Institut National Agronomique Paris-Grignon
Département des Sciences Animales

Institut National de la Recherche Agronomique
Station de Génétique Quantitative et Appliquée

THESE

présentée par

Hélène GILBERT

en vue de l'obtention du diplôme de
Docteur de l'Institut National Agronomique Paris-Grignon

Multidimensionnalité pour la détection de gènes influençant des caractères quantitatifs

Application à l'espèce porcine

Soutenue le 31 janvier 2003 devant le jury composé de :

Bruno Goffinet	Directeur de recherches à l'INRA	Rapporteur
Frédéric Farnir	Maître de Conférences à l'Université de Liège	Rapporteur
Dominique de Vienne	Professeur à l'Université Paris XI	Examinateur
Pascale Le Roy	Directeur de recherches à l'INRA	Directeur de thèse
Etienne Verrier	Professeur à l'INA P-G	Président

Remerciements

Ce travail de thèse a été réalisé à la Station de Génétique Quantitative et Appliquée de l'INRA de Jouy-en-Josas. A ce titre, je tiens à remercier Bernard Bibé, directeur du Département de Génétique Animale de l'INRA, et François Ménissier, directeur de la station, pour les moyens mis à ma disposition pour mener à bien ce travail.

Je remercie les membres du jury : Etienne Verrier, qui a accepté la présidence de ce jury et s'est chargé des démarches administratives, Frédéric Farnir et Bruno Goffinet, qui m'ont fait l'honneur d'être rapporteurs de ce travail, et Dominique de Vienne, qui m'a fait l'honneur d'en être examinateur.

Je tiens tout particulièrement à remercier Pascale Le Roy, qui a encadré ce travail, et, malgré un emploi du temps chargé, a toujours été disponible, m'a soutenue et guidée pendant ces trois années avec le sourire et dans la bonne humeur.

Je voudrais par ailleurs remercier Jean-Pierre Bidanel, qui m'a fourni les données du protocole PORQTL analysées dans ce document.

Je remercie aussi les membres du Club des Amis des QTL, Didier Boichard, Sara Casu, Jean-Michel Elsen, Bruno Goffinet, Pascale Le Roy, Eduardo Manfredi, Carole Moreno, Miguel Pérez-Enciso, Etienne Verrier et Zulma Vitezica, dont les éclairages réguliers ont suivi et guidé ce travail.

Je tiens remercier très sincèrement Juliette Riquet, et Denis Milan, sans qui la génétique quantitative serait pour moi restée à jamais un monde hermétique d'équations inaccessibles.

Je voudrais par ailleurs remercier Serge Tignoux, Françoise Bouchain, Pierrette Gillet et Marie-Laure Le Paih pour leur disponibilité et leur gentillesse dans l'accomplissement des démarchent administratives et de reprographie, et Sylvie Nugier et Hervé Lagant pour leur faculté à surmonter les caprices informatiques les plus divers.

Mes remerciements chaleureux vont à l'ensemble du personnel de la SGQA, et tout particulièrement aux membres permanents du "café cochon", pour leur bonne humeur et leur joie de vivre : aux badmintonniens, Sophie, Vincent et Seb, qui m'ont permis de me défouler régulièrement ; à Marie-Pierre et Thierry, voisins sympatiques quoique parfois dangeureux ; à Isa, et à Stef, pour nos crises de rire; et à Anne, Didier, Laurianne (bon courage !), François, Marie-Noëlle, Jacques, Florence, Hélène, Sophie, Raquel, Gilles, Aurélie... Je voudrais enfin adresser une mention spéciale à Denis et Florence, dont les piques constituent un plaisir "rare" et inégalé !

Je ne saurais conclure sans remercier tout ceux qui m'ont soutenue et supportée depuis longtemps, pour leur présence sans faille: ma famille, AnSo et Lionel évidemment; Orel,

qui a même eu le courage de m'accompagner en vacances; Bruno, qui a toujours une oreille attentive à prêter; Rachel, qui peut être plus que d'autres a subi mes sautes d'humeur; Gaetan, Claire et Ju, qui sont toujours trop loin. Une mention spéciale aussi à Vince, GautHier et son oncle, qui plus d'une fois m'ont redonné le sourire. Et à Katia, sans qui ces remerciements ne seraient pas ce qu'ils sont.

Et à tous ceux qui ne sont pas nommés explicitement...!

Table des matières

1 Introduction **1**

1. Détection de QTL : principes 4

2. Protocoles de détection de QTL en populations animales 5

2.1. Croisements pour l'obtention d'individus hétérozygotes au QTL . . 6

2.2. Croisements pour les générations suivantes 7

2.3. Mesures réalisées 8

2.4. Critères de choix du protocole 8

3. Méthodologies de détection de QTL unicaractères et uniQTL 10

3.1. Hypothèses comparées 10

3.2. Détection marqueur par marqueur 10

3.3. Cartographie d'intervalle 13

3.4. Statistiques de test 14

3.5. Propriétés des méthodes de cartographie unicaractère uniQTL . . . 22

4. Outils disponibles pour la détection de QTL 25

4.1. Méthode unicaractère utilisée 25

4.2. Données disponibles 26

4.3. Illustration choisie 27

5. Définition du champ d'étude 30

2 Méthodes multicaractères **32**

1. Eléments de bibliographie 32

1.1. Questions posées par les méthodologies unicaractères 32

1.2. Premières méthodes multicaractères proposées 33
1.3. Généralisation 37
1.4. Développements récents 39
1.5. Génotypage d'extrêmes 42
1.6. Bilan 42
2. Méthode proposée 43
2.1. Principes de l'analyse discriminante (DA) 43
2.2. Application à la détection de QTL 44
2.3. Généralisation 46
3. Comparaisons effectuées 48
3.1. Test d'hypothèse appliqué 48
3.2. Méthodes comparées 49
3.3. Axes de comparaison 51
3.4. Dispositifs simulés 54
4. Résultats 58
4.1. Temps de calcul 58
4.2. Seuils de rejet de H0 59
4.3. Puissances et qualité des estimations 60
4.4. Discussion 76
5. Application 81
5.1. Méthodologie utilisée 81
5.2. Résultats 82
5.3. Conclusion 87

3 Méthodes multiQTL **89**

1. Introduction 89
1.1. Questions posées par les méthodologies uniQTL 89
1.2. Panorama sur les méthodes multiQTL proposées 92
1.3. Panorama sur la distribution des statistiques de test 97

1.4. Application aux populations animales 102
2. Cadre de l'étude 103
2.1. Test d'hypothèse appliqué 104
2.2. Méthodes comparées 104
2.3. Axes de comparaison 108
2.4. Dipositifs simulés 111
3. Résultats et discussion 112
3.1. Temps de calcul 112
3.2. Positions et effets estimés sous l'hypothèse uniQTL 113
3.3. Seuils de rejet de l'hypothèse nulle 115
3.4. Puissances 115
3.5. AIC 122
3.6. Qualité d'estimation 124
4. Bilan 130
5. Application 133
5.1. Analyses unicaractères 133
5.2. Analyses conjointes des 5 caractères 134
5.3. Analyses par groupe de deux caractères 134
5.4. Analyse des deux groupes de deux caractères 136
5.5. Conclusion 137

4 Bilan et Perspectives **138**

1. Introduction 138
2. Bilan général et synthèse 139
3. Propositions de stratégie 140
3.1. Détection pour un caractère 140
3.2. Détection pour un groupe de caractères 141
3.3. Niveaux de signification des statistiques de test 142
4. Perspectives 143

Chapitre 1

Introduction

L'étude des caractères d'intérêt agronomique repose sur la mise en place de modèles statistiques qui permettent de prendre en compte les effets de différents paramètres sur les caractères, et les relations entre ces caractères. Dans le modèle le plus général, l'expression phénotypique d'un caractère est considérée comme la somme de l'action d'effets fixés, du génome, et d'effets environnementaux que l'on ne sait pas identifier. Ce modèle se traduit par l'équation suivante :

$$P = \mu + G + E \tag{1.1}$$

où P est la valeur phénotypique, $\mu = \mathrm{E}(P)$ est la moyenne des phénotypes P, G est la valeur génotypique, c'est à dire l'espérance de la valeur phénotypique conditionnée par le génotype et centré sur la moyenne : $G = \mathrm{E}(P|génotype) - \mu$, et E est l'écart de la valeur phénotypique centrée à la valeur génotypique dû aux effets de milieu non-identifiables. Chacun de ces éléments est décomposable à l'infini pour la prise en compte ou non d'interactions, d'environnements variés... La partie héritable d'un phénotype réside donc dans sa valeur génotypique, qui résulte des effets des gènes polymorphes, qui créent de la variabilité entre les individus dans une population.

La modélisation de la valeur génotypique permet de la rendre manipulable, à des fins de sélection, par exemple. Le but d'un modèle génétique est de minimiser la variance de l'écart entre la valeur génétique prédite par le modèle et la valeur phénotypique disponible, donc $\mathrm{var}(E)$. De façon schématique, la valeur génétique peut être modélisée de trois grands types de façons. Le premier revient à considérer que la valeur génotypique est le résultat de la somme des actions d'un grand nombre de gènes ayant chacun un petit effet sur le caractère. Il s'agit du modèle polygénique. Si on considère au contraire que seuls quelques gènes polymorphes déterminent le caractère, le modèle est dit oligogénique. Enfin, le modèle le plus simple ne considère l'action que d'un gène unique sur le caractère. Il s'agit

du modèle gène majeur ou monogénique. Lorsque le modèle prend en compte l'existence d'un fond génétique de type polygénique et d'un ou plusieurs gènes à effets individuels "forts", on parle d'hérédité mixte du caractère.

La valeur génotypique peut ainsi être modélisée comme la synthèse de l'action d'un ou plusieurs gènes. Si l'on considère un seul gène, pour chaque couple d'allèles notés 1 et 2, la valeur génotypique qu'il détermine peut avoir deux composantes majeures : une composante additive a et une composante dominante d. a, tel que nous l'utiliserons par la suite, est défini comme la différence entre les valeurs génotypiques des homozygotes au locus. d est défini comme l'écart de la valeur génotypique des performances des hétérozygotes au locus à la moyenne des valeurs génotypiques des homozygotes. En notant $\mu_{mm'}$ la valeur génotypique associée au génotype mm', a et d s'écrivent donc respectivement $a = \mu_{11} - \mu_{22}$ et $d = \mu_{12} - (\mu_{11} + \mu_{22})/2$. On peut alors écrire la composante génétique comme la somme des composantes additive et de dominance. La modélisation de plusieurs locus permet de modéliser des interactions entre les locus, telles que des interactions épistatiques. L'effet additif de substitution des allèles au locus, que nous utiliserons par la suite, est défini comme la moitié de l'écart entre les homozygotes, soit $a/2$ (Falconer, 1989 [20]). Les valeurs seront en général présentées comme des proportions de la déviation standard phénotypique σ_p.

Dans la pratique, l'utilisation de l'un ou l'autre des modèles est conditionnée par le but recherché. Le premier but de la modélisation des caractères est leur sélection. Dans ce cadre, la valeur génotypique des caractères d'intérêt agronomique est modélisée comme la somme des effets d'une infinité de gènes ayant chacun un effet infime sur le caractère. Il s'agit du modèle infinitésimal. Ce modèle, bien que faux par essence, a permis de sélectionner de façon efficace les individus présentant les meilleures performances au cours des dernières décennies, et donc d'améliorer considérablement les niveaux de production de la majorité des espèces auquel il a été appliqué. Un exemple classique de l'efficacité de cette méthode en sélection animale réside dans la sélection de la quantité de lait chez les brebis Lacaune, multipliée par 2,5 en 25 ans (Barillet *et al.*, 1996 [4]). Cependant, son principal défaut est de ne bien sélectionner que les caractères les plus héritables, c'est à dire pour lesquels la part de la variabilité due à la valeur génotypique est élevée par rapport à la variabilité phénotypique. Les autres caractères, qui sont en général des caractères de type reproduction, sont moins facilement améliorables par ces stratégies (Ollivier, 1981 [84]).

L'identification précise de gènes déterminant un caractère donné peut aussi être recherchée. Des gènes ayant des effets importants sur certains caractères, tels que le nombre de descendants chez la brebis (Piper et Bindon, 1990a [88] : Booroola), la qualité de la viande chez le porc (Sellier, 1998 [97] : Halothane, RN), ont ainsi été récemment identifiés et caractérisés. L'accès à cette information peut permettre d'augmenter considérablement

l'efficacité de la sélection sur le caractère considéré, tout en maintenant le niveau de performances des individus sur les autres caractères (Piper et Bindon, 1990b [89]; Manfredi *et al.*, 1998 [73]). Les modèles développés pour ce type d'étude seront de type monogénique ou mixte.

Globalement, les gènes à effet fort sur les caractères qui seraient en ségrégation dans les populations de production animale, ou "gènes majeurs", sont considérés comme peu nombreux. Une des raisons avancée est la pression de sélection importante exercée depuis quelques décennies maintenant, qui a tendance à fixer les allèles favorables déterminant la plus grande part de la variabilité des caractères. Des études ont permis d'extrapoler une distribution de la fréquence des locus polymorphes déterminant un caractère d'intérêt en fonction de leur effet sur le caractère. Cette distribution est clairement de type exponentiel, avec un très grand nombre de locus présentant des effets très faibles et peu de locus déterminant de grandes déviations du caractère (figure 1, Hayes et Goddard, 2001 [31]). Les origines de cette distribution en L ont été largement explorées par Bost *et al.* (2001) [9]).

L'identification de locus dont les effets sont relativement faibles (de 0,5 à 1 écart-type phénotypique), mais dont certains seraient en ségrégation dans les génomes des espèces d'intérêt, pourraient permettre de contrebalancer de façon non négligeable les lacunes du modèle infinitésimal, par l'identification et à terme la sélection, d'allèles améliorateurs pour les caractères peu héritables (Goffinet *et al.*, 1994). Cependant, la caractérisation de tels gènes est délicate à mettre en oeuvre. En effet, l'existence d'un mélange de distributions sous-jacent à la distribution des performances, dont les composantes sont fonction des génotypes au locus d'intérêt, ne peut être identifiable que lorsque les effets des allèles sont importants (figure 2). Des méthodes basées sur la mise en évidence d'écarts par rapport à la distribution attendue, dues à la présence de gènes en ségrégation dans la population, ont été développées dans ce sens. Ces méthodes, parmi lesquelles on peut citer les tests de normalité des distributions, l'analyse de ségrégation ou l'analyse de pedigree (Le Roy *et al.*, 1989 [68]; Knott *et al.*, 1992b [49]), sont basées sur la description des distributions des performances, et peuvent intégrer de l'information sur leur transmission à travers la connaissance de la généalogie de la population. Dans la pratique, elles servent essentiellement à mettre en évidence des gènes à effet fort. Cependant, quand les effets sont faibles, l'écart des mélanges de distributions à un modèle de distribution normale est difficilement caractérisable (figure 2) sans utiliser des informations de cartographie génétique *via* des génotypages pour des marqueurs moléculaires. Il faut alors mettre en place des protocoles de production de données et des méthodes statistiques particulières. Le but de ces méthodes est de caractériser les locus, par l'estimation de leurs effets additifs (et éventuellement de dominance), de la part de variabilité du caractère qu'ils expliquent, et de leur position sur le génome: on parle alors de détection de QTL (*Quantitative Trait*

Locus).

Nous verrons dans cette première partie introductive les stratégies qui permettent de mettre en évidence les QTL, à travers la mise en place de protocoles de croisements expérimentaux entre populations, et les modèles et méthodes statistiques associés à leur étude. Nous développerons ensuite les outils disponibles au début de ce travail, en mettant en évidence les raisons qui nous ont poussés à effectuer de nouveaux développements.

1. Détection de QTL : principes

Nous avons vu que seuls les locus pour lesquels la ségrégation d'allèles dans la population entraînent des différences de performances entre les individus sont recherchés. La détection de tels gènes est donc basée sur l'analyse de la distribution des performances de descendants issus d'un individu hétérozygote au locus ayant un effet sur le caractère quantitatif. Ces individus hétérozygotes sont en effet les seuls individus informatifs quant aux effets des allèles au locus. Pour illustrer ce phénomène, nous avons représenté dans la figure 3 la distribution des performances des descendants d'un tel individu. Cependant, les statuts, hétérozygote ou homozygote, du parent et de ses descendants pour le locus d'intérêt ne sont pas connus, et pour des effets d'allèles faibles ne peuvent pas être extrapolés à partir des performances. Les marqueurs génétiques sont alors utilisés pour identifier ces statuts.

Dans la figure 4, nous avons repris la distribution des performances des descendants d'un individu hétérozygote au locus quantitatif en y ajoutant un marqueur génétique lié. Cette figure illustre le suivi de la ségrégation des allèles au locus quantitatif au cours des générations grâce à l'utilisation d'un marqueur moléculaire complètement lié. Cependant, dans la pratique, la localisation du locus quantitatif étant inconnue, sa détection passe par l'analyse de nombreux marqueurs moléculaires répartis sur le génome. De plus, les locus marqueurs doivent impérativement être polymorphes, pour mettre en évidence une liaison entre les allèles au locus quantitatif et au locus marqueur.

Les analyses de populations végétales font en général appel aux marqueurs RFLP (*Restriction Fragment Length Polymorphism*), alors qu'en génétique animale les marqueurs génétiques les plus couramment utilisés sont les marqueurs microsatellites, et de façon plus récente les SNP (*Single Nucleotide Polymorphism*). Dans les deux cas, ces marqueurs sont nombreux, bien répartis sur le génome, supposés neutres par rapport à la sélection et polymorphes. Au cours des deux dernières décennies, des cartes génétiques couvrant l'ensemble des génomes des espèces d'intérêt agronomique ont été développées (Georges et Andersson, 1996 [22]). L'existence de ces cartes a permis la mise en place de protocoles systématiques de détection de QTL dans la majorité des espèces.

Cependant, les marqueurs génétiques utilisés, bien qu'ils soient nombreux, ne permettent pas de tester toutes les positions possibles sur le génome. Dans la pratique, et en fonction des marqueurs disponibles dans les espèces, les réseaux de marqueurs génétiques utilisés pour une première phase de détection ont des écarts de 20 à 40 cM (centiMorgan) entre les marqueurs. Darvasi *et al.* (1993 [12]) ont montré qu'un réseau de marqueurs génétiques plus dense que 20 cM ne permet pas, pour les tailles de populations analysées, d'augmenter considérablement les puissances de détection des locus déterminant des caractères quantitatifs. Pour une taille de la population et un effet de gène donné, on peut en effet définir une résolution maximale du dispositif expérimental, limitée par la quantité de recombinaisons informatives disponibles dans le dispositif (Darvasi *et al.*, 1993).

Par rapport à la situation décrite au premier paragraphe, avec un locus marqueur complètement lié au locus quantitatif, l'occurrence de recombinaisons entre les locus marqueur et quantitatif pour la production de la descendance limite cependant la capacité de détection des locus quantitatifs (Soller, Brody et Genisi, 1976 [98]). Nous verrons par la suite quelles stratégies ont été développées pour tenir compte de ces recombinaisons.

Pour détecter un locus quantitatif, il faut donc :

1. Disposer d'une descendance dans laquelle les allèles au QTL sont en ségrégation et les génotypes pour un réseau de marqueurs génétiques polymorphes répartis sur le génome sont mesurés,
2. Mesurer la valeur du (ou des) caractère(s) quantitatif(s) sur chaque individu de la descendance,
3. Mettre en oeuvre des méthodes d'analyse pour rechercher les locus marqueurs dont la ségrégation est corrélée à celle du (ou des) caractère(s) quantitatif(s) et estimer les paramètres génétiques du QTL détecté (effets, position).

Nous allons développer certaines des stratégies mises en place pour l'acquisition des points 1 et 3 dans les paragraphes suivants.

2. Protocoles de détection de QTL en populations animales

Nous allons restreindre la description des protocoles de détection de QTL aux plans de croisement mis en oeuvre dans les populations d'animaux domestiques, comme par exemple l'espèce porcine. Les populations porcines présentent diverses caractéristiques

intéressantes pour la mise en place de croisements expérimentaux : l'intervalle de génération est relativement court, de l'ordre d'un an, et le nombre d'individus d'une portée (environ 10) permet d'obtenir une descendance nombreuse, mélange de plein-frères et demi-frères. La restriction à ce type de population exclut les protocoles basés sur des croisements entre lignées génétiquement fixées, telles qu'il en existe pour la majorité des végétaux d'intérêt agronomique et pour les animaux de laboratoire, et les protocoles de détection basés sur les populations de production existantes (populations *outbred*), tels que ceux utilisés pour l'espèce bovine en particulier, en raison d'intervalles de génération longs (Weller *et al.*, 1990 [106]).

Le but de la mise en place de populations expérimentales est de maximiser le taux d'hétérozygotie des locus dans la population d'étude, ainsi que le déséquilibre de liaison entre ces locus. La première étape d'un croisement expérimental est l'obtention d'un maximum d'individus hétérozygotes aux locus quantitatifs. La deuxième étape sera la production des descendants de ces individus.

2.1. Croisements pour l'obtention d'individus hétérozygotes au QTL

Le critère déterminant l'obtention d'individus hétérozygotes pour les locus déterminant des caractères quantitatifs est le choix des populations à croiser dans la première génération. Pour maximiser le nombre de locus pour lesquels les individus obtenus seront hétérozygotes, la stratégie classique est de croiser des individus appartenant à des populations très divergentes pour les caractères d'intérêt. On suppose alors que les allèles aux locus déterminant le (ou les) caractère(s) quantitatif(s) sont différents dans les deux populations, même s'ils ne sont pas fixés intra-population. La génération d'individus hétérozygotes est classiquement appelée F1, celle de leurs parents étant notée F0.

Dans l'espèce porcine, plusieurs protocoles expérimentaux ont été réalisés aux cours de ces dernières années (Bidanel et Rotschild, 2002 [7]) en croisant une race européenne sélectionnée, type Large White, avec soit une espèce sauvage (le sanglier), soit une race "exotique" (par exemple une race chinoise, telle que la Meishan). L'objectif de ces grands programmes est de détecter un maximum de QTL sur les caractères d'intérêt économique chez le porc. En effet, les principales caractéristiques des races porcines en production sont leur bonne vitesse de croissance du tissu maigre et leurs performances relativement moyennes en terme de reproduction, qui constituent les deux axes majeurs de la sélection. Le sanglier pour sa part présente des performances très médiocres pour ces types de caractères, alors que les races chinoises combinent des vitesses de croissance lentes et des taux de gras élevés, avec des performances de reproduction très bonnes (avec par exemple une prolificité multipliée par deux).

Une autre stratégie consiste à croiser des lignées divergentes obtenues par plusieurs années de sélection (sur la croissance par exemple; Paszek, 1999 [86]). Le nombre de caractères quantitatifs considérés est alors réduit à ceux qui différencient les deux lignées.

Cependant, dans un cas comme dans l'autre, les effets des allèles décrits suite à ce type de protocole sont caractéristiques des races ou des lignées croisées. Ils ne peuvent donc pas être extrapolés directement pour l'application aux populations en sélection. L'utilisation des résultats de détection peut alors se faire selon deux axes. Le premier consiste à utiliser l'information sur la localisation pour identifier et sélectionner dans les populations en sélection les allèles codant pour les déviations les plus intéressantes du caractère. La deuxième stratégie repose sur l'identification d'allèles intéressants pour le caractère dans les populations exotiques utilisées pour le croisement. Ces allèles peuvent ensuite, à l'aide de l'information moléculaire disponible, être introgressés de façon raisonnée dans une lignée sélectionnée.

2.2. Croisements pour les générations suivantes

La figure 5 décrit les deux grands types de croisement qui peuvent être utilisés pour l'obtention des descendants des F1.

2.2.1. Croisement en retour ou *back-cross* (BC)

Le back-cross repose sur le croisement des individus F1 avec des individus d'une des deux races utilisées en F0. Cette stratégie permet d'obtenir deux classes d'individus à la génération suivante (BC1), en fonction de l'origine grand-parentale des allèles aux locus déterminant les caractères quantitatifs : des "homozygotes" porteurs des allèles de la race F0 utilisée en retour, et des "hétérozygotes". Ces deux classes ont en espérance des effectifs identiques.

2.2.2. Croisement F2 ou intercross

Dans les populations animales qui nous intéressent, les individus F2 sont obtenus en croisant des individus F1 non apparentés. Cette stratégie permet d'obtenir à la génération F2 les trois classes d'individus par rapport à l'origine grand-parentale des allèles aux locus quantitatifs.

2.2.3. Générations supplémentaires

Dans les deux cas, des générations supplémentaires peuvent être produites en continuant les croisements selon les mêmes règles. Les générations sont numérotées par ordre croissant. La première étape de détection de QTL dans les populations animales repose en général sur l'analyse de données issues de la première génération, BC1 ou F2. Les générations suivantes, qui servent à produire des recombinaisons informatives dans les régions identifiées comme porteuses de locus d'intérêt, sont réservées à la mise en place de protocoles particuliers d'affinage de la détection.

2.3. Mesures réalisées

Pour la détection de QTL, les individus des générations F0 et F1 sont en général uniquement génotypés, alors que tous les individus de la génération suivante sont génotypés et phénotypés. Le relevé des performances dans ce type de populations concerne en général un maximum de caractères d'intérêt, afin de maximiser l'information exploitable en regard du coût élevé du génotypage de tous les individus pour l'ensemble du réseau de marqueurs génétiques. Lorsque les allèles sont en ségrégation, l'identification de leurs origines, en référence aux populations grand-parentales utilisées, requiert la détermination des phases gamétiques (associations alléliques portées par un même chromosome) des parents et, par la suite, le génotypage des reproducteurs F0.

2.4. Critères de choix du protocole

Le premier critère de choix de protocole repose sur la nature des individus obtenus en dernière génération. En effet, nous avons vu que deux classes d'individus sont présentes dans un BC1, alors que les trois classes sont obtenues par un croisement F2. En se reportant aux définitions des caractéristiques a et d du QTL données en introduction, il apparaît qu'un back-cross ne permettra pas de dissocier les effets additifs a des effets dominants d. De plus, si la dominance est complète, seuls les locus pour lesquels les allèles de la race utilisée en retour sont récessifs pourront être mis en évidence. En revanche, dans un croisement F2, chaque effet peut être estimé séparément, mais les hétérozygotes ne sont pas exploités pour l'estimation de l'effet additif. Par conséquent, à taille de protocole égale, la puissance de détection de QTL à partir d'un back-cross est supérieure à celle d'un croisement F2.

Des études ont permis d'estimer la puissance de détection des différents protocoles en fonction de la structure des caractères et du nombre de descendants produits (Soller, Brody et Genisi, 1976 [98]; Tanksley *et al.*, 1982 [101]). Ces analyses sont basées sur des

tests de Student ou des analyses de variance, qui permettent de détecter des différences de moyennes pour le caractère quantitatif entre les classes d'individus ayant hérité de l'un ou l'autre allèle au marqueur génétique, considéré comme complètement lié au locus quantitatif. Les résultats obtenus sont résumés dans le tableau 1.1, pour des croisements entre lignées fixées pour les allèles au QTL. Par rapport à ces résultats, la puissance de détection diminue avec l'éloignement du locus marqueur par rapport au locus quantitatif. Par exemple, pour un taux de recombinaison de 0,1 entre les locus marqueur et quantitatif, le nombre de descendants F2 nécessaires pour obtenir la même puissance de détection est de 10462 (Soller, Brody et Genisi, 1976 [98]).

TAB. 1.1 - *Nombre de descendants nécessaires à la détection d'un QTL.*

	Degré de dominance				
	-d	-d/2	0	d/2	d
BC vers la race 1	525	934	2100	8400	*
BC vers la race 2	*	8400	2100	934	525
BC combinés	2100	2100	2100	2100	2100
F2	1050	1050	1050	1050	1050

Le QTL représente 1% de la variance phénotypique globale, pour une erreur de première espèce de 5% et une puissance de détection de 10%, avec un marqueur génétique lié au locus quantitatif. Le signe positif correspond à la dominance de l'allèle de la race 1 sur celui de la race 2. *= non détectable. *D'après Soller, Brody et Genisi, 1976.*

Par ailleurs, Elsen et Le Roy (1996) [18] ont comparé l'évolution de la puissance d'un test de lod-score en fonction du nombre de familles et de descendants par famille. Ils montrent que la prise en compte des contraintes familiales pour les espèces polythoques favorise le croisement F2 par rapport au back-cross. Par ailleurs, le nombre de descendants produits par mère (facteur limitant dans les espèces où l'insémination artificielle permet de maximiser le nombre de descendants par père) doit être maximal. Enfin, le nombre de familles de père (famille de demi-frères d'un même père F1) doit être au moins égal à quatre pour maximiser les chances d'avoir au moins un individu hétérozygote en F1.

Cependant, le choix du protocole de détection est avant tout contraint par des exigences économiques. En effet, si la majorité des allèles que l'on souhaite mettre en évidence appartient à la race sélectionnée dans un croisement entre races divergentes pour les performances, on pourra être tenté de ne faire que le croisement en retour sur l'autre race. Or les performances zootechniques des animaux issus de tels back-cross sont en général très détériorées par rapport à des animaux issus des populations de production. Leur coût d'obtention est donc très élevé, et on choisira préférentiellement de réaliser un croisement F2 pour limiter les pertes.

Les cohortes de données obtenues à partir des protocoles expérimentaux décrits dans

cette partie, ou de ceux évoqués plus haut pour la détection de QTL en populations *outbred*, doivent être analysées à l'aide de méthodes statistiques spécifiques que nous allons présenter dans la partie suivante.

3. Méthodologies de détection de QTL unicaractères et uniQTL

Des méthodes statistiques pour la détection de QTL ont été développées par la majorité des grandes écoles de statistique. Nous nous concentrerons ici sur les méthodes de type paramétriques dites "classiques", laissant de côté l'apport des statistiques bayésienne et non-paramétrique en particulier. Deux hypothèses sont en général sous-jacentes : la normalité des distributions des phénotypes pour le caractère analysé, et l'égalité des variances entre génotypes au QTL. Les méthodes décrites dans cette partie ont souvent été développées dans un premier temps pour des applications aux croisements entre lignées fixées génétiquement. Cependant, leur généralisation à l'analyse de populations présentant des structures familiales est relativement simple et les procédures particulières à mettre en oeuvre seront décrites.

3.1. Hypothèses comparées

Les méthodes décrites dans cette introduction correspondent aux modèles les plus simples. Ils visent à tester l'hypothèse de l'existence d'un QTL contre celle de l'absence de QTL agissant sur un caractère, pour une région chromosomique donnée. L'hypothèse nulle est donc "il n'y a pas de QTL déterminant le caractère", testée contre l'hypothèse alternative "il y a un QTL déterminant le caractère".

3.2. Détection marqueur par marqueur

Les premières statistiques de test développées pour la mise en évidence de coségrégation d'allèles aux locus marqueur et quantitatif étaient basées sur l'analyse d'un unique marqueur, éventuellement répétée de façon déconnectée pour plusieurs marqueurs sur le génome. Cette technique correspond à celle développée par Soller, Brody et Genisi (1976 [98]) évoquée au paragraphe 2.4.. Les individus des deux classes génotypiques au marqueur, déterminées par l'allèle hérité du parent hétérozygote, sont comparés pour leurs moyennes phénotypiques. La différence des moyennes sert d'estimation à un effet de substitution d'un allèle au QTL sur l'autre. La valeur estimée est alors comparée à zéro par

un test d'analyse de variance à un facteur sous l'hypothèse de normalité des distributions des variances.

Les méthodes de détection marqueur par marqueur peuvent être interprétées comme des régressions linéaires des phénotypes sur les génotypes. Si on note y_i le phénotype et g_i la variable indicatrice du génotype d'un individu i, qui prend les valeurs 0 ou 1, les deux éléments sont supposés liés par l'équation :

$$y_i = \mu + a/2 * g_i + e \tag{1.2}$$

où $a/2$ est l'estimation de l'effet phénotypique des allèles au QTL , e est l'erreur résiduelle du modèle, supposée normalement distribuée avec une moyenne 0 et une variance σ_e^2 à estimer, et μ est la moyenne des performances. Les paramètres à estimer sont μ, a et σ_e^2. Les solutions de cette régression linéaire sont aussi les estimations du maximum de vraisemblance (*maximum likelihood estimates* MLEs) pour ces paramètres $(\hat{\mu}, \hat{a}, \hat{\sigma_e}^2)$, c'est à dire les valeurs qui maximisent la probabilité $L(\mu, a, \sigma_e^2)$ d'observer les données si le modèle considéré est correct. Dans le cas où les structures familiales ne seraient pas prises en compte, un seul paramètre de chaque catégorie serait estimé pour toute la population de n descendants. La fonction de vraisemblance sous l'hypothèse de l'action d'un QTL sur le caractère s'écrit alors :

$$L(\mu, a, \sigma_r^2) = \prod_{i=1}^{n} f(y_i - (\mu + a g_i), \sigma_e^2) \tag{1.3}$$

où $f(y, \sigma^2)$ est la densité de probabilité de la distribution normale de moyenne nulle et de variance σ^2. Les MLEs sont comparés aux MLEs pour la vraisemblance contrainte sous l'hypothèse d'absence de QTL dans la région considérée, avec $a = 0$. Les MLEs sous cette hypothèse sont aisément estimables, et notés $(\hat{\hat{\mu}}, 0, \hat{\hat{\sigma}}_r^2)$. Le test de la présence du QTL lié au marqueur génétique peut alors se résumer dans un LOD score :

$$LOD = log\left(\frac{L(\hat{\mu}, \hat{a}, \hat{\sigma}_e^2)}{L(\hat{\hat{\mu}}, 0, \hat{\hat{\sigma}}_e^2)}\right) \tag{1.4}$$

ou dans un test de maximum de vraisemblance (*likelihood ratio test* LRT) :

$$LRT = -2ln\left(\frac{L(\hat{\mu}, \hat{a}, \hat{\sigma}_e^2)}{L(\hat{\hat{\mu}}, 0, \hat{\hat{\sigma}}_e^2)}\right) \tag{1.5}$$

Ces statistiques de test quantifient la différence de chances que les données soient telles qu'elles sont sous l'hypothèse de la présence d'un QTL par rapport à l'hypothèse

d'absence de QTL. Si la statistique de test dépasse un seuil de signification prédéterminé T, l'hypothèse d'absence de QTL est rejetée. Sous l'hypothèse d'une population de grande taille, le LRT est asymptotiquement distribué selon un χ^2 à 1 degré de liberté (Kendall et Stuart, 1979 [43]), pour le test de chaque marqueur considéré séparément. La valeur seuil pour une erreur de première espèce α associée au test est donc estimable par $T = (Z_\alpha)^2$, où Z_α est définie par l'équation $Probabilité(z > Z_\alpha) = \alpha$, avec z une variable gaussienne. Pour des recherches systématiques, Soller et Brody (1978 [99]) préconisent une erreur de première espèce de 5%, c'est à dire relativement élevée, pour maximiser les chances de détection, ce qui correspond à un seuil de 3.82 pour un LRT et 0.83 pour un LOD score. Dans un test statistique, cette erreur α correspond au pourcentage de faux-positifs que l'on accepte, c'est à dire le nombre de fois où il n'y a pas de QTL mais que l'on tolère d'en accepter un. Cette erreur diminue avec l'augmentation du nombre de fois où l'on accepte de ne pas détecter un QTL existant, souvent noté β, et appelée erreur de deuxième espèce. $1 - \beta$ correspond alors à la capacité de détection d'un QTL existant, appelée puissance.

Les méthodes de détection marqueur par marqueur sont très puissantes pour la détection de QTL localisés à la position du marqueur génétique utilisé. De plus, elles ne requièrent pas l'utilisation de cartes génétiques complètes. Cependant, si le QTL n'est pas localisé sur le locus marqueur, l'estimation de l'effet est biaisée d'un facteur (1-2r), où r est le taux de recombinaison entre les locus marqueur et quantitatif. De plus, la distance entre les locus marqueur et quantitatif est inversement proportionnelle à la capacité de détection du QTL, qui est alors réduite d'un facteur $(1 - 2r)^2$. Pour obtenir la même puissance que si les deux locus étaient strictement liés, il faut alors augmenter la taille de la population de ce même facteur. Enfin, la localisation du locus quantitatif n'est pas estimée, puisque l'effet et la position sont confondus dans une même estimation de $a(1 - 2r)$. D'autre part, les auteurs utilisant ces méthodes ne prennent en général pas en compte dans l'estimation de leurs valeurs de seuils de signification le fait que les tests sont répétés pour chaque marqueur analysé et pour chaque caractère étudié. Or chacun de ces tests est l'occasion de la détection d'un faux-positif. Cette question de la multiplicité des tests sera abordée dans le même temps que celle des confusions effet/position au paragraphe 3.4..

Cependant, de nombreux auteurs ont utilisé cette démarche avec succès pour localiser des gènes à effet quantitatif. Un exemple de ce type de détection est donné par Tanksley *et al.* en 1981 pour la détection de QTL sur un backcross entre deux lignées de tomates, à l'aide de 12 locus enzymatiques.

3.3. Cartographie d'intervalle

3.3.1. Principes

Lander et Botstein, en 1989 [62], ont les premiers proposé une méthode combinant l'information disponible sur plusieurs marqueurs pour localiser les QTL. Cette méthode, largement utilisée par la suite pour la détection de QTL, est appelée cartographie d'intervalle (*interval mapping*). Elle est basée sur l'exploitation des données issues du développement de cartes génétiques relativement denses pour la majorité des espèces d'intérêt. Les individus de la population sont supposés tous génotypés pour un réseau de marqueurs génétiques répartis sur le génome. Un ensemble de marqueurs génétiques ordonnés sur un chromosome forme un groupe de liaison.

La cartographie d'intervalle est alors basée sur l'exploitation des génotypes aux marqueurs génétiques flanquant une position donnée. En considérant deux marqueurs flanquant une position donnée sur le groupe de liaison analysé, la distance de chaque marqueur à la position testée est lue sur la carte génétique disponible puis convertie en taux de recombinaison en inversant la fonction de distance de la carte (Haldane, 1919 [27]; Kosambi, 1944 [60]). La probabilité que le descendant ait reçu l'un ou l'autre des allèles au QTL d'un parent à la position testée peut alors être estimée. En général, les auteurs font l'hypothèse d'absence d'interférence entre les locus. Si l'information est disponible sur les deux parents, ou le nombre de descendants important, les phases héritées de chacun des parents peuvent ainsi être reconstituées, au moins en probabilité (probabilités de transmission). Dans la pratique, la reconstitution des phases des parents est donc souvent préliminaire à la mise en oeuvre de la statistique de test.

Considérons les descendants d'un individu porteurs des phases orientées A1Q1B1 et A2Q2B2. La figure 6 représente la probabilité qu'un descendant présentant les allèles A1 et B1 aux locus marqueurs ait reçu Q1 au milieu de l'intervalle, en fonction de la distance entre les locus marqueurs θ et de la méthode utilisée (sans information, à l'aide d'un unique marqueur, ou à l'aide des deux marqueurs flanquants). Cette représentation permet d'illustrer le gain d'information apporté par la prise en compte conjointe de marqueurs flanquants pour la détection de QTL.

La cartographie d'intervalle permet de tester de façon systématique l'hypothèse de la ségrégation d'un QTL pour toutes les positions couvertes par le réseau de marqueurs génétiques. Dans la pratique, "toutes les positions" se traduit par un balayage du génome pour des pas de 1 cM par exemple. Différentes statistiques de test ont été utilisées, qui seront explicitées dans le paragraphe suivant. La valeur des statistiques de test peut être représentée en fonction des positions sur le groupe de liaison analysé. On obtient ainsi un profil de la statistique de test en fonction de la position, et la position du QTL la plus

probable est aisément identifiable.

3.3.2. Comparaison à la détection marqueur par marqueur

La cartographie d'intervalle nécessite l'analyse de données sur moins d'individus que l'analyse marqueur par marqueur. En effet, si r est le taux de recombinaison entre les deux marqueurs flanquants, on peut montrer (voir Lander et Botstein, 1989 [62]) que le nombre de descendants requis pour une cartographie d'intervalle est $(1 - r)$ fois plus faible que pour une cartographie marqueur par marqueur. Pour des densités en marqueurs génétiques de 10, 20, 30 ou 40 cM, cela correspond respectivement à des diminutions de 9, 16, 23 et 28% du nombre de descendants.

Pour des dispositifs équivalents, la cartographie d'intervalle ne permet pas d'augmenter significativement la puissance de détection par rapport à une détection marqueur par marqueur où tous les marqueurs sont testés individuellement pour des densités en marqueurs génétiques de 20 à 10 cM quand elle est appliquée à des croisements entre lignées génétiquement fixées. Quand la densité diminue, jusqu'à 50cM, des écarts de 6 à 8% de détection apparaissent pour la détection de QTL à effets relativement faibles (0,25 σ_p) (Darvasi *et al.*, 1993 [12]) quand le QTL est localisé au milieu de l'intervalle entre deux marqueurs génétiques. Cependant, Knott *et al.* (1996 [51]) ont montré que la limite de résolution ainsi mise en évidence par Darvasi *et al.* (1993 [12]) n'apparaît pas lors de l'analyse de familles de demi-frères issues de populations où les QTL sont en ségrégation. Par ailleurs, l'avantage majeur de la cartographie d'intervalle réside dans la distinction de l'estimation des effets du QTL et de celle de la position, ce qui conduit à des estimations beaucoup plus précises des deux paramètres (Knott et Haley (1992 [50])).

3.4. Statistiques de test

Les deux types de méthodes d'estimation des paramètres, maximum de vraisemblance et méthode des moindres carrés (*via* la régression linéaire, évoquées pour les analyses marqueur par marqueur, ont été utilisées pour la cartographie d'intervalle. Cependant, dans ce cas elles ne permettent pas d'obtenir exactement les mêmes types de résultats (Kao, 2000 [42]), et seront donc traitées successivement. Les développements présentés ici seront toujours limités au cas d'un croisement entre lignées fixées pour les allèles au QTL, donc permettant en particulier des analyses sans prise en compte de structures familiales des populations.

3.4.1. LRT et régression multiple

3.4.1.1. Likelihood ratio test L'association du test de maximum de vraisemblance à la cartographie d'intervalle a été proposée par Lander et Botstein en 1989 pour l'analyse d'une population backcross issue de lignées génétiquement fixées. Un seul paramètre génétique, a, est donc estimé. Elle a été rapidement reprise par Knapp *et al.* (1990 [44]) et Knapp (1991 [45]), qui ont linéarisé la vraisemblance, et appliquée par Paterson *et al.* (1991) pour la détection de QTL chez la tomate.

On appelle $G_i(0)$ et $G_i(1)$ les probabilités que l'individu i ait reçu respectivement les génotypes 0 (hétérozygotes) ou 1 (homozygote) de ses parents à la position testée conditionnellement aux génotypes et positions des marqueurs flanquants. De la même façon, on dérive de l'équation 1.3 $f_i(0)$ et $f_i(1)$ la fonction de pénétrance $f_i(x) = f(y_i - (\mu + ax), \sigma_r^2)$, où x=0 ou 1 en fonction du génotype considéré. La vraisemblance sous l'hypothèse de la présence d'un QTL à la position testée est alors calculée par :

$$L(\mu, a, \sigma_r^2) = \prod_{i=1}^{n} [G_i(0)f_i(0) + G_i(1)f_i(1)] \tag{1.6}$$

Les MLEs $(\mu*, a*, \sigma_r*)$ correspondant à cette vraisemblance doivent être estimés en utilisant des procédures de maximisation spécifiques, telles que des algorithmes de Newton ou des algorithmes EM, car les solutions analytiques ne sont pas accessibles. Divers algorithmes ont été proposés avec le développement de la cartographie d'intervalle associée au LRT, qui, n'étant pas l'objet de l'étude réalisée, ne seront pas détaillées dans ce document (Jansen, 1992 [34]; Zeng, 1994 [115]).

Sous l'hypothèse nulle, où $a = 0$, les paramètres peuvent, dans le cas de croisements entre lignées fixées, être directement extrapolables de la distribution des performances de la génération F1.

La maximisation de la vraisemblance associée à la cartographie d'intervalle présente de nombreux avantages par rapport aux méthodes marqueur par marqueur. En particulier, les estimations des effets des allèles au QTL sont asymptotiquement non biaisées. Par ailleurs, la position la plus probable du QTL est accessible, et peut être associée à un intervalle de confiance.

La principale limite des méthodes basées sur le LRT est le développement de logiciels spécifiques requis pour la maximisation de la vraisemblance. De plus, les méthodes de maximisation sont en général demandeuses en terme de temps de calcul, ce qui peut être limitant pour des analyses systématiques.

3.4.1.2. Régression multiple Haley et Knott (1992 [28]) et Martinez et Curnow (1992 [80]) ont proposé, pour contrecarrer les limites algorithmiques des LRT, d'effectuer des régressions multiples pour détecter les QTL en cartographie d'intervalle. Cette approche est identique à la linéarisation de la vraisemblance telle que proposée par Knapp *et al.* (1990 [44]) et Knapp (1991 [45]). Ils définissent A et B les marqueurs flanquant la position testée, supposée correspondre à un QTL Q. Les développements sont réalisés pour un croisement F2 entre deux lignées fixées 1 et 2 pour les locus marqueurs et quantitatifs. On note $A_jQ_jB_j$ les haplotypes apportés par chaque lignée j pour les locus considérés. Le modèle de description des performances prend en compte la dominance. On définit $\mu + a$ la moyenne des performances des individus Q_1Q_1, $\mu + d$ la moyenne des performances des individus Q_1Q_2 et Q_2Q_1 et $\mu - a$ la moyenne des performances des individus Q_2Q_2, où μ est la moyenne des performances des homozygotes, a et d sont respectivement les déviations additives et dominantes dues aux effets des allèles au QTL. De plus, on définit r_A et r_B les taux de recombinaison respectivement entre A et Q, et B et Q. Le taux de recombinaison entre A et B est supposé connu et égal à r. Les auteurs supposent l'absence d'interférence entre les locus, ce qui permet d'établir $r = r_A + r_B - 2r_Ar_B$.

Les fréquences attendues des haplotypes pour la combinaison des trois locus en F2 sont facilement extrapolables à partir des notations décrites au paragraphe précédent. Les espérances des performances moyennes pour chaque génotype observé aux marqueurs peuvent donc être analytiquement décrites. Elles sont résumées dans le tableau 1.2 pour un croisement F2. Ces espérances peuvent être utilisées pour ajuster a, d et μ par régression multiple à chaque position séparément sous l'hypothèse de l'existence d'un QTL. Cette stratégie permet naturellement d'estimer a et d, mais aussi de calculer les sommes des carrés des résidus et des valeurs régressées et les carrés moyens correspondants. On peut ainsi calculer un test F pour tester les valeurs de a et d. La position la plus probable du QTL correspond à celle du plus petit carré moyen des résidus et aux meilleures estimations des paramètres.

Toutes les étapes décrites précédemment peuvent être réalisées à l'aide de logiciels de statistiques classiques et ne demandent pas de développements particuliers en programmation.

3.4.1.3. Comparaison des deux approches Haley et Knott (1992 [28]) montrent que dans les cas simples qu'ils ont analysés, les estimations des paramètres sont très similaires entre le test de rapport de vraisemblance et la régression multiple. Les corrélations les plus faibles entre les estimations par les deux méthodes correspondent à la détection de QTL ayant des effets faibles ou inexistants. Cependant, elles sont en général de l'ordre de 0,99 pour les estimations des paramètres et toujours supérieures à 0,96 pour les valeurs de statistique de test.

TAB. 1.2 - *Espérance des coefficients des composantes de a et d pour toutes les combinaisons de génotypes possibles dans une population F2.*
D'après Haley et Knott, 1992.

Génotype aux marqueurs génétiques	a	d
$A_1A_1B_1B_1$	$\frac{(1-r_A)^2(1-r_B)^2-r_A^2r_B^2}{(1-r)^2}$	$\frac{2r_A(1-r_A)r_B(1-r_B)}{(1-r)^2}$
$A_1A_1B_1B_2$	$\frac{(1-r_A)^2r_B(1-r_B)-r_A^2r_B(1-r_B)}{r(1-r)}$	$\frac{r_A(1-r_A)(1-r_B)^2+r_A(1-r_A)r_B^2}{r(1-r^2)}$
$A_1A_1B_2B_2$	$\frac{(1-r_A)^2r_B^2-r_A^2(1-r_B)^2}{r^2}$	$\frac{2r_A(1-r_A)r_B(1-r_B)}{r^2}$
$A_1A_2B_1B_1$	$\frac{r_A(1-r_A)(1-r_B)^2-r_A(1-r_A)r_B^2}{r(1-r)}$	$\frac{(1-r_A)^2r_B(1-r_B)+r_A^2r_B(1-r_B)}{r(1-r)}$
$A_1A_2B_1B_2$	0	$\frac{r_A^2r_B^2+r_A^2(1-r_B)^2+(1-r_A)^2r_B^2+(1-r_A)^2(1-r_B)^2}{r^2+(1-r)^2}$
$A_1A_2B_2B_2$	$\frac{r_A^2(1-r_A)r_B^2-r_A(1-r_A)(1-r_B)^2}{r(1-r)}$	$\frac{(1-r_A)^2r_B(1-r_B)+r_A^2r_B(1-r_B)}{r(1-r)}$
$A_2A_2B_1B_1$	$\frac{r_A^2(1-r_B)^2-(1-r_A)^2r_B^2}{r^2}$	$\frac{2r_A(1-r_A)r_B(1-r_B)}{r^2}$
$A_2A_2B_2B_2$	$\frac{r_A^2r_B^2-(1-r_A)^2(1-r_B)^2}{(1-r)^2}$	$\frac{2r_A(1-r_A)r_B(1-r_B)}{r(1-r)}$

Rebaï *et al.* (1995 [91]) ont par ailleurs montré analytiquement sur des croisements entre lignées *inbred* que les méthodes simplifiées, par la linéarisation (Knapp *et al.*, 1990 [44]) ou la régression multiple (Haley et Knott, 1992 [28]), sont asymptotiquement équivalentes à l'utilisation du test de maximum de vraisemblance de Lander et Botstein (1989 [62]) pour l'analyse d'un unique intervalle. Les auteurs utilisent alors l'approche de Davies (1977 [13]) pour obtenir les approximations des seuils du LRT pour un back-cross.

Cependant, Kao (2000 [42]) a montré que les deux méthodes ont des comportements différents dans certaines situations, quand les effets du (ou des) QTL sont élevés, que les QTL sont éloignés de leurs marqueurs flanquants, ou que plusieurs QTL interagissent.

Dans la suite de ce document, sauf précision particulière, les éléments développés correspondent toujours à des méthodologies de cartographie d'intervalle.

3.4.2. Distribution des statistiques de test

Les deux méthodes statistiques décrites aux paragraphes précédents sont bien connues. Les distributions asymptotiques des statistiques de test sous l'hypothèse nulle ont été largement décrites, mais sont soumises à la validation d'hypothèses parfois difficilement compatibles avec la détection de QTL.

3.4.2.1. Analyse d'un intervalle Considérons dans un premier temps un test unique pour une position donnée. Nous avons vu que la distribution du LRT où p paramètres sont fixés sous l'hypothèse particulière testée peut asymptotiquement être approchée par

un χ^2 à p degrés de liberté (Wright, 1968 [111]). Une distribution asymptotique pour la régression multiple peut de la même façon être extrapolée, sous l'hypothèse d'erreurs indépendantes et normalement distribuées. Un test de rapport de vraisemblance peut être construit sur la base des sommes des carrés des résidus (*residual sum of squares*, RSS) pour le modèle complet (RSS_{comp}) et pour le modèle réduit (sous l'hypothèse d'absence de QTL en l'occurrence : RSS_{red}) et du nombre n d'observations intégrées à l'analyse : $LRT = nlog\left(\frac{RSS_{red}}{RSS_{comp}}\right)$ (Aitkin *et al.*, 1989 [1]). Ces approximations sont soumises à l'hypothèse de distributions normales des erreurs intra-classes génotypiques au QTL égales entre les classes, ce qui n'est pas vérifié aux positions des marqueurs génétiques. Cependant, ces distributions sont proches des valeurs de statistiques de test obtenues expérimentalement (Haley et Knott, 1992 [28]).

Les approximations décrites ci-dessus correspondent aux analyses nominales de chaque position. Cependant, dans la pratique, de nombreux tests sont réalisés sur des positions très corrélées et sur de nombreux caractères. Des auteurs ont donc proposé des corrections des approximations différentes des valeurs de seuil pour prendre en compte la multiplicité des tests, en particulier quant à la définition de l'erreur de première espèce à utiliser.

3.4.2.2. Extension à l'analyse d'un chromosome

Cas extrêmes Lander et Botstein (1989 [62]) ont étudié deux cas extrêmes pour l'analyse d'un backcross : une densité élevée de marqueurs génétiques sur le génome (la corrélation entre deux tests consécutifs est de 1), et une densité très faible (les marqueurs sont tous indépendants). Dans le premier cas, ils montrent que le test varie selon le carré d'un processus de Orenstein-Uhlenbeck (Leadbetter *et al.*, 1983, [63]). Dans le deuxième cas, ils proposent une approximation de correction de Bonferroni pour des tests indépendants, qui revient à calculer la valeur seuil comme pour un test nominal, mais pour une erreur de première espèce k fois inférieure, où k est le nombre de répétitions du test.

Détermination empirique des statistiques de test Lander et Botstein (1989 [62]) proposent par ailleurs, pour des densités intermédiaires, qui correspondent aux cas rencontrés dans la pratique, ou l'analyse d'autres types de population, de réaliser des simulations de données pour obtenir une distribution empirique de la statistique de test sous l'hypothèse nulle. Le principe de cette méthode est simple : la valeur seuil utilisée pour déterminer le niveau de signification d'un test correspond à une erreur de première espèce α préalablement choisie. Le seuil correspond donc à la valeur de la distribution de la statistique de test sous l'hypothèse nulle au delà de laquelle la probabilité qu'il n'y ait pas de QTL impliqué dans le déterminisme du caractère est de $\alpha\%$. La répétition du test sur un nombre élevé de situations où il n'y a pas de QTL permet d'obtenir cette

distribution, et la valeur seuil est le quantile à $\alpha\%$ de la distribution obtenue.

Lander et Botstein (1989 [62]) proposent de simuler des données sous le modèle polygénique pour obtenir cette distribution. Cette stratégie suppose que la distribution des performances du caractère suive effectivement celle simulée, par exemple une distribution normale. Dans le cas contraire, la valeur seuil déterminée n'est pas correcte. Pour pallier à cette limite, Churchill et Doerge, en 1994 [11], ont proposé de permuter les valeurs des performances par rapport aux génotypes, afin de casser les liaisons entre allèles au QTL et expression des performances. Cette opération est répétée pour chaque analyse, et les valeurs de statistiques de test conservées et traitées de la même façon que précédemment pour déterminer la valeur seuil. Les auteurs ont validé cette stratégie par la comparaison de valeurs seuils obtenues par permutations à des valeurs seuils extrapolées de test du χ^2. Elle a par ailleurs été validée plus complètement par Doerge et Churchill (1996 [16]) qui la comparent à des approches différentes. Cette technique présente l'avantage d'être complètement non-paramétrique, et donc applicable à tous les types de données. Cependant, elle implique l'utilisation de populations de taille importante pour que les permutations permettent de balayer la plus largement possible la distribution sous l'hypothèse nulle. Dans son application à des populations à structures familiales, les permutations doivent être réalisées intra-famille pour conserver la structure généalogique étudiée, ce qui implique la disponibilité de familles de tailles importantes. Ces deux méthodes, si elles sont théoriquement bien validées, sont très demandeuses en terme de temps de calcul puisqu'elles impliquent l'estimation répétée 1000 à 10000 fois des statistiques de test. Certains auteurs se sont donc intéressés au développement d'approximations des valeurs de seuil.

Approximations des seuils Rebaï *et al.* (1994 [90]) ont généralisé leur approche développée pour l'analyse d'un unique intervalle et d'un back-cross (Rebaï *et al.*, 1995 [91]) à l'analyse d'un chromosome pour d'autres types de population. Ils comparent leurs développements analytiques à des résultats de 10000 simulations pour des erreurs de première espèce de 1 et 5%. Le problème de la généralisation de la méthode au chromosome réside dans la non-linéarité de la dérivée de la fonction utilisée à la position des marqueurs. Cependant, l'analyse de Davies (1987, [14]) montre que les approximations utilisées sont correctes si le nombre de sauts de linéarité est fini. Leurs approximations ne sont donc applicables que pour des densités en marqueurs génétiques raisonnables (de 3 à 26 marqueurs génétiques sur 100cM). Pour l'extension de la méthode à des populations où plusieurs paramètres doivent être estimés, les résultats montrent que les approximations sont utilisables pour des densités en marqueurs génétiques de 5 à 20cM. Cependant, quand le nombre de paramètres à estimer augmente, les approximations deviennent difficilement calculables analytiquement. Par ailleurs, elles sont limitées à l'analyse de populations issues de croisements entre lignées génétiquement fixées de taille supérieure à 200 individus,

et leur extrapolation à des analyses de structures familiales n'a pas été réalisée.

D'une façon générale, les approximations décrites ci-dessus correspondent à l'analyse de données issues de croisements entre lignées génétiquement fixées. Dans le cas de l'application à des populations présentant des structures familiales, les approximations deviennent d'autant plus compliquées que les familles sont différentes, en fonction du nombre de descendants par père et / ou par mère par exemple. Dans la pratique, les auteurs qui analysent de telles données estiment les seuils par permutation quand les tailles de famille sont suffisamment importantes, ou bien par simulations des données sous l'hypothèse nulle.

3.4.2.3. Niveaux de signification L'extension des valeurs de seuils à l'analyse de plusieurs chromosomes, ou du génome entier, est aisément extrapolable à partir de ceux obtenus pour H chromosomes individuels indépendants par des corrections de Bonferroni. Ce type de correction consiste à calculer pour chaque chromosome un seuil correspondant à une erreur de première espèce α_h tel que $\alpha_h = 1 - (1 - \alpha)^{1/H}$, où α est l'erreur de première espèce recherchée pour l'ensemble de l'analyse. En général, le calcul des α_h peut être approximé par $\alpha_h = \alpha/H$ étant données les valeurs de α utilisées. Cependant, certains auteurs (Rebaï *et al.*, 1994 [90]) envisagent d'adapter l'erreur de première espèce à la taille du chromosome analysé, c'est à dire au nombre de tests réalisés par chromosome.

Concernant les corrections pour le nombre de caractères analysés, qui sont en général corrélés au moins par groupe de caractères, les auteurs effectuent en général lors d'applications à des données réelles une correction de type Bonferroni. Le facteur de correction utilisé est le nombre de caractères indépendants qui devrait être retenu pour expliquer une fraction maximale (de l'ordre de 95%) de la variabilité des performances. Ce facteur peut être calculé sur la base d'une analyse en composantes principales réalisée sur la matrice de (co)variance phénotypique ou bien sur la matrice de (co)variance génétique entre les caractères connue par ailleurs. Bidanel *et al.* (2000) ont ainsi considéré 40 variables indépendantes lors de l'analyse de 92 caractères.

Etant données les faibles valeurs des erreurs de première espèce utilisées suite à ces corrections, les valeurs de seuils sont en général élevées, et les QTL d'autant plus difficiles à détecter. Lander et Kruglyak, en 1995 [61], ont proposé une nomenclature qui permette de valider la présence de QTL pour différents niveaux de signification des statistiques de test. Bien que leur nomenclature ait été développée pour une application à des analyses de pedigree humains, elle est aisément extrapolable à tout type de pedigree. Ils se basent sur la différence entre la notion de seuil nominal, défini pour l'analyse d'une position, et la notion de seuil au niveau du chromosome, qui implique la prise en compte de l'occurrence

possible de faux-positifs à toutes les positions testées. Ils distinguent ainsi :

1. La liaison suggérée (*suggestive linkage*), qui correspond à une espérance de l'occurrence d'une détection par chance pendant le balayage du génome.

2. La liaison significative (*significant linkage*), qui correspond à une espérance de l'occurrence de 0,05 détection par chance pendant le balayage du génome, c'est à dire une erreur de première espèce globale de 5%.

3. La liaison hautement significative (*highly significant linkage*), qui correspond à une espérance de l'occurrence de 0,001 détection par chance pendant le balayage du génome, c'est à dire une erreur de première espèce de 0,1%.

4. La liaison confirmée (*confirmed linkage*), qui correspond à une liaison significative pour une étude ou une combinaison d'études confirmée par une étude indépendante. Pour la confirmation, les auteurs préconisent une probabilité critique nominale de 0,01.

