Université de Montréal

Utilisation de triades cas-parents dans la régression logique : exploration d'interaction génétique

par

Steven Sanche

Département de mathématiques et de statistique

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures

en vue de l'obtention du grade de

Maître ès sciences (M.Sc.)
en statistique

décembre 2008

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Utilisation de triades cas-parents dans la régression logique : exploration d'interaction génétique

présenté par

Steven Sanche

a été évalué par un jury composé des personnes suivantes :

Jean-François Angers

(président-rapporteur)

Yves Lepage

(directeur de recherche)

Ridha Joober

(co-directeur)

Alejandro Murua

(membres du jury)

Mémoire accepté le:

SOMMAIRE

La génétique des maladies complexes peut nécessiter l'exploration de l'interaction gène-gène ou épistasie. Au cours des dernières années, plusieurs méthodes ont vu le jour dans le but d'analyser l'épistasie. Ces méthodes dont la régression logique (RL) et la réduction multifactorielle de dimensionalité (RMD), ont été développées et testées sur des plans d'échantillonnage cas/contrôles. Ces plans sont sujets à l'influence de la stratification de la population résultant possiblement en de sérieux problèmes de faux positifs ou négatifs. Les allèles non transmis des parents à l'enfant malade (échantillon de triades) ont été utilisés avec succès pour éviter ce biais lors d'études d'haplotypes pris un à un. Dans ce mémoire nous étudierons la puissance statistique de détection d'interaction ainsi que la robustesse face à la stratification de la population de la RL utilisant des échantillons de triades. Nous avons choisi la RL puisqu'elle capture la représentation intuitive de l'interaction génétique. Afin d'étudier la méthode, nous avons simulé la détection de l'interaction et l'influence de la stratification de population de la RL par 500 simulations de cas/contrôles et de triades pour chacun de trois modèles de maladie avec trois tailles d'échantillons. Par la suite, nous avons comparé la puissance statistique et la fréquence de faux positifs sous chaque plan d'échantillonnage. Il en résulte que les interactions peuvent être identifiées avec les échantillons de cas/contrôles et de triades. Par contre, une plus grande taille d'échantillons est nécessaire pour les triades lorsqu'il n'y a pas de stratification de la population. Si cette dernière est présente, la puissance des deux approches peut s'équivaloir avec l'avantage supplémentaire d'une fréquence plus basse de faux positifs pour les triades. Une application sur des données génétiques de parents et de leur enfant atteint du trouble de déficit de l'attention avec hyperactivité a permis de soulever

une interaction possiblement intéressante. En conclusion, la RL peut détecter l'interaction gène-gène pour des échantillons de triades. Dans l'étude de l'épistasie, échantillonner des triades est recommandable pour éviter les faux positifs, un problème qui continue de nuire aux progrès dans la génétique de maladies complexes.

Mots-clés : *épistasie, régression logique, triades cas-parents, stratification de population*

SUMMARY

Genetics of complex diseases may require exploration of gene by gene interactions or epistasis. In the past few years, several methods have emerged to analyze epitasis. These methods, such as Logic Regression (LR) and Multifactor Dimensionality Reduction were developed and tested using case/control sampling approaches. These approaches may result in serious problems of false positive/negative findings because of population stratification. Indeed, population stratification is a major confounder in analyzing single gene effect, and may be pernicious if stratification is present at two loci presumed to interact. Using non-transmitted parental chromosomes (sample of triads) has been used efficiently to avoid this bias in single haplotype analysis. Here, we investigate statistical power to detect interaction and robustness with regard to population stratification of LR using triads samples. We selected LR to investigate because it may capture the intuitive representation of genetic interactions. For this purpose, we have simulated the detection of interaction and sensitivity to population stratification of LR using 500 simulations of cases/control and triad samples for three models of a disease with three different sizes and compared the statistical power under each sampling scheme. For robustness, we have compared the rate of false positive findings from the case/control and triad simulated samples when population stratification is present. We found that interactions can be identified with both case/control and triad samples. However larger sample sizes are needed when using triads. If population stratification is present, the power of the two sampling schemes may be equivalent with the further advantage of detecting less false positive interactions when triads are used. The method was successfully applied to genotype data in children affected by Attention Deficit Hyperactivity Disorder.

An interaction was detected and needs further investigation. In conclusion, LR can detect gene by gene interaction in triads. Sampling triads is a recommended procedure to avoid false positive results, a problem that has plugged genetics of complex traits.

Keywords :*epistasis, logic regression, case-parents triads, population stratification*

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

REMERCIEMENTS

Je désire remercier Monsieur Ridha Joober de même que Monsieur Yves Lepage, tous deux directeurs de recherche pour le projet présenté dans ce mémoire. Grâce à Monsieur Joober et ses conseils, j'ai découvert le monde inconnu et fascinant de la statistique génétique. Le travail n'aurait également pas pu avoir une telle qualité sans les exigences et mots justes de Monsieur Lepage. Finalement, je désire remercier le soutien financier du Conseil de recherche en sciences naturelles et en génie du Canada (CRSNG).

INTRODUCTION

Encore aujourd'hui, beaucoup de maladies demeurent des mystères pour la médecine. Bien loin de connaître les phénomènes biologiques sous-jacents à ces maladies, il est souvent même difficile de déterminer pourquoi certains individus sont atteints et d'autres pas. En ce qui concerne virus et bactéries pathogènes, on dira qu'il est nécessaire pour que l'individu devienne malade qu'il ait été exposé, dans son environnement, par une source pathogène et que les défenses de son corps n'aient pas résisté suffisamment pour empêcher celle-ci de se développer. En général, les maladies ont deux composantes quant à leur apparition chez un individu : une composante environnementale et une intrinsèque à ce dernier. C'est cette dernière partie qui intéresse les généticiens étudiant une maladie puisque les gènes englobent tout ce qui est inné chez un individu.

Le sujet du présent mémoire est la statistique génétique qui se veut un outil pour l'étude des relations pouvant joindre les gènes et les caractéristiques innées de l'individu. L'objectif de ce mémoire est double : il veut initier toute personne intéressée à la statistique génétique aux concepts biologiques et mathématiques de la génétique ainsi que de présenter les résultats d'une approche nouvelle utilisant des techniques peu connues de sélection de modèle.

Le premier chapitre est introductoire. Il définit les notions principales de la génétique et le vocabulaire qui sera utilisé afin de décrire les événements aléatoires étudiés dans ce mémoire. Ce chapitre permettra aussi de voir comment s'effectuent les calculs de probabilités en génétique, partie essentielle afin de comprendre et analyser les modèles génétiques. Finalement, on décrira le problème de l'association statistique entre gène et maladie. L'objectif du chapitre est d'introduire toute notion nécessaire à la compréhension du mémoire.

Le deuxième chapitre souligne les problèmes auxquels s'attaque le présent mémoire. Le premier problème est qu'il peut exister une association entre un gène et une maladie sans pour autant que ce gène soit impliqué directement ou indirectement avec la maladie. C'est ce qui se produit lorsque la population est stratifiée. Un exemple permettra de comprendre le phénomène. Le deuxième problème est celui de l'interaction entre gènes et sera lui aussi expliqué en grands détails.

Ce n'est qu'au troisième chapitre qu'on sera en mesure d'établir l'objectif du mémoire, obtenir une méthode statistique qui affronte à la fois les deux problèmes mentionnés dans le chapitre deux. La méthode proposée se base sur la sélection de modèle et fera usage de deux méthodes permettant l'exploration de l'interaction entre gènes. On adaptera ces méthodes de façon à éviter l'influence d'une population stratifiée. On verra qu'un type d'échantillon particulier peut être utilisé efficacement en ce sens.

Le quatrième chapitre présente des simulations. Puisque la méthode n'a jamais été explorée antérieurement, il est nécessaire de la tester dans le contexte du problème. On expliquera alors comment on a effectué les simulations et on étalera les résultats obtenus. Les résultats seront discutés au chapitre suivant. Grâce à un théorème original à ce mémoire, on voudra comparer les résultats obtenus aux résultats attendus. Notamment, on expliquera les contextes où l'on croit que la méthode peut être utilisée. Tout ceci dressera le tableau pour une application sur des données réelles composées des génotypes de sujets atteints du trouble de déficit d'attention avec hyperactivité ainsi que de leurs parents. C'est le sujet du sixième chapitre. Des gènes conjointement étudiés seront-ils associés à la maladie ?

À ce point il est important de mentionner que ce mémoire contient possiblement plus d'informations sur la génétique que nécessaire. Toutefois, pour comprendre l'objectif du mémoire explicité au troisième chapitre, le lecteur doit comprendre ce qui motive le choix des méthodes statistiques, le type d'échantillonnage sélectionné ainsi que le type de situation où les méthodes s'avèrent applicables. Il est donc nécessaire de sortir du contexte mathématique et statistique et de plonger dans certaines notions de génétique.

Chapitre 1

NOTIONS DE GÉNÉTIQUE DE POPULATIONS

Un vocabulaire génétique sera utilisé tout au long du mémoire afin de décrire les événements aléatoires étudiés. C'est le sujet de ce premier chapitre. À ce propos, le statisticien ne pourra se débrouiller à l'intérieur du mémoire sans connaître les notions qui y sont définies. La première section décrit la base de la génétique. La deuxième section traite des fréquences des événements aléatoires qui seront d'intérêt : la fréquence des allèles et des génotypes ainsi que leur probabilité dans la population. Il sera aussi nécessaire de comprendre la manière dont le génotype d'un individu se forme à partir du génotype des parents, ce qui permet de formuler la plupart des modèles mathématiques utilisés dans le mémoire. L'équilibre Hardy-Weinberg et de liaison seront ensuite brièvement définis et on y fera référence plus loin afin d'émettre des suppositions sur les modèles. La dernière section traite des maladies génétiques et de l'association entre les gènes et les marqueurs génétiques. Les idées introduites proviennent de l'ouvrage de Lange (2002), livre suggéré afin d'obtenir plus de détails au sujet de la génétique de populations et des mathématiques de celle-ci. En fait, tout bon ouvrage d'introduction à la génétique contient les principaux sujets parcourus relevant de la biologie. Par contre, seuls des ouvrages spécialisés comme Ott (1999) et Terwilliger et Ott (1994) comprennent les modèles mathématiques et statistiques décrits dans ce chapitre. Il est recommandé aux non-initiés de se référer à la figure 1.1 qui illustre les termes définis à l'intérieur de la prochaine section.

1.1. DÉFINITIONS

Tout être vivant pluricellulaire possède dans chacune de ses cellules de l'acide désoxyribonucléique (**ADN**) que l'on croit être l'unique source de toute l'information nécessaire à la construction et au fonctionnement d'un être. Cette structure se présente sous la forme de longues chaînes de petites molécules appelées **nucléotides**. À chaque emplacement sur une chaîne d'ADN ou **locus**, un seul d'une possibilité de 4 nucléotides réside. Certaines suites de nucléotides servent de matrices de codage pour la construction de **protéines**, principaux outils des êtres vivants. Ces suites particulières de nucléotides portent un nom, celui de **gènes**. Un gène peut avoir plusieurs variantes, toutes codant pour une protéine semblable, mais pas toujours en tout point identique. Le nom d'une variante d'un gène est **allèle**. L'interaction entre les multiples variantes de protéines implique différents fonctionnements ainsi que différentes caractéristiques chez un individu. On appelle **phénotype**, tout caractère observable qui résulte directement ou indirectement de l'expression d'un ou plusieurs gènes ; l'expression d'un gène correspond à la production de la structure qu'il code.

Chaque individu possède 46 chromosomes ou chaînes d'ADN, distincts l'un des autres. De plus, chaque chromosome est apparié à un et un seul autre chromosome. Chez l'humain, il y a une paire de **chromosomes sexuels** et 22 paires de **chromosomes homologues**. Les chromosomes homologues possèdent aux mêmes loci (pluriel de locus) les mêmes gènes mais pas nécessairement les mêmes allèles. On appelle **haplotype** toute séquence d'allèles sur un même chromosome. La combinaison des deux variantes de gènes provenant des deux chromosomes, à un locus précis, forme le **génotype** d'un individu en ce locus. Le terme génotype s'applique également si plusieurs loci sont impliqués. L'ensemble des génotypes d'un individu forme son **génome** et détermine toutes les caractéristiques intrinsèques de celui-ci. Si on inspecte un seul locus qui peut contenir plusieurs allèles $a_1, a_2, \ldots, a_n$ pour un entier n quelconque. On dira qu'un individu est homozygote pour l'allèle a_1 si ce dernier possède sur chacun de ses deux chromosomes homologues, une copie de a_1 au locus concerné. Si l'individu possède deux allèles différents à ce locus, on dit qu'il est hétérozygote. Ces termes reviendront dans

les parties évoquant les problèmes génétiques du mémoire.

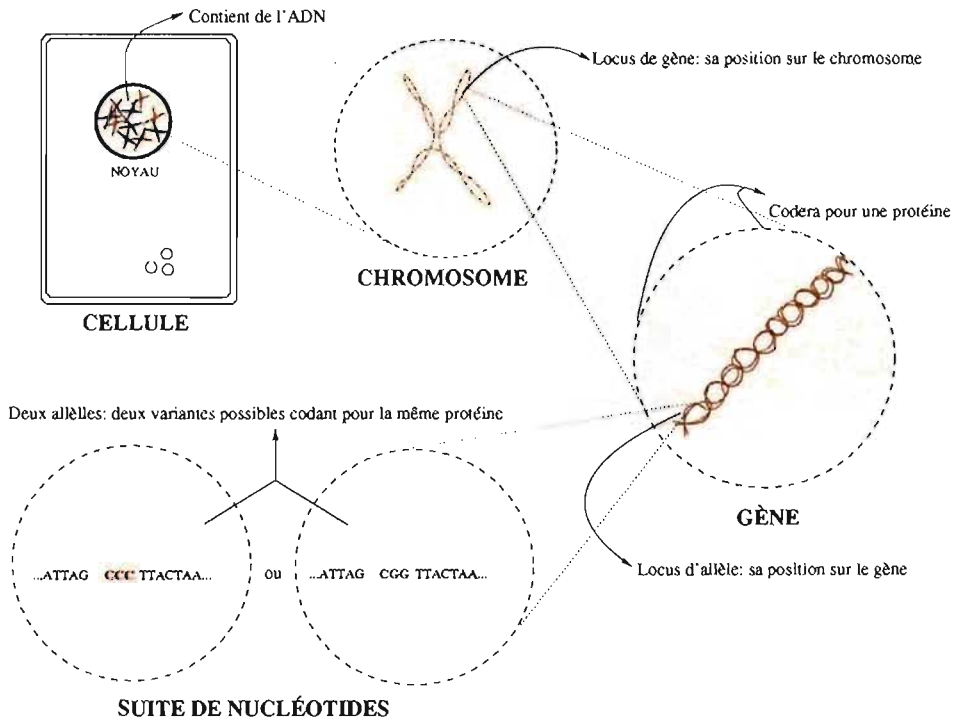

FIG. 1.1. Notions de génétique

1.2. FRÉQUENCES

Le principal événement aléatoire qui sera considéré dans ce mémoire est l'allocution du génotype des individus conditionnellement au génotype des parents et possiblement aussi conditionnellement à une ou plusieurs autres conditions de l'individu. Elle se modélise à partir de la fréquence d'allèle : sur un chromosome C, on définit la **probabilité de l'allèle** a dans la population par la probabilité qu'un chromosome C pris au hasard possède la version du gène (ou allèle) a. La notion de **fréquence d'allèle** a dans une population (fréquence de l'événement décrit à la phrase précédente) reviendra constamment dans ce mémoire. Un exemple de calcul de la **probabilité de génotype g** ou probabilité qu'un

individu pris au hasard dans la population possède le génotype g à un locus, à partir de la probabilité de chaque allèle sera fourni un peu plus loin lors de la discussion sur l'équilibre Hardy-Weinberg.

Afin de mesurer la part du hasard dans l'allocution du génome d'un individu, il est nécessaire de comprendre comment le génome se forme d'une génération à l'autre. La **méiose** est le processus de production de cellules reproductrices chez les individus. La méiose effectue des copies des 2 chromosomes appariés du génome. Les copies des chromosomes ne sont pas nécessairement conformes à cause du phénomène de **recombinaison**. La recombinaison, c'est l'échange de matériel génétique entre les 2 chromosomes homologues d'un individu. Le processus, par ailleurs illustré dans la figure 1.2, est relativement fréquent. La méiose produit deux cellules reproductrices résultant de la séparation de toutes les paires de chromosomes après copie, tel qu'illustré dans la figure 1.3 (homologues et sexuels). Bien sûr, lors de la reproduction, les paires de chromosomes sont reformées : les chromosomes de la cellule reproductrice de l'homme s'assemblent avec ceux de la cellule reproductrice de la femme.

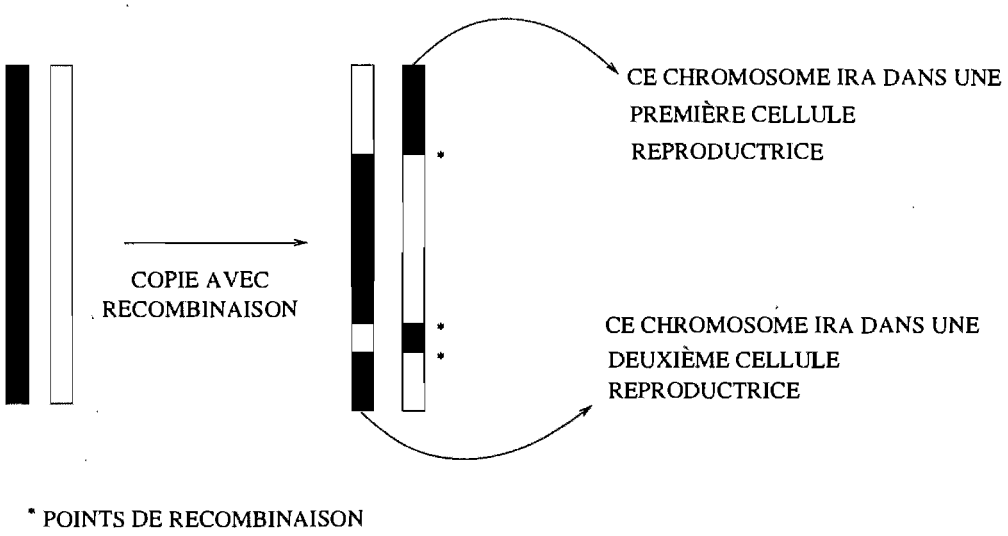

FIG. 1.2. Recombinaison

Pour formaliser le calcul de probabilité de formation de génotypes d'une génération à l'autre, quelques suppositions sont nécessaires. On suppose en tout

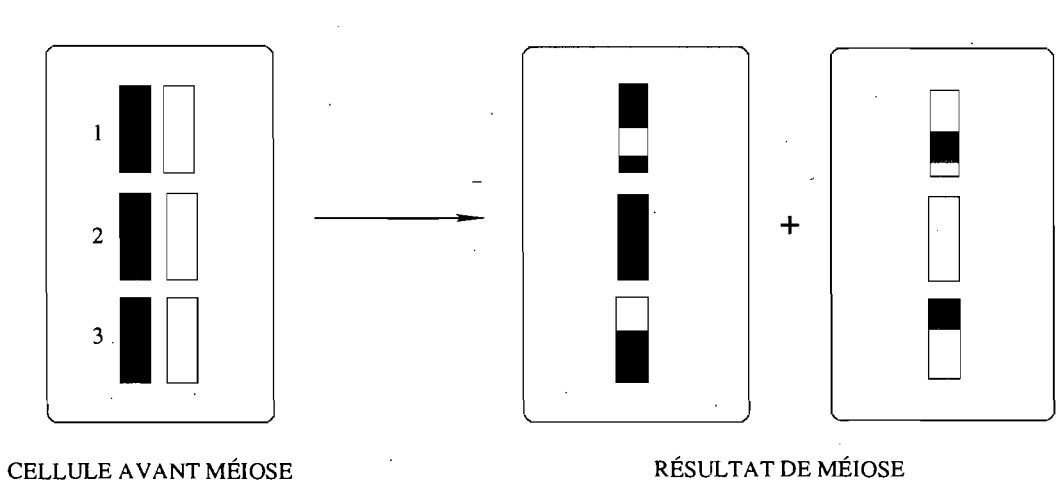

FIG. 1.3. Méiose

premier lieu que la *population étudiée est infinie* (pour des raisons d'homogénéité), que la *reproduction est aléatoire* (un individu se reproduira avec un autre individu de sexe opposé sans que le processus ne dépende des allèles que possèdent les deux individus impliqués), qu'il n'y a pas de reproduction entre membres de générations différentes (*générations discrètes*) et qu'il y a une *fréquence égale des génotypes pour les 2 sexes.* Notons que **la fréquence de génotype g** correspond au nombre d'individus de la population qui possède ce génotype.

1.2.1. Probabilité de transmission d'allèle

Dans le respect des conditions précédentes, une première probabilité importante d'un point de vue statistique et qui sera utilisée au chapitre trois et quatre correspond à la probabilité qu'un génotype particulier soit formé chez un individu par la connaissance des génotypes des parents. On note qu'un parent a exactement une chance sur deux de remettre un chromosome précis à son enfant, de sorte qu'un individu possède exactement une chance sur quatre d'avoir une combinaison précise de deux chromosomes. Tout calcul de probabilité de transmission d'un ou plusieurs allèles des parents à l'enfant se base sur cette notion. Si la transmission des allèles respectent cette loi, on dit que la **transmission est mendélienne**.

1.2.2. Probabilité de génotype et d'haplotype à partir de la probabilité des allèles

Deux types d'équilibre seront souvent mentionnés dans ce mémoire et permettent le calcul de la probabilité d'haplotypes et de génotypes à partir de la probabilité des allèles. Il sera nécessaire de retenir ce que signifient globalement l'équilibre Hardy-Weinberg pour un locus à une génération ainsi que l'équilibre de liaison entre deux locus. L'**équilibre Hardy-Weinberg** correspond à la situation où la probabilité de retrouver n'importe quel génotype se calcule en supposant l'indépendance entre les différents allèles à ce locus. Soyons plus clair, soit un locus bi-allélique d'allèles a_1 et a_2. Posons également que la probabilité de l'allèle a_1 est de 25% et donc 75% pour l'allèle a_2. Alors si l'équilibre Hardy-Weinberg est respecté, un individu pris aléatoirement de cette génération a probabilité de génotype formé de deux a_1 de $(0.25)^2$, probabilité de génotype formé d'un a_1 et d'un a_2 de $2(0.25)(0.75)$ et probabilité de génotype formé de deux a_2 est de $(0.75)^2$. Le calcul a supposé l'indépendance de l'allèle au premier chromosome de l'individu par rapport à l'allèle au second chromosome homologue. Plusieurs situations peuvent générer un déséquilibre Hardy-Weinberg pour un locus d'une génération mais elles ne seront pas mentionnées ici. Le second type d'équilibre, nommément l'**équilibre de liaison** entre deux loci, est un équilibre similaire mais pour les haplotypes au lieu des génotypes. Autrement dit, il y a déséquilibre de liaison entre deux loci si un allèle au premier de deux locus a tendance a être situé sur le même chromosome qu'un autre allèle a un deuxième locus. Donc, la présence d'un allèle particulier au premier locus nous informe sur la présence d'un second allèle situé ailleurs sur le chromosome. Plus un locus est situé physiquement près d'un deuxième locus, plus on risque d'y noter un déséquilibre de liaison. En ce sens, recenser un locus nous renseigne aussi sur les allèles autour de lui. C'est sur cette notion que se base l'idée de marqueur génétique. Un exemple de calcul de probabilité de génotype est fourni dans la figure 1.4

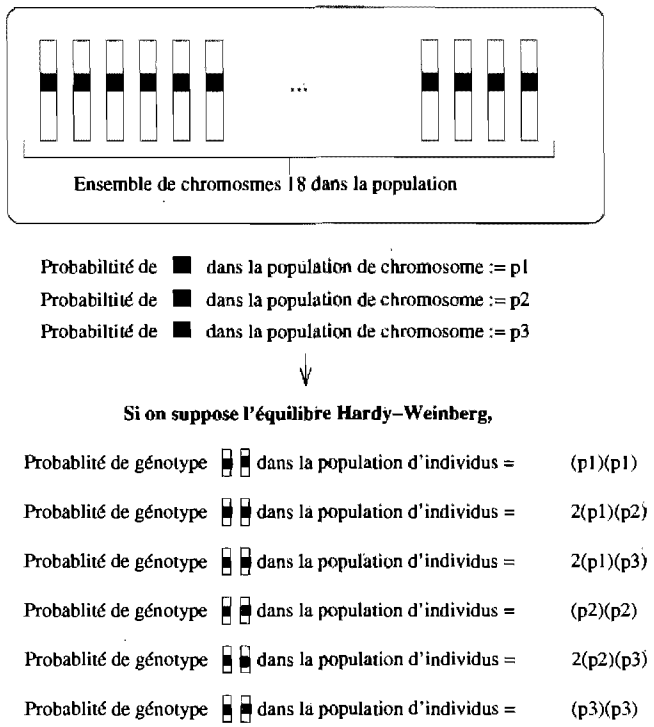

FIG. 1.4. Exemple de calcul de probabilité de génotype dans une population

1.3. MALADIES ET GÈNES

Si les gènes codent pour des protéines, que se passe-t-il si une anomalie sur un ou plusieurs de ces gènes est présente chez un individu ? Ces anomalies se répercutent directement dans la formation de protéines. Résultat : elles peuvent parfois à elles seules être la cause de maladies. Les **maladies mendéliennes** sont de telles maladies : la présence ou l'absence d'un génotype à l'emplacement d'un seul gène détermine complètement si l'individu sera atteint ou non de l'affection. Par contre, fréquemment c'est la combinaison de gènes qui amène une susceptibilité accrue chez l'individu envers une affection. C'est le cas de **maladies complexes**. Qui plus est, posséder une combinaison de gènes particulière ne détermine souvent pas complètement si l'on sera atteint de la maladie : il est parfois nécessaire d'être soumis à certaines conditions de notre environnement. Comment déterminer quels gènes sont impliqués dans la maladie ? C'est là qu'intervient la statistique génétique et les études d'association.

