabilité que ces deux haplotypes soient IBD. Pour estimer ces probabilités, Meuwissen et Goddard (2001) ont développé une approche qui permet d'obtenir la probabilité pour deux haplotypes d'être IBD, conditionnellement à l'IBS des différents marqueurs qui définissent l'haplotype. La probabilité d'IBD est celle d'un QTL supposé être au milieu de l'haplotype ; elle dépend de la taille et de l'âge de la population (supposés connus) et du taux de recombinaison entre les marqueurs. Les probabilités sont donc tout d'abord calculées pour les deux marqueurs flanquant cette position, puis les marqueurs adjacents sont intégrés de proche en proche en tenant compte des taux de recombinaison entre eux.

Ces probabilités d'IBD sont introduites dans un modèle linéaire mixte via la matrice de variance-covariance des effets des haplotypes, qui est ici la matrice des probabilités d'IBD des haplotypes. La vraisemblance de la présence d'un QTL est calculée pour chaque position centrale d'haplotypes, et elle est comparée à celle d'absence de QTL à la position testée, via un test de rapport de vraisemblances. La position maximisant ce test le long d'un chromosome est considérée comme position putative du QTL. La première application de cette méthode concerne la localisation d'un QTL de gémellité dans des populations bovines, avec un intervalle de localisation inférieur à 1 cM (Meuwissen *et al.* 2002).

III.B.2 Mutation unique dans des populations à pedigree (Farnir *et al.* 2002)

La méthode proposée ici décompose les analyses en une première partie d'analyse de liaison, puis une intégration du modèle proposé par Terwilliger afin de prendre en compte le LD. Pour

- Article 1 -

l'analyse de liaison, les hypothèses de la population à l'équilibre de Hardy-Weinberg au QTL et d'association aléatoire entre les allèles du QTL et des marqueurs sont utilisées. La vraisemblance du pedigree est alors calculée en fonction de la fréquence attendue des allèles au QTL sous un équilibre de Hardy-Weinberg, et de la probabilité du phénotype exprimé sachant l'allèle au QTL hérité du père.

Dans un second temps, l'information apportée par le LD est introduite dans les formules issues de l'analyse de liaison. Comme dans la méthode de Terwilliger (1995), la décroissance du LD est supposée être une fonction du taux de recombinaison entre le QTL et le marqueur et du temps, sous un équilibre de Hardy-Weinberg. Les fréquences des différents haplotypes obtenues sont ensuite introduites dans la formule du calcul de la vraisemblance du pedigree obtenue dans l'étape d'analyse de liaison. Comme l'haplotype fondateur est inconnu, tous les allèles du marqueur sont successivement considérés comme étant celui lié à la mutation. Les vraisemblances obtenues pour ces différents allèles sont additionnées, et une extension à plusieurs marqueurs est réalisée, comme dans la méthode de Terwilliger (1995), en multipliant les vraisemblances des différents marqueurs.

La position la plus vraisemblable est celle qui maximise la différence entre la vraisemblance calculée sous l'hypothèse H0 (pas d'effet du QTL ou absence de QTL sur le chromosome) et celle obtenue sous l'hypothèse H1 (un QTL d'effet  $\alpha$  est présent à la position testée). Cette méthode a été appliquée à la cartographie d'un gène bovin majeur affectant la quantité de matière grasse contenue dans le lait avec un intervalle de confiance de la localisation du QTL de 3 cM.

- Article 1 -

III.B.3 Etude cas–contrôle avec des données généalogiques (Zöllner et Pritchard 2004)

Cette méthode repose sur la théorie de la coalescence, c'est-à-dire sur la capacité à retracer la généalogie d'un ensemble de chromosomes au cours des générations en ayant recours à des méthodes d'échantillonnage bayésiennes. Les données sont initialement d'origine humaine, de type cas-contrôle : on postule que les haplotypes porteurs de la maladie devraient se trouver sur des branches proches dans les arbres de coalescence retraçant leurs apparentements. Deux branches proches correspondent à deux haplotypes ayant récemment eu un ancêtre commun. Ces arbres retracent donc les relations ancestrales entre tous les haplotypes ; ils diffèrent selon la position testée dans la région chromosomique. Dans un premier temps, à chacune des positions à tester, l'information apportée par les marqueurs est utilisée pour construire les arbres de coalescence de l'ensemble des haplotypes (cas et contrôles). Puis, à partir d'un échantillon d'arbres, la vraisemblance des données phénotypiques est estimée pour chaque position comme la probabilité moyenne d'observer ces phénotypes. La position retenue comme localisation du QTL est celle maximisant le rapport de cette vraisemblance par rapport à l'hypothèse nulle (pas de mutation causale dans la région étudiée).

**Conclusion**

La cartographie fine reste un enjeu majeur pour les années qui viennent. La disponibilité de marqueurs nombreux et peu coûteux est une avancée considérable mais elle ne résout pas toutes les difficultés méthodologiques. Des dispositifs nouveaux sont à définir, cherchant pour la plupart à tirer parti du déséquilibre de liaison. De nombreuses méthodes sont aujourd'hui disponibles et on constate que les résultats de cartographie fine et de caractérisation de gène

**- Article 1 -**

se multiplient. Cependant, les propriétés et l'efficacité relative de ces méthodes restent mal connues. Certaines méthodes reposent sur des hypothèses plus ou moins cruciales (histoire de la population, nature du QTL, additivité...) ou utilisent tout ou partie de l'information disponible (pedigree par exemple). Toutes les méthodes utilisant le déséquilibre de liaison populationnel font l'hypothèse du modèle de Wright Fisher et donc de l'absence de sélection. Ceci est une hypothèse particulièrement forte qu'il convient d'analyser.

**Références bibliographiques**

Abdallah J.M., Mangin B., Goffinet B., Cierco-Ayrolles C., Perez-Enciso M., A comparison between methods for linkage disequilibrium fine mapping of quantitative trait loci, *Genet. Res.* 83 (2004) 41-47.

Boichard D., Le Roy P., Levéziel H., Elsen J.M. 1998. Utilisation des marqueurs moléculaires en génétique animale. *INRA Prod. Anim.* 11 : 67-80.

Boitard S., Abdallah J., Rochambeau H. de, Cierco-Ayrolles C., Mangin B. 2006. Linkage disequilibrium interval mapping of quantitative trait loci. *BMC Genomics* 7:54-67.

Butte A., 2002. The use and analysis of microarray data. *Nature reviews* 1: 951-960.

Darvasi A., Weinreb A., Minke V., Weller J.I., Soller M., 1993. Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics* 134: 943-951.

Darvasi A., Soller M., 1995. Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics* 141: 1199-1207.

- Article 1 -

Darvasi A., 1996. Interval-specific congenic strains (ISCS): an experimental design for mapping a QTL into a 1-centimorgan interval. *Mamm. Gen.* 8: 163-167.

Darvasi A., Soller M., 1997. A Simple Method to Calculate Resolving Power and Confidence Interval of QTL Map Location. *Behavior Genetics* 27: 125-132.

Darvasi A., 1998. Experimental strategies for the genetic dissection of complex traits in animal models. *Nature Genetics* 18: 19-24.

Falconer D.S., Mackay T.F.C., 1996. *Introduction to quantitative genetics*. 4<sup>th</sup> Ed. Longman Sci. and Tech., Harlow, UK.

Farnir F., Grisart B., Coppieters W., Riquet J., Berzi P., Cambisano N., Karim L., Mni M., Moisisio S., Simon P., Wagenaar D., Vilkki J., Georges M., 2002. Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. *Genetics* 161: 275-287.

Fredriksson R., Schiöth H.B., 2005. The repertoire of G-protein-coupled receptors in fully sequenced genomes. *Molecular Pharmacology* 67: 1414-1425.

Gilbert H., Le Roy P., 2003. Comparison of three multitrait methods. *Genet. Sel. Evol.* 35: 281-304.

Hästbacka J., de la Chapelle A., Kaitila I., Sistonen P., Weaver A., Lander E., 1992. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genetics* 2: 204-211.

- Article 1 -

Korol A.B., Ronin Y.I., Kirzhner V.M., 1995. Interval mapping of quantitative trait loci employing correlated complex traits. *Genetics* 140: 1137-1147.

L'Hote D., Serres C., Laissue P., Oulmouden A., Rogel-Gaillard C., Montagutelli X., Vaiman D. 2007 Centimorgan-range one-step mapping of fertility traits using interspecific recombinant congenic mice. *Genetics*, 176: 1907-1921

Lander E.S., Botstein D., 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121: 185-199.

Lebowitz R.J., Soller M., Beckmann J.S., 1987. Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theor. Appl. Genet.* 73: 556-562.

Loudet O., Saliba-Colombani V., Camilleri C., Calenge F., Gaudon V., Koprivova A., North K.A., Kopriva S., Daniel-Vedele F. 2007. Natural variation for sulfate content in *Arabidopsis* is highly controlled by APR2. *Nature Genetics* 39: 896-900.

Lynch M., Walsh B., 1998. *Genetics and analysis of quantitative genetics*. Sinauer Associates, Inc. Publishers. Sunderland, Massachusetts, USA.

Meuwissen T.H.E., Goddard M.E., 2000. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* 155: 421-430.

Meuwissen T.H.E., Goddard M.E., 2001. Prediction of identity by descent probabilities from marker-haplotypes. *Genet. Sel. Evol.* 33: 605-634.



- Article 1 -

Meuwissen T.H.E., Karlsten A., Lien S., Olsaker I., Goddard M.E., 2002. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* 161: 373-379.

Milan D., Woloszyn N., Yerle M., Le Roy P., Bonnet M., Riquet J., Lahbib-Mansais Y., Caritez J.C., Robic A., Sellier P., Elsen J.M., Gellin J., 1996. Accurate mapping of the "acid meat" RN gene on genetic and physical maps of pig chromosome 15. *Mamm. Gen.* 7: 47-51.

Niemann-Sorensen A., Robertson A., 1961. The association between blood groups and several production characteristics in three Danish cattle breeds. *Acta Agric.Scand.* 11: 163-196.

Ollivier L., Sellier P., Monin G., 1975. Déterminisme génétique du syndrome d'hyperthermie maligne chez le porc Piétrain. *Ann. Génét. Sél. Anim.* 7: 159-166.

Schadt E.E., Lum P.Y. 2006 Reverse engineering gene networks to identify key drivers of complex disease phenotypes *J. Lipid Res.* 47: 2601-2613

Soller M., Genizi A., 1978. The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations. *Biometrics* 34: 47-55.

Spielman R.S., McGinnis R.E., Ewens W.J., 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* 52: 506-516.

- Article 1 -

Terwilliger J.D., 1995. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am. J. Hum. Genet.* 56: 777-787.

Tribout T., Iannuccelli N., Druet T., Gilbert H., Riquet J., Gueblez R., Mercat M.J., Bidanel J.P., Milan D., Le Roy P., 2008 Detection of quantitative trait loci for reproduction and production traits in Large White and French Landrace pig populations. *Genet. Sel. Evol.*, in press.

Van der Beek S., van Arendonk J.A.M., Groen A.F., 1995. Power of two- and three-generation QTL mapping experiments in an outbred population containing full-sib or half-sib families. *Theor. Appl. Genet.* 91: 1115-1124.

Weller J.I., Kashi Y., Soller M., 1990. Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *J. Dairy Sci.* 73: 2525-2537.

Zöllner S., Pritchard J.K., 2004. Coalescent-based association mapping and fine-mapping of complex trait loci. *Genetics* 169: 1071-1092.



**DEUXIEME PARTIE :**

**EVOLUTION DE LA STRUCTURE**

**DU DESEQUILIBRE DE LIAISON**

**SOUS L'INFLUENCE DE LA**

**SELECTION**



La plupart des méthodes utilisant le LD pour la cartographie fine de QTL reposent sur des hypothèses d'histoire de population variées, mais excluant toujours la possibilité de sélection. Cette hypothèse peut être critique d'une part parce que les populations ont été effectivement sélectionnées et même parfois avec des intensités très fortes depuis les 50 dernières années, d'autre part car la sélection affecte la structure du déséquilibre de liaison. Dans ce chapitre, les conséquences de cette hypothèse sur la structure du LD autour d'un QTL seront explorées.

La structure du LD entre un QTL et les locus adjacents dans des populations soumises à sélection est comparée à la structure du LD dans des populations ne subissant pas de sélection pour fournir des outils d'interprétation des résultats de cartographie de QTL produits dans les populations soumises à sélection. Cette étude, basée sur des simulations, a pour objectif d'identifier des paramètres optimaux pour la cartographie utilisant le LD, c'est-à-dire pour lesquels les locus en plus fort LD avec le QTL sont physiquement les plus proches de celui-ci.

Les populations sont produites avec un simulateur qui est présenté préalablement. Il permet de simuler des effectifs génétiques et des histoires de populations variées. La structure du LD résultant de divers scénarios de simulation est évaluée à travers l'étude des distributions des localisations des locus en LD maximum avec le QTL au cours des générations.



Deuxième partie : Evolution de la structure du déséquilibre de liaison sous l'influence de la sélection  
Partie 2.1. : Développement d'un simulateur « Linkage Disequilibrium with Several Options »  
(LDSO)

## **Partie 2.1.**

### **Développement d'un simulateur : « Linkage Disequilibrium with several options » (LDSO)**





## **Introduction**

Un simulateur de populations a été développé pour générer des données pour la cartographie de QTL issues de populations soumises aux phénomènes les plus fréquents créant et maintenant le LD dans les populations animales. Il comprend par conséquent de nombreuses options, qui correspondent à des facteurs évolutifs, populationnels ou génétiques. D'une manière générale, il permet de suivre au cours des générations l'évolution au sein d'une ou deux populations du déséquilibre de liaison (calculé par le  $D'$  ou le  $\chi^2$ ), de la consanguinité, du nombre d'allèles et/ou d'haplotypes en ségrégation, de la proportion d'IBD entre deux haplotypes au(x) QTL ou à tous les marqueurs... Le simulateur est décrit dans un article qui fait l'objet de la première partie de ce chapitre. Les détails, fichiers de paramètres et fichiers de sortie sont documentés de manière exhaustive en anglais dans la notice constituant l'annexe 1. Dans la suite du travail de thèse, seule l'influence de la sélection sur la structure du LD a été étudiée en détail, n'utilisant qu'une partie des possibilités du logiciel.

## I. Article

### **LDSO: A complete program for the simulation of pedigrees and molecular information under various evolutionary forces.**

LDSO is a complete computer program for simulations of whole diploid population histories under various scenarios based on the gene-dropping method (MacCluer *et al.*, 1986). The genetic history of one or two populations is simulated; the output files give various statistics (inbreeding rates, allele frequencies, linkage disequilibrium (LD)) on these populations at user-defined times. Many evolutionary forces that are classically found in livestock populations such as mutation, selection by truncation, bottlenecks or random drift can be taken into account. This allows the simulation of a wide-range of situations. The program is developed in Fortran 90 and self-contained.

#### **Simulated populations**

The history of the populations is subdivided into two parts. First, one or two historical populations are simulated according to the gene-dropping method. The evolutionary forces, that may take place simultaneously, are: random drift, mutation, one bottleneck and selection by truncation. The historical part can be summarized as represented on Figure 1. It may be subdivided into two parts separated by a bottleneck that may act over one or several generations. At each generation, a set of individuals among the population is chosen at random as parents of the next generation, each parental individual giving one gamete to its offspring. Selfing is allowed. Depending on the final kind of population wished, one or two

- Article 2 -

populations can be simulated, the two populations are independent, except for the mutation rates, the initial LD status and the QTL effects.

The second half of the simulation is a “pedigree known” part. Different populations can be simulated with or without familial structure. The “pedigree known” part allows the simulation of three kinds of populations:

- without any experimental design. If one population has been simulated so far, the only available option is to have random mating for the last two generations. If two populations were simulated, there are three situations of admixture: (a) the two populations mate at random, (b) an F1 population is produced, or (c) only a part of the population is F1-like, the rest of the population being produced by matings between individuals of the first population,
- a grand-daughter design which intervenes after a one population simulation,
- an F1 generation followed by an F2 generation or BC generations.

### **Evolutionary forces**

Four evolutionary forces have been considered: genetic drift, selection, one bottleneck and mutation. Genetic drift occurs because populations of limited size are simulated. The creation of each new generation is subjected to a selection rate  $s$  (ranging from 0 to 1), which implies that only the  $s$  percent of the individuals with the best phenotypic values may become parents. If the selection rate is 1 for the two populations, no phenotypes are simulated until one of the selection rates becomes lower than 1 or the pedigree is known (final population). The bottleneck may be punctual (the number of individuals is reduced at generation  $T_1$  and remains constant thereafter) or it may be progressive with a flat decrease rate between

- Article 2 -

generations  $T_1$  and  $T_2$ . These three forces generate a reduction of the number of haplotypes existing in the population.

On the other hand, mutations may also occur, with two different rates corresponding to SNP and microsatellites. Mutations on microsatellites follow the Stepwise Mutation Model (Kimura and Ohta, 1978), implying that an allele  $i$  becomes  $i-1$  or  $i+1$ . For SNP, there is only a switch in the allele number (2 becomes 1 and reciprocally).

### Genetic architecture

The number of marker loci, QTL and chromosomes are unlimited, thus providing the possibility of simulating a whole genome evolution, which can be useful for instance for simulations on genomic selection. Two kinds of allelic effects at the QTL loci are considered: they may be attributed to only one of the QTL alleles, as it is usually supposed in most of the simulation studies, or each QTL allele may have an effect on the phenotype. This is for example the case of the Prnp gene in sheep where the different alleles correspond to different resistance levels to scrapie.

Three different initial situations of Linkage Disequilibrium have been considered:

- no initial LD. The allele numbers are distributed at random with a uniform distribution of the initial allelic frequencies (at the marker and QTL loci);
- a unique mutation is introduced into the population but in a haplotype that may not be unique. This classically corresponds to a mutation in a gene intervening in a quantitative trait (corresponding to the QTL) in a population that was formerly in Linkage Equilibrium (LE). This is the initial situation supposed in many fine-mapping methods (Terwilliger 1995, Farnir, 2002);

- Article 2 -

- a unique haplotype carrying a unique QTL allele exists in the population. This would correspond to the introduction of a unique individual harbouring a new mutation in a population that was in LE.

**Simulation algorithms**

The simulations are programmed according to the gene dropping method (MacCluer *et al.*, 1986). This is a forward process: one starts with a population of founders and each generation is created with an allele transmission from each parent to its offspring following the Mendelian rules. Recombinations between homologue chromosomes are implemented *via* the Haldane recombination function (1919), which assumes no crossover interference. The principle of gene dropping makes it easy to apply selection or mutation on each generation.

Phenotypes are simulated as the sum of all QTL effects, an infinitesimal additive polygenic effect and an environmental effect. Polygenic additive and residual effects are normally distributed with mean 0 and constant variances over time. The polygenic value of an individual is computed as half the sum of the parental polygenic values plus a Mendelian sampling term of variance equal to half the initial genetic variance. QTL effects are allelic and purely additive, i.e., no dominance or epistasis is simulated. When a unique QTL allele affects the phenotype, its effect is defined by the user. When each QTL allele has an effect, they are drawn at random from a Gamma distribution (Hayes and Goddard, 2001) with parameters determined by the user. These phenotypic values are computed for populations that are submitted to selection by truncation or for the final populations.

At the beginning of the simulation, each allele is given a unique copy number (independently from the allele number, i.e. the QTL allelic state) corresponding to the founder

- Article 2 -

haplotype. It is then possible to know the IBD status at each locus (and thus the average inbreeding rate for each locus in the whole population) and the number of founder haplotypes remaining in the population at different times of the population. A mutation leads to the apparition of a new copy number, indicating that this locus is no longer IBD with that of any ancestor.

### Output

Different files are created according to the user's wishes. Some of them contain information for four different generations of the population history: the founder generation, one generation of the historical part (the last before the bottleneck), the last generation of the historical part and at the end of the pedigree known part. Such files record allelic frequencies, copy frequencies, average inbreeding at each locus, the Polymorphism Information Content (PIC) (Botstein *et al.*, 1980) of the QTL, the proportion of remaining haplotypes according to the initial number of different haplotypes or the real IBD status at each locus or at the QTL locus only. Two LD measures are computed, the  $D'$  (Lewontin, 1964; Hedrick, 1987) and the  $\chi^2$  (Yamasaki, 1977). They can be used to compute LD between loci or between one marker and a haplotype whose length is defined by the user.

Other output files are specific to the pedigree known part. These files contain the pedigree, the haplotypes or genotypes of the individuals of the two last generations, or the phenotypes of the individuals of the last generation.

- Article 2 -

**General considerations**

The time required for one simulation depends mainly on the number of requested output files, but also on the population size and on the number of simulated loci. Two populations ending with five generations of back-cross with 10 males at each generation mated each to 5 females with 20 offspring per female were simulated. When the initial population was composed of 100 individuals and 21 loci were simulated, the whole simulation (comprising the writing time for the pedigree, phenotypes, haplotypes and genotypes files) took 2 seconds; 58 seconds were needed with an initial population size of 500 individuals and 101 loci simulated.

This simulator provides a very large range of options on population history and population structures, covering a great number of situations observed in livestock populations. It is different from other simulating programs mainly through the opportunity of applying artificial selection from the first generation on, or only after a few generations. It is particularly suited to investigate the effects of selection on mapping methods or LD.

**References**

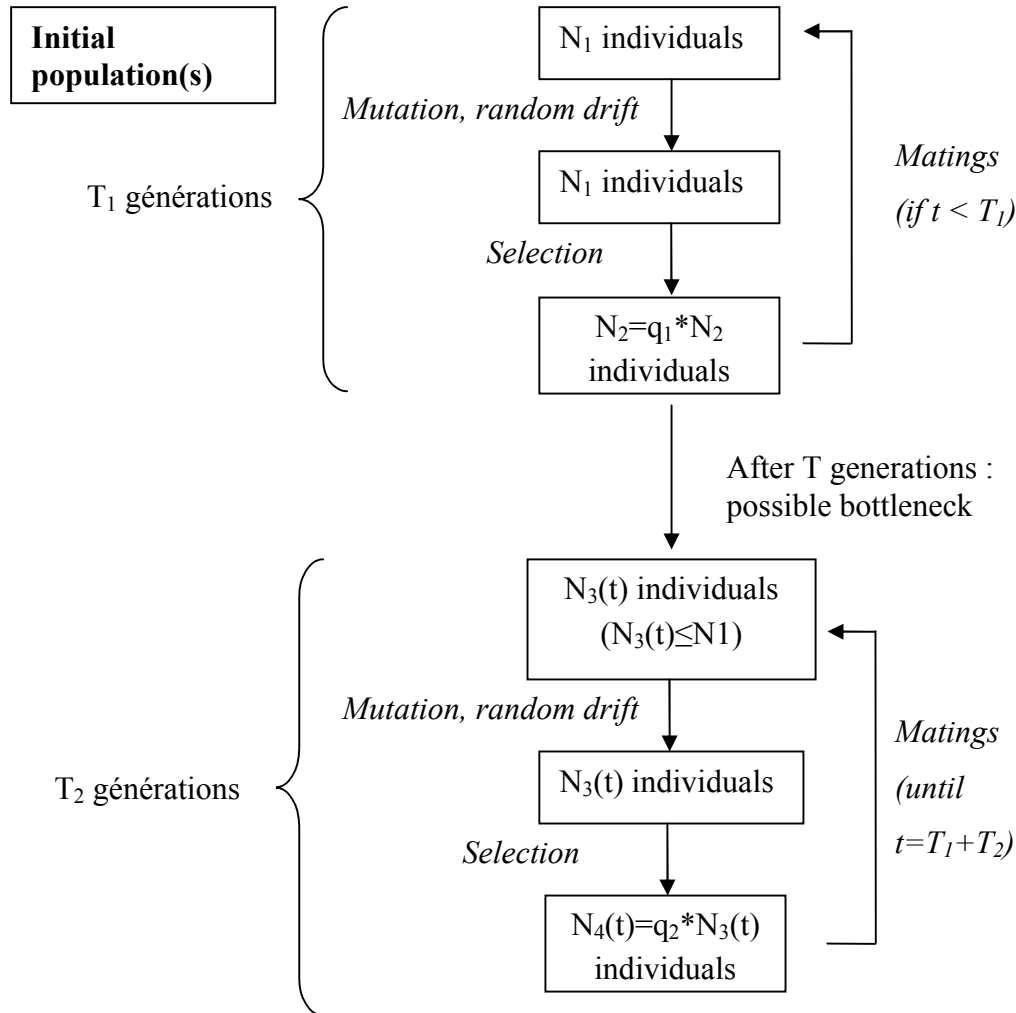
Botstein D., White R.L., Skolnick M., Davis R.W., 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32: 314-331.



- Article 2 -

- Farnir F., Grisart B., Coppieters W., Riquet J., Berzi P., Cambisano N., Karim L., Mni M., Moisisio S., Simon P., Wagenaar D., Vilkki J., Georges M., 2002. Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. *Genetics* 161: 275-287.
- Haldane J.B.S., 1919. The combination of linkage values, and the calculation of distances between loci of linked factors. *J. Genet.* 8: 299-309.
- Hayes B., Goddard M.E., 2001. The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* 33:209-229.
- Hedrick P.W., 1987. Gametic disequilibrium measures: proceed with caution. *Genetics* 117: 331-341.
- Kimura M., Ohta T., 1978. Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc. Natl. Acad. Sci. U.S.A.* 75:2868-2872.
- Lewontin R.C., 1964. On measures of gametic disequilibrium. *Genetics* 49: 49-67.
- MacCluer J.W., VandeBerg J.L., Read B., Ryder O.A., 1986. Pedigree analysis by computer simulation. *Zoo. Biol.* 5: 147-160.
- Terwilliger J.D., 1995. A powerful likelihood method for the analysis of LD between trait loci and one or more polymorphic loci. *Am. J. Hum. Genet.* 56: 777-787.
- Yamazaki T., 1977. The effects of overdominance on linkage in a multilocus system. *Genetics* 86: 227-236.

- Article 2 -



**Figure 1:** General structure of LDSO for the historical part of the simulations.  $q_1$ ,  $q_2$  : proportion of the individuals that may be parents of the next generation ( $0 < q_1 \leq 1$ ,  $0 < q_2 \leq 1$ ).  $t$  : number of generations since the population was founded.

## II. Justification du choix de simulation des haplotypes

Le principe du « gene dropping » de MacCluer *et al.*, proposé en 1986 (cf. Encadré 1), est suffisamment flexible pour permettre d'appliquer des forces évolutives variables aux populations en fonction des générations. Cette méthode est par ailleurs fréquemment utilisée pour les simulations de validation des méthodes de cartographie fine (Abdallah *et al.*, 2003 ; Meuwissen et Goddard, 2000). Elle permet également de simuler des structures de populations différentes au cours des générations, et notamment d'achever les simulations par des populations correspondant aux dispositifs expérimentaux rencontrés dans les populations animales.

### Encadré 1 : La méthode du gene dropping

Cette méthode permet de simuler l'évolution d'une population issue de fondateurs diploïdes. Connaissant un pedigree complexe, les génotypes de la population sont simulés pour chaque individu à partir de ses parents en commençant par les fondateurs. A chaque génération, chaque descendant reçoit un allèle de son père et un de sa mère à chaque locus : les parents étant diploïdes, l'allèle transmis est tiré aléatoirement. Lorsque plusieurs locus sont liés sur un même chromosome, l'allèle transmis de chaque locus dépend de l'allèle transmis au locus précédent. Sachant la distance entre deux locus L1 et L2, l'allèle transmis au descendant pour le locus L2 est issu du même chromosome que l'allèle au locus L1 s'il n'y a pas de recombinaison entre les deux locus, et de l'autre chromosome parental dans le cas contraire. Selon ces principes, il est donc possible de simuler les transmissions des allèles dans un pedigree.

Deuxième partie : Evolution de la structure du déséquilibre de liaison sous l'influence de la sélection

Partie 2.1. : Développement d'un simulateur « Linkage Disequilibrium with Several Options » (LDSO)

Une autre méthode de simulation repose sur la théorie de la coalescence, dont le principe est expliqué dans l'encadré 2. Ce principe est plutôt utilisé dans le cadre d'analyses phylogénétiques remontant sur un très grand nombre de générations (Gasbarra *et al.*, 2005). De telles simulations reconstituent les pedigrees dans un ordre anti-chronologique : elles partent de données obtenues sur la dernière génération pour reconstruire la généalogie des individus actuels. Elles permettent la prise en compte de nombreuses forces évolutives (Nordborg, 2001), mais sont moins flexibles que le gene dropping pour les forces autres que la mutation ou les variations dans la taille de la population.

**Encadré 2 : La théorie de la coalescence**

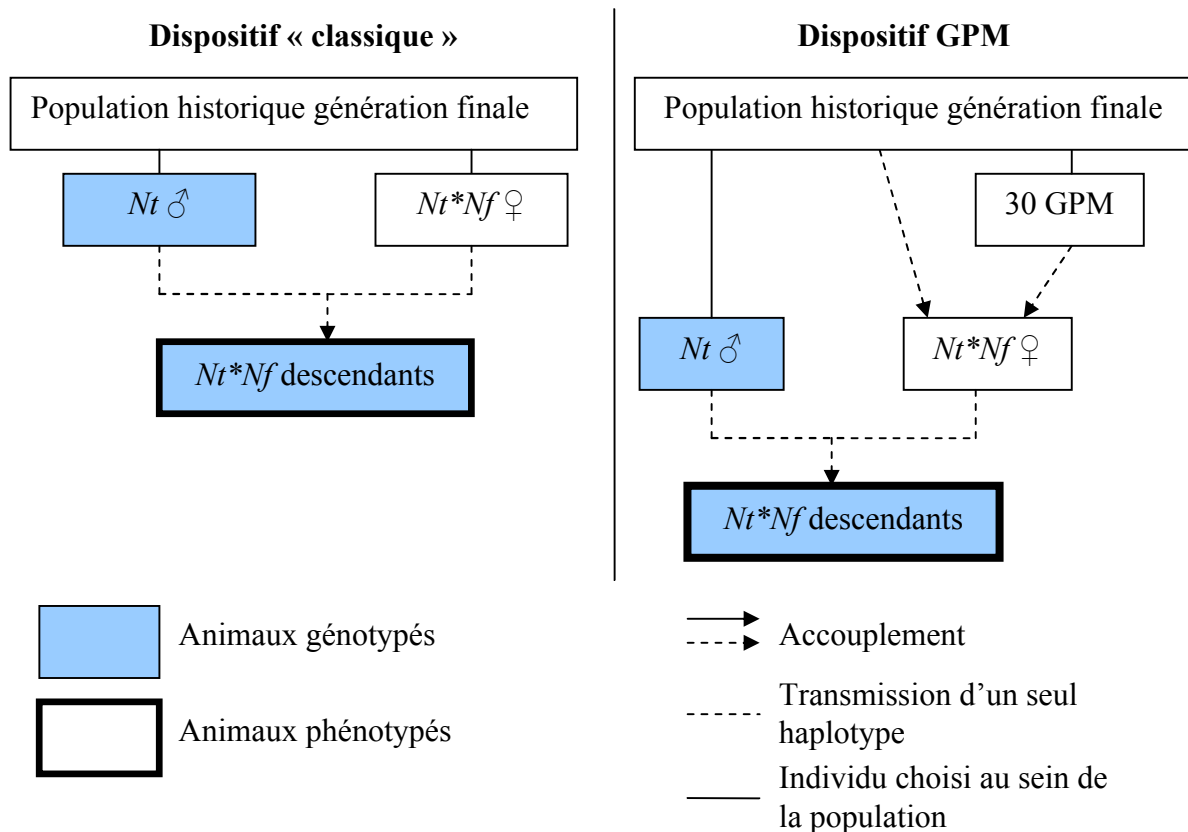
La théorie de la coalescence permet de reconstruire les pedigrees en s'appuyant sur l'observation des séquences ou des génotypes. Elle s'appuie sur la description des séquences orthologues dans la population, c'est-à-dire les variantes d'un même gène ou d'une même séquence observées chez plusieurs individus. En partant des génotypes de la dernière génération observée, les méthodes de coalescence permettent ainsi de décrire la réunion des lignées phylogénétiques (qui sont dans le cas des populations animales les familles) de séquences orthologues. Le pedigree (ou généalogie) est alors couramment représenté sous forme d'arbre. On peut finalement obtenir l'ensemble des arbres généalogiques observables pour un modèle de génétique des populations données.