A titre d'illustration, le tableau 1.3 présente les probabilités critiques pour des liaisons significatives ou suggérées pour un lod score réalisé sur des croisements entre lignées fixées de souris ou rat (taille de génome considérée : 1600cM) pour différents types de croisements.

TAB. 1.3 - *Probabilités critiques (p_c) et valeurs seuils de lod pour des liaisons suggérées ou significatives.*
Croisements entre lignées génétiquement fixées de souris ou de rat. ddl= degrés de liberté. *D'après Lander et Kruglyak, 1995*

	Taux de	Liaison suggérée		Liaison significative	
	recombinaison	p_c x 10^{-3}	seuil lod	p_c x 10^{-4}	seuil lod
BC (1 ddl)	1	3,4	1,9	1,0	3,3
F2 (1ddl, additif)	1	3,4	1,9	1,0	3,3
F2 (1ddl, récessif)	4/3	2,4	2,0	7,2	3,4
F2 (1ddl, dominant)	4/3	2,4	2,0	7,2	3,4
F2 (2ddl)	1,5	1,6	2,8	5,2 x 10^{-1}	4,3

Les auteurs conseillent d'appliquer la nomenclature présentée quelle que soit la taille de la région chromosomique analysée, de l'intervalle au génome entier, considérant le risque de fausses détections implicites que constituent les régions non analysées.

3.5. Propriétés des méthodes de cartographie unicaractère uni-QTL

3.5.1. Croisement entre lignées fixées

Les capacités de détection de QTL des méthodes décrites dans les paragraphes précédents ont été explorées par Haley et Knott (1992) et Darvasi *et al.* (1993 [12]). Haley et Knott (1992) comparent les méthodes basées sur la régression et le LRT appliquées à une population F2 en fonction de la densité en marqueurs génétiques, de la distance entre le QTL et le marqueur génétique le plus proche et de l'effet du QTL sur le caractère. Darvasi *et al.* (1993) comparent dans un premier temps les performances de détection du LRT avec cartographie d'intervalle au test-t réalisé marqueur par marqueur sur les mêmes données. Dans un deuxième temps, ils s'intéressent aux performances de détection des QTL par le LRT, en terme de puissance de détection et de précision d'estimation des paramètres. Cette étude est restreinte à l'analyse de données issues d'un back-cross entre deux lignées fixées. Les auteurs caractérisent les variations de puissance de détection et de précision des paramètres du QTL en fonction de la taille de la population utilisée, de la densité en marqueurs génétiques, de l'emplacement du QTL dans l'intervalle génétique considéré et de la valeur de l'effet de substitution du QTL.

De façon attendue, l'ensemble des performances des méthodes est conditionné par la quantité d'information disponible dans le protocole. Les performances sont donc meilleures quand la densité en marqueurs génétiques, le nombre de descendants ou l'effet du QTL augmentent, ou que le QTL est localisé sur un des marqueurs génétiques.

En général, pour des croisements entre lignées génétiquement fixées, les puissances observées sont supérieures à 50% pour un effet de substitution de 0,25 et des densités en marqueurs génétiques de 50 à 10cM. Darvasi *et al.* (1993) montrent que l'écart de puissances entre une densité supposée infinie (carte génétique saturée en marqueurs) et 20 cM n'est pas très élevé, voire inexistant. En revanche, passer de 500 à 1000 individus dans la population augmente la puissance de 25 à 30% en moyenne.

En ce qui concerne les estimations des paramètres, les MLEs sont non-biaisés, bien que la cartographie d'intervalle ait tendance à localiser le QTL sur le marqueur génétique le plus proche. La précision d'estimation des effets est en général meilleure que celle de la position.

Darvasi *et al.* (1993) ont aussi comparé les intervalles de confiance de localisation des QTL en fonction des situations. Ce paramètre est essentiellement affecté par la taille de la population et l'effet du QTL sur le caractère. Pour une population de 500 individus et un effet de 0,25 σ_p, avec une densité en marqueurs génétiques infinie, l'intervalle de confiance est de 90 cM, pour un groupe de liaison de 100 cM. Si l'effet est de 0,5 σ_p et

la population de 1000 individus, la taille de l'intervalle de confiance diminue jusqu'à 11 cM. La densité en marqueurs génétiques, pour sa part, ne réduit la taille de l'intervalle de confiance que jusqu'à une limite irréductible que l'on peut estimer en fonction de l'effet du QTL et de la taille de la population avec une densité infinie en marqueurs. Ce paramètre est aussi très influencé par la place du QTL sur le groupe de liaison, et en particulier sa proximité avec les extrémités du groupe de liaison et sa localisation dans l'intervalle entre deux marqueurs flanquants.

3.5.2. Populations à structures familiales

Knott *et al.* (1996 [51]) et Mangin *et al.* (1999 [77]) ont étudié dans le même esprit les performances de détection du LRT appliqué à des populations à structure familiale de type familles de demi-frères. L'application des stratégies développées pour l'analyse de populations issues de croisements entre lignées génétiquement fixées est relativement simple si l'on considère que les parents sont non-apparentés d'une famille à l'autre. La fonction de pénétrance est alors dépendante de paramètres différents pour chaque famille considérée, et les paramètres sont estimés intra-famille (voir 4.1. pour l'écriture complète de la vraisemblance). Les comparaisons dépendent du nombre de descendants par père de chaque famille, de la densité en marqueurs génétiques, de la position du QTL et de ses effets sur un caractère donné. En comparaison avec les croisements entre lignées fixées, les puissances sont globalement nettement moins élevées.

En ce qui concerne l'estimation de la position, un biais apparaît vers le milieu du chromosome, d'autant plus important que le QTL est éloigné du centre du chromosome. Ce biais augmente avec la diminution de la puissance de détection. En revanche, les estimations des effets ne sont pas biaisées, ou sont légèrement sous-estimées, et leurs variances d'estimations relativement faibles.

Par ailleurs, Goffinet *et al.* (1999 [26]) démontrent l'importance de la prise en compte de l'hétéroscédasticité dans le modèle quand elle existe. La différence de variance résiduelle en fonction des familles peut être due à la ségrégation de QTL autres que celui testé mais qui détermine le même caractère quantitatif. Les variances résiduelles de ce caractère pour chaque famille de père dépendent alors du génotype du père pour ces autres QTL. Dans le cadre de l'analyse de populations dans lesquelles les QTL sont en ségrégation, la prise en compte de l'hétéroscédasticité permet d'atteindre une plus grande robustesse.

3.5.3. Bilan

En résumé, la majorité des méthodes de détection de QTL disponibles à l'origine de ce travail sont basées sur la cartographie d'intervalle. Elles reposent sur la détection d'un

QTL par groupe de liaison et des analyses caractère par caractère. Or, comme nous l'avons vu dans l'introduction, la détection de QTL est essentiellement basée sur la minimisation de la variance résiduelle du modèle de détermination du caractère. On peut aisément supposer que la détermination d'un caractère dépend de plusieurs QTL, qui peuvent agir sur d'autres caractères. L'utilisation des modèles simples décrits ci-dessus néglige donc une partie de l'information potentiellement disponible. Certaines études préalables à ce travail ont en effet permis de montrer que la prise en compte de la variabilité du caractère due à d'autres portions chromosomiques dans les modèles (Rodolphe et Lefort, 1993), ou l'utilisation de l'information venant de caractères corrélés (Jiang et Zeng, 1995; Korol *et al.*, 1995), peuvent permettre d'améliorer nettement les performances des méthodes de détection de QTL. De plus, étant basée sur des modèles de détermination des caractères moins simplistes, elles peuvent permettre de mettre en évidence l'architecture génétique déterminant les variations d'un ou plusieurs caractères.

Cependant, ces travaux concernent essentiellement l'analyse de données issues de protocoles de croisement entre lignées génétiquement fixées. Comme nous l'avons vu au paragraphe précédent, l'analyse de données issues de populations dans lesquelles les allèles au QTL sont en ségrégation implique l'estimation des paramètres intra-famille. L'utilisation de modèles plus complexes multiplie encore le nombre de ces paramètres. Le but de ce travail de thèse est de développer des méthodologies adaptées à ces structures et de les comparer à des méthodologies existantes afin de prendre en compte dans les modèles de détection l'information venant de plusieurs caractères (méthodes multicaractères), ou de positions liées (méthodes multiQTL).

Les méthodologies développées ont été comparées, sur la base de données simulées, pour leur puissance de détection et leur précision d'estimation des paramètres du QTL. Nous nous sommes appuyés sur le logiciel de détection de QTL QTLMAP qui permet l'analyse unicaractère et uniQTL dans le cas de mélanges de familles de demi-frères et de plein-frères préalablement développé à l'INRA (Le Roy *et al.*, 1998). Les méthodes multidimensionnelles développées ont ensuite été utilisées à titre d'illustration sur des données réelles issues du protocole INRA PORQTL de détection de QTL chez le porc, mené de 1991 à 1999 (Bidanel *et al.*, 2000).

La méthode unicaractère uniQTL ainsi que les données réelles utilisées sont présentées dans la partie à venir.

4. Outils disponibles pour la détection de QTL

4.1. Méthode unicaractère utilisée

La méthode unicaractère uniQTL utilisée a été décrite par Le Roy *et al.* (1998, [69]). Il s'agit d'un test de maximum de vraisemblance, avec cartographie d'intervalle tel que décrit par Lander et Botstein (1989 [62]), considérant un mélange de demi et plein-frères dans les familles. La population est constituée de n familles de père (i=1,...,n), avec n_i accouplements par mâle i (j=1,...,n_i), et n_{ij} descendants par femelle ij (k=1,...,n_{ij}). Nous avons conservé ici certaines hypothèses et simplifications décrites par Elsen *et al.* (1999 [19]), Goffinet *et al.* (1999 [26]) et Mangin *et al.* (1999 [77]) qui permettent d'améliorer la robustesse et les temps de calcul du processus. En particulier, à l'issue de la reconstitution des phases des parents F1, seul le génotype (y compris la phase) le plus probable des pères est retenu. De plus, les variances résiduelles sont estimées intra-familles de père et la vraisemblance est linéarisée pour les familles de plein-frères. En reprenant les notations de Elsen *et al.* (1999 [19]), adaptées à chaque caractère l (l=1,...,p) analysé, la fonction de vraisemblance linéarisée intra-famille de plein-frères à la position x s'écrit :

$$\Lambda_l^x = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \sum_{hd_{ij}} p(hd_{ij}/\widehat{hs_i}, M_i) \prod_{k=1}^{n_{ij}} f(yp_{ijkl}/\widehat{hs_i}, hd_{ij}, M_i) \tag{1.7}$$

où M_i est l'information marqueur pour la famille du père i, $\widehat{hs_i}$ est le génotype aux marqueurs génétiques le plus probable pour le père i, hd_{ij} est le génotype aux marqueurs génétiques pour la mère ij (dans la pratique, seuls les génotypes hd_{ij} dont la probabilité de transmission est supérieure à 0,1 sont pris en compte), yp_{ijkl} est le phénotype du descendant ijk pour le caractère l.

La fonction de pénétrance f, supposée normale, s'écrit :

$$f(yp_{ijkl}/\widehat{hs_i}, hd_{ij}, M_i) = \frac{1}{\sqrt{2\pi}\sigma_{il}} \exp\left(-\frac{1}{2}\left(\frac{Y_{ijkl}}{\sigma_{il}}\right)^2\right) \tag{1.8}$$

où σ_{il}^2 est la variance résiduelle intra-famille de père i pour le caractère quantitatif l, et

$$Y_{ijkl} = yp_{ijkl} - \sum_{q_s=1}^{2} \sum_{q_d=1}^{2} p\left(d_{ijk}^x = (q_s, q_d)/\widehat{hs_i}, hd_{ij}, M_i\right) \left(\mu_{il}^{xq_s} + \mu_{ijl}^{xq_d}\right) \tag{1.9}$$

où $p\left(d_{ijk}^x = (q_s, q_d)/\widehat{hs_i}, hd_{ij}, M_i\right)$ est la probabilité de transmission du couple d'haplotypes q_s et q_d, *i.e.* la probabilité que le descendant ijk ait reçu de son père i le segment

chromosomique q_s à la position testée x (q_s=1 venant du grand-père paternel, q_s=2 venant de la grand-mère paternelle) et qu'il ait reçu de sa mère ij le segment chromosomique q_d à la position testée x (q_d=1 venant du grand-père maternel, q_d=2 venant de la grand-mère maternelle), $\mu_{il}^{xq_s}$ est la moyenne des performances des descendants ayant reçu la fraction chromosomique q_s à la position x du père i pour le caractère quantitatif l, $\mu_{ijl}^{xq_d}$ est la moyenne des performances des descendants ayant reçu la fraction chromosomique q_d à la position x de la mère ij pour le caractère quantitatif l.

μ_{il}, μ_{ijl}, α_{il}^x et α_{ijl}^x étant respectivement les moyennes et les effets de substitution moyens du QTL estimés intra-familles de père i et de mère ij pour le caractère quantitatif l , $\mu_{il}^{xq_s}$ et $\mu_{ijl}^{xq_d}$ s'écrivent : $\mu_{il}^{x1} = \mu_{il} + \alpha_{il}^x/2$, $\mu_{il}^{x2} = \mu_{il} - \alpha_{il}^x/2$, $\mu_{ijl}^{x1} = \mu_{ijl} + \alpha_{ijl}^x/2$ et $\mu_{ijl}^{x2} = \mu_{ijl} - \alpha_{ijl}^x/2$.

Pour chaque caractère l, trois paramètres (variance intra-famille, moyenne et effet de substitution du QTL) par mâle, et deux par femelle (moyenne et effet de substitution du QTL) sont estimés, soit $3n+2\sum_{i=1}^{n} n_i$ paramètres. Les variances résiduelles sont supposées indépendantes du génotype au QTL.

La maximisation de la vraisemblance est réalisée à l'aide de la procédure e04jyf de NAG, qui minimise l'opposé de la vraisemblance par un algorithme de quasi Newton.

4.2. Données disponibles

4.2.1. Animaux

Le protocole PORQTL de détection de QTL chez le porc a débuté en 1991 avec la mise en place d'un dispositif de croisement entre les races Meishan et Large White au domaine INRA du Magneraud. Les caractéristiques de reproduction de l'espèce porcine (espèce multipare avec un intervalle de génération court) permettent de mettre en place des protocoles expérimentaux sur trois générations avec une puissance de détection correcte et dans un intervalle de temps minimum. Afin de détecter un maximum de QTL sur l'ensemble du génome pour des caractères de production et de reproduction, et en raison des contraintes liées à la production d'individus 3/4 Meishan très éloignés des standards du marché français, le protocole choisi est de type F2 .

Les individus (F0) à l'origine du croisement sont phénotypiquement divergents pour la majorité des caractères d'intérêt, ce qui permet de maximiser *a priori* le déséquilibre de liaison et le taux d'hétérozygotie entre les individus porteurs d'allèles différents aux QTL. Les 6 mâles F0 sont de race Large White, couramment utilisée en Europe pour la production, et qui exprime de bonnes performances pour les caractères de production. Ils sont non-apparentés. Les 6 femelles F0 sont de race Meishan, race chinoise présentant

des performances de reproduction nettement supérieures aux races européennes et des performances médiocres pour les caractères de production. Elles sont peu apparentées. Chaque couple F0 a permis de produire 1 mâle et 4 femelles F1 en moyenne. Chacun de ces mâles a été croisé avec 3 ou 4 femelles issus d'autres couples F0, fondant ainsi des familles de père. Les croisements ont été réalisés de façon à minimiser la parenté entre les individus. Un total de 1103 F2 a été produit, sur lesquels les performances ont été mesurées (Bidanel et *al.*, 2001 [6]).

Avec la méthode unicaractère uniQTL précédemment décrite, il faut donc estimer 3*6+2*23 paramètres, soit 64 paramètres à chaque position et pour chaque caractère analysé.

4.2.2. Marqueurs génétiques

L'ensemble des animaux a été génotypé pour un réseau de 137 marqueurs. Ces marqueurs sont répartis sur le génome tous les 27 centiMorgans (cM) en moyenne (Bidanel et *al.*, 2001, [6]). Cent vingt trois marqueurs microsatellites ont été génotypés pour l'ensemble des individus, plus 13 utilisés uniquement pour certaines familles non-informatives aux marqueurs préalablement choisis dans des régions chromosomiques intéressantes. Le dernier marqueur est un marqueur du complexe majeur d'histocompatilité (SLA). Les marqueurs ont été choisis de façon à maximiser l'informativité du dispositif, en terme de position et d'hétérozygotie, ainsi que pour la qualité et la reproductibilité de leurs profils d'analyse.

Le réseau ainsi développé permet de couvrir les 18 autosomes et le chromosome X, avec de 3 (SCC 18) à 12 (SCC 7) marqueurs par chromosome. La répartition exacte des marqueurs sur le génome porcin est explicitée dans Bidanel *et al.* (2001).

4.2.3. Caractères mesurés

Les performances ont été mesurées pour 92 caractères sur l'ensemble de la génération F2. Ces caractères concernent la croissance, la qualité de carcasse, et la reproduction (mâle et femelle). La liste des caractères disponibles et une première étude réalisée avec le logiciel décrit au paragraph 4.1. sont données dans Bidanel *et al.* (2000 [5]).

4.3. Illustration choisie

Les méthodes multidimensionnelles développées dans ce travail ont été appliquées à l'analyse de 5 caractères de composition corporelle pour lesquels un QTL avait été mis en évidence sur le chromosome 7 porcin.

4.3.1 Données disponibles

4.3.1.1. Carte et marqueurs génétiques Au moment de cette étude, 10 marqueurs génétiques étaient disponibles sur le chromosome 7, possédant 7,1 allèles en moyenne, dont le SLA qui en a 11. Leur répartition sur le chromosome 7 est résumée dans la figure 7.

4.3.1.2. Caractères Les 5 caractères choisis sont les poids de panne (panne) et de bardière (bardière) à la découpe, les mesures d'épaisseur de lard dorsal sur la carcasse notées x2 (site lombaire) et x4 (site costal), et la teneur en lipides intra-musculaires du muscle long dorsal (imf pour *intra muscular fat*). Les moyennes, variances et corrélations phénotypiques entre les caractères calculées sur les données mesurées sur les F2 sont résumées dans le tableau 1.4. Les épaisseurs de lard dorsal sont les caractères les plus variables, ce qui s'explique par la grande divergence de ce caractère entre les deux races grand-parentales. Les poids de panne et de bardière, x2 et x4 sont très corrélés positivement, avec des valeurs de corrélation supérieures à 0,65. La teneur en gras intra-musculaire est en revanche corrélée légèrement négativement avec les autres caractères, de -0,15 à -0,22.

TAB. 1.4 - *Moyennes, variances et corrélations phénotypiques entre les caractères étudiés.*

Caractère	Bardière	imf	Panne	x2	x4
Moyenne	7.169	1.984	0.963	34.681	31.157
Variance	0.425	0.157	0.037	15.610	14.869
Corrélation avec					
imf	-0.147				
Panne	0.694	-0.170			
x2	0.874	-0.151	0.643		
x4	0.829	-0.227	0.661	0.853	

4.3.1.3. Population La mesure de gras intra-musculaire est lourde à mener en routine car il s'agit d'une analyse chimique de la composition du muscle. Elle n'a donc été réalisée que pour une sous-population du schéma expérimental, impliquant 4 familles de père et 16 mères. La répartition des descendants F2 dans les différentes familles est indiquée dans le tableau 1.5 pour la sous-population concernée par l'étude de ces caractères, soit un total de 236 descendants.

4.3.2. Résultats des analyses unicaractères uniQTL

Les cinq caractères ont été analysés par la méthode unicaractère uniQTL décrite au paragraphe 4.1.. Les résultats présentés ici sont légèrement différents de ceux obtenus par

TAB. 1.5 - *Répartition des F2 dans les familles de PORQTL pour l'analyse des 5 caractères.*

Père			910001		
Mère	910014	910016	910071	910084	
Nb de descendants					
par mère	28	4	15	11	
par père			58		
Père			910045		
Mère	910013	910086	910095	910096	
Nb de descendants					
par mère	14	4	14	9	
par père			41		
Père			910081		
Mère	910002	910010	910020	910072	910097
Nb de descendants					
par mère	24	27	10	14	8
par père			83		
Père			910088		
Mère	910009	910069	910074		
Nb de descendants					
par mère	15	15	24		
par père			54		

Bidanel *et al.* (2002 [8]) et Milan (2002 [82]) en raison du nombre de descendants restreint dans cette étude, et de la correction préalable des données pour la moyenne et l'écart-type phénotypique du caractère estimés sur les individus F2 dans le but de rendre les résultats comparables à ceux des méthodes multidimensionnelles. Les profils de statistique de test ainsi obtenus sont présentés figure 8. La synthèse des résultats est donnée dans le tableau 1.6. Pour l'ensemble de ces caractères, les seuils à 5% globaux sur le groupe de liaison sont de 52,72. Le poids de bardière, le taux de gras intra-musculaire, les épaisseurs de gras dorsal x2 et x4 ont des maxima de statistiques de test localisés dans une fenètre de 5 cM. Pour le poids de panne, qui présente un maximum de statistique de test de 90,26 à la position 41 cM, l'observation du profil de la statistique de test correspondant permet de mettre en évidence un maximum local de 88,27 à la position 66 cM. Ces statistiques de test sont donc toutes très fortement significatives.

Les effets de substitution des allèles aux QTL estimés pour les pères sont tous élevés, et leur orientation particulièrement intéressante. En premier lieu, les allèles portés par les mâles F1 induisent tous une déviation du caractère dans le même sens, ce qui n'est pas le cas pour tous les caractères analysés dans ce protocole (Bidanel *et al.*, 2001). De plus, un signe positif de l'effet indique un allèle entraînant une déviation vers des valeurs hautes du caractère dans la race Meishan. L'allèle permettant d'augmenter le pourcentage de gras intra-musculaire serait donc porté par les haplotypes Meishan, ainsi que l'allèle permettant

de réduire l'adiposité de la carcasse. Ces relations entre les caractères sont à l'inverse des corrélations disponibles dans la littérature (Sellier *et al.*, 1998). L'identification de ces locus pourrait donc permettre de contourner grâce aux outils de la génétique moléculaire la corrélation naturellement défavorable au sens de la sélection entre ces caractères.

TAB. 1.6 - *Analyses unicaractères uniQTL des 5 caractères de composition de carcasse sur le chromosome 7.*
Maxima des statistiques de test (MaxLRT), positions et effets estimés pour les pères

		Bardière	imf	Panne	x2	x4
Max						
	LRT	76.10	89.14	90.26	71.82	76.47
	Position (cM)	66	64	41	67	69
Effets par père						
	910001	-0.558	1.240	-0.868	-0.696	-0.612
	910045	-0.480	0.906	-0.292	-0.482	-0.426
	910081	-0.564	0.784	-0.890	-0.464	-0.608
	910088	-0.716	0.394	-1.154	-0.646	-0.542
moyen		-0.579	0.831	-0.801	-0.572	-0.547

Tous les caractères présentent donc des valeurs de statistiques de test hautement significatives pour la même région chromosomique. De plus, les profils de statistiques de test sur le chromosome sont semblables entre les caractères. Il s'agit donc d'un exemple caractéristique de situation où l'on peut s'interroger sur la nature du déterminisme génétique reliant ces 5 caractères : s'agit-il d'un seul QTL pléiotrope, de QTL physiquement liés, sur quels caractères agissent-ils? Nous avons tenté, à titre d'illustration des méthodes développées, de répondre à ces questions.

5. Définition du champ d'étude

Le travail de thèse a été segmenté en deux parties distinctes : la mise en place de méthodologies multicaractères applicables à l'analyse de données issues de croisements expérimentaux animaux, puis le développement de méthodes permettant de dissocier des QTL liés sur un groupe de liaison donné. La synthèse de ces deux axes d'étude permettra de quantifier la puissance de discrimination entre des QTL pléiotropes et des QTL liés lors de l'analyse de pedigree animaux. Nous avons donc restreint cette étude à la prise en compte d'un unique groupe de liaison dans l'analyse. La phase de caractérisation des différentes méthodes a été réalisée à partir de données simulées. Pour chacune des parties multicaractère et multiQTL, une analyse des données présentées au paragraphe précédent permettra de souligner les avantages et les limites des différentes approches, et d'améliorer

la caractérisation du déterminisme de ces caractères.

Chapitre 2

Méthodes multicaractères

1. Eléments de bibliographie

1.1. Questions posées par les méthodologies unicaractères

Les premières détections de QTL systématiques caractère par caractère sur des populations animales (par exemple Andersson *et al.*, 1994 [3]) ont permis de mettre en évidence de nombreux QTL déterminant la majorité des caractères d'intérêt (Bidanel et Rotschild, 2002 [7]). L'analyse des résultats ainsi obtenus permet de mettre en évidence des régions chromosomiques porteuses de loci à effets quantitatifs pour plusieurs caractères. Un exemple classique est la région du SLA (Swine Leukocyte Antigens) sur le chromosome 7 porcin, qui semble porteuse de gènes d'intérêt pour la majorité des caractères de croissance et de composition corporelle. A la vue de tels résultats, la question qui se pose immédiatement est de savoir si tous les caractères pour lesquels un QTL peut être détecté sont déterminés par le même locus, ou bien s'il s'agit de locus liés. Pour répondre à ce type de question, des stratégies particulières doivent être mises en oeuvre, qui permettent de regrouper l'information concernant les caractères d'intérêt dans un même test.

Les méthodologies multicaractères pour la détection de QTL peuvent par ailleurs améliorer les performances de détection de QTL en comparaison des méthodes unicaractères, par la prise en compte des corrélations entre les caractères. En effet, la détection de QTL caractère par caractère est basée sur la minimisation de la variance résiduelle associée au caractère considéré. Dans certaines circonstances, la prise en compte de la variabilité conjointe de plusieurs caractères peut permettre d'améliorer cette minimisation, en autorisant une meilleure discrimination des relations phénotype / génotype (Jiang et Zeng, 1995 [40]; Korol, 1995 [57]).

Les deux points évoqués ci-dessus représentent les principales motivations à la mise en

oeuvre de méthodologies multicaractères. Le but, à terme, est d'arriver à une meilleure compréhension des hypothèses génétiques de relation entre les caractères, telles que la liaison génétique, la pléiotropie et les interactions QTL x environnement.

Nous allons dans un premier temps explorer les différentes méthodes qui ont été proposées dans la littérature et les conclusions auxquelles elles ont permis d'aboutir. Nous présenterons ensuite les méthodes que nous avons programmées pour l'application à des populations animales et les différentes situations simulées pour lesquelles elles ont été comparées. Enfin, nous présenterons les résultats obtenus et les conclusions que l'on peut en tirer.

1.2. Premières méthodes multicaractères proposées

1.2.1. Détection marqueur par marqueur

La première méthode multicaractère approfondie a été décrite par Ronin *et al.* (1995 [93]). Il s'agit d'un test de rapport de vraisemblance sur un marqueur supposé lié au QTL. Les populations expérimentales sont supposées fixées pour les locus marqueur et quantitatif. Aucune structure familiale n'est donc prise en compte dans l'écriture de la statistique de test. Le principe de la méthode est de modéliser la coségrégation des performances de deux caractères x et y corrélés en fonction des génotypes supposés au QTL. Par rapport à une méthode unicaractère, les performances sont supposées suivre, pour chaque classe génotypique au QTL (1=QQ, 2=Qq, 3=qq), des distributions binormales de moyennes et de matrices de (co)variance à estimer. La distribution conjointe de x et y pour le génotype i au QTL, $i = 1, \ldots, 3$, $f_i(x, y)$, s'écrit alors :

$$f_i(x,y) = \frac{1}{\sqrt{2\pi\sigma_{ix}\sigma_{iy}(1-R_i^2)}}$$

$$exp\left[-\frac{1}{2(1-R_i^2)}\left[\frac{(x-\mu_{ix})^2}{\sigma_{ix}^2} - 2R_i\frac{(x-\mu_{ix})(y-\mu_{iy})}{\sigma_{ix}\sigma_{iy}} + \frac{(y-\mu_{iy})^2}{\sigma_{iy}^2}\right]\right] \qquad (2.1)$$

où μ_{ix}, μ_{iy}, σ_{ix} et σ_{iy} sont les moyennes et écart-types des caractères x et y pour chaque génotype i, et R_i est le coefficient de corrélation intra groupe i.

En notant $S_j(x, y)$ la distribution espérée des performances pour chaque classe génotypique au marqueur génétique j testé, on a donc $S_j(x, y) = \sum_{i=1}^{3} \pi_{ji}(r) f_i(x, y)$, où $\pi_{ji}(r)$ sont les proportions d'individus de chaque classe génotypique i au QTL dans la classe j au marqueur, dépendant du coefficient de recombinaison à estimer r entre les deux locus. Dans ce cas de trois classes génotypiques, 16 paramètres doivent donc être estimés.

En réalité, la pertinence du modèle est essentiellement testée pour la précision de l'estimation des paramètres, dans le cadre d'un backcross à partir de données simulées. Aucun résultat n'est donné sur la puissance de la méthode proposée. Les auteurs montrent que la précision de l'estimation des paramètres est très nettement améliorée par la prise en compte d'un caractère corrélé au caractère analysé, même si ce deuxième caractère n'est pas déterminé par le QTL. L'amélioration est d'autant plus importante que la corrélation entre les deux caractères est élevée et que l'effet du QTL sur les caractères est important. Par exemple, pour un effet de 0,5 σ_p uniquement sur le premier caractère avec une variance résiduelle de 1 pour chacun des caractères, la précision de l'estimation de l'effet est améliorée d'un facteur 3 si la corrélation entre les caractères est de 0,9. De plus, l'estimation de r, le coefficient de recombinaison entre le QTL et le locus marqueur est de la même façon améliorée, et ce d'autant plus que l'effet du QTL est élevé. Les auteurs démontrent par ailleurs l'importance de la prise en compte des inégalités de variance intra-classes génotypiques pour la qualité de la détection. Des tests ont de plus été réalisés pour la prise en compte de la dominance dans l'analyse de données issues d'un croisement F2. Les auteurs concluent que, quand elle existe, sa prise en compte permet d'améliorer nettement la précision d'estimation des paramètres.

La figure 9 représente à titre d'illustration la co-distribution de deux caractères corrélés en fonction de l'origine de la corrélation entre les caractères, avec au moins un des caractères déterminé par un QTL. Ce schéma synthétise les types de situations pour lesquelles les méthodes multicaractères permettent d'améliorer la puissance et la précision d'estimation des paramètres par rapport aux méthodes unicaractères.

Un certain nombre de conclusions atteintes dans cette étude ont été validées dans le cadre de méthodes de cartographie d'intervalle.

1.2.2. Cartographie d'intervalle

1.2.2.1. Méthodes Korol *et al.* (1995 [57]) et indépendamment Jiang et Zeng (1995 [40]) ont discuté de l'intérêt de la cartographie d'intervalle de QTL multicaractère pour arriver à des conclusions relativement similaires. Les méthodes utilisées sont semblables. Ce sont des extensions de méthodes développées précédemment : la méthode proposée par Korol *et al.* (1995 [57]) est dérivée des travaux de Ronin *et al.* (1995 [93]) sur la détection multicaractère marqueur par marqueur, alors que Jiang et Zeng (1995 [40]) appliquent la méthode de détection unicaractère et multilocus de Zeng (1994 [115]) aux stratégies multicaractères. Les performances sont modélisées pour deux caractères selon une loi binormale dont les paramètres sont à estimer (voir équation 2.1), ce qui permet de réaliser un test de rapport de vraisemblance ou de calculer un lod-score. Les auteurs ont exploré le comportement de la méthode appliquée à des populations backcross (Korol *et*

al., 1995 [57]) ou F2 (Jiang et Zeng, 1995 [40], ce qui leur permet d'inclure dans le modèle des effets de dominance) fixées pour les locus marqueurs et quantitatifs, s'intéressant en particulier à la puissance de détection et à la précision d'estimation des paramètres. Dans les deux études, malgré un certain nombre de résultats antérieurs sur les méthodes uni et multicaractères (voir Ronin *et al.*, 1995 [93]) démontrant l'importance de la prise en compte de variances différentes entre les groupes génotypiques au QTL, les matrices de (co)variance sont considérées comme égales.

1.2.2.2. Etude analytique Jiang et Zeng (1995 [40]) et Korol *et al.* (1995 [57]) ont effectué des développements analytiques permettant de cadrer les circonstances dans lesquelles la prise en compte de la distribution conjointe des caractères permet d'améliorer la puissance de détection par rapport aux méthodes unicaractères. Korol *et al.* (1995 [57]) ont basé leur analyse sur une extrapolation des calculs d'espérance de LOD score pour le dispositif ($ELOD_x = -0.5 * N * log(1 - H_x^2)$ pour le caractère x, où H_x^2 est l'héritabilité du caractère x et N le nombre d'individus de la population testée) au cas multicaractère. En notant $V = \begin{bmatrix} \sigma_x^2 & R_{xy}\sigma_x\sigma_y \\ R_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$ la matrice de (co)variance résiduelle à estimer pour les deux caractères et a_x et a_y les effets respectifs du QTL sur les caractères x et y, $H_x^2 = a_x^2/4/(a_x^2/4 + \sigma_x^2)$ pour un backcross. Les auteurs développent un H_{xy}^2 tel que $ELOD_{xy} = -0.5 * N * log(1 - H_{xy}^2)$:

$$H_{xy}^2 = 1 - \frac{\sigma_x^2\sigma_y^2(1 - R_{xy}^2)}{(\sigma_x^2 + a_x^2/4)(\sigma_y^2 + a_y^2/4) - \sigma_x^2\sigma_y^2\left[R_{xy} + a_x a_y/(4\sigma_x\sigma_y)\right]^2} \tag{2.2}$$

Ils montrent alors que la puissance de détection de la méthode multicaractère est supérieure à celle de la méthode unicaractère si $H_x^2 \leq H_{xy}^2$, ce qui correspond à $R_{xy}a_x a_y \leq 0$ avec $R_{xy} \neq 0$ et $a_y \neq 0$, et même $H_x^2 < H_{xy}^2$ si $R_{xy} \neq 0$ et $a_y = 0$. Dans la pratique, on peut donc dire que les méthodes multicaractères permettront d'améliorer la puissance de détection par rapport aux méthodes unicaractères si le produit $Pr = R_{xy}a_x a_y$ des effets du QTL sur chaque caractère et de la corrélation résiduelle entre les caractères est négatif, ou bien que l'effet du QTL sur l'un des caractère est nul avec une corrélation résiduelle non nulle. La puissance de détection peut donc être augmentée par le seul apport d'information dû à la corrélation résiduelle entre les caractères. Jiang et Zeng (1995 [40]) arrivent aux mêmes conclusions en développant l'écriture de la vraisemblance pour un QTL localisé sur un marqueur génétique. Le fait qu'un caractère résiduellement corrélé au caractère d'intérêt suffise à améliorer la puissance de détection du processus implique que l'intégration d'un caractère contribuant de façon significative à la statistique de test ne préjuge pas du fait qu'il est ou non déterminé par le QTL. Ce dernier point ne pourra être résolu que par le test du niveau de signification des effets du QTL estimés pour ce

caractère (voir Korol *et al.*, 2001 [59]).

1.2.2.3. Statistique de test Le test principal, pour une région chromosomique donnée, dans les deux méthodes présentées ci-dessus, consiste à comparer l'hypothèse nulle : "il n'y a pas de QTL agissant sur les caractères" à l'hypothèse alternative : "il y a un QTL agissant sur au moins un caractère". Jiang et Zeng (1995 [40]) se contentent d'une approximation de la distribution de la statistique de test sous H0 par un $\chi^2_{\alpha/M,2m+1}$, où α est l'erreur de première espèce voulue, M est le nombre de groupes de liaison analysés (approximation d'une correction de Bonferroni), $2m+1$ est le nombre de paramètres estimés, avec m le nombre de caractères auquel ils ajoutent 1 pour la position. Ils proposent d'estimer dans la pratique de façon plus propre les seuils par des permutations des données selon une extrapolation à plusieurs caractères de la méthode proposée par Churchill et Doerge en 1994 [11]. La permutation des données est alors réalisée en conservant les relations entre les caractères et en cassant les liaisons avec les génotypes. Korol *et al.* (1995 [57]) comparent les puissances obtenues avec un seuil calculé à partir d'un $\chi^2_{\alpha/M,ddl}$, où ddl est le nombre de degrés de liberté du test (donc similaire à celui de Jiang et Zeng (1995 [40])), et des seuils estimés par simulation des données sous l'hypothèse H0. Ils concluent à une bonne approximation par le χ^2.

Jiang et Zeng (1995 [40]) proposent de plus un test des effets pléiotropes, en faisant deux comparaisons parallèles de statistiques de test. Ils définissent H01 : "il y a un QTL pour le premier caractère à la position x", H02 : "il y a un QTL pour le deuxième caractère à la position x", H1 : "il y a un QTL pour chaque caractère à la position x". Le test pour les effets pléiotropes consiste à valider à la fois H1 contre H01 et H1 contre H02. L'écriture de la vraisemblance pour H01 et H02 est identique à celle de H1 sauf que certains paramètres sont contraints à 0. Comme les positions sont fixées, les auteurs considèrent que chacun de ces tests correspond à un χ^2_2. Ces auteurs proposent aussi un test de la pléiotropie contre deux QTL liés qui sera développé dans la partie multiQTL.

1.2.2.4. Résultats Les résultats obtenus sur des données simulées confirment ceux de Ronin *et al.* (1995 [93], voir 1.2.). Les analyses multivariées permettent d'améliorer la précision d'estimation des paramètres, et ce d'autant plus que la corrélation résiduelle entre les caractères est élevée. Les puissances de détection sont augmentées par la prise en compte des distributions conjointes des caractères. Comme montré analytiquement, les performances de la méthode sont meilleures quand le produit des effets du QTL et de la corrélation entre les caractères est négatif (majorité des cas simulés). La détection peut par ailleurs être améliorée par la prise en compte uniquement d'un caractère résiduellement corrélé au caractère d'intérêt.

Les auteurs explorent de plus la capacité de leurs méthodes à prendre en compte, et

éventuellement distinguer, la variabilité sur les caractères due à d'autres portions chromosomiques non liées. Les résultats de cet aspect seront donnés dans la partie multiQTL.

Enfin, Jiang et Zeng (1995 [40]) consacrent un point particulier de leur étude à la mise en évidence d'interactions QTL x environnement, qui peut être un des buts de la mise en place des analyses multicaractères. Les performances enregistrées dans des environnements différents pour le même caractère peuvent en effet être considérées comme des caractères différents en raison de l'interaction entre le génotype et l'environnement. Les auteurs distinguent deux types de protocoles : ceux où les mêmes individus sont utilisés pour l'enregistrement de performances dans des environnements différents, et ceux où des individus différents d'une même population sont caractérisés pour leurs performances dans différents environnements. Le modèle du premier protocole est très similaire à celui prenant en compte séparément des paires de caractères corrélés. Pour le deuxième protocole, les deux environnements sont considérés comme indépendants. La vraisemblance totale est donc égale au produit des vraisemblances calculées pour chaque environnement séparément. Le test revient à comparer les différences d'estimation des effets entre paires d'environnements à zéro. Les auteurs montrent alors que, pour un même nombre d'enregistrements de performances (et donc deux fois plus de génotypages pour le deuxième protocole), le deuxième type de protocole permet de détecter plus de QTL, alors que le premier permet plus souvent de conclure à une interaction QTL x environnement.

1.3. Généralisation

L'ensemble des méthodes proposées en 1995 a été développé pour des applications à des croisements entre lignées *inbred*. Leur extrapolation à des populations *outbred*, où les paramètres doivent être estimés intra-famille, ou à l'analyse conjointe de plus de deux caractères, entraîne l'estimation d'un nombre de paramètres très important, qui devient vite limitant d'une part pour la puissance de détection des méthodes, et d'autre part en terme de faisabilité, puisque les temps de calcul augmentent quasi exponentiellement avec le nombre de paramètres. Pour contrecarrer cette limite, Weller *et al.* (1996 [107]) ont proposé une solution originale, qui consiste à synthétiser l'information sur les liaisons entre les caractères dans une combinaison linéaire des performances, afin de ne réaliser que des analyses univariées. La méthode proposée par Weller *et al.* en 1996 [107], revient à réaliser une Analyse en Composantes Principales (*Principal Component Analysis*: PCA) sur les données phénotypiques.

Le principe de l'analyse en composantes principales est de répartir la variabilité des données dépendantes de K variables selon K axes à partir de leur matrice de (co)variance (Mardia *et al.*, 1979 [79], Chap 8) de telle façon que ces axes définissent des directions orthogonales les unes par rapport aux autres. Les combinaisons linéaires des données

obtenues à partir de ces axes sont donc indépendantes. Le premier axe représente le maximum de la variabilité synthétisable dans une combinaison linéaire des caractères, le deuxième axe le maximum de la variabilité restant après la définition du premier, et ainsi de suite. On peut montrer que ces combinaisons linéaires correspondent aux K vecteurs propres z_k, $k = 1, \ldots, K$ de la matrice de (co)variance des données, la variabilité qu'elles expliquent étant représentée par la valeur propre λ_k à laquelle chacune est associée. La proportion de variance expliquée par les k premières variables en composantes principales est donc donnée par $\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{K} \lambda_i$. On considère habituellement que seules les variables en composantes principales impliquées dans l'explicitation de la plus grande partie de la variance totale doivent être retenues dans une analyse.

La proposition de Weller *et al.* (1996 [107]) est de réaliser une PCA sur les données phénotypiques, donc sur la base de la variabilité phénotypique, et d'analyser uniquement les variables expliquant la plus grande part de la variabilité totale. Chaque variable en composantes principales est analysée par une méthode univariée de type méthode unicaractère, marqueur par marqueur. Les résultats de l'analyse de chaque variable sont alors considérés comme indépendants. Cette supposition permet d'après les auteurs d'une part de corriger de façon simple (type correction de Bonferroni) les seuils nominaux pour prendre en compte la multiplicité des variables analysées, et d'autre part de conclure à la présence de QTL liés si deux combinaisons linéaires permettent de mettre en évidence des QTL à des positions liées. Or la part que l'on cherche à maximiser lors d'une détection de QTL est celle des résidus, c'est à dire la part ne dépendant pas des relations phénotype/génotype. Par définition, cette variabilité dépend de la structure des données à la position testée. De plus, elle ne correspond pas nécessairement à la variabilité phénotypique. Mangin *et al.* (1998 [76]) ont étendu cette méthode à la cartographie d'intervalle. Ils ont mis en évidence que la proposition de Weller *et al.* (1996 [107]) de se baser sur les données phénotypiques pour détecter des QTL pléiotropes peut être valable si la structure génétique des données est relativement simple (ségrégation d'un unique QTL déterminant deux caractères par exemple), et donc la variabilité résiduelle proche de la variabilité phénotypique.

Par ailleurs, Mangin *et al.* (1998 [76]) montrent que la PCA pour détecter les QTL correspond asymptotiquement à un test de maximum de vraisemblance sur la distribution conjointe des caractères si 1) elle est réalisée sur la matrice de (co)variance résiduelle, et 2) la statistique de test est calculée comme le maximum de la somme des statistiques de test pour chacune des composantes principales à chaque position (sPCA). Si la proposition 2 est aisément réalisable, la première est quant à elle beaucoup plus limitante. En effet, la variabilité résiduelle, intra-classe génotypique au QTL, n'est pas accessible et doit théoriquement être alors estimée à chaque position. Les auteurs proposent une approximation de cette variabilité sous H0 par une analyse multivariée (pas de QTL, donc variabilité indépendante de la position testée). On obtient alors asymptotiquement une estimation

correcte de la matrice de (co)variance résiduelle. Dans la pratique, elle est cependant surestimée, puisqu'elle contient la variabilité liée à la présence du QTL. L'effet de cette surestimation n'a pas été exploré par les auteurs. De plus, cette proposition ne permet que partiellement de s'affranchir de l'aspect multivarié limitant des méthodes proposées précédemment. Cependant, étant donné que l'analyse de chaque variable est univariée, on peut aisément supposer qu'à partir d'un nombre limité de caractères à analyser des gains importants seront réalisés en terme de temps de calcul nécessaire à l'analyse multicaractère globale, par rapport à une analyse multivariée complète.

Mangin *et al.* (1998 [76]) ont extrapolé les distributions des statistiques de test de la sPCA, et par extension des méthodes multivariées, sous différentes hypothèses, à partir des résultats obtenus par Rebaï *et al.* (1995 [91]) et Mangin et Goffinet (1994a [74]) dans le cadre unicaractère pour des puissances de détection inférieures à 100%, c'est à dire des effets de QTL relativement faibles. Ces extrapolations sont réalisées à partir de l'exemple d'un backcross entre populations *outbred*. Ils confirment ainsi la relation directe entre le produit des effets du QTL sur les deux caractères et leur corrélation résiduelle avec la capacité d'amélioration des performances des méthodes unicaractères par les méthodes multicaractères (figure 10). Sur la figure, l'opposition Pr positif/négatif apparaît clairement, de même que la possibilité de gain en puissance par rapport aux méthodes unicaractères quand l'effet du QTL sur le deuxième caractère est nul. De la même façon que Jiang et Zeng (1995 [40]) le signalaient, ils montrent que la valeur de la statistique de test multicaractère est toujours supérieure ou égale à la valeur des statistiques de test unicaractères. Cependant, les seuils de signification étant plus élevés pour les analyses multicaractères du fait de l'estimation d'un plus grand nombre de paramètres, cela ne permet en aucun cas de conclure à une supériorité systématique des puissances obtenues par les méthodes multivariées sur celles obtenues par les méthodes unicaractères.

1.4. Développements récents

1.4.1. Méthodologie

Korol *et al.* (2001 [59]) ont proposé une méthode pour contrecarrer l'inflation du nombre de paramètres à estimer avec le nombre de caractères lors des analyses multicaractères, pour des applications à des croisements entre lignées *inbred*. Ils se basent sur les observations réalisées lors de leurs analyses précédentes sur les méthodes multicaractères, à savoir une meilleure cartographie des QTL par la prise en compte de la distribution conjointe des caractères quand le produit Pr est négatif, c'est à dire quand l'éloignement entre les classes génotypiques dues au QTL est augmenté par la prise en compte de caractères supplémentaires. Ils cherchent alors à transformer l'espace à 2 dimensions (dans

le cas de deux caractères) représenté par la distribution conjointe des caractères en un espace à une dimension où le rapport entre la variation entre les classes génotypiques au QTL et la variation intra-classes génotypiques au QTL est maximal. La transformation des caractères est donc spécifique de la structure génétique des performances induite par la présence du QTL, ce qui la rend très différente des propositions de Weller *et al.* (1996 [107]) et Mangin *et al.* (1998 [76]). L'avantage de cette transformation est qu'elle reflète directement la variabilité due au QTL, et qu'elle permet, quel que soit le nombre de caractères analysés, de ne réaliser qu'une analyse univariée de la variable résultante.

La transformation proposée par Korol *et al.* (2001 [59]) est spécifique de chaque intervalle entre deux marqueurs génétiques. Pour chaque intervalle, la classe génotypique au QTL de chaque individu est déterminée en fonction de son génotype aux marqueurs flanquants. Les recombinants, dont la classe génotypique varie en fonction de la localisation dans l'intervalle, sont exclus de cette étape et les auteurs font l'hypothèse d'absence de doubles recombinaisons dans l'intervalle considéré. La matrice de (co)variance intra-classes génotypiques, considérée comme égale entre les classes, est estimée et ses valeurs et vecteurs propres calculés pour réaliser une PCA. Les auteurs procèdent ensuite à des transformations d'échelle et des rotations pour ne plus obtenir qu'une unique variable synthétique des données initiales. Cette variable est ensuite analysée en chaque point de l'intervalle génétique selon une procédure univariée classique de cartographie d'intervalle.

Les auteurs ont développé la notion d'héritabilité de la variable synthétique comme le rapport de la variabilité qu'elle explique sur la variabilité totale. En notant V_G la variabilité inter-classes et V_R la variabilité intra-classe, on a ainsi $H_T^2 = V_G/(V_G + V_R)$. Ils comparent ainsi H_x^2 (voir 1.2.) et H_T^2 pour prédire le gain de puissance apporté par l'analyse de cette variable par rapport aux analyses unicaractères par un calcul de $ELOD$.

1.4.2. Résultats

Les auteurs ont par ailleurs réalisé des simulations pour caractériser leur méthode en fonction de différentes relations génétiques entre 10 caractères. Ils concluent ainsi à une bonne prédiction des valeurs moyennes de LOD obtenues par simulation grâce au $ELOD$. Les seuils ont été calculés d'une part par simulation de données sous l'hypothèse nulle (sans QTL sur le groupe de liaison), et d'autre part avec des permutations des données obtenues sous H1 (Churchill et Doerge, 1994 [11]). Ils concluent à une amélioration de la puissance de détection et de la précision d'estimation des effets et de la position du QTL quand tous les caractères sont intégrés à l'analyse. L'ampleur de l'amélioration dépend essentiellement du signe des Pr de chaque couple de caractères intégrés à l'analyse.

Korol *et al.* (2001 [59]) proposent aussi des tests à réaliser pour d'une part évaluer la signification de la contribution de chaque caractère à la puissance de détection du QTL,

et d'autre part tester la signification des effets du QTL estimés pour chaque caractère. Ces deux objectifs sont distincts puisqu'un caractère peut être intégré à l'analyse uniquement pour l'information qu'il apporte à travers sa corrélation résiduelle avec le caractère d'intérêt. Pour ces deux buts, les valeurs des performances pour le caractère d'intérêt sont ré-échantillonnées par rapport aux génotypes et aux valeurs de performances pour les autres caractères. Le nombre de fois où le LOD est supérieur à celui des données initiales sert à estimer le niveau de signification de la contribution du caractère. Si un caractère s'avère non significatif, il peut alors être exclu, et l'analyse recommencée. De la même façon, le nombre de fois où l'estimation de l'effet du QTL est supérieure à celle obtenue pour les données originales est comptabilisé et sert à évaluer la signification de l'effet estimé pour le caractère. Sur quelques cas particuliers, ils montrent ainsi que l'intégration de caractères non informatifs pour le QTL n'est pas neutre, puisqu'elle diminue la précision d'estimation des paramètres et affecte la puissance de détection.

Enfin, Korol *et al.* (2001 [59]) ont comparé les performances de leur méthode avec celles de la PCA telles que proposée par Weller *et al.* (1996 [107]) et Mangin *et al.* (1998 [76]). Ils montrent que la puissance de détection et la précision d'estimation des paramètres est nettement améliorée par la définition d'une variable synthétique spécifique de l'intervalle génétique considéré, en particulier en ce qui concerne l'estimation de la position du QTL. Cette comparaison n'a été réalisée que dans le cas d'un seul QTL déterminant les caractères, et les auteurs supposent une amélioration nettement plus importante quand la variabilité résiduelle à la position testée diffère beaucoup de la variabilité totale en raison de la présence d'autres QTL.

1.4.3. Limites de l'application à des populations animales

La principale limite de la méthode proposée par Korol *et al.* (2001 [59]) est l'exclusion des recombinants pour l'estimation de la variabilité résiduelle. En effet, les populations expérimentales animales sont souvent limitées par 1) le nombre d'individus, 2) la densité en marqueurs génétiques. Exclure les recombinants lorsque la densité de la carte génétique est en moyenne de 20 cM revient à diminuer de façon non négligeable la taille de la population pour certains intervalles génétiques, et éventuellement de façon différente en fonction des familles. La puissance de la méthode est donc fortement affectée.

De plus, le nombre de tests proposés pour tester la signification du LOD, puis la signification de la contribution de chaque caractère et la signification des effets du QTL pour chaque caractère semble dans l'application aux populations animales irréaliste en terme de besoins de calcul, même si la méthode est univariée.

1.5. Génotypage d'extrêmes

Certains auteurs ont développé des adaptations de méthodes d'affinage de la cartographie de QTL aux stratégies multicaractères, tel en particulier le génotypage d'extrêmes. Cette stratégie proposée en 1989 par Lander et Botstein ([62]), consiste à phénotyper une population d'individus plus importante que pour la cartographie classique. Pour limiter les coûts liés au génotypage de nombreux marqueurs pour tous les individus, seuls les individus présentant des performances extrêmes pour le caractère considéré, c'est à dire les individus les plus informatifs quant à la liaison génotype / phénotype, sont génotypés. Cette stratégie a deux limites essentielles : le coût de production d'un individu et / ou de sa performance (qui peuvent entrer en concurrence avec le coût en génotypages), et le nombre de caractères mesurés pour ces individus. En effet, le deuxième point est limitant si la sélection d'individus à performances extrêmes pour de nombreux caractères considérés individuellement revient à génotyper toute la population. Le même type de limite doit être pris en compte lors des stratégies multicaractères. Si la corrélation entre deux caractères est élevée, la limite est relativement faible. Si elle est faible, le génotypage d'extrêmes peut se révéler totalement inadéquat. Weller *et al.* (1998 [109]) ont proposé de sélectionner les extrêmes pour la valeur de la combinaison linéaire de caractères analysée. Cependant, si cette combinaison est intervalle-spécifique, telle que proposée par Korol *et al.* (2001 [59]), la proposition n'est plus réaliste. Henshall et Goddard (1999 [32]) ont de leur côté proposé une régression logistique qui permet de prendre en compte le génotypage d'extrêmes dans la détection de QTL multicaractère, ce qui permet de réduire les biais d'estimation des effets du QTL.

1.6. Bilan

En résumé, deux grands types de méthodes ont été proposés pour la détection multicaractère de QTL. Le premier consiste à extrapoler directement les méthodes unicaractères au cas multicaractère par l'utilisation de fonctions multinormales pour modéliser la distribution des performances. Cette méthode est efficace dans le cadre de l'application aux populations issues de croisements entre lignées *inbred*, où le nombre de paramètres à estimer par caractère est relativement faible. Cependant, si le nombre de caractères analysés simultanément augmente, ou que les données requièrent la prise en compte d'une structure familiale, le nombre de paramètres à estimer devient vite limitant.

Pour pallier l'augmentation du nombre de paramètres, Weller *et al.* (1996 [107]) et Korol *et al.* (2001 [59]) ont proposé indépendamment de synthétiser l'information sur les relations entre les caractères dans une ou un nombre limité de variables à analyser de façon univariée. La proposition de Weller *et al.* (1996 [107]) de baser la transformation

des données sur la variabilité totale limite son application à des caractères présentant des structures génétiques simples. La proposition de Korol *et al.* (2001 [59]) semble plus attractive, bien qu'elle implique l'exclusion d'une partie de l'analyse des individus recombinants dans chaque intervalle génétique.

2. Méthode proposée

La principale difficulté posée pour la prise en compte de plusieurs caractères pour la détection de QTL dans les pedigree d'animaux domestiques réside donc dans le nombre de paramètres à estimer. Afin de gérer celle-ci, nous proposons, selon un principe similaire à celui développé par Weller *et al.* (1996) et Korol *et al.* (2001), de résumer l'information concernant les relations entre les caractères dans une variable composite. L'application des techniques d'analyse discriminante nous a semblé une voie possible pour atteindre cet objectif.

2.1. Principes de l'analyse discriminante (DA)

L'analyse discriminante est une méthode statistique qui s'applique à des structures expérimentales où les individus sont répartis en groupes en fonction des valeurs prises par des variables explicatives (Mardia, 1979 [79]). Les premiers exemples simples ont reposé sur la discrimination d'individus atteints ou non d'une maladie en fonction de différentes mesures physiologiques. DA détermine, en fonction des relations entre les variables explicatives et de la répartition des individus en groupes qui en découle, les combinaisons linéaires qui maximisent le rapport de la variabilité inter-groupes sur la variabilité intra-groupes. Le but de cette méthode est double : trouver une variable synthétique qui discrimine au mieux les groupes, et donc les variables explicatives qui participent le plus à la répartition entre groupes, et trouver une variable synthétique qui permette, sur la base des réalisations des variables pour un individu, de le classer le plus pertinemment possible dans un ou l'autre groupe. C'est le premier objectif qui nous intéressera dans l'application à la détection de QTL.

La combinaison linéaire des variables est obtenue simplement. En notant $\mathbf{B}$ (*Between*) la matrice de (co)variance inter-groupes et $\mathbf{W}$ (*Within*) la matrice de (co)variance intra-groupes, la combinaison linéaire des variables qui maximise le rapport des variabilités $\mathbf{W}^{-1}\mathbf{B}$ est obtenue par le vecteur propre associé à la plus grande valeur propre de $\mathbf{W}^{-1}\mathbf{B}$. Dans la pratique, $\mathbf{W}$ est estimée soit comme une matrice égale entre les groupes, soit comme la matrice de (co)variance moyenne sur tous les groupes.

2.2. Application à la détection de QTL

L'application de l'analyse discriminante à la cartographie de QTL est basée sur la répartition des descendants en deux groupes d'haplotypes au QTL hérités de leur(s) parent(s) hétérozygote(s) à chaque position testée. On cherche alors à maximiser le rapport de la variabilité due au QTL (variabilité inter-groupes) sur la variabilité résiduelle, due à tout autre facteur (variabilité intra-groupes). La combinaison linéaire des caractères définie à partir du vecteur propre associé à la plus grande valeur propre de la matrice représentant le rapport de variabilités est donc caractéristique de l'écart de performances induit par la ségrégation des allèles au QTL. Elle caractérise donc l'effet du QTL. Un des avantages de cette méthode est que, ne modélisant pas la variabilité résiduelle, celle-ci peut être due à n'importe quel facteur. La méthode doit donc théoriquement être relativement robuste à la ségrégation d'autres QTL dans la population.

2.2.1. Cas théorique : génotype au QTL connu

Supposons qu'on connaisse de façon sûre le génotype à un QTL localisé à la position x sur le groupe de liaison pour tous les n descendants d'un individu hétérozygote au QTL Qq. Chaque descendant k est associé à une réalisation des variables $\mathbf{y}_k$. On peut définir un poids pour chacun de ces individus, p_k, tel que $\sum_{k=1}^{n} p_k = 1$ (ici, les p_k sont tous égaux à $1/n$), $\mathbf{g} = \sum_{k=1}^{n} p_k \mathbf{y}_k$ le vecteur des moyennes des variables explicatives sur l'ensemble des données (géométriquement, ce vecteur est le centre d'inertie du nuage des données) et $\mathbf{T} = \sum_{k=1}^{n} p_k \mathbf{g}\mathbf{g}'$ la matrice de (co)variance totale (phénotypique).

On peut alors répartir les descendants en deux groupes G_g, $g = 1, 2$, en fonction de l'allèle au QTL qu'ils ont reçu. On définit alors n_g le nombre d'individus du groupe g, $P_g = \sum_{k=1}^{n_g} p_k$ le poids associé au groupe g, et $\mathbf{g}_g$ le vecteur des moyennes de performances intra-groupe (géométriquement, $\mathbf{g}_g$ correspond aux coordonnées du centre d'inertie du groupe g). La variabilité inter-groupes $\mathbf{G}$ (due au QTL) est estimable par :

$$\mathbf{G} = \sum_{g=1}^{2} P_g \mathbf{g}_g \mathbf{g}_g'. \tag{2.3}$$

De la même façon, la variabilité résiduelle, intra-groupe, $\mathbf{R}_g$ pour chaque groupe g s'écrit :

$$\mathbf{R}_g = \sum_{k=1}^{n_g} p_k \mathbf{y}_k \mathbf{y}_k' = \sum_{k=1}^{n} p_k \mathbf{y}_k \mathbf{y}_k' - P_g \mathbf{g}_g \mathbf{g}_g'. \tag{2.4}$$

La variabilité intra-groupe moyenne s'écrit alors $\mathbf{R} = 1/2 \sum_{g=1}^{2} \mathbf{R}_g$. Cette option a été

choisie comme permettant de moyenner les différences potentiellement existantes entre les variabilités résiduelles des groupes. On peut montrer que $\mathbf{T} = \mathbf{G} + \mathbf{R}$, ce qui correspond bien aux définitions respectives de ces variabilités dans les modèles génétiques.

On peut par ailleurs montrer que $\mathbf{R}^{-1}\mathbf{G}$ et $\mathbf{T}^{-1}\mathbf{G}$ ont les mêmes vecteurs propres, mais associés à des valeurs propres différentes (Mardia, 1979) : si λ est valeur propre de $\mathbf{T}^{-1}\mathbf{G}$, alors $\lambda(\lambda - 1)$ est valeur propre de $\mathbf{R}^{-1}\mathbf{G}$. La matrice $\mathbf{T}$ étant en général plus accessible que la matrice $\mathbf{R}$, elle pourra être utilisée de façon préférentielle.