On parlera dans ce qui suivra du concept de **pénétrance du génotype g** que l'on définit par la probabilité qu'un individu possédant le génotype g soit affecté par celle-ci (qui s'écrit mathématiquement par $P(M = 1|G = g)$). Afin d'étudier quels sont les génotypes qui ont des pénétrances différentes, les généticiens se basent sur quelques idées (voir Lange (2002)). Ils supposent qu'une **mutation** a lieu dans une population. Notons qu'une mutation est l'altération d'un gène de façon naturelle lors de la méiose et est un phénomène relativement rare. Les généticiens supposent également que *cette mutation augmente la susceptibilité d'un individu* face à une maladie ou un trait (autre phénotype). Lorsque ce gène muté se transmet d'une génération à l'autre, il est accompagné d'allèles qui se situaient physiquement près de lui lors de la mutation : les allèles près du locus de la mutation sont en déséquilibre de liaison avec la mutation. Plus un allèle est physiquement près de la mutation sur le chromosome, plus il nécessitera du temps avant que la présence de la mutation et de cet autre allèle dans une cellule reproductrice soient des phénomènes indépendants. Afin de localiser une région du génome qui a été affectée par la mutation, les études dites d'**association** tirent profit du déséquilibre de liaison : il est possible de recenser une région du génome par la connaissance d'un seul allèle dans cette région. Les études d'association allélique utilisent ainsi des **marqueurs génétiques** : des allèles facilement reconnaissables du génome et répartis à travers celui-ci. Ces études permettent d'éviter la tâche (pour le moment) beaucoup trop fastidieuse de recenser tout le génome.

Les toutes premières étapes des études d'association servent à estimer les régions de chromosomes où pourrait être situé un allèle qui augmente le risque (ou le diminue) d'être affecté par un trait (ou maladie). Pour se faire, les analyses tirent profit d'une mesure : l'**association génétique** entre un allèle et une maladie. Afin d'illustrer le concept, soit un locus A d'un marqueur génétique sur un chromosome. Soient également deux allèles de ce marqueur notés a_1 et a_2. Ces allèles sont les deux variantes du gène situé au marqueur A. On dit alors que le marqueur A est bi-allélique. Notons que toutes les notions développées ici se généralisent au cas de plus de deux allèles. Soit également un autre locus D

pouvant abriter un allèle de susceptibilité pour le trait, que l'on notera d_1 chez les individus d'une population cible, c'est-à-dire que la présence de d_1 dans le génotype d'un individu augmente directement la susceptibilité de ce dernier face à la maladie étudiée. Alors, on dit qu'il existe une association entre le marqueur A et la maladie dans une population s'il existe une relation de dépendance entre la présence de d_1 et celle d'un des allèles du marqueur A chez les individus de cette population. Pour chaque allèle d'un marqueur, on calcule une valeur δ appelée mesure d'association, pour mesurer l'association entre ce marqueur et le marqueur de susceptibilité d_1. Si l'allèle est a_1, cette valeur correspond à

$$\delta = P_c(a_1 d_1) - P_c(a_1)P_c(d_1) \tag{1.1}$$

où on rappelle que P_c représente une probabilité de présence d'allèle dans l'ensemble des cellules reproductrices et on omet le paramètre de temps puisqu'une seule génération est observée. On dit qu'il existe une association entre le locus A et la maladie si la mesure d'association est relativement éloignée de zéro pour un des allèles du locus. Plusieurs raisons peuvent expliquer pourquoi un allèle est associé à la maladie dont notamment le déséquilibre de liaison. Si c'est le cas, on s'est rapproché d'un gène de susceptibilité pour la maladie : il existe dans la région du locus A, un allèle qui augmente la susceptibilité d'un individu pour la maladie.

Afin de mesurer l'association entre un marqueur et un allèle de susceptibilité pour la maladie, une première approche très simple consiste à tester l'hypothèse correspondant à l'égalité de la fréquence des différents allèles au marqueur chez les individus atteints de la maladie et celle des individus non-affectés par la maladie. Telle une étude de cohorte, l'idée correspond à prendre des individus au hasard (indépendamment de l'affection des individus) et de noter l'affection des individus ainsi que la fréquence des allèles au loci du marqueur (l'exposition dans les études de cohorte). La question est la suivante : existe-t-il une différence de la fréquence d'allèles entre les cas et les contrôles ? Les méthodes usuelles de biostatistique relatant une exposition à une maladie peuvent alors être utilisées afin de déterminer s'il existe une association entre la maladie et l' "exposition génétique".

Par contre, il est possible qu'il y ait association d'un marqueur avec un locus de susceptibilité pour la maladie sans qu'il n'existe de déséquilibre de liaison entre les deux loci. Il est également possible qu'il existe déséquilibre de liaison sans association. Ce premier cas est problématique puisque le but premier des études d'association est de déterminer des régions du génome abritant des gènes de susceptibilité pour la maladie. S'il existe une association génétique entre un marqueur et une maladie sans qu'il n'y ait de déséquilibre de liaison, nous n'avons pas plus d'information sur la localisation de l'allèle de susceptibilité. Également, cette méthode ne donne aucune information sur toute interaction entre gènes et leur rôle dans la maladie. Ces deux problématiques seront illustrées au prochain chapitre et motivent le présent mémoire.

Chapitre 2

PROBLÈMES : STRATIFICATION DE LA POPULATION ET INTERACTION ENTRE GÈNES

Toutes les notions de génétique nécessaire à la compréhension de ce mémoire ont été explicitées dans le chapitre précédent. Il est maintenant possible de décrire avec plus de précision le but de ce mémoire : l'exploitation de méthodes dans le but de détecter une association entre des gènes interagissant de manière à influencer la susceptibilité d'un individu porteur de ces gènes. On parlera de tests d'association, c'est-à-dire des tests statistiques pour la mesure d'association δ du chapitre précédent entre un seul allèle (ou un seul marqueur) et la maladie. Afin de concrétiser l'idée, le lecteur pourra imaginer une statistique khi-deux pour un tableau de contingence du type exposition génétique. Par exemple, dans l'étude d'un allèle a provenant d'un locus A, on pourrait recenser le nombre d'allèles a que chaque cas et chaque contrôle possède dans son génotype, c'est-à-dire zéro si le génotype de l'individu en question ne possède pas d'allèle a, un si l'individu en question est hétérozygote pour l'allèle a, et deux si celui-ci est homozygote pour l'allèle a. Un tableau de contingence se forme à partir de telles données avec deux colonnes pour l'affection de l'individu (cas ou contrôle) et trois rangées pour le nombre d'allèles a, (0, 1 ou 2). Comme on l'a abordé lors du chapitre précédent, on voudra éviter de détecter des gènes associés à la maladie qui sont éloignés physiquement d'un gène de susceptibilité (pas en déséquilibre de liaison avec un tel gène). On verra qu'un problème guette les méthodes qui font

usage d'échantillons cas-contrôle afin d'évaluer l'association : le problème de la stratification de la population. On illustrera aussi ce qui motive l'idée d'associer à une maladie un groupe de gènes conjointement plutôt qu'un seul à la fois. Pour atteindre de tels objectifs, il est nécessaire d'approfondir le problème génétique. Dans la dernière section du chapitre, on trouve un résumé des points essentiels de cette problématique qui peut sembler aride pour des non-initiés à la génétique.

2.1. Stratification de la population

Il est possible qu'il y ait une différence dans la probabilité d'allèle a entre les cas et les contrôles, c'est-à-dire une association génétique théorique entre la maladie et cet allèle ($\delta \neq 0$), sans qu'il n'y ait déséquilibre de liaison avec un allèle augmentant la susceptibilité d'affection. La **stratification** de la population peut en être la cause. Illustrons la situation par un exemple.

Soit une population hétérogène formée de deux sous-populations de même dénombrement qui ne se reproduisent qu'à l'intérieur de leur sous-population. On suppose l'équilibre Hardy-Weinberg à l'intérieur des deux sous-populations. On mesurera la grandeur de l'association par le phénotype de maladie. Le problème survient lorsque la probabilité des allèles diffère d'une sous-population à l'autre. Si c'est le cas, on dit que la population est stratifiée du point de vue des allèles. Soit un marqueur A d'allèles a_1 et a_2. Soit également une maladie due à un allèle d_1 situé sur un marqueur D situé sur un autre chromosome. Soit finalement des probabilités différentes d'allèles $P(d_1)$, $P(a_1)$ et $P(a_2)$ dans les deux sous-populations respectives selon le tableau 2.1. Admettons que la présence de l'allèle

Tab. 2.1. Tableau de fréquences alléliques dans les deux sous-populations

	Prob. de a_1	Prob. de a_2	Prob. de d_1
Population 1	0.50	0.50	0.50
Population 2	0.05	0.95	0

d_1 chez un individu provoque automatiquement la maladie. On doit tout d'abord définir les événements $N(a_1) = 2$, $N(a_1) = 1$, $N(a_1) = 0$ pour signifier que l'individu possède deux, un ou bien aucun allèle a_1 dans son génotype. Soit également

$N(d_1) = 2$, $N(d_1) = 1$ et $N(d_1) = 0$ l'analogue pour l'allèle d_1. Finalement, soit $M = 1$ ou 0 indiquant si l'individu est un individu malade ou pas et $Pop = 1$ ou 2 pour indiquer si ce dernier provient de la première ou bien de la deuxième sous-population.

Notons qu'un individu malade correspond a un individu tel que $N(d_1) = 1$ ou $N(d_1) = 2$. Dans cette situation hypothétique, pour la première population on a

$$\begin{aligned} P(N(d_1) = 1|Pop = 1) &= P(\text{hétérozygote en } d_1|Pop = 1) \\ &= 2P(d_1)(1 - P(d_1)) \\ &= 0.5 \end{aligned} \tag{2.1}$$

et

$$\begin{aligned} P(N(d_1) = 2|Pop = 1) &= P(\text{homozygote en } d_1|Pop = 1) \\ &= P(d_1)^2 \\ &= 0.25. \end{aligned} \tag{2.2}$$

Ce calcul repose sur l'exemple de calcul de probabilité de génotype à partir de la probabilité des allèles du chapitre 1. Donc, c'est 75% de la première population qui est affectée par la maladie. Pour la deuxième population, puisqu'il n'y a pas d'individu possédant l'allèle d_1, aucun individu ne peut être affecté de la maladie. Si on prend des individus au hasard dans la population composée de ces deux dernières, l'individu malade provient nécessairement de la première population. La probabilité qu'un individu sain provienne de la première population est

$$\begin{aligned} P(Pop = 1|M = 0) &= \frac{P(M = 0|Pop = 1)P(Pop = 1)}{P(M = 0)} \\ &= \frac{(0.25)(0.5)}{1 - 0.375} = 0.2 \end{aligned} \tag{2.3}$$

puisque $P(Pop = 1)$ correspond à 0.5, la proportion d'individus provenant de la première population,

$$\begin{aligned} P(M = 0|Pop = 1) &= 1 - P(N(d_1) = 1|Pop = 1) - \\ &\quad P(N(d_1) = 2|Pop = 1) \\ &= 0.25 \end{aligned} \tag{2.4}$$

et finalement

$$\begin{aligned} P(M = 1) &= P(M = 1|Pop = 1)P(Pop = 1) + \\ &\quad P(M = 1|Pop = 2)P(Pop = 2) \\ &= P(M = 1|Pop = 1)P(Pop = 1) \\ &= (0.75)(0.5) = 0.375. \end{aligned} \tag{2.5}$$

L'individu sain a donc seulement 20% de chance de provenir de cette même population. On en déduit que

$$\begin{aligned} P(N(a_1) = 1|M = 1) &= P(N(a_1) = 1, Pop = 1|M = 1) + \\ &\quad P(N(a_1) = 1, Pop = 2|M = 1) \\ &= P(N(a_1) = 1, Pop = 1|M = 1) \\ &= P(N(a_1) = 1|Pop = 1, M = 1)P(Pop = 1|M = 1). \end{aligned} \tag{2.6}$$

Puisque le fait de posséder zéro, un ou bien deux allèles a_1 n'a pas d'influence sur le risque d'être malade à l'intérieur d'une même sous-population et qu'il n'existe pas de déséquilibre de liaison entre les loci A et D dans cette population, on a finalement

$$\begin{aligned} P(N(a_1) = 1|M = 1) &= P(N(a_1) = 1|Pop = 1) \\ &= P(\text{Individu hétérozygote}|Pop = 1) \\ &= 2P(a_1)(1 - P(a_1)) \\ &= 0.5. \end{aligned} \tag{2.7}$$

Un calcul semblable nous permet d'obtenir que $P(N(a_1) = 2|M = 1) = 0.25$. On retient que la probabilité qu'un individu malade ait un ou deux allèles a_1 est de

75%. Cette même quantité se calcule pour les contrôles, on a

$$P(N(a_1) = 1|M = 0) =$$
$$P(N(a_1) = 1, Pop = 1|M = 0) + P(N(a_1) = 1, Pop = 2|M = 0)$$
$$= \sum_{i=1,2} \frac{P(M = 0|N(a_1) = 1, Pop = i)P(N(a_1) = 1|Pop = i)P(Pop = i)}{P(M = 0)}$$
$$= \sum_{i=1,2} \frac{P(M = 0|Pop = i)P(N(a_1) = 1|Pop = i)P(Pop = i)}{P(M = 0)} \quad (2.8)$$

puisque le nombre d'allèles ne donne aucune information sur la probabilité d'être malade si on connaît la population de provenance. Tous les termes ont été calculés explicitement antérieurement à l'exception de $P(N(a_1) = 1|Pop = 2)$. Cette expression se calcule facilement par la probabilité d'allèle dans la deuxième sous-population. Elle vaut

$$P(N(a_1) = 1|Pop = 2) = 2(0.95)(0.05)$$
$$= 0.0925. \quad (2.9)$$

Un calcul semblable permet d'obtenir des valeurs pour $P(N(a_1) = 2|M = 0)$. La somme de ce terme avec $P(N(a_1) = 1|M = 0)$ rapporte 0.228.

Cette différence dans la fréquence d'allèles a_1 entre cas et contrôles signifie une association entre le marqueur étudié et la maladie. Notons que le moyen de mesurer la grandeur de δ n'est pas en cause, puisqu'on montre que

$$\delta = 0.25P(a_1|Pop = 1)P(d_1|Pop = 1) + 0.25P(a_1|Pop = 2)P(d_1|Pop = 2)$$
$$- 0.25P(a_1|Pop = 1)P(d_1|Pop = 2) - 0.25P(a_1|Pop = 2)P(d_1|Pop = 1)$$
$$= 0.05625 > 0. \quad (2.10)$$

Pourtant, on n'a donné aucune précision sur la proximité du marqueur A par rapport à celui du marqueur D : il n'y aucune raison de croire qu'il y a un déséquilibre de liaison entre ces deux marqueurs dans les deux populations sous-jacentes. Si la population se reproduit de manière homogène, sans préférence pour une population ou l'autre, la fréquence des allèles aux marqueurs A ainsi

qu'au locus D tenderaient vers une valeur d'équilibre, ce qui enrayerait la stratification allélique de la population (voir Lange (2002)). À l'équilibre, les probabilités $P(a_1 = 1, Pop = 1)$ et $P(a_1 = 1, Pop = 2)$ s'équivalent de même que $P(M = 1|Pop = 1)$ et $P(M = 1|Pop = 2)$ puisque les allèles aux deux marqueurs en question sont répartis de manière homogène dans les deux sous-populations. On notera que, sous ces conditions, les équations (2.6) ainsi que (2.8) deviennent équivalentes. On peut également déduire que la stratification peut affecter les résultats des analyses mais pas si les cas et les contrôles proviennent de la même sous-population.

2.2. Interaction entre gènes

L'autre concept d'intérêt est l'interaction entre gènes. L'interaction entre loci ou gènes de loci différents, c'est ce qui se produit si l'effet d'un génotype à un premier locus (donc une combinaison de deux gènes à ce locus) dépend du génotype présent à un deuxième locus. Par exemple, on peut considérer un gène qui s'exprime et affecte la maladie mais seulement si un deuxième gène situé à un autre locus est présent ou absent, ce qui peut se produire si ce deuxième gène code pour une protéine qui inhibe la transcription du premier gène. La situation inverse est également envisageable avec un deuxième gène qui accentue la transcription d'un premier gène. Il est également possible qu'un troisième (quatrième, cinquième, etc.) gène soit impliqué dans le processus.

Un exemple d'interaction entre deux gènes à des loci différents est illustré dans le tableau 2.2. En effet, dans ce dernier tableau, on voit qu'un seul gène ne

Tab. 2.2. Phénotype de couleur de fourrure de souris en fonction de son génotype à un locus du chromosome 9

Génotype locus G → Génotype locus B ↓	g_1g_1	g_1g_2	g_2g_2
b_1b_1	Blanc	Gris	Gris
b_2b_1	Noir	Gris	Gris
b_2b_2	Noir	Gris	Gris

détermine pas entièrement la couleur de fourrure de la souris. Une souris avec au locus G au moins une copie de l'allèle g_2 aura une fourrure grise, mais une souris avec deux allèles g_1 au locus G peut avoir une fourrure blanche ou bien noire, dépendant des allèles au locus B. En fait et jusqu'à maintenant, plusieurs phénotypes ont révélé une structure d'interaction de gènes sous-jacents. Entre autres, le phénotype BOMBAY chez l'humain est l'un des phénotypes qui est le résultat direct de l'interaction entre gènes (voir Carlborg et Haley (2004)).

Du point de vue de la maladie, on dira qu'il y a interaction entre gènes situés à deux loci différents si la pénétrance du génotype au premier locus dépend du génotype au deuxième locus. De façon générale, le tableau 2.3 illustre la pénétrance d'une maladie pour deux loci bi-alléliques. Ce tableau est inspiré directement de Wilson (2001). Les génotypes sont notés de façon à ce que (a_1a_1,b_1b_2) représente, par exemple, un génotype composé de deux allèles a_1 au premier locus alors qu'au deuxième locus se situent une copie de l'allèle b_1 et une copie de l'allèle b_2. On définit $\mathrm{f} = P(M = 1|G = (a_1a_1, b_1b_1))$ ainsi que les $e^{\beta_{i,j}}$, pour i et j valant zéro, un ou deux et qui représentent les risques relatifs avec comme référence le génotype (a_1a_1, b_1b_1). Notons qu'en théorie, chaque case du tableau précédent peut abriter des valeurs différentes. En pratique par contre, on a observé jusqu'à présent que quelques structures plus restreintes de pénétrance lorsqu'il y a interaction entre gènes (voir à nouveau Carlborg et Haley (2004)).

TAB. 2.3. Pénétrance du phénotype de maladie en fonction du génotype, situation générale

Génotype au premier locus $\rightarrow$ Génotype au deuxième locus $\downarrow$	a_1a_1	a_1a_2	a_2a_2
b_1b_1	f	$\mathrm{f}e^{\beta_{0,1}}$	$\mathrm{f}e^{\beta_{0,2}}$
b_1b_2	$\mathrm{f}e^{\beta_{1,0}}$	$\mathrm{f}e^{\beta_{1,1}}$	$\mathrm{f}e^{\beta_{1,2}}$
b_2b_2	$\mathrm{f}e^{\beta_{2,0}}$	$\mathrm{f}e^{\beta_{2,1}}$	$\mathrm{f}e^{\beta_{2,2}}$

On remarque qu'en comparant les fréquences d'un seul allèle à la fois entre cas et contrôles, on ne peut pas obtenir d'information sur l'interaction. En fait, pour la plupart des méthodes actuelles, il n'est possible de détecter que des marqueurs

en déséquilibre de liaison avec un allèle de susceptibilité envers la maladie ou bien ces allèles mêmes mais il en demeure impossible de connaître la nature des interactions entre ceux-ci si elles existent. Ces méthodes étudient la possibilité qu'un seul marqueur à la fois soit en association ou en déséquilibre de liaison avec un gène de susceptibilité. On peut alors se questionner sur la pertinence de détecter l'interaction entre gènes dans l'étude de traits phénotypiques si ces méthodes permettent de détecter tous les gènes impliqués dans la maladie. Or, la littérature est submergée d'articles qui en soutiennent l'importance. Les principales raisons soulevées concernent le manque de réplicabilité d'expériences identiques, les difficultés rencontrées lors d'études sur des organismes modèles et qui se concentrent sur un seul des gènes détectés lors d'études précédentes, la capacité à détecter les gènes impliqués dans les traits (ou puissance statistique) ainsi que la compréhension des phénomènes biologiques sous-jacents. Chacun de ces points sera exploré en plus de détails dans les paragraphes qui suivent.