Le gene dropping est donc *a priori* plus adapté pour répondre à l'objectif du simulateur de permettre l'obtention de pedigrees sur peu de générations, mais pour une grande variété de dispositifs et de combinaisons de forces évolutives.

### III. Dispositifs expérimentaux simulables

A l'issue des générations « historiques » de création du LD, quelques générations sont structurées selon un dispositif expérimental de cartographie de QTL, où seuls les mâles (au nombre de  $Nt$ ) et les descendants ( $Nd$  par mère) sont génotypés. Seuls les descendants sont phénotypés pour le caractère. Dans la pratique, seules ces générations sont connues pour les apparentements.

Si une seule population historique est simulée, deux dispositifs de détection sont possibles. Ces deux types de pedigrees correspondent à des dispositifs de détection dits « protocoles filles » (Weller, 1990) (cf. partie 1.2, figure 4). Dans les deux situations, les pères sont tirés au hasard dans la génération finale de la population historique disponible, et chaque mère n'a qu'un descendant ( $Nd=1$ ). Dans le premier dispositif, les mères sont tirées au hasard dans la population et le dispositif est un ensemble de familles de demi-frères (dispositif « classique »). Dans le second cas, les mères sont issues de 30 grands-pères maternels (GPM), eux-mêmes tirés au hasard dans le pedigree, et de mères tirées au hasard (Figure 2). Cette situation renforce le degré d'apparentement entre mères.



**Figure 2 : Dispositifs pour la détection de QTL proposés à partir d'une unique population historique**

Un dispositif de deux générations avec accouplements aléatoires sans aucune structure familiale de population est également proposé.

Trois dispositifs peuvent être simulés si deux populations historiques ont été simulées simultanément (Figure 3). Ils utilisent des individus dits F1, issus du croisement entre des animaux de la dernière génération de chacune des deux populations, PHF1 et PHF2.

Deux de ces schémas correspondent à des populations animales expérimentales classiques (BC et F1-F2), utilisées notamment avec les races porcines (par exemple Bidanel *et al.* (2001) : croisement d'une race chinoise avec une race européenne). Le dispositif de type BC (Figure 3 B) est construit de la manière suivante : après la génération de production des animaux F1, les générations de ( $Nf*Nt$ ) BC sont issus d'accouplements entre  $Nf$  femelles F1 ou BCx et  $Nt$  mâles descendants de la population PHF1. Le retour n'est donc envisagé que sur les femelles, alors qu'il est en pratique généralement réalisé sur les deux sexes. Ceci est reconduit sur  $g$  générations pour produire en génération  $g+1$  ( $Nt*Nf*Nd$ ) descendants. Le

Deuxième partie : Evolution de la structure du déséquilibre de liaison sous l'influence de la sélection

Partie 2.1. : Développement d'un simulateur « Linkage Disequilibrium with Several Options » (LDSO)

dispositif F1-F2 (Figure 3C) est pour sa part construit en dernière génération par l'accouplement de mâles et de femelles issus de la population F1 pour produire ( $N_t * N_f * N_d$ ) descendants.

Enfin, le troisième dispositif, de type « introgression » (Figure 3A) correspond, à la génération ( $N-g$ ), au mélange dans une même population d'une fraction  $r$  d'individus de type F1 alors que le reste de individus (fraction  $1-r$ ) est issu de l'accouplement d'individus de la première population historique finale (PHF1). Tous ces individus sont fondateurs d'une nouvelle population (population d'introgression PI), où les accouplements aléatoires sont conduits comme décrit précédemment pendant  $g$  générations (ce paramètre est indiqué par l'utilisateur). De la dernière génération de la population d'introgression sont extraits les parents des individus phénotypés, produits par des accouplements au hasard. Ce cas de figure peut-être rencontré dans de nombreuses populations. Par exemple, à de rares occasions, des individus issus du croisement de deux races bovines ont été produits et introduits dans les schémas de sélection. Ces accouplements visaient à introduire dans une race (race A) des caractéristiques propres à l'autre race (race B) (notamment une plus grande production laitière), ou à augmenter la variabilité génétique, particulièrement pour les races à petits effectifs. On dit que l'on a introgressé des gènes de la race B dans la race A. Par exemple, l'introduction de sang de Holstein a conduit à l'absorption totale de la race Frisonne entre les années 1960 à 1980. Plus modeste, l'introduction de sang de Holstein rouge est un exemple d'introgression très partielle dans les races Montbéliarde et Abondance (Boichard *et al*, 1996). Ces pratiques sont utilisées aussi pour importer des caractéristiques phénotypiques de races et/ou lignées différentes en populations porcines ou de volailles de sélection.

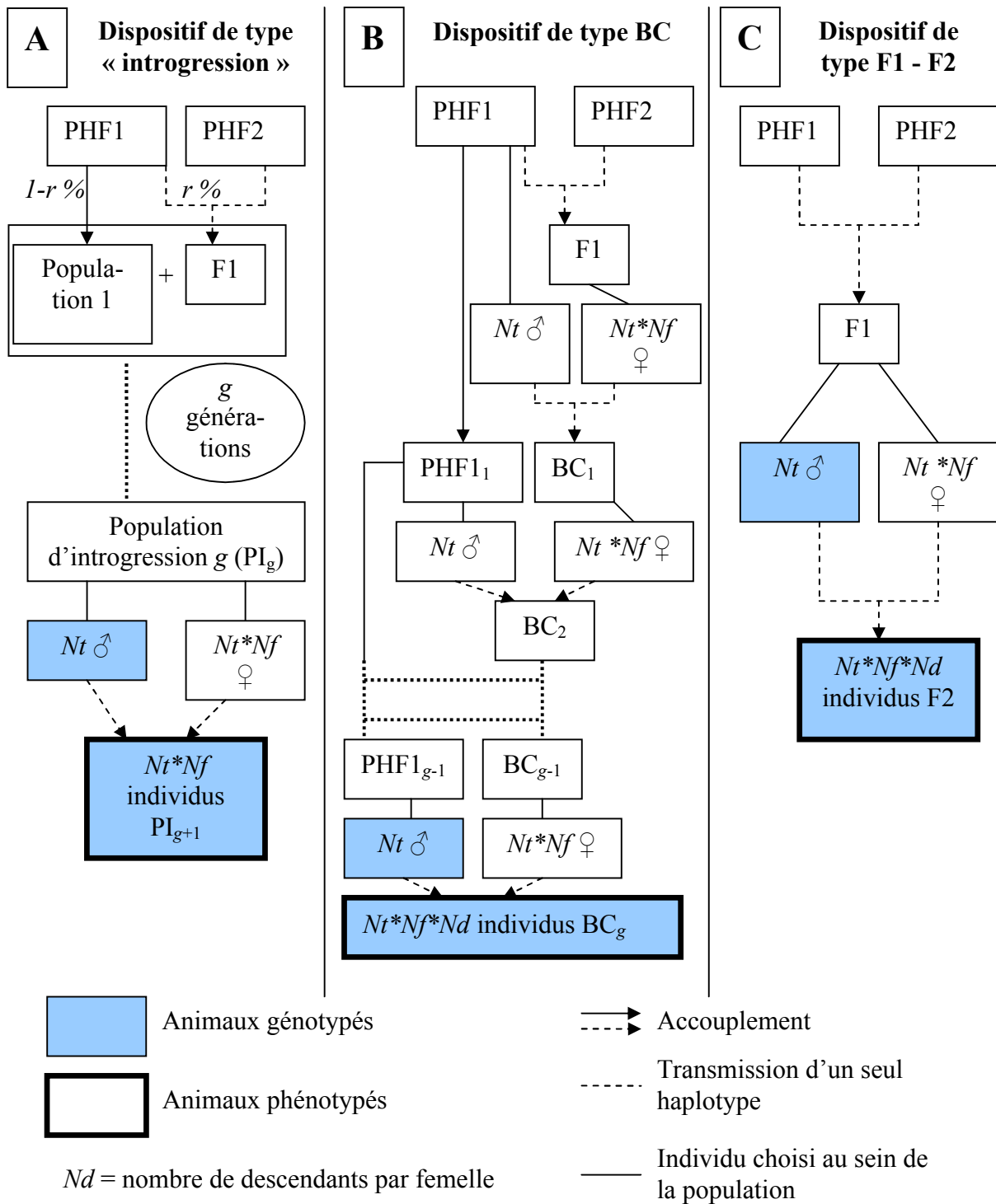


Figure 3 : Dispositifs simulés pour la cartographie de QTL issus de deux populations historiques



Deuxième partie : Evolution de la structure du déséquilibre de liaison sous l'influence de la sélection  
Partie 2.1. : Développement d'un simulateur « Linkage Disequilibrium with Several Options » (LDSO)

## **Conclusion**

Cette flexibilité et la variété des paramètres pris en compte sont les particularités de ce simulateur. Il permet non seulement de mimer des populations naturelles mais aussi des populations animales expérimentales ou sélectionnées de façon réaliste. Il est donc utilisable pour une très grande variété de situations, pour la génétique végétale aussi bien qu'animale.

Deuxième partie : Evolution de la structure du déséquilibre de liaison sous l'influence de la sélection

Partie 2.2. Structure du déséquilibre de liaison dans des populations sélectionnées

## **Partie 2.2.**

### **Structure du déséquilibre de liaison dans des populations sélectionnées**



## Introduction

L'étude du LD dans les populations animales s'est développée ces dernières années dans les populations animales réelles (Farnir *et al.*, 2000 ; Heifetz *et al.*, 2005 ; Lou *et al.*, 2003 ; McRae *et al.*, 2002 ; Odani *et al.*, 2006 ; Tozaki *et al.*, 2007, Gautier *et al.*, 2007), notamment du fait de son utilisation en cartographie fine (Abdallah *et al.*, 2004 ; Meuwissen *et al.*, 2000 ; Boitard *et al.*, 2006 ; Perez-Enciso, 2003 ; Zöllner *et al.*, 2005).

Il est important (a) d'évaluer le LD établi dans les populations, et (b) de savoir comment l'exploiter au mieux. Sur ce dernier point subsistent des questions, notamment liées à la qualité de l'information moléculaire utilisée (type de marqueurs, densité de carte, haplotypes ou marqueurs uniques) et son impact sur l'estimation du LD. Le principe de la cartographie par LD repose sur l'hypothèse que la position la plus probable du QTL est la région en déséquilibre de liaison maximum avec lui. L'objectif de cet article est d'évaluer la localisation de cette région en déséquilibre maximum avec le QTL, sous des hypothèses variées d'histoire des populations, afin de cibler les dispositifs optimaux pour la cartographie.

Il apparaît dans les différentes études menées sur les méthodes de cartographie fine (Abdallah *et al.*, 2004 ; Grapes *et al.*, 2006 ; Zhao *et al.*, 2007) que l'utilisation de l'information moléculaire sous la forme de marqueurs seuls ou d'haplotypes influe sur les résultats. Il ne semble pas y avoir de forme consensus qui soit optimale pour toutes les méthodes de cartographie. Différentes combinaisons d'information moléculaire ont donc été utilisées dans cette étude, pour déterminer si l'utilisation de un, deux ou quatre marqueurs simultanément permet de réduire l'intervalle autour du QTL dans lequel se situe en moyenne le locus en déséquilibre de liaison maximal avec le QTL.

Le simulateur présenté dans la partie précédente doit servir cet objectif. Les calculs de LD entre le QTL et les différents locus (marqueurs seuls ou haplotypes de 2 ou 4 marqueurs), effectués par le logiciel à différents âges de la population, sont directement utilisables à partir des sorties du simulateur. Différentes tailles de population, ainsi que plusieurs pressions de sélection sont simulées dans cette étude afin d'obtenir un schéma de l'évolution de la structure du LD maximal entre le QTL et les locus marqueurs de la région chromosomique au cours des générations. Les résultats majeurs sont présentés sous forme d'article, puis des résultats complémentaires sont rapportés dans une deuxième partie afin de compléter la discussion.

Deuxième partie : Evolution de la structure du déséquilibre de liaison sous l'influence de la sélection

Partie 2.2. Structure du déséquilibre de liaison dans des populations sélectionnées

**- Article 3 -**

**I. Article**

**Location of maximum linkage disequilibrium in selected populations**

Florence YTOURNEL\*, Hélène GILBERT, Tom DRUET, Didier BOICHARD

Institut national de la recherche agronomique, Station de génétique quantitative et appliquée,

78352 Jouy-en-Josas Cedex, France

Telephone number : (+33)134652819

Fax number : (+33)134652210

e-mail :florence.ytournel@jouy.inra.fr

Short title: Linkage Disequilibrium in selected populations

**- Article 3 -**

**Abstract**

To evaluate the characteristics of the locations of the maximum Linkage Disequilibrium (LD) in selected populations, we computed the maximum LD between a locus of interest (denoted QTL thereafter) and either individual markers, or 2- or 4-marker haplotypes was computed using  $D'$  and  $\chi^2$  in various simulated populations of medium size (100 to 200), with different numbers of generation and selection pressures. The mean and standard deviation of the location of the maximum LD, its concentration in a given interval surrounding the QTL and its 95% confidence interval (CI) were derived. Selection decreased mapping efficiency, i.e. increased the variability of the locations of maximum LD values and reduced its concentration around the QTL. Mapping efficiency was found to be optimal after 25 to 75 generations of population evolution and decreased thereafter, likely due to allelic fixation. This optimum was reached faster for small and/or selected populations, particularly with  $D'$ . After 100 generations, the frequency of maximum LD locations mapped within the non recombinant founder QTL segment decreased, particularly in selected populations and with  $D'$ , reflecting this decrease in efficiency.

Key words: Linkage Disequilibrium / Selection / Identity By Descent / molecular information

- Article 3 -

**1. INTRODUCTION**

Linkage Disequilibrium (LD) is a non-random allelic association between loci. Although LD between a Quantitative Trait Locus (QTL) and a genetic marker cannot be directly observed, it generates an association between marker locus segregation and the phenotype distribution which provides an efficient tool for QTL fine mapping. LD has already been used to fine-map QTLs for production traits or complex diseases [5, 9, 12, 21]. Confidence intervals of QTL location obtained from familial linkage studies are usually very large due to limited number of recombinants and could be reduced to regions of high LD that are highly conserved over many generations. LD results from mutation, random drift, migration or selection [3]. Livestock populations are commonly subjected to several of these forces which may structure LD around the QTLs.

Studies have been conducted to estimate the range of LD achieved in these populations (cattle [20, 4, 23], sheep [17], dogs [15] or commercial layer chicken [11]). Most of them used the  $D'$  measure [10, 14] and concluded to a wide-range LD, unfavourable to QTL mapping. However, Heifetz *et al.* [11] recently used the  $\chi^2$  measure [25] and found few high LD levels between widely separated or non-syntenic loci. Recently, Zenger *et al.* [26] and Gautier *et al.* (submitted) showed that even in domestic animal populations, LD is a good tool for QTL fine mapping. LD measures can thus offer different pictures for LD fine mapping methods.

On the other hand, fine-mapping methods use either marker or haplotype information. The number of possible haplotypes obtained by combining marker alleles is greater than the number of alleles of a single marker and thus provides more opportunity for a strong association with phenotypes. Studies [7, 1] have been conducted to search for the optimal

**- Article 3 -**

haplotype size for QTL fine mapping. Abdallah *et al.* [1] and Grapes *et al.* [7] suggested that 2- or 4-marker haplotypes are the most appropriate, depending on the way LD is integrated in the mapping method, with Terwilliger and Meuwissen & Goddard's methods, respectively. Some authors, however, argue that best results are obtained by individual markers [6, 27] because of their large number and their more precise localisation.

Population size and QTL effects are known to affect mapping efficiency. But other factors could also affect fine mapping. Selection increases the length of the conserved segment surrounding the QTL. The number of generations is usually considered as favourable as it tends to eliminate long-range LD, but it also affects allelic frequencies through drift. As LD estimation depends on marker polymorphism and not only on distance to the QTL, the maximum LD location may be affected. In this paper, we addressed the following questions by simulation:

- Where is located the position of maximum LD with the QTL, according to the population size, number of generations, marker density, and number of markers considered?
- How is the mapping efficiency affected by selection?
- Assuming that short-range LD derives from founder effects, can one actually expect that the location of maximum LD is found within the conserved founder segment?

To carry out this study, we focused on the LD structure in the QTL neighbourhood, concentrating on the loci having the highest LD with the QTL. LD was calculated between the QTL and either every single marker, 2- or 4-marker haplotypes. We thus a) characterized the location of maximum LD with the QTL, estimated with the  $D'$  or the  $\chi^2$  measures, and b) compared it with the conserved founder segments containing the QTL.



- Article 3 -

## 2. DESIGNS, METHODS AND SIMULATIONS

### 2.1. Simulated designs

Populations were simulated with constant sizes of 100, 150 or 200 individuals per generation and evolved over 100 generations. Generations were separate, matings were random allowing selfing. Truncation selection was applied and animals with best phenotype in proportion  $q$  (1, 0.9, or 0.8) were retained as potential parents for the next generation. The phenotype was simulated as the sum of a polygenic effect, an additive QTL effect and a residual effect. In the founder population, the QTL explained 20% of the genetic variance and the heritability of the trait was 5%. Polygenic Mendelian samplings and residual effects were normally distributed with mean 0 and constant variances over time. A QTL allelic effect  $a$  was simulated for QTL allele 1 whereas other QTL allelic effects were set to zero.

Simulated haplotypes included 20 evenly spaced markers and one QTL located in the middle of the chromosomal region (Fig. 1). Each locus, including the QTL, had 5 alleles with equal frequency in the founder population. Founder alleles were drawn at random for each locus, assuming initial linkage equilibrium. In addition to the allele number, each founder locus was given a copy number (1 to  $2N$ ,  $N$  being the number of founders), making it possible to follow up IBD segments showing no apparent recombination through generations. The one containing the QTL is called founder segment or non-recombinant segment. Allele transmissions over generations were simulated using the gene-dropping technique [16].

Two genetic maps were defined with 1 or 0.25 cM between adjacent markers, for different total segment lengths of 19 cM and 4.75 cM, respectively. Loci were numbered from -10 to -1 for the left markers, 0 for the QTL, and from 1 to 10 for the right markers. Two-marker haplotypes were noted -9 to 9 and 4-marker haplotypes were noted -8 to 8 (Fig. 1). Recombination was simulated using the Haldane mapping function [8].

- Article 3 -

Figure 1 here

2.2. Linkage disequilibrium computation

Because the study focused on the LD structure of the segment harbouring the QTL and not on the QTL effect, the QTL alleles were assumed to be known and thus were directly used to compute LD. Note that the phenotype was used only in the selection process and not for LD computation.

Two measures,  $D'$  and  $\chi^2$ , were used to compute LD between the 5-allele QTL and each marker or haplotype.  $D'$  [10] is defined as  $D' = \frac{\sum_{i=1}^I \sum_{j=1}^J p_i q_j |D'_{ij}|}{\sum_{i=1}^I \sum_{j=1}^J p_i q_j}$ , where  $I$  and  $J$  are the number of alleles at loci A and B, respectively,  $p_i$  ( $q_j$ ) is the frequency of the allele  $i$  ( $j$ ) at locus A (B) and  $|D'_{ij}|$  is the absolute value of Lewontin's [15] normalized LD measure:

$$D'_{ij} = \frac{x_{ij} - p_i q_j}{D_{\max}} = \frac{D_{ij}}{D_{\max}}, \text{ with } x_{ij} \text{ the observed frequencies of the haplotypes } A_i B_j \text{ and}$$

$$D_{\max} = \begin{cases} \min(p_i q_j, (1-p_i)(1-q_j)) & \text{if } (x_{ij} - p_i q_j) < 0 \\ \min(p_i(1-q_j), (1-p_i)q_j) & \text{if } (x_{ij} - p_i q_j) > 0. \end{cases}$$

Zhao *et al.* [28] showed that the  $\chi^2$  [25] better described the LD decline with respect to the distance between the markers. It is computed as  $\chi^2 = \frac{\chi^2}{2N(l-1)}$ , where  $l = \min(I, J)$  and

$$\chi^2 = 2N \sum_{i=1}^I \sum_{j=1}^J \frac{D_{ij}^2}{p_i q_j}.$$

For the LD calculation, the marker haplotypes were assimilated to alleles, creating a maximum number of 25 or 625 alleles for haplotypes of 2 or 4 markers, respectively. The location of a haplotype was defined as a single point in its middle.

**- Article 3 -**

2.3. Simulations

The designs were identified combining the selection rate ( $q$ ), the marker density and the population size. For a given combination, two thousand simulations were retained after excluding those where one QTL allele was fixed. At generation 0, 10, 25, 50, 60, 75 and 100, the following parameters were recorded: the frequency of the QTL alleles; the average length of the founder segment containing the QTL; the position showing the highest LD level, for both LD measures. When several positions showed the same LD level, a single position was drawn at random among them. From the distribution of the maximum LD values, we derived two statistics:

- A 95% confidence interval of the length of the segment containing both the QTL and the marker (or haplotype) in highest LD. For haplotypes, the confidence interval was first computed with reference to the distance between their position (middle of the haplotype) and the QTL position. We then added to it twice the half length of the haplotype to account for the total length of the most external haplotypes belonging to that confidence interval.
- The proportion of maximum LD observed in the QTL neighbourhood. For individual markers, the QTL neighbourhood was defined by markers -2 to 2. As in [8, 20], neighbour haplotypes were defined as the three haplotypes the closest to the QTL.

A third parameter was computed based on the length of the IBD segment containing the QTL. It was calculated using copy numbers, assuming that no recombination occurred between markers with same copy numbers, and that recombination events occurred in the middle of intervals. The average length of the IBD segment containing the QTL was computed for each design at generations 0, 10, 25, 50, 60, 75 and 100. We also recorded the percentage of

**- Article 3 -**

replicates for which the highest LD was within the non recombinant segment containing the QTL, which is expected for fine mapping.

The targeted characteristics of these parameters for LD-fine mapping would be: a zero mean and a small standard deviation of the locations of maximum LD; a high concentration of the maximum LD locations in the QTL neighbourhood; and a narrow confidence interval.

### **3. RESULTS**

The results in the figures and in the tables mainly focus on populations of 150 individuals; different population sizes are commented only when they caused notable differences.

#### **3.1. LD structure**

##### 3.1.1. General statistics on the distances between the QTL and the location of maximum LD

The distribution of the distances between the QTL and the maximum LD location (Fig. 3, generation 0) was initially uniform along the chromosomal region, in agreement with the initial linkage equilibrium assumption.

*Figure 3 here*

At any analysed generation, the mean location was close to the QTL position (0) for both LD measures and individual markers or haplotypes (Fig. 2). On the sparse map, the standard deviations initially decreased during the first 25 to 75 generations before stabilizing. They increased thereafter in selected and small populations whereas they remained quite constant in large unselected populations. The standard deviation minimum values reached by both LD measures were approximately the same but their increase rate after the optimum was much higher with  $D'$ . Measuring LD with haplotypes led to reduced standard deviations compared to single markers in all situations (i.e. 0.46 with one marker vs. 0.25 with H4 without

**- Article 3 -**

selection) but at the 100<sup>th</sup> generation with  $\chi^2$ . Situations with selection or small populations (not shown) increased the standard deviations, the effect of selection being particularly strong (i.e. 0.46 without selection vs. 0.68 when  $q=0.8$  with one marker).

Results were rather similar on the dense and sparse maps. On the dense map, however, the optimum was reached faster, that is after only 25 to 50 generations.

*Figure 2 here*

The distributions of the locations of maximum marker-QTL LD on a sparse map (Fig. 3) indicated that the highest concentration around the QTL occurred before 25 generations, with an increase in frequency around the QTL position, a small standard deviation, and a nearly zero frequency of distances larger than 7.5 cM. When  $D'$  was used, standard deviations increased after that time point as well as the frequency of DL peak at distance larger than 7.5 cM (Fig. 3A). In contrast, when using  $\chi^2$  (Fig. 3B) or when the populations were unselected (not shown), these extreme locations still had low frequencies after 100 generations. Such results were also true for haplotype-QTL LD.

3.1.2. LD concentration and 95% confidence intervals

*Figure 4 here*

The concentration of the locations of maximum LD around the QTL slightly increased with population size (Fig. 4). When LD was measured with  $D'$ , a decrease was noticed after 25 to 75 generations. The maximum concentration was reached later in unselected and/or larger populations. This decrease was faster when LD was computed with individual markers. With  $\chi^2$ , the concentrations remained stable or only slightly decreased during the last generations in unselected populations or with marker-QTL LD. With longer haplotypes, the decrease in selected populations became more noticeable. After 100 generations of evolution in selected populations, the optimal genetic information to be used depended on the LD measure: with

**- Article 3 -**

$D'$ , the concentrations were higher with 4-locus haplotypes whereas markers showed the highest values when using  $\chi^2$ . The concentrations were lowest in selected populations with both measures. For recent LD, after less than 10 generations, the concentration differences mainly relied on the haplotype length: H4 provided a greater concentration than H2, itself much better than single markers. In addition, the advantage of H4 was more pronounced on a dense map.

The maximum concentration was found to be at least as long to reach as the shortest 95% confidence interval (Tab. I and II). For most designs, the required time increased with lower marker spacing both for confidence interval and concentration.

*Tables I and II here*

There was little influence of the haplotype length on the maximum concentrations in unselected populations. However, these concentrations increased when using longer haplotypes in selected populations, with a noticeable superiority of H4. This superiority was also true for the 95% CI in selected populations.

The 95% confidence intervals reduced less than proportionally to marker spacing: with  $D'$  (respectively  $\chi^2$ ), the reduction was 25% (resp. 42%) for marker-QTL LD on an unselected population and 75% for both measures on 4-loci haplotypes with  $q=0.8$ . Therefore the length decrease was stronger with H4 and in selected populations.

**3.2. IBD segments transmitted from the founders**

The average length of the non-recombinant founder segments containing the QTL slightly increased with selection (Fig. 5), with both map densities. The segment mean length was greater for small population sizes.

*Figures 5 and 6 here*

**- Article 3 -**

Figure 6 presents the proportion of replicates where the peak of marker-QTL LD is contained in the non recombinant founder QTL segment. The following tendencies were observed:

- the number of maximum LD values located in the QTL founder segment decreased during the evolution of the population, while the length of the segment reduced,
- before 25 generations,  $D'$  and  $\chi^2$  provided similar results, but after 25 generations, the  $\chi^2$  showed better results than the  $D'$  as there were more maximum LD locations contained in the founder segment. This tendency was particularly noticeable on the dense map,
- the best agreement between the location of the maximum LD value and the QTL founder segment was observed for large and/or unselected populations for both map densities, the highest values being achieved on the dense map.

#### **4. DISCUSSION**

This study focused on the maximum LD location as estimator of the QTL position in populations selected or not and of various sizes. In order to avoid any noise due to QTL effects, we considered the optimal situation where QTL alleles were known. This optimistic situation is equivalent to a strong major gene, where the genotype is given by the phenotype.

Several parameters were used: means and standard deviations of the distances between the QTL and the maximum LD values, the confidence interval of these distances and the proportion of maximum LD fit in the interval around the QTL. The position in strongest LD with the QTL got closer to this locus over at least the first generations. Depending on the population size and the selection pressure, LD may achieve a maximum concentration before the end of the simulation process of 100 generations and this maximum concentration gets subsequently lost. However, the two LD measures considered in this study did not agree on this time point, without any optimum common to all the considered designs.

**- Article 3 -**

This pattern of optimal detection may be explained by a loss of polymorphism, up to the fixation of alleles for markers close to the QTL. This trend increases with generation number and is faster with smaller population sizes because of the random drift [3]. It is also amplified by selection, for all markers because effective population decreases, and particularly for marker loci close to a selected gene [18]. When alleles are fixed, LD is null and the maximum LD level can therefore not be located at these loci, which are generally close to the QTL. This situation explains why the peak of LD has a lower efficiency in QTL mapping in selected populations. Even before fixation, very unbalanced allelic frequencies may generate low or spurious values for LD, limiting QTL mapping efficiency.

This study also shows that using haplotypes is efficient for fine mapping purposes, particularly in selected or young populations, probably due to more informativity. However,  $\chi^2$  seems to advantage markers in eldest selected populations. Considering haplotypes is interesting to balance limited polymorphism of markers and also to delay fixation time. A possible reason for limited haplotype efficiency in selected populations is the length of the chromosomal region covered by a 4-marker haplotype, preventing any very accurate mapping and not fully compensated by a much higher number of remaining alleles.

In proportion, less maximum LD values were close to the QTL in selected or small populations. Meanwhile, the length of the QTL segment inherited from a single founder became smaller as the population grew older, leading to a higher frequency of LD maximum outside the segment. This wide-range LD contributes to a reduced mapping resolution, which can be only very partially compensated by a higher marker density.

The two measures of LD behaved very differently.  $D'$  appeared to be more sensitive to low or null allele frequencies, in agreement with previous studies [28]. Consequently,  $\chi^2$  provided better results particularly for higher generation numbers and selected or small



**- Article 3 -**

populations, although results were very similar with both measures in the first generations. At generation 100, most populations showed long-range LD measured with  $D'$ , in agreement with the results found in studies on livestock [4, 15, 17, 20, 23]. These situations are not optimal for fine mapping purposes and marker density should be strongly increased to detect older LD. In agreement with Zhao *et al.* [28], we found that  $\chi^2'$  was a better indicator of LD than  $D'$  when the measure was directly made between the marker loci and the QTL: it seemed to be more robust to selection, random-drift or molecular information content. Consequently, it provided better mapping results, as the highest LD values were more often located in the non recombinant QTL founder segment. We therefore recommend to use  $\chi^2'$  instead of  $D'$  as indicator of LD, particularly in selected populations, when computing LD with low polymorphic markers or for populations aged of more than 50 generations.

## **5. CONCLUSION**

This study was based on a commonly assumed structure for livestock breeds, with a recent bottleneck few centuries ago, a limited effective size and selection. With such a structure, blocks of haplotypes remain relatively large (2 cM). LD methods with current marker information are able to map QTL but accuracy is limited by this genetic structure. Selection, which is often omitted in the studies, generates long-range LD and an increase in haplotype block size, and therefore decreases mapping efficiency, although this loss of efficiency remains limited. Assuming that QTL alleles pre-exist to livestock breeds and are not due to recent mutations, which is likely for SNPs, it would be more beneficial to take advantage of the remaining older genetic structure existing before breed standardisation. This requires a higher marker density and would provide stronger LD in the close neighbourhood of the QTL.

Deuxième partie : Evolution de la structure du déséquilibre de liaison sous l'influence de la sélection

Partie 2.2. Structure du déséquilibre de liaison dans des populations sélectionnées

**- Article 3 -**

This approach would be enhanced by analysing several breeds simultaneously, optimally long time separated.

- Article 3 -

**REFERENCES**

- [1] Abdallah J.M., Mangin B., Goffinet B., Cierco-Ayrolles C., Perez-Enciso M., A comparison between methods for linkage disequilibrium fine mapping of quantitative trait loci, *Genet. Res.* 83 (2004) 41-47.
- [2] Boichard D., Maignel L., Verrier E., Analyse généalogique des races bovines laitières françaises, *INRA Prod. Anim.* 9 (1996) 323-335.
- [3] Falconer D.S., Mackay T.F.C., Introduction to quantitative genetics, 4<sup>th</sup> edn., Longman, Essex, 1996.
- [4] Farnir F., Coppeters W., Arranz J.J., Berzi P., Combisano N., Grisart B., Karim L., Marcq F., Moreau L., Mni M., Nezer C., Simon P., Vanmanshoven P., Wagenaar D., Georges M., Extensive genome-wide linkage disequilibrium in cattle, *Genome Res.* 10 (2000) 220-227.
- [5] Gautier M., Barcelona R.R., Fritz S., Grohs C., Druet T., Boichard D., Eggen A., Meuwissen T.H.E., Fine mapping and physical characterization of two linked quantitative trait loci affecting milk fat yield in dairy cattle on BTA26, *Genetics* 172 (2006) 425-436.
- [6] Grapes L., Dekkers J.C.M., Rotschild M.F., Fernando R.L., Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci, *Genetics* 166 (2004) 1561-1570.
- [7] Grapes L., Firat M.Z., Dekkers J.C.M., Rothschild M.F., Fernando R.L., Optimal haplotype structure for linkage disequilibrium-based fine mapping of quantitative trait loci using identity by descent, *Genetics* 172 (2006) 1955-1965.
- [8] Haldane J.B.S., The combination of linkage values, and the calculation of distances between loci of linked factors, *J. Genet.* 8 (1919) 299-309.