Une expression analytique peut être donnée de la part κ du rapport de variabilités expliquée par la combinaison linéaire associée à la plus grande valeur propre dans le cas simple de deux caractères. En notant $\sigma_r^2 = 1$ la variance résiduelle (indépendante du génotype au QTL), ρ_{12} la corrélation résiduelle entre les caractères, et σ_1 et σ_2 les variances génétiques respectivement pour le premier et le deuxième caractère dues à un QTL pléiotrope, la corrélation due au QTL étant égale à 1, κ s'écrit :

$$\kappa = \frac{\sigma_1 (\sigma_1 - \rho\sigma_2) + \sigma_2 (\sigma_2 - \rho\sigma_1)}{1 - \rho^2} \tag{2.5}$$

Cette expression montre que la variabilité prise en compte par la combinaison linéaire est spécifique de la variabilité génétique des caractères due au QTL à la position testée.

2.2.2. Cas général

Dans la pratique, les allèles reçus d'un parent au QTL par les descendants ne sont connus qu'en probabilité ($p\left(d_{ijk}^x = (q_s, q_d)/\widehat{hs_i}, hd_{ij}, M_i\right)$ décrits au paragraphe 4.1. de la partie 1). Ces probabilités peuvent être intégrées à l'analyse comme des poids (les p_k utilisés au paragraphe précédent) indiquant le degré de confiance à accorder à la présence d'un individu dans un groupe. Tous les descendants sont donc dans la pratique présents dans les deux groupes, leurs présences étant pondérées par la probabilité qu'ils aient reçu l'haplotype correspondant au groupe.

Lors de la détection de QTL, la transformation des caractères en une combinaison linéaire est réalisée pour chaque position testée comme une étape préliminaire au calcul de la vraisemblance univariée. La variable analysée à chaque position étant une combinaison linéaire différente, la vraisemblance sous l'hypothèse nulle est calculée pour chaque position. L'analyse univariée est techniquement identique à celle d'un caractère par la méthode unicaractère ou à l'analyse d'une variable en composantes principales.

Par rapport à la méthode proposée par Korol *et al.* (2001 [59]), l'étape de répartition des individus en groupes en fonction du génotype au QTL est similaire, sauf qu'elle est réalisée ici à chaque position, en intégrant les recombinants potentiels par l'utilisation

des pondérations. L'étape suivante est différente, puisque Korol *et al.* (2001 [59]) basent l'établissement de la combinaison linéaire sur une succession de transformations qui débute par une analyse en composantes principales, suivie de changements d'échelle et de rotations, alors que l'analyse discriminante regroupe toute la transformation en une étape simple à appréhender.

Par rapport à la méthode proposée par Weller *et al.* (1996 [107]), l'analyse discriminante permet de prendre en compte les spécificités de la structure génétique des caractères induites par la présence du QTL potentiel dans la définition de la combinaison linéaire. De cette façon, une seule variable doit être analysée à chaque position, ce qui doit permettre de considérablement réduire les temps de calcul nécessaires à l'analyse. D'autre part, l'analyse d'une variable unique évite d'avoir à prendre en compte des tests multiples corrélés réalisés sur des variables non indépendantes dans le calcul des seuils de signification. Cependant, les paramètres estimés caractérisent la combinaison linéaire sans retour possible aux variables de départ.

2.3. Généralisation

2.3.1. Impact de la prise en compte d'allèles en ségrégation dans la population

La méthode telle que nous l'avons décrite ci-dessus est basée sur une répartition des descendants en deux groupes d'haplotypes en fonction de l'haplotype hérité d'un parent hétérozygote. Dans la pratique, les individus de toute la population peuvent être répartis en 2 groupes en fonction de l'origine grand-parentale des allèles hérités de leur père. On considére alors que, pour l'analyse de pedigree de type F2 issus de croisements entre animaux de populations divergentes pour les caractères d'intérêt, les allèles de même origine grand-parentale ont tendance à entraîner des déviations du caractère dans le même sens. Cette restriction est cependant simpliste.

Afin d'évaluer son impact quand des allèles codant pour des déviations de sens opposés sont en ségrégation dans les populations grand-parentales, nous avons testé la possibilité de réaliser les transformations spécifiques de chaque famille de demi-frères sur quelques exemples. Une combinaison linéaire différente est alors calculée et utilisée pour chaque famille de père. Dans ce cas, on peut présumer que la taille des familles de demi-frères devra être assez importante pour que les matrices de (co)variances intra-classes génotypiques au QTL soient bien estimées. Cependant, comme nous allons le voir dans le paragraphe suivant, l'analyse discriminante sur la base de deux groupes génotypiques déterminés par le père permet de bien caractériser à la fois les effets des allèles au QTL des mâles et ceux des femelles de la génération F1. Restreindre la transformation intra-famille implique donc que si le père n'est pas hétérozygote au QTL pour une famille donnée, la combinaison

linéaire ne permettra pas de discriminer les allèles au QTL portés par les femelles hétérozygotes auxquelles il peut être accouplé, ce qui peut entraîner une perte de puissance du dispositif.

2.3.2. Impact de la prise en compte des méioses maternelles dans la formation des groupes

Un autre développement consiste à considérer non pas deux mais trois ou quatre groupes d'haplotypes, en fonction des méioses survenues chez les mères. Quand le déterminisme du QTL est strictement additif, nous allons montrer que la combinaison linéaire obtenue avec 2, 3 ou 4 groupes génotypiques est exactement identique lors de l'analyse de données issue d'un croisement F2. L'explication du phénomène est simple. Etant donné que les individus F1 sont tous issus du même croisement entre des individus porteurs d'allèles fixés au QTL, les allèles portés par les femelles F1 sont les mêmes que ceux portés par les mâles. Si la combinaison linéaire définie par la répartition des individus en fonction des allèles au QTL portés par les pères caractérise au mieux l'effet du QTL dû à ces allèles, elle caractérise de la même façon au mieux l'effet du QTL dû aux allèles des mères. On ne gagne donc pas d'information à prendre en compte la répartition des individus due aux génotypes au QTL transmis par les mères.

Dans le cas simple d'individus issus d'un croisement F2 entre lignées fixées pour les allèles Q et q au QTL, avec un déterminisme additif des caractères et une population infinie telle qu'on puisse considérer que la proportion de chaque génotype au QTL dans la descendance est $1/4$, nous allons montrer que la combinaison linéaire obtenue que l'on prenne en compte 2 (en fonction des méioses des pères), 3 (en fonction des méioses des pères et des mères, sans distinction des hétérozygotes) ou 4 (en fonction des méioses des pères et des mères, en distinguant les hétérozygotes) groupes de génotypes au QTL est identique.

Les pondérations de la combinaison linéaire sont les coordonnées du vecteur propre associé à la plus grande valeur propre de $\mathbf{R}^{-1}\mathbf{G}$, donc $\mathbf{T}^{-1}\mathbf{G}$. Etant donné que $\mathbf{T}$ est indépendante du nombre de groupes considérés, nous allons développer analytiquement l'écriture de $\mathbf{G}$ uniquement. On appelle $\mathbf{G}_G$ la matrice de (co)variance inter-groupes quand G groupes sont pris en compte dans l'analyse. Comme nous l'avons vu précédemment et en reprenant les mêmes notations (voir 2.3),

$$\mathbf{G_G} = \sum_{g=1}^{G} P_g \, \mathbf{g}_g \mathbf{g}_g{}' \tag{2.6}$$

On note α_l l'effet de substitution des allèles au QTL pour le caractère l. Les moyennes

pour les différentes classes génotypiques au QTL seront donc : QQ : $+\alpha_l$, Qq : 0, qq : $-\alpha_l$. On appelle $\mathbf{a}$ le vecteur $\{\alpha_l; l = 1, \ldots, p\}$ et $\mathbf{A}$ la matrice $\mathbf{aa}'$.

Dans le cas de deux groupes d'haplotypes transmis par le père, le vecteur $\mathbf{g}_q$ correspond au groupe déterminé par l'allèle q, $q = 1, 2$. On a alors $\mathbf{g_1} = \mathbf{a}/2 = -\mathbf{g}_2$, d'où $\mathbf{g_1 g_1}' = \mathbf{g_2 g_2}' = \mathbf{A}/4$. Avec $P_1 = P_2 = 0,5$, on obtient donc $\mathbf{G_2} = \mathbf{A}/4$.

Dans le cas de quatre groupes d'haplotypes déterminés par les méioses des pères et des mères, les vecteurs $\mathbf{g}_h$ correspondent respectivement, pour $q = 1, \ldots, 4$, aux génotypes QQ, Qq, qQ et qq. On a alors $\mathbf{g_1} = \mathbf{a} = -\mathbf{g}_4$ et $\mathbf{g_2} = \mathbf{0} = \mathbf{g}_3$, donc $\mathbf{g_1 g_1}' = \mathbf{g_2 g_2}' = \mathbf{A}$. Avec $P_q = 0,25\ \forall q$, on obtient $\mathbf{G_4} = \mathbf{A}/2 = 2\mathbf{G_2}$. L'extension au cas de trois groupes est simple et conduit aux mêmes résultats.

γ_G, la plus grande valeur propre de $\mathbf{T}^{-1}\mathbf{G}_G$ vient de la résolution de l'équation $|\mathbf{T}^{-1}\mathbf{G}_G - \gamma_G \mathbf{I}| = 0$. Il est aisé de montrer que $\gamma_4 = 2\gamma_2$. En notant $\mathbf{x}_G$ le vecteur propre associé à γ_G, il reste donc juste à combiner $(\mathbf{T}^{-1}\mathbf{G}_2 - \gamma_2\mathbf{I})\,\mathbf{x}_2 = 0$ et $(\mathbf{T}^{-1}\mathbf{G}_4 - \gamma_4\mathbf{I})\,\mathbf{x}_4 = 0$, sachant que $\gamma_4 = 2\gamma_2$ et $\mathbf{G_4} = 2\mathbf{G_2}$, pour montrer que $\mathbf{x}_2 = \mathbf{x}_4$, alors que γ_G n'explique pas la même part des rapports de variabilités en fonction du nombre de groupes considéré.

3. Comparaisons effectuées

La méthode basée sur l'analyse discriminante que nous proposons a été comparée à deux méthodes multicaractères ainsi qu'à la méthode unicaractère décrite dans la première partie (4.1.). Les propriétés étudiées pour ces méthodes sont les temps de calcul requis, les puissances de détection et les précisions d'estimation des paramètres du QTL, localisation et effets. Ces comparaisons sont basées sur une utilisation intensive de simulations de Monte-Carlo.

3.1. Test d'hypothèse appliqué

Les hypothèses que nous avons comparées sont les hypothèses classiques rencontrées pour ce type d'étude dans la littérature. L'hypothèse nulle est la même que pour les analyses unicaractères : "il n'y a pas de QTL sur le groupe de liaison pour les caractères étudiés". L'hypothèse alternative, sachant que l'effet du QTL sur chaque caractère n'est pas testé, est : "il y a un QTL sur le groupe de liaison déterminant au moins un des caractères intégrés à l'analyse". En notant a_l l'effet du QTL sur le caractère l, avec $l = 1, \ldots, p$, on a donc : H0 : $a_l = 0 \quad \forall l = 1, \ldots, p$; H1 : $\exists$ au moins un l, $l = 1, \ldots, p$, tel que $a_l \neq 0$.

3.2. Méthodes comparées

Pour des raisons de cohérence et de facilité de comparaison avec la méthode unicaractère utilisée (voir chapitre 1, paragraphe 4.1.), toutes les méthodes développées sont basées sur la maximisation du rapport de vraisemblance correspondant aux hypothèses H0 et H1.

Par rapport à la méthode unicaractère (ST), le développement de méthodes multicaractères est basé sur une modification de la fonction de pénétrance (pour prendre en compte plus de variables), ou bien sur la modification de la nature des variables analysées (remplacement des performances par des variables synthétiques). Ces méthodes sont appliquées à des performances standardisées pour éviter toute confusion entre l'échelle de mesure des performances pour un caractère donné et l'ampleur de l'effet du QTL sur ce caractère. En dehors de la méthode DA décrite au paragraphe 2. (en pratique, la transformation par l'analyse discriminante a été réalisée à l'aide de la procédure g03acf de NAG), deux méthodes multicaractères ont été envisagées.

3.2.1. Analyse multivariée (MV)

La première des méthodes multicaractères que nous avons utilisées est la méthode de référence multivariée, définie selon les mêmes principes que Jiang et Zeng en 1995 [40] et Korol *et al.* en 1995 [57], et adaptée à l'analyse de structures de pedigree familiales. La vraisemblance s'écrit alors pour p caractères :

$$\Lambda^x = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \sum_{hd_{ij}} p(hd_{ij}/\widehat{hs_i}, M_i) \prod_{k=1}^{n_{ij}} f(\mathbf{yp}_{ijk}/\widehat{hs_i}, hd_{ij}, M_i) \tag{2.7}$$

où $\mathbf{yp}_{ijk} = \{yp_{ijkl}; l = 1, \ldots, p\}$ est le vecteur des performances standardisées pour le descendant ijk.

La fonction de pénétrance est une loi multinormale de dimension p qui s'écrit, en reprenant les notations du paragraphe 4.1., chapitre 1, équation 1.9 :

$$f(\mathbf{yp}_{ijk}/\widehat{hs_i}, hd_{ij}, M_i) = \sqrt{\frac{|\mathbf{VC}_i^{-1}|}{2\pi}} \exp\left(-\frac{1}{2}\mathbf{Y}'_{ijk}\mathbf{VC}_i^{-1}\mathbf{Y}_{ijk}\right) \tag{2.8}$$

où $\mathbf{Y}_{ijk} = \{Y_{ijkl}; l = 1, \ldots, p\}$ et $\mathbf{Y}'_{ijk}$ est le vecteur transposé,

$\mathbf{VC}_i$ est la matrice de (co)variance résiduelle entre les caractères pour la famille de père i :

$$\mathbf{VC}_i = \begin{pmatrix} \sigma^2_{i1} & \rho_{12} & \rho_{13} & \dots & \rho_{1p} \\ \rho_{12} & \sigma^2_{i2} & \rho_{23} & \dots & \rho_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \rho_{3p} & \dots & \sigma^2_{ip} \end{pmatrix} \tag{2.9}$$

où $\rho_{ll'}$ ($l \neq l'$, $l = 1, \dots, p$, $l' = 1, \dots, p$) est la corrélation résiduelle entre les caractères l et l',

$\mathbf{VC}_i^{-1}$ est son inverse et $|\mathbf{VC}_i^{-1}|$ le déterminant de l'inverse.

En dehors des coefficients de corrélation $\rho_{ll'}$, les paramètres estimés par cette méthode sont les mêmes que ceux estimés par la méthode unicaractère. Cependant, ils sont tous estimés dans une seule maximisation de la vraisemblance, ce qui, étant donné le nombre de paramètres à estimer par famille, peut devenir rapidement limitant pour la puissance. Par souci de simplification et pour limiter l'inflation du nombre de paramètres à estimer avec l'augmentation du nombre de caractères, nous avons choisi d'estimer des coefficients de corrélation entre les caractères communs pour l'ensemble des familles, malgré les observations de divers auteurs (Ronin *et al.*, 1995 [93], pour l'intégralité de la matrice de (co)variance), qui montrent que négliger les différences entre les familles quand elles existent peut limiter fortement les performances de détection de la méthode. Le nombre de paramètres à estimer pour chaque maximisation de la vraisemblance est alors de $(3n + 2\sum_{i=1}^{n} n_i)$ x $p + p(p-1)/2$.

Nous avons choisi de considérer la méthode multivariée comme une méthode de référence, sachant que dans le cadre de l'application aux populations animales à structure familiale, elle devient rapidement fortement limitante en terme de temps de calcul et de puissance du fait du nombre excessif de paramètres à estimer.

3.2.2. Analyse en composantes principales (PCA)

La deuxième méthode est la transformation des données proposée par Weller *et al.* (1996 [107]) par une analyse en composantes principales réalisée sur la base des données phénotypiques. Cette méthode a été adaptée à la cartographie d'intervalle. Nous l'avons décrite dans l'introduction (paragraphe 1.3.), de même que ses limites à la détection de QTL. Nous avons réalisé la transformation sur l'ensemble de la population, c'est à dire en particulier sans tenir compte de la structure familiale. L'étape de transformation a été entièrement programmée en fortran 90.

La variable en composantes principales qui représente au mieux le QTL ne pouvant pas être déterminée *a priori*, nous avons choisi d'analyser pour chaque jeu de données l'ensemble des variables en composantes principales obtenues après transformation. Dans la suite du document, on distinguera les différentes variables en composantes principales

par un numéro, PCAl étant la variable associée à la $l^{ième}$ plus grande valeur propre. Le nombre d'analyses réalisées est donc identique au nombre d'analyses réalisées par la méthode unicaractère pour l'analyse de tous les caractères. Cependant, les paramètres estimés caractérisent les composantes principales. Un des avantages largement revendiqué par Weller *et al.* (1996 [107]) pour l'utilisation de la méthode est la capacité, par une transformation *reverse* des effets estimés pour chaque variable en composantes principales, de retrouver l'estimation des effets du QTL sur chaque caractère. Cependant, nous verrons que les maxima de statistique de test étant localisés à des positions différentes pour chaque variable, cette transformation *reverse* est difficilement envisageable, et ne peut en aucun cas prétendre caractériser l'effet du QTL détecté grâce à une variable donnée.

Suites aux observations de Mangin *et al.* (1998) (voir 1.3.), nous avons par ailleurs calculé pour chaque position la somme des statistiques de test obtenues pour chaque variable en composantes principales (sPCA), la transformation étant toujours réalisée à partir de la variabilité phénotypique. Le maximum de la statistique de test ainsi obtenu a été conservé afin de le comparer aux valeurs de statistique de test obtenues par la méthode multivariée.

3.3. Axes de comparaison

3.3.1. Temps de calcul

Partant du principe que les p caractères intégrés aux analyses multicaractères sont préalablement analysés par la méthode ST, les temps de calcul de ST retenus sont ceux qui correspondent à l'analyse des p caractères. De la même façon, l'analyse de chaque variable de PCA n'étant en réalité pas caractéristique d'un QTL, les temps de calcul correspondant à cette méthode sont ceux de l'analyse des p variables obtenues. Ces temps de calcul sont aussi ceux de sPCA. Pour DA et MV, les temps de calcul correspondent aux analyses complètes, transformation comprise pour DA.

Les temps de calcul sont évalués à l'aide de la procédure x05baf de la bibliothèque NAG. Les résultats présentés sont les moyennes des temps de calcul estimés pour une structure génétique particulière. Les temps de calcul dépendent essentiellement du nombre de paramètres à estimer. Nous ne présenterons donc ici que les résultats obtenus sur les simulations destinées à comparer l'impact du nombre de caractères analysés sur les méthodes multicaractères.

3.3.2. Seuils de rejet

Les lois de distribution des méthodes multicaractères appliquées aux populations animales n'étant pas connues, nous avons choisi d'estimer les seuils de rejet de l'hypothèse H0 par simulation. Deux mille simulations ont été réalisées sous l'hypothèse d'absence de QTL, pour chaque situation simulée pour la détection de puissance, avec des corrélations résiduelles de 0,4 entre les caractères, une étude préalable ayant permis de montrer que la valeur de la corrélation résiduelle influence peu la valeur du seuil. Pour chacune de ces 2000 simulations, le maximum du rapport de vraisemblance obtenu sur le groupe de liaison pour chaque méthode est stocké. Les vecteurs des statistiques de test sont ensuite ordonnés. Une distribution empirique de la statistique de test sous l'hypothèse nulle est alors extrapolée à l'aide de la méthode proposée par Harrel et Davis (1982). Le quantile à $\alpha\%$ est ensuite retenu comme valeur seuil pour une erreur de première espèce de $\alpha\%$. Les $\alpha\%$ simulations dont les valeurs de statistique de test dépassent la valeur seuil ainsi déterminée représentent le pourcentage de cas d'absence de QTL où l'on accepte de déclarer de façon erronée la présence d'un QTL.

L'erreur nominale de première espèce α est différente en fonction des méthodes considérées, et plus particulièrement en fonction du nombre de variables analysées par la méthode. En effet, chaque variable, lors des analyses ST ou PCA, peut permettre de détecter un QTL à la position analysée, et représente donc un risque de fausse détection à elle seule. Etant donné que nous n'avons pas simulé de structures de corrélation très élevées entre les caractères, les variables analysées sont en général peu corrélées. Nous avons donc choisi, par souci de simplification, d'effectuer des approximations de la correction de Bonferroni pour le nombre de variables. Les seuils utilisés pour ces deux méthodes sont donc estimés pour des erreurs de première espèce de α/p, où α est l'erreur globale de 5% recherchée et utilisée pour le calcul des seuils de DA et MV. Les seuils ainsi obtenus sont néanmoins légèrement plus sévères que ceux utilisés pour les deux autres méthodes du fait de la légère corrélation entre les variables analysées. Par conséquent, dans certains cas, nous présenterons les puissances obtenues avec les seuils corrigés et non corrigés, ce qui fournit une fourchette de valeurs pour les puissances de détection.

Nous avons choisi de retenir des seuils caractérisant des erreurs au niveau du groupe de liaison. L'extrapolation de seuils au niveau du génome est aisément réalisable par une correction de Bonferroni pour le nombre M de groupes de liaison analysés : $\alpha^* = 1 - (1-\alpha)^M \approx \alpha/M$, où α est le seuil que l'on veut obtenir pour chaque groupe de liaison, et α^* le seuil global que l'on doit utiliser en conséquence. Le fait d'utiliser des seuils au niveau du groupe de liaison ne modifie pas nos conclusions quant aux performances relatives des méthodes. La seule différence attendue par la prise en compte de seuils plus sévères est une diminution des puissances observées au niveau d'un dispositif plus complet.

3.3.3. Puissances

Les puissances ont été estimées par simulation de données sous le modèle d'un QTL présent sur le groupe de liaison, déterminant de 1 à 6 caractères, pour des structures de populations expérimentales variées. Les maxima du rapport de vraisemblance pour chaque méthode sur 1000 simulations ont été stockés, et le pourcentage de fois où ces statistiques dépassent les seuils pré-estimés pour chaque méthode représente leur capacité de détection d'un QTL existant, soit leur puissance.

Pour ST, sauf précision particulière, les puissances et les estimations de paramètres présentées sont celles obtenues pour un caractère donné (le premier simulé), considéré comme le caractère d'intérêt. Les autres caractères simulés ont en général les mêmes caractéristiques, et sont considérés comme source d'information complémentaire. En ce qui concerne PCA, les puissances et les estimations de paramètres présentées sont celles associées à la variable permettant d'obtenir la meilleure puissance.

3.3.4. Estimations

Pour les 1000 simulations réalisées pour l'estimation des puissances, les estimations des paramètres correspondant au maximum du rapport de vraisemblance pour chaque méthode sont stockées. Nous nous sommes intéressés aux estimations de la position du QTL et des effets du QTL estimés pour les pères.

La précision d'estimation de chacun de ces paramètres est estimée avec la somme des carrés des erreurs (SCE) des estimations sur les 1000 simulations effectuées. Pour un paramètre θ, pour lequel on simule une valeur $\tilde{\theta}$ et on estime N valeurs θ_n, $n = 1, \ldots, N$, la somme des carrés des erreurs est égale à :

$$\mathrm{SCE} = \frac{1}{N} \sum_{n=1}^{N} (\theta_n - \tilde{\theta})^2 \tag{2.10}$$

En notant σ_θ l'écart type d'estimation du paramètre et β_θ son biais, on peut par ailleurs montrer aisément que SCE=$\sigma_\theta^2 + \beta_\theta^2$. Cet estimateur de la précision d'estimation des paramètres permet donc de synthétiser dans une même valeur la variance d'estimation et son biais. Nous distinguerons néanmoins parfois les biais et les variances d'estimation dans la discussion.

Pour chaque méthode, un vecteur de 1000 positions est obtenu à l'issue des simulations réalisées pour estimer les puissances. La SCE peut donc être calculée en appliquant directement la méthode proposée ci-dessus (voir 2.10).

En ce qui concerne les effets des allèles au QTL, une estimation est obtenue pour chaque

père et pour chaque simulation. Pour les méthodes ST et MV, les effets du QTL sont directement estimés, et des vecteurs de 1000 x n effets sont donc disponibles pour chaque caractère. En général et sauf précision, les résultats présentés pour ces méthodes sont ceux obtenus pour l'analyse du premier caractère simulé. Pour les deux méthodes basées sur la transformation des données par des combinaisons linéaires, seules les estimations des effets pour la(es) combinaison(s) linéaire(s) des caractères sont disponibles. Pour rendre les estimations des effets estimés pour ces variables avec DA et PCA comparables avec celles des autres méthodes, nous avons choisi de normaliser la moyenne des estimations obtenues en fonction des pondérations des caractères. Si on appelle γ_{ln} la pondération du caractère l pour la $n^{ième}$ simulation, en reprenant les notations de l'équation 2.10, nous avons donc appliqué la transformation suivante :

$$\mathrm{SCE} = \frac{1}{N} \frac{\sum_{n=1}^{N} (\theta_n - \tilde{\theta})^2}{\sum_{n=1}^{N} \sum_{l=1}^{p} \gamma_{ln} \sum_{n=1}^{N} \sum_{l=1}^{p} \gamma_{ln}} \tag{2.11}$$

La correction, bien que grossière, doit permettre de standardiser les ordres de grandeur des précisions d'estimation entre les méthodes.

3.4. Dispositifs simulés

Les caractéristiques des méthodes ont été évaluées sur la base de données simulées variant pour la structure des familles, la densité du marquage génétique et le déterminisme génétique des caractères étudiés. A chaque simulation, les données ont été analysées par les quatre méthodes successivement après standardisation des performances.

3.4.1. Pedigree

La population expérimentale simulée est un croisement F2 issu de deux populations fixées pour les allèles au QTL. Le QTL est donc biallélique. La taille du dispositif expérimental est constante, fixée à 500 descendants issus de 10 mâles F1 non-apparentés, quel que soit le nombre de mères simulé. La génération F0 est constituée de 10 couples mâle x femelles non-apparentés qui produisent chacun un mâle F1 et 1 ou 2 femelle(s) F1. Chaque mâle F1 est donc accouplé avec 1 ou 2 femelles (qui lui sont non-apparentées) en fonction du dispositif considéré.

Le nombre de paramètres à estimer pour chaque analyse des p caractères à une position est résumé dans le tableau 2.1.

TAB. 2.1 - *Nombre de paramètres à estimer pour chaque méthode multicaractère, en fonction du nombre de caractères analysés (p) et du nombre de descendants par mère (ndm).*

ndm	25					50				
p	2	3	4	5	6	2	3	4	5	6
ST	140	210	280	350	420	100	150	200	250	300
DA	70	70	70	70	70	50	50	50	50	50
PCA	140	210	280	350	420	100	150	200	250	300
MV	141	213	286	360	435	101	153	206	260	315

3.4.2. Carte génétique et génotypes

La carte génétique simulée est de 1 Morgan, avec 9 ou 3 marqueurs génétiques régulièrement répartis sur le groupe de liaison (densité de marquage de 12,5 ou 50 cM respectivement). Les marqueurs génétiques ont 5 allèles isofréquents chacun, indépendants de leur origine grand-parentale. Les allèles pour les marqueurs génétiques sont tirés aléatoirement pour les individus de la génération F0, et leur transmission est simulée selon les lois de Mendel pour les générations suivantes.

Le QTL est simulé à la position 31 cM, ce qui correspond à peu près au milieu d'un intervalle quelle que soit la densité en marqueurs simulée. Les allèles pour le QTL sont fixés dans les populations grand-parentales, les individus de la génération F1 sont donc tous hétérozygotes au QTL.

3.4.3. Performances

Les performances sont toujours simulées pour p caractères, le QTL ayant ou non un effet sur tous les caractères. Le modèle de simulation des performances est le suivant :
$\mathbf{y} = \mu + q\mathbf{a} + \mathbf{e}$
où $\mathbf{y} = \{y_l; l = 1, \ldots, p\}$, avec y_l la performance de l'individu pour le caractère l,

$\mu = \{\mu_l; l = 1, \ldots, p\}$, avec μ_l la moyenne de la partie polygénique des performances pour le caractère l, et μ tiré dans une loi multinormale de moyenne nulle et de matrice de (co)variance V_{pol}, telle que $h^2_l = 0.2$ et $\rho_{pol} = 0$. Dans la pratique, la transmission de la partie polygénique est simulée sur les trois générations.

$\mathbf{a} = \{a_l; l = 1, \ldots, p\}$, avec a_l l'effet de substitution des allèles au QTL pour le caractère l,

q la variable indicatrice du génotype de l'individu au QTL (q=1 si le génotype hérité au QTL est QQ; q=0 si le génotype hérité au QTL est Qq et q=-1 si le génotype hérité au QTL est qq),

$\mathbf{e} = \{e_l; l = 1, \ldots, p\}$, avec e_l l'erreur pour le caractère l, où $\mathbf{e}$ suit une loi multinormale

$(\mathbf{0},\ \mathbf{V}_{res})$, avec $\mathbf{V}_{res} = \begin{pmatrix} \sigma_1^2 & \rho_{12} & \rho_{13} & \dots & \rho_{1p} \\ \rho_{12} & \sigma_2^2 & \rho_{23} & \dots & \rho_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \rho_{3p} & \dots & \sigma_p^2 \end{pmatrix}$, en fixant $\sigma_l = 1$.

L'impact de quatre types de variation sur les performances des méthodes a été exploré :

1) En fonction du nombre de caractères analysés : avec 25 descendants par mère, 1 marqueur tous les 12,5 cM, des effets du QTL de 0,50 sur chaque caractère et une corrélation résiduelle entre les deux premiers caractères de 0,4 ou -0,4, les autres corrélations résiduelles étant fixées à 0,4, le nombre de caractères simulés varie de 2 à 6.

2) En fonction de la valeur de la corrélation génétique entre les caractères : avec 25 descendants par mère, 1 marqueur tous les 12,5 cM, 2 caractères, des effets du QTL égaux de 0,50 ou 0,30 sur chaque caractère, la corrélation résiduelle entre les deux caractères varie de -0,9 à 0,6.

3) En fonction de la pléiotropie du QTL : avec 25 descendants par mère, 1 marqueur tous les 12,5 cM, 2 caractères, une corrélation résiduelle entre les deux caractères de 0,6 ou -0,6, l'effet du QTL sur le caractère 2 simulé est de 0,50 ou 0,30, alors que l'effet du QTL sur le caractère 1 est nul.

4) En fonction de l'informativité du dispositif utilisé : pour 2 caractères, des effets du QTL égaux sur chaque caractère de 0,50 ou 0,30 et une corrélation résiduelle entre les deux caractères de 0,4 ou -0,4, la densité en marqueurs génétiques est égale à 12,5 ou 50 cM, et le nombre de descendants par mère est de 25 ou 50.

Le fait de simuler pour chaque cas des situations où ρ_{12} est négatif ou positif permet, sachant que tous les effets simulés sont positifs, de comparer des cas de détection où il existe un produit Pr négatif pour un couple de caractères aux cas où tous les Pr sont positifs. Dans la suite du document, on appellera $Pr_{ll'}$ le produit de la corrélation résiduelle entre les caractères l et l' avec les effets du QTL sur ces caractères.

3.4.3. Cas particulier des allèles au QTL non fixés en F0

L'impact de la non fixation des allèles d'un QTL biallélique dans les populations grand-parentales sur la capacité de détection des méthodes a été étudié. En effet, en particulier pour les méthodes basées sur des analyses de combinaisons des caractères, les transformations sont *a priori* proposées sur l'ensemble de la population, ce qui permet de travailler sur des matrices de (co)variances estimées sur des populations de grande taille. Cependant, toutes les familles de père ne sont pas nécessairement informatives pour le génotype au QTL. L'estimation de combinaisons linéaires intégrant ces familles peut donc

limiter la puissance de détection des méthodes (voir paragraphe 2.3.).

Nous avons réalisé quelques simulations pour ce type de situations. Par rapport aux simulations décrites aux paragraphes précédents, nous avons simulé 2, 5 ou 10 mâles dans la génération F1, en simulant 25 descendants par femelle. Le nombre de descendants par père varie donc de 250 à 50.

Le QTL reste biallélique, avec des allèles Q et q. On appelle fp(q) la fréquence de l'allèle q au QTL dans la population grand-paternelle et fm(q) la fréquence de l'allèle q au QTL dans la population grand-maternelle. Deux cas de non-fixation des allèles ont été explorés :
1) Les allèles ne sont fixés dans aucune des populations grand-parentales. Les situations simulées correspondent à fp(Q)=fm(q), qui varient de 1 à 0,5. Les quatre génotypes au QTL sont donc présents dans la génération F1.
2) Les allèles ne sont fixés que dans une des populations grand-parentales. Nous avons fait varier fp(Q) de 0,25 à 1, en maintenant fm(q)=1. Les deux classes, une d'hétérozygotes et une d'homozygotes qq, sont donc présentes en génération F1. Si l'on considère les croisements F2 réalisés pour la détection de QTL dans les populations animales, cette situation est plus réaliste que la précédente. En effet, à défaut de lignées que l'on peut supposer fixées pour un grand nombre de gènes, les généticiens animaux choisissent en général de croiser des animaux issus d'une lignée commerciale sélectionnée avec des animaux issus de lignées "exotiques" (type chinois ou sanglier chez le porc) présentant des performances divergentes pour les caractères d'intérêt. On peut alors supposer que la lignée sélectionnée aura tendance à porter des allèles au QTL agissant dans une même direction pour les caractères d'intérêt, alors que les allèles au QTL de la lignée exotique devraient être moins fixés.

Les allèles au QTL des populations F0 non fixées ont été tirés aléatoirement dans une loi uniforme de paramètre variable en fonction de la fréquence respective des deux allèles dans la population. La transmission des allèles au QTL a été simulée de façon identique à celle des allèles marqueurs. Le QTL est localisé comme précédemment à la position 31 cM. Il est pléiotrope sur deux caractères, avec des effets de substitution des allèles au QTL de 0,5 sur chaque caractère. Les caractères ont une corrélation résiduelle de -0,4.

Pour l'analyse des données issues des simulations décrites dans ce paragraphe, nous avons utilisé quatre méthodes multicaractères : DA (réalisée sur l'ensemble de la population), FDA (réalisée par famille, voir paragraphe 2.3.), PCA et sPCA, ainsi que la méthode ST.

4. Résultats

4.1. Temps de calcul

Les temps de calcul ont été estimés selon la procédure décrite au paragraphe 3.3. pour chaque type de simulations réalisé. Le tableau 2.2 présente les moyennes des temps de calcul obtenus en faisant varier le nombre de caractères pris en compte dans les analyses multivariées. Comme attendu, le facteur qui influence le plus les temps de calcul est le nombre de paramètres à estimer en une maximisation de vraisemblance (MV), ou en plusieurs maximisations successives (ST et PCA).

TAB. 2.2 - *Temps de calcul moyens des méthodes multicaractères, en secondes de CPU en fonction du nombre de caractères analysés*

Nombre de caractères analysés	2	3	4	5	6
ST	2,69	3,97	5,29	6,37	7,93
DA	2,25	2,20	2,22	2,12	2,24
PCA	2,73	3,95	5,27	6,24	8,12
MV	10,52	91,68	258,02	744,56	NE

Avec des corrélations résiduelles $\rho_{12} = -0.4$ et $\rho_{ll'} = 0.4$, $l \neq l'$, $l > 2$; $l' > 3$. Cas de 25 descendants par mère et une densité en marqueurs génétiques de 1 tous les 12,5 cM. ST : méthode unicaractère; DA : analyse discriminante; PCA : analyse en composantes principales; MV : méthode multivariée; NE = NonEstimé

Cependant, l'évolution des temps de calcul en fonction de ces deux critères est différente. Pour les méthodes basées sur l'analyse d'une variable par caractère (ST et PCA), les temps de calcul augmentent de façon linéaire avec le nombre de caractères analysés. Les temps de calcul de ST sont en général légèrement inférieurs à ceux de PCA, en raison du temps nécessaire à la transformation des variables. Pour MV, en revanche, le nombre de paramètres à estimer pour chaque maximisation de la vraisemblance augmente de $3n + 2\sum_{i=1}^{n} n_i + p$ quand on passe de l'analyse de p à $p + 1$ caractères, soit dans le cas simulé de $70 + p$ paramètres. Le fait de regrouper les paramètres dans une même maximisation entraîne donc une augmentation exponentielle des temps de calcul avec l'augmentation du nombre de caractères analysés. Les temps de calcul moyens pour 6 caractères analysés avec MV ne sont ainsi pas donnés en raison du faible nombre de simulations réalisées du fait de temps de calcul excessifs.

Les temps de calcul obtenus pour MV sont clairement prohibitifs. Cependant, ceux de PCA, et par conséquent ceux de sPCA, restent raisonnables pour le nombre de caractères analysés dans cette étude. Si cette méthode peut être utilisée pour l'estimation des seuils par simulations, qui est l'étape limitante en terme de temps de calcul dans la pratique, il semble donc qu'une solution alternative puisse être envisagée si l'on tient à réaliser des

détections multivariées.

4.2. Seuils de rejet de H0

Les seuils de rejet de l'hypothèse nulle H0 ont été estimés selon la procédure décrite au paragraphe 3.3.. Les estimations sont présentées dans le tableau 2.3, d'une part en fonction du nombre de caractères inclus dans l'analyse, et d'autre part en fonction de l'informativité du dispositif utilisé. Pour les méthodes impliquant l'analyse de plusieurs variables, les seuils corrigés par une approximation de la correction de Bonferroni (ST et PCA) et non corrigés sont présentés (STv et PCAv).

Les variations des seuils non corrigés pour le nombre de variables de STv et PCAv sont uniquement imputables à l'échantillonnage des données entre les simulations. Cet aspect permet de mettre en évidence que les seuils à 5% sont estimés avec une bonne précision avec 2000 simulations.

L'impact de l'augmentation des valeurs seuils en fonction du nombre de caractères analysés, et donc du nombre de paramètres à estimer, est relativement faible quand l'augmentation du nombre de paramètres est due à une augmentation du nombre de variables analysées. Les seuils corrigés de ST et PCA augmentent d'environ 5 points de vraisemblance quand on passe de 2 à 6 caractères analysés. On peut remarquer que le même type d'augmentation est observé pour les seuils de DA, ce qui signifie que la méthode permet bien de prendre en compte l'augmentation du nombre de caractères impliqués dans la transformation, et donc l'augmentation mécanique de l'informativité du processus.

Les seuils de MV augmentent de façon beaucoup plus rapide, en raison de l'inflation du nombre de paramètres à estimer dans chaque maximisation de la vraisemblance. On peut néanmoins remarquer que l'estimation des seuils par la somme des statistiques de test de PCA à chaque position (sPCA) donne effectivement une bonne approximation des seuils de la méthode multivariée, ce qui constitue une alternative beaucoup plus raisonnable en terme de temps de calcul.

La comparaison des seuils en fonction de l'informativité des données (simulations qui correspondent au point 4) du paragraphe 3.4.) conduit à des conclusions classiques : les seuils diminuent quand le système est moins informatif, de façon similaire pour toutes les méthodes (Knott *et al.*, 1996; Mangin *et al.*, 1999). L'augmentation du nombre de paramètres à estimer, en relation avec celle du nombre de mères (le nombre total de descendants étant constant), entraîne une augmentation des seuils plus importante proportionnellement que l'augmentation de la densité du marquage génétique.

TAB. 2.3 - *Seuils de rejet de H0 des méthodes multicaractères, en fonction du nombre de caractères analysés (p), de la densité de la carte génétique et du nombre de descendants par mère (ndm).*

p	2	3	4	5	6	2		
ndm			25			25	50	
Densité (cM)			12,5			50	12,5	50
ST	60,60	62,16	64,94	65,59	64,86	57,97	44,59	41,20
STv	56,40	58,03	57,16	57,79	57,62	53,68	41,19	38,12
DA	58,87	61,16	63,11	63,75	65,02	56,08	43,20	40,71
PCA	60,74	62,42	65,69	64,17	65,71	56,64	44,51	42,01
PCAv	56,49	57,91	58,54	57,04	56,76	53,75	41,47	38,12
MV	97,44	139,61	179,12	214,99	252,89	94,07	69,97	64,98
sPCA	98,83	138,77	179,21	216,46	253,13	93,08	69,89	65,24

Avec des corrélations résiduelles $\rho_{12} = -0,4$ et $\rho_{ll'} = 0,4$, $l \neq l'$, $l > 2$; $l' > 3$, pour une erreur de test globale de 5% sur le groupe de liaison. ST : méthode unicaractère; STv : méthode unicaractère, seuils non corrigés pour le nombre de variables analysées; DA : analyse discriminante; PCA : analyse en composantes principales; PCAv : analyse en composantes principales, seuils non corrigés pour le nombre de variables analysées; MV : méthode multivariée; sPCA : somme à chaque position des statistiques de test des variables de la PCA.

4.3. Puissances et qualité des estimations

4.3.1. En fonction du nombre de caractères

Les résultats commentés dans cette partie correspondent au cas 1) des simulations décrites au paragraphe 3.4.. De deux à six caractères corrélés sont simulés, avec Pr_{12} positif ou négatif.

4.3.1.1. Puissances Les résultats concernant les comparaisons de puissances en fonction du nombre de caractères analysés sont regroupés dans le tableau 2.4. La puissance de détection de ST est comme attendu quasi indépendante de la structure génétique du système. Cependant, étant donné que les valeurs seuil sont corrigées pour le nombre de caractères propres à chaque système, elles augmentent avec le nombre de caractères analysés, entraînant une diminution de la puissance associée à ST (de 23 à 12 % de détection). Par ailleurs, un ordre de grandeur de la variabilité d'estimation des puissances peut être obtenu en observant la ligne correspondant à l'estimation des puissance STv avec des seuils non corrigés, de même qu'en comparant deux à deux les estimations de puissance réalisées pour le même nombre de caractères et des corrélations résiduelles différentes. Cette variation semble relativement importante quand l'estimation des puissances est réalisée sur la base de 1000 simulations. La hiérarchie entre les méthodes étant cependant bien conservée d'une réplication des 1000 simulations à l'autre pour les mêmes paramètres de

simulation (non présenté), ce phénomène est essentiellement imputable à la variabilité d'échantillonnage des données. Son ampleur ne conditionne donc pas la pertinence des conclusions avancées ici.

TAB. 2.4 - *Puissances de détection des méthodes multicaractères et ST en fonction du nombre de caractères simulés (p) et du signe de la corrélation résiduelle entre les deux premiers caractères (ρ_{12}).*

ρ_{12}	0.4					-0.4				
p	2	3	4	5	6	2*	3	4	5	6
ST	23,6	15,9	12,0	12,2	14,3	22,1	18,3	12,3	11,3	12,3
STv	32,4	29,0	26,5	29,9	30,3	30,1	28,5	28,1	25,7	29,8
DA	43,8	41,1	40,1	41,5	44,3	81,9	79,1	79,6	85,9	97,6
PCA	36,9	36,1	30,0	38,1	38,0	73,8	46,6	36,8	41,4	39,2
MV	29,7	27,7	24,2	22,6	23,7	66,3	52,3	48,0	58,0	NE
sPCA	32,9	27,3	22,6	16,9	20,7	68,0	46,0	42,7	44,3	57,5

Avec $\rho_{12} = 0,4$ ou $-0,4$, avec $\rho_{ll'} = 0,4$, $l \neq l'$, $l > 2$; $l' > 3$. Cas d'un marqueur génétique tous les 12,5 cM, de 25 descendants par mère et d'effets de substitution du QTL égaux sur tous les caractères à 0,50. ST : méthode unicaractère; STv : méthode unicaractère, seuils non corrigés pour le nombre de variables analysées; DA : analyse discriminante; PCA : analyse en composantes principales (résultats obtenus avec l'analyse de la variable PCA1, sauf (*) : résultats avec l'analyse de la variable PCA2); MV : méthode multivariée; sPCA : somme à chaque position des statistiques de test des variables de la PCA. NE = NonEstimé.

Dans les conditions simulées, ST n'est pas très puissante, de 23 à 11% de détection. Quand Pr_{12} est négatif, les méthodes multicaractères permettent toutes, à des degrés divers, d'améliorer la puissance de détection par la prise en compte d'autres caractères. Ce type de résultat est attendu étant donnés les résultats existant dans la littérature (Jiang et Zeng, 1995 [40]; Korol *et al.*, 1995 [57]; Mangin *et al.*, 1998 [76]). Néanmoins, l'ampleur de la différence de puissance entre les deux cas est très importante, avec des puissances qui peuvent être multipliées par deux.

La puissance de MV est affectée par l'augmentation du nombre de caractères, surtout quand Pr_{12} est positif. Elle est même inférieure aux puissances de STv pour l'analyse de plus de 3 caractères dans ce cas. sPCA, basée sur la somme des statistiques de test des variables en composantes principales à chaque position, permet d'obtenir des puissances similaires à celles de la méthode multivariée, mais légèrement inférieures. Cette différence de valeur reflète la perte d'information que l'on crée en définissant la transformation des variables à partir de la matrice de (co)variance phénotypique. Cependant, pour les structures de caractères relativement simples que nous avons simulées, la différence reste limitée.

Les puissances de détection de DA sont toujours supérieures à celles obtenues par les autres méthodes, de 10 à 70%. Entre l'analyse de deux et six caractères, la puissance de la

méthode a tendance à augmenter avec l'augmentation du nombre de caractères analysés.

Toutes les puissances représentées dans le tableau 2.4 pour PCA concernent les valeurs obtenues avec PCA1, excepté le cas de 2 caractères quand le produit Pr_{12} est négatif, où il s'agit de PCA2. Les puissances de PCA sont inférieures de 15% environ à celles de DA quand Pr_{12} est positif. Ce cas est le cas optimal pour la mise en oeuvre de PCA, avec une structure des données qui permet de regrouper un maximum de l'information sur la variabilité des caractères due au QTL selon l'axe représentant la plus grande part de variabilité phénotypique, puisque tous les paramètres caractérisant le QTL sont positifs. Il ne s'agit cependant pas des conditions décrites précédemment commme étant les plus favorables à la détection de QTL en analyse multicaractère (Jiang et Zeng, 1995 [40]; Korol *et al.*, 1995 [57]; Mangin *et al.*, 1998 [76]).

Quand le produit Pr_{12} est négatif, PCA a un comportement particulier qui illustre bien les limites de la méthode. Quand deux caractères sont analysés, la puissance de détection de PCA est quasiment égale à celle de DA. Cependant, quand le nombre de caractères augmente, la puissance de PCA diminue fortement. Pour deux caractères, la variable en composantes principales qui permet d'obtenir la plus grande puissance est celle qui représente la moins grande part de variabilité phénotypique, en raison du signe négatif de la corrélation résiduelle entre les caractères. Cet axe permet néanmoins d'expliquer toute la variabilité due au QTL. En revanche, quand on analyse plus de deux caractères, la transformation sur la base de la variabilité phénotypique ne permet pas de regrouper l'information sur la variabilité due au QTL sur un seul axe. La plus grande part de cette information se retrouve selon le premier axe, qui est donc l'axe qui permet de détecter le plus de QTL (de l'ordre de 40%), tout en négligeant la part importante de l'information qui permet d'augmenter la distance entre les groupes d'haplotypes au QTL par rapport à une analyse unicaractère. Les puissances diminuent alors jusqu'à devenir inférieures à celles de la méthode multivariée.

4.3.1.2. Précision d'estimation de la position Les résultats sur la précision d'estimation des positions sont récapitulés dans le tableau 2.5. Nous avons représenté séparément les biais et les variances d'estimation afin de montrer sur ce premier exemple leur similarité en terme d'ordre de grandeur (respectivement 0,073 et 0,063 en moyenne pour ST) et d'évolution en fonction du nombre de caractères.

Globalement, la hiérarchie des méthodes en fonction du signe de Pr_{12} et du nombre de caractères est identique à celle observée pour les puissances, avec une amélioration de la précision qui peut aller jusqu'à un facteur 50 pour 6 caractères analysés et Pr_{12} négatif pour DA. Cependant, quand le produit est positif, la méthode qui permet d'obtenir la meilleure précision d'estimation de la position est PCA, avec une amélioration d'un facteur 2,2 pour le biais et 1,5 pour la variance d'estimation. Cette différence de hiérarchie avec

TAB. 2.5 - *Biais et variances d'échantillonnage des estimations des positions pour les analyses multicaractères en fonction du nombre de caractères analysés (p) et du signe de* Pr_{12}.

ρ_{12}	0.4					-0.4				
p	2	3	4	5	6	2*	3	4	5	6
	Biais de l'estimation des positions (x 10^2)									
ST	7,17	7,48	6,09	6,89	7,89	8,27	8,41	6,69	7,06	7,33
DA	7,41	4,84	5,04	5,73	5,92	1,88	1,22	1,72	0,70	0,33
PCA	5,63	2,99	3,42	2,54	3,17	2,51	3,22	3,28	0,82	3,30
MV	6,30	7,26	9,06	8,52	9,65	3,07	3,79	3,67	1,86	NE
	Variance de l'estimation des positions(x 10^2)									
ST	6,28	6,52	5,63	6,22	6,15	6,10	7,12	6,05	6,48	6,28
DA	5,86	4,92	5,03	4,60	4,54	2,00	1,76	1,68	1,16	0,59
PCA	4,74	4,20	3,44	3,01	3,00	2,60	3,13	3,38	2,34	3,17
MV	6,17	6,58	7,13	6,99	7,20	2,66	3,65	3,69	3,03	NE

Avec $|\rho_{12}| = 0,4$, avec $\rho_{ll'} = 0,4$, $l \neq l'$, $l > 2$; $l' > 3$. Cas d'un marqueur génétique tous les 12,5 cM, 25 descendants par femelle et des effets de substitution des allèles au QTL égaux sur chaque caractère à 0,50. ST : méthode unicaractère; DA : analyse discriminante; PCA : analyse en composantes principales (résultats obtenus avec l'analyse de la variable PCA1, sauf (*) : résultats avec l'analyse de la variable PCA2); MV : méthode multivariée. NE = NonEstimé.

les puissances vient de la correction des seuils pour le nombre de variables analysées. En effet, PCA1, prise individuellement sans correction du seuil pour le nombre de variables analysées, permet quand Pr_{12} est positif d'obtenir des puissances supérieures à DA.

4.3.1.3. Précision d'estimation des effets QTL Les résultats concernant la précision d'estimation des effets QTL sont regroupés dans le tableau 2.6. Les sommes des carrés des erreurs (SCE) ont été calculées tel que décrit au paragraphe 3.3..

Pour ST, les SCE sont en moyenne de 0,03, où le carré du biais représente moins de 1% de la valeur. Ce type de résultat est couramment observé pour l'estimation d'effets relativement faible : les estimations sont peu biaisées, et les variances d'estimation faibles (Knott *et al.*, 1996 [51]; Mangin *et al.*, 1998 [76]).

Quand Pr_{12} est négatif, la hiérarchie des méthodes multicaractères par rapport à la méthode unicaractère est identique aux observations sur les puissances. Les stratégies multicaractères permettent de diviser la SCE de la méthode unicaractère par 2,5 à 4 en utilisant DA.

La hiérarchie des méthodes multicaractères est différente des cas précédents pour la précision de l'estimation des effets quand le produit Pr_{12} est positif. PCA permet de

TAB. 2.6 - *Somme des Carrés des Erreurs (SCE) des estimations des effets QTL (x 10^2) pour les méthodes multicaractères en fonction du nombre de caractères analysés (p) et du signe de Pr_{12}.*

ρ_{12}			0.4					-0.4		
p	2	3	4	5	6	2*	3	4	5	6
ST	3,04	2,99	3,08	3,03	3,04	3,00	3,07	3,07	3,02	3,04
DA	2,90	2,93	2,97	3,26	3,52	1,16	1,19	1,20	1,09	0,75
PCA	1,86	1,81	1,72	1,63	1,55	1,18	1,51	1,54	1,50	1,47
MV	2,67	2,73	2,86	2,74	2,78	2,60	2,60	2,64	2,52	UE

Avec $|\rho_{12}| = 0,4$, avec $\rho_{ll'} = 0,4$, $l \neq l'$, $l > 2$; $l' > 3$. Cas d'un marqueur génétique tous les 12,5 cM, 25 descendants par femelle et des effets de substitution des allèles au QTL égaux sur chaque caractère à 0,50. ST : méthode unicaractère; DA : analyse discriminante; PCA : analyse en composantes principales (résultats obtenus avec l'analyse de la variable PCA1, sauf (*) : résultats avec l'analyse de la variable PCA2); MV : méthode multivariée. NE = NonEstimé.

réduire la SCE de ST d'un facteur 1,5 à 2 en fonction du nombre de caractères analysés. La réduction de la SCE par MV est de l'ordre de 10%. En revanche, pour DA, les SCE obtenues sont à peine inférieures à celles de ST quand on intègre jusqu'à 4 caractères dans l'analyse, et deviennent supérieures de 7 à 14 % quand le nombre de caractères augmente encore.

Les différences observées ici s'expliquent en s'intéressant à la part de la SCE représentée par le carré du biais de l'estimation des effets. Pour MV et PCA, la part due au carré du biais est constante quel que soit le nombre de caractères analysés, de l'ordre de 1,6% (supérieur à celui observé pour ST) et 0,16 % respectivement. Pour DA, on observe un accroissement important du biais avec le nombre de caractères quand Pr_{12} est positif, le carré du biais représentant de 1,1 à 16% de la SCE. Il est donc supérieur au carré du biais observé pour ST, alors que la variance d'estimation reste comprise entre celle de ST et celle de MV. Une analyse détaillée des résultats obtenus pour DA montre que lorsque la méthode ne permet pas de discriminer les groupes d'haplotypes au QTL par la transformation, les coefficients de la combinaison linéaire ont tendance à être très faibles. La moyenne des sommes de coefficients utilisée pour corriger la SCE et la rendre comparable à celles obtenues pour ST et MV (voir paragraphe 3.3., équation 2.11) devient alors très petite, entraînant une inflation de la SCE corrigée. La correction proposée semble donc n'être vraiment pertinente que quand la méthode permet d'obtenir des puissances de détection élevées, c'est à dire pour des effets forts du QTL et/ou des Pr négatifs.

4.3.2. En fonction de la structure génétique

Les résultats commentés dans cette partie sont issus des simulations décrites au point 2) du paragraphe 3.4.. Il s'agit de deux caractères déterminés par des effets égaux du QTL, et des corrélations résiduelles variables.

4.3.2.1. Puissances Les résultats obtenus sur les puissances en fonction des valeurs de corrélation résiduelle entre les deux caractères et les effets du QTL sont synthétisés dans le tableau 2.7. Nous avons à nouveau indiqué à la fois les puissances obtenues par l'utilisation de seuils corrigés pour le nombre de variables analysées (ST) et les puissances obtenues à partir de seuils non-corrigés (STv) pour la méthode unicaractère.

TAB. 2.7 - *Puissances pour les méthodes multicaractères en fonction de la corrélation résiduelle (ρ_{12}) et des effets de substitution des allèles au QTL (α_1 ; α_2).*

ρ_{12}	-0,9*	-0,6*	-0,4*	-0,1	0	0,1	0,4	0,6
$\alpha_1 = \alpha_2$				0,5				
ST	20,6	21,6	22,1	22,1	20,8	22,4	23,6	21,9
STv	30,2	29,6	30,1	30,8	31,1	30,1	32,4	29,4
DA	100,0	95,6	81,9	62,9	55,1	50,3	43,8	25,6
PCA	100,0	94,5	73,8	57,1	48,2	42,8	36,9	28
MV	99,2	85,2	65,9	48,2	39,7	34,9	29,7	24,7
sPCA	99,4	87,4	68,0	49,9	43,7	38,3	32,9	26,7
$\alpha_1 = \alpha_2$				0,3				
ST	7,0	6,1	6,1	5,8	6,2	7,0	6,7	7,2
STv	10,3	10,8	12,3	10,2	11,1	11,1	11,3	11,6
DA	67,2	35,3	29,0	12,1	17,2	16,8	14,2	10,7
PCA	63,5	33,6	20,6	11,8	12,3	11,8	9,0	9,1
MV	51,8	28,6	18,3	10,7	11,6	12,4	9,6	9,8
sPCA	54,6	31,7	21,7	12,7	13,9	13,9	11,9	11,7

Cas d'un marqueur génétique tous les 12,5 cM et de 25 descendants par femelle. ST : méthode unicaractère; STv : méthode unicaractère, seuils non corrigés pour le nombre de variables analysées; DA : analyse discriminante; PCA : analyse en composantes principales (résultats obtenus avec l'analyse de la variable PCA1, sauf (*) : résultats avec l'analyse de la variable PCA2); MV : méthode multivariée; sPCA : somme à chaque position des statistiques de test des variables de la PCA.

Pour des effets de substitution des allèles au QTL de $0{,}5\sigma_p$, en moyenne 21,9 % des QTL simulés sont détectés par ST. Quand on passe à des effets du QTL de $0{,}3\sigma_p$, la méthode ne permet plus de détecter que 6,5 % des QTL, soit à peine plus que l'erreur de première espèce de 5%. La perte de puissance observée pour les méthodes multicaractères entre les deux cas est du même ordre de grandeur. Le facteur de division est de l'ordre de 3,5, excepté quand la corrélation résiduelle entre les caractères est très fortement négative (-0,9). La puissance ne chute alors que d'un facteur 1,7 en moyenne. Cependant,

pour cette valeur de corrélation et des effets forts du QTL, les méthodes multicaractères permettent d'atteindre 100% de puissance. La différence de chute de puissance n'est donc pas imputable à un comportement différentiel des méthodes en fonction de la valeur de la corrélation résiduelle.

La hiérarchie des méthodes multicaractères correspond à celle décrite pour les cas deux caractères dans la partie précédente (4.3.1). On observe cependant en général une légère supériorité de la puissance de sPCA sur la puissance de MV. Cette différence peut être due à la simplicité de la structure génétique des caractères, qui permet à la transformation en composantes principales de compenser le nombre de paramètres à estimer par la méthode multivariée.

Le point le plus remarquable dans cette comparaison est l'ampleur de la variation de la puissance des méthodes multicaractères en fonction de la valeur de la corrélation résiduelle, et donc du produit Pr_{12}. Quand les effets du QTL sont égaux à $0{,}5\sigma_p$, les puissances passent de 100% de détection pour une corrélation égale à -0,9 à environ 25% (soit la puissance de ST) pour une corrélation de 0,6. Quand les effets sont de $0{,}3\sigma_p$, les méthodes multicaractères perdent très rapidement en puissance de détection jusqu'à une corrélation résiduelle de -0,1. La tendance à la diminution de la puissance semble ensuite plus lente. Cependant, les niveaux de détection devenant très proches de l'erreur de première espèce considérée pour l'estimation des seuils, les variations de puissance deviennent difficilement interprétables. Comme remarqué précédemment, ce type de résultats est très largement attendu étant données les conclusions de la littérature (voir 1..2 et 1..3).

Par ailleurs, la variable en composantes principales utilisée dans cette partie est soit PCA1, soit PCA2, en fonction du rapport entre le signe de la corrélation résiduelle et celui de la corrélation phénotypique. Quand les deux vont dans le même sens ($\rho_{12} \in [-0,1;0,6]$), PCA1 permet de détecter le QTL. Dans les autres cas ($\rho_{12} \in [-0,9;-0,4]$), on utilise PCA2. Ce critère semble simple mais n'est dans la pratique pas utilisable puisque les valeurs des corrélations résiduelles ne sont pas connues.

Enfin, dans le tableau 2.7 sont présentés les résultats de simulation de caractères corrélés uniquement en raison de la présence du QTL pléiotrope (ρ_{12}=0). La prise en compte d'autres caractères corrélés au caractère d'intérêt uniquement en raison de l'action du QTL étudié permet donc d'améliorer la détection par rapport aux méthodes unicaractères.

4.3.2.2. Précision d'estimation de la position Les SCE pour l'estimation de la position sont récapitulées dans le tableau 2.8. Pour des effets du QTL sur chaque caractère respectivement de $0{,}5\sigma_p$ et $0{,}3\sigma_p$, les SCE obtenues pour ST sont en moyenne de 0,066 et 0,115.