Le premier des problèmes cité plus haut est le manque de réplicabilité d'expériences identiques. Il arrive fréquemment que des tests d'association entre un allèle ou locus à une maladie aient des résultats positifs lors d'une expérience mais se révèlent négatifs lors d'expériences reproduisant cette première. Est-ce que ce résultat était en fait un faux positif? Pas nécessairement car plus un génotype possède un ratio de pénétrance près de 1 par rapport aux génotypes différents, plus la puissance de détection de l'association définie comme la probabilité qu'une méthode détecte une association entre marqueurs (ou allèle) et la maladie, est petite pour la plupart des méthodes évaluant l'association (Wilson (2001), Zondervan et Cardon (2004)). Or il existe une relation directe entre la pénétrance marginale d'un gène (marginale puisque le gène est considéré seul dans l'analyse) et la fréquence des allèles avec lesquels celui-ci est en interaction. Afin d'illustrer ceci, posons une situation hypothétique : soient deux échantillons d'individus provenant de deux populations distinctes. Soient également deux loci en équilibre de liaison qui sont impliqués en interaction dans la maladie. On pose que le premier locus abrite deux allèles, les allèles a_1 et a_2 alors qu'au deuxième locus peuvent siéger les allèles b_1 ou b_2. Admettons que les pénétrances pour la

maladie soient telles que dans le tableau 2.4, c'est-à-dire que seule la combinaison

TAB. 2.4. Pénétrance du phénotype de maladie en fonction du génotype, situation générale

Génotype au premier locus → Génotype au deuxième locus ↓	a_1a_1	a_1a_2	a_2a_2
b_1b_1	0.5	0.1	0.1
b_1b_2	0.1	0.1	0.1
b_2b_2	0.1	0.1	0.1

d'allèles (a_1a_1,b_1b_1) est responsable de l'augmentation du risque d'être atteint de la maladie. Supposons que dans la première population, la fréquence de l'allèle b est p_{1b_1} alors que dans la deuxième population, cette probabilité est p_{2b_1}. Supposons également que les fréquences d'allèles a_1 et a_2 ne diffèrent pas entre les deux populations. Quelles seront alors les pénétrances marginales dans les deux populations ? Dans la population i, on a

$$\begin{aligned} P(M=1|G=(a_1a_1)) &= P(M=1|G=(a_1a_1,b_1b_1))P(G=b_1b_1)+ \\ &\quad P(M=1|G=(a_1a_1,b_1b_2))P(G=b_2b_1)+ \\ &\quad P(M=1|G=(a_1a_1,b_2b_2))P(G=b_2b_2) \\ &= 0.5p_{ib_1}^2 + 0.1(1-p_{ib_1}^2) \end{aligned} \tag{2.11}$$

car les probabilités $P(G=b_1b_1)$ et $P(G=b_2b_1)$ ou $P(G=b_2b_2)$ valent respectivement $p_{ib_1}^2$, $2p_{ib_1}(1-p_{ib_1})$ et $(1-p_{ib_1})^2$ dans la population i. Notons également que les probabilités d'être affecté par la maladie dans le cas de génotypes différents de (a_1a_1) sont égales dans les deux populations. Alors le ratio des pénétrances pour le génotype (a_1a_1) par rapport aux génotypes différents de (a_1a_1) sera différent entre les deux populations dû aux fréquences des allèles au deuxième locus. Ceci veut également dire que malgré que la fonction de pénétrance de la maladie dépende du génotype de manière identique dans les deux populations, la puissance de détection de ce génotype peut différer grandement si on ne considère pas les allèles du deuxième locus avec lesquels le premier locus est en interaction. La puissance des méthodes considérant un seul allèle à la fois est donc sensible à

la stratification des populations. Une méthode statistique permettant de rendre compte de l'interaction serait profitable à cet effet puisque la pénétrance étudiée ne serait plus la pénétrance marginale qui dépend de la fréquence des allèles et par conséquent, il y aurait une plus grande uniformité des résultats quant à l'effet des génotypes sur la maladie.

On comprend de ce dernier paragraphe qu'il est possible que la détection soit plus difficile si on étudie les marqueurs seuls à seuls par rapport à l'analyse de plusieurs loci en même temps puisque la pénétrance marginale est une moyenne des pénétrances conjointes. Par contre, on doit noter que la puissance (telle que définie plus haut) de la plupart des méthodes explorées dépend également du nombre d'individus dans chaque catégorie de génotype considéré : plus un génotype (marqueur seul, haplotype ou bien génotype à multilpes loci) est fréquent, plus la puissance des méthodes est grande. Donc, si le rapport de pénétrances s'éloigne de l'unité par l'étude de génotypes définis en plusieurs loci conjointement (exemple (a_1a_1,b_1b_2)), le nombre d'individus possédant ce génotype est plus petit puisque ceux-ci forment un sous-ensemble du nombre total d'individus possédant un génotype défini en un seul de ces loci (a_1a_1 pris seul). Également, considérer des loci conjointement a pour effet d'augmenter le nombre de tests effectués car le nombre de combinaisons de génotypes est bien plus grand que le nombre de génotypes. L'ajustement pour un plus grand nombre de tests a pour effet de diminuer la puissance statistique. En résumé, il est difficile de juger si la considération de loci pris conjointement dans les analyses statistiques sera profitable pour la puissance. Notons que lorsque le rapport de pénétrances marginales est très près de l'unité (par exemple lorsque l'effet d'un génotype sur la pénétrance s'oppose tout dépendant du génotype à un deuxième locus), la considération de loci pris conjointement représente inévitablement une amélioration. Toute autre situation n'est malheureusement pas aussi simple.

Finalement, une des raisons principales de l'étude des interactions est la compréhension du phénomène au niveau biologique. En effet, une meilleure compréhension de l'interaction entre gènes nous permet de cerner le rôle des protéines et nous rapproche du but principal motivant les études d'association, c'est-à-dire

de comprendre la maladie afin de cerner les protéines qui sont impliquées dans celle-ci et comment elles le sont afin de trouver une cure ou bien un soulagement aux gens atteints.

2.3. En résumé

En résumé, on a vu que dans une population formée de sous-populations avec des fréquences d'allèles différentes, un test d'association ($\delta \neq 0$) entre allèles et maladie peut être significatif sans que cet allèle ne soit près d'un allèle de susceptibilité pour la maladie. On voudra trouver un moyen d'éviter cette situation. Également, on a vu que si c'est une combinaison de génotypes à plusieurs loci qui est responsable de l'accroissement de la susceptibilité des individus, cela peut expliquer le manque de réplicabilité des tests d'association et peut influencer la puissance de détection de ceux-ci. Qui plus est, il est souhaitable pour les généticiens théoriques de connaître comment interagissent les gènes (et leurs produits), une information qui ne peut pas être connue par beaucoup de méthodes d'association génétiques. Le prochain chapitre décrira des solutions qui abordent individuellement puis conjointement les questions d'association de l'interaction entre gènes et de stratification de la population.

Chapitre 3

SOLUTIONS

Ce chapitre veut aborder des solutions aux problèmes mentionnés dans le chapitre précédent. La première section aborde le problème de la stratification de la population et propose un échantillonnage de cas et de leurs parents. Ce type d'échantillons sera utilisé dans les simulations et analyses de ce mémoire. S'en suivra une partie génétique qui vise à illustrer comment ce type d'échantillons apporte de la robustesse aux analyses. Nous allons ensuite décrire deux méthodes existantes qui ont pour but l'exploration des modèles de maladie avec une ou des interactions entre gènes et qui sont utilisées avec des échantillons cas-contrôles. Le statisticien sera intéressé par les procédures impliquées par les méthodes et possiblement moindrement aux raisons génétiques ayant motivé le choix de celles-ci. La dernière section propose une manière d'utiliser des échantillons composés de triades cas-parents dans les méthodes d'exploration de l'interaction génétique. C'est d'ailleurs l'idée-clé et l'innovation qu'apporte ce mémoire.

3.1. TRIOS CAS-PARENTS : ROBUSTESSE À LA STRATIFICATION

Afin d'obtenir une statistique qui soit moins influencée par la stratification de la population que ne l'est une mesure de l'association du même type qu'une statistique chi-deux d'un tableau de contingence avec des cas et des contrôles, un autre plan d'expérience et une autre statistique ont été proposés (voir Spielman et al. (1993)). On utilise des trios cas-parents, c'est-à-dire un malade avec ses 2 parents, desquels on investigue le génotype à différents marqueurs. La méthode est semblable à associer comme "cas" les allèles qui ont été transmis des parents

à l'enfant atteint et comme "contrôle" les allèles qui n'ont pas été transmis. En figure 3.1, on voit clairement les allèles représentés par des carrés colorés qui sont transmis et non transmis des parents à l'enfant malade.

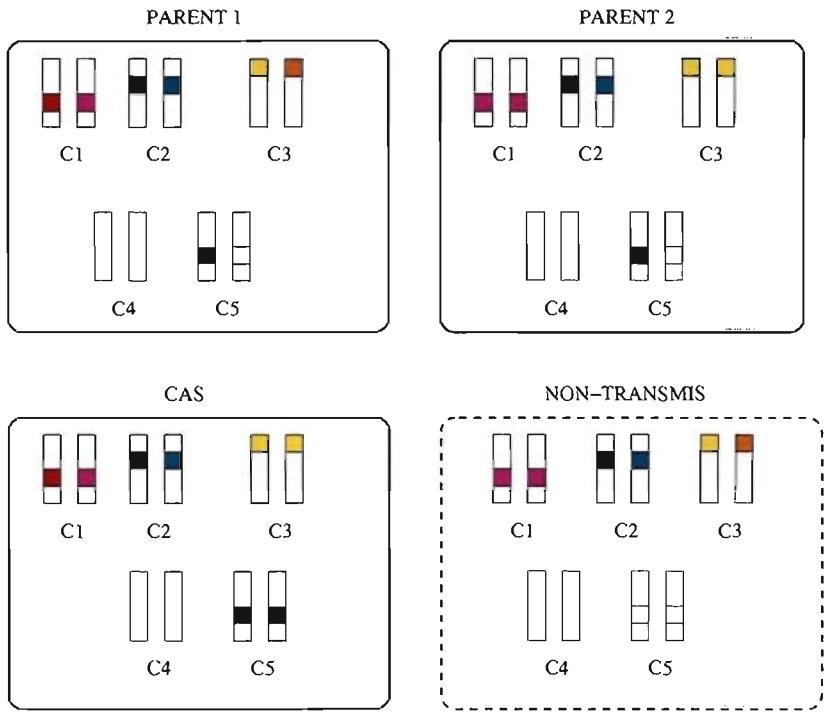

FIG. 3.1. Exemple de triades composés de cas et de ses parents avec une illustration des allèles non-transmis

Ainsi, puisque le contrôle et le cas proviennent tous deux de la même famille, donc également de la même population sous-jacente, la contribution de l'association due à la stratification de la population est réduite. On teste ensuite la différence de fréquences d'allèles entre les deux groupes par un test basé sur le khi-deux de Pearson. La méthode porte le nom de "AFBAC" en anglais pour "Affected Family-Based Controls".

Afin d'illustrer comment les échantillons cas-parents peuvent être utilisés de façon à réduire l'effet de stratification de la population, on propose une méthode élaborée par Wilcox et Weinberg (Wilcox et al. (1998)). La méthode étudie un seul allèle à la fois et un trait phénotypique. Soit C le nombre de cet allèle présent dans le génome d'un individu dans une population donnée. Soient également H et F ce même nombre chez le père et la mère de cet individu. Par exemple,

admettons qu'on étudie l'allèle a situé au locus A. Supposons également que le génotype du père est composé d'un a et d'un autre allèle alors que celui de la mère est composé de deux a au locus A. Finalement, posons que le cas issu de ces parents possède le génotype (a, a) c'est-à-dire qu'il possède deux a. Alors, ce trio fait partie de la catégorie $(H, F, C) = (1, 2, 2)$. Wilcox et Weinberg admettent comme variable aléatoire mesurée le nombre de trios (H, F, C). Notons le nombre de trios (H, F, C) par $N(H, F, C)$. Le modèle proposé par Wilcox et Weinberg est que $N(H, F, C)$ suit une distribution de Poisson avec paramètre donné par

$$\lambda := \mathbb{E}(N(H, F, C)) = \exp(\gamma(H, F) + \beta_1 \mathbb{1}_{C=1} + \beta_2 \mathbb{1}_{C=2} + \ln(2)\mathbb{1}_{H=F=C=1}) \quad (3.1)$$

où $\mathbb{E}$ symbolise l'espérance, $\mathbb{1}$ est la fonction indicatrice avec condition écrite en indice et γ est un paramètre constant pour chaque catégorie de reproduction distincte (H, F). En faisant référence à l'exemple du paragraphe précédent, la catégorie de reproduction (1,2) correspond aux triades où le père possède une copie de l'allèle a seulement dans son génotype alors que la mère en possède deux. La présence des paramètres γ permet de contrer l'effet de la stratification de population comme il sera expliqué plus loin. Notons que l'on pose l'équivalence des catégories de reproduction (H, F) et (F, H). Par exemple, ceci implique que l'on pose provenir d'une même catégorie de reproduction à la fois $(H = 2, F = 1)$ et $(H = 1, F = 2)$. Le modèle peut se résumer sous forme du tableau 3.1.

La signification des paramètres γ, β_1 et β_2 est particulièrement subtile. Leur interprétation passe par la probabilité de recenser un trio particulier (H, F, C) tel qu'on sait que l'individu en question est malade :

$$P(H, F, C|M = 1) = \frac{P(M = 1|H, F, C)}{P(M)} P(C|H, F)P(H, F). \quad (3.2)$$

Notons que la probabilité $P(C|H, F)$ s'estime directement et peut être exclue de l'estimation de ces paramètres. En effet, on a vu au chapitre 1 comment calculer la probabilité de génotype transmis des parents à l'enfant et $P(C|H, F)$ représente donc des constantes.

Envisageons deux possibilités. La première possibilité est celle de stratification allélique de la population, c'est-à-dire que la population est subdivisée en

TAB. 3.1. Fréquences espérées de trios pour un allèle

Trio (H, F, C)	Catégorie de reproduction (H, F)	ln(fréquence espérée)
222	1	$\gamma_1 + \beta_2$
212	2	$\gamma_2 + \beta_2$
211	2	$\gamma_2 + \beta_1$
122	2	$\gamma_2 + \beta_2$
121	2	$\gamma_2 + \beta_1$
201	3	$\gamma_3 + \beta_1$
021	3	$\gamma_3 + \beta_1$
112	4	$\gamma_4 + \beta_2$
111	4	$\gamma_4 + \beta_1 + \ln(2)$
110	4	γ_4
101	5	$\gamma_5 + \beta_1$
100	5	γ_5
011	5	$\gamma_5 + \beta_1$
010	5	γ_5
000	6	γ_6

sous-populations avec fréquences de l'allèle étudié différentes ainsi que probabilités d'être malade à l'intérieur de ces sous-populations différentes également. Supposons aussi que la probabilité de maladie ne dépend du nombre de copies d'allèle C que possède l'individu que par la stratification de la population c'est-à-dire par le nombre d'allèles que possède la sous-population dont il provient. Ceci signifie alors que $P(M = 1|H, F, C) = P(M = 1|H, F)$ et l'événement est indépendant du nombre d'allèles C si le nombre d'allèles de la sous-population dont il provient est fixé. En effet, fixer la catégorie de reproduction (H, F) revient à fixer une sous-population où le nombre d'allèles est constant. On utilisera cette information pour simplifier l'espérance des fréquences de catégories (H, F, C), nommément $N * P(H, F, C|M = 1)$, où la probabilité précédente peut s'exprimer sous la forme de l'équation (3.2). Dans ce cas, la valeur attendue de l'estimation de β_1 et de β_2 est 0 puisque toute la différence de fréquences entre les malades

peut être expliquée par un paramètre qui dépend de (H, F), nommément $\gamma(H, F)$, car la probabilité décrite par l'équation (3.2) ne dépend plus de C.

La deuxième possibilité est celle de la non-influence du couple (H, F) sur la détermination de la maladie de l'individu, mais plutôt du nombre d'allèles que possède l'individu malade. Ceci signifie qu'il y a une association génétique entre la maladie et l'allèle indépendamment de la stratification. C'est le type d'association génétique qu'il nous est désirable de détecter. Ceci implique directement que $P(M = 1|H, F, C)$ est équivalente à $P(M = 1|C)$ et ainsi ne dépend pas de la fréquence des catégories de reproduction (H, F). Dans ce cas, les paramètres γ ne peuvent plus expliquer totalement les fréquences dans chaque catégorie de trio (H, F, C) car il existe une différence autre qu'attendue par les probabilités de transmission seules à l'intérieur d'une même catégorie. Dans ce cas, $\exp(\gamma(H, F)) = N * P(H, F)$, $\exp(\beta_1) = \frac{P(M=1|C=1)}{P(M=1)}$ et $\exp(\beta_2) = \frac{P(M=1|C=2)}{P(M=1)}$ c'est-à-dire que $\exp(\beta_1)$ et $\exp(\beta_2)$ équivalent aux rapports de la probabilité d'être malade si on sait le nombre d'allèles de référence que possède l'individu sur la même quantité lorsque ce nombre d'allèles est inconnu. En résumé, c'est l'hypothèse nulle $\beta_1 = \beta_2 = 0$ qui permet de tester l'association entre la maladie et l'allèle étudié avec l'influence de la stratification réduite par la paramétrisation γ. Ainsi, on remarque que l'information génétique des trios cas-parents peut être utilisée efficacement afin de mesurer l'association génétique qui n'est pas due au phénomène de stratification de la population.

3.2. Sélection de modèle : exploration des interactions

Oublions pour l'instant l'aspect stratification de la population et concentrons-nous sur l'exploration des phénomènes d'interaction. Intuitivement, on peut tester l'interaction de la même manière qu'on le ferait tout aussi intuitivement en considérant les allèles seul à seul. Par exemple, si on considère une interaction de deux loci, on pourrait considérer les différences de fréquences entre les cas et les contrôles pour les catégories de génotypes possibles pour ces deux loci. Par contre, on est confronté à plusieurs problèmes en procédant ainsi : si on teste pour la non-association entre génotypes conjoints de deux, voire trois ou même quatre

loci, la quantité de tests à effectuer devient rapidement grande et les temps de calcul énormes avec les ordinateurs actuels si le nombre de marqueurs considérés est grand. En effet, si on inspecte l'association génétique pour deux loci bi-alléliques pris conjointement, ceci revient à comparer par une certaine méthode les fréquences de cas et de contrôles pour chaque génotype possible des deux loci (il y a neuf possibilités). Si le nombre de ces marqueurs bi-alléliques est N, le nombre de combinaison de deux marqueurs est $N(N-1)/2$ pour un total de génotypes dont on a à inspecter l'effet de $9 * N(N-1)/2$. C'est pourquoi il devient intéressant de considérer des méthodes de sélection de modèles et ce seront ces méthodes qui auront notre attention tout au long de ce mémoire.

3.2.1. RMD : Réduction Multifactorielle de Dimensionalité

Cette méthode effectue la sélection de modèle selon ce que notre intuition nous a dicté dans le paragraphe précédent. La méthode sera explicitée ici puisqu'utilisée dans le mémoire. Son importance dans ce mémoire est toutefois seconde à la méthode décrite dans la section suivante. Le niveau de détails des concepts statistique est conséquent au but d'application seulement de celle-ci. C'est une méthode de sélection de modèle non paramétrique (voir Ritchie et al. (2001)). L'idée de la méthode est de considérer, une à une, toutes les combinaisons de loci possibles. Afin d'illustrer la méthode, soit une étude d'association avec N cas et N contrôles. Une dizaine de loci est à l'étude. La méthode étudie une à la suite de l'autre toutes les combinaisons de 1, 2, 3, ... , 10 loci. On notera par (i,j,k), par exemple, la combinaison des loci i, j et k. Entre autres, si on regarde l'ensemble des combinaisons de deux loci, celui-ci correspond à l'ensemble $\{(i,j)|i,j \in \{1,2,...,10\}$ et $i<j\}$ alors que celui des combinaisons de trois loci est $\{(i,j,k)|i,j,k \in \{1,2,...,10\}$ et $i<j<k\}$. La réduction de dimensionalité telle qu'elle sera décrite est appliquée pour chaque combinaison de chaque taille. Pour une combinaison, il existe différents génotypes possibles. Par exemple, soit la combinaison (1,9). Soit également le locus 1 pouvant abriter des allèles a_1 et a_2 et le locus 9 pouvant abriter les allèles b_1 et b_2. Alors, les génotypes possibles pour une telle combinaison est le produit cartésien de l'ensemble de génotypes individuels au locus

1, $\{(a_1,a_1),(a_1,a_2),(a_2,a_2)\}$, avec l'ensemble de génotypes individuels au locus 9, $\{(b_1,b_1),(b_1,b_2),(b_2,b_2)\}$. À chaque élément du produit cartésien on associe l'effigie "haut risque" si le nombre parmi les N cas possédant le génotype exact aux loci 1 et 9 surpasse ou égale celui pour les N contrôles. Si c'est l'inverse qui se produit, le génotype portera plutôt l'annotation de "bas risque". La figure 3.2 explicite la catégorisation pour la combinaison de deux tels loci.

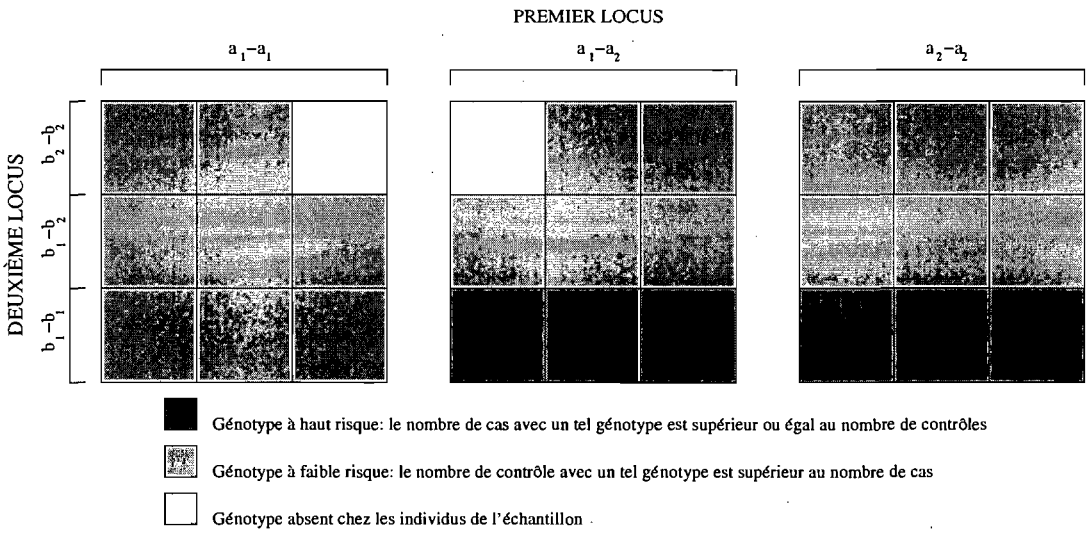

FIG. 3.2. Exemple de réduction de dimensionalité pour deux loci

Ceci ramène la structure à haute dimensionalité (plusieurs marqueurs, plusieurs génotypes aux marqueurs) à une structure à une seule dimension (génotype à haut risque/génotype à bas risque). Il résulte de chaque combinaison de loci un modèle qui classifie les individus possédant un génotype à bas risque comme non malade et un génotype à haut risque comme malade. Ce modèle sera évalué et comparé à ses pairs.

La sélection d'un modèle pour chaque taille de modèle (effet seul, combinaison de 2, 3, etc. loci) est effectuée : le modèle qui maximise le ratio cas/contrôle des génotypes à haut risque est sélectionné sur l'ensemble des modèles de même taille. Notons qu'en agissant ainsi et par définition, l'erreur de classification, définie

comme le nombre d'erreurs commises en classifiant à tort un individu comme cas ou contrôle, sera minimale pour ce modèle sur ce même ensemble des modèles.

Les modèles de chaque taille sont évalués par quatre mesures : la précision du formateur, la précision de prédiction, le test du signe et la constance de validation croisée. La précision du formateur, la précision de prédiction ainsi que la constance de validation croisée se calculent en partitionnant les données en 10 ensembles contenant des parts égales de cas et de contrôles. Chaque ensemble contenant 1/10ème des individus à l'étude porte le nom d'ensemble test et est associé à l'ensemble des 9/10ème restant qui porte le nom d'ensemble formateur. Pour chaque taille T de combinaison et en utilisant cette fois-ci les données du formateur seulement, on évalue à nouveau le meilleur modèle selon les critères du paragraphe précédent. Alors, la précision de prédiction correspond à la proportion moyenne d'individus provenant des ensembles tests qui a été bien classé par le meilleur modèle de taille T associé. La précision du formateur correspond à la même proportion pour l'ensemble formateur au lieu de l'ensemble test. Finalement, la constance de validation croisée correspond à la proportion de fois où le meilleur modèle de taille T sélectionné par le formateur correspond au meilleur modèle de taille T sur toutes les données.