**- Article 3 -**

- [9] Hästbacka J., de la Chapelle A., Kaitila I., Sistonen P., Weaver A., Lander E., Linkage disequilibrium mapping in isolated founder populations : diastrophic dysplasia in Finland, *Nat. Genet.* 2 (1992) 204-211.
- [10] Hedrick P.W., Gametic disequilibrium measures: proceed with caution, *Genetics* 117 (1987) 331-341.
- [11] Heifetz E.M., Fulton J.E., O'Sullivan N., Zhao H., Dekkers J.C.M., Soller M., Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations, *Genetics* 171 (2005) 1173-1181.
- [12] Kerem B.S., Romens J.M., Buchanan J.A., Markiewicz D., Cox T.K., Lehesjoki A., Koskiniemi J., Norio R., Tirrito S., Sistonen P., Lander E.S., de la Chapelle A., Localisation of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21: linkage disequilibrium allows high resolution mapping, *Hum. Mol. Genet.* 2 (1993) 1229-1234.
- [13] Lee S.H., Van der Werf J.H.J., The role of pedigree information in combined linkage disequilibrium and linkage mapping of quantitative trait loci in a general complex pedigree, *Genetics* 169 (2005) 455-466.
- [14] Lewontin R.C., On measures of gametic disequilibrium, *Genetics* 49 (1964) 49-67.
- [15] Lou X.Y., Todhunter R.J., Lin M., Lu Q., Wang Z., Bliss S.P., Casella G., Acland G.M., Lust G., Wu R., The extent and distribution of linkage disequilibrium in a multi-hierarchical outbred canine pedigree, *Mamm. Genome* 14 (2003) 555-564.
- [16] MacCluer J.W., VandeBerg J.L., Read B., Ryder O.A., Pedigree analysis by computer simulation, *Zoo. Biol.* 5 (1986) 147-160.
- [17] McRae A.F., McEwan J.C., Dodds K.G., Wilson T., Crawford A.M., Slate J., Linkage disequilibrium in domestic sheep, *Genetics* 160 (2002) 1113-1122.

**- Article 3 -**

- [18] Maynard Smith J., Haigh J., The hitch-hiking effect of a favourable gene, *Genet. Res. Camb.*, 23 (1974) 23-35.
- [19] Meuwissen T.H.E., Goddard M.E., Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci, *Genetics* 155 (2000) 421-430.
- [20] Odani M., Narita A., Watanabe T., Yokouchi K., Sugimoto Y., Fujita T., Oguni T., Matsumoto M., Sasaki Y., Genome-wide linkage disequilibrium in two Japanese beef cattle breeds, *Anim. Genet.* 37 (2006) 139-144.
- [21] Slatkin M., Disequilibrium mapping of quantitative-trait locus in an expanding population, *Am. J. Hum. Genet* 64 (1999) 1765-1773.
- [22] Terwilliger J.D., A powerful likelihood method for the analysis of LD between trait loci and one or more polymorphic loci, *Am. J. Hum. Genet.*, 56 (1995) 777-787.
- [23] Vallejo R.L., Li Y.L., Rogers G.W., Ashwell M.S., Genetic diversity and background linkage disequilibrium in the north American Holstein cattle population, *J. Dairy Sci.* 86 (2003) 4137-4147.
- [24] Xiong M. and Guo S.W., Fine-scale mapping of quantitative trait loci using historical recombinations, *Genetics* 145 (1997) 1201-1218.
- [25] Yamazaki T., The effects of overdominance on linkage in a multilocus system, *Genetics* 86 (1977) 227-236.
- [26] Zenger K.R., Khatkar M.S., Cavanagh J.A.L., Hawken R.J., Raadsma H.W., Genome-wide genetic diversity of Holstein friesian cattle reveals new insights into Australian and global populations variability, including impact of selection, *Anim Genet.* 38 (2007) 7-14.
- [27] Zhao H.H., Fernando R.L., Dekkers J.M.C., Power and precision of alternate methods for linkage disequilibrium mapping of quantitative trait loci, *Genetics* 175 (2007) 1975-1986.

Deuxième partie : Evolution de la structure du déséquilibre de liaison sous l'influence de la sélection

Partie 2.2. Structure du déséquilibre de liaison dans des populations sélectionnées

**- Article 3 -**

[28] Zhao H., Nettleton D., Soller M., Dekkers J.C.M., Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL, *Genet. Res. Camb.* 86 (2005) 77-87.

- Article 3 -

**Table I.** Minimum confidence interval (in cM) and corresponding number of generations in a population of 150 individuals.

LD between the QTL and Marker	q <sup>(a)</sup>	Sparse map				Dense map			
		D'		$\chi^2$		D'		$\chi^2$	
		T <sup>(b)</sup>	Min.	T <sup>(b)</sup>	Min.	T <sup>(b)</sup>	Min.	T <sup>(b)</sup>	Min.
		95 %		95 %		95 %		95 %	
		IC		IC		IC		IC	
		(cM)		(cM)		(cM)		(cM)	
	1	50	3	60	3	50	2.25	100	1.75
	0.8	25	7	25	7	25	3.75	50	3.25
H2	1	50	4	25	5	25	2.75	25	3.00
	0.8	25	5	25	6	50	1.75	60	1.75
H4	1	25	5	25	6	25	1.75	25	1.75
	0.8	10	7	10	7	25	1.75	25	1.75

<sup>(a)</sup> : proportion of the population conserved as potential parents ; <sup>(b)</sup> : number of generations (computed for 0, 10, 25, 50, 60, 75 and 100 generations).

- Article 3 -

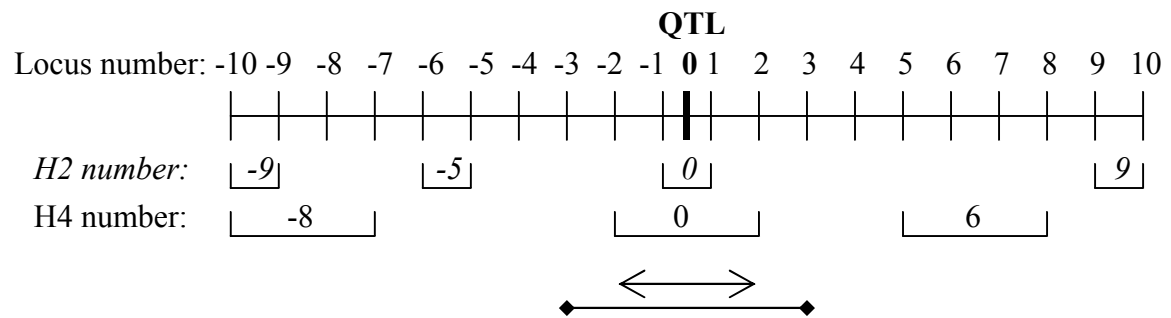
**Table II.** Maximum concentration and corresponding number of generations in a population of 150 individuals.

LD between the QTL and	q <sup>(a)</sup>	Sparse map				Dense map			
		D'		$\chi^2$		D'		$\chi^2$	
		T	Max. concen- tration (%)	T	Max. concen- tration (%)	T	Max. concen- tration (%)	T	Max. concen- tration (%)
Marker	1	50	96.2	100	96.6	75	78.3	100	82.8
	0.8	25	84.2	50	85.7	50	54.0	100	68.5
H2	1	50	96.9	60	95.8	100	90.0	100	87.7
	0.8	25	90.0	25	88.1	50	67.7	50	69.3
H4	1	25	97.8	25	97.3	60	90.2	60	90.0
	0.8	25	95.9	25	95.6	25	82.4	25	82.4

<sup>(a)</sup> : proportion of the population conserved as potential parents ; <sup>(b)</sup> : number of generations (computed for 0, 10, 25, 50, 60, 75 and 100 generations).



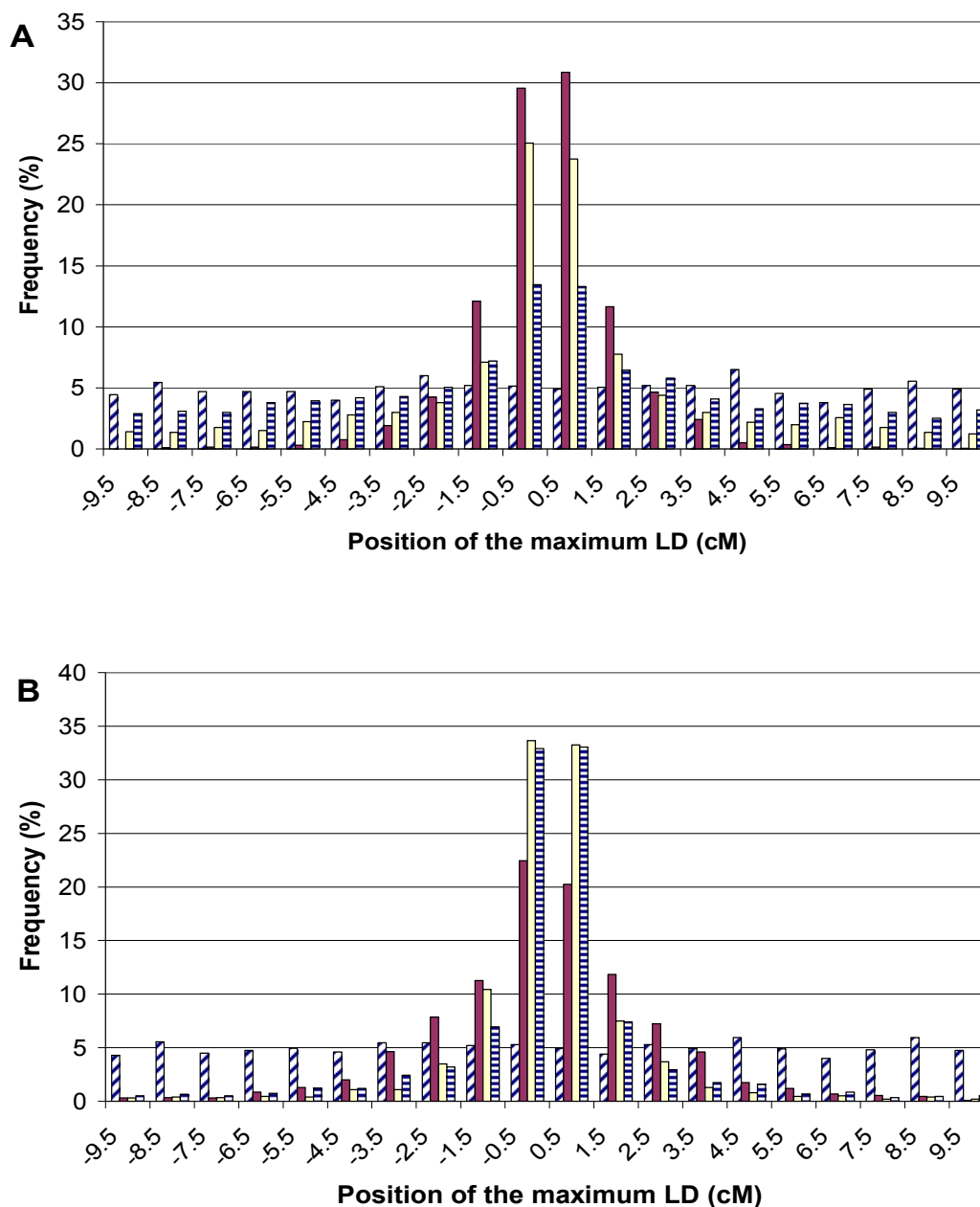
- Article 3 -



- ◆—◆ Interval studied with 4-loci haplotypes (5 cM on the sparse map, 1.25 cM on the dense map)
- ↔ Interval studied with marker and 2-loci haplotypes (3 cM on the sparse map, 0.75 cM on the dense map)

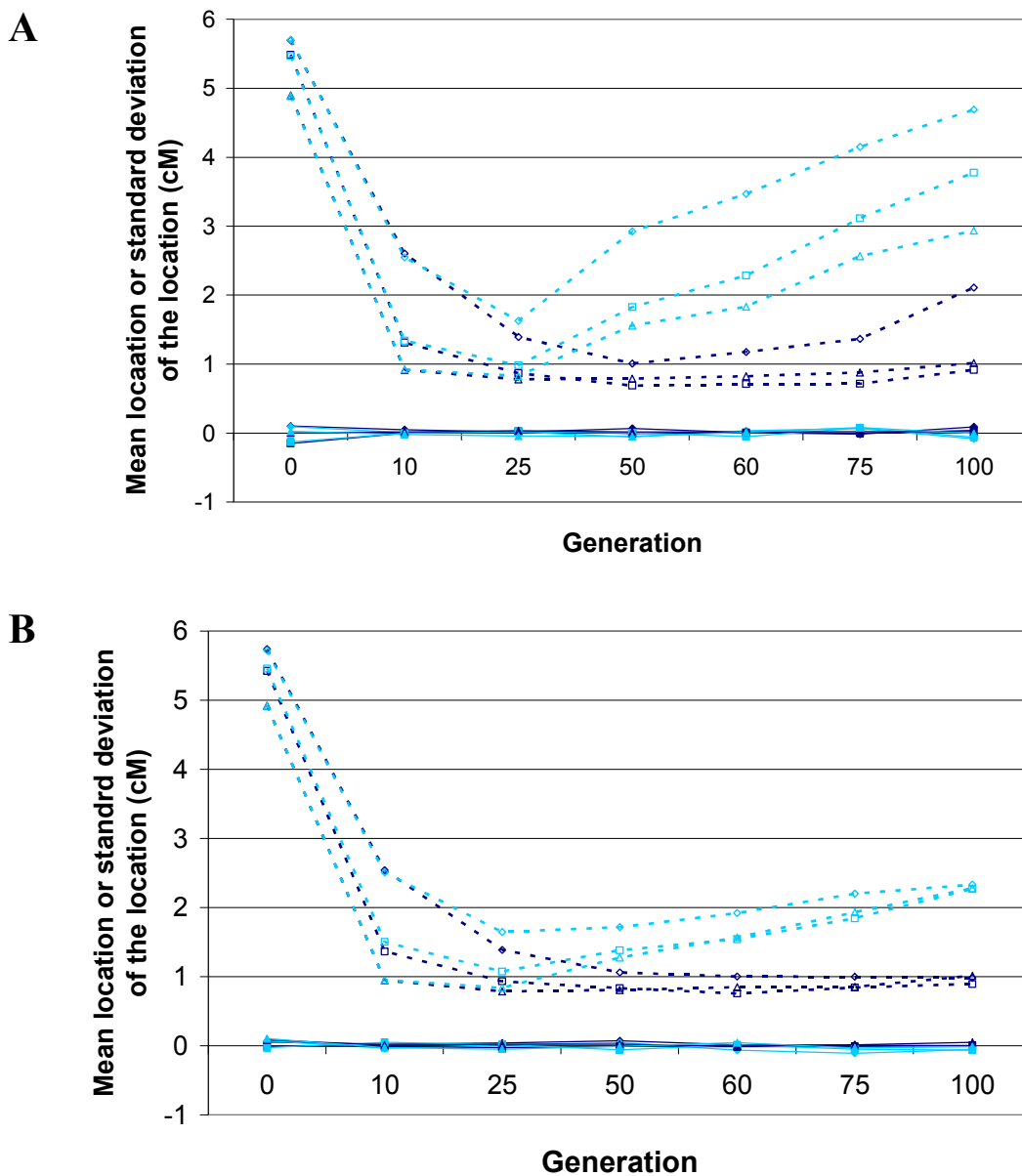
**Figure 1.** Notations for locus, haplotypes, and intervals studied.

- Article 3 -



**Figure 2.** Distribution of the positions of maximum LD between a marker and the QTL for a population of 150 individuals, using the sparse map, after 0 (diagonally striped bars), 25 (dark bars), 60 (white bars) or 100 generations (horizontally striped bars). A:  $q=0.8$ , LD measured with the  $D'$ ; B:  $q=0.8$ , LD measured with the  $\chi^2$ .

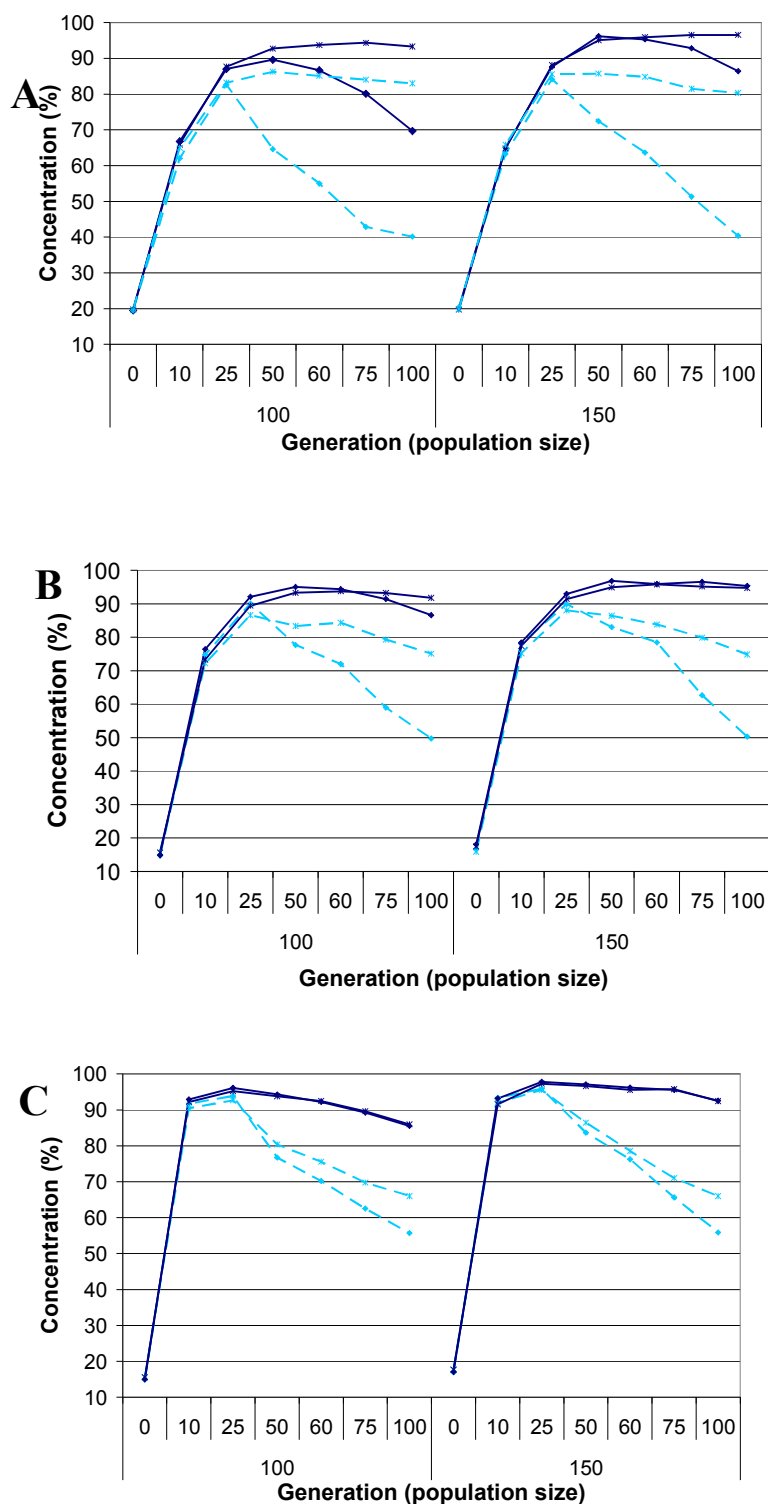
- Article 3 -



**Figure 3.** Evolution of the average distance between the QTL and the individual marker, the 2- or the 4-marker haplotype in maximum LD with it and of its standard deviation on a sparse map in a population of 150 individuals (A) when LD is measured with  $D'$ , (B) when LD is measured with  $\chi^2$ . Solid line: average distance; dotted line: standard deviation of the location.

Light lines:  $q=0.8$ ; dark lines:  $q=1$ . Diamonds: marker, squares: H2, triangles: H4.

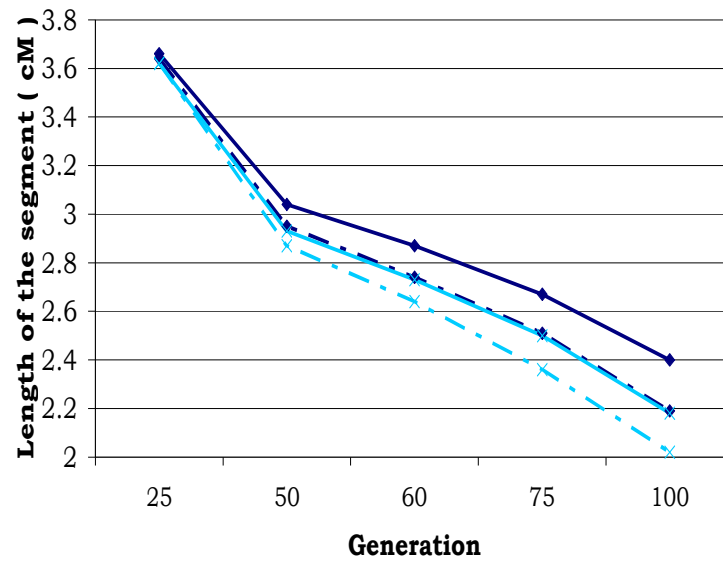
- Article 3 -



**Figure 4.** Concentration of the maximum LD around the QTL on the dense map. A: marker,

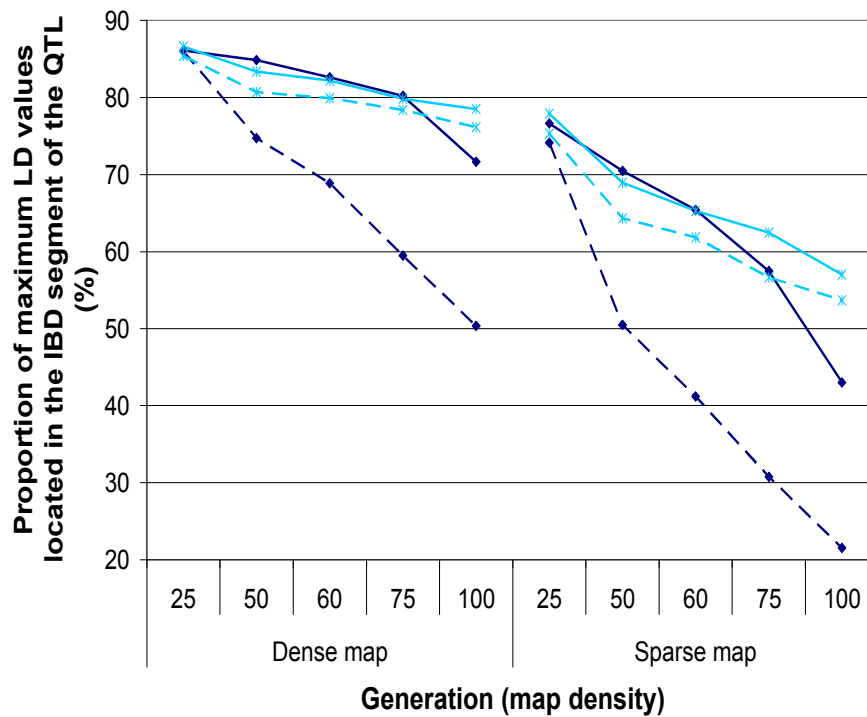
B: H2, C: H4. Diamonds:  $D'$ , crosses:  $\chi^2$ ; solid dark line:  $q=1$ , dotted light line:  $q=0.8$ .

- Article 3 -



**Figure 5.** Average length of the IBD segment containing the QTL position over the generations on a dense map. Solid line:  $q=0.8$ ; dotted line:  $q=1$ ; diamonds: 100 individuals; crosses: 200 individuals.

- Article 3 -



**Figure 6.** Percentage of replicates among 2,000 where the position of maximum marker-QTL LD belonged to the IBD segment of the QTL for a population size of 150 individuals.

Diamonds:  $D'$ , crosses:  $\chi^2$ ; solid dark line:  $q=1$ , dotted light line:  $q=0.8$ .

## II. Résultats complémentaires sur les locus en LD maximum avec le QTL et discussion finale

Ce complément rapporte les différences observées entre estimateurs du LD ainsi que l'effet de la taille de la population sur les différents critères. Le type de simulations utilisé dans cette partie, ainsi que les critères de comparaison entre situations, sont les mêmes que dans l'article.

Tous les paramètres sur la distribution des positions en LD maximum avec le QTL s'accordent sur plusieurs points, quelle que soit la mesure de LD employée :

- Un regroupement maximal des positions en plus fort LD avec le QTL, caractérisé par un écart-type de la localisation, un intervalle de confiance minimal ou une concentration maximale, sont obtenus pour de nombreux schémas simulés bien avant la centième génération, en général entre les générations 25 et 60.
- Au-delà de cette génération d'optimum, les positions détectées en LD maximal s'éloignent du QTL.
- Ce phénomène de « regroupement – écartement » est d'autant plus accentué que la population est de taille restreinte ou que la sélection appliquée est forte.
- Le regroupement est obtenu plus tardivement sur la carte dense, quelle que soit la mesure de LD utilisée ou la statistique considérée.

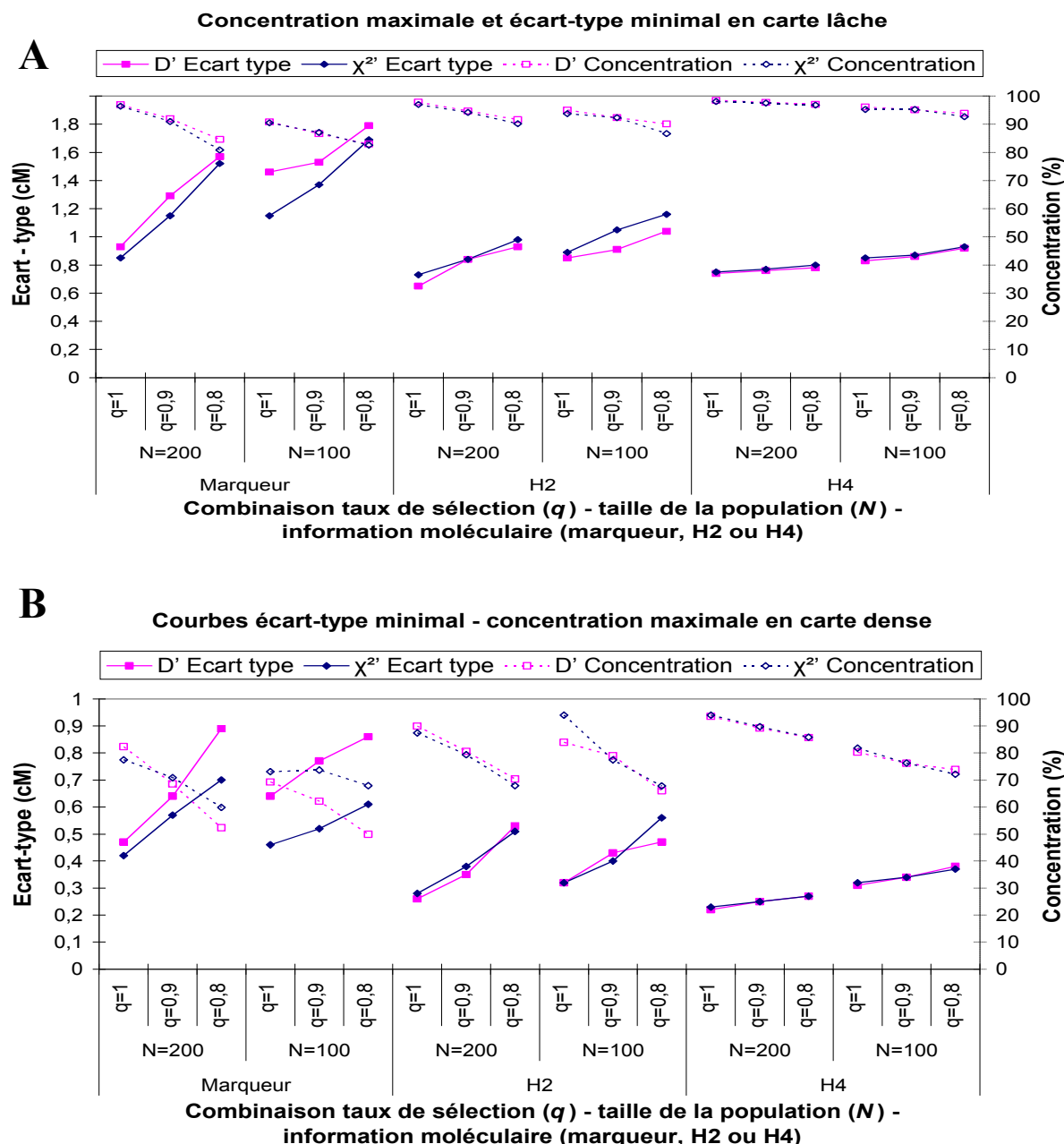
Cependant la génération du regroupement maximal diffère selon l'indicateur statistique et la mesure de LD utilisés (Tableaux 3A et 3B). Le  $D'$  indique souvent un regroupement plus précoce que le  $\chi^2$ . Ceci peut être dû à sa sensibilité à la perte allélique et donc haplotypique. En effet, les variations enregistrées entre deux positions sont moins nettes lorsque des allèles ou des haplotypes atteignent des fréquences extrêmes, le  $D'$  tendant fortement vers 1. La localisation de la position en LD maximum correspond alors à un écart entre les deux  $D'$  très faible, ce qui n'est pas le cas du  $\chi^2$ , peu sensible aux pertes alléliques. En ce qui concerne les paramètres de contrôle de la dispersion, l'intervalle de confiance à 95% est souvent le facteur le plus rapide à atteindre son minimum.