De façon attendue, la même tendance que pour les puissances se dégage, avec une

TAB. 2.8 - *Somme des Carrés des Erreurs (SCE) des estimations de la position du QTL ($x\ 10^2$) en fonction de la corrélation résiduelle (ρ_{12}) et des effets de substitution des allèles au QTL (α_1, α_2).*

ρ_{12}	-0,9*	-0,6*	-0,4*	-0,1	0	0,1	0,4	0,6
$\alpha_1 = \alpha_2$					0,5			
ST	6,91	6,56	6,78	6,76	6,08	6,58	6,79	6,10
DA	1,36	0,79	2,03	3,16	3,59	5,52	6,41	5,76
PCA	0,44	0,75	2,66	2,88	2,94	4,11	5,05	5,49
MV	0,74	1,37	2,76	4,26	4,48	5,44	6,56	7,06
$\alpha_1 = \alpha_2$					0,3			
ST	11,50	11,40	10,90	11,7	11,90	11,40	10,70	12,35
DA	4,99	6,14	8,04	9,96	8,24	10,70	11,10	11,92
PCA	2,60	5,10	7,32	8,48	8,41	9,94	10,60	12,00
MV	4,27	6,70	8,20	10,70	9,32	11,60	11,80	13,07

Cas d'un marqueur génétique tous les 12,5 cM et de 25 descendants par femelle. ST : méthode unicaractère; DA : analyse discriminante; PCA : analyse en composantes principales (résultats obtenus avec l'analyse de la variable PCA1, sauf (*) : résultats avec l'analyse de la variable PCA2); MV : méthode multivariée.

amélioration de la précision de l'estimation de la position par l'utilisation de méthodes multicaractères d'autant plus importante que la corrélation entre les caractères est fortement négative et que les effets sont élevés (dans le cadre des valeurs d'effets et de corrélation que nous avons simulé), avec des précisions d'estimation de la localisation diminuées jusqu'à plus de 10 fois.

La hiérarchie des méthodes multicaractères est très similaire à celle observée lors des comparaisons de précision d'estimation des positions en fonction du nombre de caractères simulés. PCA permet en général d'obtenir les meilleures précisions d'estimation de la position, suivie par DA puis MV.

Enfin, pour des effets du QTL relativement faibles ($0{,}3\sigma_p$), les précisions d'estimation de la position par les méthodes multicaractères n'atteignent le niveau des SCE de la méthode unicaractère que pour des corrélations résiduelles de l'ordre de 0,6, contrairement à ce que nous avons pu observer précédemment concernant les puissances.

4.3.2.3. Précision d'estimation des effets QTL Les SCE pour l'estimation des effets de substitution des allèles au QTL sont résumées dans le tableau 2.9.

Les résultats obtenus pour ST sont indépendants de la corrélation résiduelle entre les caractères, ce qui est attendu, mais aussi relativement peu affectés par la diminution des effets du QTL de 0,5 à $0{,}3\sigma_p$ (passage de 0,0297 à 0,0313 en moyenne). De façon plus surprenante, les SCE calculées sur les estimations des effets pour MV sont aussi

relativement indépendantes de la corrélation résiduelle simulée. Les SCE n'augmentent que de 11% en moyenne, quels que soient les effets simulés, quand on passe d'une corrélation de -0,9 à une corrélation de +0,6. Elles sont en moyenne inférieures de 10% aux SCE obtenues pour ST.

TAB. 2.9 - *SCE des estimations des effets de substitution au QTL (x 10^2) en fonction de la corrélation résiduelle (ρ_{12}) et des effets de substitution des allèles au QTL (α_1, α_2).*

ρ_{12}	-0,9*	-0,6*	-0,4*	-0,1	0	0,1	0,4	0,6
$\alpha_1 = \alpha_2$				0.50				
ST	2,94	2,87	2,97	3,03	2,93	3,02	3,00	2,99
DA	0,43	0,76	0,81	1,24	1,93	2,28	2,94	2,26
PCA	0,41	0,72	1,11	1,43	1,56	1,77	2,07	2,37
MV	2,38	2,45	2,65	2,75	2,73	2,83	2,82	2,92
$\alpha_1 = \alpha_2$				0.30				
ST	3,12	3,12	3,13	3,11	3,16	3,08	3,15	3,13
DA	0,60	1,34	1,30	2,35	3,75	4,52	6,38	4,87
PCA	0,46	0,85	1,19	1,66	1,72	1,89	2,27	2,58
MV	2,69	2,79	2,79	2,92	2,95	2,87	2,97	2,95

Cas d'un marqueur génétique tous les 12,5 cM et de 25 descendants par femelle. ST : méthode unicaractère; DA : analyse discriminante; PCA : analyse en composantes principales (résultats obtenus avec l'analyse de la variable PCA1, sauf (*) : résultats avec l'analyse de la variable PCA2); MV : méthode multivariée.

Pour PCA, les précisions d'estimation des effets sont directement liées à la valeur de la corrélation résiduelle simulée entre les caractères, avec une augmentation des SCE de 0,0040 à 0,0237 pour des corrélations de -0,9 à 0,6 et des effets du QTL de $0{,}5\sigma_p$, et de 0,0046 à 0,0258 quand les effets sont de $0{,}3\sigma_p$.

Pour DA, la même tendance est observée que pour la comparaison des SCE des effets estimés en fonction du nombre de caractères. Quand les puissances de détection sont relativement élevées, les SCE sont du même ordre de grandeur que celles obtenues pour PCA. Quand les puissances de détection deviennent faibles, les SCE augmentent beaucoup, en raison d'un accroissement important du biais de l'estimation, jusqu'à devenir deux fois supérieures aux SCE calculées pour ST quand les effets sont faibles et la corrélation positive et élevée. De la même façon que nous l'avions observé au paragraphe 4.3.1., cette inflation des SCE semble due à des coefficients des combinaisons linéaires très faibles quand l'analyse discriminante ne parvient pas à bien séparer les groupes d'haplotypes, ce qui conduit à la construction d'un paramètre de correction de la SCE excessivement faible.

4.3.3. En fonction de la pléiotropie du QTL

Les résultats commentés dans cette partie sont issus des simulations décrites au point 3) du paragraphe 3.4.. Il s'agit de deux caractères résiduellement corrélés, dont un seul est déterminé par le QTL (le caractère 2), avec une corrélation résiduelle positive ou négative entre les deux caractères. Nous ne présenterons ici que les puissances et les précisions d'estimation de la position du QTL, en distinguant pour la méthode unicaractère les résultats obtenus sur le caractère 1 et ceux obtenus sur le caractère 2.

TAB. 2.10 - *Puissances et SCE des positions estimées par les méthodes multicaractères en fonction du signe de la corrélation résiduelle (ρ_{12}) et de la valeur de l'effet de substitution des allèles au QTL sur le premier caractère (α_2), avec $\alpha_1 = 0$.*

	Puissance (%)				SCE des positions estimées (x 10^2)			
α_2	0,50		0,30		0,50		0,30	
ρ_{12}	-0,6	0,6	-0,6	0,6	-0,6	0,6	-0,6	0,6
ST1	3,8	2,7	2,6	2,2	14,70	15,73	16,15	15,91
STv1	6,2	4,3	4,4	4,5				
ST2	20,0	21,9	7,5	6,4	6,58	7,05	11,36	11,90
STv2	29,3	31,4	11,7	11,3				
MV	24,5	38,1	8,4	11,5	5,68	4,34	11,65	10,11
DA	35,1	54,4	11,5	15,7	7,50	5,38	12,63	10,40
PCA	17,1	34,8	6,0	8,7	7,28	5,84	11,99	11,03
sPCA	26,6	40,0	11,4	13,1				

Cas de deux caractères, un marqueur génétique tous les 12,5 cM et 25 descendants par mère. STl : méthode unicaractère pour l'analyse du $l^{ième}$ caractère; STlv : méthode unicaractère pour l'analyse du $l^{ième}$ caractère, seuils non corrigés pour le nombre de variables analysées; DA : analyse discriminante; PCA : analyse en composantes principales (résultats obtenus avec l'analyse de la variable PCA1); MV : méthode multivariée; sPCA : somme à chaque position des statistiques de test des variables de la PCA.

Les résultats obtenus sont regroupés dans le tableau 2.10. Les résultats obtenus avec ST pour le caractère 1, sur lequel le QTL n'agit pas, permettent de valider le calcul des seuils pour cette méthode, puisqu'on observe des puissances qui correspondent bien à l'erreur de première espèce de 2,5% fixée pour l'analyse de chaque caractère. De la même façon, on peut montrer que la moyenne des positions estimées est au milieu du groupe de liaison quand aucun QTL n'existe. Les résultats obtenus ici pour l'analyse du caractère 2 par ST correspondent à des situations commentées dans les paragraphes précédents, nous n'y reviendrons donc pas.

Les puissances de détection des méthodes multicaractères confirment l'intérêt potentiel de la prise en compte de caractères résiduellement corrélés avec le caractère d'intérêt pour la détection de QTL (Jiang et Zeng, 1995 [40]; Korol *et al.*, 1995 [57]; Mangin *et al.*, 1998 [76]). La seule méthode qui permette d'améliorer de façon importante la

puissance de détection par rapport à ST2 pour les 4 situations envisagées est DA, avec une multiplication de la puissance par un facteur 2,48 dans le cas le plus favorable. Les autres méthodes multicaractères sont toujours moins puissantes quand la corrélation résiduelle entre les deux caractères est négative. PCA, avec l'analyse de PCA1, a toujours une puissance de détection inférieure à l'analyse unicaractère du deuxième caractère, sauf dans le cas le plus favorable d'un effet important du QTL sur le caractère et d'une corrélation résiduelle positive. Dans ce cas particulier, la variabilité due au QTL est relativement répartie sur les deux axes de la transformation phénotypique. PCA2 permet alors de détecter une partie des QTL, avec des puissances de l'ordre de 10%.

Quand l'effet du QTL est relativement faible ($0{,}3\sigma_p$), les méthodes multicaractères permettent d'améliorer de façon importante la probabilité de détection du QTL (passage d'une probabilité de 6,4 à 15,7% de détection), mais sans atteindre des puissances très élevées.

En ce qui concerne l'amélioration de la précision de la localisation du QTL, la méthode le plus performante est MV. Quand la corrélation résiduelle est négative, MV est la seule méthode qui permet d'améliorer la précision d'estimation de la position. DA et PCA sont sensiblement équivalentes, avec des SCE 1,25 fois plus élevées en moyenne que celles de MV quand l'effet du QTL est de $0{,}5\sigma_p$, et 1,05 fois quand il est de $0{,}3\sigma_p$.

L'amélioration de la détection d'un QTL par la prise en compte d'un caractère corrélé au caractère affecté semble donc limitée à la détection de QTL ayant des effets relativement importants.

4.3.4. En fonction de l'informativité du dispositif expérimental

Les résultats commentés dans cette partie correspondent aux simulations décrites au paragraphe 4) de la partie 3.4.. Nous avons comparé les puissances de détection et les précisions d'estimation des paramètres en fonction de l'informativité du dispositif expérimental. Nous avons fait varier le nombre de femelles dans le pedigree, à taille globale de dispositif constante, ce qui modifie essentiellement le nombre de paramètres estimés, et la densité en marqueurs génétiques, pour deux niveaux d'effets du QTL égaux sur deux caractères. Pour l'ensemble de ces comparaisons, les hiérarchies entre les différentes méthodes ne sont pas modifiées par rapport aux cas d'analyse de deux caractères que nous avons développés précédemment. Nous n'y reviendrons donc pas. Les résultats sont récapitulés dans le tableau 2.11.

Les puissances des analyses unicaractères sont en moyenne multipliées par un facteur 3,5 quand on utilise des méthodes multicaractères, avec une amélioration d'autant plus importante que la quantité d'information disponible dans le système est élevée. Dans le

TAB. 2.11 - *Puissances, SCE des estimations des positions et SCE des estimations des effets QTL des méthodes multicaractères en fonction de la densité en marqueurs génétiques et du nombre de descendants par femelle (ndm).*

$\alpha_1 = \alpha_2$	0,50				0,30			
Densité (cM)	12,5		50		12,5		50	
ndm	50	25	50	25	50	25	50	25
Puissance (%)								
ST [a]	26,50	22,10	14,00	7,20	7,00	6,10	5,00	4,80
STv	38,00	30,10	19,60	13,00	12,90	12,30	9,00	8,90
DA	90,30	81,90	54,20	41,40	31,50	29,00	15,00	12,90
PCA	87,30	73,80	52,70	37,10	27,30	20,60	11,70	11,60
MV	80,80	65,90	45,70	31,70	22,60	18,30	14,00	12,80
sPCA	81,50	68,00	45,40	29,40	23,40	21,70	13,60	12,50
SCE des estimations des positions (x 10^2)								
ST	5,53	6,78	8,51	8,29	11,28	10,92	11,28	12,01
DA	1,10	2,03	3,68	4,71	7,62	8,04	9,48	9,02
PCA	1,30	2,66	3,21	4,14	6,56	7,32	8,27	8,64
MV	1,56	2,76	4,49	5,12	7,86	8,20	9,67	9,44
SCE des estimations des effets (x 10^2)								
ST	2,81	2,97	83,09	60,38	3,08	3,17	94,36	73,56
DA	1,04	1,15	30,38	31,29	1,96	2,04	69,19	58,81
PCA	1,03	1,14	41,65	38,39	1,24	1,18	60,64	53,15
MV	2,43	2,58	51,22	43,64	2,79	2,84	68,13	53,21

Cas $\alpha_1 = \alpha_2$ pour le QTL pléiotrope sur deux caractères, avec $\rho_{12} = -0,4$. ST : méthode unicaractère; STv : méthode unicaractère, seuils non corrigés pour le nombre de variables analysées; DA : analyse discriminante; PCA : analyse en composantes principales (résultats obtenus avec l'analyse de la variable PCA2); MV : méthode multivariée; sPCA : somme à chaque position des statistiques de test des variables de la PCA.

cas où les effets du QTL sont faibles ($0,3\sigma_p$) et les marqueurs génétiques peu denses (1 tous les 50 cM), les méthodes multicaractères permettent de passer de puissances de l'ordre de l'erreur de première espèce avec ST à des puissances supérieures à 10%, ce qui reste relativement faible.

L'amélioration de la précision d'estimation de la position est très nettement dépendante de la densité en marqueurs génétiques, avec des facteurs de gains de 3,4 (1,6) quand les effets sont de $0,5\sigma_p$ ($0,3\sigma_p$), alors qu'ils ne sont que de 2,3 (1,4) pour des marqueurs génétiques moins denses. Les trois méthodes multicaractères permettent des gains sensiblement équivalents, avec toujours une meilleure précision pour les méthodes basées sur l'analyse de combinaisons linéaires des caractères.

La réduction des SCE des estimations des effets QTL obtenus avec ST par l'utilisation de méthodes multicaractères est surtout importante quand les cartes génétiques sont denses. Le facteur de réduction est alors de 2,6 environ pour DA et PCA, alors qu'il est de

1,6 quand la carte génétique est moins dense. En ce qui concerne l'analyse discriminante, on n'observe pas ici d'augmentation du biais d'estimation des effets avec la diminution de la puissance. MV réduit les SCE de 10 à 40 % par rapport à ST.

En résumé, les méthodes multicaractères permettent en général d'améliorer les résultats obtenus avec la méthode unicaractère. Cependant, cette amélioration est beaucoup plus importante quand la carte génétique est relativement dense.

4.3.5. En fonction du degré de fixation des allèles au QTL

Les résultats commentés dans cette partie sont issus des simulations réalisées selon les conditions décrites au paragraphe 3.4.. Nous ne commenterons ici que les évolutions de la puissance et de la précision d'estimation de la position en fonction du nombre de descendants par père (ndp) et de la fréquence des allèles au QTL dans les populations grand-parentales.

Etant données les caractéristiques des caractères simulés, seuls les résultats obtenus pour PCA2 pour l'analyse en composantes principales seront présentés.

4.3.5.1. Allèles non fixés dans les deux populations grand-parentales Dans cette partie, les fréquences des allèles minoritaires sont égales dans les deux populations F0. Nous avons donc fp(q)=fm(Q)=1-fp(Q)=1-fm(q) $\in [0;0,5]$ (de 0,5 à 1, les résultats sont symétriques car les effets des allèles au QTL sont strictement additifs).

Nous avons représenté figure 11 l'évolution de la fréquence des classes de génotypes au QTL dans la génération F1 correspondant à l'éventail des cas simulés. Le génotype attendu en F1 est Qq si l'on fait l'hypothèse d'allèles fixés au QTL en F0. Quand les fréquences des allèles minoritaires en F0 augmentent jusqu'à 0,5, la fréquence des hétérozygotes décroît presque linéairement jusqu'à atteindre 0,5 pour les deux classes d'hétérozygotes confondues. Ce type de situation doit en principe perturber la qualité de détection de DA. En effet, cette méthode repose sur la répartition des individus en classes génotypiques au QTL. Cette stratégie implique en particulier que les allèles au QTL soient différents en fonction de leur origine grand-parentale, pour créer une variabilité inter-classes génotypiques identifiable pour la transformation. Dans le cas de fréquences des génotypes au QTL toutes égales dans la génération F1, on s'attend donc à ce que les QTL soient très mal détectés.

En ce qui concerne PCA, la transformation est basée sur la forme globale du nuage de points initial. Or l'augmentation de la fréquence des homozygotes en F1 "tasse" le nuage, mais ne le déforme pas dans une direction préférentielle. On ne s'attend donc pas à une perte de puissance supérieure à celle due à la diminution de la quantité d'information

disponible dans le dispositif. Comme référence, nous avons utilisé d'une part ST, et d'autre part sPCA, qui, comme montré précédemment et confirmé sur quelques exemples présentés ici, permet dans les cas de déterminisme génétique simple des caractères de reproduire de façon fidèle le comportement de MV.

a. Puissance L'ensemble des résultats obtenus sur la puissance de détection quand les allèles au QTL ne sont pas fixés dans les populations grand-parentales est résumé dans le tableau 2.12.

Pour ST, l'augmentation du nombre de descendants par père ne permet d'améliorer la puissance de détection que de 40 % quelle que soit la fréquence des allèles au QTL en F0. En revanche, l'augmentation de la fréquence des homozygotes au QTL à la génération F1 entraîne une diminution d'un facteur 2,3 en moyenne de la capacité de détection des QTL, avec par exemple une chute de 20,3% de détection quand les allèles au QTL sont fixés en F0, à 8,7% quand les allèles sont isofréquents dans les populations grand-parentales.

D'une façon générale, la chute de puissance des méthodes multicaractères avec l'augmentation de la fréquence des homozygotes en F1 est du même ordre de grandeur en proportion que pour ST. Cependant, les méthodes multicaractères permettent toujours de détecter environ trois fois plus de QTL que ST.

Pour toutes les méthodes multicaractères, l'augmentation du nombre de descendants par père permet d'améliorer nettement la puissance de détection. L'amélioration est d'autant plus prononcée que la proportion d'allèles minoritaires dans les populations F0 augmente. L'augmentation de puissance la plus importante est observée pour FDA, ce qui s'explique aisément par une meilleure estimation des matrices de (co)variance intra-familles avec l'augmentation du nombre de descendants par famille.

Quand les allèles au QTL sont fixés en F0, DA permet de détecter un maximum de QTL, suivie par PCA. sPCA permet de détecter moins de QTL que les autres méthodes multicaractères, avec 5 à 10% de détection en moins. FDA permet des niveaux de détection équivalents à DA et PCA quand le nombre de descendants par père est élevé, et des niveaux de détection équivalents à ceux de sPCA quand *ndp* est faible. Cette différence indique qu'un nombre minimum de descendants par père doit être disponible pour une estimation précise des matrices de (co)variance intra-famille.

Quand la fréquence des homozygotes augmente en F1, la hiérarchie entre les méthodes multicaractères se modifie progressivement, et les écarts de puissance entre les méthodes multicaractères se réduisent. En effet, la puissance de DA se dégrade plus vite que celles des autres méthodes. Pour des fréquences des génotypes au QTL égales en F1, les puissances de DA sont donc les plus faibles de quelques % (sauf pour ndp=50 où sPCA est inférieure). A l'inverse, les puissances de FDA sont moins rapidement dégradées que celles des autres

TAB. 2.12 - *Puissances des méthodes multicaractères en fonction de la fréquence des allèles au QTL dans les populations grand-parentales (fp(q)=fm(Q)) et du nombre de descendants par père (ndp).*

fp(q)=fm(q)	0,5			0,625			0,75			0,875			1		
ndp	50	100	250	50	100	250	50	100	250	50	100	250	50	100	250
ST	8,7	11,9	11,6	8,1	13,5	13,0	11,4	15,7	15,6	13,9	20,2	21,5	20,3	26,1	26,6
DA	31,0	40,4	46,1	37,4	45,0	52,5	51,0	59,2	61,6	68,7	76,3	74,8	89,1	92,4	95,4
FDA	29,9	41,5	48,4	34,3	48,5	51,6	44,5	55,0	59,1	54,7	69,0	73,9	74,5	83,5	94,3
PCA	36,2	47,9	47,9	42,8	51,2	54,5	50,8	60,4	62,4	67,5	76,5	74,6	86,4	91,4	94,8
sPCA	32,1	40,7	43,2	35,6	43,2	48,2	44,9	49,3	54,1	55,8	66,6	68,4	75,1	80,6	87,1

Cas de 1 marqueur génétique tous les 12,5 cM, 25 descendants par mère, $\alpha_1 = \alpha_2 = 0,5$ pour le QTL pléiotrope sur deux caractères, avec $\rho_{12} = -0,4$. ST : méthode unicaractère; DA : analyse discriminante; FDA : analyse discriminante, transformation par famille de demi-frères; PCA : analyse en composantes principales (résultats obtenus avec l'analyse de la variable PCA2); sPCA : somme à chaque position des statistiques de test des variables de la PCA.

TAB. 2.13 - *SCE des estimations des positions des méthodes multicaractères en fonction de la fréquence des allèles au QTL dans les populations grand-parentales (fp(q)=fm(Q)) et du nombre de descendants par père (ndp).*

fp(q)	0,5			0,75			1		
ndp	50	100	250	50	100	250	50	100	250
DA	0,08	0,08	0,07	0,06	0,06	0,05	0,01	0,02	0,01
FDA	0,08	0,07	0,06	0,06	0,06	0,05	0,03	0,02	0,02
PCA	0,06	0,05	0,05	0,05	0,04	0,03	0,02	0,02	0,01
sPCA	0,07	0,07	0,07	0,06	0,06	0,06	0,03	0,02	0,02

Cas de 1 marqueur génétique tous les 12,5 cM, 25 descendants par mère, $\alpha_1 = \alpha_2 = 0,5$ pour le QTL pléiotrope sur deux caractères, avec $\rho_{12} = -0,4$. ST : méthode unicaractère; DA : analyse discriminante; FDA : analyse discriminante, transformation par famille de demi-frères; PCA : analyse en composantes principales (résultats obtenus avec l'analyse de la variable PCA2); sPCA : somme à chaque position des statistiques de test des variables de la PCA.

méthodes. Quand les fréquences des allèles sont égales en F0, elles sont supérieures ou égales au niveau de détection de DA quel que soit le nombre de descendants par père.

Par rapport à l'analyse discriminante réalisée sur l'ensemble de la population, FDA devient plus puissante uniquement lorsque la fréquence des homozygotes en F1 se rapproche de la moitié de la population et que le nombre de descendants par père est supérieur à 100. Cette dernière restriction semble être une condition à la bonne détection des QTL quand la transformation est réalisée intra-famille de père. Cela revient à estimer les matrices de (co)variances intra-classes sur 50 individus en moyenne.

Il semble donc que les méthodes basées sur la transformation des caractères à partir des caractéristiques de l'ensemble de la population soient relativement robustes à la non fixation des allèles dans les deux populations grand-parentales. Dans le cas particulier que nous venons de décrire, le bon maintien de la puissance de DA avec l'augmentation de la fréquence des homozygotes s'explique sans doute par le caractère fini de la population. En effet, les pedigree simulés contiennent de 2 à 10 pères. La variance d'échantillonnage entre simulations des génotypes des mâles F1 doit être importante, et les variabilités intra et inter-classes génotypiques au QTL s'éloignent donc de leur espérance, conservant des différences entre classes génotypiques exploitables.

b. SCE des estimations des positions Nous ne présentons les résultats obtenus que pour trois fréquences fp(Q)=fm(q)=1; 0,75; 0,5 (tableau 2.13). Les tendances générales d'évolution observées pour les puissances sont directement extrapolables aux SCE des estimations des positions. Une différence mérite cependant d'être soulignée. Quand la fréquence des allèles au QTL minoritaires en F0 augmente, la méthode qui permet d'obtenir les estimations de positions les plus précises est PCA, avec une réduction de plus de 20% des SCE par rapport aux autres méthodes.

4.3.5.2. Allèles non fixés dans une population grand-parentale Dans cette partie, les allèles au QTL de la population grand-paternelle ne sont pas fixés, avec fp(Q)$\in$ $[0,25;1]$. Dans la population des femelles F0, l'allèle q est fixé. Cette situation implique que deux génotypes au QTL existent à la génération F1, des hétérozygotes et des homozygotes qq. Nous avons représenté à la figure 12 les fréquences de génotypes au QTL engendrées à la génération F2 pour la gamme de fréquences alléliques simulées. Par rapport à la situation d'allèles fixés en F0, il existe très rapidement un excès d'individus porteurs du génotype qq en F2. Cet excès doit théoriquement détériorer les performances de détection de PCA, puisque cela implique une déformation du nuage des performances vers la moyenne des homozygotes qq.

Nous commenterons surtout ici les points remarquables par rapport à la situation

décrite au paragraphe précédent. L'ensemble des résultats est donné dans les tableaux 2.14 et 2.15.

a. Puissance Les puissances des méthodes étudiées sont en moyenne divisées par un facteur 4,8 entre des fréquences d'allèles extrêmes, de façon relativement homogène entre les méthodes. Cette constatation implique en particulier que la puissance de PCA ne semble pas plus affectée que celle des autres méthodes par la déformation du nuage de performances à la génération F2, malgré un léger décalage.

La hiérarchie des méthodes multicaractères n'est pas notablement modifiée par la diminution de fp(Q). L'utilisation de l'analyse discriminante réalisée intra-famille de demi-frères (FDA) ne permet en particulier pas d'améliorer la puissance de détection de l'analyse discriminante réalisée sur l'ensemble de la population (DA). Les niveaux de détection qu'elle permet d'atteindre ne sont encore une fois équivalents à ceux de DA que quand le nombre de descendants par père est grand (ndp=250).

De la même façon que précédemment, on constate un resserrement des puissances des méthodes multicaractères avec la diminution de fp(Q). Les différences les plus importantes, de l'ordre de 20% de la puissance de DA quand les allèles sont fixés sont réduites à 1 à 2% quand fp(Q)=0,25.

Quelles que soient les méthodes considérées et la fréquence de Q dans la population grand-paternelle, les méthodes multicaractères permettent toujours d'améliorer nettement la puissance de détection par rapport à ST, d'un facteur supérieur à 3,1 pour DA qui reste la méthode la plus puissante quelle que soit la situation.

b. SCE des estimations des positions Les évolutions des SCE observées quand fp(Q) diminue sont très semblables à celles des puissances, à ceci près que PCA devient plus précise que les autres méthodes quand les allèles ne sont pas fixés, comme nous avons pu l'observer dans la situation décrite au paragraphe 4.3..5.2.b.

4.4. Discussion

Nous avons comparé dans cette partie trois méthodes multicaractères pour un panel de situations relativement large. Nous avons ainsi pu confirmer, pour l'ensemble de ces méthodes, la caractéristique principale des méthodes multicaractères, à savoir leur capacité à d'autant mieux détecter des QTL quand les caractères ajoutés à l'analyse permettent d'augmenter la distance entre les nuages de performances associés aux génotypes au QTL (Korol *et al.*, 1995). Ceci se traduit dans la pratique par des situations où le produit des effets du QTL sur deux caractères avec leur corrélation résiduelle est négatif (Jiang et

TAB. 2.14 - *Puissances des méthodes multicaractères en fonction de la fréquence de l'allèle Q au QTL dans la population grand-paternelle (fp(Q), avec fm(q)=1) et du nombre de descendants par père (ndp).*

fp(Q)	0,25			0,50			0,75			1		
ndp	50	100	250	50	100	250	50	100	250	50	100	250
ST	4,2	6,6	6,4	10,0	12,1	14,6	14,0	16,9	20,8	20,3	26,1	26,6
DA	15,1	20,5	25,7	40,1	47,4	51,2	69,4	72,6	73,8	89,1	92,4	95,4
FDA	13,0	18,8	23,7	31,4	40,3	49,0	54,9	64,1	71,6	74,5	83,5	94,3
PCA	14,2	20,2	22,9	38,1	46,7	50,5	68,6	71,8	72,7	86,4	91,4	94,8
sPCA	15,3	18,5	21,8	35,3	38,1	46,0	56,5	59,9	64,3	75,1	80,6	87,1

Cas de 1 marqueur génétique tous les 12,5 cM, 25 descendants par mère, $\alpha_1 = \alpha_2 = 0,5$ pour le QTL pléiotrope sur deux caractères, avec $\rho_{12} = -0,4$. ST : méthode unicaractère; DA : analyse discriminante; FDA : analyse discriminante, transformation par famille de demi-frères; PCA : analyse en composantes principales (résultats obtenus avec l'analyse de la variable PCA2); sPCA : somme à chaque position des statistiques de test des variables de la PCA.

TAB. 2.15 - *SCE des estimations des positions des méthodes multicaractères en fonction de la fréquence de l'allèle Q au QTL dans la population grand-paternelle (fp(Q), avec fm(q)=1) et du nombre de descendants par père (ndp).*

fp(q)=fm(q)	0,25			0,50			0,75			1		
ndp	50	100	250	50	100	250	50	100	250	50	100	250
ST	0,1275	0,1266	0,1205	0,1001	0,0994	0,0970	0,0835	0,0730	0,0797	0,0725	0,0692	0,0525
DA	0,1114	0,1015	0,1055	0,0648	0,0616	0,0655	0,0365	0,0342	0,0299	0,0187	0,0132	0,0113
FDA	0,1170	0,1028	0,1061	0,0680	0,0681	0,0691	0,0446	0,0361	0,0361	0,0293	0,0202	0,0160
PCA	0,0949	0,0887	0,0912	0,0535	0,0565	0,0638	0,0315	0,0233	0,0257	0,0202	0,0169	0,0110
sPCA	0,1112	0,1042	0,1075	0,0705	0,0646	0,0680	0,0414	0,0386	0,0354	0,0283	0,0235	0,0193

Cas de 1 marqueur génétique tous les 12,5 cM, 25 descendants par mère, $\alpha_1 = \alpha_2 = 0,5$ pour le QTL pléiotrope sur deux caractères, avec $\rho_{12} = -0,4$. ST : méthode unicaractère; DA : analyse discriminante; FDA : analyse discriminante, transformation par famille de demi-frères; PCA : analyse en composantes principales (résultats obtenus avec l'analyse de la variable PCA2); sPCA : somme à chaque position des statistiques de test des variables de la PCA.

Zeng, 1955; Mangin *et al.*, 1998). Toutefois, l'application directe de cette règle est difficilement envisageable, puisqu'elle nécessite la connaissance des valeurs des effets du QTL et de la corrélation résiduelle. Cependant, il s'agit de situations intéressantes dans la pratique. En effet, si l'on considère par exemple deux caractères d'intérêt que l'on souhaite sélectionner dans des directions différentes, pour lesquels l'effet du QTL pléiotrope est positif et qui sont négativement corrélés résiduellement, la capacité à dissocier des QTL liés de QTL pléiotropes devient cruciale, et ne peut être atteinte que si l'on peut détecter le QTL pléiotrope.

Les méthodes multicaractères permettent en général d'améliorer la puissance de détection par rapport aux méthodes unicaractères, ainsi que la précision d'estimation des paramètres. En particulier, l'estimation de la position est en général peu précise par les méthodes unicaractères quand les effets du QTL sont faibles (voir paragraphe 3.5., chapitre 1). L'utilisation des méthodes multicaractères permet de réduire nettement la variation d'estimation de ce paramètre. Dans des objectifs de sélection assistée par marqueurs ou de cartographie fine de QTL, ce dernier élément est crucial.

Nous avons par ailleurs pu confirmer que la seule corrélation résiduelle entre les caractères est suffisante pour que le gain d'information soit exploitable et traduit en terme de gain de puissance par les méthodes multicaractères. Ce type de détection semble néanmoins devoir être limité à la détection de QTL présentant des effets élevés sur le caractère d'intérêt, QTL donc potentiellement bien détectables par les méthodes unicaractères. L'utilisation des méthodes multicaractères permet néanmoins d'affiner l'estimation des paramètres du QTL, tels que sa position et ses effets.

Enfin, nous avons testé pour les méthodes proposées la robustesse à la non fixation des allèles au QTL dans la génération F0. L'ensemble des méthodes, dans le cadre simple de l'analyse de deux caractères que nous avons exploré, est particulièrement robuste à cette hypothèse. Il s'agit là d'un aspect important puisque la majorité des populations expérimentales animales d'intérêt agronomique, même issues de croisements entre lignées ou races divergentes, ne garantit pas la fixation des allèles au QTL dans le dispositif, et de nombreux exemples issus des résultats de détection ont confirmé la légèreté d'une telle hypothèse (par exemple Lipkin *et al.*, 1998 [70]; Bidanel *et al.*, 2000 [5]).

Cette étude nous a par ailleurs permis de mettre en évidence que l'analyse discriminante peut être réalisée avec la même puissance sur l'ensemble de la population ou intra-familles de père si le nombre de descendants par père est supérieur à 100. Cette constatation permet d'envisager des applications à des schémas de détection de QTL de type filles ou petites-filles (Weller *et al.*, 1990 [106]). Ce type de schéma est surtout utilisé dans les populations animales où la création de populations expérimentales est difficile, souvent en raison d'intervalles de génération longs, mais où de grandes cohortes de don-

nées sont disponibles au niveau de la population de production. De tels protocoles sont en particulier souvent utilisés pour la détection de QTL chez les bovins laitiers.

Cependant, les caractéristiques dégagées ci-dessus sont d'ampleur variable en fonction de la méthode multicaractère considérée. Nous avons pu mettre en évidence que l'application de la méthode multivariée à la détection de QTL multicaractère dans des populations animales est inenvisageable en routine du fait des temps de calcul qu'elle requiert. En effet, les distributions des statistiques de test correspondantes étant difficilement appréhendables analytiquement, l'estimation de seuils de signification sera dans la pratique empirique, soit à l'aide de permutations des données, soit à l'aide de simulations. Ce type de stratégie demande l'analyse répétée de nombreux jeux de données, ce qui est incompatible avec la mise en oeuvre de méthodes très exigeantes en terme de temps de calcul. Néanmoins, nous avons pu montrer que, dans le cadre des simulations que nous avons réalisées, les seuils de la méthode multivariée sont très bien estimés par la maximisation de la somme des statistiques de test des variables en composantes principales à chaque position. En effet, sous réserve que le modèle de détermination des caractères soit simple et sous l'hypothèse nulle d'absence de QTL sur le groupe de liaison étudié, la matrice de (co)variance résiduelle est identifiable à la matrice de (co)variance phénotypique. L'utilisation de la méthode multivariée pour détecter des QTL peut donc être combinée avec des calculs de seuils par sPCA. Cependant, la méthode multivariée est aussi la méthode qui permet d'obtenir les puissances de détection les moins élevées dans la majorité des situations. Ce résultat est une conséquence directe d'un nombre de paramètres élevé à estimer pour des tailles de dispositifs relativement faibles. Cette méthode est néanmoins la seule parmi celle que nous avons explorées qui permet d'obtenir les estimations directes des effets de substitution des QTL et des corrélations résiduelles entre les caractères.

En ce qui concerne les deux méthodes multicaractères basées sur l'analyse de combinaisons linéaires des caractères, nous avons pu montrer qu'elles ont des comportements différents en fonction des situations analysées. L'analyse en composantes principales réalisée à partir des données phénotypiques a été décrite analytiquement par Mangin *et al.* (1998 [76]) dans le cas de croisements enrte lignées *inbred*. Nous avons pu confirmer ici que cette méthode permet de détecter de façon puissante et précise des QTL pléiotropes quand le déterminisme et les relations entre les caractères sont simples. Des simulations réalisées pour p caractères, avec des combinaisons d'effets et de corrélations résiduelles plus compliquées que celles présentées dans ce document, ont permis de montrer que la composante principale expliquant la variabilité due au QTL n'est pas prédictible si on ne connait pas la structure génétique des caractères sous-jacente. En terme d'estimation de la position, PCA permet souvent d'obtenir les estimations les plus précises pour l'étude simultanée de deux caractères.

L'analyse discriminante permet quant à elle de détecter des QTL de façon puissante

quel que soit le déterminisme des caractères. De plus, l'estimation de la position est précise par rapport aux autres méthodes multicaractères. Cette méthode présente par ailleurs l'avantage de ne pas être plus exigeante en terme de temps de calcul avec l'augmentation du nombre de caractères analysés. En revanche, et de la même façon que la PCA, elle ne permet pas d'estimer les effets du QTL sur chaque caractère. Nous avons donc dû corriger le paramètre de calcul de la précision d'estimation des effets de la combinaison linéaire pour la rendre comparable à la précision de l'estimation des effets des autres méthodes. Cette correction est très dépendante de la capacité de la méthode à proposer une discrimination des classes génotypiques au QTL pertinente, ce qui limite la portée de nos conclusions quant à ce paramètre. De plus, d'un point de vue pratique, l'accès aux effets du QTL sur chaque caractère analysé est déterminant.

Dans la pratique, étant donné qu'un des avantages principaux des méthodes multicaractères est de détecter des QTL qui ne sont pas détectables par les méthodes unicaractères, on doit envisager de réaliser, pour des groupes de caractères d'intérêt, des balayages complets du génome avec les méthodes multicaractères. Une stratégie pourrait consister en la réalisation du balayage avec l'analyse discriminante, qui permettra dans un premier temps d'identifier de façon puissante et précise les régions chromosomiques d'intérêt. Une analyse plus fine de ces régions pourra alors être entreprise à l'aide de méthodes multivariées pour estimer les effets de QTL correspondants. Cette stratégie peut permettre de tester les significations des contributions de chaque caractère à la vraisemblance, et les significations des effets du QTL sur chaque caractère selon la procédure proposée par Korol *et al.* (2001 [59]), en restreignant les tests aux seules positions détectées lors de l'analyse préliminaire.

5. Application

Les données utilisées ont été décrites au paragraphe 4.3. de l'introduction. Nous avons alors pu voir que pour les 5 caractères de composition corporelle analysés, les statistiques de test présentent des profils très semblables sur le chromosome. La question ici est donc de savoir si l'on peut mettre en évidence un déterminisme pléiotrope pour certains de ces caractères. Cette analyse est une première étape pour la mise en évidence d'une architecture de QTL déterminant les caractères de composition corporelle dans cette région chromosomique. Les conclusions ici ne seront donc que partielles, et seront complétées dans la partie suivante par l'utilisation de méthodes permettant de mettre en évidence des QTL génétiquement liés.

Sur cet exemple, nous décrirons la méthodologie utilisée pour sélectionner les groupes de caractères à retenir dans l'analyse. Nous présenterons ensuite les résultats obtenus, que nous discuterons. Enfin, le traitement de ce cas particulier sera l'occasion de vérifier la réalisation de certaines conclusions de l'étude précédente.

5.1. Méthodologie utilisée

5.1.1. Analyse conjointe et sélection des caractères

L'analyse a été réalisée séquentiellement. L'analyse discriminante, d'après les résultats présentés précédemment, est la méthode qui permet d'obtenir la plus grande puissance de détection quelles que soient les relations entre les caractères et leur déterminisme. Nous avons donc choisi de baser la sélection des caractères sur leurs pondérations dans la combinaison linéaire obtenue au maximum de la statistique de test pour DA.

Comme nous l'avons vu, l'analyse discriminante peut être réalisée de différentes façons. Les choix résident d'une part dans la distinction ou non des familles de pères pour l'estimation des combinaisons linéaires, et d'autre part dans la détermination du nombre de groupes génotypiques à considérer pour le calcul de la combinaison linéaire. Dans une première étape, nous avons donc comparé les statistiques de test obtenues pour des combinaisons linéaires estimées intra-famile de père ou sur l'ensemble de la population. La deuxième étape s'intéresse quant à elle au nombre de groupes génotypiques à prendre en compte dans l'analyse discriminante. Nous avons en effet montré que la prise en compte des méioses femelles est superflue quand le déterminisme des caractères est additif et que les allèles au QTL sont fixés. Ces hypothèses ne peuvent ici pas être validées *a priori*.

L'étape suivante réside dans la sélection du groupe de caractères permettant d'obtenir la statistique de test la plus significative. Cette sélection a été réalisée sur la base de la contribution de chaque caractère à la combinaison linéaire au maximum de la statistique

de test pour l'analyse discriminante :
a) Le caractère contribuant le moins à la statistique de test pour l'analyse de p caractères est enlevé de l'analyse,
b) une nouvelle statistique de test est calculée pour $p-1$ caractères,
c) si le niveau de signification de la statistique de test pour $p-1$ caractères n'est pas diminué par rapport à p caractères, un nouveau caractère est retiré de l'analyse sur la base de sa contribution à la combinaison linéaire,
d) la séquence recommence à l'étape a), jusqu'à sélection d'un groupe de p_v caractères permettant d'obtenir une statistique de test plus significative que $p_v - 1$ caractères, ou l'épuisement du stock de caractères.

Pour chaque étape, le niveau de signification empirique de la statistique de test est calculé en utilisant 2000 simulations de la même façon que décrit au paragraphe 3.3.. Une fois un groupe de caractères sélectionné, une analyse multivariée a été réalisée dans la région mise en évidence par l'analyse discriminante afin d'estimer les effets du QTL sur chacun des caractères et les corrélations résiduelles entre les caractères.

5.1.2. Analyses complémentaires

Afin d'illustrer les différences mises en évidence précédemment entre les méthodes multicaractères, chaque méthode a été utilisée pour l'analyse des groupes de caractères sélectionnés. Ces résultats seront présentés dans un deuxième temps. Pour chaque méthode, des seuils de signification ont été estimés par simulation comme décrit précédemment, en fonction du nombre de caractères analysés. D'une façon générale, seuls les seuils de signification à 5% seront indiqués. Les statistiques de test obtenues ont souvent des niveaux de signification beaucoup plus faibles, difficilement estimables avec précision sur la base de 2000 simulations.

5.2. Résultats

5.2.1. Analyses préliminaires des 5 caractères

5.2.1.1. Analyse intra-famille ou sur la population Les résultats des analyses discriminantes sur la population en ne considérant que les groupes d'haplotypes paternels ou par famille de père sont donnés dans les deux premières colonnes du tableau 2.16. Les seuils de signification à 5% correspondants sont indiqués. Les maxima des deux statistiques de test présentent des valeurs très proches et largement significatives au seuil de 5%. Cependant, le maximum de l'analyse par population, bien que légèrement inférieur à celui de l'analyse par père, est nettement plus significatif, du fait d'un nombre de degrés de

liberté plus faible. Nous avons donc choisi dans la suite de cette étude de travailler avec le calcul des combinaisons linéaires sur la population. Etant donnée la structure de la population (41 à 75 descendants par père), ce type de résultat était attendu.

5.2.1.2. Détermination du nombre de groupes génotypiques à considérer Les maxima des statistiques de test obtenues pour l'analyse de 2 à 4 groupes génotypiques déterminés à l'échelle de la population et les positions correspondantes sont donnés dans le tableau 2.16. Comme nous l'avons vu précédemment, les seuils de signification ont tendance à augmenter avec le nombre de groupes considérés, mais sont sensiblement équivalents. La statistique de test la plus significative est donc ici obtenue par l'analyse de 2 groupes génotypiques.

TAB. 2.16 - *Analyse discriminante des 5 caractères.*

	Nombre de groupes d'haplotypes / Type d'analyse			
	2/père	2/pop	3/pop	4/pop
Seuils (5%)	69,29	51,63	55,74	57,42
Max LRT	158,58	149,77	145,23	145,91
Position	65	65	66	66

Seuils, maxima des statistiques de test et positions correspondantes en fonction du nombre de groupes génotypiques considérés. /père= transformation intra-famille de père. /pop= transformation sur la population.

L'analyse discriminante retenue pour le reste de l'étude correspond donc au calcul de combinaisons linéaires moyennes sur la population globale, avec 2 groupes génotypiques dépendant de l'origine grand-parentale des haplotypes hérités des pères.

5.2.2. Sélection des caractères

Les résultats de la sélection d'un groupe de caractères significatif sur la base de la contribution de chaque caractère à la combinaison linéaire sont donnés dans le tableau 2.17. Les seuils à 5% correspondants sont indiqués.

Le premier caractère enlevé est le poids de bardière, qui ne participe qu'à 0,36% de la combinaison linéaire. Le second caractère est l'épaisseur de lard x4, qui y participe pour 5,40%. L'épaisseur de lard x2 est ensuite soustraite à l'analyse. Elle représente alors 9,10% de la combinaison linéaire. La teneur en gras intramusculaire et le poids de panne représentent alors respectivement 48,9 et 51,1% de la combinaison linéaire finale sur deux caractères. Le niveau de signification de la statistique de test ne change pas que l'on analyse les 5 caractères ou seulement le gras intramusculaire et le poids de panne. On peut donc conclure que les trois caractères enlevés à l'analyse n'apportent pas d'information sur

TAB. 2.17 - *Sélection des caractères par l'analyse discriminante.*

Nombre de caractères		5	4	3	2
Seuils (5%)		51,63	49,63	48,74	46,82
Max	LRT	149,77	149,77	149,87	146,90
	Position	65	65	65	65
Pondération des caractères					
	imf	-0,716	-0,716	-0,708	-0,717
	x2	0,204	0,203	0,137	
	x4	-0,088	-0,091		
	bardière	-0,006			
	panne	0,678	0,677	0,661	0,750

le QTL. L'évolution des profils de statistiques de test le long du chromosome correspondant à cette sélection pour l'analyse discriminante est représentée à la figure 13. Le profil global est très peu modifié quelles que soient les positions analysées.

TAB. 2.18 - *Effets estimés par père et effets moyens pour l'analyse conjointe de imf et panne.*

Numéro du père	910001	910045	910081	910088	moyen
imf	1,230	0,916	0,776	0,404	0,832
panne	-0,632	-0,406	-0,714	-1,200	-0,738

Les effets moyens par famille de père estimés pour ces deux caractères par la méthode multivariée sont respectivement de 0,83 et -0,74 pour imf et panne. Les effets estimés intra-père pour chaque famille de père sont donnés dans le tableau 2.18. Les effets intra-famille de père sont par construction estimés beaucoup plus précisément que les effets intra-famille de mère, étant données les tailles respectives des familles. Ces derniers ne sont donc pas présentés ici. Les valeurs estimées intra-père sont relativement homogènes d'un père à l'autre. On peut donc supposer que les allèles issus de chaque race, s'ils ne sont pas nécessairement fixés, déterminent les deux caractères dans le même sens. Un signe négatif correspond à une augmentation de la valeur du caractère due à un allèle d'origine Large White. Les effets ici sont donc opposés, sur deux caractères de composition de carcasse pour lesquels des estimations dans le même sens sont attendues (chapitre 1, paragraphe 4.3.).

La méthode multivariée permet par ailleurs d'estimer la corrélation résiduelle entre les deux caractères après prise en compte du QTL. Elle est ici de 0,054, alors que la corrélation phénotypique estimée était de -0,23 (4.3..2.). Ce QTL semble donc avoir un rôle important dans la structuration génétique des relations entre les deux caractères.

Nous avons par ailleurs analysé séparément le groupe de trois caractères supprimés lors de la sélection. Ces trois caractères permettent d'obtenir une statistique de test largement

TAB. 2.19 - *Sélection des caractères complémentaires par l'analyse discriminante.*

Nombre de caractères		3	2
Seuils (5%)		48,74	46,82
Max	LRT	82,11	80,10
	Position	68	68
Pondération des caractères			
	x2	0,311	0,400
	x4	0,507	0,680
	bardière	0,274	

significative à 1% avec l'analyse discriminante, avec un maximum de 82,11 de la statistique de test (tableau 2.19). De même que précédemment, nous avons soustrait à l'analyse le caractère participant le moins à la combinaison linéaire, soit bardière. Le maximum de la statistique de test est alors de 80,10, à la position 68 cM. Ce maximum représente, compte tenu du caractère enlevé, un niveau de signification équivalent à celui de l'analyse des trois caractères. Nous pouvons donc envisager de traiter de façon séparée les épaisseurs de lard et le poids de bardière. Le profil de vraisemblance obtenu sur le chromosome diffère fortement de celui obtenu pour le groupe de caractères précédent. La position du maximum est légèrement décalée par rapport à celle du groupe de caractères précédent. Ceci suggère la présence de deux QTL différents sur le chromosome, plus un pour le poids de bardière. Cependant, cette hypothèse, à ce stade de l'étude, ne peut pas être testée.

Les effets moyens estimés intra-famille de père sont respectivement de 0,57 et 0,59 sur les épaisseurs de lard x2 et x4. La corrélation résiduelle estimée est de 0,81, ce qui est relativement proche des corrélations phénotypiques connues entre ces deux caractères. Les effets des allèles Large White et Meishan sont donc ici dans le même sens, et de taille équivalente, sur tous les caractères.

5.2.3. Analyses complémentaires

Le détail des maxima et des positions des statistiques de test obtenu pour chaque méthode multicaractère au cours des deux phases de sélection de caractères est donné dans le tableau 2.20 pour l'analyse multivariée (MV) et le tableau 2.21 pour l'analyse en composantes principales (PCA). Ces caractéristiques soulignent certaines des observations réalisées à partir des simulations quant au comportement relatif de ces méthodes multicaractères. L'évolution des profils de statistiques de test correspondant à la première phase de sélection est représentée dans la figure 13. Seul le profil de la variable en composante principale permettant d'obtenir la statistique de test la plus élevée est représenté.

Les statistiques de test de MV diminuent avec la diminution du nombre de caractères

TAB. 2.20 - *Analyses réalisées avec la méthode multivariée.*

			1ère sélection des caractères			2ème sélection des caractères	
Nombre de caractères		5	4	3	2	3	2
Seuil (5%)		161,04	131,68	104,63	75,92	104,63	75,92
Max	LRT	248.98	221.13	201.68	168.82	114.37	91,34
	Position	64	65	65	65	64	67
Effet moyen estimé	imf	0.836	0.836	0.836	0.836		
	x2	-0.600	-0.600	-0.600		-0.550	-0,693
	x4	-0.606	-0.606			-0.570	-0,586
	bardière	-0.590				-0.540	
	panne	-0.740	-0.740	-0.740	-0.740		

analysés, ce qui correspond à la diminution du nombre de paramètres estimés. Cependant, celles-ci sont toujours très significatives au seuil de 5%.

Au contraire, le maximum de la statistique de test la plus significative par l'analyse en composantes principales augmente avec la diminution du nombre de caractères analysés. Pour 5 caractères, deux variables en composantes principales permettent d'obtenir des statistiques de test significatives pour la région de 62 à 67 cM. On peut donc supposer que la variabilité due au QTL mis en évidence par l'analyse discriminante est répartie sur plusieurs axes en composantes principales phénotypiques, ce qui limite la capacité de détection de la méthode. En contrepartie, on ne peut pas conclure à la présence de deux QTL distincts dans cette région. Pour l'analyse finale des deux caractères imf et panne, la statistique de test de PCA permet d'obtenir un maximum très proche en terme de valeur, et donc de niveau de signification, de celle de l'analyse discriminante. Cet aspect est illustré par la superposition des profils de statistique de test pour ces méthodes lors de l'analyse de deux caractères. Cependant, les pondérations des caractères dans les variables en composantes principales sont toujours très différentes des pondérations obtenues avec l'analyse discriminante.

5.2.4. Remarque sur l'estimation des effets

Sur cet exemple, nous avons effectué la transformation des effets moyens estimés intra-famille de père par la méthode multivariée à l'aide de la pondération de la combinaison linéaire de l'analyse discriminante d'une part, et de la combinaison des pondérations de la transformation en composantes principales d'autre part. Dans les deux cas, et quel que soit le nombre de caractères analysés, les estimations d'effets pour DA et PCA ainsi obtenues sont très proches de celles estimées directement par la maximisation de la vraisemblance à la position correspondante. Cette remarque suggère que les effets sont bien estimés

par les méthodes basées sur des combinaisons linéaires, du moins quand un QTL est en ségrégation dans la population.

5.3. Conclusion

Les trois groupes de caractères discriminés correspondent à des catégories de gras différentes : panne et imf caractérisent plutôt le gras "interne", alors que x2, x4 et bardière caractérisent plutôt le gras "de couverture", les épaisseurs de lard étant des mesures particulières, différentes du poids de bardière qui dépend plus de la longueur de l'animal. Les groupes de caractères obtenus sont donc phénotypiquement intéressants dans une optique de sélection : le gras de couverture, génétiquement très lié au taux de muscle de la carcasse, a été efficacement sélectionné dans les années passées, et l'on considère généralement qu'une limite est atteinte sur ce type de caractères pour la sélection. En revanche, le gras intra-musculaire présente actuellement un intérêt majeur. En effet, il est associé aux qualités sensorielles de la viande, mais il est difficile à sélectionner en raison de la difficulté à phénotyper les animaux en routine. La mise en évidence de marqueurs moléculaires spécifiques de ce type de caractère est déterminante pour la mise en place d'une sélection rapide et efficace. Or l'objectif d'augmenter imf tout en diminuant ou stabilisant le gras de couverture semble aller dans le sens des effets QTL estimés ici. De plus, les effets mis en évidence indiquent une action favorable sur un caractère des allèles d'origine Large White, pour lesquelles la sélection est techniquement plus simple à réaliser que l'introgression d'allèles Meishan.

TAB. 2.21 - *Analyses en composantes principales.*

				1ère sélection des caractères			2ème sélection des caractères	
Nombre de caractères			5	4	3	2	3	2
Seuil (5%)			53,16	50,65	50,33	47,68	50,33	47,68
Max	LRT	PCA1	100.10	108.00	122.40	147.00	82.30	79,88
	Position		67	67	66	65	67	68
	LRT	PCA2	65.15	59.45	52.15	43.05	24.97	30,35
	Position		62	61	61	37	13	14
	LRT	PCA3	52.47	48.32	46.66		36.47	
	Position		34	34	31		04	
	LRT	PCA4	26.20	31.16				
	Position		11	147				
	LRT	PCA5	36.23					
	Position		139					
Pondération	PCA1	imf	-0.147	-0.206	-0.302	-0.707		
		x2	0.502	0.578	0.672		0.572	0,707
		x4	0.515	0.588			0.582	-0,707
		bard	0.512				0.577	
		panne	0.446	0.527	0.676	0.707		
	PCA2	imf	0.986	0.977	0.953	0.707		
		x2	0.108	0.168	0.229		0.787	0,707
		x4	0.023	0.077			-0.193	0,707
		bard	0.116				-0.586	
		panne	0.043	0.111	0.198	0.707		
	PCA3	imf	0.022	0.002	0.022			
		x2	-0.318	-0.409	-0.704		-0.229	
		x4	-0.293	-0.352			0.789	
		bard	-0.161				-0.569	
		panne	0.887	0.842	0.710			
	PCA4	imf	-0.121	-0.065				
		x2	0.764	-0.666				
		x4	-0.235	0.723				
		bard	-0.594					
		panne	0.890	-0.031				
	PCA5	imf	0.074					
		x2	-0.226					
		x4	0.770					
		bard	-0.588					
		panne	0.065					

PCAl=analyse de la variable en composantes principales associée à la $l^{ième}$ valeur propre

Chapitre 3

Méthodes multiQTL

1. Introduction

1.1. Questions posées par les méthodologies uniQTL

1.1.1. Cas de base : unicaractère/uniQTL

Les premières stratégies de détection de QTL étaient basées sur des modèles simples, avec un QTL détecté par marqueur génétique analysé. La faiblesse de ce type de modèle a rapidement été mise en évidence quant à la pertinence des tests réalisés (MacMillan et Robertson, 1974), puis lors de la parution des premiers résultats de recherche systématiques de QTL (Andersson *et al.* 1994 [3]; Georges *et al.*, 1995 [23] pour se limiter aux animaux de ferme). Les analyses uniQTL peuvent ainsi être fortement influencées par la présence de plusieurs QTL agissant sur le même caractère. La robustesse et la puissance des méthodes sont alors directement atteintes, ainsi que la qualité d'estimation de la position et des effets du QTL. En effet, le principe même de la détection de QTL repose sur la minimisation de la variance résiduelle du modèle $y = q_1 + r_1$, où y est le phénotype, q_1 l'effet du QTL et r_1 la résiduelle. Si r_1 peut, du fait de la présence d'un deuxième QTL, se décomposer en $r_1 = q_2 + r_2$, sa minimisation est par essence limitée, et la qualité de la détection s'en ressent. Le modèle de base de la détection multiQTL sera donc une décomposition de y en $\sum_{t=1}^{Q} q_t + r$.

En fonction de la nature de la relation entre les QTL agissant sur le caractère considéré, différents problèmes peuvent apparaître. Dans le cas le plus simple où deux QTL sont indépendants, la prise en compte simultanée de ces QTL permet de réduire la proportion de variance résiduelle dans le modèle par rapport à la proportion de variance due au(x) QTL. La puissance du modèle et la qualité d'estimation des paramètres sont

par conséquent améliorées mécaniquement. Si les deux QTL interagissent, en raison de phénomènes d'épistasie et/ou parce qu'ils sont physiquement liés, la prise en compte de cette interaction permet d'améliorer considérablement la qualité de la détection en raison de l'augmentation de la pertinence du modèle sous-jacent (Lander et Botstein, 1989, [62]; Haley et Knott, 1992 [28]; Zeng, 1994 [115]; Martinez et Curnow, 1992 [80]; Rodolphe et Lefort, 1993 [92]; Dizier *et al.* 1993 [15]). Mather et Jinks, en 1982 [81] ont défini les différents types d'épistasie qui peuvent être rencontrés selon les relations entre les effets des QTL, récapitulés dans le tableau 3.1. En fonction du cas étudié, l'influence de la relation entre deux QTL sur la qualité de leur détection est très variable.

TAB. 3.1 - *Différents modèles d'épistasie.*

Nature de l'épistasie	Relation entre les paramètres
Complémentaire	$a_1 = a_2 = d_1 = d_2 = aa_{12} = ad_{12} = ad_{21} = dd_{12}$
Duplicate	$a_1 = a_2 = d_1 = d_2 = -aa_{12} = -ad_{12} = -ad_{21} = -dd_{12}$
Dominante	$a_1 = d_1 \neq a_2, a_2 = d_2 = -aa_{12} = -ad_{12} = -ad_{21} = -dd_{12}$
Récessive	$a_1 = d_1 \neq a_2, a_2 = d_2 = aa_{12} = ad_{12} = ad_{21} = dd_{12}$
Inhibitrice	$a_1 = a_2 = d_1 = -d_2 = -aa_{12} = -ad_{12} = -ad_{21} = -dd_{12}$

a_i est l'effet additif du QTL i; d_i est l'effet de dominance du QTL i, xy_{ij} est l'interaction des différents effets x_i et y_j. *D'après Mather et Jinks, 1982.*

Deux cas particuliers de détérioration de la qualité de détection des QTL lors d'analyses uniQTL ont été longuement commentés, en particulier par Haley et Knott (1992 [28]) et Luo et Kearsey (1992 [72]) sur des résultats de simulations, et par Martinez et Curnow (1992 [80]) sur le plan analytique. Ils concernent l'analyse d'un groupe de liaison portant deux QTL agissant sur le même caractère. Ces QTL sont physiquement liés et ont uniquement des effets additifs et du même ordre de grandeur sur le caractère. Les effets peuvent être en addition, les deux QTL sont alors en phase, ou en répulsion, les allèles ayant des effets dans le même sens étant sur des haplotypes différents (figure 14). Dans le premier cas, un QTL dit fantôme entre les deux QTL existants peut être détecté (figure 15). Ce type de résultat met immédiatement en défaut la proposition de Lander et Botstein (1989 [62]) de fixer la position d'un premier QTL détecté lors d'une analyse uniQTL pour réitérer une analyse avec covariables selon les effets et position issus de la première. Dans le deuxième cas, la détection uniQTL détecte mal les QTL, les deux s'absorbant l'un l'autre lors de la co-transmission d'allèles à effets contradictoires.

De plus, Visscher et Haley (1996 [104]) et Liu et Dekkers (1998 [71]), respectivement sur des croisements entre lignées consanguines et dans des populations outbred, ont voulu comparer l'influence du choix d'une hypothèse nulle classique en détection de QTL uniQTL : "il n'y a pas de QTL sur le groupe de liaison", avec le choix d'une hypothèse nulle correspondant au modèle infinitésimal habituellement utilisé en génétique quantitative :

"il y a de nombreux locus influençant chacun de façon infime la détermination du caractère sur le groupe de liaison". Ils ont alors montré que les distributions des statistiques de test sous l'hypothèse d'un locus à effet important ou celle de la présence de nombreux locus à très faibles effets peuvent être très proches. Cette confusion sur le nombre de locus impliqués dans la détermination d'un caractère peut entraîner une sous estimation importante du seuil de rejet de l'hypothèse nulle classique utilisée dans les analyses uniQTL, conduisant de façon erronée à la conclusion de la présence d'un QTL. La prise en compte de locus supplémentaires semble donc cruciale pour la caractérisation précise du déterminisme génétique des caractères.