Les trois mesures du paragraphe précédent veulent évaluer si l'ajout de complexité correspondant à considérer un modèle de plus grande taille a permis de détecter un véritable effet de génotype ou bien n'ajuste que l'erreur expérimentale. L'idée derrière la constance de validation croisée est que le "vrai signal" sera présent à travers les données peu importe comment elles sont divisées. Une trop faible constance de validation croisée est donc un indicateur d'ajustement du modèle à l'erreur expérimentale. De plus et en général, la précision de prédiction a tendance à augmenter lorsque la taille du modèle s'approche de la taille du "vrai" modèle (voir Ripley (1996)). Finalement, on peut effectuer un test de signal pour chaque modèle de la manière suivante. La valeur-p est calculée en évaluant la grandeur de la constance de validation croisée avec sa distribution sous l'hypothèse nulle de non-association entre génotype et maladie. Cette distribution s'obtient en effectuant 1000 permutations de l'assignation des cas et des

contrôles et en calculant à chaque fois la constance de validation croisée sur ces données permutées. Tous ces calculs sont effectués par un programme informatique portant le nom de MDR (Moore et al. (2006)). Un exemple de données et de sortie du programme informatique MDR sera fourni plus loin.

La réduction multifactorielle de dimensionalité est non-paramétrique et les modèles sélectionnés doivent être interprétés selon le classement de chaque génotype selon qu'ils sont à haut ou à bas risque pour la maladie, chose qui peut s'avérer difficile. À cet effet, une mesure de l'entropie est fournie par le logiciel, c'est une mesure qu'on ne décrira pas avec précision dans ce mémoire, on réfère le lecteur à l'article de Jakulin et Bratko (2003) pour plus de détails. On utilisera une valeur qui est fonction de l'entropie et qui mesure le gain d'information à considérer conjointement par rapport à isolément chaque locus. C'est cette quantité qui permet d'évaluer si les données suggèrent une interaction entre loci.

Voici maintenant un exemple d'application de la méthode par le logiciel MDR. Le logiciel prend comme entrée un fichier comprenant le génotype des individus et son statut d'affection (cas ou contrôle). Un exemple de quelques lignes d'un tel fichier est donné immédiatement :

```
...
1 2 1 2 2 1 1 0 1 0 0 1 0 1 1
1 2 0 1 1 2 1 0 2 2 1 1 0 1 1
0 0 2 0 1 0 2 2 2 0 1 0 1 1 0
1 1 2 1 0 2 2 1 1 1 0 1 2 0 0
...
```

Dans cet exemple, chaque ligne correspond à un individu et il y a 15 colonnes. La dernière colonne représente le statut d'affection de l'individu et les quatorze premières représentent chacune un marqueur. Un statut d'affection de zéro signifie que l'individu en question est un contrôle alors qu'un un représente un cas. Chaque valeur associée à un marqueur représente un génotype possible de celui-ci. Si on considère un marqueur d'un seul nucléotide pouvant être "A" ou bien "G", le zéro pourrait représenter le génotype "AA", l'unité signifie le génotype "AG" et finalement un deux indiquerait le génotype "GG".

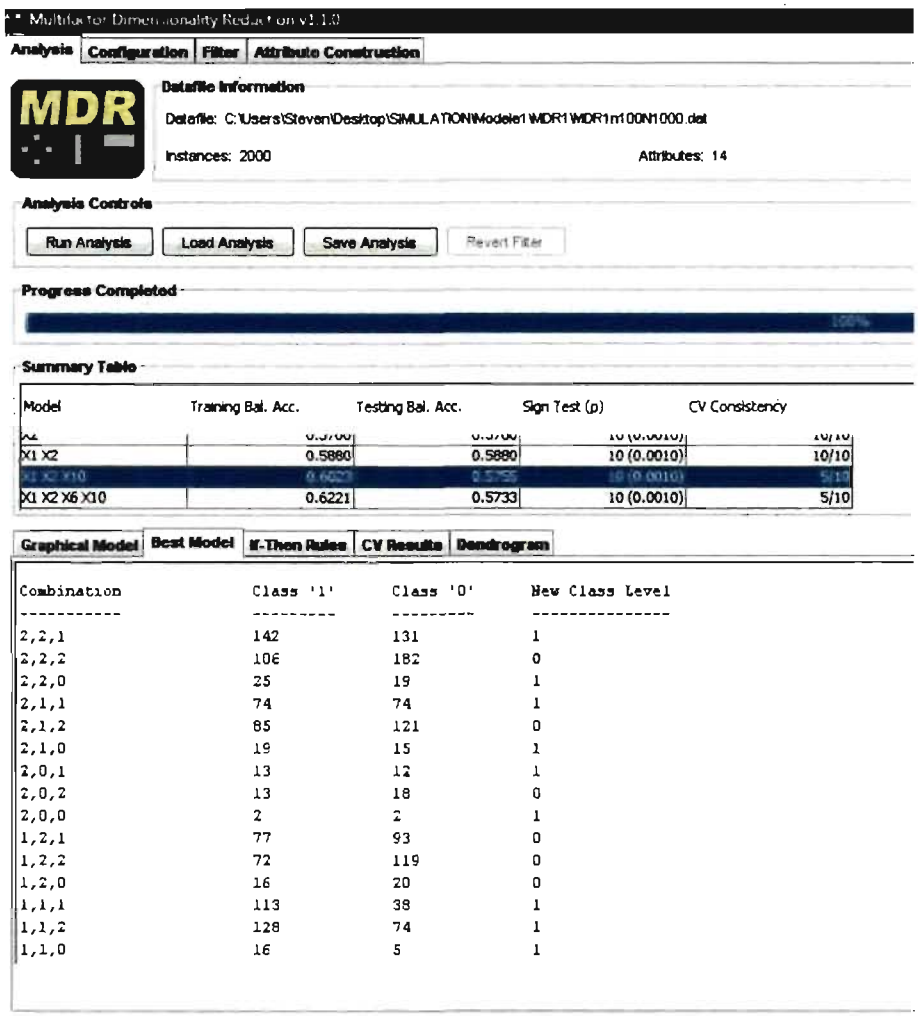

FIG. 3.3. Sortie MDR : les précision de formateur, précision de l'ensemble test, test du signe et constance de validation croisée pour les modèles sélectionnés de chaque taille

La sortie du programme a l'allure de la figure 3.3 pour un fichier de données de 1000 cas et de 1000 contrôles. Un ensemble de quatorze marqueurs bi-alléliques a été utilisé. Dans la figure, on voit dans le haut l'information sur le fichier lu par le programme ("Datafile Information"). Dans la section un peu plus bas portant le nom "Analysis Control", on peut ajuster des paramètres de l'analyse, ce qui ne sera pas discuté ici. Les résultats de l'analyse sont présentés plus bas avec un tableau ("Summary Table") où sont présentés en ordre les meilleurs modèles de chaque taille ("Model"), la précision de formateur associée ("Training Bal.

Accuracy"), la précision de prédiction ("Testing Bal. Accuracy"), un test du signal qui n'est pas discuté ici ("Sign Test (p)") et finalement la constance de validation croisée ("CV Consistency").

Encore plus bas on voit plusieurs onglets qui ont des valeurs différentes pour les différents modèles. Dans la figure 3.3, l'onglet sélectionné est "Best Model" et le modèle inspecté est celui composé des variables X1, X2 et X10. On y voit la combinaison d'allèles pour les trois marqueurs. Par exemple, en première ligne du tableau on voit que le génotype correspond à X1 = 2, X2 = 2 et X10 = 1 avec les valeurs définies plus haut. Le nombre d'individus possédant ces génotypes qui sont dans la classe de valeur 1 (les sujets atteints de la maladie) est présenté sous la colonne "Class '1'" tandis que le nombre de contrôles avec un tel génotype est présenté sous la colonne "Class '0'". La dernière colonne indique sous quelle catégorie le modèle assigne un individu possédant ce génotype aux marqueurs du modèle. Sous l'onglet "Graphical Model", on voit la figure 3.4 qui montre graphiquement ce qui est présenté sous l'onglet "Best Model". Dans cette figure, le modèle illustré est formé des deux marqueurs X1 et X2. On voit que chaque case représente un génotype particulier. Par exemple, la case en haut à droite représente le génotype X1 = 2 et X2 = 0. Dans cette case, on voit deux barres. La barre de gauche représente le nombre de cas possédant le génotype de cette case tandis que celle de droite représente le nombre de contrôles. La couleur de la case indique si le nombre de cas surpasse ou égale le nombre de contrôle et est donc un génotype à haut risque. Le dernier onglet d'intérêt est "Dendogram" où l'on présente les mesures d'entropies graphiquement ou sous forme de tableau tel qu'en la figure 3.5. Les valeurs intéressantes pour l'analyse sont les mesures I(A ;B ;C) qui est positive lorsque les données suggèrent une interaction entre les marqueurs. Les mesures H() sont des mesures d'entropies.

Le choix du modèle final dépend de critères subjectifs. Par exemple, on peut se fixer une constance de validation croisée minimale et choisir dans l'ensemble de modèles respectant ce critère le modèle avec la plus grande précision de prédiction. Ensuite, pour juger de l'interaction entre les variables du modèle sélectionné, on inspecte les valeurs I(A ;B ;C) et conclut en une interaction si la valeur est plus

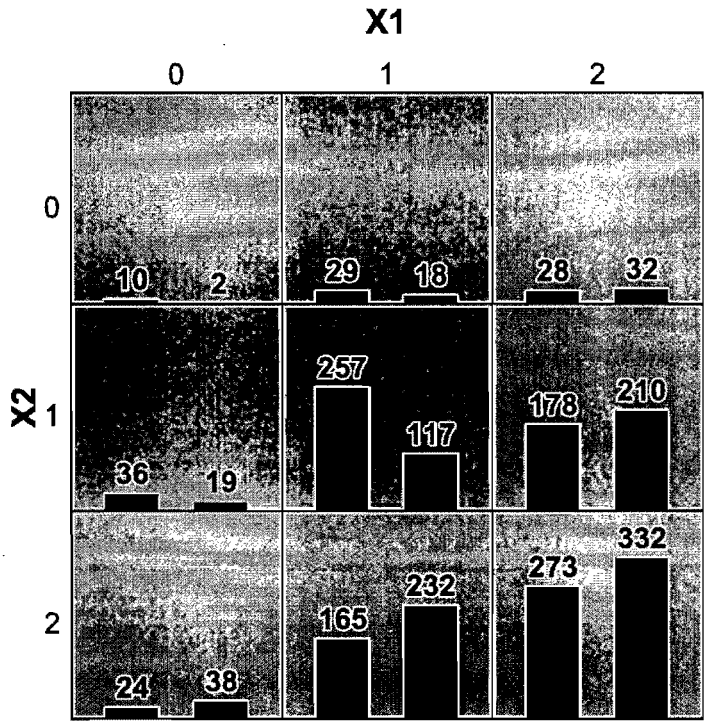

FIG. 3.4. Sortie MDR : la réduction de dimensionalité pour le modèle de combinaison X1 et X2

Single Attribute Values:

Attribute	H(A)	H(A|C)	I(A;C)
X1	1.2699	1.2633	0.0066
X2	1.2542	1.2400	0.0143
X6	1.2332	1.2324	0.0007
X10	1.3145	1.3065	0.0080

Pairwise Values:

Attribute A	Attribute B	H(AB)	H(AB|C)	I(A;B)	I(A;B;C)	I(AB;C)
X1	X2	0.8016	0.7708	1.7225	0.0100	0.0308
X1	X6	0.9791	0.9704	1.5240	0.0014	0.0087
X1	X10	0.8407	0.8277	1.7436	-0.0015	0.0131
X2	X6	0.9970	0.9828	1.4903	-0.0008	0.0142
X2	X10	0.9960	0.9814	1.5727	-0.0076	0.0147
X6	X10	0.9956	0.9868	1.5520	0.0000	0.0088

FIG. 3.5. Sortie MDR : la mesure de l'interaction par l'entropie

grande que k, une valeur aussi subjective. Dans le cas de l'exemple présenté par les figures 3.3, 3.4 et 3.5, on aurait pu se donner une valeur de constance de validation croisée minimale de 0.7 (on rappelle que cette constante se situe à la figure 3.3

sous la section intitulée "CV Consistency"), et choisir ainsi le modèle formé de la combinaison X1 et X2, soit la combinaison des deux premiers marqueurs. Une inspection de la valeur de la mesure de l'interaction nous démontre que celle-ci est positive et relativement grande (un choix possible de valeur minimale de I pourrait être 0.002). On aurait interprété des résultats dans les figures 3.3 à 3.5 que les données suggèrent une interaction entre le marqueur 1 et le marqueur 2 avec de tels critères puisque ce dernier a une très grande constance de validation croisée (10/10) et la précision de prédiction décroît lorsque la taille du modèle augmente (elle passe de 0.588 à 0.5755). De plus, la mesure de I(A ;B ;C) pour les attributs X1 et X2 est de 0.01, suggérant une interaction entre les variables. Supposons que le premier marqueur était codé de façon à ce que le génotype "AA" est représenté par 0, "AC" par 1 et "CC" par 2 et que le deuxième marqueur a exactement le même code. Alors en regardant la figure 3.4 on remarque que pour qu'il y ait un effet des gènes sur la susceptibilité d'être malade, il est nécessaire qu'un individu ait au moins un A pour le premier locus et au moins un A pour le deuxième locus.

Cet exemple permet de comprendre les outils que nous fournissent le programme MDR. Ces outils nous permettent de prendre une décision finale quant à la présence et la nature des interactions génétiques vraisemblablement suggérées par les données. On utilisera les concepts amenés ici afin de se forger des critères de décisions qui seront utiles lorsque des simulations numériques seront exécutées au quatrième chapitre. Toutes les procédures statistiques n'ont pas été expliquées en détails par soucis de concision de ce mémoire. Notons qu'on ne connaît pas l'amplitude de l'effet des génotypes et ne pouvons donc pas non plus concentrer nos efforts sur un nombre plus restreint de loci puisque tous les marqueurs du modèle final sélectionné doivent être étudiés ensemble. La conversion de l'interprétation statistique en interprétation biologique est également particulièrement difficile si le nombre de marqueurs du modèle sélectionné est grand et si les fréquences à l'intérieur d'un génotype de plusieurs loci sont petites. Pour obtenir plus d'informations, on aurait besoin d'une méthode paramétrique.

3.2.2. Régression logique

Une méthode paramétrique simple pourrait prendre la forme usuelle utilisée en biostatistique dans le contexte de maladie et d'exposition. Un exemple de modèle où on a considéré 35 facteurs d'exposition pourrait être

$$logit(P(M = 1|G = g)) = \alpha_0 + \alpha_1 X_8(g) + \alpha_2 X_{35}(g) + \alpha_3 X_8(g) \cdot X_{35}(g) \quad (3.3)$$

où à nouveau l'événement M=1 indique qu'une personne est malade et les variables prédictives X_i sont des variables dichotomiques qui représentent des niveaux de facteurs d'exposition pour les facteurs i (i $\in$ {1,2,..., 35}). On reconnaît également le troisième terme de l'expression de droite comme une expression d'interaction entre les facteurs X_8 et X_{35}. Dans le cas qui nous concerne, les facteurs dépendent du génotype des individus, de là la dépendance en g des variables X_i.

Si on considère des échantillons de cas et de contrôles, on peut penser qu'une méthode adéquate de sélection de modèle faisant intervenir des interactions puisse se présenter exactement sous la forme de l'équation (3.3) où les variables X_i représentent des génotypes à un seul locus. Ainsi, l'interaction entre les loci i et j se présente sous la forme de la sélection de termes $X_i \cdot X_j$. Une interaction triple, par extension, sera la multiplication de trois de ces facteurs s'ils concernent des loci différents. Appliquer une méthode de sélection sur l'ensemble des modèles possibles (un très grand espace) nous apporterait ainsi l'information recherchée de l'exploration des interactions. Du point de vue de la statistique, la méthode est usuelle, mais d'un point de vue génétique, plusieurs points indisposent.

Par exemple, soient deux gènes : l'un code pour une protéine P et l'autre code pour le récepteur de cette protéine dans une cellule. Admettons également qu'une mutation survienne sur le gène codant pour la première protéine P. Cette mutation affecte la conformation de la protéine et celle-ci adhère moins bien au récepteur : il peut en résulter un malfonctionnement dans un système biologique. De la même façon, une mutation peut survenir sur le gène codant le récepteur de cette protéine avec exactement ou presque le même effet. Dans ce cas, si une personne possède soit l'un ou l'autre des génotypes mutés à l'un ou l'autre locus,

elle possède presque exactement la même susceptibilité pour la maladie. Or, comment se modéliserait mathématiquement le phénomène d'interaction ? On notera $X_1 = 1$ si le génotype de l'individu possède au moins une copie de la mutation de la protéine P dans son génotype ou bien X_1=0 si ce n'est pas le cas. Soient également X_2=1 et X_2=0 les situations analogues pour un locus du récepteur de la protéine. L'équation mathématique de l'interaction entre les deux gènes prendrait alors la forme

$$logit(P(M = 1|G = g)) = \alpha_0 + \alpha X_1 + \alpha X_2 - \alpha X_1 \cdot X_2 \quad , \tag{3.4}$$

et trois de quatre termes sont nécessaires à décrire l'interaction biologique, soit les trois derniers termes. Cette fois-ci, considérons une situation différente où les deux mutations doivent être présentes pour ajouter à la susceptibilité d'un individu. Dans ce cas, la modélisation de la susceptibilité (ou probabilité d'être malade) étant connu le génotype d'un individu est :

$$logit(P(M = 1|G = g)) = \beta_0 + \beta X_1 \cdot X_2. \tag{3.5}$$

Ainsi, un type d'interaction biologique a besoin de trois termes pour être complètement décrit et un autre n'en a besoin que d'un seul. Pourtant, malgré que les deux modèles aient le même niveau de complexité biologique, ils représentent des niveaux de complexité différents du point de vue statistique. Ainsi, puisque les méthodes de sélection de modèles usuelles imposent une pénalité en fonction du nombre de paramètres, il est possible que le modèle possédant le plus grand nombre de paramètres soit plus difficile à identifier clairement. De plus il n'est pas facile de relier la forme de l'équation (3.4) avec son interprétation génétique qui a permis de former cette équation. En fait, il a été souligné que de mêmes modèles statistiques pouvaient représenter des interactions génétiques bien différentes, voir Cordell (2002). De plus, la situation se complique au niveau de l'interprétation si plus de deux loci sont en interaction. Pour faciliter l'interprétation et afin de diminuer la différence entre complexité des modèles génétiques et statistiques, on utilisera un modèle de régression tout à fait particulier.

Premièrement, on peut noter que les deux types d'interaction utilisés en exemple dans le paragraphe précédent peuvent être représentés par les expressions $X_i \wedge X_j$ et $X_i \vee X_j$ où X_i représente une variable dichotomique prenant la valeur unité s'il y a présence de l'allèle mutant dans le génotype de l'individu au locus i et où les opérateurs $\wedge$ et $\vee$ sont les opérateurs logiques respectivement "ET" et "OU". En fait, l'idée qui sous-tend la principale méthode qui sera utilisée dans le mémoire est la suivante : comme les lois de la génétique semblent intuitivement être plus près d'expressions logiques que la modélisation usuelle, on désire inclure aux modèles des opérateurs logiques. La régression de tels modèles porte le nom de régression logique (Ruczinski et al. (2003)). Décrivons plus amplement la régression logique. Notons que tout ce qui a été dit se rapporte à un allèle situé près (en déséquilibre de liaison) d'un autre allèle qui est directement impliqué dans la susceptibilité face à la maladie.

En fait, plus précisément, la régression logique est une procédure permettant l'exploration et la sélection de différents modèles où réside le type d'interaction discuté ici. En effet, elle considère (de là en provient son nom) des expressions logiques prenant la valeur 1 lorsque vraie et 0 lorsque fausse, formées à partir de variables explicatives dichotomiques. En résumé et plus spécifiquement, elle prend comme entrée des variables dichotomiques prédictives ainsi qu'une variable réponse d'un échantillon et donne en sortie un ou plusieurs modèles consistants en l'espérance de la variable réponse dépendant d'expressions logiques de variables prédictives. Illustrons cela par un exemple. On note Y la variable réponse et X_1, X_2, X_3 et X_4 des variables prédictives prenant les valeurs 0 ou 1. Un modèle choisi par la méthode pourrait avoir la forme suivante :

$$f(E(Y)) = \beta_0 + \beta_1 \cdot (X_1 \text{ et } X_2) + \beta_2 \cdot (X_3 \text{ et } (\neg X_4 \text{ ou } X_1)). \tag{3.6}$$

Dans le contexte de la sélection de modèle génétique de maladie, Y peut par exemple représenter l'événement {l'individu est atteint de la maladie} et prendre les valeurs 1 ou 0 alors que f est la fonction logistique. Les variables X_i en chaque marqueur i peuvent, par exemple, représenter les événements {l'individu a au moins une copie du premier allèle au marqueur i}. Dans l'équation précédente et selon cette définition, le modèle de maladie implique une interaction entre le

premier et le deuxième marqueur et une autre entre les premier, troisième et quatrième marqueurs. En effet et par exemple, si le signe du paramètre β_1 est positif, la présence d'une copie du premier allèle au marqueur 1 et d'une copie du premier allèle au marqueur 2 augmente le risque d'être atteint de la maladie.

Dans ce qui suivra, on utilisera des arbres pour représenter les expressions logiques dont chaque élément en forme ce que l'on nommera les feuilles. La figure 3.6 représente les deux arbres impliqués dans l'équation précédente. Les feuilles sont les variables dichotomiques X_1, X_2, X_3 et X_4 aux extrémités des branches des arbres. Ainsi, le premier arbre de notre exemple possède deux feuilles alors que le deuxième en possède trois ; le modèle est composé de 5 feuilles. L'emplacement de chaque opérateur ou feuille d'un arbre est appelé un noeud. On remarque que chaque noeud possède soit 0 ou deux noeuds directement adjacents à lui.

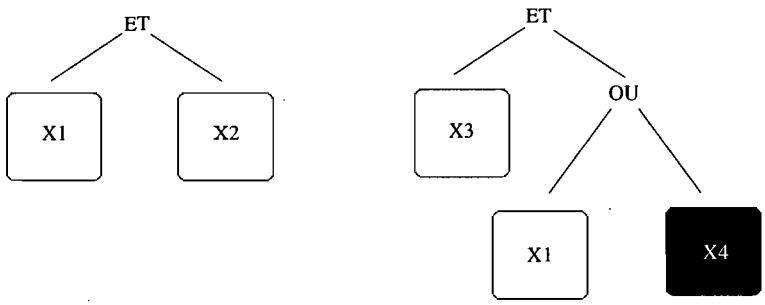

FIG. 3.6. Représentation des expressions logiques

Afin d'explorer la très grande possibilité de modèles, Ruczinski, Kooperberg et Leblanc (Ruczinski et al. (2003)) proposent une approche basée sur la construction et la modification progressive d'arbres logiques d'un ensemble restreint d'opérations possibles. C'est par ces modifications qu'il est possible de manoeuvrer à l'intérieur de l'espace des modèles possibles. Ces modifications sont illustrées dans la figure 3.7 tirée directement de la thèse de doctorat de Ruczinski (2003) et peuvent être énumérées :

1- le remplacement d'une feuille d'un arbre par une autre (fig. 3.7, Alternate Leaf),

2- le remplacement d'un opérateur ET par un opérateur OU et vice-versa (fig.