<i>Carte lâche A</i>	Taille de la population	Sélection	D'			$\chi^2$		
			Ecart type	95% CI	Concentration	Ecart type	95% CI	Concentration
Marqueur	200	1	60	25	<b>75</b>	<b>100</b>	75	75
		0.8	25	<b>50</b>	25	25	25	<b>50</b>
	100	1	25	25	<b>50</b>	<b>75</b>	25	<b>75</b>
		0.8	<b>25</b>	<b>25</b>	<b>25</b>	<b>50</b>	25	<b>50</b>
H2	200	1	<b>60</b>	25	<b>60</b>	60	25	<b>75</b>
		0.8	<b>25</b>	<b>25</b>	<b>25</b>	25	25	25
	100	1	<b>50</b>	25	<b>50</b>	<b>60</b>	25	<b>60</b>
		0.8	<b>25</b>	<b>25</b>	<b>25</b>	<b>25</b>	<b>25</b>	<b>25</b>
H4	200	1	<b>50</b>	25	25	<b>50</b>	25	25
		0.8	<b>25</b>	<b>25</b>	<b>25</b>	<b>25</b>	10	<b>25</b>
	100	1	<b>25</b>	10	<b>25</b>	<b>25</b>	10	<b>25</b>
		0.8	<b>25</b>	10	<b>25</b>	<b>25</b>	10	<b>25</b>

<i>Carte dense B</i>	Taille de la population	Sélection	D'			$\chi^2$		
			Ecart type	95% CI	Concentration	Ecart type	95% CI	Concentration
Marqueur	200	1	<b>100</b>	75	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
		0.8	<b>50</b>	25	<b>50</b>	<b>100</b>	25	<b>100</b>
	100	1	50	50	<b>60</b>	<b>100</b>	<b>100</b>	<b>100</b>
		0.8	<b>25</b>	<b>25</b>	<b>25</b>	<b>100</b>	50	<b>100</b>
H2	200	1	75	75	<b>100</b>	<b>100</b>	50	<b>100</b>
		0.8	25	25	<b>50</b>	50	25	<b>60</b>
	100	1	50	50	<b>60</b>	75	75	<b>100</b>
		0.8	<b>25</b>	<b>25</b>	<b>25</b>	<b>50</b>	25	<b>50</b>
H4	200	1	<b>50</b>	25	<b>50</b>	75	25	<b>75</b>
		0.8	<b>25</b>	<b>25</b>	<b>25</b>	<b>25</b>	<b>25</b>	<b>25</b>
	100	1	<b>60</b>	25	50	<b>50</b>	<b>50</b>	<b>50</b>
		0.8	<b>25</b>	<b>25</b>	<b>25</b>	<b>25</b>	<b>25</b>	<b>25</b>

**Tableaux 3 : Nombre de générations nécessaires pour atteindre l'optimum de chacun des indicateurs de regroupement du locus en LD maximum avec le QTL autour de celui-ci (écart-type de la localisation, intervalle de confiance à 95% -95% CI- et concentration). A : valeurs obtenues avec la carte lâche ; B : valeurs obtenues sur la carte dense. En gras : génération la plus tardive d'obtention du regroupement maximal**





**Figure 2.1.7 : Ecarts-types minimaux et concentrations maximales pour les différents designs. A : en carte lâche ; B : en carte dense.**

L'intervalle de confiance à 95% est aussi l'indicateur le moins sensible à la taille de la population (et donc à la dérive) ainsi qu'à la sélection (Tableau 4), même s'il montre une tendance à l'augmentation si la population est sélectionnée et/ou de petite taille. La concentration des maxima autour du QTL et l'écart-type de l'estimation de la localisation du locus en LD maximum avec le QTL sont pour leur part nettement affectés par ces deux facteurs (Figures 2.1.7 A et B). La concentration augmente avec la taille de la population, et quand la population n'est pas sélectionnée. Dans les mêmes conditions, l'écart-type

Deuxième partie : Evolution de la structure du déséquilibre de liaison sous l'influence de la sélection

Partie 2.2. Structure du déséquilibre de liaison dans des populations sélectionnées d'estimation diminue (Figures 2.1.7). L'utilisation d'haplotypes plus longs augmente également la concentration et diminue l'écart-type de la localisation estimée du locus. Ces tendances sont observées aussi bien avec le  $D'$  qu'avec le  $\chi^2$ . En effet, dans les populations de taille restreinte et/ou soumises à la sélection, la fixation des allèles est plus rapide. Du fait de l'effet d'entraînement exercé par le QTL, les fixations sont plus rapides à proximité de ce locus (Maynard-Smith et Haigh, 1974). Une fois fixés, ces locus ne sont plus en LD avec le QTL (toutes les mesures de LD vaudront 0), et le locus détecté en LD maximal avec le QTL est alors situé plus loin du QTL. Etant donné que le QTL n'est pas fixé dans les simulations retenues pour cette étude, il paraît peu probable que tous les locus proches de celui-ci soient fixés : la localisation, quoique plus lointaine, reste dans un intervalle restreint autour du QTL, ce qui explique probablement la moindre sensibilité à la sélection de l'intervalle de confiance à 95%.

Information moléculaire	Taille de la population	Sélection	Carte lâche		Carte dense	
			$D'$	$\chi^2$	$D'$	$\chi^2$
Marqueur	200	1	3	3	1.75	2.00
		0.8	7	7	3.75	3.50
	100	1	5	6	3.00	2.75
		0.8	7	9	3.75	3.50
H2	200	1	5	5	1.25	1.75
		0.8	5	5	2.75	3.00
	100	1	5	5	1.75	1.75
		0.8	5	7	2.50	3.00
H4	200	1	5	5	1.75	1.75
		0.8	6	7	1.75	1.75
	100	1	7	7	2.25	2.00
		0.8	7	7	2.25	2.25

**Tableau 4 : Intervalles de confiance à 95% minimaux (en cM) atteints pour les différents designs**

Si les nombres de générations optimaux obtenus avec les deux mesures de LD sont en général assez similaires, l'évolution des différents indicateurs après avoir atteint cet optimum est très différente selon l'information moléculaire utilisée, et de l'indicateur du regroupement des locus en LD maximum autour du QTL. Ces évolutions différentes sont liées à la sensibilité relative des deux mesures de LD. Le  $D'$  a plus de chances de conserver sa valeur maximale de 1, atteinte dès qu'une combinaison allèle du locus marqueur – allèle du QTL est absente, lorsque des haplotypes de 4 locus sont utilisés : d'une part, la fixation des allèles est

Deuxième partie : Evolution de la structure du déséquilibre de liaison sous l'influence de la sélection

Partie 2.2. Structure du déséquilibre de liaison dans des populations sélectionnées plus lente dans ce cas de figure qu'avec des marqueurs seuls, et d'autre part ces pertes alléliques seront plus fréquentes avec des haplotypes de 4 marqueurs. Pour sa part, le  $\chi^2$  est moins sensible à ces absences, et il fera moins de différences entre les types d'information moléculaire.

La sélection affecte donc de façon importante la concentration du LD à proximité du QTL, que ce soit en terme de valeur ou de localisation de ce maximum d'association ou d'évolution de cette association. Cependant, un nombre de génération consensuel pour obtenir une association maximale n'a pas pu être dégagé. Par ailleurs, ce paramètre est incontrôlable quand il s'agit d'analyses de populations réelles. Nous pouvons *a priori* penser que les populations animales actuelles, sur lesquelles sont pratiquées les analyses de cartographie fine, sont dans la phase où les locus en LD maximum avec le QTL sont en moyenne assez loin du QTL compte tenu de leur âge, des pressions de sélection auxquelles elles sont soumises et de leur faible taille efficace, sauf peut-être les espèces à grand intervalle générationnel (bovins, équins). Cependant, la mutation, qui maintient un équilibre dans la variabilité génétique, n'a pas été envisagée dans cette étude et peut agir comme un facteur limitant voire contrecarrant les phénomènes observés dans ces simulations.

### III. Conclusion

Il résulte de cette étude que le LD actuellement disponible dans les populations soumises à sélection est sans doute peu favorable à une localisation précise des QTL, d'autant plus que leur taille efficace est limitée. En effet, les segments IBD autour du QTL sont allongés, et ils contiennent moins souvent les locus en LD maximum avec le QTL. De plus, la présence de sélection rend plus précoce le regroupement des locus en LD fort autour du QTL et accélère la dispersion de ceux-ci. Cependant, cette perte potentielle d'efficacité des méthodes de cartographie fine est probablement limitée dans les études sur les populations réelles car les locus marqueurs utilisés ont par définition plusieurs allèles. En situation réelle, il ne paraît donc pas nécessaire en première approche de supprimer les marqueurs ayant des fréquences alléliques extrêmes. En effet, si de tels marqueurs, utilisés seuls, vont mettre en évidence des locus plus éloignés du QTL, mais ayant des fréquences intermédiaires, l'utilisation d'haplotypes doit pallier à ce déséquilibre en multipliant le nombre d'allèles.

Deuxième partie : Evolution de la structure du déséquilibre de liaison sous l'influence de la sélection

Partie 2.2. Structure du déséquilibre de liaison dans des populations sélectionnées

Enfin, il est sans aucun doute très efficace, en termes de résolution de cartographie, d'analyser simultanément différentes populations ayant divergé depuis quelques dizaines ou centaines de générations, car les segments ancestraux entourant le QTL sont beaucoup plus réduits qu'intra population.



**TROISIEME PARTIE :**  
**INFLUENCE DE LA SELECTION**  
**SUR LES METHODES DE**  
**CARTOGRAPHIE FINE**



## **Partie 3.1 :**

### **Influence de la sélection sur les probabilités d'IBD**





## Introduction

L'étude précédente sur la distance entre le locus en LD maximum avec le QTL et le segment fondateur contenant le QTL a initié une étude plus approfondie sur l'influence de la sélection sur la qualité d'estimation des probabilités d'IBD. Celles-ci sont intégrées dans la méthode de cartographie fine LDLA (Linkage Disequilibrium and Linkage Analysis), devenue classique en génétique animale (Meuwissen et Goddard, 2000), comme marqueurs du LD. Ces auteurs (2001) ont développé une méthode analytique pour calculer la probabilité qu'ont deux haplotypes d'être IBD d'après leurs statuts IBS aux différents marqueurs les composant. L'efficacité de ces probabilités pour la cartographie fine a été testée dans plusieurs études (Grapes *et al.*, 2006 ; Zhao *et al.*, 2007). Cependant, deux aspects n'ont pas été envisagés, bien qu'ils puissent avoir des répercussions sur la qualité de l'identification du locus QTL :

- la capacité à distinguer les allèles IBD au QTL des allèles non IBD, et la possible implication pour réduire la taille des matrices d'IBD utilisées en cartographie fine (article),
- le lien entre les probabilités d'IBD, telles qu'estimées dans la méthode de calcul de Meuwissen et Goddard, et le LD.

Comme précédemment, nous avons utilisé les données produites par le simulateur décrit en partie 2.1. L'information sur le fondateur dont est issue la copie du QTL de chaque haplotype présent dans la population après 100 générations a été utilisée comme marqueur du vrai statut IBD entre les locus. Cette information a été combinée aux probabilités d'IBD calculées à l'aide du module fourni dans l'article de Meuwissen et Goddard (2001) pour évaluer la relation entre les probabilités d'IBD estimées et le statut vrai, présentée dans l'article de la première partie de ce chapitre. La valeur du LD maximum entre le QTL et son marqueur le plus proche a été notée, et la corrélation entre le statut IBD vrai au QTL et la probabilité d'IBD a été calculée, pour observer l'évolution de cette corrélation en fonction du niveau de LD maximum dans l'haplotype. Ces données complètent l'article pour clore la première partie. Finalement, les performances comparées pour la cartographie fine de QTL sont examinées dans la seconde partie de ce chapitre.

- Article 4 -

**I. Article**

**Comparison of estimated Identity-by-Descent probabilities and true Identity-by-Descent status of chromosomal regions in a population**

**F. Ytournel, H. Gilbert, D. Boichard**

**Summary:**

The goal of this simulation study was to evaluate the properties of Identity-by-Descent (IBD) probability estimates at a QTL location given surrounding marker information. Different genetic maps were simulated with various lengths, marker densities and markers informativity (SNP, microsatellites, or a mixture of both). The simulated populations included 100 individuals over 100 random mating generations. These populations were either unselected or subjected to selection by truncation over the last 20 generations with 80% of the individuals being possible parents of the next generation. The IBD probability estimated with haplotypes composed of 6 markers was compared to the true IBD status for all maps and selection pressure. We also looked for IBD probabilities that could be used as thresholds to cluster them while keeping a type I or a type II error rate under 5%. Non-IBD QTL were better identified than IBD ones with all designs. The discrimination ability improved with markers informativity, with dense haplotypes. Selection decreased the discrimination ability of IBD probabilities. However, it appeared that haplotype pairs with medium IBD probabilities could be clustered with type I error rates lower than 5%, even in selected populations.

- Article 4 -

**Introduction**

Linkage Disequilibrium (LD) has become of common use for QTL mapping. Meuwissen and Goddard (2000, 2001) proposed a now widely used model to take advantage of LD, fitting a random QTL effect with a covariance structure proportional to Identity-By-Descent (IBD) probabilities to provide LD information. They proposed an analytical estimation of these probabilities according to the Identity By State (IBS) status of markers for a given haplotype surrounding the putative QTL position.

Recent studies (Grapes *et al.*, 2006; Zhao *et al.*, 2007) investigated the accuracy of this method for QTL fine mapping. Basically, however, the efficiency of the method relies on its ability at discriminating IBD from non-IBD loci, which relates to the discrimination between IBS and IBD loci. This depends on one hand on the molecular information available (marker informativity and haplotype length), and on the other hand on the evolutionary forces which created or maintained the LD in the population. Limited information provides intermediate estimated IBD probabilities quite different from the true status (0 or 1) and therefore affects the covariance structure assumed for fine mapping. Additionally, a good discrimination ability makes it possible to reduce the size of the IBD matrix by gathering IBD QTL in one cluster. Such a clustering procedure more easily ensures the matrix to be positive definite and saves computation time. In this study, we investigated the impact of the discrimination ability on the inference of the IBD status at the QTL position using the IBD probabilities proposed by Meuwissen and Goddard (2001), based on simulated data.

**I. Methods**

I.1. Simulated designs

The populations were composed of 100 individuals and evolved over 100 generations. Matings were random, selfing being allowed. Selection, when applied, occurred only over the

- Article 4 -

last 20 generations. In this case, the 80% animals with the best phenotypes were retained as potential parents for the next generation. Designs with selection are denoted *s* in the following. The phenotypes were simulated as the sum of an additive QTL effect, a polygenic effect and a residual effect. Residual and within-family Mendelian sampling effects were normally distributed with 0 mean and constant variances over time. The initial heritability of the trait was 5%. A bi-allelic QTL with a 0.2 initial frequency of the favourable allele explained 20% of the genetic variance in the founder population.

I.2. Genetic maps

The genetic maps were simulated with six markers and a QTL always located in the middle of the haplotype (Figure 1). The map length was comprised between 0.5 cM and 1.0 cM, depending on the interval between the QTL and its two adjacent markers and on the distribution of the markers. The maps were either composed of SNP (A, B and C), microsatellites only (A-M), or a mixture of 4 SNP and 2 microsatellite markers (A-V and B-V). The IBD probabilities were computed at the QTL position according to the analytical formula of Meuwissen and Goddard (2001), assuming that the marker haplotypes are known. Founder alleles were drawn at random, assuming initial linkage equilibrium. Thus, haplotypes may not be unique and there may be founder haplotypes being IBS at all marker loci but not at the QTL locus.

*Figure 1 here*

I.3. IBD probability accuracy

One thousand replicates were obtained for each simulated design, the IBD status of the QTL being recorded over the generations. Each founder QTL locus was initially given a copy number (1 to 2N, N being the number of founders), which was transmitted from the parents to

- Article 4 -

their offspring at each generation and was thus used to distinguish IBD from non IBD locus. The IBD probabilities were computed for each pair of haplotypes and were then distributed into ten classes, with class 1 corresponding to IBD probabilities between 0 and 0.1 up to class 10 corresponding to IBD probabilities over 0.9. The IBD and non-IBD occurrences were then distinguished into each class.

Probability thresholds T1 to T10 were defined every 0.1 steps between 0 and 0.9, to set up a power test for the inference of the IBD status based on the IBD probabilities. The type I and the type II errors corresponding to each threshold were then computed. The type I error for threshold  $i$  corresponded to the probability for a haplotype pair of probability greater than  $T_i$  of being considered IBD while it is not. It was thus defined for  $i, i=1,10$  as

$$\alpha_i = \frac{\sum_{k=i}^{10} n_k(IBD)}{\sum_{k=i}^{10} [n_k(IBD) + n_k(nonIBD)]}$$

with  $n_k(IBD)$  and  $n_k(nonIBD)$  being respectively the number

of true IBD and true non-IBD pairs in class  $k$ . A threshold value for a 5% type-I error test was defined as the lowest  $T_i$  with an error rate lower than 5%.

The type II error for threshold  $i$  was computed as the probability for a haplotype pair of probability lower than  $T_i$  of being considered non-IBD while it is: for  $i, i=1,10$ ,

$$\beta_i = \frac{\sum_{k=1}^i n_k(IBD)}{\sum_{k=1}^{10} n_k(IBD)}$$

The power was finally calculated as one minus the type II error rate. A threshold value for a 5% type II error corresponded to the largest  $T_i$  with  $\beta_i$  immediately lower than 5%.

- Article 4 -

**II. Results**

II.1. Distribution of the IBD probabilities

For all genetic maps, a great majority of the highest probabilities corresponded to pairs of haplotypes truly IBD at the QTL locus, while the lowest probabilities were mostly due to non-IBD pairs of haplotypes (Figure 2). For example, with design A-M, 98.3% of the haplotype pairs were non-IBD in class 1 and 99.8% of the pairs in class 10 were IBD.

Non-IBD QTL were always better discriminated than IBD ones. For instance, with map A, 84.7% of the non-IBD QTL pairs had probabilities lower than 0.1 while 49.2% of the IBD pairs had probabilities over 0.9. The map density in the QTL neighbourhood improved the discrimination ability (A vs. C), with more non-IBD loci with low probabilities and more IBD loci with high probabilities. Using haplotypes composed only of microsatellites also increased the discrimination ability (A vs. A-M), but the maps composed of a mixture of microsatellites and SNP showed results similar to those of maps composed only of SNP (A vs. A-V).

Selection strongly reduced the frequencies of the highest class of IBD probabilities (28% less IBD QTL pairs in class 10 between A and As). This coincided with the increase of the frequencies of the truly IBD QTL in the classes 5 to 8 and thus to the overall frequency of these classes of probabilities. On the other hand the distribution of the probabilities of non-IBD QTL pairs changed little.

II.2. Thresholds for type I and type II error rates

It clearly appeared that 5% type I errors were achieved with low thresholds for IBD probabilities (Figure 2), particularly when the population was subjected to selection, with denser maps for and/or maps composed only of microsatellites. The use of two microsatellites had little influence. The associated powers were frequently over 50%, and selection helped in increasing these powers.

- Article 4 -

The probability threshold for a 5% type II error rate was rather similar across all designs. Once more, selection led to a reduction of these thresholds (under 0.1). However, such low thresholds were achieved at the expense of high type I error rates, except on the densest map with microsatellite markers without selection (A-M).

### III. Discussion and conclusion

This study aimed at evaluating the ability of IBD probabilities to infer the IBD status at a QTL given the marker information. A practical implication concerns the definition of clustering thresholds. Gathering IBD QTL in the same cluster reduces the IBD matrix size, makes it positive definite more easily and reduces computation time.

In this study, the distributions of IBD probabilities are influenced by both marker informativity and marker density around the tested position. However, for all studied designs, the discrimination between IBD loci and non-IBD loci by IBD probabilities was good, particularly for non-IBD QTL, with very low chances of observing false-positives: in the U-shaped distribution of the probabilities, very low frequencies were obtained in the middle-range IBD probabilities. Maps composed only of microsatellites provided the most discriminant distributions of IBD probabilities, *i.e.* with the greatest proportions of IBD QTL with IBD probabilities over 0.9 and non-IBD QTL with probabilities under 0.1. As these dense haplotypes of microsatellites are rather unrealistic, we checked the properties of haplotypes composed of both microsatellites and SNP to provide additional haplotype variability compared to SNP haplotypes. However, this did not increase the discrimination ability of IBD loci, maybe because of the internal disequilibrium of information it created. It suggested that the best results may be obtained with dense, balanced and polymorphic haplotypes, related to high SNP densities. Further simulations were conducted with 4- or 10-marker haplotypes (data not shown). The same tendencies were observed, with best



- Article 4 -

discrimination ability obtained on dense, balanced and polymorphic haplotypes. However, it appeared that there was little difference between 4- and 10-marker haplotypes on map A, indicating that the furthest markers did not bring much to the discrimination ability after a certain amount of information has been provided.

When individual selection was applied, the proportion of pairs of haplotypes IBD at the QTL locus among all haplotype pairs increased. The frequencies of IBD probabilities comprised between 0.5 and 0.8 strongly increased, corresponding to a decrease of the frequency of IBD QTL with probabilities over 0.9. According to this study, it is usually possible to set a probability threshold around 0.5 with a type I error rate of 5%. Such a low threshold is supported by few occurrences of non-IBD QTL with IBD probabilities over 0.3, and also by a reduced number of haplotypes segregating in the selected populations. Haplotypes are thus more frequently IBD, even when only the two flanking markers are IBS. In selected populations, the frequencies of class 5 increased dramatically compared to class 4, which happened only in class 9 or 10 compared to their relative previous class in unselected populations. It is noteworthy that estimated probabilities do not exactly behave as probabilities. In class 5 for instance, the majority of pairs are IBD whereas the expected proportion would have been around 50%.

The population size and number of generations used in this study correspond to those which had been used by Meuwissen and Goddard (2001) and are used by many authors when applying Meuwissen's and Goddard's mapping strategy (*e.g.* Meuwissen *et al.*, 2002; Gautier *et al.*, 2006). We chose to apply a moderate selection intensity on only 20 generations to avoid too much fixation at the QTL locus. However, it clearly appeared that even such a limited selection intensity had a deleterious impact on the distributions of IBD frequencies with regard to their discrimination ability.

- Article 4 -

One should also notice that we used the correct population size and number of generations to estimate the IBD probabilities. However, in the selected populations, population size does not correspond to the effective population size, but it should not affect our results as it seems that the computational method is robust to these assumptions (Meuwissen and Goddard, 2001). IBD probabilities were computed only at the QTL locus. When computed on the marker bracket next to the QTL, the discrimination ability was slightly worse as expected, but further simulations would be required at farther locations from the QTL to get a more general view of the distributions of IBD probabilities depending on the distance between the QTL and the middle of the haplotype.

The overall distribution of the IBD probabilities leads to a strategy for QTL clustering in the IBD probability matrix and to the definition of the corresponding probability thresholds, while maintaining a reasonable type I error rate. The improvement in terms of mapping accuracy through such a strategy is currently underway. The objective is twofold : reducing the size of the covariance matrix and then computation time and memory by keeping at least the same mapping efficiency ; but also improving the mapping efficiency by improving the covariance matrix by better accounting for the true IBD or not IBD status of the QTL pairs.

**References**

Gautier M., Barcelona R.R., Fritz S., Grohs C., Druet T. Boichard D., Eggen A., Meuwissen T.H., 2006. Fine mapping and physical characterization of two linked quantitative trait loci affecting milk fat yield in dairy cattle on BTA26. *Genetics* 172(1):425-436.

Grapes L., Firat M.Z., Dekkers J.M.C., Rotschild M.F., Fernando R.L., 2006. Optimal haplotype structure for linkage disequilibrium-based fine mapping of quantitative trait loci using identity by descent. *Genetics*, 172: 1955-1965.

- Article 4 -

Meuwissen T.H., Karlsten A., Lien S., Olsaker I., Goddard M.E., 2002. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* 161(1):373-379.

Meuwissen T.H.E., Goddard M.E., 2001. Prediction of identity by descent probabilities from marker haplotypes. *Genet. Sel. Evol.* 33: 605-634.

Meuwissen T.H.E., Goddard M.E., 2000. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics*, 155: 421-430.

Zhao H.H., Fernando R.L., Dekkers J.C.M., 2007. Power and precision of alternate methods for linkage disequilibrium mapping of quantitative trait loci. *Genetics*, 175: 1975-1986.

- Article 4 -

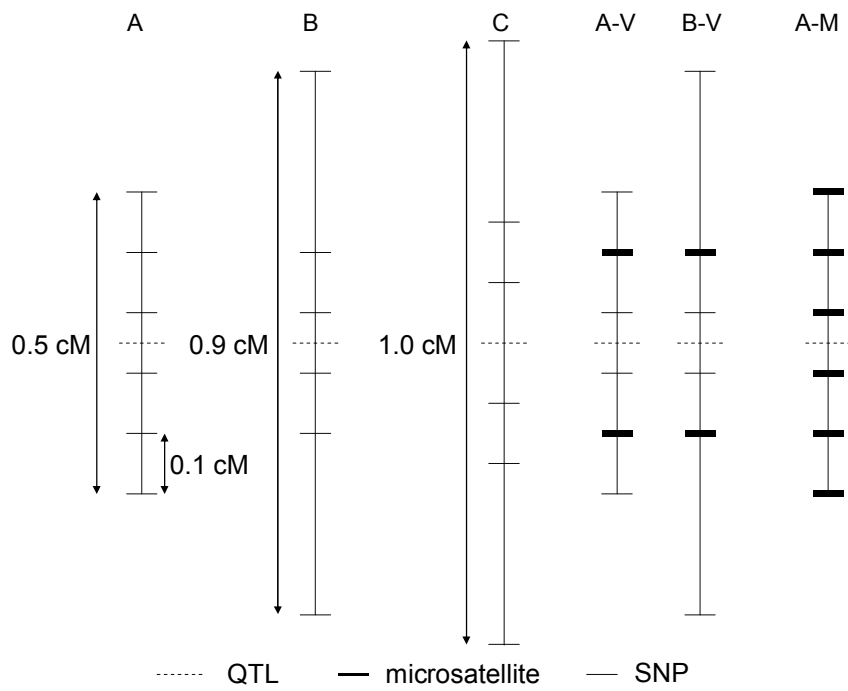


Figure 1: Different maps considered in the study.

- Article 4 -

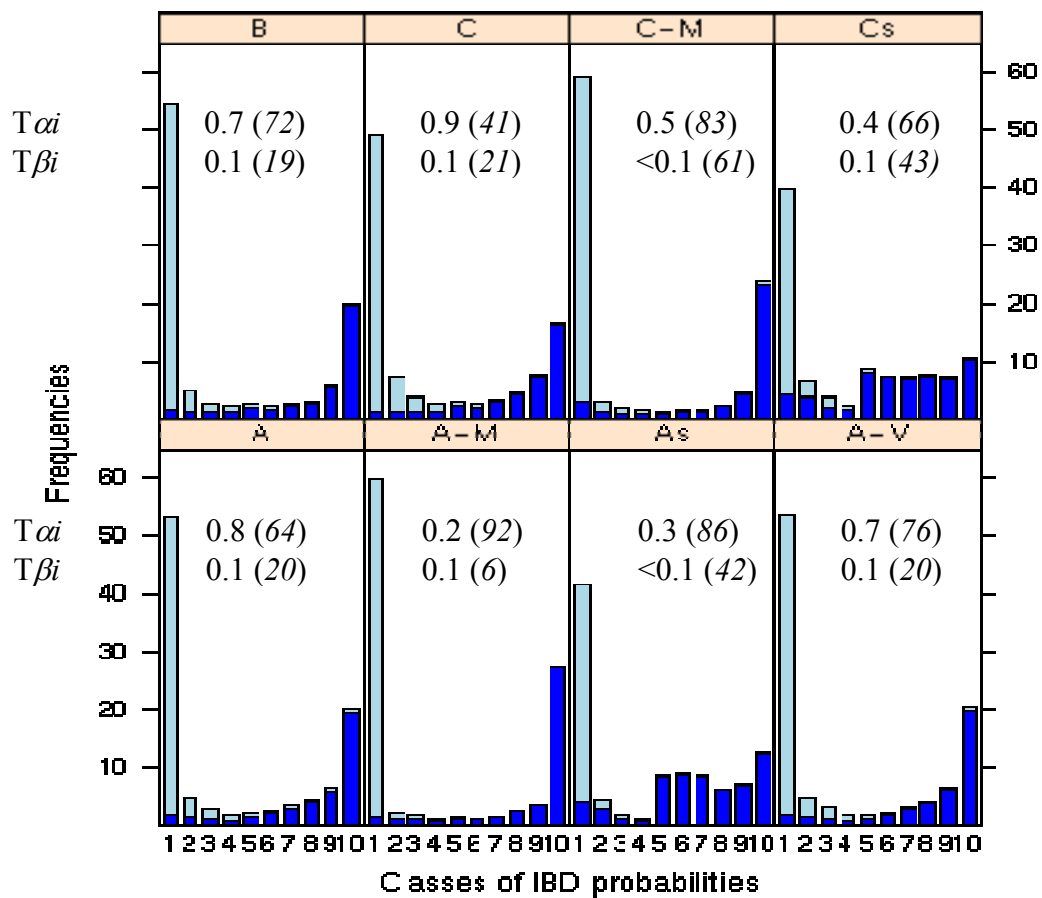


Figure 2: Distributions of the IBD probability frequencies for IBD QTL (dark blue) or non-IBD QTL (light blue), according to the haplotype length, the marker density around the QTL, the marker information and selection pressure. For each design,  $T_{\alpha i}$  = threshold for a 5% type I error and its associated power (%),  $T_{\beta i}$  = threshold for a 5% type II error and its associated type I error (%).

## **II. Lien entre l'IBD et le LD**

### **II.A. Critère de comparaison**

Les simulations sont conduites de la façon présentée dans l'article précédent. Les notations des dispositifs sont également reprises.

La corrélation entre la probabilité d'IBD et le statut IBD réel au QTL pour chaque couple d'haplotypes a été calculée pour chaque simulation. Le LD entre le QTL et le marqueur le plus proche a été subdivisé en 10 classes de la même manière que les probabilités IBD précédemment. Les corrélations ont ensuite été réparties dans ces classes et la moyenne intra-classe de LD est rapportée.

### **II.B. Résultats**

La corrélation entre le statut IBD réel au QTL et la probabilité d'IBD augmente de manière continue, mais avec une faible pente, avec un LD mesuré entre le QTL et son marqueur le plus proche croissant, et ce pour tous les dispositifs (Tableau 3.1.1, résultats uniquement présentés pour les deux classes extrêmes de LD). Il faut cependant relativiser les observations suivantes car chaque classe de LD regroupe un faible nombre de simulations, surtout la classe de LD supérieur à 0.9 (seulement 50 à 70 simulations).

Les corrélations sont très élevées pour tous les dispositifs, dépassant les 0,75 quelle que soit la classe du LD entre le QTL et son marqueur le plus proche. Il semble y avoir une légère augmentation de la corrélation avec la densité en marqueurs de l'haplotype (A vs. B vs. C), de même que lorsque tous les marqueurs de l'haplotype sont multi-alléliques.

Dispositif	Corrélation moyenne entre les probabilités d'IBD et le statut IBD réel du QTL, en fonction de la valeur de LD	
	0,0 < LD < 0,1	0,9 < LD ≤ 1,0
C	0,80	0,87
Cs	0,75	0,86
C-M	0,82	0,92
B	0,82	0,87
B-V	0,82	0,87
Bs-V	0,78	0,85
B-M	0,86	0,93
A	0,85	0,93
As	0,80	0,86
A-V	0,86	0,92
A-M	0,91	0,94
6As-M	0,83	0,87

**Tableau 3.1.1 : Corrélations moyennes entre le statut IBD réel du QTL et la probabilité d'IBD pour un LD inférieur à 0.1 ou supérieur à 0.9, avec des haplotypes de 6 marqueurs**

### III. Conclusion

Il résulte de cette étude que les probabilités d'IBD sont affectées par la sélection, qui affaiblit leur capacité discriminatoire sur les QTL réellement IBD. Cependant, quel que soit le niveau de LD existant au sein de l'haplotype et le régime de sélection, ces probabilités restent très fortement corrélées au statut IBD réel et en sont donc un bon indicateur. Cette forte corrélation est aussi et surtout due au statut IBS des marqueurs qui est l'information qui sera utilisée lors du calcul des probabilités d'IBD, que les marqueurs soient IBS et IBD ou seulement IBS.

De plus, l'augmentation de la longueur des haplotypes semble ne pas accroître la discrimination entre haplotypes IBD et non IBD par les probabilités, dès lors qu'ils sont composés de 6 marqueurs également espacés. L'utilisation d'haplotypes trop longs semble pénaliser l'estimation du statut IBD dans les populations sélectionnées puisque la corrélation entre les probabilités et le statut IBD réel diminue (résultats non présentés). L'étude de

Grapes *et al.* (2006) conclut que des haplotypes de 4 ou 6 marqueurs sont optimaux en termes de précision de la localisation du QTL lors dans le cadre de la méthode de cartographie fine proposée par Meuwissen et Goddard (2000). L'utilisation d'haplotypes composés de 6 marqueurs régulièrement espacés semble donc faire consensus. Finalement, de tels haplotypes devraient permettre d'avoir une bonne précision de résultats de cartographie fine, après avoir éventuellement procédé à une binarisation des valeurs d'IBD en fonction des probabilités : au-dessus d'un certain seuil de probabilité, on considère les QTL (ou positions testées) IBD, en-dessous, ils sont considérés non IBD.





**Partie 3.2 :**

**Robustesse des méthodes de cartographie fine**

**à la sélection**



## Introduction

Il existe un grand nombre de facteurs pour lesquels la robustesse des méthodes de cartographie fine n'a pas, ou peu, été évaluée :

- les événements relevant de l'histoire de la population, tels que les mutations, les goulets d'étranglement ou les processus de sélection,
- et les événements relevant de l'échantillonnage des données, telles que la présence de données manquantes.