1.1.2. Cas multicaractère : liaison de QTL agissant chacun sur un caractère

Un deuxième type de problématique est directement lié à la mise en place de méthodes multicaractères. En effet, dans la pratique, la distinction entre un locus agissant sur deux caractères et deux locus agissant chacun sur un caractère peut être stratégiquement cruciale si l'on espère casser, ou au contraire conserver, la liaison entre les QTL codant pour ces deux caractères. On doit donc distinguer un QTL pléiotrope de deux QTL liés. De plus, comme illustré par la figure 16, la prise en compte des effets de deux QTL sur les caractères peut permettre de diminuer fortement la variance résiduelle du modèle de détermination des caractères.

De plus, Ronin *et al.* (1999 [94]) montrent, sur des explicitations analytiques du $ELOD$ pour la comparaison des trois hypothèses "il n'y a pas de QTL", "il y a un QTL" et "il y a deux QTL", que la prise en compte d'un caractère résiduellement corrélé à un premier caractère déterminé par deux QTL sur le groupe de liaison permet d'améliorer les performances de détection par rapport à des analyses uniQTL ou unicaractère. Cependant, ces développements sont restreints à des cas simples de populations dihaploïdes, avec un ou deux marqueurs génétiques pris en compte dans l'analyse, et sur la base d'estimations asymptotiques du $ELOD$. Elles ont été confirmées par des simulations pour le même type de population.

Enfin, Korol *et al.* (2001 [59]) ont testé le comportement de leur méthode multicaractère uniQTL (voir chapitre 2) dans le cas où d'autres QTL sont en ségrégation sur le même groupe de liaison. En comparaison avec les cas où un seul QTL est en ségrégation, ils concluent logiquement à une diminution de la puissance de détection et de la précision d'estimation de la position. Ils conseillent alors de prendre en compte l'existence de ces QTL supplémentaires à travers la définition de cofacteurs (comme Zeng, 1994 [115] ou Jansen et Stam, 1994 [37]).

1.2. Panorama sur les méthodes multiQTL proposées

L'ensemble des grands types de méthodes de détection uniQTL (maximum de vraisemblance, régression multiple, méthodes non paramétriques ou bayésiennes) a donné lieu à des développements multiQTL. Nous commenterons essentiellement les méthodes basées sur les deux premières approches.

1.2.1. Premières méthodes proposées

Historiquement, les méthodes multiQTL ont été développées avant les méthodes multicaractères. Les premières propositions ont été faites par Lander et Botstein (1989 [62]), Haley et Knott (1992 [28]), Martinez et Curnow (1992 [80]) et Rodolphe et Lefort (1993 [92]) pour la détection de deux QTL liés, et par Zeng (1994 [115], 1993 [114]) et Jansen *et al.* (1993 [35], 1994a [36], 1994b [37], 1996 [38]) pour la détection de QTL répartis sur l'ensemble du génome. Ces méthodes ont été essentiellement proposées pour des applications à des croisements entre lignées *inbred*.

1.2.1.2. QTL sur des chromosomes différents Zeng (1994 [115]) et Jansen *et al.* (1993 [35], 1994a [36], 1994b [37], 1996 [38]) ont proposé indépendamment des méthodes similaires permettant de prendre en compte le déterminisme du caractère dû à d'autres régions chromosomiques que la position étudiée, afin de réduire la variance résiduelle et d'améliorer les tests. Ces méthodes sont nommées CIM, pour Combined Interval Mapping pour la méthode de Zeng et MQM pour Multiple QTL Mapping pour celle de Jansen *et al.*.

Le CIM est donc une méthode univariée. Elle consiste à intégrer l'effet d'autres régions chromosomiques dans l'analyse en utilisant les marqueurs flanquant ces régions comme cofacteurs dans une régression linéaire multiple. Les auteurs définissent le modèle des performances pour un croisement entre lignées inbred de la façon suivante :

$$y_i = b_0 + \sum_{h=1}^{t} b_h x_{ih} + \epsilon_i \tag{3.1}$$

pour chaque individu $i = 1, \ldots, n$, et t marqueurs génétiques,

avec x_{ih} la variable indicatrice du génotype au $h^{ième}$ marqueur pour l'individu i,

b_0 la moyenne du modèle,

b_h le coefficient de régression partielle du phénotype y sur le $h^{ième}$ marqueur conditionné aux autres marqueurs,

ϵ_i l'erreur aléatoire normalement distribuée avec une moyenne nulle.

Les auteurs ont procédé à une étude poussée des mécanismes déterminant leur méthode, et ont dégagé quatre grandes propriétés qui la conditionnent :

- si les QTL ont tous des effets additifs, l'espérance du coefficient de régression partielle du caractère sur un marqueur ne dépend que des QTL localisés dans l'intervalle formé des deux marqueurs adjacents et n'est pas affectée par les QTL situés ailleurs,

- si on conditionne pour les marqueurs non liés dans la régression multiple, la variance d'échantillonnage de la statistique de test est réduite par le contrôle d'une partie de la variance génétique résiduelle, ce qui augmente donc la puissance de détection des QTL,

- si on conditionne sur des marqueurs liés, les chances d'interférence de QTL liés sur les tests d'hypothèse et l'estimation des paramètres sont réduites, mais on peut augmenter la variance d'échantillonnage de la statistique de test,

- deux coefficients de régression multiple du caractère sont généralement non corrélés, sauf si les marqueurs sont adjacents.

Pour la prise en compte, grâce à des cofacteurs, de QTL présents sur d'autres chromosomes, la faiblesse de la précision d'estimation de la position du QTL devient moins importante que quand les deux QTL sont physiquement liés, mais l'effet des QTL en cofacteur doit être bien estimé.

Zeng (1994 [115]) ne propose par contre pas de méthode particulière pour sélectionner les marqueurs à intégrer en cofacteurs, bien qu'il évoque la possibilité d'une sélection itérative de type *stepwise*, mais qui peut être longue, ou l'intégration comme cofacteurs de marqueurs régulièrement espacés sur le génome, les autres servant à la détection de QTL (Zeng, 1995). Jansen (1994b [37]) propose quant à lui une sélection des marqueurs basée sur le Critère d'Information d'Akaike (AIC, Akaike, 1974 [2]) de type *backward*, stratégie appliquée par Jansen et Stam (1994a [36]). Dans des stratégies de même type, Knott *et al.* (1998 [52]) proposent de ne sélectionner que les cofacteurs les plus significatifs pour chaque groupe de liaison, avec une élimination *backward* des cofacteurs non significatifs ensuite. Ce type de stratégie revient, à l'étape de cartographie finale du QTL, à n'effectuer qu'un test univarié à une position, puisque tous les auteurs conseillent, pour comparer des hypothèses consistantes, de garder dans les modèles des hypothèses nulle et alternative le même réseau de cofacteurs. Les distributions de statistiques de test classiques extrapolées du χ^2 sont donc en général appliquées.

Zeng (1994 [115]) a aussi discuté l'intérêt de ce type d'approche pour détecter un QTL proche d'un cofacteur représentant un QTL sur le même chromosome. Il conclut à l'efficacité de la méthode dans les cas où les deux QTL sont dans des intervalles au moins séparés par un autre intervalle. Cette approche a été plus développée dans la continuité du travail de Zeng (1994 [115]) par Jiang et Zeng (1995 [40]) pour la discrimination d'un QTL pléiotrope et de deux QTL liés. En testant l'égalité des positions pour lesquelles

des QTL sont détectés sur deux caractères différents, sur la base de données simulées, ils mettent en évidence que la distinction entre deux QTL liés et un QTL pléiotrope est d'autant plus puissante que la densité en marqueurs génétiques est élevée.

1.2.1.2. Deux QTL liés Lander et Botstein (1989 [62]), dans le premier article proposant la cartographie d'intervalle comme alternative à la cartographie marqueur par marqueur, envisagent la détection de deux QTL en deux temps pour ne pas perdre l'unidimensionnalité du modèle uniQTL. La première étape est une cartographie de QTL uniQTL. La position et les effets du QTL ainsi détectés sont ensuite intégrés en covariable dans le modèle pour réaliser la deuxième détection. Cependant, la fixation de la position du premier QTL dans le deuxième modèle est limitée d'abord par la relativement faible précision de l'estimation de ce paramètre, et ensuite par le risque de prendre en compte dans le deuxième modèle non pas un vrai QTL, mais un QTL fantôme localisé entre les deux vrais QTL. La deuxième détection est donc probablement fortement biaisée dès le départ. De plus, le premier QTL pris en compte peut être un faux QTL au sens de l'erreur de première espèce. Celle-ci doit donc être fixée à un niveau faible dans le calcul des seuils pour la détection du premier QTL pour éviter de fausser complètement l'analyse. Ces remarques sont généralisables à toutes les méthodes basées sur des algorithmes itératifs de recherche. Nous verrons plus loin (1.2..2.) comment limiter leur influence.

Suite à ces constations, Haley et Knott (1992 [28]) et Martinez et Curnow (1992 [80]) ont proposé, dans le cadre de stratégies basées sur la régression multiple essentiellement pour des raisons de rapidité d'exécution, de cartographier simultanément les deux QTL. Cette proposition revient à ajouter l'effet d'un QTL dans le modèle de base, et éventuellement les interactions entre les deux QTL. Le nombre de paramètres estimés augmente donc, ainsi que le nombre de positions testées, puisqu'il faut alors tester tous les couples de positions possibles sur le groupe de liaison. Cette méthode, si elle a été validée par les auteurs, est limitée par sa réduction à la comparaison 1 contre 2 QTL. En effet, la prise en compte de QTL supplémentaires reviendrait à une augmentation très importante des temps de calcul à consacrer à cette hypothèse. Cependant, dans le cadre de cette question, peu de nouveaux développements de méthodes ont depuis été réalisés.

Dans le cadre de deux QTL liés sur un même chromosome, de Koning *et al.* (1998 [55]) ont montré que lorsque les cartes génétiques sont relativement peu denses, on ne peut pas prétendre distinguer plus de deux QTL liés sur un même groupe de liaison. En effet, le nombre de recombinaisons nécessaires à la mise en évidence de deux QTL liés est plus important que pour la cartographie d'un seul QTL, et impose l'utilisation d'une carte génétique relativement dense pour bien identifier les recombinants.

Rodolphe et Lefort, en 1993 [92], ont proposé une méthode adaptée à la détection de QTL liés pour des croisements entre lignées génétiquement fixées. Cette méthode est

basée sur la régression. Contrairement aux méthodes séquentielles destinées à la détection de QTL non liés, la méthode décrite par ces auteurs permet de détecter simultanément plusieurs QTL. Ils effectuent des développements analytiques pour montrer que les effets d'un QTL ne sont partagés que sur ses deux marqueurs flanquants. Ils montrent ainsi que deux QTL localisés dans le même intervalle génétique sont difficilement séparables. La méthode développée semble très efficace, en raison de propriétés asymptotiques connues. Cependant, les développements proposés sont spécifiques de croisements entre lignées génétiquement fixées, et leur application aux populations animales semble difficile.

Les développements multicaractères sont relativement récents et seront explicités dans la partie suivante.

1.2.2. Développements récents

1.2.2.1. MIM Multiple Interval Mapping Le MIM est une extension du CIM à la détection simultanée d'un nombre quelconque de QTL interagissant (Kao *et al.*. 1999 [41]). Sur la base du modèle défini pour le CIM (équation 3.1), le modèle du MIM peut s'extrapoler pour t QTL de la façon suivante :

$$y_i = b_0 + \sum_{h=1}^{t} \left(a_h x_{ih}^* + \sum_{k \neq h}^{t} \delta_{kh} w_{kh} x_{ih}^* x_{ik}^* \right) + \epsilon_i \qquad (3.2)$$

pour chaque individu $i = 1, \ldots, n$, et t QTL pris en compte dans le modèle,
avec x_{ih}^* la variable indicatrice du génotype au $h^{ième}$ QTL pour l'individu i,
a_h l'effet principal du QTL h,
δ_{kh} la variable indicatrice de l'épistasie entre les QTL h et k,
w_{kh} l'effet d'épistasie entre les QTL h et k,
ϵ_i l'erreur aléatoire normalement distribuée avec une moyenne nulle.

Il s'agit d'une méthode itérative de type sélection *stepwise*, avec des tests de rapport de vraisemblance à chaque itération. Le modèle de base est un modèle sans QTL. A chaque itération, le modèle peut être augmenté ou diminué d'un ou plusieurs QTL à la fois. La première étape consiste à déterminer des seuils de signification pour l'entrée et la conservation de QTL du modèle (SVE : *significant value for entry* et SVS : *significant value for staying* respectivement). Dans une deuxième étape, la contribution du QTL potentiel à la variation quantitative globale du caractère est calculée à chaque position. La recherche est donc dans un premier temps unidimensionnelle. Si pour certaines positions cette contribution est supérieure à SVE, elles sont sélectionnées et intégrées au modèle (sauf quand on teste $t = 1$, où on intégrera la position contribuant le plus à la variation du caractère pour éviter un arrêt prématuré de la procédure). Dans une troisième étape, les contributions

de chacun des QTL détectés à l'explicitation de la variation totale sont comparées à SVS, et les positions présentant des contributions inférieures à cette valeur sont éliminées. Ceci permet en particulier de corriger l'incorporation précoce de QTL fantômes. Les auteurs proposent enfin, dans une dernière étape, une recherche multidimensionnelle restreinte aux zones identifiées comme porteuses de QTL pour estimer conjointement les effets et les positions des QTL, estimations potentiellement plus précises que lors de la recherche itérative unidimensionnelle.

La pertinence des effets d'épistasie intégrés au modèle est testée séparément avec le MIM par un test de rapport de vraisemblance, pour lequel les auteurs conseillent d'utiliser un $\chi^2_{l,\alpha/k}$, où l est le nombre de QTL dans le modèle, α le niveau de signification recherché et k le nombre potentiel d'effets d'épistasie entre deux QTL dans un modèle à l QTL ($k = l(l-1)/2$). Ce test contient donc une approximation de la correction de Bonferroni pour le nombre d'effets d'épistasie intégrés.

La question de la discrimination entre QTL liés et QTL fantôme n'est pas très approfondie par les auteurs, si ce n'est qu'ils conseillent, pour les régions pour lesquelles il semblerait qu'il existe des QTL liés, de comparer l'hypothèse un QTL à celle de plusieurs QTL restreinte à cette seule région.

Grâce aux seuils de statistique de test d'entrée et de conservation des QTL dans le modèle, l'augmentation de la quantité d'information disponible est gérée afin d'éliminer des cofacteurs ou des QTL fantômes en cours d'analyse. SVE et SVS sont extrapolés à partir de seuils obtenus pour la cartographie d'intervalle 1 QTL, en ajoutant des corrections de Bonferroni. Cette approche semble très efficace, même si elle ne traduit pas forcément la totalité de la complexité du modèle. Cette méthode a été appliquée à l'analyse de données de 12 groupes de liaison sur le pin pour un backcross entre lignées fixées génétiquement (Kao *et al.*. 1999 [41]).

1.2.2.2. Algorithme génétique Carlborg *et al.* (2000 [10]) et Nakamichi *et al.* (2001, [83]) se sont indépendamment intéressés à la réduction mécanique des temps de calcul dans les situations de recherche simultanée de plusieurs QTL sur l'ensemble du génome. Dans le premier cas, la détection de QTL repose sur une extension de la méthode de Jansen et Stam (1994 [36]) à deux QTL. Dans le deuxième cas, la méthode de détection de QTL est dérivée du MIM et appliquée à m QTL. Les auteurs proposent de remplacer le balayage systématique des vecteurs de positions possibles par un algorithme de recherche pseudo-aléatoire, l'algorithme génétique (AG). Ce type d'algorithme, très efficace, permet de réduire considérablement les temps de calcul. Pour comparer les différents modèles entre eux et ne retenir que les plus pertinents, ces deux auteurs proposent l'utilisation du Critère d'Akaike [2], qui permet une comparaison rapide des modèles, qui ne sont pas nécessairement emboîtés ici.

1.2.2.3. Méthode des moindres carrés Knott et Haley (2000 [53]), de façon assez équivalente à Wu *et al.* (1999, [113]) dans le cadre de l'analyse de données sur le riz, ont proposé d'étendre la stratégie unicaractère proposée par Haley et Knott (1992, [28]) et Haley *et al.* (1994 [29]) à l'approche multicaractère multiQTL dans le cadre de l'analyse d'une F2. Le modèle de base est décrit par $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$, où $\mathbf{Y}$ est la matrice n x p contenant les p valeurs de performances pour les n individus, $\mathbf{X}$ est la matrice de correspondance n x v contenant les niveaux d'effets fixés, les covariables et les fonctions de probabilités des génotypes pour la(es) position(s) considérée(s), v étant le nombre de variables explicatives, $\mathbf{B}$ est la matrice v x p contenant les estimations pour chaque caractère des effets fixés, des covariables et des effets des génotypes, et $\mathbf{E}$ est la matrice n x p contenant les erreurs pour chaque caractère et chaque individu, supposées multinormalement distribuées. Pour un modèle strictement additif, la matrice $\mathbf{X}$ contient une colonne supplémentaire par QTL intégré à l'analyse. Les solutions de l'équation sont obtenues, comme pour une régression multiple classique, par $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. La matrice de la somme des carrés des résidus (*residual sums of squares* **RSS**) peut alors être obtenue avec $\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}\mathbf{X}'\mathbf{Y}$ pour les différents modèles. Les auteurs restreignent en réalité leur démonstration au cas de l'analyse de deux caractères.

1.3. Panorama sur la distribution des statistiques de test

1.3.1. Modèles et hypothèses comparées

En fonction du but recherché et des méthodes employées, le type d'hypothèses comparées est différent.

1.3.1.1. Plusieurs QTL agissant sur un caractère D'une façon générale, le test consiste à comparer H0 : $y = \sum_{t=1}^{Q-1} q_t + r$ à H1 $y = \sum_{t=1}^{Q} q_t + r$, donc à tester $q_Q = 0$. En théorie, ce test revient donc à ajouter la prise en compte d'une nouvelle position dans le test, donc tester un nouvel effet. Cependant, la majorité des modèles développés tentent de prendre en compte les interactions possibles entre les QTL. Le nombre de paramètres à estimer augmente alors rapidement, et les auteurs conseillent parfois, pour éviter la surparamétrisation, de limiter dans un premier temps la recherche au seul effet du QTL, et seulement si il est confirmé de tester la présence d'interactions avec les autres QTL précédemment détectés. De la même façon, pour limiter la détection précoce de QTL fantômes, certains auteurs conseillent de retester, à chaque ajout d'un QTL, la pertinence de la présence des autres, à travers par exemple la contribution du QTL à la statistique de test globale (Kao *et al.*, 1999 [41]).

1.3.1.2. Deux QTL liés contre un QTL pléiotrope D'une façon générale, dans le test de deux QTL liés contre un QTL pléiotrope, les hypothèses sont en général limitées à l'analyse de deux caractères. L'hypothèse nulle est celle d'un QTL pléiotrope, avec $\mathbf{y} = \mathbf{q} + \mathbf{r}$, où $\mathbf{y} = (y_1; y_2)$, $\mathbf{q} = (q_1; q_2)$, où q_l est l'effet du QTL sur le caractère l, et $\mathbf{r} = (r_1; r_2)$, et l'hypothèse générale est celle de deux QTL liés, avec $\mathbf{y} = \mathbf{q} + \mathbf{r}$, où $\mathbf{y} = (y_1; y_2)$, $\mathbf{q} = (q^1; q^2)$, où q^l est l'effet du QTL l sur le caractère l, et $\mathbf{r} = (r_1; r_2)$. Entre les deux hypothèses, le seul paramètre supplémentaire est donc une position. L'extension de ce type de test à plusieurs caractères est relativement simple, même si elle oblige à multiplier le nombre de tests. Cette hypothèse nulle est différente des hypothèses nulles classiques où l'on teste l'absence de QTL, et requiert une attention particulière. En particulier, la présence du premier QTL implique que les tests de permutation (Doerge et Churchill, 1996 [16]) qui ont pu être proposés ne sont plus applicables.

Knott et Haley (2000 [53]) proposent une succession de tests qui permet d'aboutir à la distinction entre un QTL pléiotrope et deux QTL liés agissant chacun sur un caractère. Le premier consiste à comparer un modèle avec deux caractères et sans QTL avec un modèle avec deux caractères et un QTL pléiotrope. Ce test correspondrait à un χ^2 à p x ddl, où ddl est le nombre de degrés de liberté du modèle avec QTL, si la procédure n'impliquait pas de réaliser de nombreux tests corrélés. Les auteurs proposent donc d'utiliser un test de permutations (Churchill et Doerge, 1994 [11]) pour générer la distribution de la statistique de test sous l'hypothèse nulle.

Un deuxième test consiste à comparer le modèle sans QTL avec celui de deux QTL liés agissant chacun sur un caractère. Pour ce second modèle, les auteurs utilisent les estimations des paramètres à la position la plus probable obtenue par les analyses uniQTL unicaractère séparées pour calculer directement $\hat{\mathbf{B}}_1$. **RSS** est alors obtenue par $(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_1)'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_1)$. De la même façon que précédemment, ce test correspond à un χ^2 à p x ddl, où ddl est le nombre de degrés de liberté du modèle avec QTL.

Le dernier test consiste en la comparaison des modèles "1 QTL pléiotrope" contre "2 QTL liés agissant chacun sur un caractère". Nous développerons dans la partie suivante la méthode utilisée pour tester la signification des statistiques de test ainsi obtenues.

Les auteurs comparent dans différents cas les puissances et précisions relatives de deux stratégies : 1) envisager la discrimination 1 QTL pléiotrope contre deux QTL liés après avoir validé l'hypothèse d'un QTL pléiotrope contre pas de QTL, ou 2) envisager la discrimination 1 QTL pléiotrope contre deux QTL liés après avoir validé l'hypothèse de deux QTL liés contre pas de QTL. En fonction du modèle de détermination des performances, les résultats sont différents, et le choix de la stratégie à adopter reste donc difficile. De plus, les auteurs limitent le test 1 QTL pléiotrope contre deux QTL liés aux seuls cas significatifs pour le test préliminaire, ce qui tend à limiter la puissance réelle des tests,

en particulier quand les effets de deux QTL liés sont en opposition, et donc souvent non détectés par les méthodes uniQTL.

1.3.2. Statistiques de test

Chaque comparaison de modèles correspond quasiment à une statistique de test différente, en fonction des vecteurs de positions considérés, des interactions prises en compte, et du nombre de caractères impliqués. Les statistiques de test développées dans le cadre des études multiQTL sont donc généralement mal connues, et les critères de discrimination entre les hypothèses souvent empiriques. De nombreux auteurs ont dans un premier temps limité leurs analyses à la double comparaison de l'hypothèse d'un QTL pléiotrope avec l'analyse uniQTL et unicaractère pour chacun des caractères. La présence d'un QTL pléiotrope est alors acceptée quand les deux tests sont significatifs. L'assise statistique de ces tests est malgré tout limitée, même s'ils peuvent donner une idée sur l'opportunité de poursuivre plus loin l'analyse. De tels tests sont en général appliqués à des balayages de l'ensemble du génome, où seuls les effets estimés dans les modèles complets seront complètement testés (Carlborg *et al.*, 2000 [10]; Nakamichi *et al.*, 2001 [83]).

Les méthodes basées sur des recherches itératives, qui impliquent des tests nombreux et répétés, font souvent appel à des critères d'information, le seul utilisé étant celui d'Akaike [2] dans la pratique, ou à des extrapolations des seuils obtenus pour les méthodes uniQTL. Ces méthodes, si elles sont moins bien caractérisées statistiquement que l'utilisation d'une statistique de test classique, sont en général moins sévères , et permettent de sélectionner les modèles les plus complets en limitant le taux de faux négatifs.

D'une façon générale, les statistiques de test tendant à se compliquer beaucoup avec la multiplication du nombre de paramètres pris en compte, certains auteurs préfèrent se ramener à une connaissance empirique de la statistique de test grâce à des permutations (Churchill et Doerge, 1994 [11], qui n'est pas applicable à un test de QTL pléiotrope contre deux QTL liés), ou à des simulations. Certains auteurs (Wu et Li, 1994 [112]; Whittaker *et al.* 1996 [110]) ont proposé de valider les hypothèses "deux QTL" grâce à des tests de χ^2 à 2 degrés de liberté. Cependant, Goffinet et Mangin (1998, [25]) montrent que leur justification n'est pas valable, et conseillent de se reporter sur l'estimation des seuils par simulations.

Jiang et Zeng (1995 [40]) ont suggéré que le LOD score obtenu suite à une comparaison 1 QTL contre 2 QTL suive un χ^2 à un degré de liberté (pour la position supplémentaire) sous l'hypothèse nulle de pléiotropie. Cependant, Van Ooijen (1992 [85]) et Mangin *et al.* (1994 [75]) ont montré que cette approximation était loin de la distribution réelle quand l'effet du QTL est fort ou moyen.

Doerge et Churchill, en 1996 ([16]), reprenant la technique de calcul de la distribution de la statistique de test proposée par Churchill et Doerge en 1994 [11], l'adaptent à la cartographie de plusieurs locus localisés sur des groupes de liaison différents. Il s'agit d'une stratégie basée sur des permutations des performances des descendants par rapport aux génotypages. L'approche de Churchill et Doerge (1994, [11]) permet, en cassant les liaisons existant entre les valeurs de performances et les génotypages, de reproduire la distribution de la statistique de test sous l'hypothèse nulle sans faire la moindre hypothèse sur la forme de la distribution des performances. Elle s'oppose en cela aux méthodes de calcul de seuils basées sur des simulations, qui nécessairement supposent une distribution sous jacente des valeurs génétiques pour chacune des classes génotypiques au QTL.

Doerge et Churchill (1996 [16]) proposent deux adaptations de cette méthode au calcul de seuils pour la détection itérative de plusieurs QTL. Les deux stratégies commencent par une détection uniQTL et un calcul de seuil tels que présenté dans Churchill et Doerge (1994, [11]). La première méthode, appelée seuil empirique conditionné (*conditional empirical threshold*: CET), consiste à effectuer une deuxième itération, par exemple avec une méthode à covariables. Les individus sont ensuite répartis en 2 (pour un backcross) ou 3 (pour une F2) classes génotypiques en fonction de leur génotype au premier QTL. Les permutations sont réalisées intra-classes génotypiques, ce qui permet de respecter la prise en compte du premier QTL dans les permutations tout en cassant les liaisons marqueur/phénotype dues au deuxième QTL potentiel. D'autres itérations peuvent être ajoutées pour chaque ajout de QTL en subdivisant à chaque tour les classes génotypiques conditionnellement aux génotypes aux QTL précédemment identifiés. Les auteurs réservent chaque nouvelle itération aux groupes de liaison sur lesquels des QTL n'ont pas encore été détectés, pour éviter l'inclusion de marqueurs liés aux marqueurs servant à conditionner la statistique de test, afin de limiter l'accroissement de la valeur des seuils (Zeng, 1994 [115]). Cette première méthode est à nouveau complètement non-paramétrique.

La deuxième méthode, seuil empirique résiduel (*residual empirical threshold*: RET), est semi-paramétrique. Elle repose sur la modélisation des performances en fonction du premier QTL, par exemple comme ayant des effets uniquement additifs. Les performances de chaque individu sont alors corrigées pour l'effet du premier QTL en fonction de leur génotype au QTL. Les performances analysées pour la détection d'un nouveau QTL sont les résidus issus de cette correction, toute nouvelle association performance/marqueur étant supposée due à un autre locus. Les seuils sont alors calculés selon la procédure standard présentée dans Churchill et Doerge (1994 [11]). Cette deuxième méthode est semi-paramétrique, donc plus fortement conditionnée par la prise en compte d'un modèle de ségrégation des performances pertinent dans l'analyse. Cependant, elle évite la subdivision successive des classes génotypiques, qui peut être rapidement limitée par le nombre

d'individus disponibles dans le schéma expérimental.

Les deux types de tests, comme tous les tests de type itératifs proposés, reposent fortement sur l'hypothèse que les QTL détectés dans les itérations précédant l'itération courante sont de vrais QTL et non pas de faux positifs. Les auteurs ont validé leurs méthodes d'une part par une explicitation analytique des tests réalisés, essentiellement basée sur la théorie de Lehman (1986, [66]) et Schmoyer (1994 [96]), et des simulations pour la détection de 2 QTL sur quatre groupes de liaison, ou trois QTL dont deux sur le même groupe de liaison pour le RET. Ces simulations ont permis de mettre en évidence des comportements différents des deux méthodes proposées, avec en général une puissance plus élevée pour le RET pour détecter le bon nombre de QTL quand il y en a 0 ou 1, mais moins élevée que le CET quand il y a en a deux. Les deux méthodes de Doerge et Churchill (1996 [16]) ne permettent cependant pas de détecter des QTL liés, puisque les groupes de liaison portant un QTL précédemment détecté sont exclus des nouvelles itérations.

Suite aux travaux sur l'utilisation du bootstrap de Visscher *et al.* (1996 [104]) et Lebreton et Visscher (1998 [64]), Lebreton *et al.* (1998 [65]) ont proposé l'utilisation de bootstrap non paramétriques pour estimer des intervalles de confiance des positions de deux QTL liés, ce qui permettrait de rejeter l'hypothèse d'un QTL pléiotrope si la position estimée pour celui-ci ne se trouve pas dans la zone de recouvrement. Cette méthode présente l'avantage de s'affranchir totalement de la notion de distribution de la statistique de test sous l'hypothèse nulle, mais oblige à réitérer les analyses à chaque bootstrap. Par rapport aux auteurs précédemment cités, elle permet de plus de réduire le biais d'estimation des positions et semble relativement robuste pour la gamme de paramètres testée par les auteurs.

Knott et Haley (2000 [53]), reprenant une proposition de Walling *et al.* (1998, [105]) initialement destinée à la caractérisation des intervalles de confiance des positions estimées, ont suggéré d'utiliser les estimations des paramètres obtenues avec le modèle complet (QTL pléiotrope) pour simuler les données sous l'hypothèse nulle et ainsi obtenir la distribution de la statistique de test correspondante. Ils proposent, pour chaque analyse de données, d'effectuer des simulations à partir des paramètres estimés (effets et position du QTL pléiotrope) et d'estimer un seuil particulier. Cette stratégie, particulièrement exigeante en terme de calculs, est cependant bien validée. Ils comparent cette stratégie avec le bootstrap non paramétrique proposé par Lebreton *et al.* (1998 [65]), qui utilise la notion de recouvrement des intervalles de confiance définis pour chacun des QTL liés pour trancher entre les deux hypothèses. Cette stratégie est *a priori* parfaitement adaptable à la caractérisation de l'hypothèse nulle "il y a un QTL agissant sur un caractère" dans le cas unicaractère.

1.4. Application aux populations animales

L'ensemble des méthodes présentées ci-dessus ont été développées dans un objectif d'application à des croisements entre lignées *inbred*, donc en particulier sans tenir compte d'éventuelles structures familiales.

Certaines extensions des méthodes présentées ont cependant été réalisées pour l'application à des croisements outbred (Jansen *et al.*, 1996 et 1998 [38] [39]; Kao *et al.*, 1999 [41]). Ces méthodes exigent néanmoins souvent l'estimation d'un nombre de paramètres très important pour chaque famille, du fait de l'estimation de nombreux cofacteurs. Or Sakamoto *et al.* (1986 [95]) ont montré que le nombre de paramètres d'un modèle ne doit pas dépasser le double de la racine carrée du nombre d'observations, ce qui correspond à 15 cofacteurs pour 50 individus d'une famille, ce qui est largement insuffisant pour prendre en compte l'ensemble du génome. Jansen *et al.* (1998 [39]) ont ainsi développé une méthode qu'ils réservent à l'analyse d'un unique groupe de liaison, applicable aux pedigree complexes des espèces animales, mais difficilement extrapolable à la prise en compte d'un plus grand nombre de groupes de liaison en raison des temps de calcul requis.

La seule étude complète à ce jour de méthodes multiQTL applicables à des populations animales est celle de de Koning *et al.* (2001b [56]). Elle a été développée pour l'application à un protocole bovin laitier de type petites filles. Il s'agit d'une méthode itérative, basée sur la cartographie d'intervalle développée par Knott *et al.* (1996 [51]). Cette méthode est proche, dans le principe, de celle proposée par Doerge et Churchill (1996 [16]) pour l'obtention de seuils empiriques sur les résidus, mais elle ne fixe pas les effets des QTL mis en cofacteurs.

Le schéma global des différentes étapes est donné figure 17. La première étape est une analyse uniQTL qui permet d'identifier les régions candidates à l'intégration comme cofacteurs dans l'analyse. Dans une deuxième étape, les positions correspondantes sont intégrées en cofacteur, et leurs effets sont réestimés conjointement par régression linéaire multiple. La troisième étape consiste à ajuster les données phénotypiques pour les effets ainsi estimés. L'analyse est alors reprise à la première étape pour détecter d'éventuelles nouvelles régions candidates, en excluant du modèle les cofacteurs localisés sur le groupe de liaison analysé, pour maximiser la puissance de détection du modèle (Zeng, 1994 [115]; Doerge et Churchill, 1996 [16]). Cette méthode est donc destinée à l'identification simultanée de QTL non liés.

Les auteurs proposent différents types de seuils de signification en fonction de l'étape réalisée. Les seuils sont calculés par permutation selon la méthode proposée par Churchill et Doerge (1994 [11]) et Doerge et Churchill (1996 [16]), et appliquée à diverses structures de population de demi frères (Spelman *et al.*, 1996 [100]; Vilkki *et al.*, 1997 [102]) pour chaque chromosome. Des seuils à 5% au niveau du chromosome sont utilisés pour intégrer

un QTL potentiel comme cofacteur dans l'analyse, alors qu'une correction de Bonferroni appliquée aux seuils nominaux permet d'obtenir des seuils au niveau du génome, utilisés pour rejeter l'absence de QTL selon la terminologie de QTL "suggestif" ou "significatif" décrite par Lander et Kruglyak (1995 [61]). Les seuils utilisés pour l'intégration des cofacteurs sont donc nettement moins sévères que ceux pour la cartographie des QTL, comme suggéré par Spelman *et al.* en 1996 [100].

La méthode de de Koning *et al.* (2001 [56]) présente l'avantage de réaliser un maximum d'analyses unidimensionnelles, pour lesquelles les seuils sont estimés à chaque itération par permutations. Les auteurs ont validé leur approche par une application à des données finlandaises de production laitière, permettant de mettre en évidence conjointement 8 QTL affectant la production laitière, dont trois n'avaient pas été détectés par les analyses uniQTL. Cependant, elle reste très demandeuse en terme de temps de calcul, et l'ajustement des phénotypes aux effets trouvés a soulevé certaines critiques quand à l'efficacité de l'algorithme de recherche. De plus, cette méthode n'a pas donné lieu à une validation théorique très importante, et certains points, tels que l'impact de l'adaptation des seuils au nombre de cofacteurs significatifs pour chaque chromosome ou l'ajustement à chaque famille du nombre de cofacteurs pris en compte (ici les familles non informatives sont intégrées, considérant que cela crée éventuellement du bruit de fond mais pas de biais), mériteraient d'être approfondis.

D'une façon générale, il semble donc que l'ensemble des méthodes décrites est très demandeur en terme de temps de calcul. Nous avons donc choisi d'orienter ce travail sur la mise en place de méthodes "rapides" de détection, qui puissent permettre d'identifier des zones chromosomiques qui pourront ensuite être analysées de façon exhaustives grâce à des méthodes longues mais plus complètes (Hoeschele *et al.*, 1997, [33]).

2. Cadre de l'étude

L'objectif principal de ce travail est la discrimination de deux QTL liés et d'un QTL pléiotrope. Nous nous sommes donc limités à la détection de plusieurs QTL sur un seul groupe de liaison, dans les cas d'analyses unicaractère et multicaractère. Ce type d'approche, comme nous l'avons vu précédemment, exclut d'emblée la possibilité d'utiliser les méthodes à cofacteurs en raison du manque de précision de l'estimation de la position du premier QTL détecté et de la possibilité de détecter des QTL fantômes. Il faudra donc, pour chaque méthode décrite, tester l'ensemble des couples de positions possibles.

Bien que nous ne l'envisagions pas ici, il faut cependant souligner l'intérêt non négligeable des approches multiQTL envisageant plusieurs groupes de liaison à la fois, en particulier dans le cadre de l'exploration de données issues de populations animales sé-

lectionnées. On peut en effet observer dans les familles de taille restreinte l'apparition de co-ségrégation d'allèles de différents QTL sur des chromosomes différents uniquement par chance (Farnir *et al.*, 2000 [21]).

2.1. Test d'hypothèse appliqué

Les études présentées ici se sont limitées à l'analyse de un ou deux caractères simultanément. Dans le cas d'un caractère, l'hypothèse nulle est "il y a un QTL sur le groupe de liaison", comparée à l'hypothèse générale "il y a deux QTL sur le groupe de liaison". Quand deux caractères sont analysés simultanément, l'hypothèse nulle est "il y a un QTL pléiotrope sur le groupe de liaison", comparée à deux types d'hypothèses générales : soit "il y a deux QTL sur le groupe de liaison, chacun agissant sur un caractère", soit "il y a deux QTL pléiotropes sur le groupe de liaison".

Dans les deux cas, la statistique de test est un rapport de maximum de vraisemblance. Si on note $max(\Lambda_x)$ le maximum de la vraisemblance sous H0, trouvé à la position x, et $max(\Lambda_{\mathbf{x}})$ le maximum de la vraisemblance sous l'hypothèse générale, trouvé pour le couple de position $\mathbf{x} = (x_1; x_2)$, la statistique de test s'écrit :

$$MLRT = -2\ln(max(\Lambda_x)/max(\Lambda_{\mathbf{x}})) \tag{3.3}$$

Dans les deux cas, que l'on cherche à discriminer deux QTL sur un même caractère ou agissant sur deux caractères, on peut supposer que les analyses uniQTL ont déjà été réalisées, et donc que $max(\Lambda_x)$ est connu. Ces vraisemblances ont toutes été décrites dans la partie sur les analyses multicaractères et nous ne reviendrons pas dessus dans cette partie.

2.2. Méthodes comparées

2.2.1. Analyse unicaractère (ST2)

La méthode unicaractère multiQTL que nous avons utilisée est la retranscription sur deux positions de la méthode uniQTL précédemment présentée. Cette approche correspond à celle préconisée par Haley et Knott (1992 [28]) et Martinez et Curnow (1992 [80]), mais appliquée au test de maximum de vraisemblance pour des mélanges de demi et de plein frères.

2.2.1.1. Ecriture de la vraisemblance La vraisemblance s'écrit de la même façon que pour une analyse uniQTL (voir paragraphe 4.1., équation 1.9) :

$$\Lambda^{\mathbf{x}} = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \sum_{hd_{ij}} p(hd_{ij}/\widehat{hs_i}, M_i) \prod_{k=1}^{n_{ij}} f(yp_ijkl/\widehat{hs_i}, hd_{ij}, M_i) \tag{3.4}$$

mais avec

$$Y_{ijkl} = yp_{ijkl} - \sum_{\mathbf{q_s}=(1,1)}^{(2,2)} \sum_{\mathbf{q_d}=(1,1)}^{(2,2)} p\left(d^{\mathbf{x}}_{ijk} = (\mathbf{q_s}, \mathbf{q_d})/\widehat{hs_i}, hd_{ij}, M_i\right) \left(\mu_{il}^{\mathbf{xq_s}} + \mu_{ijl}^{\mathbf{xq_d}}\right) \tag{3.5}$$

où $p\left(d^{\mathbf{x}}_{ijk} = (\mathbf{q_s}, \mathbf{q_d})/\widehat{hs_i}, hd_{ij}, M_i\right)$ est la probabilité de transmission du couple d'haplotypes $\mathbf{q_s} = (q1_s, q2_s)$ et du couple d'haplotypes $\mathbf{q_d} = (q1_d, q2_d)$, probabilité que le descendant ijk ait reçu de son père le segment chromosomique $q1_s$ à la position testée $x1$ et $q2_s$ à la position testée $x2$, avec $\mathbf{x} = (x1; x2)$ (qr_s=1 si qr_s vient du grand-père paternel, qr_s=2 si qr_s vient de la grand-mère paternelle, $r = 1, 2$) et que le descendant ait reçu de sa mère le segment chromosomique $q1_d$ à la position testée $x1$ et $q2_d$ à la position testée $x2$ (qr_d=1 si qr_d vient du grand-père maternel , qr_d=2 si qr_d vient de la grand-mère maternelle, $r = 1, 2$),

$\mu_{il}^{\mathbf{xq_s}}$ est la moyenne des performances des descendants qui ont reçu les segments chromosomiques $q1_s$ et $q2_s$ au couple de positions $\mathbf{x} = (x1; x2)$ du père i pour le caractère quantitatif l ,

$\mu_{ijl}^{\mathbf{xq_d}}$ est la moyenne des performances des descendants qui ont reçu les segments chromosomiques $q1_d$ et $q2_d$ au couple de position $\mathbf{x} = (x1; x2)$ de la mère ij pour le caractère quantitatif l.

μ_{il}, μ_{ijl}, α_{il}^{x1}, α_{il}^{x2}, α_{ijl}^{x1} et α_{ijl}^{x2} étant respectivement les moyennes et les effets de substitution moyens de chaque QTL pour le caractère l, intra-famille de demi-frères et intra-famille de plein-frères, $\mu_{il}^{\mathbf{xq_s}}$ et $\mu_{ijl}^{\mathbf{xq_d}}$ peuvent s'écrire :

$$\begin{array}{ll}
\mu_{il}^{\mathbf{x}(1,1)} = \mu_{il} + \alpha_{il}^{x1}/2 + \alpha_{il}^{x2}/2, & \mu_{il}^{\mathbf{x}(1,2)} = \mu_{il} + \alpha_{il}^{x1}/2 - \alpha_{il}^{x2}/2, \\
\mu_{il}^{\mathbf{x}(2,1)} = \mu_{il} - \alpha_{il}^{x1}/2 + \alpha_{il}^{x2}/2, & \mu_{il}^{\mathbf{x}(2,2)} = \mu_{il} - \alpha_{il}^{x1}/2 - \alpha_{il}^{x2}/2, \\
\mu_{ijl}^{\mathbf{x}(1,1)} = \mu_{ijl} + \alpha_{ijl}^{x1}/2 + \alpha_{ijl}^{x2}/2 & \mu_{ijl}^{\mathbf{x}(1,2)} = \mu_{ijl} + \alpha_{ijl}^{x1}/2 - \alpha_{ijl}^{x2}/2 \\
\mu_{ijl}^{\mathbf{x}(2,1)} = \mu_{ijl} - \alpha_{ijl}^{x1}/2 + \alpha_{ijl}^{x2}/2 & \mu_{ijl}^{\mathbf{x}(2,2)} = \mu_{ijl} - \alpha_{ijl}^{x1}/2 - \alpha_{ijl}^{x2}/2 \ .
\end{array}$$

Le nombre de paramètres à estimer est donc augmenté d'un effet par père et par mère pour le QTL supplémentaire, soit $4n + 3\sum_{i=1}^{n} n_i$, plus la position, par rapport à une analyse uniQTL.

2.2.1.2. Linéarisation de la vraisemblance La linéarisation de la vraisemblance pour une position x est donnée par Goffinet *et al.* (1999, [26]). Elle consiste à transformer

$$\sum_{q_s=1}^{2} \sum_{q_d=1}^{2} p\left(d^{x}_{ijk} = (q_s, q_d)/\widehat{hs_i}, hd_{ij}, M_i\right) \left(\mu_{il}^{xq_s} + \mu_{ijl}^{xq_d}\right)$$

$= p_1(\mu^x_{il}-a^x_{il}+\mu^x_{ijl}-a^x_{ijl})+p_2(\mu^x_{il}-a^x_{il}+\mu^x_{ijl}+a^x_{ijl})+p_3(\mu^x_{il}+a^x_{il}+\mu^x_{ijl}-a^x_{ijl})+p_4(\mu^x_{il}+a^x_{il}+\mu^x_{ijl}+a^x_{ijl})$

en $\mu^x_{il} + \mu^x_{ijl} + p^x_p a^x_{il} + p^x_m a^x_{ijl}$,
avec $p^x_p = -p_1 - p_2 + p_3 + p_4$, $p^x_m = -p_1 + p_2 - p_3 + p_4$, $a^x_{il} = \alpha^x_{il}/2$ et $a^x_{ijl} = \alpha^x_{ijl}/2$.

Dans le cas multiQTL, pour le couple de positions $\mathbf{x} = (x_1; x_2)$,

$$\sum_{\mathbf{q_s}=(1,1)}^{(2,2)} \sum_{\mathbf{q_d}=(1,1)}^{(2,2)} p\left(d^{\mathbf{x}}_{ijk} = (\mathbf{q_s}, \mathbf{q_d})/\widehat{hs}_i, hd_{ij}, M_i\right)\left(\mu^{\mathbf{xq_s}}_{il} + \mu^{\mathbf{xq_d}}_{ijl}\right)$$

$$\begin{aligned}
&= p_1(\mu_{il} + a^{x1}_{il} + a^{x2}_{il} + \mu_{ijl} + a^{x1}_{ijl} + a^{x2}_{ijl}) + p_2(\mu_{il} + a^{x1}_{il} + a^{x2}_{il} + \mu_{ijl} + a^{x1}_{ijl} - a^{x2}_{ijl}) \\
&+p_3(\mu_{il} + a^{x1}_{il} + a^{x2}_{il} + \mu_{ijl} - a^{x1}_{ijl} + a^{x2}_{ijl}) + p_4(\mu_{il} + a^{x1}_{il} + a^{x2}_{il} + \mu_{ijl} - a^{x1}_{ijl} - a^{x2}_{ijl}) \\
&+p_5(\mu_{il} + a^{x1}_{il} - a^{x2}_{il} + \mu_{ijl} + a^{x1}_{ijl} + a^{x2}_{ijl}) + p_6(\mu_{il} + a^{x1}_{il} - a^{x2}_{il} + \mu_{ijl} + a^{x1}_{ijl} - a^{x2}_{ijl}) \\
&+p_7(\mu_{il} + a^{x1}_{il} - a^{x2}_{il} + \mu_{ijl} - a^{x1}_{ijl} + a^{x2}_{ijl}) + p_8(\mu_{il} + a^{x1}_{il} - a^{x2}_{il} + \mu_{ijl} - a^{x1}_{ijl} - a^{x2}_{ijl}) \\
&+p_9(\mu_{il} - a^{x1}_{il} + a^{x2}_{il} + \mu_{ijl} + a^{x1}_{ijl} + a^{x2}_{ijl}) + p_{10}(\mu_{il} - a^{x1}_{il} + a^{x2}_{il} + \mu_{ijl} + a^{x1}_{ijl} - a^{x2}_{ijl}) \\
&+p_{11}(\mu_{il} - a^{x1}_{il} + a^{x2}_{il} + \mu_{ijl} - a^{x1}_{ijl} + a^{x2}_{ijl}) + p_{12}(\mu_{il} - a^{x1}_{il} + a^{x2}_{il} + \mu_{ijl} - a^{x1}_{ijl} - a^{x2}_{ijl}) \\
&+p_{13}(\mu_{il} - a^{x1}_{il} - a^{x2}_{il} + \mu_{ijl} + a^{x1}_{ijl} + a^{x2}_{ijl}) + p_{14}(\mu_{il} - a^{x1}_{il} - a^{x2}_{il} + \mu_{ijl} + a^{x1}_{ijl} - a^{x2}_{ijl}) \\
&+p_{15}(\mu_{il} - a^{x1}_{il} - a^{x2}_{il} + \mu_{ijl} - a^{x1}_{ijl} + a^{x2}_{ijl}) + p_{16}(\mu_{il} - a^{x1}_{il} - a^{x2}_{il} + \mu_{ijl} - a^{x1}_{ijl} - a^{x2}_{ijl}) \qquad (3.6)
\end{aligned}$$

En notant $ps_1 = p_1+p_2+p_3+p_4+p_5+p_6+p_7+p_8-p_9-p_{10}-p_{11}-p_{12}-p_{13}-p_{14}-p_{15}-p_{16}$
$ps_2 = p_1 + p_2 + p_3 + p_4 - p_5 - p_6 - p_7 - p_8 + p_9 + p_{10} + p_{11} + p_{12} - p_{13} - p_{14} - p_{15} - p_{16}$
$ps_3 = p_1 + p_2 - p_3 - p_4 + p_5 + p_6 - p_7 - p_8 + p_9 + p_{10} - p_{11} - p_{12} + p_{13} + p_{14} - p_{15} - p_{16}$
$ps_4 = p_1 - p_2 + p_3 - p_4 + p_5 - p_6 + p_7 - p_8 + p_9 - p_{10} + p_{11} - p_{12} + p_{13} - p_{14} + p_{15} - p_{16}$

on a donc :

$$Y_{ijkl} = yp_{ijkl} - \left(\mu^{\mathbf{x}}_{il} + \mu^{\mathbf{x}}_{ijl} + ps_1 a^{x1}_{il} + ps_2 a^{x2}_{il} + ps_3 a^{x1}_{ijl} + ps_4 a^{x2}_{ijl}\right) \qquad (3.7)$$

En reprenant les p_i des notations unicaractères indicées d'un r pour le QTL r, on obtient donc facilement : $ps_1 = p_{11} + p_{21} - p_{31} - p_{41}$, $ps_2 = p_{12} + p_{22} - p_{32} - p_{42}$,
$ps_3 = p_{11} - p_{21} + p_{31} - p_{41}$, $ps_4 = p_{12} - p_{22} + p_{32} - p_{42}$.

2.2.2. Analyses multicaractères

2.2.2.1. Analyses multivariées (MV2) Comme dans le cas uniQTL, la première méthode multicaractère est multivariée. Elle prend en compte deux positions possibles sur le groupe de liaison. En dehors de la linéarisation des probabilités de transmission des haplotypes, qui s'écrit comme au paragraphe précédent, la vraisemblance s'écrit donc de la même façon que pour le cas uniQTL :

$$f(\mathbf{yp}_{ijk}/\widehat{hs}_i, hd_{ij}, M_i) = \sqrt{\frac{|\mathbf{VC}_i^{-1}|}{2\pi}} \exp\left(-\frac{1}{2}\mathbf{Y}'_{ijk}\mathbf{VC}_i^{-1}\mathbf{Y}_{ijk}\right) \tag{3.8}$$

où $\mathbf{Y}_{ijk} = \{Y_{ijkl}; l = 1, \ldots, p\}$ tels que définis dans l'équation (3.7).

Par caractère, on a donc un effet de plus à estimer par rapport au cas uniQTL multicaractère, plus 1 pour le nombre de positions, soit $(4n + 3\sum_{i=1}^{n} n_i)p + p(p-1)/2 + 2$ paramètres, où p est le nombre de caractères intégrés à l'analyse.

Pour tester des QTL liés agissant sur deux caractères différents, on peut choisir de ne rendre les effets sur chaque caractère dépendant que d'une position sur les deux (MV2r), ce qui réduit le nombre de paramètres à estimer pour deux caractères à $(3n + 2\sum_{i=1}^{n} n_i) * 2 + 2 + 1$, identique, par rapport au test uniQTL, à la position supplémentaire près, mais à tester pour tous les couples de positions.

2.2.2.2. Analyse discriminante appliquée à deux positions (DA2) L'analyse discriminante consiste, dans son application à la détection de QTL uniQTL, à déterminer la combinaison linéaire qui maximise le rapport de la variation due au QTL sur la variation résiduelle, à partir des performances pondérées de leur probabilité d'avoir reçu un haplotype paternel donné à la position testée pour le QTL. L'extension à la détection de deux QTL revient à ne plus considérer les deux haplotypes paternels à une position, mais les quatre couples d'haplotypes pour le couple de positions testé (figure 16). Les pondérations des performances dans chacun des quatre groupes sont à nouveau dérivées des $p\left(d_{ijk}^{\mathbf{x}} = (\mathbf{q_s}, \mathbf{q_d})/\widehat{hs}_i, hd_{ij}, M_i\right)$. Par cette stratégie, on obtient non plus une, mais trois combinaisons linéaires des caractères, étant donné que quatre groupes servent à la définition du schéma. Cependant, par définition, seule celle expliquant le plus grand rapport de variabilités sert à détecter les QTL.

De la même façon que pour les stratégies uniQTL (équation 2.5), on peut calculer la valeur propre correspondant au vecteur propre définissant la combinaison linéaire de deux caractères déterminés par des QTL pléiotropes liés. Cependant, la corrélation génétique ρ_g entre les caractères n'est plus égale à 1 du fait de la distance entre les deux QTL déterminant les caractères. En reprenant les notations de l'équation 2.5, la valeur propre

λ est donc la plus grande solution de l'équation :

$$\left[\lambda(1-\rho^2)\right]\left[\lambda(1-\rho^2)-\left[\sigma_1^2+\sigma_2^2-2\rho\rho_g\sigma_1\sigma_2\right]\right]+\sigma_1^2\sigma_2^2\left[1+\rho^2\rho_g^2-\rho^2-\rho_g^2\right]=0 \quad (3.9)$$

La fonction de vraisemblance à maximiser est identique à la fonction unicaractère (équation 3.4), où les paramètres sont estimés pour la combinaison linéaire des caractères spécifique du couple de positions testé. De la même façon que pour l'analyse multiQTL d'un caractère, un effet par père et par mère, ainsi qu'une position, doivent donc être estimés en plus par rapport à l'hypothèse nulle.

2.3. Axes de comparaison

Les méthodes unicaractère et multicaractères uniQTL seront présentées dans cette partie comme référence de comparaison des résultats obtenus avec les méthodes équivalentes mais multiQTL. Pour les méthodes uniQTL, les temps de calcul, les puissances et les précisions d'estimation des paramètres sont estimés de la même façon que présenté dans la partie multicaractère. Ceci implique en particulier que les puissances des méthodes multicaractères ne sont pas directement comparables à celles des méthodes multiQTL, puisqu'elles ne correspondent pas aux mêmes tests. Les comparaisons seront alors du type "sur les x% de QTL détectés par les méthodes uniQTL, y% seront attribués à l'action de deux QTL".

Nous présenterons, quelles que soient les situations simulées, les résultats obtenus avec toutes les méthodes utilisées. En particulier, les analyses multicaractères ont été appliquées aux cas unicaractères et inversement.

2.3.1. Temps de calcul

Les tests multiQTL consistent à calculer $MLRT = -2\ln(max(\Lambda_x)/max(\Lambda_{\mathbf{x}}))$, où x est la position du maximum de la vraisemblance sous l'hypothèse de la présence d'un seul QTL, et $\mathbf{x}$ est le couple de positions au maximum de vraisemblance sous l'hypothèse alternative de deux QTL. Partant du principe que les tests uniQTL sont réalisés préalablement à l'exploration multiQTL, seul le temps nécessaire à la maximisation de la vraisemblance sous l'hypothèse de deux QTL est considéré dans cette partie.

De la même façon que dans la partie unicaractère, les temps de calcul comparés correspondent au temps nécessaire à l'analyse de la même quantité d'information. De ce fait, les temps de calcul nécessaires à la détection multicaractère multiQTL sont comparés

aux temps de calcul nécessaires pour analyser chacun des caractères avec les méthodes unicaractères multiQTL.

2.3.2. Seuils de rejet de l'hypothèse nulle et puissances

En raison de tailles de famille relativement faibles (25 descendants), les seuils ne peuvent pas être estimés par permutation. La distribution des statistiques de test des méthodes multiQTL a donc été estimée en extrapolant la méthode préconisée par Knott et Haley (2000 [53]) et validée par Korol *et al.* (1998 [58]). Ceux-ci proposent, pour chaque simulation réalisée sous l'hypothèse de deux QTL pour calculer la puissance, d'utiliser les estimations des paramètres (position et effets) estimés sous l'hypothèse nulle d'un seul QTL, pour calculer par simulations un seuil spécifique à cette simulation. Dans le cadre du croisement F2 qu'ils étudient, cette méthode est longue mais reste réaliste en terme de temps de calcul. Pour son application à des pedigree présentant des structures familiales, l'estimation d'un seuil spécifique à chaque simulation est inenvisageable. Nous avons choisi de généraliser cette méthode à l'utilisation des positions et des effets moyens estimés par le modèle uniQTL sur l'ensemble des simulations réalisées avec un modèle deux QTL pour estimer les puissances. Ces estimations sont utilisées comme paramètres de simulation sous l'hypothèse nulle d'un seul QTL. Les seuils sont estimés sur la base de 200 simulations. La distribution de la statistique de test ainsi obtenue est alors utilisée pour déterminer un seuil de signification à 5% qui est appliqué à l'ensemble des simulations réalisées avec le modèle deux QTL.

De la même façon que pour les analyses multicaractères, les seuils des analyses univariées sont corrigés pour le nombre de variables analysées par une correction de Bonferroni. Knott et Haley (2000, [53]) ont montré que, sous réserve que la corrélation résiduelle entre les caractères soit relativement faible, la correction de Bonferroni dans ce cadre permet d'obtenir les mêmes seuils que si l'on retenait le maximum des statistiques de test pour chacune des variables et de ne pas corriger pour le nombre de variables.

Pour le calcul des seuils, 200 simulations ont été réalisées, et 100 pour l'estimation des puissances. Ces valeurs sont très inférieures au nombre de simulations réalisées pour estimer les seuils et les puissances des méthodes multicaractères, en raison de temps de calcul beaucoup plus importants (voir paragraphe 3.1.), en particulier pour les méthodes multivariées.

2.3.3. Critère d'Information d'Akaike (AIC)

Nous avons testé à titre de comparaison le calcul des niveaux de signification par le Critère d'Information d'Akaike ([2], AIC). Ce critère d'information s'écrit de façon simple :

- pour une analyse uniQTL : $AIC_x = -2max(\Lambda_x) + 2npar$,
- pour une analyse multiQTL : $AIC_{\mathbf{x}} = -2max(\Lambda_{\mathbf{x}}) + 2npar$,

où $npar$ est le nombre de paramètres du modèle.

Parmi tous les modèles comparés, celui qui a le plus petit critère d'information a le pouvoir prédictif le plus élevé (Akaike, 1974 [2]; Sakamoto *et al.*, 1986 [95]). Nous avons calculé AIC pour chaque simulation réalisée pour le calcul des puissances. Le nombre de fois où chaque modèle est le plus pertinent sur la base du critère d'information d'Akaike est alors calculé.

2.3.4. Estimations

En ce qui concerne les positions des QTL, les méthodes multiQTL telles que nous les avons programmées permettent d'estimer le maximum de vraisemblance pour un couple de positions. Pour chaque simulation sous l'hypothèse de la présence de deux QTL, les deux positions estimées sont ordonnées, afin d'obtenir un vecteur de 200 positions par QTL. Pour chaque QTL, la précision d'estimation de la position est calculée, de la même façon que dans les analyses uniQTL, comme la somme des carrés des erreurs d'estimation (voir équation 2.10), la position simulée étant connue.

On remarquera que cette approche donne une borne supérieure de la variance d'estimation des positions, puisqu'une estimation précise de la position du premier QTL ne sera pas prise en compte comme telle si la position estimée du deuxième QTL est "avant" sur le groupe de liaison.

Pour ce qui est des effets des allèles au QTL, pour chaque caractère et chaque QTL, nous disposons d'un vecteur de 10x100 estimations des effets père pour les méthodes ST2, MV2 et MV2r. La précision d'estimation des effets paternels est calculée pour chaque caractère de la même façon que pour les analyses uniQTL comme la somme des carrés des erreurs d'estimation, l'effet simulé étant connu.