3.7, Alternate Operator),

3- l'ajout d'une branche à un noeud n'étant pas une feuille : à l'emplacement du noeud, ce qu'il y a sous ce noeud est conservé à droite du noeud, on ajoute une branche à gauche que l'on connecte par un opérateur ET ou bien OU. Ce qui est ajouté à la fin de la branche de gauche est une feuille, donc une variable prédictive. L'opération inverse est également considérée, c'est-à-dire qu'on enlève une branche d'un noeud qui est un opérateur (fig. 3.7, Grow Branch / Prune Branch),

4- la division de feuilles : à l'emplacement d'une feuille, on rajoute une feuille et l'on connecte ces deux dernières par un opérateur. L'opération inverse la suppression de feuille fait également partie des opérations possibles (fig. 3.7, Split Leaf / Delete Leaf).

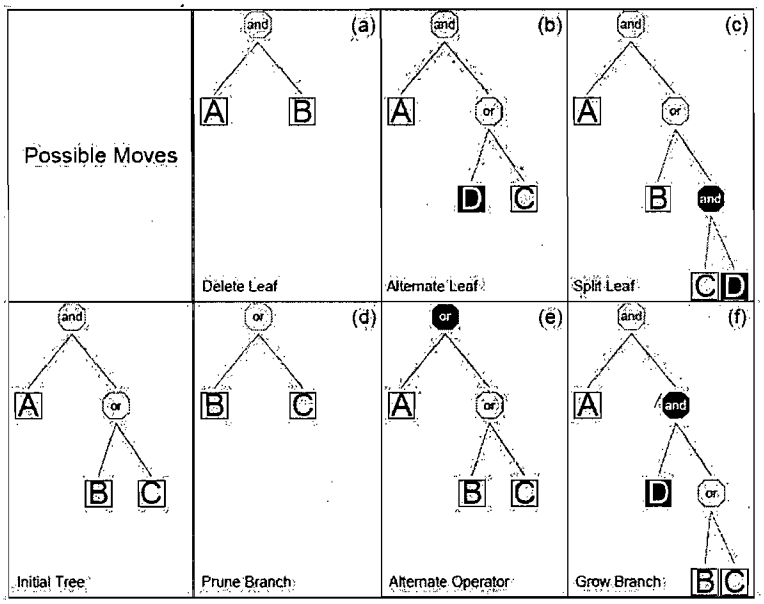

FIG. 3.7. Opérations sur les arbres logiques

On peut montrer (ce qui ne sera pas fait ici) que chaque expression logique peut être construite par de telles modifications sur n'importe quel arbre logique (Ruczinski et al. (2003)).

À ce point, nous n'avons pas parlé de fonction d'ajustement de données ou bien d'estimation de paramètres. En fait, tous les paramètres d'équations telles

que l'équation (3.6) sont estimés par les procédures habituelles de régression c'est-à-dire que pour une régression avec fonction f valant la fonction identité, l'estimation s'effectue en minimisant la somme de carrés de résidus alors que pour une fonction f logistique, on minimise plutôt la déviance binomiale (Agresti (1990)). On rappelle que la déviance prend la forme de

$$D(y|\theta) = -2(\log(P(y|\hat{\theta})) - \log(P(y|\theta = y))) \tag{3.7}$$

où D est la fonction de déviance, la variable y représente l'ensemble des variables dépendantes mesurées, θ représente les paramètres du modèle et $P(y|\theta)$ est la fonction de probabilité du modèle (la fonction de probabilité binomiale dans le cas de modèle de ce type) telle l'équation (3.5). La déviance permet également la comparaison entre les différents modèles ; le score le plus bas implique un meilleur ajustement des données. Au prochain paragraphe on verra comment s'effectue la sélection.

La première étape choisit le modèle de régression initial : ce sera celui consistant en l'espérance de la variable réponse comme fonction d'une seule variable prédictive. La variable prédictive sélectionnée correspond à celle impliquant un score (déviance dans le cas de régression logistique du modèle) le plus bas. Cette seule feuille forme le premier arbre. On calcule alors le résultat de la fonction score pour tous les modèles issus des opérations possibles (telles qu'illustrées dans la figure 3.7 de cet arbre. L'opération qui implique un modèle de régression logistique (d'expressions logiques) avec le score le plus bas est effectuée : le modèle résultant est adopté pour les étapes suivantes. On réitère ces étapes jusqu'à ce que le score ne puisse plus être amélioré par une de ces manoeuvres. On débute alors un autre arbre (toujours si ceci implique une amélioration du score). Le modèle final est adopté lorsque le score le plus bas de tous les modèles explorés par la méthode est atteint. L'estimation des paramètres, le calcul de la déviance de chaque modèle ainsi que tous les processus de sélection de modèle sont effectués par une librairie, nommée LogicReg, du langage informatique R. R a été développé par la compagnie Bell Laboratories par John Chambers et ses collègues.

Notons que cette sélection n'assure pas l'obtention du modèle au score optimal sur l'ensemble des modèles possibles puisque le meilleur modèle obtenu ainsi

pourrait encore être amélioré (être modifié afin d'en améliorer son score) par la combinaison de deux opérations et non d'une seule. De plus, par simulations numériques et par l'utilisation d'une multitude de fonctions scores, les auteurs remarquent que cette méthode n'est pas optimale puisque le modèle final est souvent trop grand c'est-à-dire que le modèle final adopté possède trop de feuilles par rapport au modèle simulé. On voudrait également considérer un ensemble de modèles qui ajuste bien les données, puisque le 'vrai' modèle n'a pas nécessairement le score le plus bas sur les données mesurées.

La méthode proposée afin de sélectionner un ensemble de modèles à scores intéressants pour le chercheur combine la régression logique avec la sélection de modèle par chaînes de Markov Monte Carlo (que l'on notera CMMC par souci de concision du texte) (voir Kooperberg et Ruczinski (2005)). Cette méthode nous permet de passer d'un modèle à l'autre, à nouveau par les opérations décrites plus haut, avec une certaine probabilité d'acceptation du nouveau modèle, telle une chaîne de Markov dont les états sont en fait les modèles. L'idée est que les modèles les plus fréquemment "visités" durant le déroulement de la chaîne de Markov sont les modèles les plus intéressants à étudier. La sélection de modèles est bayésienne, c'est-à-dire qu'elle utilisera cette fois-ci des probabilités a priori sur la taille des modèles (identifié plus tôt comme le nombre de feuilles au total sur tous les arbres logiques) ainsi que sur les modèles à l'intérieur des ensembles des modèles à taille constante. Notons que la distribution des probabilités à l'intérieur d'un de ces ensembles est choisie comme uniforme sur tous les modèles que cet ensemble contient. Aucune probabilité a priori n'est imposée sur les valeurs des coefficients de régression : ils sont à nouveau estimés par des méthodes de maximum de vraisemblance puisque ce n'est pas la valeur de ces coefficients qui nous importe mais plutôt la structure du modèle.

La transition entre les modèles est assurée par une chaîne de Markov à sauts réversibles (Green (1995)) dont on décrit ici les grandes lignes. Pour le modèle Monte Carlo de l'itération i, MC_i, on calcule la probabilité a priori de ce modèle ainsi que la vraisemblance des mesures selon le modèle logistique de la distribution binomiale et respectivement notées p_i et l_i. On choisit ensuite l'un des

arbres du modèle MC_i au hasard. Une opération sur l'ensemble des opérations possibles décrites plus haut est sélectionnée avec probabilités 10/23 d'alternance de feuille, 1/23 d'alternance d'opérateur et de 3/23 pour les 4 autres opérations possibles. Ces probabilités ont été choisies par les auteurs et possèdent des avantages de sélection de modèles (Kooperberg et Ruczinski (2005)). En fonction de l'opération retenue, il est possible que l'on doive choisir aléatoirement un opérateur (l'opérateur $\wedge$ ou bien $\vee$) ou une variable prédictive avec probabilités égales. La probabilité du choix d'opération adopté est notée q_i et sera utile plus loin. Il résulte de l'opération un nouveau modèle, duquel on calcule la vraisemblance binomiale avec modèle logistique l_{i+1}, la probabilité a priori p_{i+1} et la probabilité du choix d'opération permettant de revenir au modèle initial q_{i+1}. La probabilité de transition du modèle initial au nouveau modèle est définie par $min(1, r)$ où :

$$r = \frac{p_{i+1}}{p_i} \frac{q_{i+1}}{q_i} \frac{l_{i+1}}{l_i}. \tag{3.8}$$

Si la transition est acceptée, le modèle de l'itération $i + 1$ est le modèle issu de l'opération. Si la transition est refusée, le modèle de l'itération $i+1$ est le modèle initial MC_i. Notons que seule la régression logique par chaîne de Markov à saut réversible sera utilisée dans les simulations.

La même librairie du logiciel R nommée LogicReg permet d'effectuer toutes les procédures précédemment citées : le calcul de probabilités a priori et a posteriori et autres probabilités, le choix des opérations et le recensement des modèles visités. Le logiciel prend comme argument un nombre maximal d'arbres permis, un nombre maximal de feuilles, un modèle initial ainsi que la valeur établie pour les probabilités a priori de chaque modèle. Ensuite, le logiciel effectue le calcul des probabilités d'accepter ou de refuser une opération sur l'arbre initial, tel que décrit plus haut, et en conséquence adopte ou pas le nouveau modèle. Il effectue un certain nombre de ces itérations sans noter les modèles visités. Les auteurs de Kooperberg et Ruczinski (2005) ont arrêté le choix de ce nombre d'itérations à 10,000. Ces premiers modèles sont jugés peu importants puisqu'ils ne servent qu'à approcher le score optimal sur l'ensemble des modèles possibles. Après ces 10,000 itérations, pour chaque modèle de chaque itération on note le modèle

visité. Après un nombre d'itérations subjectivement grand (Kooperberg et Ruczinski (2005) utilisent 5,000,000 d'itérations), la procédure se termine et nous obtenons en sortie les fréquences d'apparition de chaque variable prédictive ainsi que des couples et des trios de celles-ci. C'est sur ces quantités que s'effectueront les analyses statistiques.

L'allure du fichier utilisé en entrée du programme de régression logique diffère de celui du logiciel MDR. Les lignes du fichier d'entrée de la section précédente deviennent pour la régression logique :

```
...
1 0 1 1 1 0 1 1 1 1 1 0 1 0 1
1 0 1 1 0 0 1 0 1 0 1 1 1 0 1
0 0 0 0 1 1 0 0 1 0 0 0 1 1 0
1 0 1 0 1 1 1 0 0 0 1 1 1 1 0
...
```

où la dernière colonne décrit l'état de la maladie pour l'individu représenté par la ligne, c'est-à-dire 1 pour un individu malade et 0 si ce n'est pas le cas, et où deux colonnes sont utilisées pour représenter chaque marqueur bi-allélique. Remarquons que seuls septs marqueurs sont inclus ici, puisque quatorze marqueurs auraient impliqués vingt-huit entrées, ce qui est encombrant pour un exemple. Ces colonnes forment un couple de variables dichotomiques. Rappelons que seul ce type de variable peut être utilisé dans la régression logique. Soit l'allèle a_1 du premier marqueur i choisi comme allèle de référence pour ce marqueur (ce choix est arbitraire). En la ligne k, une valeur de 1 à la (2i-1)ème colonne indique que le kième individu possède au moins 1 allèle a_1, sans quoi cette valeur est nulle. Une valeur de 1 à la (2i)ème colonne de cette même ligne indique que le kième individu possède 2 allèles a_1, sans quoi cette valeur est nulle. Ainsi l'individu homozygote pour l'allèle a_1 aura en colonne 2i-1 et 2i les valeurs 1 et 1 respectivement, l'individu hétérozygote aura en colonne 2i-1 et 2i les valeurs respectives 1 et 0 et celui qui ne possède pas d'allèle a_1 aura en colonne 2i-1 et 2i les valeurs respectives 0 et 0.

En exemple, on peut utiliser les mêmes données que celles ayant servi dans la section sur la réduction multifactorielle de dimensionalité et effectuer la régression logique Monte Carlo. La chaîne de Markov comportera 5,010,000 d'itérations (une itération par possibilité de transition entre modèles) en omettant d'écrire le résultat des 10,000 premières. La distribution a priori des grandeurs de modèles visités correspond à $P(S = i) = (\frac{1}{2})^i$ où S est la grandeur de modèle. On rappelle que la grandeur d'un modèle correspond au nombre de feuilles au total sur tous les arbres logiques. Le choix du paramètre de la loi géométrique n'a pas eu de grande influence sur les modèles sélectionnés de Kooperberg et Ruczinski (2005), de sorte qu'une seule de ces valeurs a été sélectionnée. On impose un maximum de trois arbres et neuf feuilles aux modèles pour d'autant plus encourager la parcimonie. Ce choix du nombre de termes et du nombres de variables impliquées est inspiré de Kooperberg et Ruczinski (2005). Par variable prédictive, on obtient le nombre de fois dont celle-ci faisait partie du modèle pour les 5,000,000 d'itérations. Ceci est représenté par la figure 3.8. Le nombre équivalent pour un couple de variables faisant partie du même arbre est représenté par un code de couleur en figure 3.9 : plus la couleur de la case s'éloigne du rouge, plus ce nombre est grand. Notons que ces nombres exacts peuvent également être obtenus. On ne montrera pas ici le nombre de fois qu'un trio de variables s'est retrouvé dans le même arbre, mais notons que cette quantité peut également s'obtenir de la sortie de logiciel. Voici comment pourrait s'interpréter les résultats. Dans la figure 3.8, on voit que les variables V2, V4 et V20 sont les variables les plus régulièrement incluses dans les modèles explorés par la chaîne de Markov. Toute autre variable est incluse dans moins de 15% des modèles et ne sont pas considérés (on s'est donné une valeur de 15% comme point de césure subjectif). La figure 3.9 montre que seuls les variables V2 et V4 sont fréquemment sélectionnées conjointement dans les modèles. À la lumière des résultats précédents, on aurait interprété que les données suggèrent une interaction entre les variables 2 et 4 (donc les 2 premiers marqueurs) et un effet simple de la variable 20 (marqueur 10). Le meilleur modèle comprenant ces deux arbres et trois feuilles prend la forme de la figure 3.10.

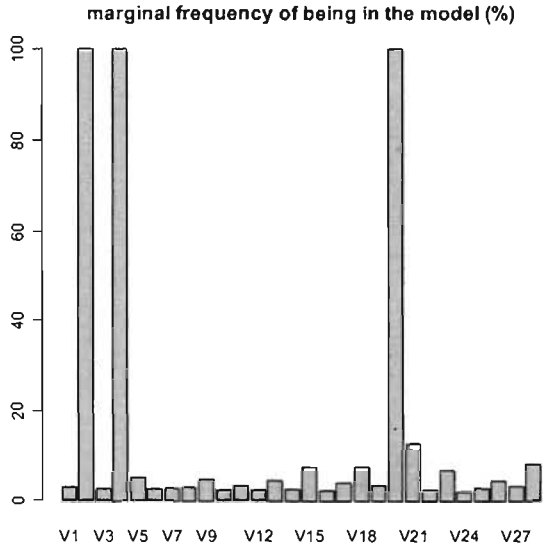

FIG. 3.8. Exemple de régression logique Monte Carlo : fréquence de sélection des variables prédictives

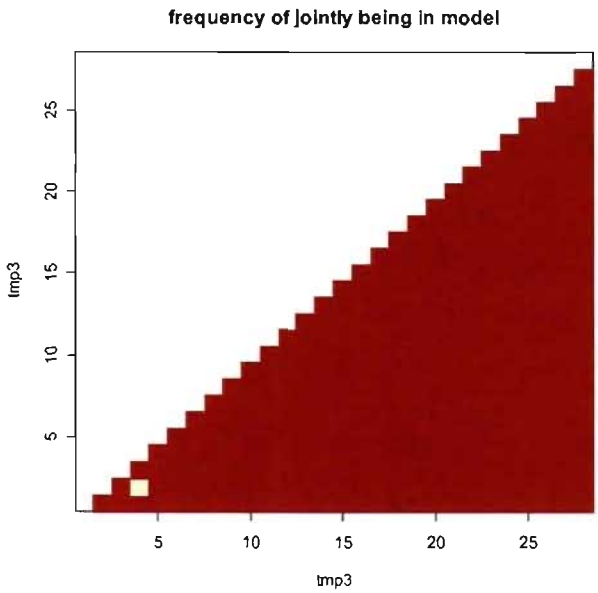

FIG. 3.9. Exemple de régression logique Monte Carlo : fréquence de sélection de couples de variables prédictives

Une remarque importante concernant à la fois la réduction multifactorielle de dimensionalité et la régression logique est qu'elles comparent les fréquences allélliques entre les cas et les contrôles. Or on a vu que des différences peuvent survenir dans ces fréquences dues à la stratification allèlique de la population étudiée. Si cette différence de fréquence peut engendrer des conclusions fautives quant aux

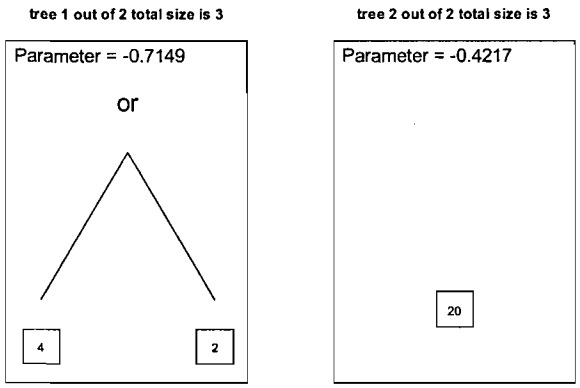

FIG. 3.10. Exemple de régression logique Monte Carlo : le modèle sélectionné avec la forme de l'interaction

rôles de certains marqueurs lorsque l'on considère les marqueurs seul à seul, on peut supposer qu'elle le sera au moins tout autant dans la considération additionnelle d'interaction. Dans le cas de marqueurs étudiés singulièrement, l'usage de trios cas-parents apportait une robustesse face au problème de stratification de la population. Est-il alors possible d'exploiter les logiciels de régression logique et de réduction multifactorielle de dimensionalité afin d'explorer les interactions par l'usage d'échantillons de trios cas-parents ? C'est le sujet de la prochaine section.

3.3. À LA CROISÉE DES CHEMINS

Le but de ce mémoire est de se prémunir d'une méthode nous permettant d'utiliser des trios cas-parents afin d'explorer les interactions génétiques. Pour y arriver, on étudiera en premier lieu la vraisemblance des mesures puisqu'elle prend une forme tout à fait particulière et semblable aux échantillons cas et contrôles.

3.3.1. Vraisemblance

On utilise la probabilité qu'un individu possède un génotype particulier si l'on sait que l'enfant est malade (on a pris les cas selon que ces derniers soient malades) et on conditionne la probabilité par le génotype des parents, afin de contrer l'effet de stratification de population. On regarde ainsi la distorsion de l'équilibre de transmission tel qu'il devrait persister s'il n'y a pas de lien entre la maladie et l'allèle situé au marqueur. Sous ces conditions, une application du théorème de Bayes nous permet de démontrer que la probabilité de mesurer le

génotype g_c en fonction du génotype du père g_h et de la mère g_f et du fait que l'enfant en question est atteint (événement que l'on dénote $M = 1$) est :

$$P(g_c|g_h, g_f, M = 1) = \frac{P(M = 1|g_c, g_h, g_f) \cdot P(g_c|g_h, g_f) \cdot P(g_h, g_f)}{\sum_{g \in G} P(M = 1|g, g_h, g_f) \cdot P(g|g_h, g_f) \cdot P(g_h, g_f)} \quad (3.9)$$

où G représente l'ensemble des génotypes possibles pour l'enfant étant connu le génotype des parents. Notons que contrairement à plus haut, on conditionne sur le génotype des parents. De là on effectue quelques suppositions. Premièrement, on suppose comme plus haut que l'événement "un cas est malade" (M=1) est indépendant du génotype des parents, lorsque l'on connaît le génotype de l'enfant. De plus, dans la population générale, la probabilité $P(g_c|g_h, g_f)$ suit communément les proportions mendéliennes (transmission de chacun des deux chromosomes, donc de chaque allèle ou haplotype, équiprobables pour les deux parents), et représentent des constantes pour la fonction de vraisemblance. En effet, pour tout trio de génotypes (g_c, g_h, g_f), il existe une unique constante K(g_c, g_h, g_f) tel que :

$$P(g_c|g_h, g_f, M = 1) = K \frac{P(M = 1|g_c)}{\sum_{g=G} P(M = 1|g)} \quad (3.10)$$

et ce, pour tout modèle $P(M|g)$. Notons que les probabilités $P(g|g_h, g_f)$ ne diffère pas entre génotypes si on discerne les génotypes hétérozygotes entre eux. Par exemple, soient les allèles (ou haplotypes) de nomenclatures "1" et "2". Soient également les génotypes du père et de la mère d'un individu malade qui est $g_h = (1, 2) = g_f$ c'est-à-dire les génotypes des deux parents sont formés de la combinaison des haplotypes "1" et "2". Dans ce cas, les génotypes possibles pour l'individu issu de ces parents (sans recombinaison) sont $(1, 1)$, $(1, 2)$ ou bien $(2, 2)$ avec probabilités respectives de 25%, 50% et 25% sous des proportions mendéliennes. Dans ce cas et pour le trio formé de ces génotypes pour les parents et du génotype $(1, 2)$ (par exemple) pour l'individu affecté, la contribution à la vraisemblance du trio est :

$$P(g_c = (1, 2)|g_h = (1, 2), g_f = (1, 2), M = 1) = \frac{2P(M = 1|g_c = (1, 2))}{\frac{1}{4}(\sum_{g^*=(1,1),(1,2),(2,1),(2,2)} P(M = 1|g = g^*))} \quad (3.11)$$

où il ne devient important de discerner le génotype $(1,2)$ du génotype $(2,1)$ que pour le calcul du dénominateur. La présence du terme $\frac{1}{4}$ signifie que toutes les probabilités de transmissions $P(g_c|g_h,g_f)$ sont identiques et respectent les proportions mendéliennes. Notons qu'aucune supposition sur le modèle $P(M|g)$ n'a été imposée. Comme notre estimation de modèle se fera d'abord et avant tout par la vraisemblance, on peut laisser tomber cette constante qui devient sans intérêt pour l'étude en question. Le modèle log-linéaire étudié se réfère aux termes de risque d'être atteint de la maladie, ce qui s'exprime mathématiquement comme :

$$log\Big(\frac{P(M|g)}{P(M|g_0)}\Big) = X'(g)\beta \tag{3.12}$$

où X représente un vecteur codant pour le génotype g et qui dépend de l'expression des gènes. Par définition, un allèle d'expression dominante s'exprimera en phénotype si au moins une copie de cet allèle est présente dans le génotype d'un individu. Un allèle d'expression récessive a besoin de deux copies pour s'exprimer chez un individu. Un allèle d'expression additive a un effet additif avec le nombre de cet allèle. Par exemple, soit un allèle d'expression dominante, alors un individu dont le génotype g possède au moins une copie de cet allèle au marqueur j aura comme jième élément de $X(g)$ la valeur 1. La vraisemblance pour la modélisation précédente prend donc la forme suivante (où on omet la constante K discutée plus haut) :

$$L = \frac{\exp(X_i'\beta)}{\sum_{X_*}\exp(X_*\beta)} \tag{3.13}$$

où X_* code pour chaque génotype à l'intérieur de l'ensemble des génotypes possibles. Par exemple, si N_a représente le nombre d'allèles à un marqueur particulier, le nombre de génotypes possible à ce marqueur correspond à $N_a(N_a-1)/2$. Alors, $X_*(g)$ pourrait être un vecteur avec $N_a(N_a-1)/2-1$ zéros et un 1 correspondant à la position désignée du génotype g sur l'ensemble des $N_a(N_a-1)/2$ positions du vecteur qui code pour chaque génotype. Le vecteur X_* possède une seule position par marqueur bi-allélique : on choisit un allèle de référence et on lui choisit une expression (dominante, récessive ou additive) qui décrit complètement les génotypes et leurs expressions.