Or la sélection des populations influence la structure du LD entre les locus (partie 2.2), elle devrait donc affecter l'ensemble des méthodes de cartographie fine utilisant cette information. L'étude précédente (partie 3.1) a par exemple confirmé que l'estimation des probabilités d'IBD, proposées par Meuwissen et Goddard (2001), est sensible à la sélection. L'impact de la modification de la distribution des probabilités d'IBD sous l'effet de la sélection n'a cependant pas été évaluée pour la cartographie jusqu'à présent. Pour autant, certains QTL ont pu être cartographiés avec des intervalles de confiance réduits grâce à ces méthodes dans des populations d'animaux de rente (taux de matière grasse du lait (Farnir *et al.* 2002, Meuwissen et Goddard, 2004), taux de gémellité (Meuwissen *et al.*, 2002), ou adiposité intramusculaire (Uleberg *et al.*, 2005)).

L'impact de la sélection sur la précision de cartographie des QTL à l'aide de méthodologies utilisant le LD est étudié dans cette partie, sur la base de données simulées. Les relations entre position détectée pour la cartographie de QTL et position du locus en LD maximum avec le QTL sont analysées. On considère ici que la région a été préalablement identifiée comme porteuse d'un QTL, la question de la puissance des méthodes n'est donc pas abordée. Cinq approches de cartographie sont choisies, qui correspondent à des exploitations d'information moléculaire apportée par un marqueur ou des haplotypes de deux ou quatre marqueurs. Si ces méthodes supposent des origines différentes pour le LD observé, elles ont toutes été développées pour des populations à l'équilibre de Hardy-Weinberg. Cette hypothèse n'est cependant pas respectée dès lors qu'une pression de sélection oriente les accouplements entre individus.

## I. Matériel et méthodes

### I.A. Populations simulées

L'histoire des populations est divisée en deux parties distinctes (cf. partie 2.1). Les 100 premières générations servent à établir le LD à partir d'une génération fondatrice supposée en équilibre de liaison. Le pedigree de ces premières générations est inconnu lors de la détection de QTL. Ces générations historiques sont composées de 200 individus, les accouplements sont aléatoires, avec éventuellement de l'autofécondation. Au cours de ces 100 générations, trois scénarios de sélection sont envisagés :

- les populations ne sont jamais soumises à la sélection. Ce scénario est dénoté NS (Non Sélection) par la suite,
- les populations sont sélectionnées pendant les 20 dernières générations avec une pression de sélection faible, les 80% des individus ayant les meilleurs phénotypes ( $q=0,8$ ) pouvant être choisis comme parents de la génération suivante (scénario Sélection Simple, noté SS ),
- les populations sont soumises à une sélection faible pendant les générations 81 à 90 ( $q=0,8$ ), puis à une sélection plus forte pendant les générations 91 à 100, où les 50% des individus ayant les meilleurs phénotypes ( $q=0,5$ ) peuvent être choisis comme parents de la génération suivante. Ce scénario est noté DS (Double Sélection).

Le modèle et les paramètres des simulations des phénotypes sont les mêmes que dans les études précédentes (parties 2.2 et 3.1) avec un QTL expliquant 17% de la variance génétique initiale. Le phénotype est simulé comme la somme de l'effet du QTL, d'un effet polygénique et d'un effet d'environnement.

La partie connue du pedigree pour la détection de QTL est constituée de familles, définies sur deux générations par un parent commun et ses descendants. Les haplotypes des pères et de leurs descendants sont supposés connus avec certitude, ainsi que les phénotypes des descendants. Ceci correspond à un échantillonnage de familles de père, de grande taille et où les descendants seraient connus sur performance propre ou grâce aux performances relevées sur une descendance nombreuse (au cas d'un dispositif « petites-filles » (Weller *et al.*, 1990), figure 4 de la partie 1.2.). Quinze familles de 100 demi-frères/sœurs sont simulées pour les détections, soit une population de 15 pères + 1500 descendants génotypés, selon le schéma présenté en partie 2.1, figure 2 « dispositif GPM ».

### **I.B. Cartes génétiques simulées**

Deux types de cartes génétiques ont été simulés. Une première (carte lâche) est composée de 20 marqueurs couvrant une région chromosomique de 19 cM, l'intervalle entre marqueurs étant de 1 cM. Dans la seconde (carte dense), 70 marqueurs couvrent une région de 6,9 cM, soit 0,1 cM entre les marqueurs. Pour chacune des cartes, le QTL simulé est au milieu de la région chromosomique simulée, soit à 9,5 cM et 3,45 cM du premier marqueur de la carte lâche et de la carte dense respectivement.

Les marqueurs de la première carte ont 5 allèles (type microsatellites) dans la majorité des simulations mais certains dispositifs sans sélection ont été simulés avec des marqueurs bi-alléliques (type SNP). La carte la plus dense est composée uniquement de marqueurs de type SNP. Dans tous les cas, les allèles sont équiréquents dans la génération fondatrice. Pour le QTL, 5 allèles sont simulés de la même façon, un seul affectant le phénotype.

### **I.C. Méthodes de cartographie fine**

Les 5 méthodes de cartographie retenues comparent toutes l'hypothèse alternative H1 « un QTL est en ségrégation à la position testée » contre l'hypothèse nulle H0 « il n'y a pas de ségrégation de QTL à la position ».

Régression sur les marqueurs (méthode dénommée « Régression » par la suite): La valeur phénotypique de l'individu  $i$  ( $y_i$ ) est régressée sur le nombre de copies  $x_i$  qu'il possède de l'allèle  $k$  du marqueur M selon le modèle  $y_i = x_0 + \sum_k b_k x_{ik} + e_i$ , avec  $x_0$  une constante,  $b_k$

le coefficient de régression de l'allèle  $k$  au marqueur M, et  $e_i$  l'erreur résiduelle associée à l'individu  $i$ . La statistique F pour tester l'association du marqueur M avec un QTL est obtenue en comparant le modèle précédent au modèle sous H0  $y_i = x_0 + e_i$ . La comparaison des p-values (probabilités que la valeur de F soit observée en l'absence d'association entre le marqueur et un QTL (Weisberg 1985)) permet de déterminer la position du QTL, le marqueur ayant la plus petite p-value étant retenu comme position putative du QTL.

Méthode de Terwilliger (1995), adaptée à la détection de QTL (Abdallah et al. 2004) (méthode dénommée « Terwilliger » par la suite): Cette méthode suppose l'existence d'une mutation unique  $t$  générations avant la population observée. Le QTL est par conséquent bi-allélique, avec l'allèle favorable Q ayant une fréquence actuelle  $p_Q$ . La vraisemblance que l'allèle  $k$  au marqueur ait été initialement associé à l'allèle Q au QTL vaut

$$L_k = \prod_{i=1}^N \left( \sum_{l=1}^3 P_{ki}(g_l) \phi(y_i; \mu_l, \sigma^2) \right), \text{ avec } N \text{ le nombre d'individus phénotypés, } P_{ki}(g_l) \text{ la}$$

probabilité pour l'individu  $i$  d'avoir le génotype  $g_l$  ( $l=1$  correspond au génotype QQ au QTL,  $l=2$  à Qq et  $l=3$  à qq) conditionnellement à l'association de l'allèle marqueur  $k$  avec le QTL,  $\phi(\cdot)$  la fonction de densité d'une distribution normale,  $\sigma^2$  la variance des phénotypes intra-génotype (supposée commune aux trois génotypes) et  $\mu_l$  la moyenne phénotypique pour le génotype  $g_l$ . Ces moyennes peuvent être décomposées comme indiqué par Falconer (1996) comme :  $\mu_1 = \mu + a$ ,  $\mu_2 = \mu + d$  et  $\mu_3 = \mu - a$ , avec  $\mu$  la moyenne des génotypes homozygotes au QTL,  $a$  l'effet de substitution et  $d$  l'effet de dominance. Les  $P_{ki}(g_l)$  sont calculées sous l'hypothèse d'équilibre de Hardy-Weinberg, les probabilités des allèles au QTL conditionnellement à l'allèle au QTL étant obtenues selon la formule de Terwilliger (1995). Elles sont exprimées avec un paramètre  $\lambda$  qui représente l'excès d'association de l'allèle marqueur  $k$  dans les chromosomes porteurs de l'allèle Q. L'allèle marqueur originellement associé au QTL est inconnu. La vraisemblance qu'un marqueur montre une association préférentielle avec le QTL est obtenue en sommant les vraisemblances  $L_k$  pour tous ses allèles, pondérées par la fréquence de chaque allèle dans la population observée. Cette vraisemblance est maximisée pour les paramètres  $\lambda$ ,  $p_Q$ ,  $\mu$ ,  $a$ ,  $d$  et  $\sigma^2$ . Le LRT est calculé comme  $LRT = -2 \ln \left( \frac{\max L(H_0)}{\max L(H_1)} \right)$  avec  $L(H_0)$  la vraisemblance évaluée sous l'hypothèse nulle de l'absence de LD, où  $\lambda=0$ , et  $L(H_1)$  celle sous l'hypothèse alternative, où  $\lambda>0$ .

L'association d'un marqueur unique avec le QTL a été étendue aux haplotypes de marqueurs. On a alors une vraisemblance approchée obtenue en multipliant les vraisemblances des marqueurs composant l'haplotype. Le LRT est calculé comme pour un marqueur unique et la position du maximum du LRT est retenue comme position putative du QTL. Nous avons utilisé dans cette étude des haplotypes de deux marqueurs, supposés permettre la localisation la plus précise par cette méthode (Abdallah *et al.* 2004).

Méthode HapIM (Boitard *et al.*, 2006) : Cette méthode suppose l'apparition d'une mutation unique créant un déséquilibre complet entre l'haplotype fondateur et la mutation. Cette mutation se serait produite  $t$  générations avant la population observée. Comme dans la méthode précédente dont elle dérive, les phénotypes sont modélisés comme un mélange de

lois normales dont les proportions dépendent des moyennes des lois normales de la vraisemblance et de l'information. Ici, l'information moléculaire est apportée par les deux marqueurs flanquant la position étudiée : les fréquences des haplotypes à deux marqueurs sont utilisées pour calculer la probabilité des génotypes au QTL conditionnellement aux deux haplotypes de chaque individu. Ces fréquences, qui évoluent de manière stochastique au cours des générations, sont approximées par leur espérance. La vraisemblance de chaque position putative est alors calculée conditionnellement aux phénotypes et aux informations génotypiques comme précédemment. Chaque position est testée avec un rapport de vraisemblances de la même manière que la méthode de Terwilliger. La position maximisant cet test est considérée comme la position putative du QTL.

Méthode de Meuwissen et Goddard (2000, 2001) (dénotée « LDLA » par la suite): Comme décrit en I.2, cette méthode repose sur l'estimation des composantes de variance des effets de QTL par REML avec un modèle linéaire mixte. Pour chaque QTL, le modèle comprend deux effets gamétiques par individu, la corrélation entre les  $2N$  effets étant égale à leur probabilité IBD. La probabilité IBD est calculée :

- entre les QTL fondateurs par la méthode de Meuwissen et Goddard (2001),
- entre les QTL des animaux du pedigree connu en fonction des probabilités de transmission des haplotypes conditionnellement à l'information marqueur et l'apparentement des fondateurs.

Cette méthode suppose que la population initiale était en équilibre de liaison. La méthodologie mise en place dans notre étude est similaire à celle décrite par Kim et Georges (2002) et Blott *et al.* (2003). Nous avons retenu la recommandation de Grapes *et al.* (2006) d'utiliser des haplotypes de 4 marqueurs pour obtenir la cartographie la plus précise.

Un modèle linéaire mixte est appliqué pour estimer les effets haplotypiques :  $\mathbf{y} = \mathbf{1}\boldsymbol{\mu} + \mathbf{Z}\mathbf{h} + \mathbf{W}\mathbf{u} + \mathbf{e}$ , avec  $\mathbf{y}$  le vecteur des phénotypes,  $\boldsymbol{\mu}$  une constante,  $\mathbf{h}$  le vecteur des effets des haplotypes,  $\mathbf{u}$  le vecteur des effets polygéniques, et  $\mathbf{W}$  et  $\mathbf{Z}$  des matrices d'incidence connues. Elles correspondent en effet à la matrice des variances-covariances des effets haplotypiques, avec  $\mathbf{Var}(\mathbf{h}) = \sigma_h^2 \mathbf{H}_p$  où  $\sigma_h^2$  est la variance haplotypique et  $\mathbf{H}_p$  est la matrice de probabilité IBD des haplotypes. Ce modèle est comparé au modèle  $\mathbf{y} = \mathbf{1}\boldsymbol{\mu} + \mathbf{W}\mathbf{u} + \mathbf{e}$  (sous l'hypothèse nulle d'absence d'effet des haplotypes). La présence d'un QTL est évaluée par un test de rapport de maximum de vraisemblances. La maximisation des



vraisemblances pour obtenir les estimations est effectuée avec un algorithme AI-REML (Jensen *et al.*, 1996).

Clustering des probabilités d'IBD (méthode dénotée « cluster » par la suite) : Pour éviter des problèmes calculatoires liés à des matrices d'effets haplotypiques non définies positives, il paraît intéressant d'essayer de regrouper les haplotypes selon leurs probabilités d'IBD. D'après la partie 3.1, il est possible de regrouper les haplotypes effectivement IBD sur la base de probabilités IBD élevées en contrôlant le risque de mauvais regroupement. Nous proposons ici de regrouper des haplotypes IBD selon une méthode de clustering qui ne requiert pas de valeur seuil pour le regroupement. C'est le cas des algorithmes de clustering flou comme celui décrit par Kaufman et Rousseeuw en 1990. Cet algorithme repose sur des calculs de dissimilarités obtenues pour chaque paire d'haplotypes comme un moins la probabilité qu'ils soient IBD à la position centrale (probabilité d'IBD de Meuwissen et Goddard, 2001). Il permet de déterminer la probabilité d'un objet – ici un groupe d'haplotypes IBS- d'appartenir à un cluster donné, la somme des probabilités sur tous les clusters pour un individu étant égale à 1. Cet algorithme est robuste aux tailles de cluster variables, ce qui permet d'absorber la structuration des haplotypes de la population. Du fait de la sélection, le nombre final de groupes d'haplotypes IBS porteurs de l'allèle favorable du QTL est souvent faible, mais des variants issus de recombinaisons peuvent persister dans la population. Ces variants n'appartiennent alors à aucun cluster majeur et restent isolés. L'algorithme de Kaufman et Rousseeuw qui a été retenu, en minimisant les sommes de dissimilarités, est plus robuste aux objets isolés que les méthodes de regroupement flou basées sur les sommes des carrés (Trauwaert 1987). La fonction objectif à minimiser s'écrit

$$C = \sum_{v=1}^k \frac{\sum_{i=1}^n \sum_{j=1}^n u_{iv}^2 u_{jv}^2 d(i, j)}{2 \sum_{j=1}^n u_{jv}^2} \text{ avec } u_{iv} \text{ la probabilité d'appartenance de l'haplotype } i \text{ au cluster } v,$$

$d(i, j)$  la dissimilarité entre  $i$  et  $j$ ,  $k$  le nombre de clusters à former, et  $n$  le nombre d'haplotypes. Le fait d'élever au carré les probabilités d'appartenance des objets au cluster est un compromis entre l'obtention d'une plus grande variabilité de ces probabilités et la rapidité de convergence de l'algorithme. L'algorithme procède en effectuant une répartition initiale aléatoire des objets au sein des clusters puis en itérant jusqu'à convergence.

La cohérence des résultats de clustering obtenus peut être évaluée pour chaque objet par un calcul de silhouette. La silhouette d'un objet (Rousseeuw, 1987) mesure la qualité de son

intégration au cluster auquel il appartient. L'ensemble de ces silhouettes peut être représenté sur un graphique, ce qui permet d'évaluer quels clusters sont cohérents et lesquels le sont moins. Pour chaque objet  $i$ , les silhouettes sont obtenues de la manière suivante :

- calculer la dissimilarité moyenne de l'objet  $o$  avec tous les autres objets appartenant au même cluster que lui (noté  $A$ ). Cette dissimilarité prend la valeur  $a(o)$  ;
- calculer pour chacun des autres clusters la dissimilarité moyenne entre l'objet  $o$  et les objets du cluster. Retenir le cluster pour lequel cette dissimilarité moyenne est minimale : ce cluster (noté  $C$ ) est appelé « cluster voisin de l'objet  $o$  » et la dissimilarité correspondante est notée  $b(o)$  ;
- calculer la silhouette comme :  $s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}$ .

La silhouette de chaque objet est comprise entre -1 et 1. Ainsi, lorsque  $s$  tend vers 1, cela signifie que la dissimilarité moyenne entre  $i$  et les autres objets de  $A$  est nettement plus faible que celle entre  $o$  et les objets composant  $C$  : l'objet  $o$  est assigné correctement. Au contraire, si  $s$  est proche de -1, la dissimilarité moyenne entre  $o$  et les autres objets de  $A$  est nettement supérieure à celle entre  $o$  et les objets composant  $C$  et l'objet  $o$  est mal assigné.

Une fois la silhouette de chaque objet calculée, la cohérence du clustering est évaluée par la silhouette moyenne de chaque cluster et la silhouette du jeu de données entier (notée SC). L'algorithme étant répété pour chaque nombre de clusters demandé, le regroupement est optimal pour le nombre de clusters avec lequel la silhouette du jeu de données complet est maximisée. Une évaluation de la qualité du clustering général peut être celle présentée dans le tableau 3.2.I (Kaufman et Rousseeuw, 1990).

SC	Interprétation possible
> 0,70	Une forte structure a été trouvée.
0,51 – 0,70	Une structure a été mise en évidence.
0,26 – 0,50	Il y a une faible structure du jeu de données ; celle-ci pourrait être un artéfact.
< 0,26	Aucune structure n'a été trouvée.

**Tableau 3.2.1 : Interprétation possible des silhouettes obtenues sur des jeux de données complets.**

Le clustering des probabilités d'IBD était très fortement marqué par les groupes d'haplotypes ayant des probabilités d'IBD très élevées (>0,9), avec des silhouettes supérieures à 0,5. Par

contre, des groupes de probabilités d'IBD plus faibles mais cohérents restaient isolés en première approche. En excluant des données les groupes IBS assignés à un cluster de manière forte, une nouvelle étape de clustering sur les données restantes permet de rassembler les structures plus faibles avec des silhouettes de cluster fortes. Il a donc été décidé d'itérer sur les données en utilisant le seuil de 0,51 pour assigner ou non un haplotype à un cluster. Lorsque plus aucun haplotype n'avait de silhouette supérieure à 0,51, tous les haplotypes restants étaient mis dans un même cluster. Ce clustering a été réalisé grâce au logiciel de Peter Rousseeuw qui est librement disponible sur Internet.

Une fois ce clustering réalisé, une régression a été appliquée sur les clusters obtenus selon le modèle  $y_o = x_0 + \sum_k b_k x_{ok} + e_o$ , avec  $x_0$  une constante,  $b_k$  le coefficient de régression du cluster  $o$  au locus M (correspondant au milieu de l'haplotype), et  $e_o$  l'erreur résiduelle associée à l'individu  $o$ . La statistique F pour tester l'association du cluster d'haplotypes C avec un QTL est obtenue en comparant le modèle précédent au modèle sous H0  $y_o = x_0 + e_o$ . La comparaison des p-values permet de déterminer la position du QTL : le marqueur ayant la plus petite p-value est retenu comme position putative du QTL.

Pour les méthodes supposant que les individus phénotypés ne sont pas apparentés (toutes les méthodes présentées, sauf celle de Meuwissen et Goddard), les phénotypes ont été corrigés pour l'effet père avant de procéder à la cartographie.

#### **I.D. Evaluation de la précision de cartographie des méthodes**

Pour chaque dispositif, défini par la combinaison densité de la carte, information moléculaire et sélection, 1000 simulations ont été réalisées. Pour nous rapprocher des situations réelles des populations expérimentales où sont effectivement détectés des QTL, nous n'avons conservé que les simulations où au moins 4 grands-pères étaient hétérozygotes, soient un peu plus d'un tiers des familles simulées.

Nous avons testé la précision de cartographie selon deux critères se rapportant à la position estimée pour le QTL par chaque approche :

- le biais, c'est-à-dire l'écart moyen à la position réelle du QTL. Un test de la différence entre les moyennes des positions estimées avec chacune des méthodes et la position

réelle du QTL a été réalisé, pour chaque scénario, avec un seuil pour l'erreur de 1<sup>ère</sup> espèce  $\alpha=5\%$ .

- La variance de la position estimée selon les différentes méthodes mais aussi selon la carte génétique et le scénario évolutif des populations. Ces variances ont été comparées avec un test F, avec un seuil pour l'erreur de 1<sup>ère</sup> espèce  $\alpha=5\%$  (a) pour chaque paire de méthodes ou (b) pour une méthode et une carte génétique données, entre les différents scénarios évolutifs des populations.

### **I.E. Evaluation de la concordance entre LD maximum et position putative du QTL**

Selon la démarche présentée dans la partie 2.2, la position des locus marqueurs (ou des haplotypes de 2 ou 4 marqueurs) en LD maximum avec le QTL a été enregistrée, la position des haplotypes (H2 correspondant aux haplotypes de deux marqueurs et H4 à ceux de 4 marqueurs) étant mesurée en les assimilant à un locus ponctuel situé en leur centre. Comme dans la partie 2.2, le LD a été évalué par le  $D'$  (Lewontin 1964) et le  $\chi^2$  (Yamazaki 1977). Deux descripteurs de la concordance entre la localisation putative du QTL et le locus en LD maximum avec le QTL ont été calculés :

- Le pourcentage de positions estimées pour le QTL qui co-localisent exactement avec le locus en plus fort LD avec le QTL, identifié par le  $D'$  ou le  $\chi^2$ ;
- La moyenne de la distance entre les deux locus et son écart-type, notée « ecartd » pour le  $D'$  et « écartk » pour le  $\chi^2$ .

Pour ces comparaisons, la position en LD maximal avec le QTL est donc comparée à la position estimée du QTL pour tous les niveaux d'information marqueur, même s'ils ne correspondent pas à l'information utilisée pour la cartographie.

## **II. Résultats**

### **II.A. Comparaison des méthodes**

Il s'agit ici de présenter les résultats obtenus avec chaque méthode pour un dispositif (densité de la carte – information moléculaire – sélection) donné.

Les résultats pour la carte lâche sont présentés dans le Tableau 3.2.2, et ceux de la carte dense sont dans le tableau 3.2.3. Pour le dispositif DS, aussi bien sur la carte lâche que sur la carte dense, le QTL n'est resté polymorphe que dans environ  $\frac{1}{4}$  des simulations. Les résultats sont donc fournis à titre indicatif et fournissent seulement une tendance.

Dispositif (Nombre de simulations)		Méthode				
		Régression	Terwilliger	HapIM	LDLA	Cluster
NS (809)	biais (cM)	-0,05	0,09	0,08	0,05	-0,01
	$\sigma_{\text{pos}}$ (cM)	1,86 <sup>a</sup>	2,16 <sup>b</sup>	2,04 <sup>b</sup>	2,65 <sup>c</sup>	2,09 <sup>b</sup>
SS (750)	biais (cM)	-0,08	0,04	-0,08	0,04	0,06
	$\sigma_{\text{pos}}$ (cM)	1,78 <sup>a</sup>	2,30 <sup>c</sup>	2,13 <sup>b</sup>	2,02 <sup>b</sup>	2,68 <sup>d</sup>
DS (248)	biais (cM)	0,01	0,23	0,35	<b>0,78</b>	<b>0,61</b>
	$\sigma_{\text{pos}}$ (cM)	3,25 <sup>a</sup>	4,33 <sup>b</sup>	4,98 <sup>c</sup>	4,05 <sup>b</sup>	4,11 <sup>b</sup>
NS – SNP (809)	biais (cM)	-0,10	0,10	-0,01	0,05	-0,07
	$\sigma_{\text{pos}}$ (cM)	3,04 <sup>bc</sup>	3,16 <sup>c</sup>	2,86 <sup>b</sup>	2,39 <sup>a</sup>	3,00 <sup>bc</sup>

**Tableau 3.2.2 : Biais et écarts-types ( $\sigma_{\text{pos}}$ ) de la localisation estimée pour le QTL sur la carte lâche, marqueurs à 5 allèles (sauf NS – SNP où les marqueurs sont bi-alléliques). La position vraie du QTL est 9.5 cM. <sup>a, b, c, d</sup> : comparaison des variances : une même lettre indique des variances égales, la variance associée à «<sup>a</sup>» est significativement inférieure à celle associée à «<sup>b</sup>» etc. En gras : biais significativement différent de 0.**

Sur la carte lâche, aucun biais systématique n'a pu être mis en évidence pour la localisation du QTL, sauf pour les méthodes LDLA et Cluster dans le scénario DS (valeurs en gras dans le tableau 3.2.2). Par contre, des différences significatives entre les variances des positions estimées par les différentes méthodes ont été mises en évidence. Ainsi, dans les trois dispositifs utilisant des marqueurs multi-alléliques, la régression permet d'obtenir le plus faible écart-type d'estimation, que la population soit soumise à la sélection ou non. La méthode LDLA fournit la position moyenne la plus éloignée de la position du QTL en l'absence de sélection, et la variance est supérieure à celles obtenues avec les quatre autres méthodes. En revanche, si la population est soumise à sélection, la variance de la localisation devient inférieure ou équivalente à celle des autres méthodes sauf la régression.

Lorsque des marqueurs bi-alléliques sont utilisés avec un marqueur tous les centiMorgans, la méthode LDLA présente en revanche le plus faible écart-type, Terwilliger étant la méthode avec la plus grande variabilité dans la localisation du QTL.

Par ailleurs, toutes les méthodes sont sensibles au type d'information moléculaire apportée (NS vs. NS – SNP), la réduction du nombre d'allèles aux marqueurs entraînant une diminution de la précision, sauf pour la méthode LDLA (Tableau 3.2.2).

Dispositif (Nombre de simulations)		Méthode				
		Régression	Terwilliger	HapIM	LDLA	Cluster
NS (810)	biais (cM)	-0,02	0,04	0,01	-0,02	-0,02
	$\sigma_{\text{pos}}$ (cM)	0,82 <sup>c</sup>	0,93 <sup>d</sup>	0,66 <sup>b</sup>	0,44 <sup>a</sup>	0,96 <sup>d</sup>
SS (536)	biais (cM)	-0,01	0,01	-0,02	0,01	0,06
	$\sigma_{\text{pos}}$ (cM)	0,75 <sup>c</sup>	0,92 <sup>d</sup>	0,63 <sup>b</sup>	0,37 <sup>a</sup>	0,96 <sup>d</sup>
DS (211)	biais (cM)	0,10	0,02	0,10	<b>0,11</b>	<b>0,26</b>
	$\sigma_{\text{pos}}$ (cM)	0,79 <sup>b</sup>	0,84 <sup>b</sup>	0,79 <sup>b</sup>	0,48 <sup>a</sup>	1,07 <sup>c</sup>

**Tableau 3.2.3 : Biais et écarts-types ( $\sigma_{\text{pos}}$ ) de la localisation estimée pour le QTL sur la carte dense, marqueurs à 2 allèles. La position vraie du QTL est 3,45 cM. <sup>a, b, c, d</sup> : comparaison des variances : une même lettre indique des variances égales, la variance associée à «<sup>a</sup>» est significativement inférieure à celle associée à «<sup>b</sup>» etc. En gras : biais significativement différent de 0.**

Sur la carte dense, aucun biais n'est détecté pour la localisation moyenne du QTL, sauf, comme sur la carte lâche, pour le dispositif avec une double sélection avec les méthodes LDLA et Cluster (valeurs en gras dans le tableau 3.2.3). Compte tenu du faible nombre de simulations conservées, ces résultats sont à confirmer.

La méthode LDLA présente les variances d'estimation des positions les plus faibles pour tous les dispositifs avec cette forte densité de marqueurs, la méthode de clustering étant celle caractérisée par les plus fortes, de même que la méthode de Terwilliger (sauf pour le dispositif avec deux pressions de sélection).

## II.B. Sensibilité des méthodes à la sélection

Les variances des positions estimées ont été comparées afin d'évaluer la sensibilité des méthodes à la sélection, en prenant la population non sélectionnée comme référence. Les résultats sont résumés dans le tableau 3.2.4.

		Méthode				
		Régression	Terwilliger	HapIM	LDLA	Cluster
Carte lâche	NS	A	A	A	A	A
	SS	A	A	A	B	B
	DS	B	B	B	C	C
	NS - SNP	B	C	C	D	D
Carte dense	NS	A	A	A	A	A
	SS	B	A	A	B	A
	DS	AB	A	B	A	A

**Tableau 3.2.4 : Comparaison des variances des positions estimées pour chaque méthode en fonction des pressions de sélection. Des lettres identiques correspondent à des variances équivalentes.**

Sur la carte lâche, les méthodes LDLA et Cluster sont sensibles à la sélection, même avec une faible pression et peu de générations concernées. Par contre, toutes les méthodes perdent en précision dès lors qu'une plus forte sélection est appliquée (scénario DS).

Avec une forte densité de marqueurs, la méthode Cluster devient insensible à la sélection, alors que la régression le devient. La méthode de Terwilliger a également une variance inchangée malgré la sélection, HapIM n'étant affectée que par une forte pression de sélection. La régression et la méthode LDLA semblent avoir des variances équivalentes lorsqu'elles sont appliquées à des populations non sélectionnées ou soumises à une forte pression de sélection dans ce contexte de carte dense.

### **II.C. Concordance LD – cartographie du QTL**

La concordance entre la localisation estimée du QTL et celle de la position en LD maximum avec le QTL est présentée dans le tableau 3.2.5, avec le LD mesuré par le  $\chi^2$ . Le  $D'$  et le  $\chi^2$  s'accordent généralement sur la méthode ayant le plus fort ou le plus faible taux de concordance entre les deux positions (résultats non présentés).

Information moléculaire	Méthode	Carte lâche				Carte dense		
		NS	SS	DS	NS - SNP	NS	SS	DS
Marqueur	Régression	41,1	30,3	22,7	16,4	4,3	2,6	<b>4,3</b>
	HapIM	46,5	41,0	22,7	21,9	4,0	3,4	0,5
	Terwilliger	<b>52,4</b>	45,5	27,6	20,8	3,5	2,4	1,9
	LDLA	46,7	40,9	24,0	<b>29,0</b>	<b>7,1</b>	<b>3,5</b>	1,9
	Cluster	49,5	<b>47,8</b>	<b>31,4</b>	23,4	6,0	3,3	2,5
H2	Régression	18,8	20,5	11,3	11,6	4,7	4,7	5,2
	HapIM	26,8	25,3	13,8	16,4	7,0	7,9	5,7
	Terwilliger	30,4	28,5	17,7	17,2	7,0	6,2	4,7
	LDLA	<b>46,4</b>	28,9	18,6	<b>26,9</b>	<b>11,0</b>	<b>8,9</b>	<b>6,5</b>
	Cluster	38,5	<b>33,4</b>	<b>22,1</b>	25,5	7,6	5,3	3,7
H4	Régression	<b>27,6</b>	<b>26,2</b>	<b>19,7</b>	<b>18,9</b>	6,4	<b>7,8</b>	5,2
	HapIM	19,5	18,9	8,9	15,7	8,0	6,4	6,6
	Terwilliger	16,2	16,8	11,3	13,7	4,9	6,2	4,7
	LDLA	14,8	13,5	6,9	12,7	<b>10,8</b>	6,4	<b>7,0</b>
	Cluster	10,4	11,2	6,9	10,7	4,6	4,9	4,9

**Tableau 3.2.5 : Proportion de concordance (en %) entre la position du locus en LD maximal avec le QTL (mesuré par le  $\chi^2$ ) et la position estimée du QTL. En jaune : information moléculaire utilisée par la méthode. En rouge (respectivement gras) : plus faible (respectivement fort) taux de concordance pour une combinaison scénario évolutif – mesure de LD.**

Sur la carte lâche, en règle générale, les taux de concordance diminuent avec l'augmentation du nombre de marqueurs pris en compte pour le calcul du LD, ainsi qu'avec une pression de sélection croissante. Ils sont généralement compris entre 30 et 45% pour des LD calculés avec un marqueur unique ou un haplotype de 2 marqueurs dans des populations non sélectionnées, mais ils sont plus faibles (10 à 20%) si des haplotypes de 4 marqueurs sont utilisés pour évaluer le LD dans des populations fortement sélectionnées. La régression est la méthode pour laquelle la concordance est généralement la plus faible si l'information moléculaire est apportée par un ou deux marqueurs. Par contre, elle est la plus performante si l'information sur le LD est apportée par un haplotype de 4 marqueurs, bien que la proportion de concordance soit plus réduite qu'avec un LD mesuré entre un marqueur et le QTL. Les méthodes pour lesquelles le plus de concordances sont enregistrées lorsque l'information est apportée par un ou deux marqueurs sont les méthodes LDLA et Cluster. Par contre, cette dernière est celle qui présente les résultats les plus faibles si le LD est calculé entre le QTL et des haplotypes de 4 marqueurs.