Pour DA2, les effets de chaque QTL sur chaque caractère ne sont pas estimables. Nous ne disposons que de l'estimation d'un effet du premier QTL et d'un effet du deuxième QTL sur la combinaison linéaire. Pour calculer les SCE des estimations, nous avons dû nous baser sur un effet de référence théorique ec_{rs} de chaque QTL r sur la combinaison linéaire pour chaque simulation s. Cet effet est calculé en fonction de a_{lr} l'effet simulé du QTL r sur le caractère l et γ_{ls} la pondération du caractère l pour la $s^{ième}$ simulation : $ec_{rs} = \sum_{l=1}^{p} \gamma_{ls} a_{lr}$. Un effet moyen de référence est alors calculable comme $ec_r = \frac{1}{S} \sum_{n=1}^{S} ec_{rs}$ pour chaque QTL. En notant $\hat{ac}_{rs}$ l'effet du QTL r estimé pour la combinaison linéaire, le calcul du SCE peut alors être réalisé comme :

$$\text{SCE} = \frac{1}{S}\sum_{s=1}^{S}(\hat{ac}_{rs} - ec_r)^2 \tag{3.10}$$

2.4. Dipositifs simulés

2.4.1. Pedigree et cartes génétiques

Différents types de simulations ont été réalisés en fonction des hypothèses testées. Le modèle de schéma expérimental est le même que pour les simulations uniQTL (3.4., chapitre 2). Etant donnés les temps de calcul des méthodes multiQTL, nous avons cependant restreint cette étude à un cas unique de pedigree, et à une densité unique en marqueurs génétiques. La population d'étude est issue d'un croisement F2 entre lignées fixées pour les allèles au QTL. Les individus F1 sont donc tous hétérozygotes au QTL. Dix mâles et 10 femelles sont simulés dans la génération F0, permettant de produire 10 mâles F1 croisés chacun avec 2 femelles F1. Chaque mâle F1 produit 50 descendants, permettant d'obtenir une taille totale de dispositif de 500 descendants. Le groupe de liaison mesure 1 Morgan, et contient 9 marqueurs génétiques régulièrement espacés, qui ont chacun 5 allèles (non fixés en F0) équiprobables (figure 18).

La simulation des pedigree et de la transmission des allèles marqueurs et quantitatifs au cours des générations est réalisée comme dans la partie multicaractère.

2.4.2. Performances

Pour le calcul des seuils, les performances sont simulées sous l'hypothèse d'un unique QTL, de la même façon que dans la partie multicaractère (3.4.).

Pour le calcul des puissances, deux QTL sont simulés, déterminant un ou plusieurs caractères. Le premier QTL est fixé à la position 31 cM (sauf dans un cas particulier), qui correspond à la position simulée pour les études uniQTL. Le deuxième QTL peut avoir 4 positions différentes (figure 18) : 81, 57, 43, ou 34 (auquel cas le premier est décalé à la position 29 cM) cM. Ces différents cas correspondent à des situations particulières commentées dans la littérature. Quand les QTL sont séparés par 50 cM, la ségrégation des allèles au QTL est indépendante. La position 57 cM du second QTL correspond à deux QTL séparés uniquement par un intervalle de deux marqueurs génétiques. La position 43 cM correspond à deux QTL dans deux intervalles adjacents, situation connue pour réduire fortement les capacités de détection des méthodes. Enfin, avec le couple de positions (29; 34) cM, on cherche à séparer deux QTL présents dans le même intervalle génétique, ce qui est *a priori* très difficile (Rodolphe et Lefort, 1993 [92]).

Les performances sont toujours simulées pour deux caractères, le(s) QTL ayant ou non un effet sur le deuxième caractère. Le modèle de simulation des performances est le suivant : $\mathbf{y} = \mu + \mathbf{Q}'\mathbf{a} + \mathbf{e}$

où $\mathbf{y} = \{y_l; l = 1, 2\}$, avec y_l la performance de l'individu pour le caractère l,

$\mu = \{\mu_l; l = 1, 2\}$, avec μ_l la moyenne de la partie polygénique des performances pour le caractère l, et μ tiré dans une loi multinormale (MVN) de moyenne nulle et de matrice de (co)variance V_{pol}, telle que $h_l^2 = 0,2$ et $\rho_{pol} = 0$,

$\mathbf{a} = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix}$, avec a_{lr} l'effet de substitution des allèles au $r_{ième}$ QTL pour le caractère l,

$\mathbf{Q} = [q_1\ q_2]$, avec q_r la variable indicatrice du génotype de l'individu au $r_{ième}$ QTL (q_r=1 si le génotype hérité au QTL est QQ; q_r=0 si le génotype hérité au QTL est Qq et q_r=-1 si le génotype hérité au QTL est qq),

$\mathbf{e} = \{e_l; l = 1, 2\}$, avec e_l l'erreur pour le caractère l, où $\mathbf{e}$ suit une MVN($\mathbf{0}$, $\mathbf{V}_{res}$), avec $\mathbf{V}_{res} = \begin{bmatrix} 1 & \rho_{res} \\ \rho_{res} & 1 \end{bmatrix}$, où $\rho_{res} = -0.4$.

Dans la pratique, la transmission de la partie polygénique est simulée sur les trois générations.

Différentes combinaisons des effets de chaque QTL sur chaque caractère ont été simulées. Elles sont résumées dans le tableau 3.2. L'ampleur de l'effet de substitution de référence a été choisie suite à quelques études préliminaires, pour lesquelles on peut montrer qu'un effet de $0{,}5\sigma_p$ de chaque QTL sur le caractère ne permet souvent pas de distinguer les deux QTL pour les tailles d'échantillon envisagées. Les cas 1 et 3 correspondent donc à des cas unicaractères, où le deuxième caractère n'est corrélé au premier que résiduellement. Pour le cas 1, les allèles aux QTL sont en phase, alors que pour le cas 3, ils sont en répulsion. Les cas 7 et 8 correspondent à deux QTL codant chacun pour un caractère différent. Les autres cas sont des combinaisons de ces situations sur les deux caractères.

3. Résultats et discussion

3.1. Temps de calcul

Le tableau 3.3 résume les moyennes des temps de calcul obtenus pour les 100+200 simulations réalisées avec les méthodes multiQTL multivariées. Elles ne concernent que l'analyse de 2 caractères. Les méthodes uniQTL sont reprises à titre de référence. Les valeurs des temps de calcul sont modifiées par rapport au chapitre sur les méthodes multicaractères uniQTL en raison de modification des options de compilation du logiciel

TAB. 3.2 - *Résumé des effets simulés pour chaque QTL sur chaque caractère lors des analyses multiQTL.* $a = 0,5$.

Cas	Caractère	QTL1	QTL2
1	1	$+a$	$+a$
	2	0	0
2	1	$+a$	$+a$
	2	$+a$	$+a$
3	1	$+a$	-a
	2	0	0
4	1	$+a$	-a
	2	$+a$	-a
5	1	$+a$	$+a$
	2	-a	-a
6	1	$+a$	-a
	2	-a	$+a$
7	1	$+a$	0
	2	0	$+a$
8	1	$+a$	0
	2	0	-a

lors du développement des méthodes multiQTL.

Quelle que soit la méthode utilisée, les temps de calcul sont mécaniquement augmentés par rapport aux méthodes uniQTL par l'analyse de tous les couples de positions. Les temps de calcul des analyses discriminantes uni et multiQTL sont ainsi linéairement liés par le nombre de tests réalisés, car le nombre de paramètres à estimer à chaque test demeure inchangé.

Pour les méthodes multivariées, l'augmentation du temps de calcul par rapport à MV est aussi liée à l'augmentation du nombre de paramètres à estimer. En considérant que pour toutes ces méthodes le nombre de paramètres dépend du nombre de caractères analysés, les méthodes multivariées sont donc difficilement utilisables en routine.

TAB. 3.3 - *Moyenne des temps de calcul nécessaires à l'analyse d'un jeu de données par les méthodes multiQTL.*

Méthode	ST	ST2	DA	DA2	MV	MV2	MV2r	PCA
CPU (secondes)	1,57	64,57	1,02	35,72	3,46	172,68	77,38	1,58

3.2. Positions et effets estimés sous l'hypothèse uniQTL

Les valeurs des paramètres estimées sous l'hypothèse uniQTL quand deux QTL sont simulés pour un caractère sont récapitulées dans le tableau 3.4. Il s'agit des moyennes des estimations obtenues pour le caractère 1 pour les cas correspondants lors des calculs de

puissance (soit 400 à 800 simulations). Trois types de cas peuvent être distingués :
1) Les deux QTL agissent sur le caractère avec des effets en phase (les allèles codant pour la même déviation du caractère sont portés par le même haplotype).
2) Les deux QTL agissent sur le caractère avec des effets en répulsion (les allèles codant pour des déviations du caractère en sens opposé sont portés par le même haplotype).
3) Un seul QTL agit sur le caractère (en l'occurrence QTL1).

Ces trois situations correspondent à des comportements différents de la méthode uni-QTL. Quand les QTL sont en phase, les effets s'additionnent sur la portion chromosomique. Le nombre de recombinaisons qui détériorent la liaison entre les allèles est proportionnel à la distance entre les deux QTL. Les effets estimés sont donc plus importants quand les QTL sont proches. A l'inverse, quand les QTL sont en répulsion, les effets des deux QTL, égaux en valeur absolue, ont tendance à s'annuler. Enfin, quand chaque QTL agit sur un caractère différent, les estimations des effets et des positions sont similaires à celles observées précédemment, et sont indépendantes de la position du deuxième QTL, qui agit sur l'autre caractère.

TAB. 3.4 - *Effets et positions estimés lors des analyses uniQTL en fonction des effets des deux QTL.*

		Moyenne des positions			Moyenne des effets		
a_1		0,50	0,50	0,50	0,50	0,50	0,50
a_2		0,50	-0,50	0,00	0,50	-0,50	0,00
Position	81	0,580	0,571	0,319	0,634	-0,025	0,467
QTL2	57	0,439	0,455	0,325	0,767	-0,003	0,460
(cM)	43	0,383	0,484	0,314	0,793	-0,010	0,472
	34	0,285	0,513	0,276	0,827	-0,010	0,471

a_r est l'effet du QTL r.

Les valeurs moyennes d'estimation des positions correspondent aux moyennes des positions simulées quand les allèles des deux QTL sont en phase. En effet, pour un deuxième QTL localisé à 0,81, 0,57, 0,43 et 0,34 cM, les positions moyennes estimées sont respectivement 0,56, 0,44, 0,37 et 0,315. Quand les allèles sont en opposition de phase, les estimations de position sont similaires à la moyenne des deux positions si les deux QTL sont relativement éloignés (de 50 et 26 cM). Pour les deux autres cas, l'estimation de la position se rapproche du milieu du chromosome. Cette tendance traduit le biais d'estimation de la position que nous avons mis en évidence précédemment dans les cas de puissances de détection faibles.

Les variations d'estimations des effets en fonction des positions et des effets des QTL permettent de mettre en évidence l'importance de la prise en compte des valeurs estimées pour l'estimation des seuils de signification du test d'hypothèse 2 QTL contre 1 QTL, en particulier quand les allèles au QTL ne sont pas en opposition de phase.

Les quelques résultats présentés dans ce tableau permettent aussi de confirmer les observations réalisées précédemment (Haley et Knott, 1992 [28]; Luo et Kearsey, 1992 [72]) : la détection potentielle de QTL fantômes entre les positions des deux QTL quand les allèles sont en phase, avec la surestimation corollaire des effets, l'absence de détection des QTL quand ils sont en répulsion et que les amplitudes des effets sont similaires. Les biais d'estimation, même par rapport aux paramètres théoriques moyens, peuvent ainsi être relativement élevés. Cependant, étant données les valeurs d'effets simulées pour chaque QTL, quand les allèles sont en phase les méthodes uniQTL permettent en général de détecter l'un des deux QTL à la position simulée. Les variances d'estimation des paramètres ont donc tendance à être relativement élevées pour les méthodes uniQTL.

3.3. Seuils de rejet de l'hypothèse nulle

Les seuils présentés ici concernent les méthodes multiQTL, soit ST2 pour la méthode unicaractère, et DA2, MV2 et MV2r pour les méthodes multicaractères. Les seuils ont été estimés indépendamment pour 6 (1, 2, 3, 5, 7, 8) des cas de relation entre les caractères décrits dans le tableau 3.2 et chacun des 4 couples de positions. Etant données les estimations des effets et des positions sous l'hypothèse uniQTL pour les cas 3, 4 et 6 (tableau 3.4), les estimations de seuils obtenues pour le cas 3 ont été utilisées pour les calculs de puissance des cas 4 et 6.

Les résultats sont synthétisés dans le tableau 3.5. Dans la pratique, les seuils sont relativement similaires d'un couple de positions simulé pour les QTL à l'autre. En effet, en ce qui concerne les cas où les allèles aux QTL sont en opposition de phase (cas 3, 4 et 6), ou les cas où chaque QTL agit sur un seul caractère (cas 7 et 8), les paramètres utilisés pour les calculs de seuil sont très similaires. Quand les allèles aux QTL sont en phase (cas 1, 2 et 5), les seuils ont tendance à augmenter avec le rapprochement des QTL, ce qui traduit l'utilisation d'effets plus élevés pour la simulation des QTL. L'augmentation représente 10 à 50% pour ST2, DA2 et MV2. Les augmentations les plus importantes sont observées pour DA2, entre les positions 81 et 43 cM pour le deuxième QTL.

3.4. Puissances

Les résultats sur les puissances sont synthétisés dans le tableau 3.6. Ils regroupent à la fois les puissances obtenues pour les méthodes uniQTL (test "il n'y a pas de QTL" contre "il y a un QTL") et ceux des méthodes multiQTL (test "il y a un QTL" contre "il y a deux QTL"). Dans les deux cas, les puissances des méthodes ont été évaluées à partir de simulations réalisées sous l'hypothèse 2 QTL.

TAB. 3.5 - *Seuils estimés pour les méthodes multiQTL.*

a_{11} a_{21}	a 0	a a	a 0	a -a	a 0	a 0	a 0	a a	a 0	a -a	a 0	a 0
a_{12} a_{22}	a 0	a a	-a 0	a -a	0 a	0 -a	a 0	a a	-a 0	a -a	0 a	0 -a
Position QTL2				81 cM						57 cM		
ST2	71,97	63,65	57,64	67,78	65,90	67,67	69,84	67,50	63,69	67,08	65,28	66,14
MV2r	24,24	8,48	19,80	5,81	15,84	7,76	26,31	11,31	16,21	5,74	14,86	11,97
MV2	111,31	109,02	99,11	109,53	109,88	109,95	110,44	112,68	99,32	111,78	105,61	107,12
DA2	91,71	123,40	103,62	98,46	101,36	85,66	102,05	133,31	111,74	109,28	94,55	85,42
Position QTL2				43 cM						34 cM		
ST2	65,26	65,86	60,35	63,71	65,58	65,58	75,87	70,10	57,88	71,40	68,82	67,40
MV2r	28,94	7,53	15,90	5,19	12,05	9,65	27,01	9,99	19,48	4,56	10,71	9,75
MV2	110,18	111,19	100,58	107,48	108,82	105,22	110,79	117,75	99,93	110,00	106,89	110,11
DA2	114,78	164,02	99,53	121,60	100,35	84,16	108,90	150,82	107,06	118,84	106,94	85,07

a_{lr} = effet de substitution du QTL r sur le caractère l, a=0,5σ_p. ST2 : méthode unicaractère; MV2r : méthode multivariée restreinte à la détection d'un effet par QTL; MV2 : méthode multivariée complète; DA2 : analyse discriminante.

3.4.1. Méthodes uniQTL

Conformément aux résultats établis précédemment par différents auteurs, on constate une différence très nette de comportement de ces méthodes en fonction de la phase des allèles aux QTL (Haley et Knott, 1992 [28]; Luo et Kearsey, 1992 [72]; Martinez et Curnow, 1992 [80]; Korol *et al.*, 1998 [58]). Quelle que soit la distance entre les deux QTL, quand les allèles sont en phase (cas 2, 5, 1) les méthodes uniQTL permettent de détecter un QTL dans 100% des cas, en surestimant les effets et en biaisant la position vers le milieu de l'intervalle entre les QTL (voir tableau 3.4).

Quand les allèles sont en opposition de phase (cas 4, 6, 3), les méthodes uniQTL permettent de détecter un QTL avec une puissance correcte uniquement quand les QTL sont éloignés de 50 cM. Les puissances chutent très rapidement ensuite, vers des valeurs de l'ordre de 15% quand les QTL sont à moins de 12 cM. Ces conditions permettent par ailleurs de dégager une hiérarchie entre les cas analysés : le QTL fantôme est mieux détecté quand les allèles d'un QTL agissent dans le même sens sur les deux caractères (cas 4) que quand ils sont en sens contraire (cas 6). Les puissances quand les QTL n'agissent que sur un des caractères sont inférieures (cas 3). Enfin, l'analyse discriminante permet dans tous les cas d'obtenir des puissances de détection supérieures à celles de l'analyse unicaractère, même quand les QTL n'agissent que sur un caractère.

Quand chaque QTL agit sur un seul caractère (cas 7 et 8), les puissances de détection des méthodes uniQTL sont toutes égales à 100%, sauf pour la méthode unicaractère (autour de 98%).

La hiérarchie entre les méthodes, que l'on ne peut mettre en évidence que dans les cas où les allèles sont en opposition de phase, correspond à celle évoquée au chapitre précédent. Les puissances sont supérieures pour les méthodes multicaractères quand les effets globaux des QTL sur les caractères sont tels que le produit des effets avec la corrélation résiduelle entre les caractères est négatif. De la même façon, les puissances sont améliorées si un caractère résiduellement corrélé au caractère d'intérêt est pris en compte dans l'analyse, même si ce nouveau caractère n'est pas déterminé par la région chromosomique analysée. La hiérarchie entre les méthodes est systématiquement retrouvée, sauf quand les deux QTL sont très proches, où l'analyse en composantes principales semble plus puissante. En contrepartie, cette méthode est la seule qui ne permette pas d'atteindre systématiquement des puissances de 100% dans les deux autres conditions.

3.4.2. Méthodes multiQTL

Dans la suite de la présentation des résultats des analyses multiQTL, nous reprendrons systématiquement la segmentation en trois conditions : allèles en phase (cas 2, 5, 1), allèles

TAB. 3.6 - *Puissances pour les méthodes multiQTL.*

a_{11} a_{21}	a 0	a a	a 0	a a	a -a	a -a	a 0	a 0	a 0	a a	a 0	a a	a -a	a -a	a 0	a 0
a_{12} a_{22}	a 0	a a	-a 0	-a -a	a -a	-a a	0 a	0 -a	a 0	a a	-a 0	-a -a	a -a	-a a	0 a	0 -a
Position QTL2				81 cM								57 cM				
ST	100	100	63	58	100	61	94	98	100	100	9	9	100	8	96	94
MV	100	100	56	98	100	71	100	98	100	100	16	44	100	16	100	99
DA	100	100	77	100	100	91	100	100	100	100	16	61	100	21	100	100
PCA	100	54	30	99	100	84	100	94	100	87	8	51	100	10	100	100
ST2	42	65	90	91	55	78	3	2	6	10	30	23	7	27	0	1
MV2r	1	75	3	27	5	0	96	100	3	11	9	19	1	5	67	81
MV2	50	100	86	100	84	90	64	71	8	33	48	97	11	59	19	38
DA2	62	99	20	100	84	37	5	16	11	47	5	28	26	4	10	10
Position QTL2				43 cM								34 cM				
ST	100	100	3	6	100	6	94	100	100	100	6	2	100	1	97	97
MV	100	100	8	13	100	8	100	99	100	100	5	8	100	5	100	100
DA	100	100	11	17	100	10	100	100	100	100	6	4	100	3	100	100
PCA	100	98	5	10	100	4	100	100	100	100	2	3	100	2	100	100
ST2	7	6	9	6	12	13	4	1	1	2	6	7	2	4	1	0
MV2r	0	11	6	5	2	5	35	51	4	0	3	0	6	1	2	0
MV2	3	17	15	39	18	16	11	17	3	8	9	5	7	4	7	4
DA2	9	4	8	5	25	9	3	12	21	19	4	5	26	4	6	9

a_{lr} = effet de substitution du QTL r sur le caractère l, a=0,5σ_p. Méthodes uniQTL : ST : méthode unicaractère; DA : analyse discriminante; PCA : analyse en composantes principales; MV : méthode multivariée. Méthodes multiQTL : ST2 : méthode unicaractère; MV2r : méthode multivariée restreinte à la détection d'un effet par QTL; MV2 : méthode multivariée complète; DA2 : analyse discriminante.

en opposition de phase (cas 4, 6, 3) et chaque QTL déterminant un caractère (cas 7, 8).

3.4.2.1. Quand les allèles sont en phase Les puissances de détection des méthodes multiQTL chutent très vite avec le rapprochement des deux QTL, passant de 99 à 100% de détection pour DA2 et MV2 quand ils sont séparés par 50 cM, à 47 et 33% respectivement quand ils ne sont plus séparés que par 26 cM pour le cas 2. Pour ces deux écarts entre les QTL, les méthodes multiQTL sont nettement plus puissantes pour le cas 2 que pour le cas 5, où elles sont plus puissantes que pour le cas 1. Cette hiérarchie correspond à celle mise en évidence au paragraphe précédent pour les méthodes uniQTL : la détection est plus puissante quand le produit des effets de chaque QTL sur chaque caractère avec la corrélation résiduelle est négatif, et elle est plus puissante si les deux caractères sont déterminés par les QTL.

Quand les QTL sont séparés par 12 cM, les puissances de détection sont inférieures à 20% pour MV2 et ST2, et elles deviennent inférieures à 10% quand aucun marqueur génétique ne sépare les deux QTL. Pour DA2, les puissances pour ces deux écarts restent supérieures, voir au-dessus de 20% quand les deux QTL sont très proches.

Dans l'ensemble de ces cas, l'analyse discriminante sur deux positions est la méthode la plus puissante, avec des puissances supérieures en moyenne de 40% relativement à celles de MV2. ST2 est moins puissante que ces deux méthodes, ce qui confirme l'avantage à prendre en compte l'ensemble des caractères corrélés dans l'analyse, même quand les QTL ne déterminent que l'un des caractères.

MV2r est comparativement beaucoup moins puissante en général, ce qui est attendu. En effet, cette méthode est basée sur un modèle de détermination des caractères totalement faux ici. Ce résultat est valable, pour les mêmes raisons, pour l'analyse des cas en opposition de phase.

Dans le cas d'allèles en phase, et pour les différents effets de QTL simulés, les méthodes uniQTL permettent toujours de détecter un QTL. L'utilisation des méthodes multiQTL permet, quand la détection est significative, de rejeter l'hypothèse d'un QTL pléiotrope au profit de deux QTL liés. Si les QTL sont suffisamment distants, l'hypothèse de QTL liés est acceptée dans plus de 60% des cas. Les méthodes multiQTL permettent donc ici d'affiner nettement les connaissances sur le déterminisme des caractères.

3.4.2.2. Quand les allèles sont en opposition de phase Les puissances de MV2 et ST2 sont nettement supérieures aux cas des allèles en phase, de l'ordre de 40% pour MV2 et 80% pour ST2. Cependant, MV2 reste la méthode la plus puissante. Cette différence par rapport aux cas des allèles en phase s'explique par des seuils de signification nettement moins élevés du fait d'effets simulés très faibles. Les moyennes des maxima des statistiques

de test sont néanmoins généralement plus faibles dans les cas d'allèles en répulsion par rapport aux cas d'allèles en phase.

DA2, au contraire, perd beaucoup en puissance comparé aux cas des allèles en phase, avec des puissances inférieures de 40%. Cette différence s'explique par les effets simulés pour chaque caractère, et les moyennes attendues des performances (tableau 3.7) pour chacun des 4 groupes haplotypiques utilisés pour réaliser la transformation par l'analyse discriminante. Quand les allèles sont en phase, les deux groupes d'haplotypes non-recombinants correspondent à des moyennes de performances attendues différentes, alors que les moyennes attendues sont identiques pour les deux groupes de recombinants. Quand les allèles sont en opposition de phase, les moyennes attendues sont identiques pour les groupes non recombinants et différentes pour les groupes recombinants. La capacité de détection des deux QTL est donc par essence fortement diminuée.

Les puissances de détection de MV2 restent par ailleurs très élevées, de 100 à 50%, tant que les QTL sont séparés de plus de 25 cM, et supérieures à 15% quand ils sont séparés par un marqueur génétique. En revanche, l'ensemble des méthodes présente des puissances inférieures à 10% quand les deux QTL sont dans le même intervalle.

De la même façon que précédemment, les cas où les produits des effets des QTL sur chaque caractère avec la corrélation résiduelle sont négatifs (cas 4) permettent une meilleure détection que quand les produits sont positifs (cas 6). De même, la détection de deux QTL n'agissant que sur un caractère est moins puissante que pour les cas précédents (cas 3). Elle est plus puissante avec des méthodes multicaractères qu'avec ST2 quand les caractères sont corrélés résiduellement.

Les cas d'allèles en opposition de phase correspondent à des puissances plus faibles avec les méthodes uniQTL qu'avec les méthodes multiQTL si les deux QTL sont séparés par au moins un marqueur génétique.

L'examen individuel des simulations donnant lieu à des statistiques de test significatives montre que les cas où les méthodes uniQTL sont significatives à 5% ne correspondent pas nécessairement à ceux où les méthodes multiQTL sont significatives à 5%. Les simulations coïncident quasiment complètement pour le cas 4 quand les QTL sont séparés par plus de 26 cM. Dans tous les autres cas, et pour les autres distances entre QTL, 5 à 10% des simulations significatives avec les méthodes uniQTL ne sont pas significatives avec les méthodes multiQTL. On ne peut donc pas envisager de réaliser les méthodes multiQTL uniquement sur les résultats significatifs pour les méthodes uniQTL sans accepter d'augmenter l'erreur de seconde espèce du test de mise en évidence de QTL liés.

3.4.2.3. Quand chaque QTL n'agit que sur un caractère La méthode qui permet d'obtenir la plus grande puissance de détection est MV2r. Les puissances décroissent de

TAB. 3.7 - *Effets des portions chromosomiques utilisables pour la transformation par DA2 dans les différents groupes d'haplotypes en fonction des cas.*

		a_{11}	a_{12}	a_{21}	a_{22}
Cas 1	Effets simulés	a	a	0	0
	Q1Q2	a	a	0	0
	Q1q2	a	-a	0	0
	q1Q2	-a	a	0	0
	q1q2	-a	-a	0	0
Cas 2	Effets simulés	a	a	a	a
	Q1Q2	a	a	a	a
	Q1q2	a	-a	a	-a
	q1Q2	-a	a	-a	a
	q1q2	-a	-a	-a	-a
Cas 3	Effets simulés	a	-a	0	0
	Q1Q2	a	-a	0	0
	Q1q2	a	a	0	0
	q1Q2	-a	-a	0	0
	q1q2	-a	a	0	0
Cas 4	Effets simulés	a	-a	a	-a
	Q1Q2	a	-a	a	-a
	Q1q2	a	a	a	a
	q1Q2	-a	-a	-a	-a
	q1q2	-a	a	-a	a
Cas 5	Effets simulés	a	a	-a	-a
	Q1Q2	a	a	-a	-a
	Q1q2	a	-a	-a	a
	q1Q2	-a	a	a	-a
	q1q2	-a	-a	a	a
Cas 6	Effets simulés	a	-a	-a	a
	Q1Q2	a	-a	-a	a
	Q1q2	a	a	a	a
	q1Q2	-a	-a	-a	-a
	q1q2	-a	a	a	-a
Cas 7	Effets simulés	a	0	0	a
	Q1Q2	a	0	0	a
	Q1q2	a	0	0	-a
	q1Q2	-a	0	0	a
	q1q2	-a	0	0	-a
Cas 8	Effets simulés	a	0	0	-a
	Q1Q2	a	0	0	-a
	Q1q2	a	0	0	a
	q1Q2	-a	0	0	-a
	q1q2	-a	0	0	a

a_{lr} = effet de substitution du QTL r sur le caractère l. Q_r et q_r indiquent les deux allèles au QTL r.

96 à 35% pour les cas 7 quand les QTL sont séparés respectivement de 50 et 12 cM. Ces puissances sont supérieures de 15% quand les effets des QTL sont opposés sur les deux caractères (cas 8). Cependant, quand les deux QTL sont dans le même intervalle, les puissances pour toutes les méthodes chutent au niveau de l'erreur de première espèce.

Les puissances de MV2 sont inférieures à celles de MV2r de 30 à 70% en fonction des situations et présentent le même type de profil en fonction des rapports des effets des QTL sur chaque caractère.

L'analyse discriminante est la méthode la moins puissante, avec 10% de détection en moyenne. Chaque caractère déterminant un QTL, la notion de combinaison linéaire des caractères qui caractérise le mieux les effets combinés des QTL devient en effet caduque.

Dans le cas spécifique de deux QTL agissant chacun sur un caractère, l'utilisation de méthodes multiQTL permet donc, si les deux QTL sont suffisamment éloignés, de rejeter l'hypothèse de pléiotropie dans plus de 65% des cas avec la méthode multivariée restreinte.

3.4.3. Discussion

Nous avons montré que la puissance de détection de la méthode unicaractère multiQTL est augmentée par les méthodes multicaractères par la prise en compte de caractères au moins résiduellement corrélés, que les allèles aux QTL soient en couplage ou en répulsion. Cette observation avait déjà été réalisée sur l'analyse de résultats de simulations de lignées fixées par Knott et Haley (2000 [53]) d'une part et Korol *et al.* (1998 [58]) d'autre part.

De plus, nous confirmons également que la distinction entre QTL liés et QTL pléiotrope est d'autant plus puissante que la distance entre les deux QTL est élevée et que les allèles aux QTL sont en répulsion. Cependant, nous avons pu montrer que cette dernière constatation n'est pas valable pour l'analyse discriminante quand les allèles sont en opposition de phase si les effets des deux QTL sur le caractère sont égaux en valeur absolue.

Enfin, Knott et Haley (2000) rappellent que la détection est d'autant plus précise et puissante que le modèle d'analyse correspond au modèle de simulation, ce qui est vérifié pour les cas que nous avons analysés, en particulier les cas 7 et 8 où l'analyse multivariée restreinte est beaucoup plus performante que les autres méthodes.

3.5. AIC

Etant donnés les temps de calcul requis par les analyses multiQTL, nous avons envisagé l'utilisation du critère d'information d'Akaike pour sélectionner les modèles les plus

pertinents. Pour chaque simulation réalisée pour l'estimation des puissances et chaque méthode, le critère d'Akaike a été calculé. Deux types de comparaison ont ensuite été réalisés : dans une première approche, les AIC de l'ensemble des méthodes uni et multi-QTL ont été comparés, et dans une deuxième approche les AIC des méthodes uniQTL et multiQTL ont été comparés séparément. Sur les 100 simulations réalisées pour estimer les puissances dans chaque cas, nous avons alors comptabilisé pour chaque méthode le nombre de fois où elle représentait le meilleur modèle sur la base de AIC.

Pour la première approche, dans la majorité des cas, la méthode offrant le modèle le plus pertinent est l'analyse en composantes principales. Dans le tableau 3.8, seuls les résultats atypiques par rapport à cette situation sont synthétisés. La variable retenue ici est uniquement celle permettant d'obtenir la statistique de test la plus significative dans chaque cas. Les quatre exceptions désignent dans la majorité des cas la méthode DA2 comme étant la plus pertinente. Il s'agit de cas de type 2, où les deux QTL sont en général bien différenciés par les méthodes multiQTL du fait d'effets élevés de chaque QTL sur chaque caractère.

Dans notre seconde approche, les AIC des méthodes uni et multiQTL sont considérés séparément. Pour l'ensemble des cas, l'analyse discriminante est la méthode qui offre le critère d'Akaike le plus faible parmi les méthodes multiQTL. La seconde méthode est toujours ST2, même quand chaque QTL ne détermine qu'un caractère, où cette méthode ne correspond pas du tout au modèle de simulation des performances. En ce qui concerne les méthodes uniQTL, dans les cas où PCA n'est pas systématiquement la méthode permettant d'obtenir le meilleur modèle, l'analyse discriminante correspond au modèle le plus pertinent.

La comparaison des résultats sur le critère d'information d'Akaike à ceux obtenus pour les puissances laisse perplexe. Les deux méthodes les plus fréquemment plébiscitées par AIC ne sont pas, de loin, celles qui permettent d'obtenir des puissances élevées dans la majorité des cas. Cependant, ce sont des méthodes basées sur des combinaisons linéaires des caractères, qui permettent donc de réduire le nombre de paramètres à estimer par rapport aux méthodes équivalentes multivariées. Mais l'élimination de ces méthodes pour la comparaison des critères d'Akaike des méthodes multiQTL, par exemple, entraîne le même type de résultat : la méthode retenue est celle qui a le moins de paramètres à estimer, et qui ne correspond que rarement au modèle complet le plus pertinent.

TAB. 3.8 - *Répartition des minima des AIC pour chaque série de 100 simulations.* Cas où PCA ne représente pas tous les minima lors de la comparaison de toutes les méthodes.

a_{11} a_{21}	a a	a a	a a	a a	a a
a_{12} a_{22}	a a	a a	-a -a	a a	a a
Position QTL2	81 cM	57 cM		43 cM	34 cM
Toutes méthodes					
ST	0	0	0	0	0
MV	0	0	0	0	0
DA	0	0	0	0	0
PCA	0	13	99	2	0
ST2	0	0	0	0	0
MV2r	0	0	0	0	0
MV2	0	0	0	0	0
DA2	100	87	1	98	100
UniQTL					
ST	0	0	0	0	0
MV	0	0	0	0	0
DA	90	63	0	91	90
PCA	10	37	100	9	10
sPCA	0	0	0	0	0
MultiQTL					
ST2	0	0	0	0	0
MV2r	0	0	0	0	0
MV2	0	0	0	0	0
DA2	100	100	100	100	100

a_{lr} = effet de substitution du QTL r sur le caractère l, a=0,5σ_p. Méthodes uniQTL : ST : méthode unicaractère; DA : analyse discriminante; PCA : analyse en composantes principales; MV : méthode multivariée. Méthodes multiQTL : ST2 : méthode unicaractère; MV2r : méthode multivariée restreinte à la détection d'un effet par QTL; MV2 : méthode multivariée complète; DA2 : analyse discriminante.

3.6. Qualité d'estimation

3.6.1. Estimation des positions des QTL

Les moyennes des estimations des positions sont représentées pour chaque méthode multiQTL figure 19. Les SCE des positions sont synthétisées dans le tableau 3.9. La hiérarchie entre les cas 1, 2 et 5 d'une part, et 3, 4 et 6 d'autre part décrite précédemment est en général retrouvée pour les SCE des estimations des positions.

Quand les QTL sont éloignés de plus de 25 cM, les positions sont bien estimées pour ST2 et DA2 (en dehors des cas 7 et 8 où le modèle ne correspond pas aux données), et MV2, que ce soit en terme de biais ou de variance d'estimation. Les différences sont très

TAB. 3.9 - *SCE des estimations des positions ($x\ 10^2$) pour chaque QTL par les méthodes multiQTL.*

$a_{11}\ a_{21}$	a 0	a a	a 0	a a	a -a	a -a	a 0	a 0	a 0	a a	a 0	a a	a -a	a -a	a 0	a 0
$a_{12}\ a_{22}$	a 0	a a	-a 0	-a -a	a -a	-a a	0 a	0 -a	a 0	a a	-a 0	-a -a	a -a	-a a	0 a	0 -a
Position QTL2				81 cM								57 cM				
$ST2_1$	1,33	0,70	0,82	0,72	1,03	1,06	2,15	2,25	1,97	1,56	1,35	1,79	2,20	1,65	2,22	1,73
$ST2_2$	1,08	0,51	0,95	1,19	1,14	0,90	6,45	6,87	3,42	2,93	1,76	2,39	3,09	1,55	7,58	5,77
$MV2r_1$	9,57	2,05	12,61	6,06	9,40	14,19	0,32	0,34	1,69	1,25	5,99	5,71	1,83	8,43	0,41	0,29
$MV2r_2$	4,85	1,66	8,98	6,37	9,35	14,49	0,45	0,23	5,85	1,37	8,76	6,53	1,64	9,45	0,42	0,29
$MV2_1$	1,38	0,22	1,16	0,25	0,51	0,69	0,71	0,64	2,18	1,08	2,16	0,32	1,97	1,71	1,61	1,44
$MV2_2$	1,85	0,80	1,68	0,74	1,31	1,21	1,70	1,65	3,84	1,85	3,06	0,69	3,27	2,20	3,56	2,55
$DA2_1$	1,26	0,09	1,16	0,12	0,41	0,61	2,84	8,97	2,63	1,80	2,56	0,25	2,35	3,15	2,80	4,72
$DA2_2$	0,83	0,12	0,84	0,10	0,40	0,91	3,39	8,19	5,01	2,41	2,11	0,22	4,87	3,42	5,14	6,54
Position QTL2				43 cM								34 cM				
$ST2_1$	1,83	1,98	3,82	5,16	1,80	3,81	2,22	2,03	1,13	1,38	7,06	8,38	1,48	5,83	1,44	1,61
$ST2_2$	6,21	6,41	9,48	9,46	5,01	6,36	5,38	6,74	8,47	8,99	17,23	17,91	8,59	17,61	7,31	5,44
$MV2r_1$	0,64	0,45	9,18	9,40	0,66	10,07	0,29	0,31	0,06	0,08	11,77	11,58	0,06	10,64	0,20	0,20
$MV2r_2$	11,22	0,18	12,71	12,20	0,33	12,75	1,02	0,35	13,17	0,24	18,37	19,77	0,29	19,91	0,51	0,48
$MV2_1$	1,24	1,48	4,49	2,56	1,79	3,28	1,72	1,71	1,42	1,20	6,50	6,53	1,18	6,97	1,54	1,75
$MV2_2$	6,98	4,29	8,60	4,87	5,36	5,89	5,23	6,52	9,22	6,97	16,56	18,69	10,87	19,00	9,73	11,55
$DA2_1$	1,94	1,58	5,78	1,67	2,41	5,03	2,67	4,00	1,23	0,78	8,38	7,12	0,71	7,43	1,22	2,31
$DA2_2$	12,57	9,16	10,49	3,41	9,78	9,84	9,93	9,43	18,79	18,81	16,90	17,10	18,85	17,50	17,23	15,44

a_{lr} = effet de substitution du QTL r sur le caractère l, a=0,5σ_p. $ST2_r$: méthode unicaractère (dans les cas 7 et 8, la position de référence considérée est la moyenne des positions des deux QTL), QTL r; $MV2r_r$: méthode multivariée restreinte à la détection d'un effet par QTL, QTL r; $MV2_r$: méthode multivariée complète, QTL r; $DA2_r$: analyse discriminante, QTL r.

faibles entre les cas d'allèles en phase ou en opposition de phase. Les SCE des positions sont inférieures à 0,03. Pour ces distances entre QTL, les SCE des estimations des positions des trois méthodes sont sensiblement équivalentes.

Quand les deux QTL se rapprochent, les différences entre les cas où les allèles sont en phase et en opposition apparaissent, et les estimations des positions des deux QTL ne se comportent pas de la même façon. Les SCE des estimations de la position du premier QTL quand les allèles sont en phase restent faibles pour DA2, MV2 et ST2. Cependant, un biais apparaît vers l'extrémité du chromosome la plus proche, de 5 à 10 cM. Pour les estimations de la position du deuxième QTL, les SCE augmentent rapidement, jusqu'à atteindre 0,10 pour MV2 et ST2, et 0,19 pour DA2. Les SCE augmentent donc plus pour DA2. Ces augmentations sont essentiellement dues à une augmentation du biais de l'estimation, qui tend à maintenir un écart entre les QTL détectés de 30 cM. Un phénomène similaire se produit quand les allèles sont en opposition de phase, mais dans des proportions équivalentes pour MV2, ST2 et DA2. De plus, les SCE des estimations de la position du premier QTL augmentent aussi avec la diminution de la distance entre les QTL, mais dans une moindre proportion.

La méthode multivariée restreinte (MV2r) ne permet d'obtenir des estimations précises des paramètres que quand elle est appliquée à des simulations d'un QTL déterminant chaque caractère. Dans les autres cas, cette méthode se comporte essentiellement comme une méthode uniQTL sur chaque caractère, confondant les positions des deux QTL et détectant des QTL fantômes. Les SCE sont alors en général inférieures à 0,005, et ne sont pas augmentées par le rapprochement des deux QTL. Cette méthode permet alors d'obtenir les estimations les plus précises. Dans les cas 7 et 8, MV2 permet d'obtenir des estimations précises des positions si les QTL sont très éloignés, alors que les estimations obtenues avec DA2 ont tendance a être plus biaisées. Quand les QTL se rapprochent, le comportement des estimations des positions par ces deux méthodes est à nouveau différent en fonction du QTL considéré, avec des estimations plus précises pour le premier QTL que pour le second.

Dans l'ensemble de ces cas, ST2 ne permet jamais d'obtenir les estimations les plus précises. Cette méthode est au mieux équivalente à MV2, même pour la détection de QTL n'agissant que sur un caractère.

3.6.2. Estimations des effets des QTL

Les valeurs absolues des moyennes des effets QTL intra-famille de père estimées sont représentées figure 20. Ces valeurs sont directement interprétables comme des effets estimés pour chaque caractère pour MV2 , ST2 et MV2r. En ce qui concerne DA2, les effets sont estimés pour la combinaison linéaire des caractères à la position. Un seul effet par

position est donc estimé et n'est pas directement comparable aux valeurs estimées pour les autres méthodes.

Les SCE des estimations des effets pères sont synthétisées dans le tableau 3.10. Les effets simulés pour les méthodes MV2, ST2 et MV2r sont des effets vrais, alors que pour DA2 il s'agit de SCE corrigées tel que décrit au paragraphe 2.3..

3.6.2.1. Quand les allèles sont en phase Si les deux QTL sont éloignés de plus de 25 cM, un léger biais apparaît dans l'estimation des effets, mais les estimations restent précises pour MV2 et ST2. Les SCE sont inférieures à 0,045 pour des écarts entre QTL de 50 cM et 0,08 pour 26 cM. Quand les QTL se rapprochent, les estimations des effets sur les deux QTL n'ont plus le même comportement. Le biais diminue pour les estimations des effets du premier QTL, conduisant à des estimations d'effets plus élevées, alors qu'il augmente pour les effets du deuxième QTL, réduisant les estimations à la moitié de la valeur espérée. Ce phénomène est à mettre en relation avec les biais d'estimation des positions. Le décalage vers l'extrémité du chromosome de l'estimation de la position du second QTL tend à diminuer l'effet qui lui est associé. La différence est alors attribuée au premier QTL. La somme des effets estimés pour les deux QTL est en effet globalement constante. Les SCE des estimations des effets QTL sont en revanche comparables pour les deux QTL. Elles sont inférieures à 0,04 quand 50 cM séparent les deux QTL, et augmentent jusqu'à 0,125 quand aucun marqueur génétique ne sépare les QTL. Cependant, ces remarques sont essentiellement valables pour l'estimation d'effets existants. Dans le cas 1, où les QTL agissent sur un seul caractère, les effets inexistants sont toujours estimés par MV2 comme quasi nuls, sans création de biais avec le rapprochement des QTL.

3.6.2.2. Quand les allèles sont en opposition de phase Les estimations des effets ont des comportements très différents. Quand les deux QTL sont éloignés de 50 cM, les estimations des effets sont moins biaisées que dans le cas précédent, mais légèrement plus variables, ce qui conduit à des SCE un peu supérieures. Quand les QTL se rapprochent, le biais, et par conséquent les SCE, ont tendance à augmenter, diminuant les valeurs d'effet estimées. Les différences entre les estimations des effets du premier et du second QTL sont moins marquées que dans le cas précédent mais se retrouvent pour une distance de 12 cM entre les QTL. En revanche, quand les deux QTL ne sont plus séparés par un marqueur génétique, les estimations de tous les effets sont quasiment nulles. Les méthodes multiQTL, en l'absence de marqueurs qui permettent de mettre en évidence les recombinaisons, se comportent alors comme des méthodes uniQTL. Les effets des deux QTL sur chaque caractère s'annulent et ne sont plus distinguables. Les SCE peuvent alors atteindre 0,4.

Dans ces deux cas, MV2r se comporte comme une méthode uniQTL pour l'estimation des effets, avec des effets très surestimés quand les allèles sont en phase, et sous-estimés

TAB. 3.10 - *SCE des estimations des effets pères pour chaque QTL par les méthodes multiQTL.*

a_{11} a_{21}	a 0	a a	a 0	a a	a -a	a -a	a 0	a 0	a 0	a a	a 0	a a	a -a	a -a	a 0	a 0
a_{12} a_{22}	a 0	a a	-a 0	-a -a	a -a	-a a	0 a	0 -a	a 0	a a	-a 0	-a -a	a -a	-a a	0 a	0 -a
Position QTL2				81								57				
$ST2_1^1$	3,76	3,07	3,95	3,56	3,14	7,17	13,78	17,64	8,33	6,28	27,78	16,71	14,18	13,81	12,15	21,83
$ST2_2^1$	4,40	3,13	3,85	4,31	3,13	7,08	14,39	17,13	8,81	6,66	27,84	15,97	8,71	14,16	11,48	20,49
$MV22_1^1$	3,41	2,75	6,16	2,77	3,11	3,25	3,52	3,31	8,11	5,17	34,70	14,28	8,37	13,73	9,34	12,22
$MV22_2^1$	3,39	2,77	6,30	2,97	2,97	3,25	3,71	3,37	8,09	5,35	33,45	13,88	7,39	14,25	8,84	13,28
$MV22_1^2$	4,48	2,70	5,09	3,02	2,95	3,35	3,25	3,27	7,80	5,71	22,82	11,15	6,68	18,00	9,99	13,80
$MV22_2^2$	4,13	2,83	5,22	3,19	3,10	3,24	3,58	3,41	7,66	5,63	22,60	13,08	6,39	19,11	9,95	12,77
$MV2r_1^1$	2,58	3,10	9,95	6,44	2,58	11,36	2,16	1,99	3,63	3,71	8,25	7,41	3,56	9,67	2,33	2,18
$MV2r_2^2$	2,98	3,17	3,13	17,84	2,66	11,97	8,03	7,59	3,02	3,85	2,77	13,20	3,57	10,45	7,77	7,55
$AD2_1$ cor	8,86	12,17	14,12	14,76	8,75	17,24	10,09	30,62	12,98	14,65	40,15	27,09	17,35	30,98	13,01	23,15
$AD2_2$ cor	9,25	12,34	14,89	14,79	9,75	17,48	12,20	29,33	11,58	15,51	39,11	27,33	12,15	31,03	12,83	24,84
Position QTL2				43								34				
$ST2_1^1$	10,32	14,45	23,31	19,05	10,19	27,58	12,55	14,04	14,73	12,58	25,08	31,38	14,25	20,47	15,55	9,39
$ST2_2^1$	10,98	14,29	23,58	19,02	11,50	28,25	12,37	14,47	16,88	13,07	24,99	33,15	14,35	24,79	15,08	9,56
$MV22_1^1$	11,21	10,14	22,89	28,88	9,69	39,13	13,14	14,39	12,44	13,34	27,95	19,16	10,46	21,60	11,46	9,86
$MV22_2^1$	11,48	9,81	24,00	27,33	10,83	37,33	13,69	14,98	13,65	14,09	27,46	20,74	12,19	22,17	10,93	9,11
$MV22_1^2$	11,44	13,30	14,39	36,41	9,42	38,98	11,85	16,42	12,19	14,80	17,92	19,65	10,38	22,30	14,95	11,61
$MV22_2^2$	10,76	12,24	16,15	36,08	10,50	35,84	11,72	16,86	11,93	15,53	17,55	20,15	11,98	23,69	14,71	12,04
$MV2r_1^1$	3,76	4,03	9,01	9,03	4,06	9,35	2,25	2,27	4,40	4,40	9,32	9,28	4,44	9,32	2,45	2,36
$MV2r_2^2$	2,90	4,30	3,12	10,88	3,85	9,83	7,56	7,95	3,06	4,39	2,85	9,12	4,38	9,08	7,92	7,87
$AD2_1$ cor	18,59	24,45	28,71	48,43	19,00	29,86	19,21	30,99	21,84	39,53	33,77	23,40	29,75	26,49	23,51	19,94
$AD2_2$ cor	10,02	14,35	28,35	47,93	12,70	30,26	17,95	28,35	10,45	10,73	39,93	24,03	11,54	25,63	16,05	14.45

a_{lr} = effet de substitution du QTL r sur le caractère l, a=0,5σ_p. $ST2_r^1$: méthode unicaractère, QTL r, caractère 1; $MV2r_r^l$: méthode multivariée restreinte à la détection d'un effet par QTL, QTL r, caractère l; $MV2_r^l$: méthode multivariée complète, QTL r, caractère l; $DA2_r$: analyse discriminante, QTL r.

quand ils sont en opposition.

3.6.2.3. Quand chaque QTL détermine un caractère Quelle que soit la distance entre les QTL, les biais et les SCE des estimations des positions obtenues avec MV2r sont faibles, respectivement inférieurs à 0,03 et 0,04. Pour MV2, les estimations présentent des profils équivalents à MV2r quand les QTL sont séparés de 50 cM, mais leur biais et leur variance augmentent quand ils se rapprochent.

Les effets estimés pour DA2 sont aussi récapitulés à titre d'illustration dans la figure 20 et le tableau 3.10. Ces valeurs correspondent aux estimations des effets pour les combinaisons linéaires des caractères. Leurs valeurs et leurs évolutions en fonction des situations ne sont donc pas directement comparables entre elles, ni à celles obtenues pour les autres méthodes.

3.6.3. Discussion

Les positions et les effets des QTL sont estimés avec précision quand les QTL sont séparés par plus de 25 cM et/ou un intervalle génétique. La bonne estimation des paramètres par les méthodes multivariées avait déjà été constatée par Knott et Haley (2000), Korol *et al.* (1998) et Nakamishi *et al.* (2001), avec une diminution des biais et des variances quand les QTL sont au moins séparés par un intervalle génétique.

Quand les QTL sont plus près, l'information sur le nombre de recombinaisons entre les QTL est restreinte à l'information disponible sur un marqueur génétique ou uniquement dépendante de la distance entre les positions testées quand les deux QTL sont dans le même intervalle. Dans ces derniers cas, les méthodes multiQTL permettent d'obtenir des statistiques de test plus significatives en créant un biais sur les positions des QTL et les effets, qui permet d'exploiter plus d'information sur les recombinaisons survenues entre les QTL. Cette stratégie, si elle permet d'augmenter une certaine part de l'information, est nécessairement réalisée au détriment de la précision d'estimation des paramètres, et entraîne une perte de puissance plus importante quand les deux QTL deviennent difficilement différentiables sur la base des recombinaisons.

On peut aisément supposer que la limite de séparation dépend à la fois de la densité en marqueurs génétiques informatifs entre les deux QTL, qui permet de mettre en évidence les recombinaisons, et de la distance entre les QTL, qui permet qu'un nombre exploitable de recombinaisons se produise au cours de l'obtention de la dernière génération d'individus.

Nous avons pu voir que le biais n'est pas symétrique sur le premier et le second QTL. Or les méthodologies de détection ne doivent *a priori* pas favoriser l'une ou l'autre position. Cependant, les QTL ont été simulés dans la première moitié du groupe de liaison. Le

décalage possible des positions est donc plus important pour le deuxième QTL que pour le premier. Nous avons vérifié cette hypothèse sur des exemples où les QTL sont simulés au milieu du groupe de liaison. Deux cas ont été simulés, qui correspondent aux cas 2 précédents, distants de 12 ou 5 cM. Dans la première configuration, les QTL sont aux positions 44 et 56 cM, séparés par un marqueur génétique à la position 50 cM. Dans la deuxième configuration, les QTL sont dans le même intervalle génétique, aux positions 43 et 48 cM. Les résultats des biais d'estimation des paramètres pour ces deux configurations et leurs équivalents non centrés sont récapitulés dans le tableau 3.11. Ils sont du même ordre de grandeur pour les deux QTL distants de 12 cM et centrés. Quand les QTL sont distants de 5 cM et légèrement décalés par rapport au milieu du groupe de liaison, ils sont inférieurs à la configuration pour laquelle les deux QTL sont dans le deuxième intervalle génétique, en particulier en ce qui concerne les positions. Ces observations n'avaient pas été réalisées par les auteurs précédemment cités. Cependant, les cas analysés par ces auteurs reposaient sur l'analyse de QTL centrés sur les groupes de liaison analysés, ce qui est attendu d'après les résultats du tableau 3.11, ou bien sur l'analyse de QTL éloignés sur le groupe de liaison (Nakamishi *et al.*, 2001). Quand ces derniers auteurs cherchent à détecter des QTL très proches, ils sont à peu près centrés sur le groupe de liaison.

4. Bilan

Nous avons comparé trois types de méthodes multiQTL. La première est multicaractère et basée sur l'analyse d'une combinaison linéaire spécifique à chaque couple de positions analysé (DA2). La seconde méthode est basée sur l'écriture complète de la vraisemblance unicaractère sur deux positions (ST2). Enfin, le troisième type de méthode réside dans l'explicitation complète de la vraisemblance multivariée pour l'analyse de deux caractères sur tous les couples de positions (MV2). Ce dernier type de méthode présente l'avantage d'être adaptable à tous les modèles que l'on cherche à tester. Cependant, le modèle complet entraîne l'estimation d'un très grand nombre de paramètres, ce qui le rend extrêmement lourd en terme de temps de calcul.

L'ensemble des résultats obtenus sur les simulations de pedigree de type animaux conduit à des conclusions similaires à celles obtenues pour des simulations comparables sur des croisements entre lignées *inbred*. De la même façon que Korol *et al.* (1998), Knott et Haley (2000) et Nakamishi (2001), nous avons pu montrer que la puissance et les estimations de paramètres sont d'autant plus élevées que les QTL sont éloignés. Un intervalle génétique est nécessaire entre les deux QTL pour bien identifier les recombinaisons entre les deux QTL, et obtenir des puissances élevées en même temps que des estimations des paramètres précises. Dans les autres cas, un biais apparaît sur les estimations des positions, et de façon corrélée sur les estimations des effets, pour maintenir, entre les po-

TAB. 3.11 - *Biais d'estimation des effets et des positions par les méthodes multiQTL en fonction de la localisation des QTL simulés.*

Ecart entre les QTL	12 cM		5 cM	
Positions	centrées	non centrées	centrées	non centrées
Positions (Biais x 10^2)				
$ST2_1$	9,10	3,75	12,70	7,05
$ST2_2$	7,35	14,00	11,05	17,50
$MV2r_1$	3,70	5,60	3,85	1,25
$MV2r_2$	3,55	2,50	8,00	4,65
$MV2_1$	5,40	2,75	11,30	6,15
$MV2_2$	6,80	10,20	8,60	14,20
$DA2_1$	12,80	2,20	13,55	2,85
$DA2_2$	12,75	20,35	19,85	35,50
Effets (Biais x 10^2)				
$ST2_1^1$	2,90	-1,57	0,64	1,82
$ST2_2^1$	4,47	-6,81	7,30	-10,42
$MV2_1^1$	3,04	-1,91	-1,81	4,00
$MV2_2^1$	4,35	-6,38	9,99	-12,69
$MV2_1^2$	3,90	-1,82	0,58	5,78
$MV2_2^2$	4,65	-6,60	8,56	-14,27
$MV2r_1^1$	-16,19	14,76	-16,61	15,91
$MV2r_2^2$	-15,44	15,53	-16,29	16,58

$ST2_r^1$: méthode unicaractère, QTL r, caractère 1; $MV2r_r^l$: méthode multivariée restreinte à la détection d'un effet par QTL, QTL r, caractère l; $MV2_r^l$: méthode multivariée complète, QTL r, caractère l; $DA2_r$: analyse discriminante, QTL r.

sitions donnant lieu aux statistiques de test les plus élevées, une quantité d'information exploitable pour la détection.

Par ailleurs, de même que ces auteurs, nous montrons ici que l'intégration à l'analyse de caractères corrélés permet d'augmenter les puissances et de préciser les estimations des paramètres. Ces améliorations sont d'autant plus flagrantes que le produit des effets des portions chromosomiques sur chaque caractère avec la corrélation entre les caractères est négatif. De la même façon que pour les méthodes uniQTL, on peut supposer que cette loi est généralisable à tous les couples de caractères dans le cas de l'analyse de plus de deux caractères.

Des analyses préliminaires avaient par ailleurs permis de mettre en évidence que seuls les QTL liés présentant des effets élevés sont discriminables. Dans ces conditions, quand les allèles aux QTL sont en phase, toutes les méthodes uniQTL permettent de conclure à la présence d'au moins un QTL. De la même façon, quand chaque QTL n'agit que sur un caractère, les détections uniQTL multicaractères permettent toutes de conclure à la présence d'au moins un locus pour la détermination des caractères. L'utilisation des

méthodes multiQTL permet alors de séparer deux QTL liés s'il sont suffisamment éloignés. En revanche, quand les allèles aux QTL sont en opposition de phase, les puissances de détection des méthodes uniQTL sont très dépendantes de la distance entre les QTL. Les QTL séparés par plus de 50 cM sont mis en évidence dans plus de 60% des cas. Dès que les QTL se rapprochent, la puissance diminue à moins de 50%. Les puissances de détection des méthodes multiQTL sont supérieures à ces valeurs tant qu'au moins un marqueur génétique sépare les deux QTL. Haley et Knott (2000) ont ainsi montré qu'en fonction du déterminisme réel des caractères analysés, le type de méthode à utiliser en priorité pour l'analyse d'un groupe de caractères doit être différent. Ils conseillent de baser le choix des tests sur les relations connues entre les caractères, en particulier l'ampleur de la corrélation génétique entre les caractères. Cependant, une absence de corrélation génétique peut traduire l'existence de QTL liés présentant des allèles en répulsion pour les deux caractères. Une telle stratégie limite donc les ambitions de la détection, en particulier pour les caractères pour lesquels on chercherait à casser une liaison génétique.

En ce qui concerne les méthodes utilisées, la méthode multivariée complète est particulièrement puissante et précise pour l'estimation des paramètres dans les situations les plus favorables. La méthode restreinte à la détection d'un QTL par caractère analysé n'est efficace que dans l'application à des simulations correspondant à ce modèle. Les estimations de paramètres sont alors particulièrement précises et peu biaisées, quelle que soit la distance entre les QTL. Enfin, la méthode basée sur l'analyse d'une combinaison de caractères est puissante si les effets des QTL ne s'annulent pas et si les QTL déterminent les deux caractères. En première approche, il semble donc que la détection de régions chromosomiques d'intérêt puisse être réalisée uniquement à l'aide de l'analyse discriminante et de l'analyse multivariée restreinte, pour éviter l'utilisation dans des procédures systématiques de méthodes requérant des temps de calcul très élevés. Les différents modèles issus de la méthode multivariée complète peuvent alors être réservés à l'analyse particulière ultérieure de ces portions chromosomiques.

Nous avons par ailleurs testé les procédures d'algorithme génétique telles que présentées par Nakamishi *et al.* (2001). Les procédures mises en oeuvre sont parfaitement efficaces pour la détection des régions impliquées dans la détermination des QTL et permettent, dans les conditions simulées, d'obtenir les mêmes résultats que les procédures de parcours systématique des couples de positions. Cependant, le nombre de tests à réaliser est augmenté du fait de la structure de l'algorithme génétique. Dans le cas de la détection de QTL liés sur un groupe de liaison de taille modérée, l'AG ne permet donc pas d'améliorer les temps de calcul requis. Ces résultats préliminaires permettent en revanche d'envisager des utilisations efficaces dans le cadre de la prise en compte de QTL sur des groupes de liaison différents dans les modèles, du fait de l'augmentation du nombre de tests à réaliser pour un balayage systématique du génome.

5. Application

Les analyses multiQTL des données porcines ont été réalisées dans la continuité des résultats obtenus avec les analyses multicaractères uniQTL. Nous avions pu alors supposer la présence de deux à trois QTL différents localisés dans la même région chromosomique, qui semblent affecter des groupes de caractères distincts. Un point particulier des analyses multiQTL consistera donc à vérifier l'hypothèse de deux QTL liés déterminant des groupes de caractères différents. Cependant, un ensemble d'analyses complémentaires a été réalisé qui seront aussi présentées. Un certain nombre de sous-modèles de la méthode multivariée complète ont été utilisés. L'écriture de la vraisemblance correspondante est facilement extrapolable à partir de celle de la méthode complète et ne sera pas explicitée dans cette partie.

Un ensemble de tests, à partir des modèles simples que nous avons comparés dans la partie précédente, a été réalisé sur les caractères, en exploitant l'information issue des analyses multicaractères précédemment réalisées. Chaque caractère a été analysé séparément, puis tous les caractères conjointement. Cette première étape a permis de valider la pertinence des analyses multiQTL pour ces groupes de caractères. Dans un deuxième temps, des modèles simples ont été appliqués séparément à chacun des groupes de deux caractères identifiés grâce aux analyses multicaractères. Enfin, nous avons testé l'hypothèse de deux QTL agissant chacun sur un groupe de caractères.