On peut montrer que la fonction de vraisemblance modifiée (3.13) est équivalente à celle provenant d'une étude cas-contrôle avec, pour chaque génotype-cas, trois génotypes-contrôles associés qui sont les autres génotypes possibles provenant des parents (Self et al. (1991)). Pour être encore plus précis, ceci correspond à la vraisemblance d'un modèle de régression logistique conditionnelle avec l'assortiment 1-3 pour 1 génotype cas avec 3 génotypes contrôles. Les vraisemblances étant équivalentes, le fait de considérer l'échantillon soutiré de cas et de leurs parents comme un échantillon de cas ($M = 1$) associés chacun à trois contrôles ($M = 0$) est équivalent, avec les mêmes probabilités de maladie étant connu le génotype $P(M = 1|g)$. Toute estimation ou inférence sur ces probabilités peut être basée sur une régression logistique conditionnelle. C'est sur cette idée que se base l'adaptation de trios cas-parents.

3.3.2. Adaptation de données de trios cas-parents

L'idée est de transformer les données de trios afin d'obtenir une forme semblable à celle d'échantillons cas-contrôles. La proposition de ce mémoire est la suivante, on pose comme "cas" l'individu atteint de la maladie auquel est associé son propre génotype et on pose comme "contrôle" ou "pseudocontrôle" le pseudo-individu possédant le génotype non transmis des parents à l'enfant atteint. Une approche similaire a déjà été proposée au niveau allélique seulement (revoir en début du chapitre, la méthode "AFBAC"). Ici, l'approche est généralisée dans le cas d'études de plusieurs gènes possiblement en interaction. Même si du moins intuitivement, une telle approche semble intéressante et plausible, on se devra de tester l'idée dans un cadre de simulation. Ce sera le sujet du prochain chapitre qui expliquera toutes les démarches entamées afin de tester cette idée. Des données sous forme de trios cas-parents ainsi que sous la forme de cas et de contrôles seront simulées. Une comparaison des résultats obtenus formera l'essentiel des réponses nécessaires à ce mémoire.

Chapitre 4

RÉSULTATS ATTENDUS ET DE SIMULATION

Les échantillons cas-pseudocontrôle, forgés à partir de triades de cas et de leurs parents, n'ont jamais servi dans une procédure d'étude de multiples loci considérés conjointement. De plus, la considération d'un tel échantillon comme un échantillon cas-contrôle indépendant a certainement des répercussions sur les analyses. L'ampleur des répercussions sur les procédures statistiques reste incertaine à ce point. Dans ce chapitre, nous voudrons connaître s'il existe une relation théorique entre les deux types de tirages d'échantillons, soit le tirage d'un échantillon cas-pseudocontrôle (provenant de triades cas-parents) et le tirage d'un échantillon cas-contrôle. Ceci permettra de juger de la forme des biais impliqués par l'adaptation des triades. Comme le rapprochement des situations de prélèvement d'échantillons de cas-pseudocontrôle par rapport aux échantillons cas-contrôle n'est pas mathématiquement parfait (nous le verrons un peu plus loin), il sera nécessaire d'effectuer des simulations numériques que nous analyserons avec le logiciel MDR ainsi que par la régression logique Monte Carlo. Ceci nous permettra de vérifier les résultats théoriques ainsi que le comportement des procédures sur des données de ce type.

4.1. JUSTIFICATION THÉORIQUE

Dans cette section, nous obtiendrons deux résultats relatifs au rapprochement existant entre le prélèvement d'un échantillon cas-contrôle et celui d'un échantillon cas-parents duquel on forme des cas et pseudocontrôles. Ce rapprochement permettra de juger de la plausibilité de l'étude d'échantillons cas-pseudocontrôle

par une méthode qui gère les données comme si elles provenaient en réalité de cas et de contrôles indépendants. La similarité existant entre les deux types d'échantillon sera suffisante pour former des hypothèses quant aux résultats de l'utilisation des méthodes de la RMD et la régression logique Monte Cartlo. Notamment, on pourra juger du caractère de détectabilité d'interaction génétique ainsi que de la robustesse face à la stratification allélique de la population. Pour l'instant, inspectons la similarité des prélèvements d'échantillons.

L'analyse du rapprochement des prélèvement d'échantillons nécessite quelques suppositions. Premièrement, on supposera que la population étudiée est formée de sous-populations infinies à l'intérieur desquelles la reproduction est aléatoire. On supposera également que les deux dernières générations de ces sous-populations ont des fréquences de génotypes équivalentes. À l'intérieur des sous-populations, il n'existe pas de dépendance de transmission entre les différents chromosomes non-homologues lorsqu'aucune autre caractéristique que le génotype n'est connue sur l'individu issu de la transmission (la transmission suit les proportions mendéliennes). Nous supposerons finalement qu'à l'intérieur de ces sous-populations, la possession des deux chromosomes homologues d'un individu pris aléatoirement représentent deux événements indépendants (équilibre Hardy-Weinberg).

Certaines variables ont également besoin d'être définies. Soit $M_c = 1$ l'événement indiquant qu'un individu duquel est soutiré un pseudocontrôle est malade. Soit également $T = [T_{ij}]$ et $U = [U_{ij}]$ pour $i \in \{1, 2\}$ et $j \in \{1, ...k\}$, où k représente le nombre de paires de chromosomes sur lesquelles un marqueur génétique sera apposé, les matrices aléatoires des génotypes transmis et non-transmis pour au cas malade. Les éléments T_{ij} représentent les séquences alléliques du ième chromosome j transmis à un individu que l'on considère comme cas alors que les éléments U_{ij} sont les équivalents pour les pseudocontrôles. Dans ce qui suivra, l'usage des caractères minuscules sera réservé aux valeurs mesurées d'une variable du même caractère majuscule. Pour l'instant, on veut connaître $P(U = u|M_c = 1)$, ou les probabilités de génotypes d'un pseudocontrôle.

Théorème 1. *Dans le respect des conditions précédentes, le génotype du pseudocontrôle a probabilité équivalente à celle d'un individu pris aléatoirement de la sous-population dont le cas associé est tiré.*

DÉMONSTRATION. Premièrement, on a que la probabilité de génotype d'un pseudocontrôle associé à un individu malade est équivalente à

$$P(U=u|M_c=1)=\sum_{\mathbf{t},s}\Big(\frac{P(M_c=1|T=t,U=u,S=s)}{P(M_c=1)}\times P(T=t,U=u|S=s)\times P(S=s)\Big) \tag{4.1}$$

où S représente l'événement aléatoire d'appartenance de l'individu dont on soutire un pseudocontrôle à une sous-population et où on a utilisé la décomposition de Bayes. La probabilité que cet individu soit malade dépend seulement de son génotype G ainsi que de son appartenance à une sous-population, sans lien avec le processus de transmission. Sans information quant au pseudocontrôle, l'événement $M_c = 1$ est donc équivalent à $M = 1$ d'un individu quelconque. Nous obtenons alors

$$P(U=u|M_c=1)=\sum_{\mathbf{t},s}\Big(\frac{P(M=1|G=t,S=s)}{P(M=1)}\times P(T=t,U=u|S=s)\times P(S=s)\Big). \tag{4.2}$$

Notons que nous avons également supposé que les événements de transmission de chromosomes non-homologues sont indépendants, de sorte que

$$P(T=t,U=u|S=s)=\prod_{j=1}^{k}P(T_j=t_j,U_j=u_j|S=s) \tag{4.3}$$

où T_j et U_j sont des vecteurs représentant la jième paire de chromosomes homologues. En ce sens, T_j, U_j et leurs caractères minuscules associés représentent la jième colonne des matrices correspondantes. Nous devons maintenant faire intervenir les génotypes parentaux. Soit G_1 et G_2 les matrices de ces génotypes pour

le premier et le deuxième parent. Alors l'expression

$$P(T=t,U=u|S=s)= \prod_{j=1}^{k}\sum_{g_1,g_2}\Big(P(T_j=t_j,U_j=u_j|S=s,G_{1j}=g_{1j},G_{2j}=g_{2j}) \times P(G_{1j}=g_{1j},G_{2j}=g_{2j}|S=s)\Big) \tag{4.4}$$

avec G_{1j}, G_{2j} définies de manière analogue à T_j et U_j transmis et non-transmis. Dans ce qui suivra, nous concentrerons nos développements mathématiques à une seule paire de chromosomes homologues j. Également, supposons pour l'instant que les chromosomes transmis et non-transmis sont distincts et sont notés c_1, c_2, c_3 et c_4. Comme la transmission n'est pas dépendante de l'appartenance à une sous-population lorsque le génotype des parents est connu, on peut négliger cette dernière variable (S) de l'expression évaluant les probabilités de transmission. On poursuit,

$$\begin{aligned}\sum_{g_1,g_2}\Big(&P(T_j=(c_1,c_2),U_j=(c_3,c_4)|G_{1j}=g_{1j},G_{2j}=g_{2j})\times \\ &P(G_{1j}=g_{1j},G_{2j}=g_{2j}|S=s)\Big) \\ =&\frac{1}{4}\times P(G_{1j}=(c_1,c_3),G_{2j}=(c_2,c_4)|S=s)+ \\ &\frac{1}{4}\times P(G_{1j}=(c_1,c_4),G_{2j}=(c_2,c_3)|S=s)+ \\ &\frac{1}{4}\times P(G_{1j}=(c_2,c_3),G_{2j}=(c_1,c_4)|S=s)+ \\ &\frac{1}{4}\times P(G_{1j}=(c_2,c_4),G_{2j}=(c_1,c_3)|S=s). \end{aligned} \tag{4.5}$$

La reproduction aléatoire entre parents (indépendant du génotype de ceux-ci), implique que les variables G_{j1} et G_{j2} sont indépendantes et peuvent s'exprimer par G_j pour un individu pris aléatoirement dans la sous-population. Finalement, l'indépendance des événements de possession de deux chromosomes homologues

pour un individu permet d'obtenir une relation simple, c'est-à-dire

$$\sum_{g_1,g_2} \Big(P(T_j = (c_1, c_2), U_j = (c_3, c_4) | G_{1j} = g_{1j}, G_{2j} = g_{2j}) \times P(G_{1j} = g_{1j}, G_{2j} = g_{2j} | S = s) \Big)$$
$$= P(G = (c_1, c_2) | S = s) P(G = (c_3, c_4) | S = s). \quad (4.6)$$

Il peut être prouvé que le résultat reste inchangé pour toute combinaison de chromosomes homologues, c'est-à-dire

$$P(T_j = t_j, U_j = u_j | S = s) = P(G = t_j | S = s) \times P(G = u_j | S = s) \quad (4.7)$$

Ce résultat se généralise pour des génotypes à multiples chromosomes par indépendance des événements reliés à des chromosomes non-homologues. L'expression (4.1) devient

$$P(U = u | M_c = 1) = \sum_{\mathbf{t},s} \frac{P(M = 1 | G = t, S = s)}{P(M = 1)} \times P(G = t | S = s) \times P(G = u | S = s) \times P(S = s) \quad (4.8)$$
$$= \sum_{s} P(S = s | M = 1) \times P(G = u | S = s). \quad (4.9)$$

□

À l'intérieur d'une population homogène, l'absence de sous-population implique que $P(U = u | M_c = 1)$ vaut simplement $P(G = u)$. Dans ce cas, c'est dire que de soutirer des pseudocontrôles de la population étudiée est équivalent à soutirer des individus aléatoirement à l'intérieur de la population. La prochaine conclusion nécessite ce résultat et la prochaine définition.

Définition 1. *Soient deux populations spécifiques, on dira que* ***ces deux populations possèdent une même structure de modèle de maladie*** *pour les différents génotypes si la probabilité qu'un individu affecté par la maladie étant donné son génotype dans la première des deux populations* $P(M = 1|G)$ *et la même mesure pour la seconde population* $P^*(M = 1|G)$ *respectent la règle suivante : pour deux génotypes* g_1 *et* g_2, $P(M = 1|G = g_1) > P(M = 1|G = g_2) \iff$

$P^*(M = 1|G = g_1) > P^*(M = 1|G = g_2)$.

Par exemple, soit un modèle de maladie s'exprimant par

$$\text{logit}(P(M = 1|G = g)) = -2 + 0.5X_{10}(g) + 1(X_1(g) \wedge X_2(g)) \tag{4.10}$$

dans une première population et où les fonctions X_1, X_2 et X_{10} sont définis comme au chapitre trois. Une deuxième population possède la même structure de maladie si et seulement si son modèle de maladie s'exprime par la même équation mais avec des paramètres différents. Le prochain résultat s'applique aux populations homogènes (sans sous-population) et si les conditions ayant permis de prouver le théorême précédent tiennent toujours.

Théorème 2. *Dans le respect des conditions précédentes, pour chaque modèle de probabilité de maladie $P(M = 1|G = g)$ associé à une population et pour chaque échantillon de triades cas-parents duquel on forme un pseudocontrôle à partir des gènes non-transmis, il existe une seconde population avec modèle de probabilité de maladie défini par $P^*(M = 1|G = g)$ possédant la même structure que la première population, de laquelle l'échantillon cas-contrôle identique à l'échantillon cas-pseudocontrôle possède la même probabilité d'être prélevé.*

DÉMONSTRATION. Un échantillon composé de N_1 cas et pseudocontrôles pris d'une première population avec probabilité d'être affecté par la maladie $P(M = 1|G = g)$, ordonné de sorte que les notations des génotypes de cas sont g_1 à g_{N_1} et que les génotypes g_{N_1+1} à g_{2N_1} correspondent aux génotypes de pseudocontôles, est prélevé avec probabilité

$$\prod_{i=1}^{N_1} P(G = g_i|M = 1) \prod_{j=N_1+1}^{2N_1} P(G = g_j). \tag{4.11}$$

Pour une seconde population avec probabilité d'être affecté par la maladie $P^*(M = 1|G = g)$, l'échantillon cas-contrôle équivalent est tiré avec probabilité

$$\prod_{i=1}^{N_1} P^*(G = g_i|M = 1) \prod_{j=N_1+1}^{2N_1} P^*(G = g_j|M = 0). \tag{4.12}$$

Dans le cas particulier où

$$P^*(G = g_i|M = 1) = P(G = g_i|M = 1) \tag{4.13}$$

et

$$P^*(G = g_i|M = 0) = P(G = g_i) \tag{4.14}$$

alors les deux échantillons sont prélevés avec la même probabilité. Posant $K := P(M = 1)$ la probabilité qu'un individu pris aléatoirement dans la première population soit affecté par la maladie, nous obtenons que la probabilité de génotype dans la seconde population vaut

$$P^*(G = g_i) = KP(G = g_i|M = 1) + (1 - K)P(G = g_i) \tag{4.15}$$

qui est une mesure de probabilité valable. Qui plus est, la probabilité qu'un individu soit malade dans cette seconde population est

$$P^*(M = 1|G = g_i) = P(M = 1|G = g_i)\Big(\frac{K}{KP(M = 1|G = g_i) + (1 - K)P(M = 1)}\Big). \tag{4.16}$$

et respecte $P(M = 1|G = g_i) > P(M = 1|G = g_j) \iff P^*(M = 1|G = g_i) > P^*(M = 1|G = g_j)$ lorsque $0 < K < 1$.

□

Notons qu'on peut obtenir une expression simple du rapport de cotes pour un génotype particulier dans la population virtuelle. Pour un génotype particulier g_i, on rappelle que le rapport de cotes correspond à

$$RC(g_i) := \frac{P(M = 1|G = g_i)P(M = 0|G \neq g_i)}{P(M = 0|G = g_i)P(M = 1|G \neq g_i)}. \tag{4.17}$$

En usant du fait que le logarithme du rapport de cotes pour un génotype équivaut à la différence d'une fonction logistique des probabilités $P(M = 1|G = g_i)$ et $P(M = 1|G \neq g_i)$, une manipulation de 4.16 nous permet d'obtenir

$$\begin{aligned} log(RC(g_i)) &= logit(P^*(M = 1|G = g_i)) - logit(P^*(M = 1|G \neq g_i)) \\ &= log\Big(\frac{P(M = 1|G = g_i)}{P(M = 1|G \neq g_i)}\Big) \end{aligned} \tag{4.18}$$

et donc que le rapport de cotes de la population virtuelle équivaut au risque relatif de la population réellement étudiée. Ce dernier résultat ainsi que les deux théorêmes seront les principaux arguments permettant de juger de l'applicabilité des méthodes de la RMD et de la régression logique Monte Carlo.

4.2. Interprétation et résultats attendus

Dans les paragraphes précédents, nous avons effectué le rapprochement entre un échantillon cas-pseudocontrôle tiré de la population étudiée et forgé à partir des génotypes transmis et non-transmis des parents au cas et un échantillon cas-contrôle "équivalent" provenant d'une population que nous avons nous-mêmes définie, mais qui n'existe pas en réalité. Dans ce qui suivra, nous discuterons des différences possibles entre les échantillons cas-pseudocontôle et les échantillons cas-contrôle tous deux tirés de la même population étudiée. Il est important de noter cette disgression afin d'éviter une éventuelle confusion dans l'interprétation des résultats. Chaque fois, la provenance de l'échantillon cas-contrôle dont on discute sera explicitée.

Lorsqu'il n'y a pas de stratification allélique de la population et que la population est relativement homogène, nous jugeons que toute méthode statistique utilisant un échantillon cas-pseudocontrôle tel un échantillon cas-contrôle indépendant n'apportera pas de résultat biaisé par rapport à la structure du modèle. Plus spécifiquement, on a vu des précédents théorêmes que le prélèvement d'un échantillon cas-pseudocontrôle est un événement identique (du point de vue des probabilités) au prélèvement d'un échantillon cas-contrôle indépendant provenant d'une population différente. De façon plus importante, la structure du modèle de la maladie de cette population n'existant pas en réalité est la même que dans la population étudiée. Ceci veut donc dire que si la méthode statistique est convergente et non-biaisée pour les échantillons cas-contrôle indépendants, la méthode devra détecter la bonne structure de la maladie avec excellente probabilité pour un échantillon suffisamment grand. Par contre, le rapport de cotes pour un génotype augmentant la susceptibilité d'un individu pour la maladie dans le contexte de la population virtuelle associée aux trios cas-parents est toujours plus petit que

le rapport de cotes de la population réelle. Comme le rapport de cotes du génotype est l'un des facteurs les plus importants afin de détecter l'effet d'un génotype (voir Zondervan et Cardon (2004)), on interprète que la capacité à détecter des effets avec des échantillons cas-parents risque d'être plus petite que dans le cas d'échantillons cas-contrôles tirés de cette même population.

Lorsqu'il y a stratification allélique de la population, la situation est quelque peu différente. La similitude entre les deux types de prélèvements d'échantillons n'est pas parfaite ; lorsqu'il y a plusieurs sous-populations, il ne peut exister de population de laquelle prendre un échantillon de cas et de contrôles indépendant est un procédé équivalent. Par contre, l'indépendance surgit de la connaissance d'une seule variable, c'est-à-dire la connaissance de l'appartenance à une sous-population. En fait, dans ce cas, un échantillon cas-pseudocontrôle provenant de triades cas-parents est équivalent à un mélange de jusqu'à N_s (N_s équivalent au nombre de sous-populations distinctes à l'intérieur de la population) échantillons cas-contrôle indépendants provenant tous de populations possédant la même structure de modèle de maladie. Nous conjecturons donc que l'application de la régression logique Monte Carlo ou bien de la RMD sur un mélange d'échantillons cas-contrôles de même structure de maladie (donc sur un échantillon cas-pseudocontrôle avec sous-populations) devrait être convergente et non-biaisée quant à la structure du modèle de maladie. Qui plus est, nous croyons que la méthode statistique appliquée aux échantillons cas-pseudocontrôle sera plus robuste par rapport à la stratification allélique de la population que si la méthode avait été appliquée à un échantillon cas-contrôle. La raison de ceci revient à la discussion du troisième chapitre : il n'y a pas de déséquilibre de provenance des cas et pseudocontrôles par rapport aux sous-populations. Finalement, en ce qui concerne la grandeur de l'échantillon requis pour obtenir des "puissances" de détection semblables d'une interaction génétique, on s'attend à ce qu'elle soit plus semblable (par rapport aux échantillons cas-contrôle) que dans la situation d'absence de sous-population. En effet, malgré que le rapport de cotes soit moindre pour les échantillons cas-pseudocontrôle, ceux-ci sont plus robuste aux

effets négatifs que la stratification pourrait avoir sur la détectabilité d'un gène de susceptibilité pour la maladie. Le tableau 4.1 résume nos attentes a priori.

TAB. 4.1. Résultats attendus d'application de régression logique ou RMD

Population	Structure ?	Converge ?	Puissance	Robustesse
Non-strat.	Non-biaisée	Oui	$<$ Cas-Contrôle	-
Strat.	Non-biaisée	Oui (conject.)	$\approx$ Cas-Contrôle	$>$ Cas-Contrôle

Nous nous doutons donc que la régression logique Monte Carlo ainsi que la RMD seront utiles à la détection de l'interaction génétique. Afin de vérifier si la structure du modèle est détectée dans les deux situations de populations homogènes et hétérogènes, c'est-à-dire que la nature de l'effet d'un génotype particulier reste le même, il sera nécessaire d'effectuer des simulations numériques. Il est également pertinent de vérifier comment la régression logique Monte Carlo ainsi que la RMD détectent les effets des interactions impliqués dans la susceptibilité face à la maladie. Comme la méthode ne considère pas les dépendances entre les cas et les pseudos-contrôles comme la vraisemblance (3.13) le suggère, il sera important d'étudier si cette inexactitude influence grandement les analyses. Finalement, est-ce que le nombre de triades nécessaire à la détection d'interaction génétique sera grand ? Est-ce que la robustesse face à la stratification allélique de la population est plus grande pour les échantillons cas-pseudocontrôle tirés de triades par rapport aux échantillons cas-contrôle pris directement de la population étudiée ? La prochaine section répondra à ces questions.

4.3. SIMULATION DE TRIOS CAS-PARENTS

Nous devons simuler des données de modèles connus afin d'en forger des échantillons. Trois modèles de maladie simples, similaires à ceux des simulations de Kooperberg et Ruczinski (2005), seront utilisés lors des simulations. Ces échantillons devront ensuite être analysés par la régression logique Monte Carlo. On suivra les procédures établies dans l'article de Kooperberg et Ruczinski (2005) afin d'effectuer l'exploration de modèles par la régression logique. Une analyse à

moins grande échelle sera effectuée utilisant la RMD. Mais tout d'abord, inspectons l'algorithme de simulation de trios cas-parents.

Le génotype recensé des individus comprend 14 marqueurs puisque ce nombre représente un compromis entre la quantité et le temps de simulation. On supposera que tous les quatorze marqueurs sont bialléliques ou du moins, qu'un seul des allèles par marqueur est susceptible d'être impliqué dans la maladie. Comme dans Kooperberg et Ruczinski (2005), on posera que l'un des deux allèles a une fréquence fixée à 25% pour chaque marqueur. Dans ce qui suit on utilisera les variables de notation M, G pour signifier l'état d'affection d'un individu et son génotype respectivement. De plus, on utilisera les variables $X_i^u(G)$ et $X_i^d(G)$ qui prennent comme valeur l'unité si le génotype est composé de un ou bien de deux copies respectivement d'un des allèles (déterminé au départ) du marqueur i. On réfère le lecteur au troisième chapitre pour un exemple de la fonction dichotomique X prenant comme argument un génotype d'un individu. Les trois modèles de maladies qui seront simulés sont les suivants :

Modèle 1 :

- Une seule population dont la reproduction est aléatoire à l'intérieur de celle-ci.
- $\text{logit}(\text{P}(M = 1 \mid \text{Génotype} = g)) = -2 + 0.5X_{10}^d(g) + 1(X_1^d(g) \wedge X_2^d(g))$.
- La probabilité d'être atteint de la maladie est calculable numériquement (on en verra la façon dans ce qui suivra) et vaut 17.9%.
- Le risque relatif d'être atteint de la maladie, soit $P(M = 1|G)/P(M = 1|\neg G)$, vaut 2.15 pour l'interaction considérée.