Sur la carte dense, les taux de concordance relevés sont très faibles (rarement supérieurs à 10%). Les taux de concordance diminuent néanmoins avec une pression de sélection croissante. Pour les trois scénarios évolutifs, la méthode LDLA est celle qui indique le plus fréquemment la localisation putative du QTL à l'emplacement du locus en LD maximum avec le QTL, surtout si le LD est évalué entre un marqueur seul et le QTL et est estimé avec le  $\chi^2$ . Aucune méthode ne se démarque comme étant celle présentant le moins de concordance entre les positions des deux locus.

La distance moyenne entre le QTL putatif et le locus en LD maximum avec le QTL estimé avec le  $\chi^2$ , ainsi que l'écart-type de cette distance, sont présentés dans le Tableau 3.2.6. Cette distance augmente avec la pression de sélection mais diminue avec la densité de la carte, de même que les écarts-types. Avec la carte lâche, l'écart moyen entre les deux locus dans une population non sélectionnée est de l'ordre de une à deux fois la distance entre deux marqueurs, mais il devient fréquemment supérieur à trois fois cette distance dans les populations sélectionnées. Avec une carte dense, dans des populations non sélectionnées, cet écart est de l'ordre de 10 intervalles entre marqueurs si le LD est évalué entre un marqueur et le QTL, et de 6 à 7 intervalles si le LD est évalué entre le QTL et un haplotype de 4 marqueurs.

Deux méthodes semblent minimiser cette distance et ses variations : la régression, quand il s'agit de la carte lâche avec des marqueurs multi-alléliques, et la méthode LDLA si des marqueurs bi-alléliques sont utilisés, et ce indépendamment de la sélection et de la densité en marqueurs (NS-SNP). En revanche, sur la carte lâche, les méthodes pour lesquelles la distance entre les locus est maximale varient selon l'intensité de la sélection. HapIM apparaît comme la méthode pour laquelle la distance entre les deux locus, ainsi que son écart-type, est maximale dans le schéma avec la plus forte pression de sélection. Sur la carte dense, la régression est la méthode qui s'avère la moins précise pour tous les scénarios évolutifs.

Lorsque des SNP sont utilisées, pour chaque méthode, les écarts les plus faibles entre le locus en LD maximum avec le QTL et la position estimée du QTL, ainsi que les écarts-types les plus faibles, sont observés lorsque l'information pour le calcul du LD provient d'haplotypes de 4 marqueurs. Cette tendance est aussi observable avec le scénario DS sur la carte lâche. Les distances et écarts types les plus faibles sur la carte lâche sont en général trouvés avec les informations marqueurs qui sont utilisées par la méthode pour la cartographie, sauf pour la régression sur un marqueur unique.

Les résultats obtenus avec le D' sont très similaires (résultats non présentés).

Troisième partie : Influence de la sélection sur les méthodes de cartographie fine  
Partie 3.2. Robustesse des méthodes de cartographie fine à la sélection

Information moléculaire	Méthode de cartographie	Carte lâche								Carte dense					
		NS		SS		DS		NS - SNP		NS		SS		DS	
		$\mu$ (cM)	$\sigma$ (cM)	$\mu$ (cM)	$\sigma$ (cM)	$\mu$ (cM)	$\sigma$ (cM)	$\mu$ (cM)	$\sigma$ (cM)	$\mu$ (cM)	$\sigma$ (cM)	$\mu$ (cM)	$\sigma$ (cM)	$\mu$ (cM)	$\sigma$ (cM)
Marqueur	Régression	<b>1,15</b>	<b>1,63</b>	<b>1,54</b>	<b>1,95</b>	<b>2,88</b>	<b>3,09</b>	3,31	3,13	1,11	1,11	1,38	1,21	1,77	<b>1,21</b>
	HapIM	1,41	1,81	1,65	1,96	4,31	3,60	3,24	3,12	1,02	1,07	1,34	1,23	1,75	1,28
	Terwilliger	1,51	1,90	1,74	2,05	3,60	3,10	3,40	3,10	1,14	1,11	1,42	1,24	1,86	1,34
	LDLA	1,56	1,65	2,12	2,23	3,54	3,16	<b>3,04</b>	<b>2,86</b>	<b>0,91</b>	<b>1,04</b>	<b>1,27</b>	<b>1,17</b>	<b>1,72</b>	<b>1,21</b>
	Cluster	1,90	2,19	1,78	1,97	3,44	3,36	3,65	3,25	1,25	1,21	1,70	1,49	1,99	1,56
H2	Régression	<b>1,34</b>	<b>1,65</b>	<b>1,50</b>	<b>1,81</b>	<b>2,83</b>	<b>3,05</b>	2,93	2,71	0,82	0,77	0,79	0,77	0,75	0,75
	HapIM	1,46	1,74	1,68	1,95	4,19	3,51	2,82	2,61	0,71	0,71	0,71	0,69	0,78	0,80
	Terwilliger	1,57	1,87	1,82	2,07	3,70	3,12	3,00	2,77	0,87	0,84	0,89	0,85	0,81	0,78
	LDLA	1,62	1,67	2,08	2,22	3,41	3,07	<b>2,71</b>	<b>2,38</b>	<b>0,57</b>	<b>0,62</b>	<b>0,61</b>	<b>0,63</b>	<b>0,63</b>	<b>0,66</b>
	Cluster	1,52	1,74	2,14	2,32	3,46	3,24	3,06	2,74	0,87	0,86	1,09	1,04	1,16	1,13
H4	Régression	1,55	<b>1,63</b>	<b>1,58</b>	<b>1,73</b>	<b>2,82</b>	<b>2,67</b>	2,98	2,76	0,71	0,69	0,72	0,71	0,77	0,69
	HapIM	1,56	1,77	1,76	1,94	4,24	3,24	2,84	2,75	0,60	0,62	0,66	0,66	0,74	0,69
	Terwilliger	1,63	1,90	1,83	2,04	3,56	2,97	3,01	2,87	0,75	0,77	0,83	0,80	0,81	0,72
	LDLA	<b>1,49</b>	1,82	2,09	2,23	3,22	2,90	<b>2,58</b>	<b>2,52</b>	<b>0,45</b>	<b>0,52</b>	<b>0,51</b>	<b>0,52</b>	<b>0,54</b>	<b>0,52</b>
	Cluster	1,55	1,81	2,09	2,32	3,13	3,00	2,95	2,85	0,81	0,84	1,03	1,05	1,15	1,10

Tableau 3.2.6 : Distance moyenne ( $\mu$ ) entre le locus en LD maximum avec le QTL et la position putative du QTL et écart-type ( $\sigma$ ) de cette distance. LD évalué par le  $\chi^2$  entre un marqueur ou un haplotype de 2 ou 4 marqueurs. En jaune : information moléculaire utilisée par la méthode. En rouge (respectivement gras) : plus faible (respectivement forte) valeur de la variable statistique ( $\sigma$  ou  $\mu$ ) pour une combinaison scénario évolutif – mesure de LD.

### III. Discussion

Cette étude vise à comparer la précision de la localisation d'un QTL, estimée avec 5 méthodes de cartographie fine, sous différents scénarios évolutifs définis par des pressions de sélection différentes. Les performances de cartographie ont été reliées à la correspondance possible entre le locus ainsi identifié comme position putative du QTL et le locus en LD maximum avec le QTL. La précision physique de la localisation est en effet un paramètre très important en vue d'une approche moléculaire de type gène-candidat pour identifier le gène responsable du QTL. Cependant, certaines utilisations, comme la sélection assistée par marqueurs, peuvent se contenter d'un locus en fort déséquilibre avec le QTL, qui permette de suivre les allèles au QTL dans la population, dès lors que le déséquilibre se maintient entre générations.

Toutes les méthodes testées dans cette étude permettent de localiser en moyenne le QTL à sa position réelle. Cependant, le QTL était situé à milieu de la région chromosomique simulée ; un biais pourrait apparaître s'il était excentré. La variabilité d'estimation de la localisation est cependant différente selon les méthodes. Sur une carte avec des marqueurs multi-alléliques espacés de 1 cM, la régression apparaît être la méthode la plus précise, et ce quel que soit le régime de sélection auquel la population est soumise, ce qui est en accord avec les résultats de Zhao *et al.* (2007). Dès que des marqueurs bi-alléliques sont utilisés, la méthode LDLA devient la méthode ayant la plus grande précision de localisation. Lorsque les marqueurs présentent de nombreux allèles, ceux-ci sont plus informatifs et la régression utilise efficacement cette information. Lorsque les marqueurs n'ont que deux allèles, la régression sur un seul marqueur est moins performante. La combinaison de plus de marqueurs permet d'avoir plus d'information, ce qui peut expliquer l'avantage de la méthode LDLA qui utilise simultanément 4 marqueurs. Il est cependant à noter que la méthode de clustering proposée, qui elle aussi utilise l'information apportée par 4 marqueurs, présente une précision de localisation plus faible que la régression, même sur carte dense avec des marqueurs bi-alléliques. Ceci peut être dû à des regroupements d'haplotypes très similaires d'un point de vue IBS, mais non IBD et ayant des effets différents sur le phénotype. Dès lors, l'effet d'un cluster d'haplotypes devient difficile à identifier. Ceci justifierait aussi la perte de précision de cette méthode par rapport à la méthode LDLA. La moindre efficacité de la méthode LDLA sur la carte lâche par rapport à la régression peut être due à un apport de bruit de fond par la partie analyse de liaison de la méthode. En effet, il a été démontré dans la partie 2.2 que la

longueur moyenne du segment IBD contenant le QTL était de l'ordre de 2,5 cM. Or, sur la carte lâche, la longueur d'un haplotype de 4 marqueurs excède cette longueur. Dès lors, l'information sur le LD utilisée dans le calcul des probabilités IBD peut être brouillée par l'information de liaison. Sur la carte dense, la longueur d'un haplotype de 4 marqueurs est en revanche très nettement inférieure à la longueur du segment IBD fondateur contenant le QTL. Les résultats obtenus ne sont pas en accord avec ceux obtenus par Zhao *et al.* (2007) qui ont trouvé que, quelle que soit la densité de la carte et/ou le type de marqueurs (bi- ou multi-alléliques), la régression était la méthode la plus efficace. Plusieurs raisons peuvent être envisagées, la principale étant que nous avons ici utilisé des données issues de structures familiales. Les données exploitées par la régression sont plus redondantes que celles utilisées dans l'étude de Zhao *et al.* où les individus devaient être moins apparentés. Une autre raison pourrait être l'informativité des marqueurs de notre étude (notamment les SNP) qui, contrairement à leur étude, n'est pas contrôlée : certains des marqueurs utilisés ici peuvent avoir des fréquences alléliques très déséquilibrées, qui apportent peu d'information à la régression.

Cette étude démontre que, si la carte génétique est assez peu dense, toutes les méthodes ici utilisées sont sensibles à la pression de sélection qui fait diminuer leur précision de cartographie. Cependant, la régression, HapIM et Terwilliger ne perdent en précision que si une plus forte pression de sélection est appliquée, ce qui est le cas dans les populations d'animaux de rente. En carte dense, la cartographie semble moins sensible à la sélection, et, sous réserve du faible nombre de données obtenues pour le scénario évolutif avec une forte pression de sélection, il semblerait que certaines méthodes conservent les mêmes performances qu'avec une population non sélectionnée. Ceci pourrait être dû à la très forte réduction du nombre d'haplotypes présents dans la population : peu d'haplotypes ne sont pas porteurs de l'allèle avantageux du QTL, et parmi ceux qui présentent cet allèle peu d'haplotypes différents sont IBS. Par rapport à l'étude présentée dans la partie 3.1, il apparaît que la variation dans la distribution des probabilités d'IBD sous l'effet de la sélection influence bien la précision de la cartographie fine pour des pressions de sélection faible (différence significative pour la méthode LDLA entre les scénarios NS et SS en carte dense, ce qui correspondait à la densité de l'étude de la partie 3.1.).

Suite à ces résultats, il semblerait qu'une stratégie en deux temps permette d'avoir une localisation fine du QTL quelle qu'ait été l'histoire de la population :

- dans un premier temps affiner la localisation du QTL à l'aide d'une régression marqueur par marqueur sur une carte assez peu densément couverte mais avec des marqueurs multi-alléliques,
- puis appliquer la méthode LDLA sur une carte densément couverte par des marqueurs bi-alléliques.

Il faut noter que la population initiale était en équilibre de liaison, et que, par conséquent, l'allèle favorable du QTL était présent initialement dans plusieurs haplotypes, ce qui ne correspond pas aux hypothèses formulées dans les méthodes HapIM et Terwilliger. Ceci pourrait expliquer leur précision plus faible que la méthode LDLA, qui repose sur les hypothèses de la population simulée, à la sélection près.

La concordance entre le locus en LD maximum et la position identifiée comme celle du QTL est un indicateur de l'utilisation du LD par les méthodes de cartographie fine. La régression est la méthode qui indique ainsi le plus souvent comme QTL un marqueur unique faisant en fait partie de l'haplotype de 4 marqueurs en LD maximum avec le QTL. Mais il est plus intéressant d'avoir l'inverse, c'est-à-dire une méthode pointant finement un locus, en faisant éventuellement usage de l'information de plusieurs marqueurs. C'est le cas avec la méthode LDLA dès que l'on utilise des marqueurs bi-alléliques : celle-ci désignera comme localisation du QTL un locus unique situé en son milieu, bien qu'ayant exploité l'information apportée par 4 marqueurs simultanément.

Cependant, les fréquences de concordance exacte sont faibles, particulièrement sur des cartes génétiques denses. La distance moyenne entre le locus indiqué comme QTL et celui en LD maximum avec le vrai QTL excède généralement le centiMorgan, sauf sur une carte densément couverte si le LD est mesuré entre un haplotype et le QTL. De plus, l'utilisation de marqueurs bi-alléliques pour la régression augmente considérablement l'écart moyen entre les deux locus. L'information de LD apportée par des haplotypes de 4 marqueurs ne ramène pas cet écart au niveau de celui mesuré entre la position estimée du QTL et le marqueur multi-allélique en LD maximum avec le QTL. Il ne suffit pas d'augmenter l'information moléculaire pour l'estimation du LD maximum lorsqu'on dispose de marqueurs bi-alléliques, il faut passer à une carte dense pour réduire cet écart.

Cette différence de localisation est probablement due à la différence de nature de l'information utilisée : la mesure du LD ne s'appuie que sur l'information moléculaire (dont, dans nos calculs, l'information au QTL qui est, en réalité, une inconnue), alors que les

méthodes de cartographie mettent en relation l'information moléculaire et l'information phénotypique.

#### **IV. Conclusion**

Les méthodes de cartographie fine sont d'autant plus sensibles à la sélection que la densité en marqueurs de la carte sur laquelle elles sont appliquées est faible. Il semble aussi que les hypothèses sur lesquelles reposent les modèles d'évolution du déséquilibre de liaison dans les populations influent sur la précision de la cartographie. Des données complémentaires à cette étude pourraient être obtenue en comparant les méthodes sur des données où le QTL est issu d'une unique mutation causale.

La méthode « cluster » dont l'objectif était de cumuler l'efficacité de la régression (démontrée dans les analyses monomarqueurs) et l'approche IBD pour synthétiser l'informatique haplotypique, ne confirme malheureusement pas les propriétés espérées.

La méthode LDLA (Meuwissen et Goddard 2000) est la méthode la plus performante sur une carte dense avec des marqueurs bi-alléliques, même si elle est sensible à la sélection, permettant d'atteindre des résolutions inférieures au cM.. L'utilisation d'haplotypes permet de compenser la faible informativité des marqueurs, qui pénalise la régression. Malgré cela, les locus où le QTL est cartographié sont rarement ceux en LD maximum avec le QTL.



**QUATRIEME PARTIE :**  
**DISCUSSION GENERALE ET**  
**PERSPECTIVES**





## Introduction

La cartographie de gènes est une thématique de recherche qui s'est fortement développée au cours des deux dernières décennies, et la mise en place de méthodes statistiques de cartographie fine a pris de l'importance depuis le début des années 1990. Tout ceci a été rendu possible grâce à l'avènement de marqueurs moléculaires de plus en plus nombreux qui couvrent maintenant densément les génomes. Cette thèse s'inscrit dans cette évolution et vise essentiellement à caractériser le comportement de différentes méthodes dans les situations usuelles de génétique animale. En effet, les méthodes sont déjà nombreuses, mais, bien qu'applicables à des populations animales, elles n'ont pas été testées en confrontation aux forces évolutives auxquelles sont soumises ces populations, notamment la sélection. Ce test s'avère nécessaire pour optimiser le choix de méthode à privilégier pour la cartographie et éventuellement mettre en place des correctifs adaptés à la sélection.

La démarche de thèse a été construite en deux étapes. Il s'agissait tout d'abord de développer un outil permettant de s'ajuster aux histoires évolutives des populations animales domestiques. Le simulateur a permis de caractériser la localisation des locus marqueurs en déséquilibre d'association maximum avec le QTL. Cette étude repose sur l'hypothèse que les méthodes utilisant le LD vont localiser le QTL à la position en LD maximum avec lui.

Dans un deuxième temps les méthodes de cartographie fine ont été abordées. Celle de Meuwissen et Goddard (2000) tend à devenir la méthode de référence en matière de cartographie fine en génétique animale, d'une part car elle combine liaison et déséquilibre de liaison avec le gain de robustesse que cela implique, d'autre part car elle repose sur des ingrédients bien maîtrisés par la communauté scientifique des généticiens animaux. Elle a recours à des probabilités d'IBD pour tracer les associations préférentielles entre marqueurs. En conséquence, le comportement des distributions de ces probabilités, telles qu'estimées avec la méthode de calcul de Meuwissen et Goddard (2001), a été étudié pour des populations soumises à la sélection. Enfin, la précision de la localisation du QTL obtenue avec différentes méthodes de cartographie fine a été comparée sur des données simulées de populations sélectionnées.

Dans cette discussion générale seront discutés non seulement les apports de cette thèse en ce qui concerne le déséquilibre de liaison et la cartographie fine de QTL, mais aussi les pistes qui n'ont pu être abordées au cours de ce travail.

## I. Réalisme des simulations

### I.A. Adéquation aux populations réelles

Le travail effectué sur le LD et sur la cartographie fine repose sur le simulateur qui a été développé. Celui-ci devait dans un premier temps permettre de conduire ces études sur des structures de populations de type « petites-filles », c'est-à-dire issues d'une seule population. Compte tenu de la variété de situations rencontrées en populations animales de rente, ses capacités ont été étendues afin de permettre des études ultérieures sur des populations expérimentales issues de croisements entre deux populations ou deux races, classiquement utilisés en volaille et en porcins.

Il est envisageable que les populations actuelles soient issues de plusieurs fusions de populations au cours de leur histoire. C'est notamment le cas des populations expérimentales à 3 ou 4 voies (voire plus) que l'on peut rencontrer en volaille (Abplanalp 1967) ou dans des populations murines (Galsworthy 2002). Le simulateur ne permet pas la création de tels dispositifs, et il pourrait être intéressant de l'étendre à de telles situations.

Il a été supposé dans l'ensemble des études réalisées qu'il n'y a pas de LD dans la population fondatrice, ce qui est une hypothèse très forte. Dès lors, tout le LD créé résulte de la dérive et, s'il y en a, de la sélection. Or, il est probable qu'un certain nombre des QTL aujourd'hui cartographiés soient issus d'une mutation ponctuelle chez un individu. Si l'effet du QTL créé est fort sur un caractère sélectionné, on peut supposer que cette mutation est relativement récente, sans quoi elle aurait eu de grandes chances d'être fixée, à moins d'être liée à un autre gène qui soit contre sélectionné, auquel cas la liaison génétique est liée à une pléiotropie défavorable. Si elle reste en ségrégation dans la population, le pedigree peut parfois permettre d'identifier l'individu chez qui est survenue la mutation causale. Ce fut notamment le cas chez les bovins laitiers en race Holstein avec le gène BLAD, qui provoque un défaut d'adhérence des leucocytes chez les bovins, dont l'origine a été trouvée chez un taureau (Grobet *et al.* 1991). Par contre, si elle a un effet modéré à faible, elle peut s'être maintenue et propagée dans la population, sans avoir d'effet net au niveau phénotypique qui permette de l'identifier. On peut penser que les phénotypes observés aujourd'hui ont une composante génétique qui mélange à la fois des mutations anciennes et des mutations récentes, qui peuvent être issues d'une situation de LD complet ou de plusieurs mutations qui ont été réunies dans une même population lors de mélanges.

Le simulateur développé permet d'appliquer une grande variété de forces évolutives (dérive, mutation, goulet d'étranglement et sélection), toutes présentes dans les populations d'animaux de rente. Le type de sélection appliqué est de type sélection « massale » ou phénotypique, c'est-à-dire que seuls les individus possédant les meilleurs phénotypes pour le caractère considéré sont retenus. Cette méthode de sélection a été largement prépondérante dans les populations animales. Cependant, les animaux de rente sont depuis longtemps sélectionnés sur plusieurs phénotypes simultanément (docilité et production par exemple à l'origine). Ces caractères peuvent avoir des corrélations génétiques négatives, ce qui implique de trouver un équilibre entre les caractères sélectionnés. Ceci peut s'avérer complexe avec plus de deux caractères considérés simultanément dans le cas d'une sélection massale, car cela nécessite d'introduire des priorités (ou pondérations) entre les caractères. C'est le principe de l'index de sélection (Hazel 1943) dont les pondérations dépendent de critères économiques ou sociaux (Olesen *et al.* 2000).

De plus, la sélection n'est aujourd'hui plus effectuée sur la valeur phénotypique de l'individu, mais sur sa valeur génétique estimée pour différents caractères. Ces valeurs sont calculées grâce à la procédure de la meilleure prédiction linéaire non biaisée (Best Linear Unbiased Prediction, BLUP ; Henderson 1976) qui permet l'estimation simultanée d'effets fixes et d'effets aléatoires. Cette méthodologie, appliquée sur différents caractères combinés ensuite en un index, devrait être appliquée en fin de simulation pour approcher au plus près de la situation réelle, voir idéalement proposer une estimation BLUP multicaractères, permettant de prendre en compte les corrélations entre caractères dans l'estimation des valeurs génétiques des individus.

La prise en compte d'une sélection phénotypique unicaractère est une première approximation pour accéder ultérieurement à des simulations plus complètes et plus proches de la situation actuelle.

Par ailleurs, les générations simulées sont non chevauchantes, c'est-à-dire que les individus de la génération  $g$  ne peuvent être parents que d'individus de la génération  $g+1$ . Ceci n'est pas vrai dans certaines populations animales de rente, où il n'est pas rare qu'un individu ayant une forte valeur génétique soit choisi comme reproducteur par des éleveurs pendant un nombre d'années dépassant l'intervalle de génération de l'espèce considérée. Cette tendance a été renforcée dans les espèces à grand intervalle générationnel par le développement de l'insémination animale (qui permet de conserver la semence mâle) et de la conservation d'embryons (qui peuvent être transplantés quelques années après leur création).

Enfin, le pedigree généré pour la cartographie dans les simulations n'est enregistré que pour les deux dernières générations, alors que, dans un grand nombre d'espèces, on peut tracer plus loin l'ascendance des animaux actuels. Ainsi, par exemple, on peut remonter jusqu'à 11 générations en race bovine Pie Rouge des Plaines (Boichard *et al.* 1996), ou 16 en race porcine Piétrain (Maignel *et al.* 1998). La connaissance d'une généalogie sur plus de deux générations, avec la possibilité d'une généalogie incomplète pour les générations les plus anciennes, permettrait de s'approcher plus exactement de structures de données réelles. De telles données, sur 4 à 5 générations, peuvent en effet améliorer considérablement la précision d'estimation des effets des haplotypes dans une population complexe (Sölkner *et al.* 2007), et par conséquent la précision de cartographie.

### **I.B. Adéquation aux données génétiques réelles**

Les deux grands types de marqueurs utilisés dans les populations d'animaux de rente sont pris en compte dans ce simulateur avec les types de mutation qui leur sont classiquement associés. En effet, pour les marqueurs de type SNP, les mutations sont rares et souvent uniques à une base donnée, ce qui explique qu'ils soient bialléliques. Dans le simulateur, une mutation transforme un allèle en l'autre. Pour les marqueurs multi-alléliques (de type microsatellites), le Stepwise Mutation Model (Kimura et Ohta 1978) a été adopté. Les taux de mutation sont fixés pour chaque type de marqueur.

A l'échelle d'un génome complet, il est connu que certaines régions chromosomiques sont sujettes à plus de recombinaisons que d'autres ; elles sont appelées « hot spot » de recombinaison (Shiroishi *et al.* 1993). De plus, il apparaît que certaines régions chromosomiques concentrent plus de mutations que d'autres. Si on voulait simuler un génome complet, il pourrait alors être intéressant de mettre en place ces types de phénomènes. En effet, les hot spots de recombinaison contribuent à une augmentation du nombre de recombinaisons pour une distance physique donnée, ce qui devrait augmenter la résolution de cartographie atteinte.

Un autre paramètre qui n'a pas été implémenté bien qu'il soit courant dans les données réelles est la présence de données manquantes. Dans toutes les simulations réalisées, les haplotypes étaient parfaitement déterminés (c'est-à-dire que les phases parentales étaient connues avec certitude) et ce pour tous les marqueurs. Si la première hypothèse est assez souvent vérifiée (dès lors que les individus ont suffisamment de descendants), la seconde l'est

assez rarement, et entraîne des pertes d'information importantes quant aux ségrégations d'haplotypes dans la population, affectant les calculs de LD et de probabilités d'IBD.

Une dernière voie pour permettre aux simulations réalisées de s'approcher au plus près de cas réels serait de partir de données génomiques réelles sur lesquelles serait appliqué un simulateur de coalescence. Les haplotypes fondateurs ainsi obtenus pourraient être introduits dans le simulateur de gene dropping afin d'obtenir une plus grande diversité d'haplotypes que celle observée. Ceci pourrait permettre d'estimer des puissances de dispositifs, des seuils de significativité des tests et éventuellement de valider des stratégies de cartographie sur des données obtenues à partir d'observations réelles.

## **II. Effets de la sélection**

Grâce au simulateur développé, les effets de la sélection sur différents paramètres intervenant dans la cartographie fine ont pu être étudiés. La sensibilité du LD et des probabilités d'IBD à la sélection a été étudiée dans un premier temps, pour dans un deuxième temps mieux interpréter l'impact de la sélection sur les méthodes de cartographie fine.

### **II.A. LD et lien LD - IBD**

Le déséquilibre de liaison est devenu un outil clef en génétique animale, tant pour son utilisation en matière de cartographie fine (Xiong et Guo 1997) que pour ses applications, notamment en sélection assistée par marqueurs (Dekkers 2004). Il a beaucoup été étudié d'un point de vue théorique, mais les premières estimations réelles de l'étendue du LD dans les populations animales de rente datent seulement de l'an 2000 (Farnir *et al.* 2000). La mise en évidence de LD à grande distance fait écho à l'effet attendu des forces évolutives présentes dans ces populations, notamment la dérive génétique et la sélection. A ce jour, aucune étude n'avait été menée pour évaluer l'influence de la sélection sur la structure du LD, et notamment sur la localisation du locus en LD maximum avec un QTL soumis à sélection. Cette position devrait, si les histoires de population étaient simples, coïncider avec la localisation déterminée pour un QTL si seul le LD est utilisé.

La partie 2.2 a permis de démontrer que la sélection influence effectivement la structure du LD, et notamment la localisation des locus en LD maximum avec le QTL. L'effet principal, mais négatif, de la sélection en vue d'une utilisation en cartographie fine est qu'une

plus faible proportion de ces locus se concentre autour de lui. Ceci pourrait expliquer le LD à forte distance trouvé dans les études de populations réelles (Farnir *et al.* 2000, Vallejo *et al.* 2003, Heifetz *et al.* 2005, Harmegnies *et al.* 2007). A l'effet de la sélection s'ajoutent probablement ceux de l'effectif génétique faible des populations et, sans doute, d'un apparemment plus important que supposé des animaux analysés. La plus longue carte génétique ici simulée couvrait 19 cM symétriquement répartis autour du QTL, soit une distance maximale entre le QTL et les marqueurs de 9,5 cM. Cette distance reste faible par rapport aux 20 cM de distance maximale à laquelle un LD significatif entre marqueurs était obtenu dans l'étude de Heifetz *et al.* (2005), qui pourtant concluait à la plus faible distance à laquelle du LD était observé parmi les études menées dans des populations réelles.

Malgré la baisse de ce regroupement à proximité immédiate du QTL, les locus en LD maximum restent compris dans un intervalle restreint, comme l'indiquent les intervalles de confiance à 95% des positions des locus en LD maximum avec le QTL. La pression de sélection qui a été appliquée dans cette étude était faible, et ce pour limiter les taux de fixation aux marqueurs. Dans les populations réelles, la sélection repose sur plusieurs caractères, qui sont *a priori* dans des régions chromosomiques différentes, ce qui peut contribuer à étendre le LD. De plus, des phénomènes génétiques (mutation, fitness) participent probablement au maintien d'une variabilité génétique, qui permet d'identifier des marqueurs tout au long du génome, en particulier à proximité de gènes sélectionnés. Des locus en LD à grande distance ont été identifiés dans les populations réelles, mais les données de Heifetz *et al.* (2005) vont dans le sens des observations faites ici à partir des simulations : ils trouvent que 85% des locus en LD sont distants de moins de 10 cM.

Dans l'ensemble des études conduites dans ce travail apparaît également la notion d'IBD. L'une des questions sous-jacentes au croisement de ces deux informations, qui sont liées mais pas équivalentes, est de connaître la force de ce lien. Il ressort de l'ensemble des études présentées dans ce document que la notion d'IBD et celle de DL ne sont en fait que faiblement liées. En effet, il a été démontré que :

- le locus en LD maximum avec le QTL n'appartient pas systématiquement au segment IBD du QTL,
- la corrélation entre la probabilité d'IBD (calculée selon la méthode de Meuwissen et Goddard, 2001) et le statut réel d'IBD au QTL augmente seulement légèrement avec le LD, alors qu'elle est élevée même avec très peu de LD au sein de l'haplotype centré sur le QTL.

La disjonction entre le maximum de LD et le segment IBD contenant le QTL peut être due au caractère fluctuant des fréquences alléliques des marqueurs à proximité du QTL. Il est établi que les valeurs de LD mesurées entre deux locus ne sont pas fonction uniquement de la distance entre eux, mais aussi des fréquences alléliques à chaque locus et des fréquences observées pour les haplotypes. La dérive affecte fortement et de manière aléatoire ces fréquences alléliques. De ce fait, un locus en LD maximum avec un autre locus n'est pas nécessairement le plus proche. Au contraire, la probabilité de deux locus d'être IBD est indépendante de ces fréquences alléliques. Elle dépend de la transmission des haplotypes au cours des générations, et des recombinaisons qui ont pu intervenir. De ce fait, plus on est proche d'un locus, plus la probabilité d'être IBD à ce locus est grande pour deux haplotypes issus d'un même ancêtre. Cette probabilité est donc une fonction décroissante du temps ; elle ne dépend pas de la fréquence allélique mais de l'origine des haplotypes.