Pour la majorité de ces analyses, seules 200 simulations ont été réalisées pour l'estimation des seuils de rejet de H0 lors des tests multiQTL. Afin d'augmenter la précision d'estimation des quantiles empiriques, nous avons appliqué aux vecteurs des maxima de statistiques de test la méthode proposée par Harrel et Davis (1982), qui permet d'extrapoler la distribution de la statistique de test. Cependant, étant donné le faible nombre de simulations réalisées, nous ne prétendrons pas fournir les niveaux de signification spécifiques de chaque statistique de test calculée. Seuls les seuils à 5% seront donnés pour chaque méthode.

5.1. Analyses unicaractères

La méthode unicaractère multiQTL (2.2.1.) a été appliquée à chaque caractère exploité dans cette analyse. Les seuils ont été estimés sur la base des effets et des positions estimés lors des analyses uniQTL unicaractères (paragraphe 4.3., chapitre 1) pour chacun de ces caractères à l'aide de 1000 simulations. Une approximation de la correction de Bonferroni a permis de prendre en compte le nombre de variables analysées. Le seuil global à 5% pour l'ensemble de ces caractères est ainsi de 51 environ. Les effets estimés par la méthode uniQTL étant relativement proches en valeur absolue pour tous ces caractères, les valeurs

seuils ne diffèrent pas beaucoup. Les maxima des statistiques de test obtenus pour chaque caractère ainsi que les positions et les seuils correspondants sont résumés dans le tableau 3.12.

TAB. 3.12 - *Maxima des statistiques de test et positions correspondantes pour l'analyse multiQTL unicaractère.*

	Bardière	imf	Panne	x2	x4
maxLRT	41,71	36,14	54,65	48,56	43,00
QTL1 (cM)	69	48	1	69	56
QTL2 (cM)	134	62	67	147	65
Seuil 5%	51,70	51,63	49,96	51,70	51,70

Seule la statistique de test du poids de panne est significative pour un test à 5%. Le maximum de vraisemblance correspond aux positions 1 et 67 cM, avec des effets QTL moyens de -0,26 et -0,57 respectivement pour chaque position. L'évolution de cette statistique de test en fonction des couples de positions pour la région chromosomique d'intérêt est représentée en trois dimensions figure 21. Les résultats de l'analyse uniQTL unicaractère de panne localisaient un QTL à la position 41 cM avec un gros effet, avec un intervalle de confiance représentant la moitié du groupe de liaison. Il semble donc que le QTL mis en évidence lors de ces analyses corresponde à un QTL fantôme.

5.2. Analyses conjointes des 5 caractères

L'analyse discriminante a permis de réaliser une analyse multiQTL prenant en compte l'ensemble des caractères conjointement. Le maximum de la statistique de test vaut 89,43 pour le couple de positions 0 et 66 cM. Le seuil à 5% correspondant est de 98,93. Cette statistique de test n'est donc pas significative pour l'analyse de l'ensemble des caractères conjointement.

Pour chacun des deux groupes de deux caractères mis en évidence avec les analyses multicaractères, nous avons testé un ensemble de modèles imbriqués.

5.3. Analyses par groupe de deux caractères

5.3.1. imf+panne

Le premier groupe de caractères analysé regroupe la teneur en gras intramusculaire et le poids de panne. Nous avons d'abord testé l'hypothèse 1 QTL pléiotrope contre 1 QTL

agissant sur chaque caractère (test 1). Pour ce test, le maximum de la statistique de test est de 1,62, pour le couple de positions 41 et 61 cM. Le QTL en position 41 cM correspond au poids de panne et celui en position 61 cM au taux de gras intra-musculaire. Cependant, la statistique de test est très loin d'être significative, avec un seuil de signification à 5% de 13,12.

Nous avons alors testé l'hypothèse 1 QTL pléiotrope contre 2 QTL pléiotropes (test 2). Les maxima de statistique de test obtenus avec l'analyse discriminante ou la méthode multivariée complète sont significatifs à moins de 1%, avec des statistiques de test respectivement de 91,48 et 84,96, pour le même couple de positions 0 et 66 cM. Les seuils à 1% sont respectivement de 73,18 et 79,52. Les effets estimés respectivement pour le premier et le deuxième QTL, grâce à la méthode multivariée, sont alors de -0,14 et 0,78 pour imf et -0,23 et -0,58 pour panne. Pour l'analyse discriminante, la statistique de test présente une évolution en fonction des positions et un niveau de signification très similaires à ceux obtenus pour l'analyse conjointe des 5 caractères. Ces derniers résultats suggèrent très clairement la présence de deux QTL affectant les deux caractères.

La synthèse des résultats précédents, unicaractères et multicaractères, permet d'envisager la présence de deux QTL, l'un à l'extrémité proximale du chromosome, qui ne déterminerait que le poids de panne, et l'autre vers la position 67 cM qui déterminerait les deux caractères imf et panne. Cette hypothèse est soutenue par le fait que deux QTL sont mis en évidence avec les analyses unicaractères pour le poids de panne uniquement, et que les effets du premier QTL estimés pour imf par MV2 sont plus faibles. Nous avons donc testé :
1) l'hypothèse d'un QTL pléiotrope contre deux QTL dont un serait pléiotrope et l'autre n'agirait que sur le poids de panne (test 3),
2) l'hypothèse de deux QTL dont un serait pléiotrope et l'autre n'agirait que sur le poids de panne contre deux QTL pléiotropes (test 4).
Pour ce dernier test, l'hypothèse nulle n'est donc plus celle d'un QTL pléiotrope, et la simulation de données pour l'estimation des seuils a été adaptée en conséquence.

Pour le test 3, le maximum du rapport de vraisemblance est obtenu pour les positions 0 et 66 cM, avec une statistique de test de 55,98, significative à 1%. Les effets estimés sont respectivement de -0,22 et -0,59 pour le premier et le deuxième QTL pour le poids de panne, et 0,83 pour le deuxième QTL sur imf. La corrélation résiduelle entre les deux caractères est estimée à -0,19, alors qu'avec les analyses uniQTL elle était de -0,054. L'évolution de la statistique de test dans ces régions est représentée figure 22. Pour le test 4, le maximum est obtenu aux mêmes positions, mais n'est pas significatif à 5%. La valeur du maximum est de 28,73, pour un seuil à 5% de 42,66.

L'ensemble de ces tests permet d'accepter l'hypothèse de l'existence d'un QTL en

début de chromosome agissant sur le poids de panne, lié à un QTL, localisé vers la position 66 cM, qui agirait à la fois sur le poids de panne et sur la teneur en gras intramusculaire.

5.3.2. Analyses x2+x4

Le test d'un QTL agissant sur chaque caractère (test 1) sur les épaisseurs de lard dorsal permet d'obtenir un maximum de la statistique de test pour l'analyse d'une position unique, en 65 cM. La vraisemblance en cette position est donc égale pour les deux modèles comparés, ce qui conduit à un rapport de vraisemblance nul.

Le maximum de la statistique de test obtenu avec MV2 (test 2) est très significatif, avec une valeur de 75,37 pour les positions 18 et 68 cM pour une valeur de seuil à un pour mille de 63,72. Les effets estimés pour le premier et le deuxième QTL sont de -0,27 et -0,32 pour x2, et -0,28 et -0,36 pour x4. L'estimation de la corrélation résiduelle est de 0,85. L'évolution de la statistique de test dans ces régions est représentée figure 23. En ce qui concerne DA2, le maximum de la statistique de test est légèrement inférieur au seuil à 5%, et localisé aux positions 69 et 140 cM.

Nous avons par ailleurs réalisé pour ces caractères les tests 3 et 4 explicités au paragraphe précédent. Le test 3 permet d'obtenir un maximum de la statistique de test de 48,02 pour le couple de positions 17 et 68 cM, tout juste significatif à 5%. Les effets estimés sont alors de 0,04 pour le premier QTL et -0,49 pour le deuxième QTL sur x2, et de -0,52 pour le deuxième QTL sur x4. La corrélation résiduelle estimée est alors de 0,85. Le test 4 permet d'obtenir pour les positions 40 et 70 cM un maximum de la statistique de test de 27,23, non significatif.

L'interprétation des résultats multiQTL concernant les analyses des épaisseurs de lard dorsal est plus délicate que celle du groupe de caractères précédemment étudié. La seule hypothèse non rejetée est l'hypothèse la plus générale de deux QTL pléiotropes. La statistique de test est alors très significative et les estimations d'effets équilibrées entre les QTL et les caractères.

5.4. Analyse des deux groupes de deux caractères

Le dernier test réalisé est celui de l'existence d'un QTL pléiotrope sur les 4 caractères, contre l'hypothèse générale de deux QTL pléiotropes agissant chacun sur deux caractères différents, le poids de panne et la teneur en gras intra musculaire d'une part, et les épaisseurs de lard dorsal d'autre part. Le maximum de la statistique de test obtenu est à peine supérieur à 0, pour des positions très proches. Ce résultat n'est donc pas significatif. Cependant les estimations des positions obtenues lors de la détection séparée de deux QTL

pléiotropes (voir paragraphe 5., chapitre 2) sont de 65 et 68 cM pour ces deux groupes de caractères. On peut donc supposer que si deux QTL sont réellement en ségrégation dans cette région, leur distinction par les méthodes utilisées ici est très peu probable.

5.5. Conclusion

En conclusion, nous avons pu mettre en évidence la présence d'au moins deux QTL agissant sur des caractéristiques de gras interne sur le chromosome 7 porcin. En parallèle, les résultats obtenus suggèrent fortement la présence d'un ou deux QTL liés aux précédents agissant sur des caractéristiques de gras externe. La figure 24 résume les positions des QTL ainsi détectés sur les différents caractères.

Comme nous l'avons vu au paragraphe 5., chapitre 2, la distinction de ces deux types de gras est importante pour la production. L'identification, grâce aux données moléculaires des individus porteurs d'haplotypes favorables pour l'un ou l'autre type de gras, est stratégiquement cruciale, et peut désormais être envisagée. Pour cartographier plus précisément ces QTL, de nouvelles stratégies doivent être abordées. La meilleure définition des régions chromosomiques à explorer peut permettre d'envisager une densification en marqueurs génétiques ciblée sur ces régions afin de déterminer des haplotypes communs aux individus hétérozygotes. A terme, le clonage fonctionnel et/ou positionnel de ces QTL pourra alors être entrepris pour des portions chromosomiques de taille abordable.

Chapitre 4

Bilan et Perspectives

1. Introduction

L'objet de ce travail de thèse était de développer des méthodologies de détection de QTL multidimensionnelles adaptées à l'analyse de données d'espèces animales d'intérêt agronomique. Contrairement aux populations disponibles pour les espèces végétales ou chez les animaux de laboratoires, la majorité des populations de production animale ne permet pas de travailler sur des lignées fixées ou considérées comme telles. De ce fait, les données doivent être analysées intra-familles de plein-frères et/ou de demi-frères. L'extrapolation des méthodologies applicables aux croisements entre lignées fixées est alors techniquement possible, quoique très exigeante en terme de temps de calcul, en raison de l'augmentation importante du nombre de paramètres à estimer.

A l'origine de ce travail, une méthode multicaractère applicable aux populations animales existait dans la littérature (Weller *et al.*, 1996), et aucune méthode multiQTL. Or de nombreux auteurs avaient préalablement souligné les avantages de la prise en compte, d'une part de caractères corrélés (Korol *et al.*, 1995; Jiang et Zeng, 1995; Mangin *et al.*, 1998), et d'autre part de positions liées (Haley et Knott, 1992; Martinez et Curnow, 1992), pour affiner, voire corriger, les résultats de détections réalisées selon le modèle de base unicaractère uniQTL. Le développement et la caractérisation de méthodologies spécifiques aux populations animales devenait de ce fait nécessaire.

Nous avons choisi de restreindre cette étude à une question particulière : la discrimination entre un QTL pléiotrope et deux QTL liés. Hors l'exigence de cohérence du travail réalisé, cette restriction correspond à un objectif pratique : la caractérisation d'un segment chromosomique que l'on souhaiterait utiliser pour la sélection assistée par marqueurs.

2. Bilan général et synthèse

Deux types de méthodologies ont été comparées : les méthodes, en général multivariées, existantes pour l'analyse de données issues de lignées génétiquement fixées, et des méthodes fondées sur l'analyse de variables synthétisant sous forme de combinaisons linéaires l'information relative aux relations entre les caractères.

Comme attendu, les méthodes de type multivariées se sont avérées particulièrement exigeantes en terme de temps de calcul. Or les distributions des statistiques de test sont généralement inconnues, ce qui implique l'analyse répétée de données simulées ou permutées pour leurs caractérisations empiriques. Les temps de calcul sont donc un élément crucial de l'applicabilité en routine des méthodes multidimentionnelles. Cependant, pour des performances de détection équivalentes ou supérieures, les méthodes de détection basées sur l'analyse de variables synthétiques développées pour cette étude permettent de ramener les temps de calcul à des niveaux abordables en routine pour la sélection de régions chromosomiques d'intérêt.

Pour l'ensemble de ces méthodes, nous avons caractérisé et quantifié les avantages de l'intégration d'informations supplémentaires pour la détection de QTL. L'utilisation de méthodes multicaractères permet d'augmenter à la fois la puissance de détection et la précision de l'estimation des paramètres par rapport aux méthodes unicaractères dans certaines conditions. Ces conditions sont similaires à celles mises en évidence par les auteurs ayant exploré les méthodes applicables aux données issues de lignées génétiquement fixées (Korol *et al.*, 1995; Jiang et Zeng, 1995; Mangin *et al.*, 1998). Les gains sont d'autant plus importants qu'il existe parmi les caractères analysés des couples de caractères pour lesquels le produit des effets du QTL (ou de la portion chromosomique considérée) avec leur corrélation résiduelle est négatif. Ce critère est cependant difficile à manipuler pour sélectionner les caractères à analyser conjointement, puisque les paramètres cités ne sont pas connus avant analyse.

De façon similaire, nous avons pu confirmer la possibilité de séparer deux QTL liés si ceux-ci sont séparés par un nombre suffisant de marqueurs génétiques informatifs. Les méthodes multiQTL ne permettent de détecter que des QTL présentant des effets élevés sur les caractères (Knott et Haley, 2000; Korol *et al.*, 1998). Cependant, ces dernières méthodologies sont particulièrement cruciales, en particulier lorsque deux QTL liés présentent des allèles en phase pour un caractère (ou un groupe de caractères). Ils entraînent alors la détection d'un QTL "fantôme" entre ceux-ci par les méthodes uniQTL. Un tel QTL peut sembler déterminant pour la caractérisation du caractère (ou du groupe de caractères) en question, puisque les effets estimés sont une combinaison des effets de chacun des QTL. Or la cartographie fine de la région chromosomique sur la base de tels résultats est hasardeuse, puisque fondée sur des estimations extrémement biaisées des paramètres.

Il semble donc sage de préconiser le test de 2 QTL liés de façon systématique avant de poursuivre des investigations coûteuses.

Un avantage non négligeable des méthodologies développées pour cette étude réside dans la possibilité de détecter des QTL non détectés par les méthodes les plus simples. Ces QTL présentent nécessairement en moyenne des effets moins forts que des QTL qui apparaîtraient dès les détections uniQTL/unicaractères. Cependant, il existe de nombreux types de caractères pour lesquels les analyses simples ne permettent pas de détecter beaucoup de QTL. Il s'agit par exemple de caractères ayant des héritabilités faibles, tels que la majorité des caractères de reproduction femelle, ou de caractères difficiles à mesurer, pour lesquels peu de données sont disponibles, nécessitant l'abattage des animaux ou des mesures coûteuses, telles que la mesure de la teneur en lipides intra-musculaires. Pour ces caractères, la détection de QTL présentant des effets plus faibles peut être cruciale. En effet, leur sélection par les méthodes quantitatives classiques est plus longue et difficile que pour les caractères bien héritables et/ou facilement mesurables. La sélection assistée par marqueurs d'allèles améliorant ce type de caractères présente alors un attrait majeur pour l'application à la production.

3. Propositions de stratégie

La synthèse des résultats présentés permet d'envisager des stratégies d'analyses systématiques qui exploitent les avantages combinés des différentes méthodes utilisées. Différents objectifs peuvent être envisagés :

1) la caractérisation d'une région particulière (pour un ou plusieurs caractères) pour laquelle la présence d'un ou plusieurs QTL éventuellement pléiotrope(s) est suspectée,

2) la caractérisation d'une relation particulière entre des caractères, pour lesquels la présence d'un ou plusieurs QTL pléiotrope(s) est suspectée,

3) la recherche de QTL pour un groupe de caractères donné pour lequel les analyses unicaractères uniQTL donnent peu de résultats, en raison par exemple d'héritabilités faibles.

Nous allons proposer dans la suite de cette partie des méthodes combinant la sélection de régions chromosomiques à explorer et de groupes de caractères à analyser.

3.1. Détection pour un caractère

La détection de QTL liés déterminant un unique caractère est simple à mettre en oeuvre. Dans un premier temps, pour toutes les régions chromosomiques d'intérêt, les

méthodes unicaractères uniQTL et multiQTL sont appliquées. Deux séries de simulations sont alors nécessaires pour estimer les niveaux de signification correspondant aux maxima des statistiques de test obtenus. La première série de simulations est réalisée avec des performances qui ne sont pas déterminées par un QTL. Elle permet d'obtenir les niveaux de signification pour la méthode uniQTL. Une deuxième série de simulations est nécessaire, où les estimations des effets et des positions pour chaque groupe de liaison par la méthode uniQTL sont utilisées pour simuler un QTL déterminant le caractère, afin d'estimer les niveaux de signification de la méthode multiQTL.

3.2. Détection pour un groupe de caractères

Comme nous avons pu le voir, les méthodes multicaractères sont exigeantes en terme de temps de calcul. La première étape de ce type d'analyse est donc la sélection rigoureuse d'un groupe de caractères. Ces caractères peuvent appartenir à une famille commune, tels que les caractères de composition corporelle que nous avons analysés pour cette étude, ou présenter des relations antagonistes à caractériser, tels que des caractères d'engraissement et de croissance. Un autre critère de sélection peut résider dans la mise en évidence préalable pour certains caractères de régions chromosomiques similaires portant des QTL. Le nombre de caractères intégrés à l'analyse doit cependant rester raisonnable pour permettre une estimation correcte des paramètres. La limite dépend naturellement de la quantité d'information disponible dans le dispositif (nombre d'individus informatifs, densité et qualité des marqueurs génétiques), mais une borne maximum de 10 caractères semble raisonnable.

Nous avons représenté dans la figure 25 une méthode itérative qui permet de sélectionner les régions chromosomiques à analyser de façon détaillée. Cette méthode repose sur une sélection de type *backward* des caractères informatifs pour chaque région sur la base de leur pondération dans la combinaison linéaire des caractères, telle que nous l'avons réalisée lors de l'analyse des données d'engraissement. Cette stratégie est généralisée à l'utilisation des méthodes multiQTL. Nous avons privilégié l'utilisation de l'analyse discriminante dans cette étape de sélection, afin de limiter les temps de calcul. La méthode multivariée uniQTL est réservée à l'estimation des paramètres nécessaires à la simulation d'un QTL pléiotrope pour le calcul des niveaux de signification. Quand seuls deux caractères subsistent pour un groupe de liaison donné, il est possible d'ajouter la méthode multivariée restreinte à la détection d'un QTL par caractère. Pour cette étape de sélection des régions chromosomiques et des caractères, les niveaux de signification à utiliser peuvent être relativement laxistes. En effet, il s'agit ici d'une étape de "débroussaillage" des données, qui va conduire à la sélection de quelques régions d'intérêt analysées complètement par la suite.

A l'issue de ces sélections, seules les régions chromosomiques présentant des statistiques de test significatives donnent lieu à des comparaisons de modèles plus fins. Pour cette dernière étape, la sélection de plus de trois caractères entraîne la comparaison potentielle d'un nombre important de sous-modèles. Plutôt que de réaliser des comparaisons systématiques, il semble raisonnable à cette étape de sélectionner quelques sous-modèles plausibles sur la base des analyses globales disponibles (analyses unicaractères et multicaractères multiQTL) et de ne tester en priorité que ceux-ci. Les niveaux de signification à considérer ici devront être adaptés à l'utilisation potentielle des résultats.

La figure 26 résume la stratégie à adopter pour la sélection d'un groupe de caractères pour la détection de QTL pléiotropes. Il convient cependant de garder à l'esprit que les tests réalisés à la dernière étape ne permettent de valider la présence d'un QTL pléiotrope que s'ils sont associés à des tests les comparant à des modèles de QTL liés.

Quelle que soit la stratégie adoptée, la dernière étape doit consister à tester la signification des effets estimés pour chaque caractère. Certains caractères peuvent en effet n'apporter d'information que du fait de leur corrélation résiduelle avec les caractères déterminés par le(s) QTL détecté(s). Dans la pratique, cette étape, quoique longue, ne diffère pas de celle réalisée pour le test des effets détectés par les méthodes uniQTL unicaractères. Korol *et al.* (2001) proposent par exemple une série de tests de permutation en ce sens.

3.3. Niveaux de signification des statistiques de test

Dans tous les cas, les niveaux de signification des statistiques de test doivent être estimés en tenant compte du nombre de tests réalisés. Au moins 1000 simulations sont recommandées pour l'obtention de valeurs précises. L'utilisation de logiciels permettant d'extrapoler la distribution de la statistique et les valeurs de quantiles peut améliorer ces estimations (par exemple d'après la méthode proposée par Harrell et Davis en 1982). Nous ne recommanderons pas ici l'utilisation de critères tels que le Critère d'Information d'Akaike pour le choix du modèle le plus pertinent, étant donnés la faible pertinence des résultats que nous avons obtenu avec cette stratégie lors de notre étude. Cependant, d'autres critères d'information plus adaptés méritent peut être une attention particulière.

La sélection du modèle le plus pertinent pour chaque groupe de liaison requiert le choix d'un seuil de signification minimum au niveau du génome pour les statistiques de test. Pour la sélection itérative de caractères et/ou de régions chromosomiques, les seuils de signification tolérés doivent être suffisament laxiste pour permettre de ne pas rejeter systématiquement les statistiques de test, en tolérant la détection de faux positifs. Un seuil de 5% au niveau du génome semble alors raisonnable.

Pour une application directe vers la cartographie fine de QTL ou la sélection assistée

par marqueur, cette valeur seuil doit permettre de sélectionner le meilleur modèle avec le moins de possibilités de se tromper (faible erreur de première espèce) mais sans rejeter systématiquement le modèle le plus complet (faible erreur de deuxième espèce). Ces deux critères sont antagonistes, et peuvent conduire au choix de seuils supérieurs à 1% ou 0,1% au niveau du génome. Cependant, ils assurent l'application de techniques lourdes et coûteuses aux quelques QTL ayant les plus gros effets sur les caractères, seuls QTL dont la caractérisation et l'utilisation peuvent être stratégiques.

4. Perspectives

Les méthodologies caractérisées pour ce travail ont permis de valider l'utilisation de modèles plus complets que les modèles classiques pour l'application aux pedigree présentant des structures familiales. Nous avons ici négligé la question de la gestion des données manquantes. Aucune étude n'a été menée jusqu'à présent pour quantifier son impact sur la qualité de détection des méthodes multicaractères. En ce qui concerne les méthodes basées sur des transformations de variables, les individus présentant un phénotype manquant pour l'un au moins des caractères semblent *a priori* devoir être exclus de l'analyse, afin de ne pas créer de biais vers les caractères les plus représentés. D'une façon générale, il semble en première approche prudent de généraliser cette attitude à toutes les méthodologies multicaractères.

Nous avons par ailleurs limité le choix des méthodes fondées sur l'analyse de variables synthétiques à une unique stratégie dérivée de l'analyse discriminante. Une autre stratégie pourrait consister à réaliser une transformation en composantes principales des données spécifique de chaque position testée, à partir de la matrice de variance covariance résiduelle des données estimée d'après une définition de groupes haplotypiques semblables à ceux de l'analyse discriminante. Cette stratégie, contrairement à l'ACP utilisée dans ce travail, permettrait de réaliser une transformation spécifique du ou des QTL présents. Une seule variable pourrait être analysée en routine, choisie selon l'objectif classique de minimisation de la variance résiduelle du modèle. En contrepartie, l'analyse, univariée, de toutes les variables permettrait d'effectuer une transformation *reverse* des effets synthétiques estimés pour s'affranchir totalement des méthodes multivariées.

Une extension possible, et aisément réalisable, des méthodes présentées ici réside dans la prise en compte de positions non liées dans le dispositif. Pour limiter le nombre de tests à réaliser, les algorithmes génétiques peuvent être utilisés comme algorithme de recherche du maximum de vraisemblance sur un ensemble de groupes de liaison (Nakamishi *et al.*, 2000; Carlborg *et al.*, 2001). Les paramètres de l'exploration par l'algorithme génétique (taille de la population, critère d'arrêt de la recherche, défintion de sous-populations...)

restent néanmoins à caractériser. Des méthodes itératives de type *stepwise*, qui permettent de sélectionner les régions chromosomiques à intégrer dans des analyses à cofacteurs, peuvent par ailleurs être envisagées (de Koning *et al.*, 2001; Kao *et al.*, 1999). Pour ce type de stratégies, l'utilisation de l'analyse discriminante doit être modifiée. La division des descendants en groupes haplotypiques ne peut en effet pas être réalisée pour plus de trois positions considérées simultanément, afin de conserver une estimation précise des matrices de variance covariance. Dans le même esprit que Churchill et Doerge (1996) pour les calculs de seuils par permutation, les résidus des transformations précédentes peuvent par exemple être utilisés à chaque étape pendant la phase de sélection des régions chromosomiques.

Enfin, nombre de stratégies développées pour la détection de QTL non liés considèrent l'éventualité de la prise en compte d'interactions de type épistastiques entre locus (Kao *et al.*, 1999; Knott *et al.*, 1998). De telles stratégies sont techniquement extrapolables aux méthodes multivariées explicitées dans ce document. Cependant, de la même façon que pour les stratégies évoquées au paragraphe précédent, le nombre de paramètres à estimer augmente alors encore. Seuls les plus gros effets pourront à nouveau être discriminés. Il convient donc, à ce stade de complication des modèles, et étant donnée la taille généralement modeste des dispositifs animaux, de rester réaliste quant aux niveaux de complexité que l'on souhaite expliciter et aux moyens mis en oeuvre pour y parvenir. Des approches complémentaires, et peut-être plus judicieuses à ce stade d'affinage de la cartographie, peuvent alors être envisagées, telles que les comparaisons d'haplotypes (Riquet *et al.*, 1999), ou le clonage positionnel ou fonctionnel de gènes.

Bibliographie

[1] Aitkin, M., Anderson, D., Francis, B. et Hinde, J., 1989. *Statistical modelling in GLM.* Oxford University Press, Oxford.

[2] Akaike H., 1974. *A new look at the statistical model identification.* IEEE Trans. Auto. Control 19, p 716-723.

[3] Andersson, L., Haley, C.S., Ellegren, H., Knott, S.A., Johansson, M., Andersson, K., Andersson-Eklund, L., Edfors-Lilja, I., Fredholm, M., Hansson, I., Hakansso, J. et Lundstrom, K., 1994. Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Science*, 263, p 1771-1774.

[4] Barillet, F., Boichard, D., Astruc, J.M. et Bonaïti, B., 1996. Validation of estimated genetic trend in French Lacaune dairy sheep evaluation. 30ème session d'ICAR, Veldhoven, Hollande, 23-28 June 1996, EAAP Publication 87, p 291-298.

[5] Bidanel, J.P., Milan, D., Iannuccelli, N., Amigues, Y., Boscher, M.Y., Bourgeois, F., Caritez, J.C., Gruand, J., Le Roy, P., Lagant, H., Bonneau, M., Lefaucjeur, L., Mourot, J., Prunier, A., Désautés, C., Mormède, P., Renard, C., Vaiman, M., Robic, A., Gellin, J., Ollivier, L. et Chevalet, C., 2000, Détection de locus à effets quantitatifs dans le croisement entre races Large White et Meishan : Résultats et perspectives. *32ème Journées de la Recherche Porcine en France*, 32, p 369-383.

[6] 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. et Chevalet C., 2001. Detection of quantitative trait loci for growth and fatness in pigs. *Genet Sel Evol*, 33, p 289-309.

[7] Bidanel, J.P. et Rotschild, M., 2002a. Current status of quantitative trait locus mapping in pigs. *Pig News & Information*, 23, p 39N-53N.

[8] Bidanel J.P., Milan D., Renard C., Gruand J., Mourot J., 2002b. Detection of quantitative trait loci for intramuscular fat content and lipogenic enzyme activities in Meishan x Large White F2 pigs, in: Proceedings of the 7th World Congress Applied to Livestock Production. CD ROM communication n03-13.

[9] Bost, B., de Vienne, D., Hospital, F., Moreau, L. et Dillmann, C., 2001. Genetic and nongenetic bases for the L-shaped distribution of quantitative trait loci effects. *Genetics*, 157, p 1773-1787.

[10] Carlborg, O., Andersson, L. et Kinghorn, B., 2000. The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. *Genetics*, 155, p 2003-2010.

[11] Churchill, G.A. et Doerge, R.W., 1994. Empirical threshold values for quantitative trait mapping. *Genetics*, 138, p 963-971.

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

[13] Davies, R.B., 1977. Hypothesis testing when nuisance parameter is present only under the alternative hypothese. *Biometrika*, 64, p 247-254.

[14] Davies, R.B., 1987. Hypothesis testing when nuisance parameter is present only under the alternative hypothese. *Biometrika*, 74, p 33-43.

[15] Dizier, M.H., Bonaiti-Pellie, C. et Clerget-Darpoux, F., 1993, Conclusions of segregation analysis for family data generated under two-locus models. *Am J Hum Genet*, 53, p 1338-1346.

[16] Doerge, R.W. et Churchill, G.A., 1996. Permutation tests for multiple loci affecting a quantitative character. *Genetics*, 142, p 285-294.

[17] Elsen, J.M. et Le Roy, P., 1995. Optimal design for the detection of a major gene segregation in crosses between 2 pure lines. *Genet Sel Evol*, 27, p 275-285.

[18] Elsen, J.M. et Le Roy, P., 1996. Recherche d'un marqueur génétique d'un gene à effet majeur. Séminaire : Planification expérimentale en génétique animale, Saint-Martin-de-Ré, 02/04 avril 1996.

[19] Elsen J.M., Mangin B., Goffinet B., Boichard D. et Le Roy P., 1999. Alternative models for QTL detection in livestock. I. General introduction. *Genet. Sel. Evol.*, 31, p 213-224.

[20] Falconer, D.S., 1989. *Introduction to quantitative genetics*, 3ème Ed., Longman, UK.

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

[22] Georges, M. et Andersson, L., 1996. Livestock genomics comes to age. *Genome Res*, 6, p 907-921.

[23] Georges, M., Nielsen, D., Mackinnon, M., Mishra, A., Okimoto, R., Pasquino, A.T., Sargeant, L.S., Sorensen, A., Steele, M.R., Zhao, X., Womack, J.E. et Hoeschele, I., 1995. Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics*, 139, p 907-920.

[24] Goffinet, B., Beckmann, J., Boichard, D., Causse, M., Charcosset, A., Chevalet, C., Christophe, C., Colleau, J.J., Demenais, F., Durel, C.E., Elsen, J.M., Foulley, J.L., Gallais, A., Gotz, K.U., Hospital, F., Kremer, A., Lorieux, M., Lefort-Buson, M., Le Roy, P., Loisel, P., Mangin, B., Maurice, A., Perrier, X., Pons, O., Rebaï, A., Rodolphe, F., San Cristobal, M. et Vu Tien Khang, J., 1994. Méthodes mathématiques pour l'étude des gènes contrôllant des caractères quantitatifs. *Genet Sel Evol*, 26, suppl 1, p 9s-20s.

[25] Goffinet, B. et Mangin, B., 1998. Comparing methods to detect more than one QTL on a chromosome. *Theor Appl Genet*, 96, p 628-633.

[26] Goffinet B., Le Roy P., Boichard D., Elsen J.M. et Mangin B., 1999. Alternative models for QTL detection in livestock. III. Heteroskedastic model and models corresponding to several distributions of the QTL effects. *Genet Sel Evol*, 31, p 341-350.

[27] Haldane, J.B.S., 1919. The combination of linkage values, and the calculation of distance between the loci of linked factors. *J Genet*, 8, p 299-309.

[28] Haley C.S. et Knott S.A., 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69, p 315-324.

[29] Haley, C.S., Knott, S.A. et Elsen, J.M., 1994. Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics*, 136, p 1195-1207.

[30] Harrel F.E. et Davis C.E., 1982. A new distribution-free quantile estimator. *Biometrika*, 69, p 635-640.

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

[32] Henshall, J.M. et Goddard, M.E., 1999. Multiple-trait mapping of quantitative trait loci after selective genotyping using logistic regression. *Genetics*, 151, p 885-894.

[33] Hoeschele, I., Uimari, P., Grignola, F.E., Zhang, Q. et Gage, K.M., 1997. Advances in statistical methods to map quantitative trait loci in outbred populations. *Genetics*, 147, p 1445-1457.

[34] Jansen, R.C., 1992. A general mixture model for mapping quantitative trait loci by using molecular markers. *Theor Appl Genet*, 85, p 252-260.

[35] Jansen, R.C., 1993. Interval mapping of multiple quantitative trait loci. *Genetics*, 135, p 205-211.

[36] Jansen, R.C. et Stam, P., 1994. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, 136, p 1447-1455.

[37] Jansen, R.C., 1994. Controlling the type I and type II errors in mapping quantitative trait loci. *Genetics*, 138, p 871-881.

[38] Jansen, R.C., 1996. A general Monte Carlo method for mapping multiple quantitative trait loci. *Genetics*, 142, p 305-311.

[39] Jansen, R.C., Johnson, D.L. et van Arendonk J.A.M., 1998. A mixture model approach to the mapping of quantitative trait loci in complex populations with an application to multiple cattle families. *Genetics*, 148, p 391-399.

[40] Jiang, C. et Zeng, Z.B., 1995. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, 140, p 1111-1127.

[41] Kao, C.H., Zeng, Z.B. et Teasdale, R.D., 1999. Multiple interval mapping for quantitative trait loci. *Genetics*, 152, p 1203-1216.

[42] Kao C.H., 2000. On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. *Genetics*, 156, p 855-865.

[43] Kendall, M. et Stuart, A., 1979. *The advanced theory of statistics*, vol 2. Griffin, London.

[44] Knapp S.J., Bridges W. et Birked D., 1990. Mapping quantitative trait loci using molecular marker linkage maps. *Theor Appl Gen*, 79, p 583-592.

[45] Knapp, S.J., 1991. Using molecular markers to map multiple quantitative trait loci : models for backcross, recombinant inbred, and doubled haploid progeny. *Theor Appl Genet*, 81, p 333-338.

[46] Knapp, M., 1998. Alternative test for linkage between two loci. *Genet Epidemiol*, 15, p 511.

[47] Knapp, M., 1999. A note on power approximations for the transmission/disequilibrium test. *Am J Hum Genet*, 64, p 1177-1185.

[48] Knott, S.A. et Haley, C.S., 1992. Maximum likelihood mapping of quantitative trait loci using full-sib families. *Genetics*, 132, p 1211-1222.

[49] Knott, S.A., Haley, C.S. et Thompson, R., 1992. Methods of segregation analysis for animal breeding data: a comparison of power. *Heredity*, 68, p 299-311.

[50] Knott, S.A., Haley, C.S. et Thompson, R., 1992. Methods of segregation analysis for animal breeding data : parameter estimates. *Heredity*, 68, p 313-320.

[51] Knott S.A., Elsen J.M. et Haley C.S., 1996. Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. *Theor. Appl. Genet.*, 93, p 71-80.

[52] Knott, S.A., Marklund, L., Haley, C.S., Andersson, K., Davies, W., Ellegren, H., Fredholm, M., Hansson, I., Hoyheim, B., Lundstrom, K., Moller, M. et Andersson, L., 1998. Multiple marker mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. *Genetics*, 149, p 1069-1080.

[53] Knott, S.A. et Haley, C.S., 2000. Multitrait least squares for quantitative trait loci detection. *Genetics*, 156, p 899-911.

[54] Knott, S.A., Nystrom, P.E., Andersson-Eklund, L., Stern, S., Marklund, L., Andersson, L. et Haley, C.S., 2002. Approaches to interval mapping of QTL in a multigeneration pedigree: the example of porcine chromosome 4. *Anim Genet*, 33, p 26-32.

[55] de Koning D.J., Visscher, P.M., Knott, S.A. et Haley C.S., 1998. A strategy for QTL detection in half-sib populations. *Animal Science*, 67, p 257-268.

[56] de Koning, D.J., Schulmant, N.F., Elo, K., Moisio, S., Kinos, R., Vilkki, J. et Maki-Tanila, A., 2001. Mapping of multiple quantitative trait loci by simple regression in half-sib designs. *J Anim Sci*, 79, p 616-622.

[57] Korol, A.B., Ronin, Y.I. et Kirzhner, V.M., 1995. Interval mapping of quantitative trait loci employing correlated trait complexes. *Genetics*, 140, p 1137-1147.

[58] Korol, A.B., Ronin, Y.I. et Nevo, E., 1998. Approximate analysis of QTL-environment interaction with no limits on the number of environments. *Genetics*, 148, p 2015-2028.

[59] Korol, A.B., Ronin, Y.I., Itskovich, A.M., Peng, J. et Nevo, E., 2001. Enhanced efficiency of quantitative trait loci mapping analysis based on multivariate complexes of quantitative traits. *Genetics*, 157, p 1789-1803.

[60] Kosambi, D.D., 1944. The estimation of map distance from recombination distributions. *Proc Natl Acad Sci*, USA, 1978, 75, p 6332-6336.

[61] Lander, E.S. et Kruglyak, L., 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet*, 11, p 241-247.

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

[63] Leadbetter, M.R., Lindgren, G. et Rootzen, H., 1983. *Extremes and related properties of random sequences and processes*. Springer, New York.

[64] Lebreton, C.M. et Visscher, P.M., 1998. Empirical nonparametric bootstrap strategies in quantitative trait loci mapping: conditioning on the genetic model. *Genetics*, 148, p 525-535.

[65] Lebreton, C.M., Visscher, P.M., Haley, C.S., Semikhodskii, A. et Quarrie, S.A., 1998. A nonparametric bootstrap method for testing close linkage vs. pleiotropy of coincident quantitative trait loci. *Genetics*, 150, p 931-943.

[66] Lehman, E.C., 1986. *Testing Statistical Hypotheses*, Ed.2 John Wiley & Sons, New York.

[67] Le Roy, P., Elsen, J.M. et Knott, S.A., 1989. Comparison of four statistical methods for detection of a major gene in a progeny test design. *Genet Sel Evol*, 21, p 341-357.

[68] Le Roy, P. et Elsen, J.M., 1992. Simple test statistics for major gene detection: a numerical comparison. *Theor Appl Genet*, 83, p 635-644.

[69] Le Roy, P., Elsen, J.M., Boichard, D., Mangin ,B., Bidanel, J.P. et Goffinet, B., 11-16 janvier 1998. An algorithm for QTL detection in mixture of full and half sib families. 6ième World Congress of Genetic Applied to Livestock Production, vol. 26, University of New England, Armidale, p 257-260.

[70] Lipkin, E., Mosig, M.O., Darvasi, A., Ezra, E., Shalom, A., Friedmann, A. et Soller, M., 1998. Quantitative trait locus mapping in dairy cattle by means of selective milk DNA pooling using dinucleotide microsatellite markers: analysis of milk protein percentage. *Genetics*, 149, p 1557-1567.

[71] Liu, Z. et Dekkers, J.C., 1998. Least squares interval mapping of quantitative trait loci under the infinitesimal genetic model in outbred populations. *Genetics*, 148, p 495-505.

[72] Luo, Z.W. et Kearsey, M.J., 1992. Interval mapping of quantitative trait loci in an F2 population. *Heredity*, 69, p 236-242.

[73] Manfredi, E., Barbieri, M., Fournet, F. et Elsen, J.M., 1998. A dynamic deterministic model to evaluate breeding strategies under mixed inheritance. *Genet Sel Evol*, 30, 127-148.

[74] Mangin, B. et Goffinet, B., 1994a. Statistical testing in genetic linkage under heterogeneity. *Biometrics*, 50, p 308.

[75] Mangin, B., Goffinet, B. et Rebai, A., 1994b. Constructing confidence intervals for QTL location. *Genetics*, 138, p 1301-1308.

[76] Mangin B., Thoquet P. et Grimsley N.H., 1998. Pleiotropic QTL analysis. *Biometrics*, 54, p 88-99.

[77] Mangin B., Goffinet B., Le Roy P., Boichard D. et Elsen J.M., 1999. Alternative models for QTL detection in livestock. II. Likelihood approximations and sire marker genotype estimations. *Genet Sel Evol*, 31, p 225-237.

[78] Mangin B., Thoquet P., Olivier J. et Grimsley N.H., 1999. Temporal and multiple quantitative trait loci analyses of resistance to bacterial wilt in tomato permit the resolution of linked loci. *Genetics*, 151, p 1165-1172.

[79] Mardia K.V., Kent J.T. et Bibby J.M., 1979. *Multivariate analysis, Discriminant Analysis*. Academic Press, London, p 300-332.

[80] Martínez O. et Curnow, R.N., 1992. Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theor Appl Genet*, 85, p 480-488.

[81] Mather, K. et Jinks, J.L., 1982. *Biometrical Geneitcs* (3ème Ed.), Chapman and Hall, Londres.

[82] Milan, D., Bidanel, J.P., Iannuccelli, N., Riquet, J., Amigues, Y., Gruand, J., Le Roy, P., Renard, C., Chevalet, C., 2002. Detection of quantitative trait loci for carcass composition traits in pigs. *Genet. Sel. Evol.*, 34, p 705-728.

[83] Nakamichi, R., Ukai, Y. et Kishino, H., 2001. Detection of closely linked multiple quantitative trait loci using a genetic algorithm. *Genetics*, 158, p 463-475.

[84] Ollivier, L., 1981. *Eléments de Génétique Quantitative*. Manson, Paris.

[85] van Ooijen, J.W., 1992. Accuracy of mapping quantitative trait loci in autogamous species. *Theor Appl Genet*, 84, p 803-811.

[86] Pazsek, A.A., Wilkie, P.J., Flickinger, G.H., Rohrer, G.A., Alexander, L.J., Beattie, C.W. et Shook, L.B., 1999. Interval mapping of growth in divergent swine cross. *Mamm Genom*, 10, 117-122.

[87] Paterson, A.H., Damon, S., Hewitt, J.D., Zamir, D., Rabinowitch, H.D., Lincoln, S.E., Lander, E.S. et Tanksley, S.D., 1991. Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. *Genetics*, 127, p 181-197.

[88] Piper, L.R. et Bindon, B.M., 1990a. *The Booroola gene, Fec^B, in Australia*, 2ième congrès international Major genes for reproduction in sheep, Toulouse, France. Les Colloques, INRA, 57, p 43-45.

[89] Piper, L.R. et Bindon, B.M., 1990b. *Strategies for utilization of a major gene for prolificacy in sheep.*, 2ième congrès international Major genes for reproduction in sheep, Toulouse, France. Les Colloques, INRA, 57, p 399-408.

[90] Rebai, A., Goffinet, B. et Mangin, B., 1994. Approximate thresholds of interval mapping tests for QTL detection. *Genetics*, 138, p 235-240.

[91] Rebai, A., Goffinet, B. et Mangin, B., 1995. Comparing power of different methods for QTL detection. *Biometrics*, 51, p 87-99.

[92] Rodolphe, F. et Lefort, M., 1993. A multi-marker model for detecting chromosomal segments displaying QTL activity. *Genetics*, 134, p 1277-1288.

[93] Ronin Y.I., Kirzhner V.M. et Korol A.B., 1995. Linkage between loci of quantitative traits and marker loci : multitrait analysis with a single marker. *Theor Appl Genet*, 90, p 776-786.

[94] Ronin, Y.I., Korol, A.B. et Nevo, E., 1999. Single- and multiple-trait mapping analysis of linked quantitative trait loci. Some asymptotic analytical approximations. *Genetics*, 151, p 387-396.

[95] Sakamoto, Y., Ishiguro, M. et Kitagawa, G., 1986. *Akaike Information Criterion*. KTK Scientific Publishers, Tokyo.

[96] Schmoyer, R.L., 1994. Permutation for correlation in regression errors. *JASA*, 89, p 1507-1516.

[97] Sellier, P., 1998. *Genetics of meat and carcass traits*, The genetics of the pigs, Ed. M.F. Rothschild & A. Ruvinsky, p 463-510.

[98] Soller, M., Brody, T. et Genizi, A., 1976. On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosse between inbred lines. *Theor Appl Genet*, 47, p 35-39.

[99] Soller, M. et Genizy, 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, p 47-55.

[100] Spelman, R.J., Coppieters, W., Karim, L., van Arendonk, J.A. et Bovenhuis, H., 1996. Quantitative trait loci analysis for five milk production traits on chromosome six in the Dutch Holstein-Friesian population. *Genetics*, 144, p 1799-808.

[101] Tanksley, S.D., Medina-Filho, H. et Rick, C.M., 1982. Use of naturally-occuring enzyme variation to detect and map genes controlling quantitative traits in an interspecific back-cross of tomato. *Heredity*, 49, p 11-25.

[102] Vilkki, H.J., de Koning, D.J., Elo, K., Velmala, R. et Maki-Tanila, A., 1997. Multiple marker mapping of quantitative trait loci of Finnish dairy cattle by regression. *J Dairy Sci*, 80, p 198-204.

[103] Visscher, P. et Haley, C., 2001. True and false positive peaks in genomewide scans: The long and the short of it. *Genet Epidemiol*, 20, p 409-14.

[104] Visscher, P.M., Haley, C.S. et Thompson, R., 1996. Marker-assisted introgression in backcross breeding programs. *Genetics*, 144, p 1923-1932.

[105] Walling, G.A., Archibald, A.L., Cattermole, J.A., Downing, A.C., Finlayson, H.A., Nicholson, D., Visscher, P.M., Walker, C.A. et Haley, C.S., 1998. Mapping of quantitative trait loci on porcine chromosome 4. *Anim Genet*, 29, p 415-424.

[106] Weller, J.I., Kashi, Y. et 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, p 2525-2537.

[107] Weller J.I., Wiggans G.R., VanRaden P.M. et Ron M., 1996. Application of a canonical transformation to detection of quantitative trait loci with the aid of genetic markers in a multi-trait experiment. *Theor Appl Genet*, 92, p 998-1002.

[108] Weller J.I., Song J.Z., Ronin Y.I. et Korol A.B., 1997. Designs and solutions to multiple trait comparisons. *Animal Biotechnology*, 8, p 107-122.

[109] Weller, J.I., Song, J.Z., Heyen, D.W., Lewin, H.A. et Ron, M., 1998. A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics*, 150, p 1699-1706.

[110] Whittaker, J.C., Thompson, R. et Visscher, P.M., 1996. On the mapping of QTL by regression of phenotype on marker-type. *Heredity*, 77, p 23-32.

[111] Wright, S., 1968. *Evolution and the genetics of populations*, vol 1, Genetic and Biometric Foundations. University of Chicago Press, Chicago.

[112] Wu, W.R. et Li, W.M., 1994. A new approach for mappping quantitative trait loci using complete genetic marker linkage maps. *Theor Appl Genet*, 89, p 535-539.

[113] Wu W.R., Li W.M., Tang D.Z., Lu H.R. et Worland A.J., 1999. Time-related mapping of quantitative trait loci underlying tiller number in rice. *Genetics*, 151, p 297-303.

[114] Zeng Z.B., 1993. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc Natl Acad Sci U.S.A.*, 90, p 10972-10976.

[115] Zeng, Z.B., 1994. Precision mapping of quantitative trait loci. *Genetics*, 136, p 1457-1468.

[116] Zeng, Z.B., Kao, C.H. et Basten, C.J., 1999. Estimating the genetic architecture of quantitative traits, *Genet Res Camb*, 74, p 279-289.

Liste des tableaux

1.1 Nombre de descendants nécessaires à la détection d'un QTL. 9

1.2 Espérance des coefficients des composantes de a et d pour toutes les combinaisons de génotypes possibles dans une population F2. 17

1.3 Probabilités critiques (p_c) et valeurs seuils de lod pour des liaisons suggérées ou significatives. 21

1.4 Moyennes, variances et corrélations phénotypiques entre les caractères étudiés. 28

1.5 Répartition des F2 dans les familles de PORQTL pour l'analyse des 5 caractères. 29

1.6 Analyses unicaractères uniQTL des 5 caractères de composition de carcasse sur le chromosome 7. 30

2.1 Nombre de paramètres à estimer pour chaque méthode multicaractère, en fonction du nombre de caractères analysés (p) et du nombre de descendants par mère (ndm). 55

2.2 Temps de calcul moyens des méthodes multicaractères, en secondes de CPU en fonction du nombre de caractères analysés 58

2.3 Seuils de rejet de H0 des méthodes multicaractères, en fonction du nombre de caractères analysés (p), de la densité de la carte génétique et du nombre de descendants par mère (ndm). 60

2.4 Puissances de détection des méthodes multicaractères et ST en fonction du nombre de caractères simulés (p) et du signe de la corrélation résiduelle entre les deux premiers caractères (ρ_{12}). 61

2.5 Biais et variances d'échantillonnage des estimations des positions pour les analyses multicaractères en fonction du nombre de caractères analysés (p) et du signe de Pr_{12}. 63

2.6 Somme des Carrés des Erreurs (SCE) des estimations des effets QTL (x 10^2) pour les méthodes multicaractères en fonction du nombre de caractères analysés (p) et du signe de Pr_{12}. 64

2.7 Puissances pour les méthodes multicaractères en fonction de la corrélation résiduelle (ρ_{12}) et des effets de substitution des allèles au QTL (α_1 ; α_2). . . 65

2.8 Somme des Carrés des Erreurs (SCE) des estimations de la position du QTL (x 10^2) en fonction de la corrélation résiduelle (ρ_{12}) et des effets de substitution des allèles au QTL (α_1, α_2). 67

2.9 SCE des estimations des effets de substitution au QTL (x 10^2) en fonction de la corrélation résiduelle (ρ_{12}) et des effets de substitution des allèles au QTL (α_1, α_2). 68

2.10 Puissances et SCE des positions estimées par les méthodes multicaractères en fonction du signe de la corrélation résiduelle (ρ_{12}) et de la valeur de l'effet de substitution des allèles au QTL sur le premier caractère (α_2), avec α_1 =0 . 69

2.11 Puissances, SCE des estimations des positions et SCE des estimations des effets QTL des méthodes multicaractères en fonction de la densité en marqueurs génétiques et du nombre de descendants par femelle (ndm). 71

2.12 Puissances des méthodes multicaractères en fonction de la fréquence des allèles au QTL dans les populations grand-parentales (fp(q)=fm(Q)) et du nombre de descendants par père (ndp). 74

2.13 SCE des estimations des positions des méthodes multicaractères en fonction de la fréquence des allèles au QTL dans les populations grand-parentales (fp(q)=fm(Q)) et du nombre de descendants par père (ndp). 74

2.14 Puissances des méthodes multicaractères en fonction de la fréquence de l'allèle Q au QTL dans la population grand-paternelle (fp(Q), avec fm(q)=1) et du nombre de descendants par père (ndp). 77

2.15 SCE des estimations des positions des méthodes multicaractères en fonction de la fréquence de l'allèle Q au QTL dans la population grand-paternelle (fp(Q), avec fm(q)=1) et du nombre de descendants par père (ndp). 77

2.16 Analyse discriminante des 5 caractères. 83

2.17 Sélection des caractères par l'analyse discriminante. 84

2.18 Effets estimés par père et effets moyens pour l'analyse conjointe de imf et panne. 84

2.19 Sélection des caractères complémentaires par l'analyse discriminante. . . . 85

2.20 Analyses réalisées avec la méthode multivariée. 86

2.21 Analyses en composantes principales. 88

3.1 Différents modèles d'épistasie. 90

3.2 Résumé des effets simulés pour chaque QTL sur chaque caractère lors des analyses multiQTL. $a = 0,5$. 113

3.3 Moyenne des temps de calcul nécessaires à l'analyse d'un jeu de données par les méthodes multiQTL. 113

3.4 Effets et positions estimés lors des analyses uniQTL en fonction des effets des deux QTL. 114

3.5 Seuils estimés pour les méthodes multiQTL. 116

3.6 Puissances pour les méthodes multiQTL. 118

3.7 Effets des portions chromosomiques utilisables pour la transformation par DA2 dans les différents groupes d'haplotypes en fonction des cas. 121

3.8 Répartition des minima des AIC pour chaque série de 100 simulations. . . 124

3.9 SCE des estimations des positions (x 10^2) pour chaque QTL par les méthodes multiQTL. 125

3.10 SCE des estimations des effets pères pour chaque QTL par les méthodes multiQTL. 128

3.11 Biais d'estimation des effets et des positions par les méthodes multiQTL en fonction de la localisation des QTL simulés. 131

3.12 Maxima des statistiques de test et positions correspondantes pour l'analyse multiQTL unicaractère. 134

Table des figures

1	Distribution du nombre de locus dans un génome en fonction de leurs effets (σ_p).	3
2	Distribution des performances lorsqu'un QTL est en ségrégation dans une population.	3
3	Ségrégation des allèles Q1 et Q2 au QTL dans la descendance d'un individu hétérozygote.	4
4	Suivi de la ségrégation des allèles au QTL à l'aide d'un marqueur.	4
5	Deux principaux dispositifs expérimentaux utilisés pour la détection de QTL.	7
6	Principe de la cartographie d'intervalle.	13
7	Carte génétique utilisée pour le chromosome 7 porcin.	27
8	Analyses des données de l'exemple par les méthodes unicaractère uniQTL.	29
9	Exemples de répartition des performances pour deux caractères dans des situations dans lesquelles les méthodes multicaractères permettent d'améliorer l'efficacité des analyses.	34
10	Pourcentages de gain relatif de puissance entre une analyse deux caractères et une analyse unicaractère, en fonction de la corrélation résiduelle et du rapport des effets du QTL.	39
11	Fréquence des génotypes en F1 en fonction des fréquences alléliques dans les populations F0.	72
12	Fréquence des génotypes au QTL dans la génération F2 quand fp(q) varie, fm(q) étant fixé à 1.	75
13	Méthodes multicaractères : sélection du groupe de caractères permettant d'obtenir la statistique de test la plus significative.	85

14	Allèles en phase ou en répulsion.	90
15	Détection d'un QTL fantôme quand deux QTL sont en phase sur le même chromosome.	90
16	Exemples de répartition des descendants d'un père F1 en fonction des couples d'allèles reçus au QTL.	91
17	Schéma des décisions présidant aux itérations dans le processus de détection de QTL multiQTL .	102
18	Répartition des QTL dans les simulations 2 QTL.	111
19	Moyennes des estimations des positions par les méthodes multiQTL.	124
20	Moyennes des estimations des effets QTL par les méthodes multiQTL.	126
21	Analyse multiQTL du poids de panne.	134
22	Analyse multiQTL de panne et imf.	135
23	Analyse multiQTL de x2 et x4.	136
24	Positions des QTL mis en évidence sur le chromosome 7 porcin pour 5 caractères de composition corporelle.	137
25	Stratégie de sélection combinée des caractères et des régions pour la mise en évidence de QTL pléiotropes ou liés.	141
26	Stratégie de sélection des régions chromosomiques à effet pléiotrope.	142

Multidimensionnalité pour la détection de gènes affectant les caractères quantitatifs. Application à l'espèce porcine.

Ce travail a pour but de développer des méthodes de détection de locus affectant les caractères quantitatifs, appelés QTL, à partir de l'information disponible sur des caractères corrélés et/ou des positions liées, chez les animaux d'élevage.

Les méthodologies ont été dans un premier temps caractérisées pour leurs puissances et leurs précisions d'estimation des paramètres (positions et effets des QTL) à partir de données simulées. Nous avons développé d'une part des méthodes multivariées, extrapolées de techniques décrites pour l'analyse de données issues de croisements entre populations supposées génétiquement fixées, et d'autre part des méthodes synthétiques univariées, développées à l'occasion de ce travail. Ces dernières méthodes permettent de synthétiser l'information due à la présence du (des) QTL déterminant plusieurs caractères dans une unique variable, combinaison linéaire des caractères. Le nombre de paramètres à estimer est ainsi indépendant du nombre de caractères étudiés, permettant de réduire fortement les temps de calcul par rapport aux méthodes multivariées. La stratégie retenue repose sur des techniques d'analyse discriminante. Pour chaque vecteur de positions testé, des groupes de descendants sont créés en fonction de la probabilité que les individus aient reçu l'un ou l'autre haplotype de leur père. Les matrices de (co)variance génétique et résiduelle spécifiques de la présence du (des) QTL peuvent alors être estimées. La transformation linéaire permet de maximiser le rapport de ces deux variabilités.

Les méthodes basées sur l'analyse de variables synthétiques permettent en général d'obtenir des résultats équivalents, voire meilleurs, que les stratégies multivariées. Seule l'estimation des effets des QTL et de la corrélation résiduelle entre les caractères reste inaccessible par ces méthodes. Une stratégie itérative basée sur l'analyse de variables synthétiques pour la sélection des caractères et des régions chromosomiques à analyser par les méthodes multivariées est proposée. Par ailleurs, nous avons quantifié les apports des méthodologies multidimensionnelles pour la cartographie des QTL par rapport aux méthodes unidimensionnelles. Dans la majorité des cas, la puissance et la précision d'estimation des paramètres sont nettement améliorées. De plus, nous avons pu montrer qu'un QTL pléiotrope peut être discriminé de deux QTL liés, s'ils sont relativement distants.

Ces méthodologies ont été appliquées à la détection de QTL déterminant cinq caractères de composition corporelle chez le porc sur le chromosome 7. Deux groupes de QTL déterminant des types de gras différents, le gras interne et le gras externe, ont ainsi été discriminés. Pour chacun de ces groupes, les analyses multiQTL ont permis d'identifier au moins deux régions chromosomiques distinctes déterminant les caractères.

Mots-clef: QTL / multicaractère / multilocus / familles / maximum de vraisemblance / simulations

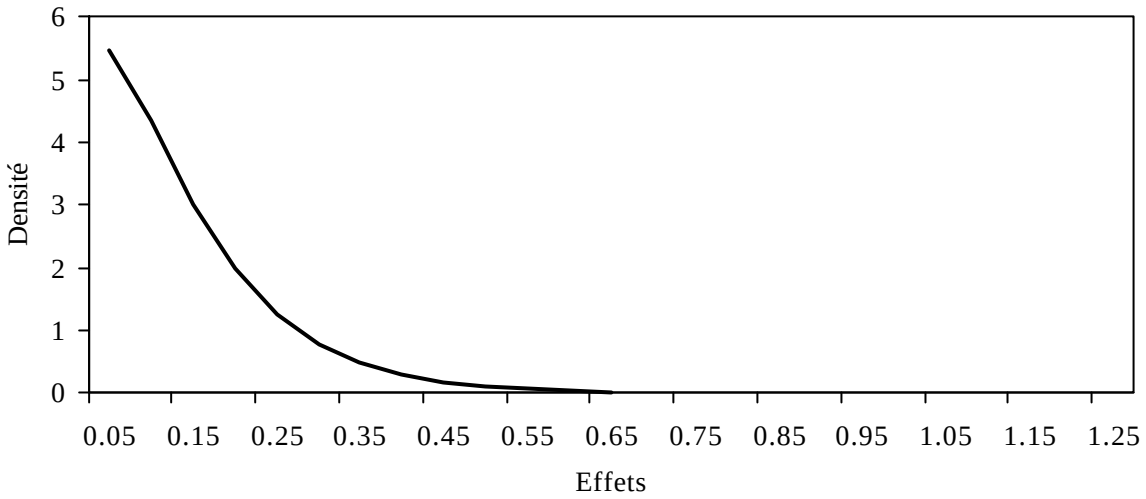


Figure 1: Distribution du nombre de locus dans un génome en fonction de leurs effets (s_p) (d'après Hayes et Goddard, 2001).