Modèle 2 :

- Une seule population dont la reproduction est aléatoire à l'intérieur de celle-ci.
- $\text{logit}(\text{P}(M = 1 \mid \text{Génotype} = g)) = -2 + 1.5(X_1^u(g) \wedge X_2^u(g) \wedge \neg X_9^u(g))$.
- La probabilité d'être atteint de la maladie vaut 14.7%.
- Le risque relatif d'être atteint de la maladie vaut 3.17 pour l'interaction triple.

Modèle 3 :

- On considère de la ségrégation de population, c'est-à-dire une population formée de deux sous-populations avec différentes fréquences alléliques et qui se reproduisent à l'intérieur d'elles-mêmes seulement.
- La population étudiée est formée de 50% de la sous-population 1 et de 50% de la sous-population 2.
- $\text{logit}(\text{P}(M = 1 \mid \text{Géno.} = g, Sous - pop.1)) = -2 + 1(X_5^u(g) \wedge X_{14}^u(g))$
et $\text{logit}(\text{P}(M = 1 \mid \text{Géno.} = g, Sous - pop.2)) = -3 + 1(X_5^u(g) \wedge X_{14}^u(g))$.
- La probabilité d'être atteint de la maladie est : 14.8% (sous-pop. 1) et 7.13% (sous-pop. 2).
- Le risque relatif d'être atteint de la maladie vaut 2.25 dans la première sous-population et de 2.51 pour la deuxième sous-population.
- Les probabilités d'allèles dans les deux sous-populations sont présentées dans le tableau 4.2.

On simulera 500 jeux de données de N=500, 750 puis 1000 trios cas-parents pour le premier modèle. Également et pour le même modèle de maladie, on simulera 500 jeux de données de N=500, 750 et 1000 cas et contrôles. Pour les deux autres modèles, la procédure est similaire, mais cette fois-ci on considère des tailles échantillonnales de N=250, 500 et puis 750. Ces tailles échantillonnales ont été déterminées de manière à obtenir des probabilités grandes de détection de l'interaction pour la plus grande taille échantillonnale.

4.3.1. Choix des modèles de simulations

Le choix des modèles est une question importante puisqu'il influence la capacité des méthodes à détecter des effets d'interaction. En effet, plus une interaction a un grand effet (grand risque relatif ou rapport de cotes) et plus cette interaction est fréquente dans la population, plus il est facile de détecter cette interaction dans les analyses. En ce sens, les modèles ont été sélectionnés en fonction de leur risque relatif : on désirait simuler des modèles avec des risques relatifs de gènes ou de combinaisons variant entre 2 et 4. Si le risque relatif est plus bas que 2, les effets deviennent très difficiles à détecter (Ruczinski et al. (2003)) : il est possible que la taille de l'échantillon nécessaire à l'obtention d'une puissance de détection

TAB. 4.2. Probabilités à l'intérieur des deux sous-populations des premiers de deux allèles pour les différents marqueurs (troisième modèle).

Marqueur	Sous-pop. 1	Sous-pop. 2
1	0.45	0.50
2	0.70	0.40
3	0.55	0.75
4	0.65	0.80
5	0.75	0.65
6	0.50	0.75
7	0.60	0.50
8	0.65	0.30
9	0.65	0.55
10	0.65	0.35
11	0.80	0.75
12	0.90	0.45
13	0.75	0.75
14	0.75	0.65

raisonnable soit relativement très grande pour les triades cas-parents et grande pour les échantillons cas-contrôles. Notons que des allèles avec des risques relatifs plus petits que 2 sont très communs dans la littérature mais que ce risque relatif peut devenir plus grand si un effet d'une interaction avec cet allèle est considéré, tel qu'on l'a observé au troisième chapitre. De plus, dans l'optique que des allèles relativement communs sont au coeur des maladies plus communes (connu sous le nom de "common disease - common variant hypothesis" dans la littérature, Lohmueller et al. (2003)), un risque relatif de plus de 4 est peu réaliste.

Lors des simulations, des allèles plus rares appartenant à 25% de la population de gamètes ne sont presque pas simulés. Dans les études cas-contrôles, il a été observé que plus un allèle impliqué dans la maladie est rare dans la population, plus il devient difficile de détecter son effet par la régression logique (Zondervan

et Cardon (2004)), et ce, de manière exponentielle par rapport à la taille échantillonnale nécessaire pour obtenir une puissance considérable. On s'attendrait donc à ce que des allèles rares impliqués dans la maladie soient d'autant sinon plus difficiles à détecter par l'usage de cas et pseudo-contrôles. Or, la détection de telles interactions peut s'avérer pertinente du point de vue de la compréhension de phénomènes biologiques. Qui plus est, une interaction peut posséder des fréquences plus importantes dans différentes populations de sorte que sa détection dans un maximum de populations est souhaitable. Donc, les raisons de l'omission des allèles plus rares sont purement techniques et pas d'intérêts scientifiques.

Concernant le choix des modèles de simulation, notons que le nombre de variables possédant un effet dans les modèles de maladies $P(M = 1|G = g)$ est peu élevé par rapport à ce qui est soupçonné dans la littérature pour les maladies complexes (Zondervan et Cardon (2004)). On croit également que ce nombre peut affecter la fréquence de détection des effets en général de méthodes telle la régression logique puisqu'il devient difficile de manoeuvrer dans l'espace des modèles par les opérations sur les arbres logiques décrites au deuxième chapitre. C'est également le cas si le nombre de marqueurs étudiés au départ de la procédure est grand : il y a beaucoup plus de modèles à explorer et la possibilité que des effets de génotype soient détectés par chance grandit. Donc, intuitivement, la fréquence de détection des effets chez les triades diminue sensiblement avec l'augmentation du nombre de loci impliqués dans la maladie ainsi que du nombre de marqueurs considérés c'est-à-dire que dans ces cas, on a besoin d'une plus grande taille échantillonnale pour détecter plus fréquemment l'effet des génotypes. L'ampleur de l'effet de telles variables sur la fréquence de détection des effets devrait être étudiée par des simulations numériques ultérieurement à ce mémoire.

4.3.2. Algorithme de simulation

La première étape de l'algorithme de simulation est simple et assigne la probabilité des allèles de chaque marqueur dans la population. Par la suite on crée une matrice contenant tous les génotypes possibles dans la population ; chaque ligne représente un génotype possible. Plus spécifiquement, à chaque marqueur

est associée une position vectorielle (le ième marqueur est associé à la ième colonne). La valeur du jième marqueur du iième génotype (en position (i,j) de la matrice des génotypes) correspond au nombre d'un premier allèle désigné, donc vaut 0, 1 ou bien 2. Comme les marqueurs sont bi-alléliques, cette notation est suffisante pour décrire l'ensemble des génotypes possibles de la population. Pour un génotype de 14 marqueurs, un génotype possible se décrit par exemple par le vecteur suivant : (0,2,1,2,0,2,2,1,1,1,1,2,0,1). De ce vecteur, on déduit que l'individu possède 2 copies du premier allèle pour les marqueurs 2, 4, 6, 7 et 12. Comme on l'a indiqué plus tôt, la fréquence dans la population de tous les premiers allèles pour chaque marqueur est fixée à 25% pour les deux premiers modèles. Ceci veut dire que la probabilité d'un allèle pris au hasard pour le marqueur 12 dans la population est de 25% pour le premier allèle désigné et de 75% pour l'autre allèle. Le cas du troisième modèle sera discuté plus loin. Soit g_i désignant le génotype d'un individu pour l'ième marqueur. Alors, la probabilité de ce génotype dans la population est désignée par $P(G = g_i)$. Le rassemblement de tous les g_i correspond au génotype en chaque marqueur pour un individu. La probabilité des génotypes dans la population est simple à calculer si on suppose qu'il y a équilibre de liaison entre les marqueurs et équilibre d'Hardy-Weinberg pour les allèles de chaque marqueur. Celle-ci correspond alors à

$$P(G = g_1, g_2, ..., g_{14}) = P(g_1) \times P(g_2) \times ... \times P(g_{14}) \tag{4.19}$$

et la probabilité de génotype d'un marqueur est calculée depuis la fréquence des allèles puisque les haplotypes sont en équilibre Hardy-Weinberg. Par exemple, que pour un génotype hétérozygote au marqueur 5, la probabilité de ce génotype dans la population est $P(g_5) = 2*0.25*0.75 = 37.5$ %(on réfère le lecteur au premier chapitre pour plus de détails.). Pour chaque ligne de génotypes de la matrice discutée plus haut, on calcule numériquement la probabilité dans la population d'avoir obtenu ce génotype par l'équation (4.19). En ce qui concerne le troisième modèle, puisqu'il y a deux sous-populations distinctes, on doit calculer les probabilités des génotypes de chaque sous-population des fréquences propres à celles-ci et ainsi calculer deux vecteurs de probabilités.

La deuxième étape consiste à calculer les probabilités $P(M = 1|G = g)$ pour chaque génotype possible de la population. Pour chaque modèle de maladie, ces probabilités se calculent directement par la formule propre au modèle et établie plus haut. Il s'en suit un autre vecteur de probabilités pour chaque génotype (deux vecteurs pour le troisième modèle, un pour chaque sous-population). Afin d'assigner les génotypes des malades de l'échantillon pour les deux premiers modèles de maladie, il est nécessaire de calculer la probabilité $P(G|M)$. En effet, on simule un échantillon rétrospectif : on suppose que l'on choisit les malades en fonction de leur phénotype de maladie. Le calcul de la probabilité $P(G, M = 1)$ est seulement la multiplication de $P(M = 1|G)P(G)$. Pour estimer la probabilité d'être malade, il s'agit seulement de sommer sur tous les génotypes possibles. On en déduit également $P(G|M = 1)$. La simulation du génotype des malades revient à un échantillonnage avec comme probabilité qu'un génotype g soit choisi équivalent à $P(G|M = 1)$. Pour le troisième modèle, le choix de génotype de maladie passe par la sous-population. La probabilité de génotype selon le statut de malade de ce dernier, soit $P(G = g|M = 1)$, s'obtient de la somme sur toutes les sous-populations i de $P(S = i, G = g|M = 1) = P(S = i|M = 1)P(G = g|S = i, M = 1)$. Afin d'obtenir un échantillon numérique respectant ces probabilités, on peut choisir aléatoirement une sous-population selon la probabilité qu'un individu malade provienne de cette sous-population et par la suite, choisir le génotype de l'individu selon la connaissance de son statut de malade et de sous-population.

Pour obtenir le génotype des pseudocontrôles, la situation est simple : s'il n'y a pas de stratification de la population, la probabilité du génotype non transmis est exactement la probabilité du génotype dans la population, sans égard à la maladie (voir l'équation (4.9)). Par contre, quand il y a hétérogénéité à l'intérieur de la population, le génotype du pseudo-individu correspond à un génotype pris au hasard dans une population formée d'une proportion $P(S|M = 1)$ pour chaque population S. On utilisera ces résultats afin de simuler les données. On simulera N cas de la population des malades et, en proportion équivalente aux proportions de malades dans chaque sous-population, on prendra au hasard N individus de ces sous-populations pour en former les pseudos-contrôles.

Afin de comparer les résultats obtenus par la régression logique et le MDR sur des cas et leurs parents, il sera également nécessaire d'effectuer des simulations de données de cas et de contrôles, choisis indépendamment. Le choix des cas est identique à ce qui a été décrit précédemment alors que le choix des contrôles est une procédure quasi identique au choix des cas mais en remplaçant $P(G|M = 1)$ par, bien sûr, $P(G|M = 0)$. Une fois tous les jeux de données simulés, chacun de ceux-ci sera analysé par la régression logique Monte Carlo ainsi que par réduction multifactorielle de dimensionalité.

4.4. Application de la régression logique Monte Carlo et de la RMD

On suivra la méthode utilisée dans le chapitre trois de ce mémoire pour l'exemple de régression logique. Faisons un petit rappel des paramètres d'utilisation. Une chaîne de Markov Monte Carlo de 5000000 d'itérations sera produite pour chaque jeu de données. La distribution a priori des grandeurs de modèles visités correspond à $P(g = i) = \frac{1}{2}^i$ où g est la grandeur de modèle. Également, on omet des résultats les 10000 premières itérations. On impose un maximum de trois arbres et neuf feuilles aux modèles pour d'autant plus encourager la parcimonie. Nous utiliserons également le logiciel prévu à l'effet de la méthode MDR afin d'analyser tous les résultats de simulations obtenus. L'objectif d'une telle manoeuvre est double : montrer que l'applicabilité de l'utilisation de l'entier génome non transmis comme pseudocontrôle s'étend en dehors du contexte de la régression logique et également de consolider les résultats obtenus pour les différents modèles. L'analyse des résultats se fera tels les exemples du chapitre trois.

4.5. Résultats

4.5.1. Régression Logique

Pour chaque modèle, on présentera les résultats de la régression logique au départ et ceux de la réduction multifactorielle de dimensionalité par la suite. Pour la régression logique, on présentera les résultats sous forme de proportion

d'échantillons où a été détectée l'interaction modélisée. Pour un échantillon, la détection d'une interaction est déclarée si la proportion de modèles visités par la chaîne de Markov qui incluent l'interaction dans un même arbre logique, surpasse un certain seuil pouvant varier entre zéro et un. Comme cette détection dépend d'un seuil, on examinera en premier lieu un graphique mettant en relation la proportion de détection et le seuil, pour chaque taille échantillonnale. Ces graphiques sont obtenus par l'utilisation de la sortie du logiciel de régression logique et des fonctions du logiciel R. On s'attend à ce que la proportion de détection augmente avec la taille échantillonnale et diminue avec l'augmentation du seuil. Pour examiner les possibles biais ainsi que l'effet de la stratification de population, on devra également tenir compte de la présence de faux positifs.

Soit i le seuil en pourcentage de détection, alors pour l'effet d'un gène sans interaction, il y aura détection de celui-ci si la proportion de modèles l'incluant dans n'importe quel arbre surpasse $i \in \{1\%, ..., 100\%\}$. Dans l'exemple du chapitre 3, si on se donne un seuil de détection de 15%, on aurait jugé que les variables V2, V4 et V20 ont été détectées, tel que le montre la figure 3.8. Pour un effet d'interaction double (triple), il y aura détection de celui-ci si la proportion de modèles incluant les deux variables prédictives dans n'importe quel arbre surpasse $i - 5\%$ $(i - 10\%)$. On pose des seuils de détection plus bas pour les interactions pour tenir compte du fait qu'il existe un beaucoup plus grand nombre possible de celles-ci qu'il y a de composantes simples : il est beaucoup plus probable qu'une variable soit incluse par chance dans un modèle qu'une interaction. Dans l'exemple du troisième chapitre, on a indiqué qu'il est possible d'extraire des résultats de la régression logique Monte Carlo, la proportion de fois qu'une interaction fait parti du modèle des 5,000,000 d'itérations. On aurait alors vu que les variables V2 et V4 sont répertoriés comme étant en interaction une proportion de fois plus grande que 10% : l'interaction entre V2 et V4 est jugée détectée pour le seuil $i = 15\%$.

Notons qu'il est difficile de suivre le nombre exact de faux positifs. En effet, soit une variable V_1 qui se retrouve toujours dans le même arbre avec deux autres variables V_2 et V_3 une proportion de fois surpassant le seuil i. On comptera un faux

positif correspondant à une interaction triple. Or, cette variable se retrouve nécessairement également dans le même arbre avec une autre des deux autres variables de l'interaction triple (soit V_2 ou V_3) et ce, une proportion de fois surpassant le seuil i. Ceci veut dire que l'on comptera deux interactions doubles en plus de l'interaction triple, comme faux positifs. Finalement, la variable V_1 se retrouvera globalement une proportion de fois surpassant le seuil, pour un faux positif additionnel d'effet simple. Or, seul l'effet triple est véritablement faux positif. Ainsi, si le nombre de faux positifs est inexact en ce sens, la comparaison de ce nombre pour les différentes tailles échantillonnales et entre les plans d'échantillonnage est tout de même instructive puisque le genre de situation décrit précédemment peut survenir qu'importe l'échantillon analysé.

Les résultats sont détaillés de la façon suivante, les figures 4.1 à 4.3 correspondent aux résultats pour le premier modèle, les figures 4.4 à 4.6 sont les résultats du deuxième modèle et les trois dernières figures, soit les figures 4.7 à 4.9 représentent le troisième modèle. Les figures 4.1, 4.2, 4.4, 4.5, 4.7 et 4.8 possèdent tous trois des valeurs de seuil en absisse et une valeur de proportion de modèles ou l'interaction a été détectée (en fonction du seuil), moyennée sur les 500 échantillons, en ordonnée. On permet la présence de faux positifs pour les figures 4.1, 4.4 et 4.7 mais pas pour les figures 4.2, 4.5 et 4.8. Finalement, les autres figures rapportent le nombre moyen de faux-positifs détectés (en fonction du seuil). Les modèles 1 et 2 représentent des situations similaires : il n'y a pas de stratification allélique de la population, et les résultats seront donc traités ensemble. Le troisième modèle est un modèle avec stratification allélique à l'intérieur de la population et sera analysé en dernier.

Sans stratification de la population, on veut inspecter si la régression logique Monte Carlo semble convergente pour la détection de l'interaction modélisée, si le nombre de triades nécessaires à la détection est plus petite plus grande ou semblable au cas d'échantillons cas-contrôle et finalement si une forme de biais empêche le modèle sans faux positif de tendre vers le modèle à la structure véritable. En ce qui concerne la convergence des deux premiers modèles, les figures 4.1, 4.3, 4.4 et 4.6 sont unanimes : si on se fixe un seuil de détection particulier, plus

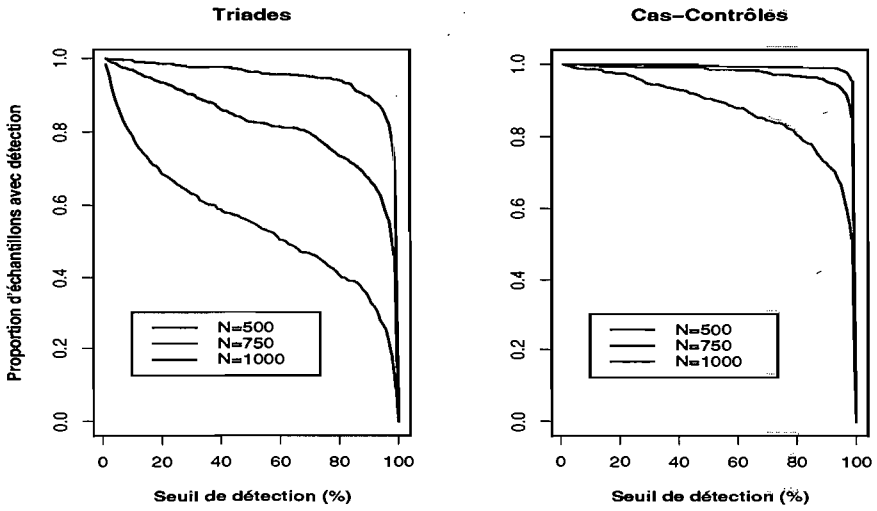

FIG. 4.1. Détection de l'interaction double de génotypes en fonction du seuil pour le premier modèle

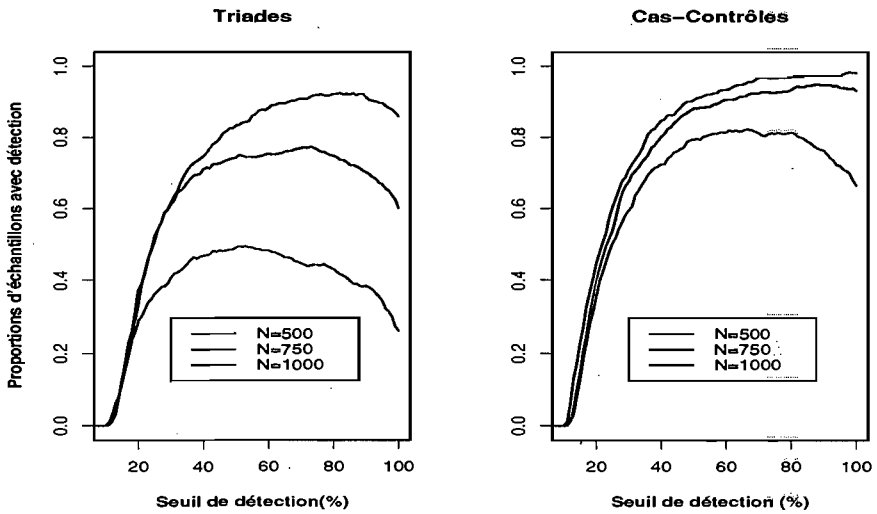

FIG. 4.2. Détection de l'interaction double de génotypes sans présence de faux positifs en fonction du seuil pour le premier modèle

le nombre de triades augmente, plus la probabilité que l'interaction soit détectée grandit et ce, sans égard au nombre de faux positifs. Ceci semble indiquer une forme de convergence. Qui plus est, cette probabilité de détection semble tendre vers 1 pour un large spectre de seuils de détection. Ces mêmes figures nous font remarquer une autre chose : le nombre de triades nécessaire afin d'obtenir une

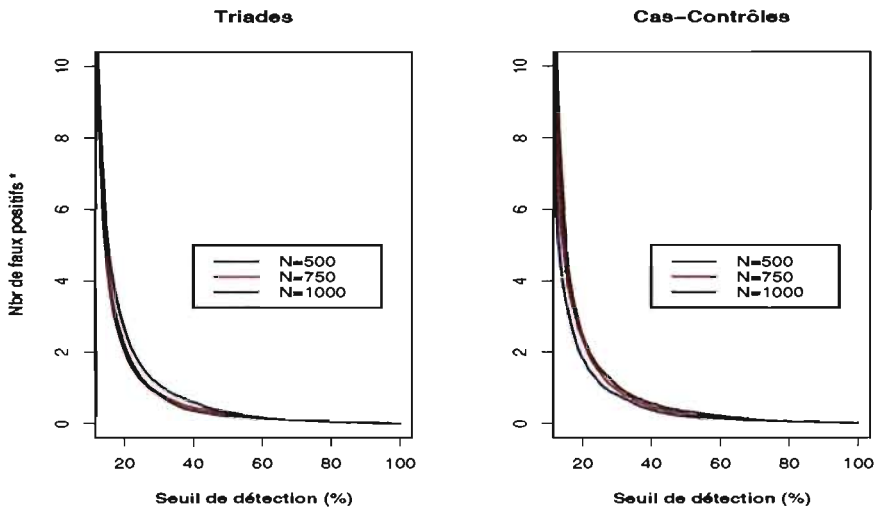

FIG. 4.3. Nombre de faux positifs tels que définis au premier paragraphe de la présente section en fonction du seuil pour le premier modèle

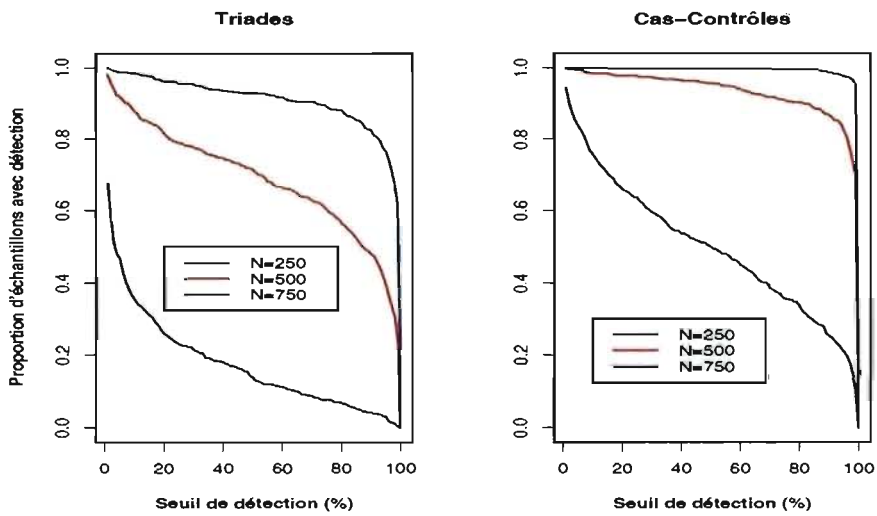

FIG. 4.4. Détection de l'interaction triple de génotypes en fonction du seuil pour le deuxième modèle

probabilité de détection de l'interaction similaire aux échantillons cas-contrôle correspond environ au nombre de ces contrôles additionné de 250. Ceci nous indique que la "puissance" de détection pour les échantillons cas-pseudocontrôle est plus petite que pour les échantillons cas-contrôles. Le comportement du nombre de