De plus, il apparaît que la sélection affecte ce lien entre le LD et l'IBD en diminuant la proportion de locus en LD maximum qui appartiennent au segment fondateur contenant le QTL, mais aussi en diminuant la corrélation entre la probabilité d'IBD et le statut réel d'IBD au QTL pour un niveau de LD donné. Cependant, ces observations sont à nuancer quant à leurs conséquences pour les méthodes de cartographie fine utilisant les probabilités d'IBD. En effet, la baisse de la proportion de locus en LD maximum avec le QTL contenus dans le segment fondateur contenant le QTL peut être due à des fréquences alléliques extrêmes dans ce segment, qui pénalisent le LD. Toutefois, si la carte est pourvue d'une forte densité de marqueurs, cette proportion reste supérieure à 50%, alors même que la longueur du segment IBD ne dépend pas de la densité de la carte génétique et ne varie que peu sous l'effet de la sélection. Par conséquent, on peut supposer qu'en utilisant une carte suffisamment dense, les locus en LD maximum avec le QTL seront presque tous dans le segment IBD du QTL.

La corrélation entre les probabilités d'IBD et le statut IBD réel est dépendante du calcul de la probabilité, qui repose essentiellement sur le statut IBS des marqueurs. Or les populations simulées tout au long de ce travail sont initialement en équilibre de liaison, c'est-à-dire qu'un même allèle du QTL peut être porté par plusieurs haplotypes. Lorsque la population est soumise à la sélection, le nombre d'haplotypes présents dans la population chute fortement. Mais ceux qui subsistent sont plus fréquemment IBD au QTL, et ils peuvent résulter de recombinaisons entre haplotypes qui ne sont pas nécessairement IBS aux marqueurs proches du QTL. Les probabilités d'IBD diminuent donc, alors que les QTL sont bien IBD. Par conséquent, pour pouvoir estimer les effets phénotypiques des haplotypes, il convient



essentiellement de bien pouvoir individualiser les groupes d'haplotypes, ce que laisse présager la bonne capacité discriminatoire des probabilités d'IBD (partie 3.1).

## **II.B. Cartographie**

Du fait de l'influence de la sélection à la fois sur le LD (partie 2.2) et sur les probabilités d'IBD (partie 3.1), il semblait très probable que les méthodes de cartographie fine soient influencées par cette force évolutive. Ceci pouvait en plus être accentué par les hypothèses sous-jacentes aux modèles utilisés par ces méthodes, puisque toutes utilisent l'équilibre de Hardy-Weinberg pour estimer les fréquences alléliques ou haplotypiques, ce qui est par définition faux dans des populations sélectionnées.

Malgré la disjonction entre les probabilités d'IBD et le LD, la méthode de cartographie fine utilisant ces probabilités s'est avérée être celle avec le plus de concordance entre les locus sur une carte dense de marqueurs bi-alléliques, même si la proportion de concordances est très faible. La régression est par contre la méthode offrant le plus de concordances avec des marqueurs multi-alléliques sur une carte lâche. Ces concordances maximales correspondent aux situations où l'écart entre la position en LD maximum avec le QTL et la position estimée du QTL est minimisée. L'écart entre ces deux locus, même lorsqu'il est minimal, reste important, de l'ordre de la distance entre 2 marqueurs sur une carte lâche et entre 5 marqueurs sur une carte dense. Ceci signifie que les méthodes de cartographie fine identifient comme QTL des locus différents de ceux qui sont en LD maximum avec le QTL. Comme aucune des deux méthodes ne présente de biais de la localisation moyenne par rapport à la position réelle (partie 2.2 pour le LD et 3.2 pour les méthodes de cartographie fine), cet écart pourrait être dû à l'adjonction de l'information phénotypique, aux modèles statistiques ou à la sélection d'haplotypes lors de la formation des deux dernières générations.

Avec un QTL simulé au centre de groupes de liaison équilibrés en matière d'information marqueurs, les localisations moyennes du QTL estimées par les différentes méthodes n'ont pas montré de biais. Par contre, leurs variances d'estimations sont affectées par la sélection : celles-ci augmentent avec la pression de sélection, mais des variances d'estimation significativement différentes de la situation de référence (population non sélectionnée) apparaissent pour des pressions de sélection différentes selon la méthode, et cela de manière plus accentuée sur une carte où les marqueurs sont espacés de 1 cM par rapport à une carte dix fois plus dense. La régression est l'une des méthodes les moins sensibles à la

sélection, alors que la méthode LDLA perd en précision de localisation dès que la population est sélectionnée, même faiblement, mais elle reste la plus précise sur une carte de forte densité. Les deux méthodes développées à partir de celle de Terwilliger semblent un peu plus sensibles à la sélection que la régression, mais moins que la méthode LDLA. Ces méthodes étaient par ailleurs défavorisées par rapport aux précédentes du fait de la population initiale simulée en équilibre de liaison, ce qui s'oppose à l'hypothèse d'une mutation causale unique sur laquelle elles reposent. En ce qui concerne la méthode proposée, qui regroupe les haplotypes selon leurs probabilités d'IBD, elle se montre aussi sensible à la sélection que la méthode LDLA sur la carte lâche, mais moins sur la carte dense. Elle est néanmoins moins précise que la méthode LDLA dans tous les scénarios étudiés, sans doute influencée par une structure d'haplotypes IBS défavorable à des regroupements optimaux.

Enfin, contrairement à l'étude de Zhao *et al.* (2007), mais conformément à celle de Grapes *et al.* (2004), il est apparu que la régression localise moins précisément le QTL que la méthode LDLA lorsque les marqueurs sont bi-alléliques, donc peu informatifs. La différence entre l'étude de Zhao *et al.* et les deux autres est que, dans leur étude, les marqueurs ayant des fréquences extrêmes étaient supprimés, ce qui est généralement le cas dans les études réelles. Cependant, les résultats obtenus dans la partie 2.2 montrent que la perte d'informativité des marqueurs, qui influe sur l'identification des locus en LD maximum avec le QTL, peut être compensée par l'utilisation d'haplotypes. En effet, utiliser ces haplotypes, comme le fait la méthode LDLA, permet d'augmenter l'information apportée pour l'utilisation du LD, ce qui augmente le nombre de situations où le locus en LD maximum avec le QTL est identifié dans un intervalle restreint autour du QTL (partie 2.2). Cette augmentation semble encore plus efficace lorsque la longueur de l'haplotype est inférieure à la longueur du segment IBD fondateur qui contient le QTL, comme le montrent les comparaisons entre cartes denses et lâches.

De cette étude découle une possibilité de stratégie pour affiner la localisation des QTL : débiter par une régression marqueur par marqueur avec une carte peu dense composée de marqueurs multi-alléliques, puis, sur l'intervalle de localisation indiqué par la régression, appliquer la méthode LDLA avec une grande densité de marqueurs bi-alléliques.

Cependant, pour vérifier tout à fait le caractère optimal de la méthode LDLA en carte très dense, il semblerait intéressant de la comparer aux autres méthodes sur des données de populations sélectionnées où le QTL est issu d'une unique mutation causale. La sensibilité de

la méthode LDLA à la sélection pourrait être due à la sensibilité des probabilités d'IBD. Or d'autres méthodes existent pour calculer de telles probabilités, qui sont une extension du coefficient de parenté. Ce coefficient est la probabilité que deux gènes d'un même locus, choisis aléatoirement chez deux individus, soient IBD. Il a pendant longtemps été calculé à partir du pedigree uniquement (Malécot 1948). Puis des méthodes ont été développées pour l'estimer à l'aide de données de marqueurs, sans utilisation conjointe de l'information du pedigree et de plusieurs marqueurs liés (Lynch et Ritland 1999 ; Thomas et Hill 2000 ; Wang 2002 ; Leutenegger *et al.* 2003). Pourtant, ces deux éléments apportent de l'information supplémentaire sur les probabilités d'IBD, car des individus apparentés ont plus de chances d'avoir un ou des gènes IBD, et l'utilisation de plusieurs marqueurs liés apporte plus d'information sur le statut IBD d'un locus donné que plusieurs marqueurs qui ne lui sont pas liés (Hernandez-Sanchez *et al.* 2005). Le calcul des probabilités d'IBD proposé par Meuwissen et Goddard en 2001 utilise ces deux informations, tout comme la méthode proposée par Hernández-Sánchez *et al.* en 2006.

Hernández-Sánchez *et al.* (2006) utilise une régression pour prédire la probabilité d'IBD à un locus  $L$  en fonction de l'information aux marqueurs proches et de l'histoire de la population. Ils ont intégré cette probabilité à une estimation des composantes de la variance pour détecter un QTL, conformément à Meuwissen et Goddard (2000). Les différences principales sont que la méthode de Meuwissen et Goddard prend en compte toute l'information apportée par l'haplotype, alors que celle de Hernández-Sánchez *et al.* (2006) n'utilise au plus qu'une paire de marqueurs. De plus, la méthode de Hernández-Sánchez *et al.* (2006) prend en compte le type de population (monoïque ou dioïque, avec ou sans autofécondation, avec ou sans structure familiale), et peut utiliser l'information génotypique ou haplotypique, ce qui affranchit du problème des reconstructions de phases. Enfin, cette méthode ne suppose pas que les fondateurs sont non apparentés. Hernández-Sánchez *et al.* (2006) ont comparé leur méthode à celle de Meuwissen et Goddard (2000 2001) en termes de puissance et de robustesse aux hypothèses sur la population initiale. Ils ont trouvé que leur méthode était plus robuste aux hypothèses sur le LD initial ou la taille de la population, mais moins à celles sur les fréquences initiales des allèles. Il pourrait être intéressant de tester la robustesse de leurs probabilités à la sélection de la même manière que ce qui a été réalisé avec les probabilités de Meuwissen et Goddard, et de comparer la précision de la cartographie fine obtenue avec une méthode des composantes de variance appliquée à des populations sélectionnées, notamment avec de faibles densités de carte.

Différentes catégories de méthodes de cartographie fine ont été utilisées dans l'étude de comparaison, mais les méthodes bayésiennes, telles que celles développées par Perez-Enciso (2003) ou Zöllner et Pritchard (2005), en ont été exclues. La raison principale à ce choix est le temps de calcul nécessaire pour chaque analyse avec de telles méthodes. En effet, elles procèdent par échantillonnages successifs pour estimer tous les paramètres du modèle, qui sont en général beaucoup plus nombreux que ceux estimés dans les méthodes utilisées dans cette thèse. En théorie, les modèles de coalescence peuvent intégrer la mutation, la sélection naturelle ou des taux de recombinaison différents entre les régions. Cependant, la sélection artificielle, telle que pratiquée dans les populations animales n'est que rarement évoquée. De plus, la méthode de Perez-Enciso (2003) est sensible à l'hypothèse de la structure initiale du LD (Perez-Enciso 2003), comme semblent l'être les méthodes HapIM et Terwilliger, et ne semble donc pas un bon candidat pour trouver une méthode optimale. La méthode de Zöllner et Pritchard (2005) demanderait à être testée sur des données issues de populations sélectionnées simulées.

Enfin, la comparaison des méthodes portait sur leur précision, étude qui n'est possible que grâce à la répétition de simulations. Avec des données réelles, une seule analyse est réalisable, et le QTL est cartographié dans l'intervalle montrant la plus grande probabilité de le contenir, avec des seuils déterminés grâce à des permutations, du bootstrapping ou un LOD drop-off. Avoir des courbes lissées rend cette détermination visuellement plus évidente. Or, parmi les méthodes étudiées, seule la méthode LDLA fournit de tels profils, grâce à la combinaison d'information haplotypiques et de l'analyse de liaison intra-famille (données non présentées), qui semblent plus engageants pour les personnes utilisant les méthodes de cartographie fine dans un contexte appliqué, ce qui lui confère un avantage supplémentaire.

## **Conclusion générale**

Ce travail visait à apporter des éléments sur l'influence de la sélection sur le LD et la cartographie fine de QTL. Cet objectif a été abordé en deux étapes :

- l'étude de l'influence de la sélection sur la structure du LD au travers de la localisation des locus en LD maximum avec le QTL,
- puis l'analyse des comportements des méthodes de cartographie fine lorsqu'elles sont appliquées sur des populations sélectionnées.

Il ressort de ce travail que la sélection, même si la pression sélective est faible :

- contribue à l'extension des régions en déséquilibre de liaison maximal avec le QTL,
- affaiblit la capacité discriminante des probabilités d'IBD proposées par Meuwissen et Goddard en 2001,
- et diminue la précision des méthodes de cartographie fine, surtout si la carte utilisée n'est pas densément couverte.

L'intérêt de l'utilisation d'haplotypes, tant pour l'évaluation de l'étendue du LD qu'à des fins de cartographie fine, a également été mise en évidence. En effet, les haplotypes permettent d'augmenter l'information utilisée, ce qui s'avère particulièrement efficace sur des cartes denses lorsqu'on emploie une méthode combinant le LD et l'analyse de liaison, où les haplotypes utilisés sont majoritairement issus directement des haplotypes fondateurs. Les cartes à très hautes densités devraient donc permettre de localiser très finement des QTL, même dans des populations sélectionnées.





# Références bibliographiques





**Abdallah J.M., Mangin B., Goffinet B., Cierco-Ayrolles C., Perez-Enciso M., 2004.** A comparison between methods for linkage disequilibrium fine-mapping of quantitative trait loci. *Genet. Res.* 83: 41-47.

**Bengtsson B.O., Thompson G., 1981.** Measuring the strength of associations between HLA antigens and diseases. *Tissue Antigens* 18: 356-363.

**Bidanel J.P., Milan D., Iannuccelli N., Amigues Y., Boscher M.Y., Bourgeois F., Caritez J.C., Gruand J., Le Roy P., Lagant H., Quintanilla R., Renard C., Gellin J., Ollivier L., Chevalet C., 2001.** Detection of quantitative trait loci for growth and fatness in pigs. *Genet. Sel. Evol.* 33: 289-309.

**Blott S., Kim J.J., Moisis S., Schmidt-Küntzel A., Cornet A., Berzi P., Cambiaso N., Ford C., Grisard B., Johnson D., Karim L., Simon P., Snell R., Spelman R., Wong J., Vilkki J., Georges M., Farnir F., Coppeters W., 2003.** Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain on the bovine growth hormone receptor is associated with a major gene effect on milk yield and composition. *Genetics* 163: 253-266.

**Bodmer W.F., 1986.** Human genetics : the molecular challenge. *Cold Spring Harbor Symp. Quant. Biol.* LI : 1-13.

**Boichard D., Maignel L., Verrier E., 1996.** Analyse généalogique des races bovines laitières françaises. *INRA Prod. Anim.* 9 : 323-335.

**Boitard S., Abdallah J., de Rochambeau H., Cierco-Ayrolles C., Mangin B., 2006.** Linkage disequilibrium interval mapping of quantitative trait loci. *BMC Genomics* 16: 54-67.

**Cierco-Ayrolles C., Abdallah J., Boitard S., Chikhi L., de Rochambeau H., Tsitrone A., Veyrieras J.B., Mangin B., 2004.** On linkage disequilibrium measures: methods and applications, in *Recent Research Developments in Genetics and Breeding*, Research SignPost, Kerala, India, p.151-180.

**Dekkers J.C., 2004.** Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *J. Anim. Sci.* 82: E-suppl: E313-328.

**Farnir F., Coppieters W., Arranz J.J., Berzi P., Combisano N., Grisart B., Karim L., Marcq F., Moreau L., Mni M., Nezer C., Simon P., Vanmanshoven P., Wagenaar D., Farnir F., Grisart B., Coppieters W., Riquet J., Berzi P., Cambisano N., Karim L., Mni M., Moisisio S., Simon P., Wagenaar D., Vilkki J., Georges M., 2002.** Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. *Genetics* 161: 275-287.

**Farnir F., Coppieters W., Arranz J.J., Berzi P., Combisano N., Grisart B., Karim L., Marcq F., Moreau L., Mni M., Nezer C., Simon P., Vanmanshoven P., Wagenaar D., Georges M., 2000.** Extensive genome-wide linkage disequilibrium in cattle. *Genome Res.* 10: 220-227.

**Falconer D.S., Mackay T.F.C., 1996.** *Introduction to quantitative genetics*, fourth edition. Longmans Green, Harlow, Essex, UK.

**Felsenstein J., 1965.** The effect of linkage on directional selection. *Genetics* 52: 349-363.

**Gasbarra D., Sillanpää M.J., Arjas E., 2005.** Backward simulation of ancestors of sampled individuals. *Theor. Pop. Biol.* 67: 75-83.

**Gautier M., Faraut T., Moazami-Goudarzi K., Navratil V., Foglio M., Grohs C., Boland A., Garnier J.G., Boichard D., Lathrop G.M., Gut I.G., Eggen A. 2007.** Genetic and haplotypic structure in 14 European and African cattle breeds. *Genetics* 177: 1059-1070.

**Grapes L., Firat M.Z., Dekkers J.M.C., Rotschild M.F., Fernando R.L., 2006.** Optimal haplotype structure for linkage disequilibrium-based fine mapping of quantitative trait loci using identity by descent. *Genetics* 172: 1955-1965.

**Grapes L., Dekkers J.C.M., Rothschild M.F., Fernando R.L., 2004.** Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. *Genetics* 166: 1561-1570.

**Grobet L., Charlier C., Hanset R., 1991.** Le diagnostic génomique de la BLAD (*Bovine Leucocyte Adhesion Deficiency*) et ses applications. In : *Biotechnologies du diagnostic et de la prévention des maladies animales*. Ed. AUPELF-UREF, John Libbey Eurotest. Paris, 1994.

**Harmegnies N., Farnir F., Davin F., Buys N., Georges M., Coppieters W., 2006.** Measuring the extent of linkage disequilibrium in commercial pig populations. *Animal Genetics* 37: 225-231.

**Hayes B.J., Visscher P.M., McPartlan H.C., Goddard M.E., 2003.** Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* 13: 635-643.

**Hayes B., Goddard M.E., 2001.** The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* 33: 209-229.

**Hedrick P.W., 1987.** Gametic disequilibrium measures: proceed with caution. *Genetics* 117: 331-341.

**Heifetz E.M., Fulton J.E., O'Sullivan N., Zhao H., Dekkers J.C.M., Soller M., 2005.** Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. *Genetics* 171: 1173-1181.

**Henderson C.R., 1976.** A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* 32: 69-83.

**Hill W.G., Robertson A., 1966.** The effects of linkage on limits to artificial selection. *Genet. Res.* 7: 44-57.

**Hill W.G., Robertson A., 1968.** The effects of inbreeding at loci with heterozygote advantage. *Genetics* 60: 615-628.

**Jennings H.S., 1917.** The numerical results of diverse systems of breeding, with respect to two pairs of characters, linked or independent, with special relation to the effects of linkage. *Genetics* 2: 99-155.

**Jensen J., Mantysaari E.A., Madsen P., Thompson R., 1996.** Residual maximum likelihood estimation of (co)variance components in multivariate mixed linear models using average information. *J. Ind. Soc. Agric. Stat. (Golden Jub. No.)* 215-236.

**Kaufman L., Rousseeuw P.J., 1990.** Finding groups in data: An introduction to cluster analysis, edited by J. Wiley and Sons, inc., New York, USA.

- Khatkar M.S., Collins A., Cavanagh J.A.L., Hawken R.J., Hobbs M., Zenger K.R., Barris W., McClintock A.E., Thompson P.C., Nicholas F.W., Raadsma H.W., 2006.** A first-generation metric linkage disequilibrium map of bovine chromosome 6. *Genetics* 174: 79-85.
- Kim J.J., Georges M., 2002.** 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. *Asian-Aust. J. Anim. Sci.* 15: 1250-1256.
- Leutenegger L.A., Prum B., Génin E., Verny C., Lemainque A., Clerget-Darpoux F., Thompson E.A., 2003.** Estimation of the inbreeding coefficient through the use of genomic data. *Am. J. Hum. Genet.* 73:516-523.
- Lewontin R.C., 1988.** On measures of gametic disequilibrium. *Genetics* 120: 849-852.
- Lewontin R.C., 1964.** The Interaction of selection and linkage. I. general considerations; heterotic models. *Genetics* 49: 49-67.
- Lou X.Y., Todhunter R.J., Lin M., Lu Q., Wang Z., Bliss S.P., Casella G., Acland G.M., Lust G., Wu R., 2003.** The extent and distribution of linkage disequilibrium in a multi-hierarchical outbred canine pedigree. *Mamm. Genome* 14: 555-564.
- Lynch M., Ritland K., 1999.** Estimation of pairwise relatedness with molecular markers. *Genetics* 152: 1753-1766.
- MacCluer J.W., VandeBerg J.L., Read B., Ryder O.A., 1986.** Pedigree analysis by computer simulation. *Zoo. Biol.* 5: 147-160.
- Maignel L., Tribout T., Boichard D., Bidanel J.P., Guéblez R., 1998.** Analyse de la variabilité génétique des races porcines Large White, Landrace Français et Piétrain, sur la base de l'information généalogique. *Journées Rech. Porcine en France* 30: 109-116.
- Malécot G., 1948.** *Les mathématiques de l'hérédité*. Masson, Paris.
- Maynard Smith J., Haigh J., 1974.** The hitch-hiking effect of a favourable gene. *Genet. Res. Camb.* 23: 23-25.

- McPeck M.S., Strahs A.**, 1999. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am. J. Hum. Genet.* 65: 858-875.
- McRae A.F., McEwan J.C., Dodds K.G., Wilson T., Crawford A.M., Slate J.**, 2002. Linkage disequilibrium in domestic sheep. *Genetics* 160: 1113-1122.
- Meuwissen T.H.E., Karlsten A., Lien S., Olsaker I., Goddard M.E.**, 2002. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* 161: 373-379.
- Meuwissen T.H.E., Goddard M.E.**, 2001. Prediction of identity by descent probabilities from marker-haplotypes. *Genet. Sel. Evol.* 33: 605-634.
- Meuwissen T.H.E., Goddard M.E.**, 2000. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* 155: 421-430.
- Nei M., Li W.H.**, 1980. Non-random association between electromorphs and inversion chromosomes in finite populations. *Genet Res* 35: 65-83.
- Nei M.**, 1967. Modification of linkage intensity by natural selection. *Genetics* 57: 625-641.
- Nordborg M.**, 2001. Coalescent theory, in *Handbook of statistical genetics, 2<sup>nd</sup> edition*, edited by Balding D.J., Bishop M.J. and Cannings C., John Wiley and sons, West Sussex, UK, pp. 179-212
- Odani M., Narita A., Watanabe T., Yokouchi K., Sugimoto Y., Fujita T., Oguni T., Matsumoto M., Sasaki Y.**, 2006. Genome-wide linkage disequilibrium in two Japanese beef cattle breeds. *Anim. Genet.* 37: 139-144.
- Palhière I., Barillet F., Astruc J.M., Aguerre X., Belloc J.P., Briois M., Fregeat G., Bibé B., Rochambeau H., Boichard D.**, 2000. Analyse de la variabilité génétique des races ovines laitières Basco-Béarnaise, Lacaune et Manech à partir des informations généalogiques. *Renc. Rech. Ruminants* 7: 153-156.
- Perez-Enciso M.**, 2003. Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a bayesian unified framework. *Genetics* 163: 1497-1510.

- Pritchard J.K., Przeworski M.,** 2001. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* 69: 1-14.
- Qian D., Beckmann L.,** 2002. Minimum-recombinant haplotyping in pedigrees. *Am. J. Hum. Genet.* 70: 1434-1445.
- Rousseuw P.J.,** 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comp. Appl. Math.* 20:53-65.
- Schneider S, Roessli D, Excoffier L,** 2000. ARLEQUIN v. 2000: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Geneva.
- Sölkner J., Tier B., Crump R., Moser G., Thompson P., Raadsma H.,** 2007. A comparison of different regression methods for genomic-assisted prediction of genetic values in dairy cattle. 58<sup>th</sup> Annual Conference of the European Association for Animal Production, Dublin, Ireland 26-29.8.2007, Session code G18.2.
- Terwilliger J.D.,** 1995. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am. J. Hum. Genet.* 56: 777-787.
- Thomas S.C., Hill W.G.,** 2000. Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics* 155: 1961-1972.
- Thompson E.A.,** 2003. Estimation of the inbreeding coefficient through use of genomic data. *Am. J. Hum. Genet.* 73: 516-523.
- Tozaki T., Hirota K., Hasegawa T., Tomita M., Kurosawa M.,** 2005. Prospects for whole genome linkage disequilibrium mapping in thoroughbreds. *Gene* 14: 127-132.
- Tozaki T., Hirota K.I., Hasegawa T., Ishida N., Tobe T.,** 2007. Whole-genome linkage disequilibrium screening for complex traits in horses. *Mol. Genet. Genom.* 277: 663-672.
- Trauwaert E.,** 1987.  $L_1$  in fuzzy clustering, in *Statistical Data Analysis Based on the  $L_1$  norm*, edited by Y. Dodge, Elsevier/North-Holland, Amsterdam, pp. 417-426.

- Vallejo, R.L., Li Y.L., Rogers G.W., Ashwell M.S.,** 2003. Genetic diversity and background linkage disequilibrium in the north American Holstein cattle population. *J. Dairy Sci.* 86: 4137-4147.
- Vallejo R.L., Bacon L.D., Liu H.C., Witter R.L., Groenen M.A.M, Hillel J., Cheng H.H.,** 1998. Genetic mapping of quantitative trait loci affecting susceptibility to Marek's disease virus induced tumors in F2 intercross chickens. *Genetics* 148: 349-360.
- Wang J.,** 2002. An estimator for pairwise relatedness using molecular markers. *Genetics* 160: 1203-1215.
- Weir B.S., Cockerham C.C.,** 1969. Group inbreeding with two linked loci. *Genetics* 63: 71-742.
- Weir B., Cockerham,** 1974. Behaviour of pairs of loci in finite monoecious populations. *Theor. Pop. Biol.* 6: 323-354.
- Weiss K.W., Clark A.G.,** 2002. Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics* 18: 19-24.
- Weller J.I., Kashi Y., Soller M.,** 1990. Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle, *J. Dairy Sci.* 73: 2525-2537.
- Weisberg S.,** 1985. *Applied linear regression*, 2<sup>nd</sup> edn. New York : John Wiley.
- Windig J.J., Meuwissen T.H.E.,** 2004. Rapid haplotype reconstruction in pedigrees with dense marker maps. *J. Anim. Breed. Genet.* 121 : 26-39.
- Xiong M., Guo S.W.,** 1997. Fine-scale mapping of quantitative trait loci using historical recombinations. *Genetics* 145: 1201-1218.
- Yamazaki T.,** 1977. The effects of overdominance on linkage in a multilocus system. *Genetics* 86: 227-236.
- Yule G.U.,** 1900. On the association of attributes in statistics. *Philos. Trans. R. Soc. Lond. A.* 194: 257-319.



**Zhao H.H., Fernando R.L., Dekkers J.C.M., 2007.** Power and precision of alternate methods for linkage disequilibrium mapping of quantitative trait loci. *Genetics* 175: 1975-1986.

**Zöllner S, Pritchard J.K., 2005.** Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* 169: 1071-1092.

**Glossaire**

<b>Mot ou expression français</b>	<b>Traduction anglaise</b>
Dérive génétique	Random drift
Déséquilibre de liaison	Linkage Disequilibrium
Taille efficace de population	Effective population size
Schéma « grand père – petite fille »	Grand-daughter design
Analyse de liaison	Linkage analysis
Locus quantitatif	Quantitative trait locus (QTL)
Mélange de populations	Population admixture
Population animale de rente	Livestock population
Haplotype fondateur	Founder haplotype
Dispositif	Design
Génotypage sélectif	Selective genotyping
Variance résiduelle	Residual variance
Variance polygénique	Polygenic variance
Intervalle de confiance	Confidence interval
Recombinaisons historiques	Historical recombinations
Etude cas-contrôle	Case-control study
Identité par descendance	Identity By Descent (IBD)
Identité par état	Identity By State (IBS)
Vraisemblance	Likelihood
Test de rapport de vraisemblances	Likelihood Ratio Test (LRT)
Sélection massale	Selection by truncation
Goulet d'étranglement	Bottleneck
Méthode de cartographie fine	Fine-mapping method
Carte génétique	Genetic map



# ANNEXES



**Annexe 1:**  
**Notice of the simulation program**  
**“Linkage Disequilibrium with several**  
**options”**

## INTRODUCTION:

This program aims at simulating populations submitted to different evolutionary histories. It also provides information about a few indicators on the population.

This note should help in using it. We shall remark before that, because of the great number of files this program can generate, it may fail. One should first think about the information he needs and exclude the production of unused files.

## I. GENERAL INFORMATIONS

The population history is subdivided into two parts:

- ➔ the first one (“historical” part) which is known from the user only through some allelic frequencies, inbreeding coefficients or the Linkage Disequilibrium (LD) in the population (before and after a possible bottleneck).
- ➔ the second one where the pedigree is known as the phenotypes of the individuals from the last generation and the information from the previous part.

This program allows the user to simulate several QTLs located on different chromosomal regions (which can be unlinked). The map density is determined by the user and the markers can be unequally spaced. The population size has to be determined at several moments of the simulations. For further details, see parts 1b and 1c. At the beginning of the simulation, each allele is given a copy number corresponding to the number of the founder haplotype it belongs to. The number of alleles for each marker and each QTL is determined by the user.

The possibility of punctual mutations is taken into account. Each mutation creates a new copy and the new allele corresponding to this copy follows the Stepwise Mutation Model (Kimura and Ohta, 1978) if the marker is a microsatellite: it has one of the numbers adjoining to the number it had before the mutation occurred. For example, if the allele number was 2, it becomes 3 or 1 (if it had been 1, it could only have become 2). In this case, new allele numbers may be created. For SNP, the allele number only moves to the other (1 becomes 2 and 2 becomes 1).

The populations are simulated following the gene-dropping method (MacCluer *et al*, 1986).

### 1. Isolated populations

One or two populations can be simulated. They are independent for all the parameters except those entered in the file “general” (see the INFILES part).

#### a. Initial disequilibrium

The populations may be created in different ways, depending on the type of initial disequilibrium wished.

- No initial disequilibrium (option 0 for *deseq*, see INFILES)

The alleles have the same initial frequency and are attributed at random to each individual for each locus. The initial situation is therefore exactly the same for all the simulated populations from an allelic point of view but not for the generated haplotypes as

alleles are associated at random. Such a situation corresponds to the one used by Meuwissen and Goddard (2000) in their fine-mapping method.

- Partial disequilibrium (option 1 for *deseq*, see INFILES)

One of the alleles at the QTL position(s) is represented by only one copy in the initial population. The rest of the haplotype is attributed at random for all other loci. The haplotype composed of all the makers (but not the QTL) may not be unique while the whole haplotype (including the QTL) is. Such a situation classically corresponds to a mutation in one of the individuals of a population.

- Total disequilibrium (option 2 for *deseq*, see INFILES)

There also is only one copy of the QTL allele but the haplotype containing this allele is also unique. This may biologically correspond to a migration from one individual into a new population.