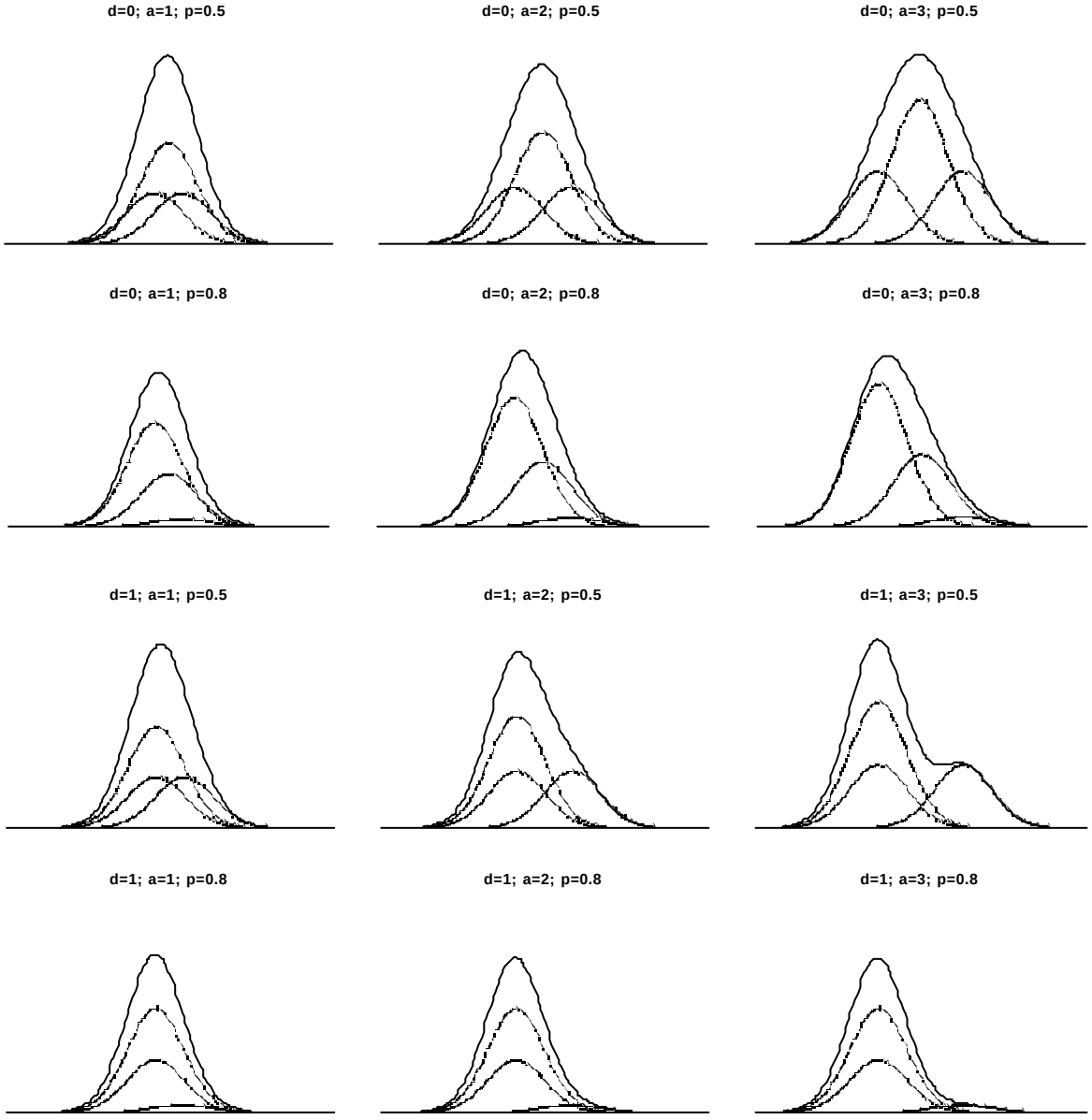


Figure 2 : Distribution des performances lorsqu'un QTL est en ségrégation dans une population.
Cas d'un QTL biallélique, d'effet additif a et dominant d, avec une fréquence p de l'allèle dominant.

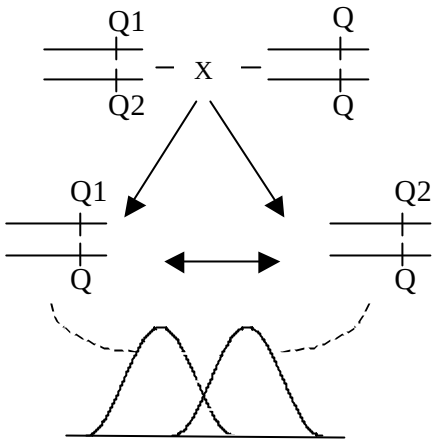


Figure 3 : Ségrégation des allèles Q1 et Q2 au QTL dans la descendance d'un individu hétérozygote.

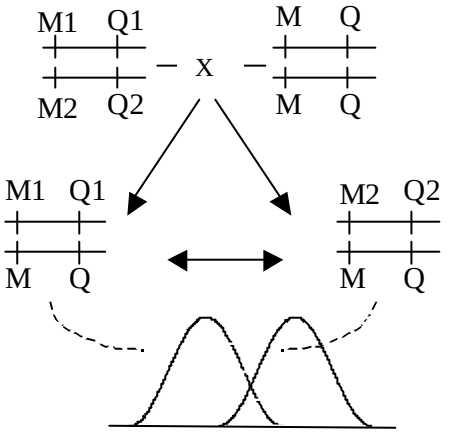
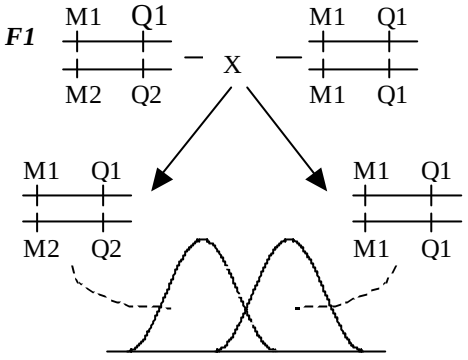
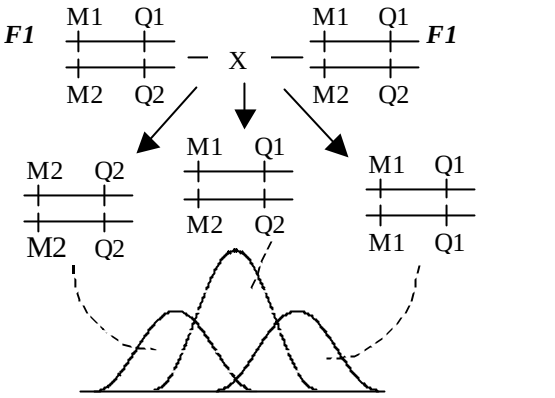


Figure 4 : Suivi de la ségrégation des allèles au QTL à l'aide d'un marqueur.
Deux allèles au QTL (Q1 et Q2) et au marqueur M (M1 et M2). Cas d'un marqueur complètement lié.



Croisement *backcross* sur la lignée parentale porteuse des allèles notés 1.



Croisement *intercross* entre deux individus F1

Figure 5 : Deux principaux dispositifs expérimentaux utilisés pour la détection de QTL.

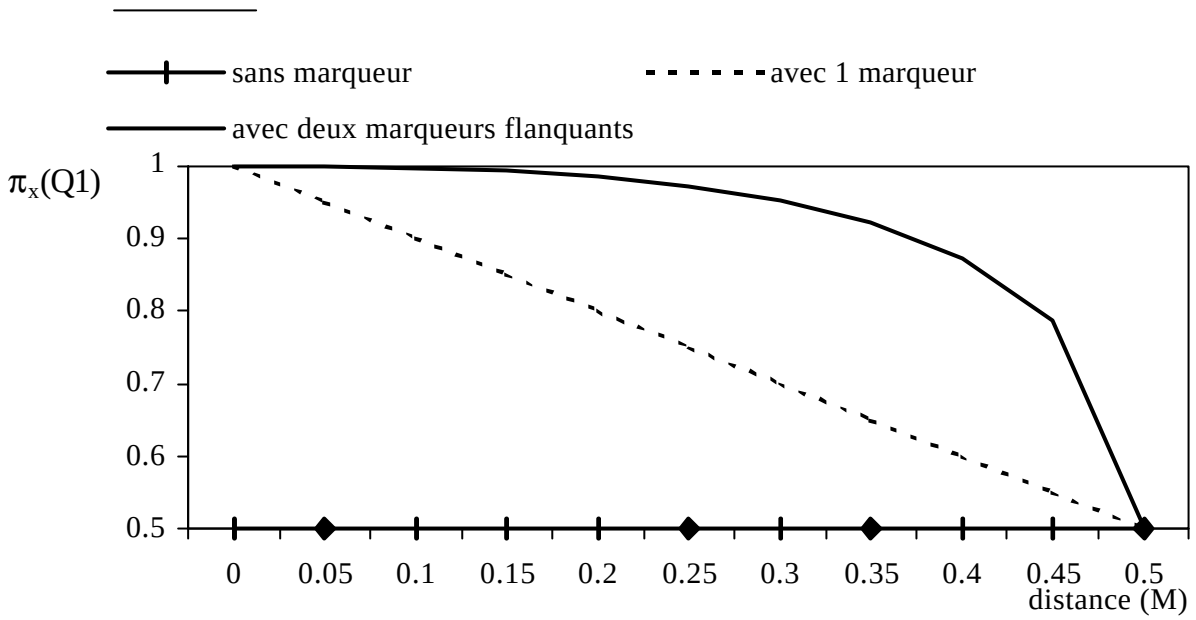


Figure 6 : Principe de cartographie d'intervalle.
Distribution de la probabilité de transmission de l'allèle Q1 à la position x en fonction de l'information à

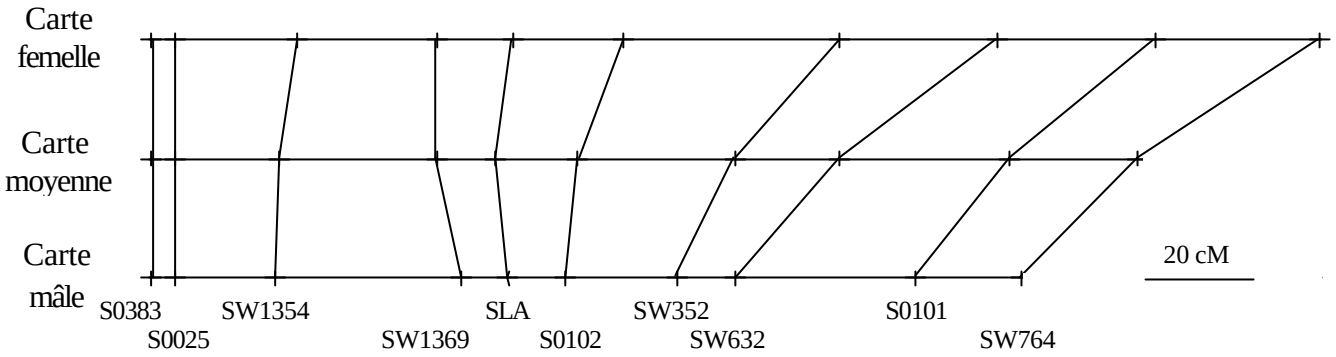


Figure 7 : Carte génétique utilisée pour le chromosome 7 porcin

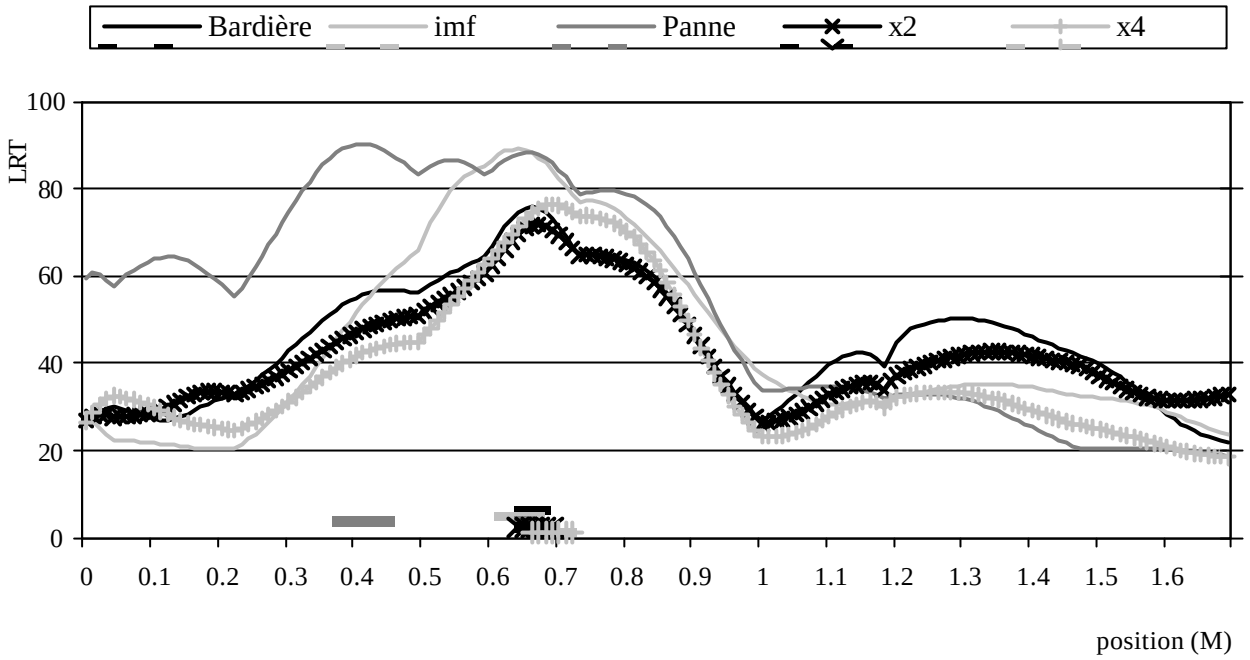


Figure 8 : Analyse des données de l'exemple par les méthodes unicaractère uniQTL.
Profils de vraisemblance obtenus pour les 5 caractères de composition corporelle sur le chromosome 7 porcin. Les lignes horizontales correspondent à une extrapolation des intervalles de confiance des

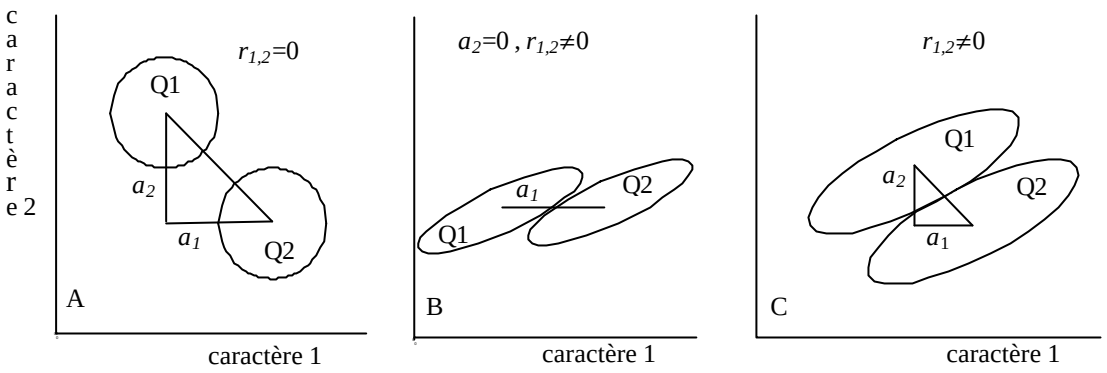


Figure 9 : Exemples de répartition des performances pour deux caractères dans des situations dans lesquelles les méthodes multicaractères permettent d'améliorer l'efficacité des analyses.
Au moins un caractère (caractère 1) est déterminé par un QTL biallélique Q1Q2. A : grâce aux effets pléï otropes du QTL; B : grâce à la corrélation $r_{1,2}$ entre les caractères due à d'autres facteurs (gènes, environnement…) que le QTL recherché ; C : grâce à la combinaison des deux effets précédents. a_i /2 = effet de substitution des allèles au QTL sur le caractère *i*.

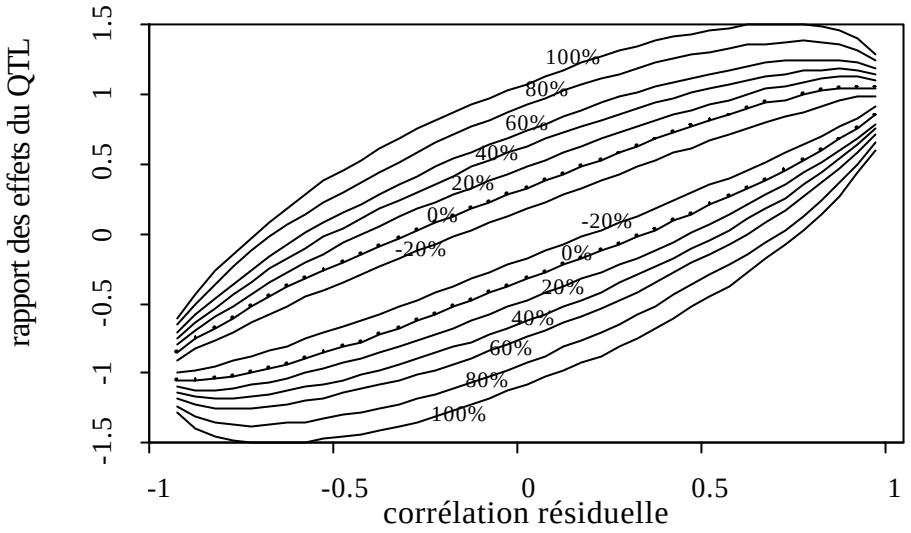


Figure 10 **:** Pourcentages de gain relatif de puissance entre une analyse deux caractères et une analyse unicaractère en fonction de la corrélation résiduelle et du rapport des effets du QTL.
Cas d'un effet de substitution standardisé des allèles au QTL sur le premier caractère de 0,3, d'une erreur de première espèce de 0,5%, de marqueurs répartis sur 100 cM tous les 20 cM et d'un backcross entre *et al.*, 1998).

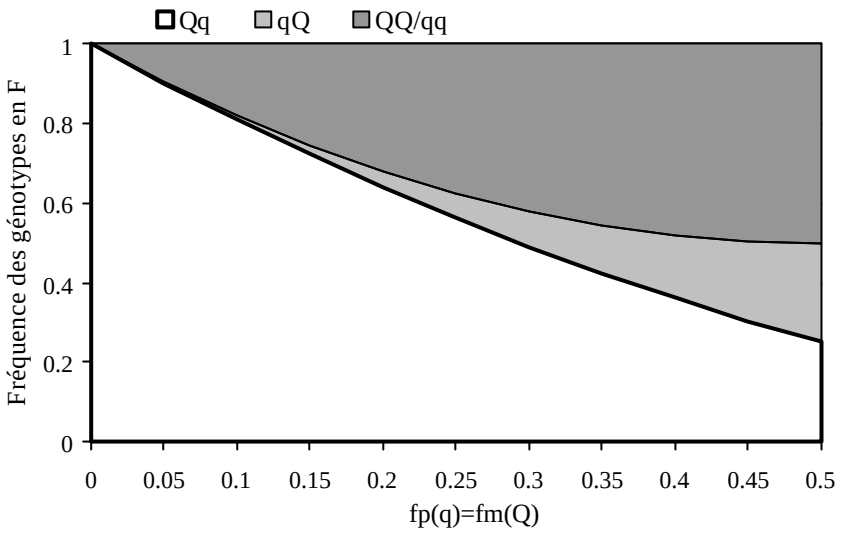


Figure 11 : Fréquence des génotypes en F1 en fonction des fréquence alléliques dans les populations F0.
Cas où la fréquence de l'allèle q chez les grand-pères (fp(q)) est égale à la fréquence de Q chez les grands-mères (fm(Q))

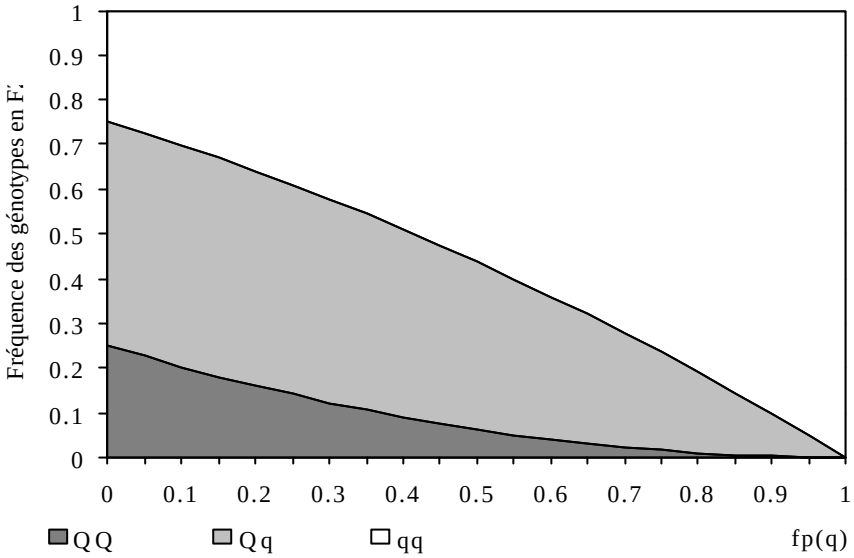


Figure 12 : Fréquence des génotypes au QTL dans la génération F2 quand fp(q) varie, fm(q) étant fixé à 1.

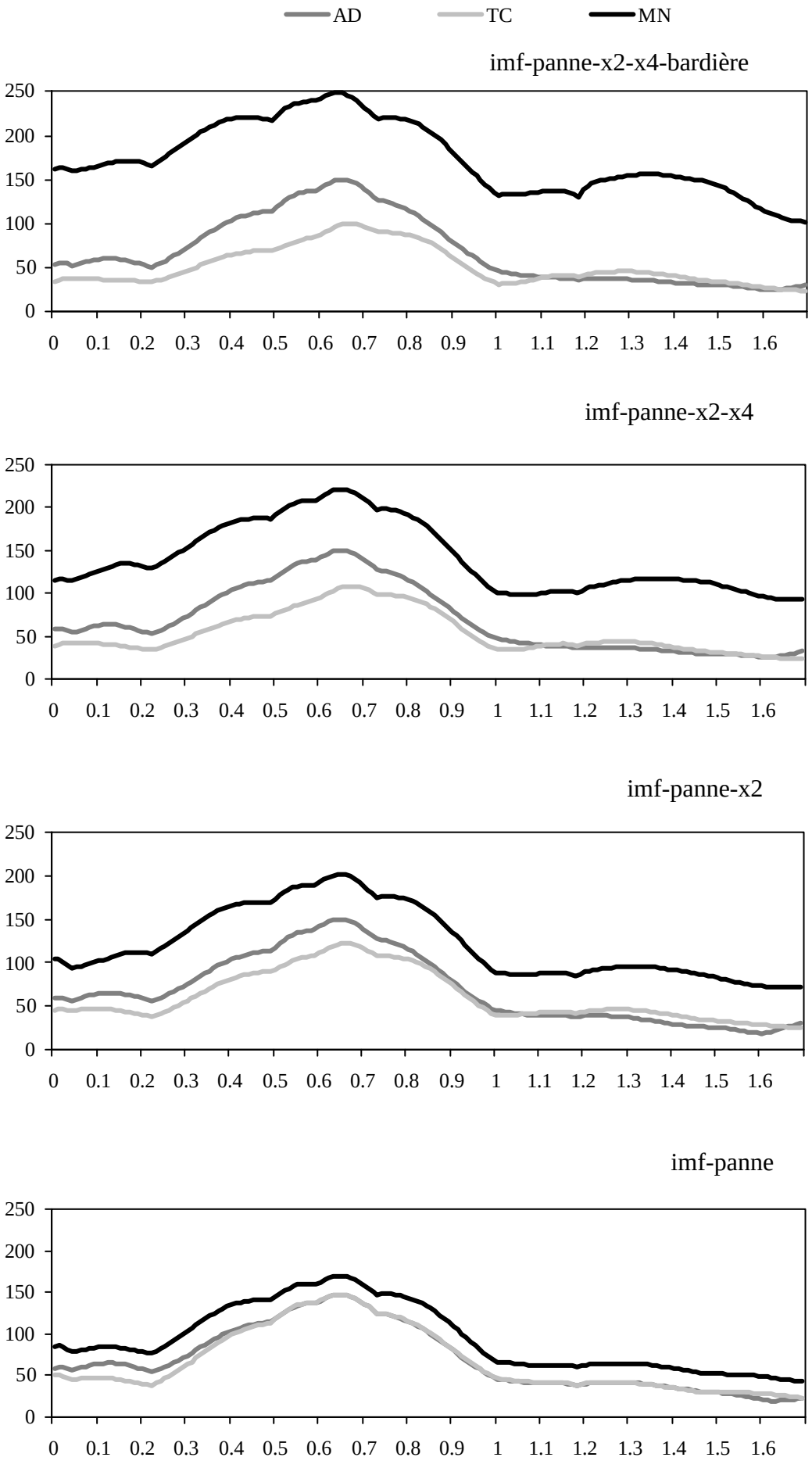


Figure 13 : Méthodes multicaractères : sélection du groupe de caractères permettant d'obtenir la statistique de test la plus significative.

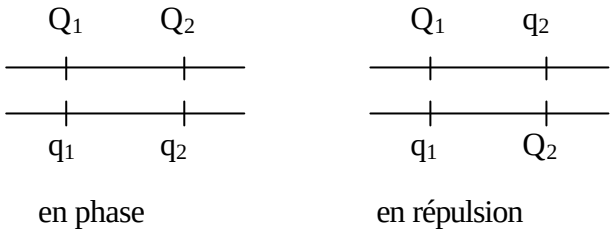


Figure 14 : Allèles en phase ou en répulsion.
Q_r et q_r représentent les deux allèles paternel pour le $r^{ième}$ QTL.

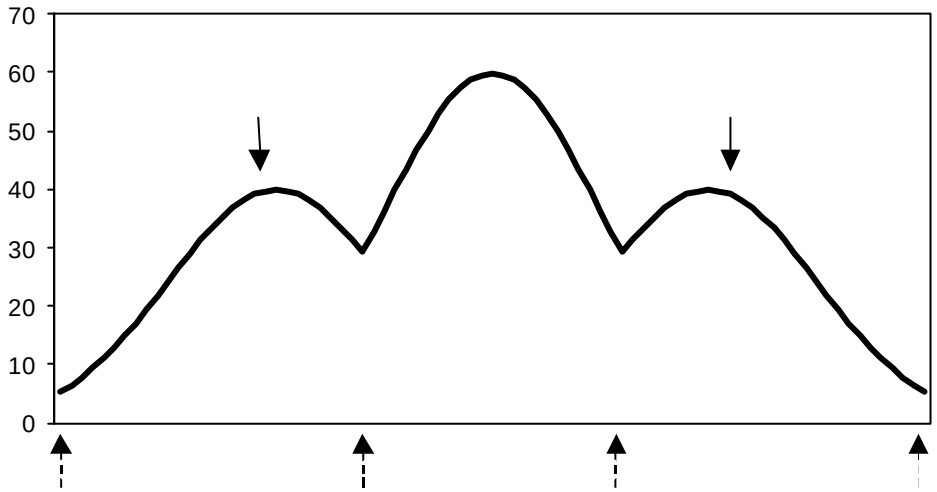


Figure 15 : Détection d'un QTL fantôme quand deux QTL sont en phase sur le même chromosome.
Les positions réelles des deux QTL sont indiquées par des flèches pleines, les positions des marqueurs génétiques par des flèches en pointillés (d'après Martinez et Curnow, 1992).

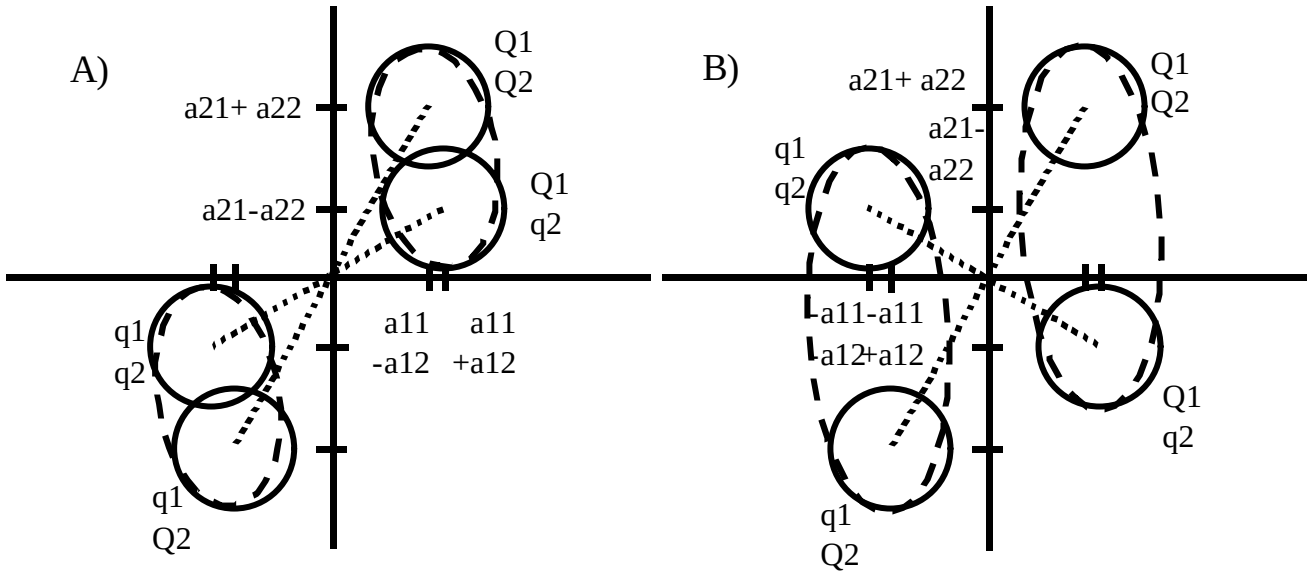


Figure 16 : Exemples de répartition des descendants d'un père en fonction des QTL.

Les deux axes représentent chacun un caractère (horizontal = caractère 1; vertical = caractère 2). Q_r et q_r représentent les deux allèles paternels pour le r QTL. a_{lr} est l'effet de substitution.

Cas A : tous les effets de substitution sont positifs. Cas B : l'effet du deuxième QTL sur le premier caractère est négatif. Les ellipses en trait pointillés représentent la répartition des descendants que l'on obtiendrait en ne considérant que la première position.

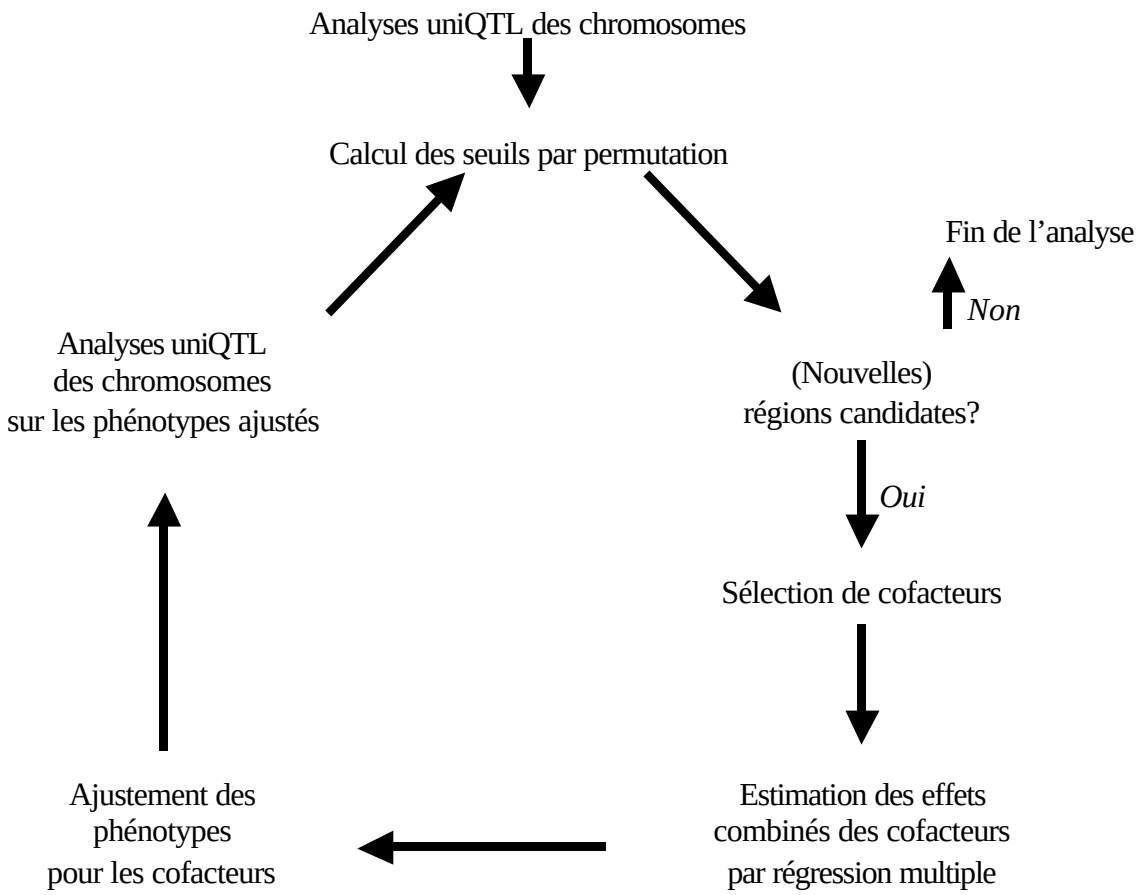


Figure 17 : Schéma des décisions présidant aux itérations dans le processus de détection de QTL *et al.*, 2001)

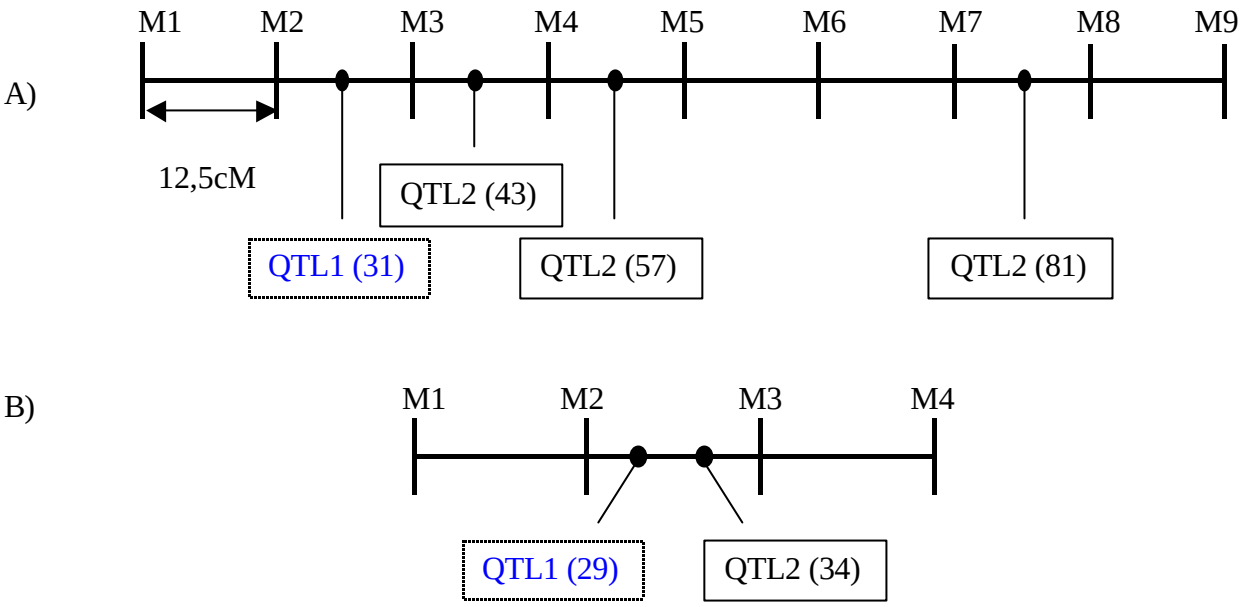


Figure 18 : Répartition des QTL dans les simulations 2 QTL.
M_i est le $i^{ème}$ marqueur sur le groupe de liaison. La distance M1M9 est de 1 Morgan. Les positions des QTL sont données entre parenthèses. Cas A : le premier QTL est toujours à la position 31 cM et le deuxième peut être en position 81, 57, ou 43 cM. Cas B : ils sont respectivement à 29 et 34 cM.

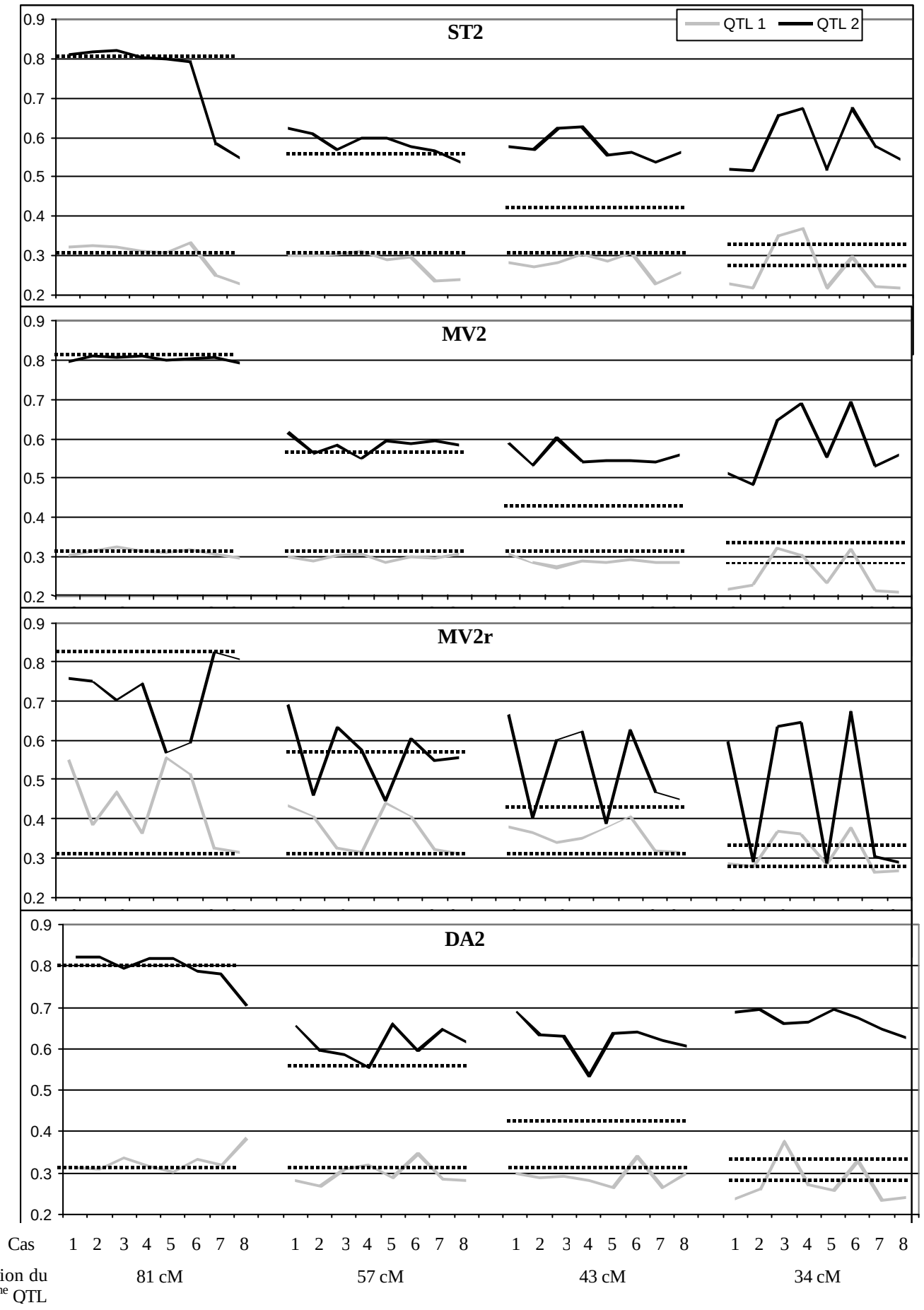


Figure 19 : Moyennes des estimations des positions par les méthodes multiQTL.
ST2 : méthode unicaractère ; MV2 : méthode multicaractère complète ; MV2r : méthode multicaractère restreinte ; AD2 : analyse discriminante.

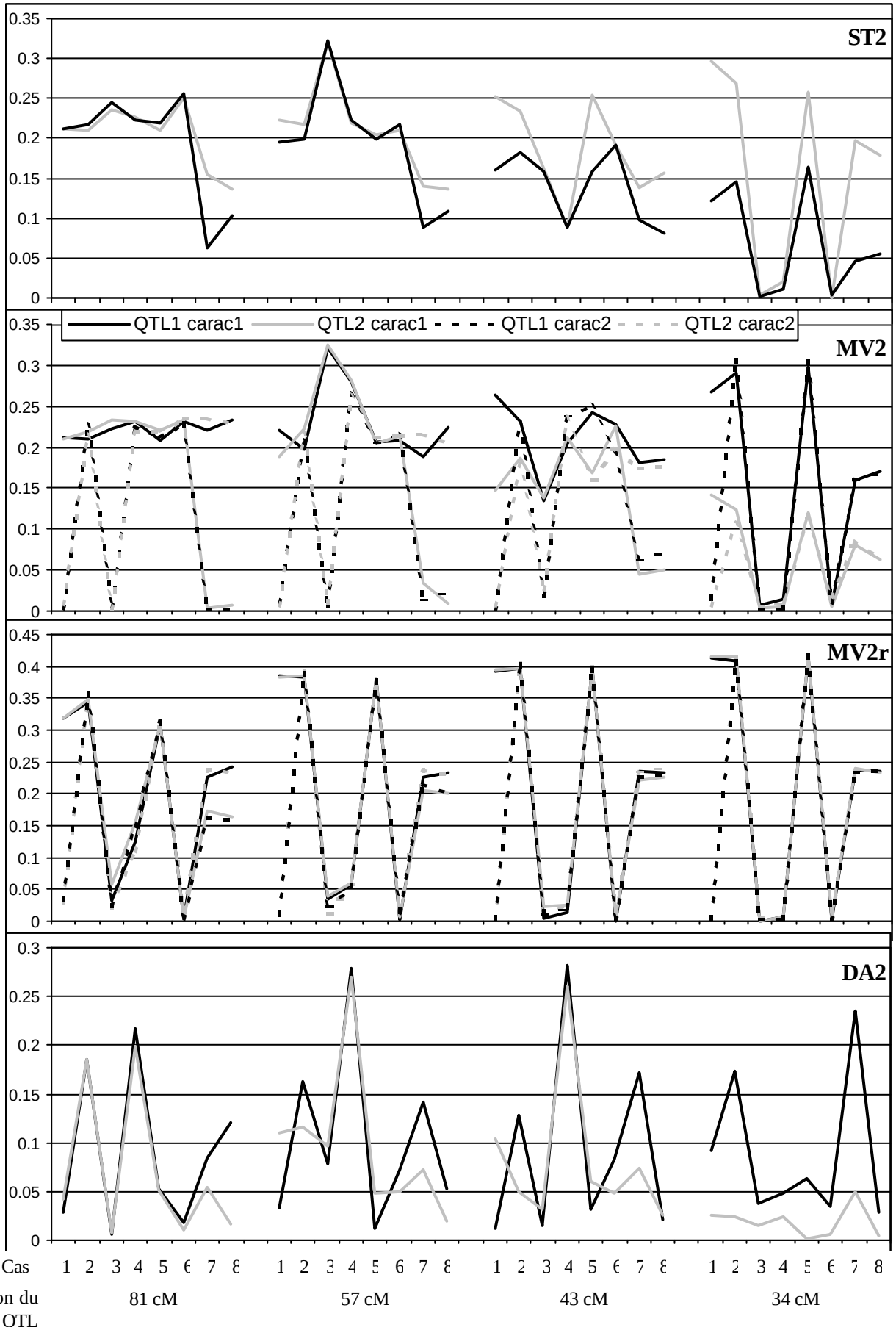


Figure 20 : Moyennes des estimations des effets QTL par les méthodes multiQTL.
L'effet de référence est la moitié de l'effet de substitution, soit 0,25 ou 0 en fonction des cas.
ST2 : méthode unicaractère ; MV2 : méthode multicaractère complète ; MV2r : méthode multicaractère restreinte ; AD2 : analyse discriminante.

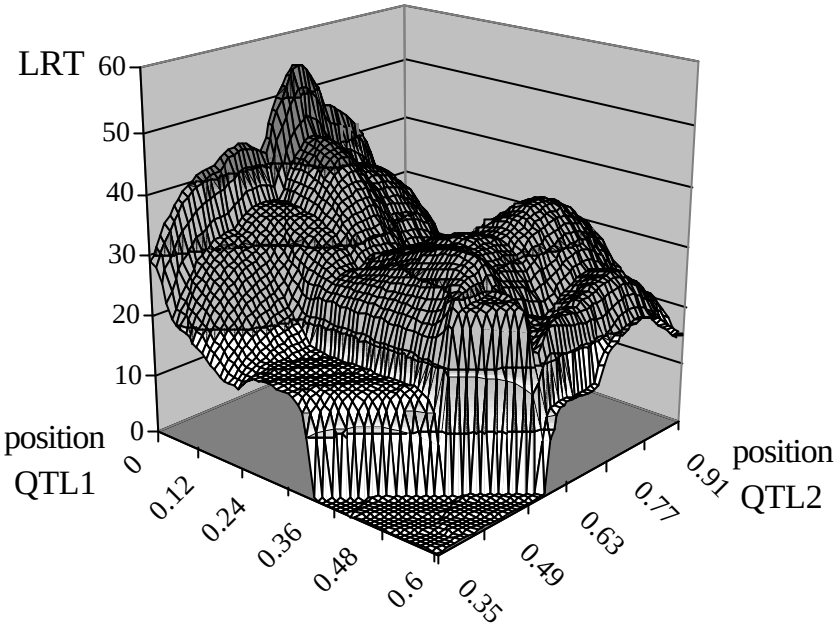


Figure 21 : Analyse multiQTL du poids de panne. Profil de vraisemblance obtenu avec la méthode ST2.

panne + imf

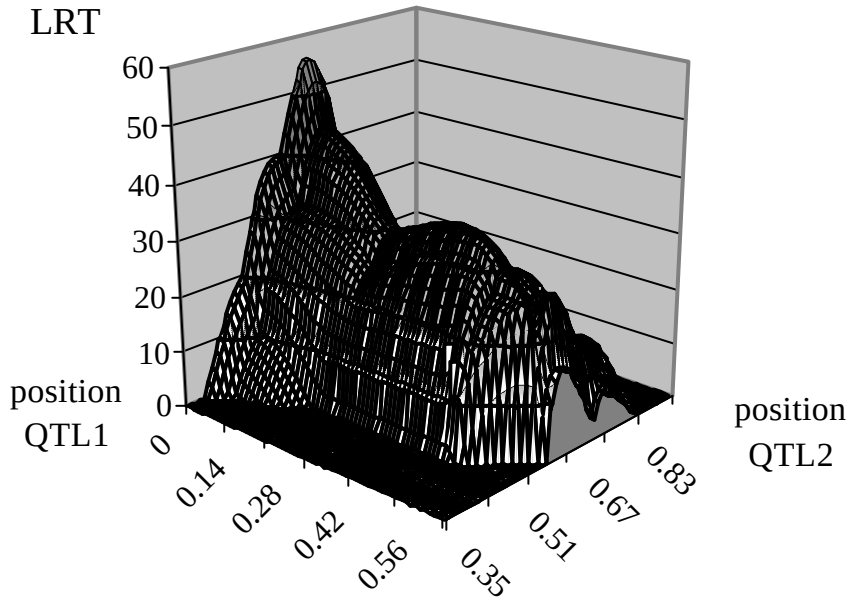


Figure 22 : Analyse multiQTL de panne et imf. Profil de vraisemblance obtenu avec la méthode MV2.
Modèles 2 QTL pour panne et 1 QTL pour imf contre 1 QTL pléiotrope.

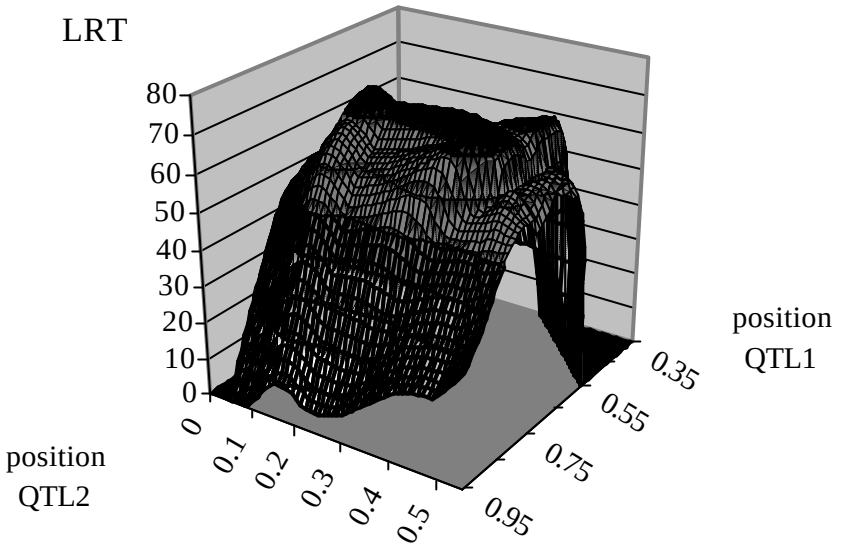


Figure 23 : Analyse multiQTL de x2 et x4. Profil de vraisemblance obtenu avec la méthode MV2. Modèles 2 QTL pléiotropes contre 1 QTL pléiotrope.

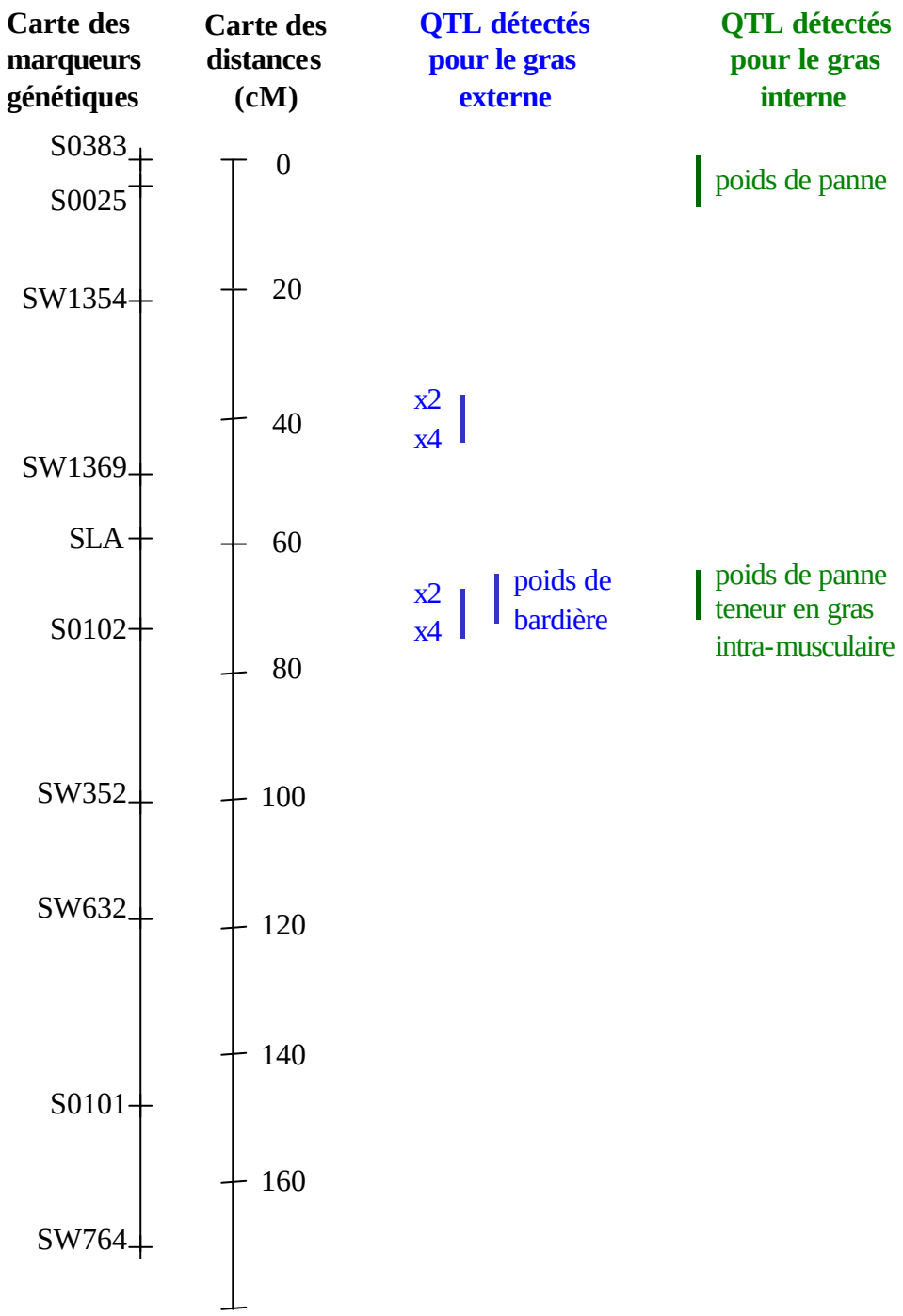


Figure 24 : Positions des QTL mis en évidence sur le chromosome 7 porcin pour 5 caractères de composition corporelle.

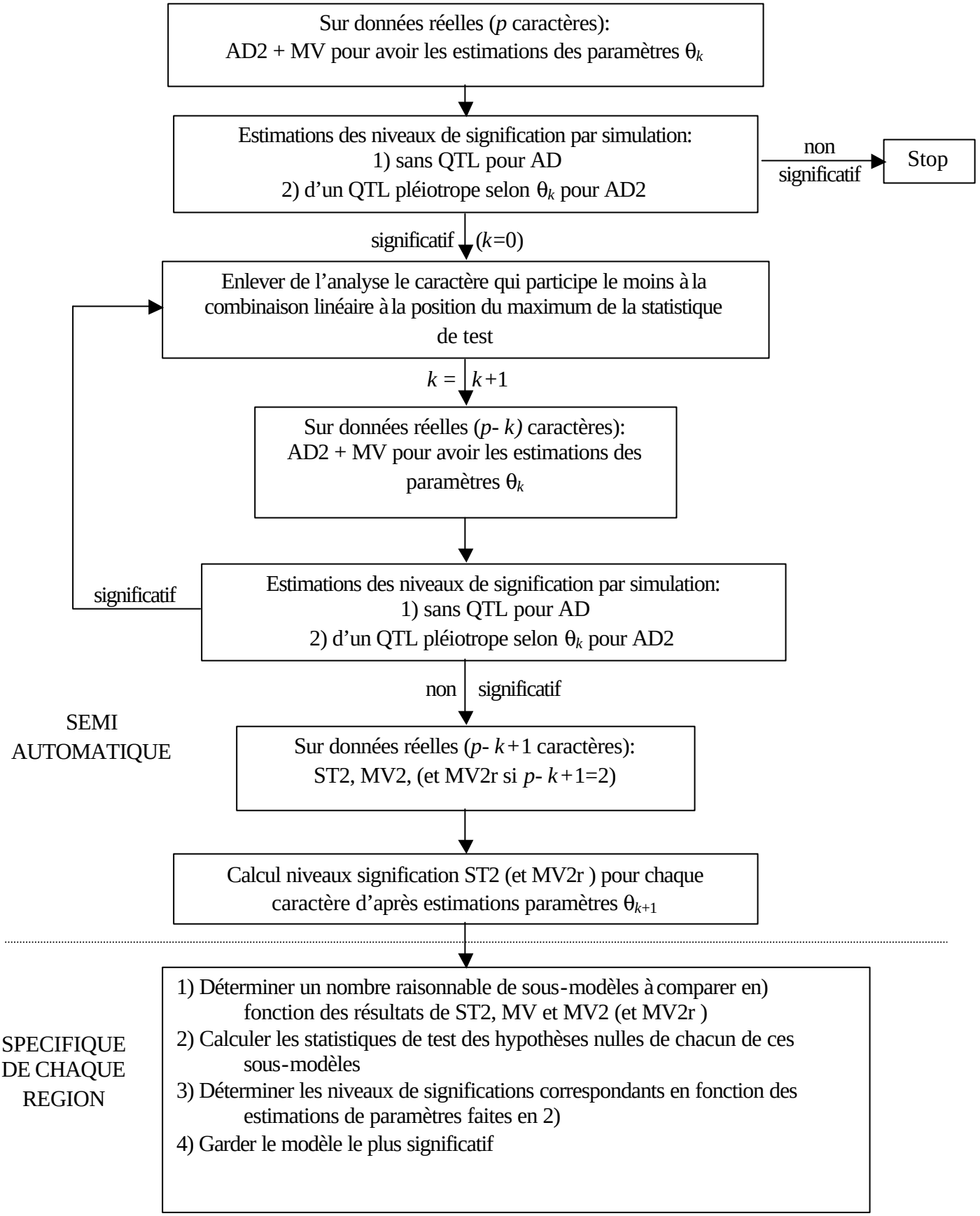


Figure 25 : Stratégie de sélection combinée des caractères et des régions pour la mise en évidence

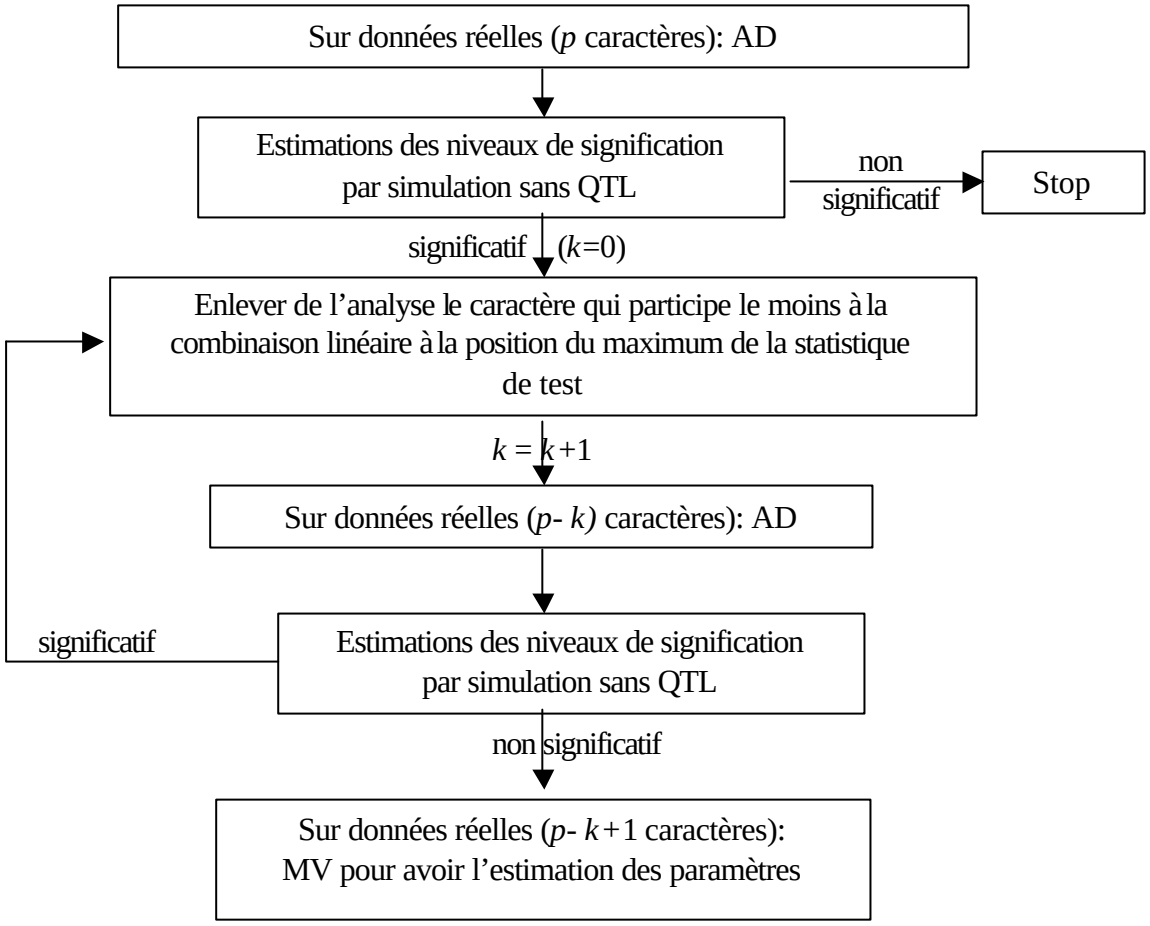


Figure 26 : Stratégie de sélection des régions chromosomiques à effet pléiotrope.