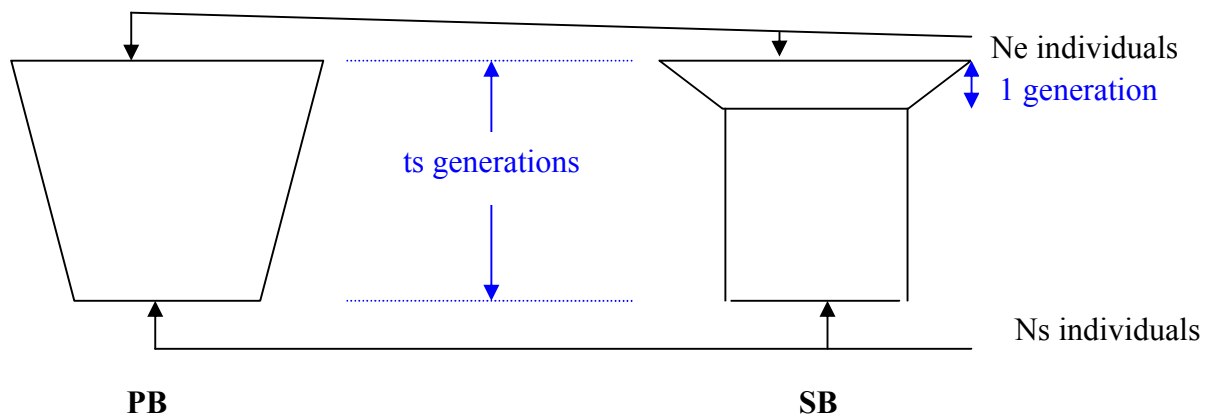
### b. Bottleneck

A bottleneck can be simulated. Two different scenarios are possible:

- It occurs on one generation and afterwards the population size remains constant (Sudden Bottleneck, SB);
- It takes place over more than one generation (Progressive Bottleneck, PB). The decrease is constant over all generations. During the last generation before the bottleneck, the population size is  $N_e$ ; at the end, it is  $N_s$ . The bottleneck takes place over  $T_s$  generations, leading to the equation at the generation  $t$  :

$$N(t) = N_e + (N_s - N_e) / t.$$

The two possibilities can be summarized by the following scheme:



### c. Selection

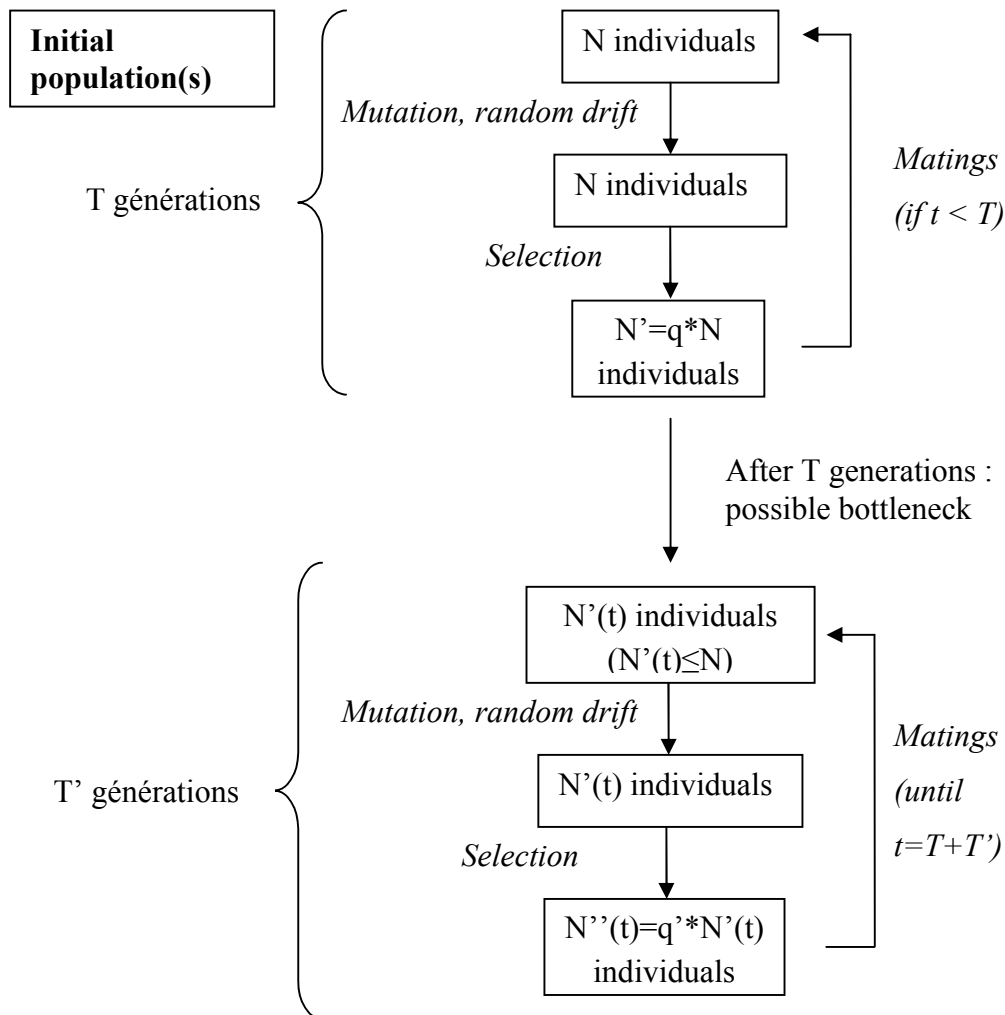
Phenotypes are simulated the following way:

$y_j = \sum_{i=1}^I q_{ij} + \phi_j + u_j + e_j$  where  $I$  is the number of QTLs,  $q_{ij}$  is the effect of the QTL  $i$  for the individual  $j$ ,  $u_j$  is a polygenic effect following a normal distribution with mean 0 and  $e_j$  is an environmental effect drawn from a normal distribution with mean 0. For each generation, the offspring receives half of the genetic value (polygenic effect and *alea de meiose*) of each of his parents. The QTL effect is the sum of the two alleles carried by the individual at the QTL. When an unique allele affects the phenotype, its value is determined by the user in the file "general". When each allele has an effect, they are drawn at random in a Gamma distribution, which parameters are given in the file "general", according to Hayes and Goddard (2001).



The simulated populations can be submitted to selection by truncation. The real number seized by the user (between 0 and 1) corresponds to the proportion of the individuals having the best phenotypes in generation  $T$  that may become parents of the generation  $T+1$ . When there is simultaneously a progressive bottleneck, the number of potential parents is calculated with the population size at generation  $T$ . Two selection rates are required for each population, one to be used before the bottleneck and one after. If the rate 1 is entered for the beginning of one of the two populations, no phenotypes are simulated in this population until the rate becomes lower than 1 or the pedigree is known (final population).

The overall evolution of the population may be represented as on Figure 1.



**Figure 1:** General structure of the program for the historical part of the simulations

## 2. Final population

It may be an admixture of the two previous populations or the continuation of the history of the first population (when only one has been simulated previously).

When two populations are simulated, three types of admixture may occur (under the option *typ\_pop=0*):

- The parents are chosen at random among all the individuals coming indifferently from the two populations;

- An F1 followed by an F2 population, meaning that all the sires come from the first population while all the dams come from the second population. There may be a few generations of back-crosses after them;
- A few individuals coming from the second population are used as parents to generate the first generation of this population.

Afterwards, other generations can be generated. The pedigree is recorded for the two last generations and the recorded phenotypes are those from the last generation.

If *typ\_pop=2*, two populations must have been simulated previously. The population created by the admixture of the two populations is an F1. An F2 generation is then simulated and back-cross generations over the first population can be created after these two generations. The pedigree is recorded for the two last generations and the recorded phenotypes are those from the last generation.

Finally, if *typ\_pop=1*, only one historical population is required. The family structure is a grand-daughter design. The recorded pedigree is composed of the grand-parents and the parents, the recorded phenotypes being the parental ones.

The new population can also be submitted to selection with a new selection rate (defined as previously). From the final populations are also known: the haplotypes and genotypes from the individuals of the 2 last generations or the number of heterozygous sires.

**II. INFILES:****1. Param**

This file provides the two seeds for simulating the two populations. They should be written on the same line.

**2. General**

This file provides the general information. If two populations are simulated, these are the common information for the both populations.

Data read by the program	type of data	name	information
2	integer	<i>nbpop</i>	number of populations to be simulated
0	real	<i>mean</i>	population mean of the phenotype
1	integer	<i>nqtl</i>	number of QTL(s)
21	integer	<i>nmq</i>	total number of loci to be simulated (including the number of QTLs)
0 (or 1)	integer	<i>plus_eff</i>	0 → one QTL allele 1 → each allele of the QTL has an additive value (different from 0). This value is drawn from a gamma distribution (Hayes B. and Goddard M.E., 2001).
11 (0.5)	integer, real	<i>posqtl, valqtl</i>	position of the QTL – effect of the QTL (if <i>plus_eff</i> equals 0) } <i>nqtl</i> lines
4.56	real	<i>alpha</i>	shape parameter of the gamma distribution (used only if <i>plus_eff</i> =1)
11.5	real	<i>beta</i>	1/scale parameter of the gamma distribution (used only if <i>plus_eff</i> =1)
0.3162	real	<i>polyg</i>	standard deviation of the polygenic effect
1.5248	real	<i>envir</i>	standard deviation of the environmental effect
0.0001	real	<i>tmutMST</i>	mutation rate of the microsatellite markers
0.0000001	real	<i>tmutSNP</i>	mutation rate of the SNP
0.01	real	<i>dist</i>	distance between positions 1 and 2
...	...	...	...
0.01	real	<i>dist</i>	distance between positions 20 and 21
0 2	integer, integer	<i>mut, typ</i>	mut = 0 if the marker cannot mutate, 1 else; typ = 1 if the marker is a microsatellite, 2 if it is a SNP
0	integer	<i>deseq</i>	type of initial disequilibrium: - =0 if all alleles are equally represented and distributed at random

- =1 if 1 of the QTL alleles (the one with an effect) is unique but attributed at random to an haplotype that may exist for other individuals
  - =2 if 1 of the QTL alleles (the one with an effect) is unique and the haplotype containing this allele also
- number of markers defining the haplotype to calculate the D' (Hedrick, 1987,

2 integer *tl\_hap*  
Lewontin, 1964) between the haplotypes and markers

### 3. Populations

Two separated populations may be simulated at the same time. They may have some distinct initial parameters, that is why two file are requested: pop1 and pop2. However, these two files have the same structure, except the last line that is specific to pop2.

## a. Initial population(s) (pop1 and pop2)

Data read	type of data	<i>name</i>	information
100 25	integer, integer	<i>ne1, t1</i>	population size and number of generations to be simulated before the possible bottleneck
100 25	integer, integer	<i>ns1, ts1</i>	population size and number of generations to be simulated after the possible bottleneck (if <i>ns1</i> < <i>ne1</i> ). <b>Be careful: the population size can only be reduced, not increased!</b>
5	integer	<i>nball</i>	number of alleles for each locus (QTL included)
[ If <i>deseq</i> =1, for each marker			
1	} integers	<i>nbporteur</i>	repartition of the alleles for each locus. <b>Be careful: the sum must be equal to twice <i>ne1</i> and there must be one of the numbers equal to one (for each QTL) ]</b>
49			
50			
50			
50	} integers	<i>quant1, quant1b</i>	Proportion of individuals conserved as potential parents for the next generation before ( <i>quant1</i> ) and after ( <i>quant1b</i> ) the possible bottleneck
1.0 1.0			
1	integer	<i>soud1</i>	type of bottleneck. <i>Soud1</i> =1 if the bottleneck occurs in one generation, =2 if the decrease of the population size occurs over <i>ns1</i> generations.
0	integer	<i>qtl_coupl</i>	takes the value 1 if the user wants the two first simulated QTLs to be on the same haplotype, 0 if the user wants them to be on two distinct haplotypes
[ For pop2 only			
0	integer	<i>mmhaplo</i>	if <i>deseq</i> =1 or 2 and the user wants the haplotypes containing the QTL to be the same in the two populations. ]

## b. Final population (popfin)

Data read	type of data	<i>name</i>	information
0	integer	<i>typ_pop</i>	corresponds to the type of population obtained. If <i>typ_pop</i> = 0, then the population has no definite type. There still may be selfing. If <i>typ_pop</i> = 1 then the population is “grand-daughter designed”. Finally, if <i>typ_pop</i> = 2, the population is an F2 type population with the possibility of back-crosses afterwards.
[ If <i>typ_pop</i> = 0			
1	integer	<i>orig</i>	corresponds to the type of admixture wished. If <i>orig</i> = 1 then there is random-mating among all individuals, independently from the population they come from; if <i>orig</i> = 2 then an F1 is produced as each offspring has one parent that

belongs to the first population and one from the second population. If *orig* = 3 then only a part of the offspring is an F1, the rest being created as offspring from individuals from the 1<sup>st</sup> population only.

				{ If <i>orig</i> = 3 only
0.9	real	<i>pcage1</i>	proportion of the offspring that are F1 individuals }	
50	integer	<i>nr</i>	number of individuals created at each generation	
3	integer	<i>tp</i>	number of generations ]	
[ If <i>typ_pop</i> =2				
20	integer	<i>tail_port</i>	litter size	
2	integer	<i>bc</i>	number of back-cross generations ]	
15	integer	<i>nt</i>	number of sires	
20	integer	<i>nf</i>	number of dams per sire (corresponding in most cases to the number of offspring per sire,	
except if <i>typ_pop</i> =2 where the number of offspring is <i>nt*nf*tail_port</i> )				
0.8	real	<i>quant3</i>	Proportion of individuals conserved as potential parents for the next generation	

#### 4. Param-Ter

This file is requested to produce “Ter” on outfile. The information to be provided is:

Data read	type of data	<i>name</i>	name in QTLIM.f	} 4 lines
y	character	<i>anal</i>	option1 to option4	
0.005	real	<i>sc_dist</i>	scan_distance	
0.000001	real	<i>tol</i>	tolerence	
10000	integer	<i>maxcal</i>	maxcal	
0.1	real	<i>qtl_fq</i>	QTL_freq	
0.1	real	<i>mu</i>	Mu	
0.5	real	<i>add</i>	a	
0.5	real	<i>dom</i>	d	
1.	real	<i>res_var</i>	Res_Var	
0.1	real	<i>lambda</i>	Lambda	
100	integer	<i>tps</i>	t	

## **5. Fichiers**

This file allows the user to select which of the output files he needs. Other files such as pedigree, haplotype, genotype or performance files are always generated at the end of the program.

### III. OUTFILES

#### 1. Files “on demand”

Two measures of the LD are computed and recorded in some of the following files: the  $D'$  (Lewontin, 1964; Hedrick, 1987) and the  $\chi^2$  (Yamazaki, 1978).

Name of the read variable	Name of the output file	Contains
$f\_dl$	simld	<ul style="list-style-type: none"> <li>– two markers between which LD is computed</li> <li>– <math>D'</math> value</li> <li>– <math>\chi^2</math> value</li> <li>– frequency of the allele of the QTL(s) affecting the phenotype</li> <li>– population size</li> <li>– generation number</li> </ul>
$f\_ibdav$	ibd-av	<ul style="list-style-type: none"> <li>– length of each segment originated from an unique founder (<b>One line for each segment → the file may be heavy !</b>)</li> <li>– population number</li> <li>– frequency of the allele of the QTL(s) affecting the phenotype</li> <li>– seed value</li> </ul>
$f\_ibdqv$	ibdqtl-av	<ul style="list-style-type: none"> <li>– QTL number</li> <li>– length of the segment issued from an unique founder and containing the QTL</li> <li>– frequency of the allele of the QTL(s) affecting the phenotype</li> <li>– seed value</li> </ul>
$f\_ibdap$	ibd-ap	Same as $f\_ibdav$
$f\_ibdqv$	ibdqtl-ap	Same as $f\_ibdqv$
$f\_ibdfin$	ibd-fin	Same as $f\_ibdav$
$f\_ibdqvfin$	ibdqtl-fin	Same as $f\_ibdqv$
$f\_csg$	consang	<ul style="list-style-type: none"> <li>– locus number</li> <li>– average inbreeding at this locus</li> <li>– variance of the inbreeding for this locus</li> <li>– frequency of the allele of the QTL(s) affecting the phenotype</li> <li>– population size</li> <li>– generation number</li> </ul>
$f\_hPIC$	haplo-PIC	<ul style="list-style-type: none"> <li>– population size</li> <li>– generation number</li> <li>– proportion of individuals conserved as potential parents for the next generation</li> <li>– number of haplotypes in the population (haplotypes defined without considering the QTL loci)</li> <li>– proportion of remaining haplotypes relatively to the initial number of haplotypes (defined without considering the QTL loci)</li> </ul>



Annexe 1 : Notice du simulateur

		<ul style="list-style-type: none"> <li>– number of haplotypes in the population (haplotypes defined considering the QTL loci)</li> <li>– proportion of remaining haplotypes relatively to the initial number of haplotypes (defined considering the QTL loci)</li> <li>– frequency of the allele of the QTL(s) affecting the phenotype</li> <li>– PIC of the QTL(s)</li> <li>– population number</li> </ul>
<i>f_fqall</i>	simfqall	<ul style="list-style-type: none"> <li>– population number</li> <li>– locus number</li> <li>– allele</li> <li>– allele frequency</li> <li>– number of the QTL(s) allele affecting the phenotype</li> <li>– population size</li> <li>– generation number</li> </ul>
<i>f_fqcop</i>	simfqcop	<ul style="list-style-type: none"> <li>– locus number</li> <li>– copy number</li> <li>– copy frequency</li> <li>– population number</li> <li>– population size</li> <li>– generation number</li> </ul>
<i>f_mk</i>	marqueur	File that can be directly used for fine-mapping purposes with Tom Druet's version of Meuwissen and Goddard's method
<i>f_Ter</i>	Ter	File that can be directly used for fine-mapping purposes with QTLIM.f
<i>f_deshap</i>	dprimax-mhap	<ul style="list-style-type: none"> <li>– position of the maximum D' (pmax) obtained between the QTL and an haplotype</li> <li>– position of the maximum <math>\chi^2</math> (kmax) obtained between the QTL and an haplotype</li> <li>– value of the maximum D' obtained between the QTL and an haplotype</li> <li>– value of the maximum <math>\chi^2</math> obtained between the QTL and an haplotype</li> <li>– number of haplotypes (defined by the allele numbers) at pmax</li> <li>– number of haplotypes (defined by the allele numbers) at kmax</li> <li>– number of remaining alleles at the QTL</li> <li>– frequency of the allele of the QTL(s) affecting the phenotype</li> <li>– average inbreeding at the QTL locus</li> <li>– number of haplotypes (defined by the copy numbers) at pmax</li> <li>– number of haplotypes (defined by the copy numbers) at kmax</li> <li>– proportion of remaining haplotypes relatively</li> </ul>

		<ul style="list-style-type: none"> <li>to the initial number of haplotypes (defined without considering the QTL loci)</li> <li>– proportion of remaining haplotypes relatively to the initial number of haplotypes (defined considering the QTL loci)</li> <li>– population size</li> <li>– generation number</li> <li>– proportion of individuals conserved as potential parents for the next generation</li> <li>– PIC at the QTL locus</li> <li>– number of simulations where the middle of the haplotype harbouring the highest <math>D'</math> with the QTL is in the segment originating from an unique founder containing the QTL locus</li> <li>– number of simulations where the middle of the haplotype harbouring the highest <math>\chi^2</math> with the QTL is in the segment originating from an unique founder containing the QTL locus</li> </ul>
<i>f_dprimmax</i>	dprimax-general	<ul style="list-style-type: none"> <li>– position of the maximum <math>D'</math> (pmax) obtained between the QTL and a marker</li> <li>– value of the maximum <math>D'</math> obtained between the QTL and a marker</li> <li>– position of the maximum <math>\chi^2</math> (kmax) obtained between the QTL and a marker</li> <li>– value of the maximum <math>\chi^2</math> obtained between the QTL and a marker</li> <li>– number of remaining alleles at pmax</li> <li>– number of remaining alleles at kmax</li> <li>– number of remaining alleles at the QTL locus</li> <li>– frequency of the allele of the QTL(s) affecting the phenotype</li> <li>– average inbreeding at the QTL locus</li> <li>– average inbreeding at pmax</li> <li>– average inbreeding at kmax</li> <li>– proportion of remaining haplotypes relatively to the initial number of haplotypes (defined without considering the QTL loci)</li> <li>– proportion of remaining haplotypes relatively to the initial number of haplotypes (defined considering the QTL loci)</li> <li>– population size</li> <li>– generation number</li> <li>– proportion of individuals conserved as potential parents for the next generation</li> <li>– PIC at the QTL locus</li> <li>– deviation from an uniform distribution at pmax (computed as a <math>\chi^2</math>)</li> <li>– deviation from an uniform distribution at kmax (computed as a <math>\chi^2</math>)</li> <li>– seed value</li> </ul>

		<ul style="list-style-type: none"> <li>– number of simulations where pmax is in the segment originating from an unique founder containing the QTL locus</li> <li>– number of simulations where kmax is in the segment originating from an unique founder containing the QTL locus</li> </ul>
--	--	--

## 2. Other outfiles

### a. **simped**

This file provides the pedigree for the last generation which is printed as follows:

individual number    sire number    dam number

For the parents, the sire and dam numbers are coded as 0. The file is structured by family of sire if *typ\_pop* equals 1 or 2.

### b. **simhaplo**

Each line is composed the following way:

i        j        M<sub>1ji</sub>   M<sub>1ji</sub>   M<sub>2ji</sub>   M<sub>2ji</sub>   ...    M<sub>(nmq)ji</sub>

with    i        the individual's number

          j        the haplotype number for the individual i

          M<sub>kij</sub>    the allele at locus k on the j<sup>th</sup> haplotype of individual i

The genotypes at all loci are written in this file. The alleles are coded as numbers.

### c. **simtypage**

Each line is composed the following way:

i        M<sub>11i</sub>   M<sub>12i</sub>   M<sub>21i</sub>   M<sub>22i</sub>   ...    M<sub>(nmq-nqt)2i</sub>

with    i        the individual's number

          M<sub>kli</sub>    the allele at locus k on the l<sup>th</sup> haplotype of individual i

There are (nmq-nqt) loci considered as the QTL loci are not printed in this file. The alleles are coded with letters.

### d. **phasout**

Each line is composed of the following:

i        j        M<sub>1ji</sub>   M<sub>1ji</sub>   M<sub>2ji</sub>   M<sub>2ji</sub>   ...    M<sub>(nmq-nqt)ji</sub>

with    i        the individual's number

          j        the haplotype number for the individual i

          M<sub>kij</sub>    the allele at locus k on the j<sup>th</sup> haplotype of individual i

There are (nmq-nqt) loci considered as the QTL loci are not printed in this file. The alleles are coded as letters.

### e. **simperf**

The phenotypes are printed in this file. The lines give the following information:

Number of traits simulated (only 1 possible) individual's number            weight  
phenotype of the individual

This file is formatted so that it can be used directly with Tom Druet's program for fine-mapping.

**f. heterozygotes**

This file contains the number of sires that are heterozygous at the QTL loci. The provided information is:

QTL number    number of sires heterozygous at this QTL    frequency of the allele of the QTL affecting the phenotype (calculated only among the sires)

**g. Ter**

The file is structured as follows:

number of phenotyped individuals  
number of marker loci (corresponding to nmq-nqtl)  
scan distance \*  
tolerance \*  
maximum number of iterations \*  
distances between the marker loci (one per line)  
QTL frequency \*  
mean \*  
additive effect of the QTL \*  
dominance effect of the QTL \*  
common within genotype variance \*  
association parameter \*  
number of generations \*  
genotypes (one haplotype per line)  
phenotypes (one per line)  
analyses to be performed with QTLIM.f \*

\* : provided in the infile "param-Ter".

**REFERENCES:**

- Hayes B. and Goddard M.E.**, 2001. *The distribution of the effects of genes affecting quantitative traits in livestock*. Genet. Sel. Evol., 33(3): 209-229.
- Hedrick P.**, 1987. *Gametic disequilibrium measures: proceed with caution*, Genetics, 117: 331-341.
- Kimura M. and Ohta T.**, 1978. *Stepwise mutation model and distribution of allelic frequencies in a finite population*. Proc. Natl. Acad. Sci. USA, 75(6): 2868-2872.
- Lewontin R.C.**, 1964. *On measures of gametic disequilibrium*, Genetics, 49: 49-67.
- MacCluer J.W., VandeBerg J.L., Read B., Ryder O.A.**, 1986. *Pedigree analysis by computer simulation*, Zoo. Biol., 5: 147-160.
- Meuwissen T.H.E., Goddard M.E.**, 2000. *Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci*, Genetics 155: 421-430.
- Yamazaki T.**, 1977. *The effects of overdominance on linkage in a multilocus system*, Genetics, 86: 227-236.



## Formations suivies durant la thèse

2004

- Journée sur le déséquilibre de liaison. 1 jour, Toulouse.
- Module ABIÉS « Write Right ». 3 jours, Paris. Cours dispensé par D. White.

2005

- Cours Supérieurs d'Amélioration Génétique des Animaux Domestiques, session « Amélioration génétique des ruminants laitiers ». 5 jours, Rennes.
- Formation à UNIX. 1 jour, Jouy-en-Josas. Cours dispensé par C. Caron.
- Cours du Département de Génétique Animale sur les méthodes MCMC. 4 jours, La Londe les Maures. Cours dispensé par C. Guihenneuc.
- Séminaire des thésards du Département de Génétique Animale. 1,5 jour, Toulouse.

2006

- Module ABIÉS « La thèse, au service de mon projet professionnel et personnel ». 3 jours, INA PG (Paris). Cours dispensé par E. Birlouez.
- « Computational techniques in animal breeding ». 5 jours, Jouy-en-Josas. Cours dispensé par I. Misztal.
- « Après la Loi d'Orientation Agricole, quelle organisation et quelle réglementation pour la sélection animale ? ». 1 jour, INA PG (Paris).
- Séminaire des thésards du Département de Génétique Animale. 1,5 jour, Limoges.

2007

- Module ABIÉS « Préparation à l'insertion professionnelle : objectif premier emploi ». 3 jours, AgroParisTech. Cours dispensé par E. Birlouez.
- Séminaire des thésards du Département de Génétique Animale. 2 jours, Jouy-en-Josas.



Par définition le déséquilibre de liaison (*Linkage Disequilibrium*, LD) décrit les associations préférentielles entre allèles de deux locus. Ce concept est devenu un outil indispensable pour la cartographie fine de locus quantitatifs (QTL), par l'identification de déséquilibres d'associations entre allèles à un locus marqueur (ou à un ensemble de locus marqueurs) et à un locus impliqué dans la variation d'un caractère quantitatif. La création et l'intensité du LD sont dépendantes des forces évolutives qui ont construit la population. Parmi ces forces, la dérive génétique et la sélection sont particulièrement actives dans les populations d'animaux de rente. Cette thèse a pour but d'étudier l'influence de la sélection sur la structure du déséquilibre de liaison autour d'un locus quantitatif, ainsi que son impact sur la précision de cartographie fine des QTL.

Un logiciel de simulation de populations a été développé dans le cadre de la thèse. A partir d'une population en équilibre de liaison, il permet de générer du LD dans des générations dites historiques, grâce à différentes forces évolutives. La détection de QTL est appliquée aux générations suivantes, de généalogie connue. Pour ces dernières générations, les principaux dispositifs de détection de QTL de génétique animale sont décrits dans le simulateur. Les données exploitées dans cette thèse sont issues de ce logiciel.

Le LD a été mesuré par le  $D'$  et le  $\chi^2$ . L'information moléculaire est apportée soit par un marqueur unique, soit par des haplotypes de 2 ou 4 marqueurs. La localisation du locus en LD maximum avec le QTL en fonction de l'âge de la population a été retenue pour décrire la structuration du LD autour du QTL. Au cours des générations, un phénomène de concentration des localisations autour du QTL, puis de déconcentration, apparaît. La distance génétique dans laquelle sont situés 95% des locus en LD maximum avec le QTL vaut 5 à 7 cM pour des marqueurs distants de 1 cM. La sélection joue un rôle négatif sur la localisation du maximum du LD : elle augmente la longueur de cet intervalle de 0 à 4 cM, par une réduction de la diversité génétique dans la région co-sélectionnée autour du QTL. Ceci pourrait impliquer une localisation moins précise du QTL par les méthodes utilisant le LD pour la cartographie.

La thèse porte sur l'influence de la sélection sur les méthodes de cartographie fine. L'une des méthodes les plus couramment utilisées en génétique animale exploite le LD dans une estimation de probabilités d'identité par descendance des locus (*Identity By Descent*, IBD). Afin d'évaluer la qualité d'estimation de ces probabilités, leur distribution a été explorée en fonction du statut IBD connu pour les allèles au QTL. Dans les populations sélectionnées, une augmentation de la fréquence des probabilités comprises entre 0,5 et 0,8 est observée par rapport aux populations non sélectionnées, affectant les locus qui sont IBD au QTL. Ce travail n'a pas permis, cependant, de proposer une règle de détermination d'un seuil de probabilité permettant de regrouper 95% des locus IBD avec un risque d'erreur inférieur à 5%. Un tel seuil permettrait de rassembler les haplotypes en un nombre limité de groupes d'haplotypes IBD, pour réduire les problèmes calculatoires lors de l'étape de cartographie par estimation des composantes de la variance classiquement réalisée.

La précision pour la cartographie de QTL de cinq méthodologies qui exploitent des informations pedigree et moléculaires différentes pour la cartographie a été comparée. Selon les données simulées, la méthode de cartographie optimale varie : la régression linéaire marqueur par marqueur fournit les meilleurs résultats lorsque les marqueurs sont multi-alléliques, alors que la méthode estimant les composantes de la variance à partir des probabilités IBD calculée sur des haplotypes de 4 marqueurs est la plus précise avec des marqueurs bi-alléliques. Toutes les cartographies sont moins précises dans des populations sélectionnées, mais la méthode la plus précise reste inchangée.

La sélection influe donc sur la structure du LD à proximité des QTL, et par conséquent sur la précision des méthodes de cartographie fine, en augmentant la distance entre le QTL ( $a$ ) et le locus en LD le plus fort avec celui-ci et ( $b$ ) entre le QTL et sa position estimée.

*Mots-clés : déséquilibre de liaison, cartographie fine de QTL, sélection, marqueurs génétiques, haplotypes, identité par descendance.*



Linkage disequilibrium (LD) is due to non random associations between alleles at two loci. It has become a classical tool to fine map loci implied in quantitative trait (Quantitative Trait Loci, QTL) determinism, through identification of the maxima of LD between alleles of a marker locus (or a group of marker loci) and a locus involved in the variability of a quantitative trait. The creation and intensity of LD evolves according to the evolutionary forces affecting the population. Among these forces, random drift and selection are particularly present in livestock populations. This PhD thesis aimed to study the influence of selection on the structure of LD around a QTL, as well as its impact on the precision on the fine mapping of QTL.

A software has been developed to simulate the evolution of populations. Starting from a population in linkage equilibrium, LD due to evolutionary forces is created over historical generations. QTL detection is applied to the next generations where the pedigree is known. The main experimental designs applied in livestock populations are implemented in the software. All data used for the further analysis of this work were obtained from this simulation program.

We first analyzed LD in the neighbourhood of the QTL by recording the location of the marker in maximum LD with the QTL. LD was estimated with two classical measures:  $D'$  and  $\chi^2$ . Molecular information was either provided by single markers or by haplotypes composed of two or four markers. The concentration of the loci in maximum LD around the QTL first increased over the generations, before reducing. For populations of 100 or 200 individuals observed after 100 generations, the 95% confidence interval of the position in maximum LD with the QTL is 5 to 7 cM for markers separated by 1 cM. Selection augments the length of this segment of 0 to 4 cM, mainly because of the loss of the genetic variability in the co-selected region located around the QTL. This implies that selection reduces the efficiency of fine mapping when methods using LD are employed.

The influence of selection on the accuracy of fine mapping methods has been studied in a second step. One of the most classical methods in animal genetics estimates the probabilities of Identity By Descent (IBD) of pairs of loci to use LD. We investigated the quality of the estimation of these probabilities in relation to the real IBD status of the QTL. IBD probabilities estimated between QTL being truly IBD were lower when the populations were selected, with an increase of the frequencies of the probabilities comprised between 0.5 and 0.8. In all the simulated designs, the probabilities of non IBD QTL remain close to 0. But it has not been proved possible to define a rule to determine a threshold allowing for grouping 95% of the IBD QTL with an error rate lower than 5%. With such a threshold, a limited number of clusters of haplotypes could be established, such that the computing difficulties when fine mapping the QTL with methods using the estimation of the variance components as classically done could be reduced.

The accuracy for fine mapping purposes of five methods that use different kinds of molecular information and include the pedigree information or not was compared. The optimal method differs depending on the simulated data: the linear regression on a single marker provided the most accurate results when multi-allelic markers are used, while the method of analysis of the variance components using 4-loci haplotypes was the most precise with bi-allelic markers. Selection affected the fine mapping accuracy but does not influence the ranking of the methods.

To sum up, selection influences both the LD structure in the QTL neighbourhood and the fine mapping accuracy by increasing the distance between the QTL and (a) the locus in maximum LD with it and (b) the QTL and its estimated position.

*Key words: linkage disequilibrium, QTL fine mapping, selection, genetic markers, haplotypes, identity by descent.*