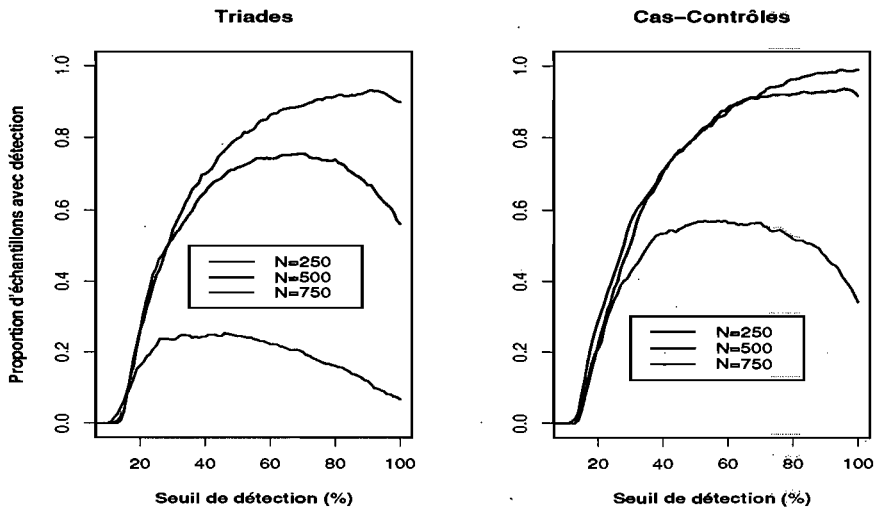

FIG. 4.5. Détection de l'interaction triple de génotypes sans présence de faux positifs en fonction du seuil pour le deuxième modèle

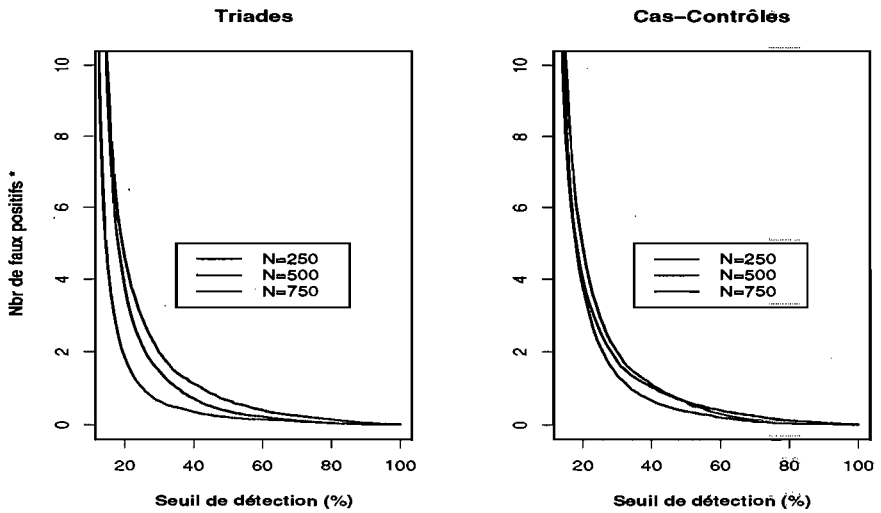

FIG. 4.6. Nombre de faux positifs tels que définis au premier paragraphe de la présente section en fonction du seuil pour le deuxième modèle

faux positifs avec l'augmentation de la taille échantillonnale, tels que présenté aux figures 4.3 et 4.6 n'est pas tel qu'attendu pour les échantillons cas-pseudocontrôles. On aurait espéré que le nombre moyen de faux positifs diminue toujours avec la taille échantillonnale. Ce n'est pas ce qui a été observé. Dans les deux modèles,

le nombre moyen de faux positifs a augmenté avec la taille échantillonnale pour les triades. Pour les échantillons cas-contrôles, une situation semblable se produit pour le deuxième modèle. Ceci pourrait être dû à des détections partielles d'effets; on a remarqué par simulations préliminaires (mais sans qu'on puisse le démontrer) que si une variable comprise dans le modèle de maladie est sélectionnée dans un arbre logique avec une grande récurrence dans la chaîne de Markov, il y a un plus grand risque qu'une variable non impliquée dans la maladie soit ajoutée à cet arbre par chance que cette même variable soit incluse dans un nouvel arbre. Dans tous les cas, la situation des faux positifs n'est pas exclusive aux échantillons de triades, ce qui signifie que le problème ne réside probablement pas dans l'échantillon, mais plutôt à l'intérieur même de la régression logique Monte Carlo. Dans tous les cas, l'augmentation du nombre de faux positifs n'a pas affecté la probabilité de détection de modèles sans faux positif, comme le démontre les figures 4.2 et 4.5.

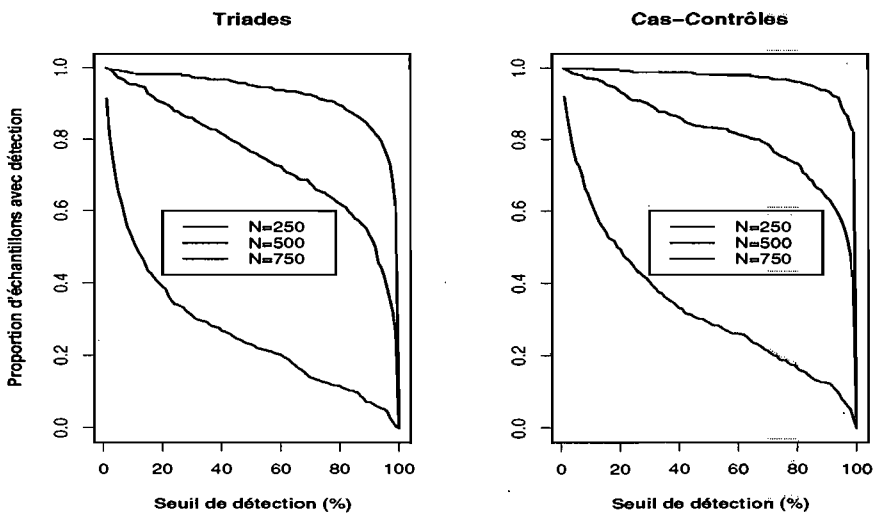

FIG. 4.7. Détection de l'interaction double de génotypes en fonction du seuil pour le troisième modèle

En ce qui concerne le troisième modèle, le modèle avec stratification allélique dans la population. Pour les échantillons de triades, les figures 4.7 et 4.9 portent à la même conclusion que pour les deux premiers modèles concernant la convergence : la probabilité de détection semble tendre vers un avec l'augmentation

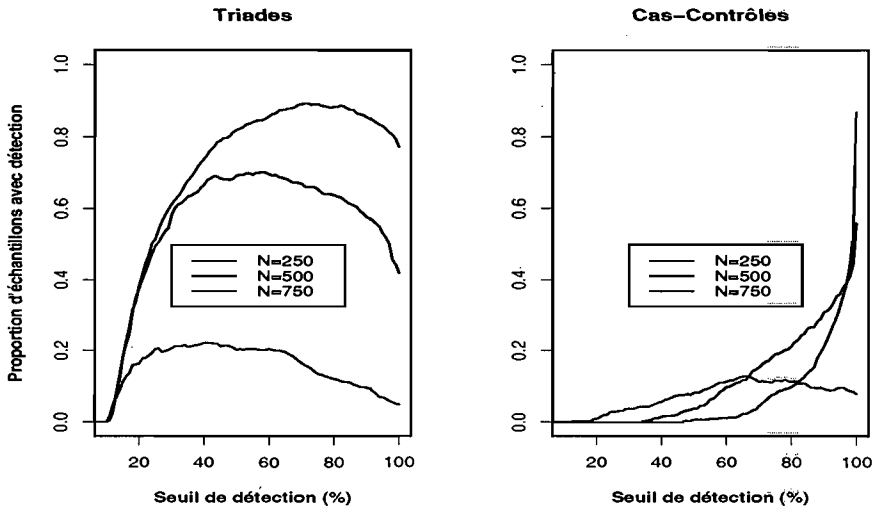

FIG. 4.8. Détection de l'interaction double de génotypes sans présence de faux positifs en fonction du seuil pour le troisième modèle

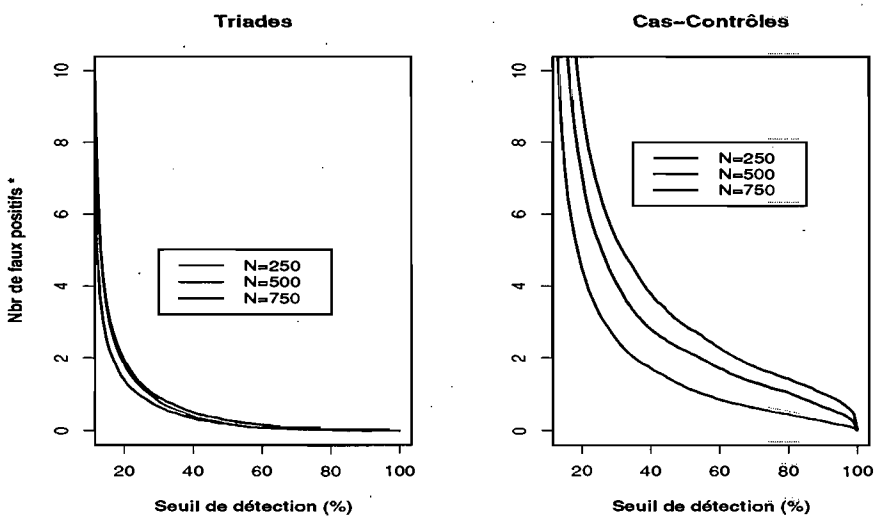

FIG. 4.9. Nombre de faux positifs tels que définis au premier paragraphe de la présente section en fonction du seuil pour le troisième modèle

de la taille échantillonnale. En fait, la situation du troisième modèle est pratiquement identique aux situations des deux premiers modèles, à l'exception des échantillons cas-contrôles. Vous l'aurez deviné, la régression logique Monte Carlo

sur les échantillons cas-contrôle ont été moins robuste face à la stratification allélique de la population. En ce qui concerne les fréquences d'allèles qui ne sont pas impliqués directement ou indirectement dans la maladie, la différence attendue est nulle entre cas et pseudos-contrôles alors qu'elle ne l'est pas entre cas et contrôles (puisque provenant de sous-populations différentes). En conséquence le comportement du nombre de faux positifs est différent par rapport aux deux premiers modèles (voir figure 4.9) et surtout, la probabilité de détection du vrai modèle sans faux positif décroît avec la taille échantillonnale pour les échantillons cas-contrôles alors qu'elle croît pour les échantillons de triades (figure 4.8). Cette conclusion est importante puisque le principal objectif des méthodes faisant usage de triades est d'obtenir une probabilité plus faible de résultat trompeur lorsqu'il y a stratification de la population. La puissance de détection, lorsque l'on permet les faux positifs, est semblable entre les deux types d'échantillons, tel que le démontre la figure 4.7. Par contre, beaucoup de variables peuvent influencer la puissance de détection. En ce sens, il n'est pas difficile d'imaginer des situations où soit les échantillons cas-contrôles ou bien de triades se comportent mieux au niveau de la détectabilité sans tenir compte des faux positifs. On n'a qu'à imaginer une situation où la stratification allélique de la population augmente les probabilités d'obtenir un génotype de susceptibilité pour la maladie. Le nombre de faux positifs peut avoir eu une influence sur la capacité de détection des interactions mais il n'est pas possible de juger si c'est la différence de fréquences des allèles aux marqueurs impliqués dans la maladie ou celle aux marqueurs qui ne sont pas impliqués (et qui engendre des faux positifs) qui a l'influence la plus marquante sur les fréquences de détection.

4.5.2. MDR : Réduction multifactorielle de dimensionalité

Les résultats peuvent se résumer en un tableau pour chaque modèle. On a jugé qu'une interaction double était détectée si elle faisait partie d'un modèle sélectionné par la méthode MDR avec une constance de validation croisée de plus de 5/10 et si le calcul de l'entropie pour la présence des deux variables en interaction est cohérent avec une interaction (voir Jakulin et Bratko (2003)). En ce

qui concerne l'interaction triple, on la considère comme détectée si deux des trois interactions doubles sous-jacentes à une interaction triple sont détectées (donc en interaction les uns avec les autres). Seuls 100 des 500 expériences numériques seront analysés par la méthode MDR puisque l'analyse de celle-ci ne peut pas être automatisée. On présentera des tableaux de la proportion d'échantillons dont le modèle sélectionné par la méthode contient l'interaction. Un total de 21 heures ont été nécessaires à l'analyse des 600 échantillons numériques. Le calcul du nombre de faux positifs aurait été trop fastidieux pour de telles analyses et est omis. Dans le chapitre de discussion, il sera écrit quelques remarques quant aux tableaux présentés immédiatement.

TAB. 4.3. Proportion de détection pour les échantillons numériques - Modèle 1

N	Prop. Triades	Prop. Cas-Contrôles
500	47/100	77/100
750	63/100	86/100
1000	75/100	93/100

TAB. 4.4. Proportion de détection pour les échantillons numériques - Modèle 2

N	Prop. Triades	Prop. Cas-Contrôles
250	15/100	17/100
500	22/100	44/100
750	42/100	59/100

TAB. 4.5. Proportion de détection pour les échantillons numériques - Modèle 3

N	Prop. Triades	Prop. Cas-Contrôles
250	22/100	24/100
500	47/100	49/100
750	60/100	46/100

La méthode RMD, quant à elle, ne peut être jugée que pour sa capacité de détection de l'interaction pour les raisons décrites un peu plus haut dans ce chapitre. Des résultats tout à fait similaires avec la régression logique sont observés dans les tableaux 4.3, 4.4 et 4.5. Par contre, si la proportion d'échantillons permettant d'interpréter en un effet de l'interaction sur la maladie est plus petite pour la méthode RMD, il ne faut pas en juger pour autant que la régression logique est plus efficace que la RMD. En effet, les critères de sélection établis subjectivement dans ce chapitre ne sont pas comparables entre les méthodes. Finalement, sans pouvoir baser cette proposition sur une statistique, le nombre de variables sélectionnées pour les échantillons de 750 triades pour le troisième modèle, semblait en moyenne être plus bas que le nombre de variables associées aux échantillons de 750 cas et 750 contrôles, suggérant une robustesse accrue des triades pour la stratification.

Un résumé des résultats de simulations obtenus est fourni dans le tableau 4.6. Nous suggérons au lecteur de comparer ce tableau avec le tableau 4.1. À noter qu'un total d'environ 1080 heures de calculs pour un processeur de 2 GHz a été nécessaire afin d'obtenir les résultats numériques.

TAB. 4.6. Résultats obtenus d'application de régression logique ou RMD

Modèle	Structure ?	Converge ?	Puissance	Robustesse
Mod.1 (RL)	Non-biaisée	Oui	$<$ Cas-Contrôle	-
Mod.2 (RL)	Non-biaisée	Oui	$<$ Cas-Contrôle	-
Mod.3 (RL)	Non-biaisée	Oui	$\approx$ Cas-Contrôle	$>$ Cas-Contrôle
Mod.1 (RMD)	-	Oui	-	-
Mod.2 (RMD)	-	Oui	-	-
Mod.3 (RMD)	-	Oui	-	-

Chapitre 5

APPLICATION SUR DES DONNÉES RÉELLES

5.1. Considérations quant au champ d'application

Comme les résultats satisfont nos attentes, on discutera des différents plans d'expériences où la méthode proposée dans ce mémoire peut s'avérer utile. Plusieurs autres plans d'expériences n'ont pas été testés dans ce mémoire, mais l'extension à un certain nombre d'entre eux est naturelle. À ce point, on fera appel à plusieurs éléments du premier chapitre qui n'ont pas été visités régulièrement.

Les deux dernières sources de variation que l'on va considérer dans cette section sont le déséquilibre de liaison entre marqueurs ainsi que l'importance que peut avoir la phase d'haplotype. Dans les données sur le déficit d'attention et d'hyperactivité ces situations se produisent. En ce sens, cette discussion prépare l'analyse de données liées au trouble du déficit de l'attention et d'hyperactivité. Ces deux phénomènes n'ont jamais été considérés lors des simulations. Notons que cette partie sera plus ardue au statisticien, mais elle fait appel à des concepts définis au premier chapitre. Le statisticien pourra ainsi seulement se référer aux conclusions du prochain développement.

Dans la modélisation, notons qu'on a toujours supposé qu'il y avait équilibre de liaison entre les marqueurs (rappelons que cela signifie que la présence d'un génotype à un des marqueurs n'apporte aucune information sur le génotype à un deuxième marqueur). Si ce n'est pas le cas, c'est-à-dire qu'il y a une corrélation entre les allèles de loci différents, à quel genre de résultats devrions-nous nous attendre de la régression logique ? Afin de se faire une idée grossière de

la réponse à cette question, supposons deux loci bi-alléliques en déséquilibre de liaison. Par exemple, le premier locus peut abriter deux allèles que l'on nomme "bleu" et "rouge" alors que le deuxième locus abrite les allèles "vert" et "jaune". Le tableau 5.1 illustre les probabilités conjointes d'allèles au premier et deuxième locus pour un même haplotype pour la population de gamètes des parents. Selon ce tableau, la probabilité d'allèles "bleu" et "rouge" au premier locus est de 60% et de 40% respectivement alors que pour le deuxième locus, "vert" et "jaune" ont probabilités 50% et 50%. On voit qu'il y a déséquilibre de liaison puisque par exemple la présence de "rouge", au premier locus nous indique qu'il y a plus de chance qu'il y ait "vert" au deuxième locus, alors que sans déséquilibre de liaison on aurait conclu qu'il y a autant de chance que "vert" ou "jaune" occupe le deuxième locus. Supposons par la suite trois situations : la maladie affecte autant

TAB. 5.1. Probabilités d'haplotypes

haplotype	Probabilité dans la pop. de gamètes
(bleu,vert)	20%
(bleu,jaune)	40%
(rouge,vert)	30%
(rouge,jaune)	10%

les individus possédant n'importe quelle combinaison d'haplotype (notamment, ni le premier locus, ni le deuxième locus, n'est en déséquilibre de liaison avec un locus de susceptibilité pour la maladie), la maladie affecte plus les individus possédant un génotype particulier à un locus en déséquilibre de liaison avec deux marqueurs étudiés et finalement la maladie affecte plus les individus possédant un génotype particulier à un marqueur et un autre génotype particulier à un deuxième marqueur en déséquilibre de liaison avec le premier. Étudions les résultats attendus de génotypes d'individus pour la population de cas ainsi que de pseudocontrôles.

Le premier cas est relativement facile à étudier : si le génotype de l'individu aux deux loci n'augmente pas le risque de maladie, alors $P(M = 1|G = g) = P(M = 1)$ (où G est la variable représentant le génotype d'un individu aux deux

loci) de sorte que $P(G = g|M = 1) = P(M = 1|G = g) * P(G = g)/P(M = 1) = P(G = g)$ et on ne s'attend pas à observer de différence de génotypes d'un individu malade à un individu qui ne l'est pas ou bien d'un individu pris aléatoirement de la population. Si par chance un allèle est associé à la maladie, alors il est fort possible qu'un deuxième allèle avec lequel il est en déséquilibre de liaison soit aussi associé à la maladie. En effet, s'il existe une différence de fréquences (entre cas et pseudocontrôles ou cas et contrôles) pour le premier locus pour un allèle, il y aura différence de fréquences pour le deuxième locus avec l'allèle auquel il est le plus fortement associé. Par contre, les auteurs Kooperberg et Ruczinski (2005) ont observé qu'un faux positif risque de ne pas être sélectionné avec une variable d'un locus duquel il est en déséquilibre de liaison dans les résultats de régression logique. Par contre, il est possible qu'un génotype d'allèle auquel il est associé soit sélectionné également, mais pas dans un même modèle. La situation où un seul locus est impliqué dans la maladie est semblable au cas précédent c'est-à-dire au cas d'association par chance; la détection a probabilité d'apparaître sous forme du locus 1 ou bien du locus 2 mais pas dans le même arbre logique, donc pas en interaction. Finalement, la situation où les deux loci sont impliqués différemment dans la maladie et ce, en interaction biologique, il est plausible d'estimer que la fréquence d'interaction soit plus élevée dans la population des cas par rapport à la population de pseudocontrôles. Par contre, l'ampleur de l'effet du déséquilibre de liaison sur la détectabilité reste un sujet qui devrait être plus profondément étudié, à la fois pour la régression logique et la réduction multifactorielle de dimensionalité.

Pour l'instant, on conclura qu'il faudrait porter une attention particulière aux résultats de sélection de modèles lorsqu'il y a déséquilibre de liaison. Il n'y a pas de raison de croire que ce déséquilibre de liaison affecte différemment une étude cas-contrôles et une étude de triades puisque ces derniers échantillons sont comparables aux échantillons cas-contrôles pris de populations similaires, tel que discuté plus tôt. Donc, des résultats de Kooperberg et Ruczinski (2005) et de la présente discussion, on interprètera que si les résultats obtenus impliquent un ou plusieurs marqueurs en déséquilibre de liaison, sans interaction entre ceux-ci, alors

il est possible qu'un seul marqueur soit impliqué dans la maladie. Également, on interprètera que si les résultats obtenus impliquent un ou plusieurs marqueurs en déséquilibre de liaison, mais cette fois-ci en interaction, une analyse plus approfondie de l'interaction est souhaitable.

Dans la modélisation, on n'a jamais considéré la possibilité que la phase d'un haplotype peut avoir une importance sur l'expression de la susceptibilité d'un individu face à la maladie. Afin de présenter ce dernier concept, considérons deux marqueurs sur le même gène. Comme ils sont sur le même gène, ceci implique qu'ils forment tous deux une partie du code qui sert de matrice pour la même structure ou protéine. Plus spécifiquement, deux allèles sur le même haplotype seront transcrits ensemble s'ils codent pour la même protéine alors que ce ne sera pas le cas s'ils codent pour des protéines différentes. Par exemple (voir figure 5.1), soit 2 loci bi-alléliques sur le même gène. Pour les besoins de la cause, posons à nouveau "bleu" et "rouge" les allèles du premier locus et "vert" et "jaune" les allèles du deuxième locus. Les 4 haplotypes servent de matrice à une même protéine qui prendra une conformation similaire mais pas nécessairement équivalente puisque le code diffère d'un haplotype à l'autre. Or, la conformation de la protéine est importante pour sa fonction. Par exemple, si on considère qu'une protéine a comme rôle de se lier à une structure grâce à sa partie encerclée dans la figure 5.1, alors seul le deuxième haplotype est dysfonctionnel.

Mais si on ne connaît que le génotype non phasé hétérozygote pour les deux loci, il n'y a aucune façon de savoir si l'individu possède l'haplotype dysfonctionnel. En effet, si un individu possède les génotypes (bleu, rouge) et (vert, jaune) aux loci 1 et 2 respectivement, alors il y a possibilité de deux combinaisons de deux haplotypes phasés. La première combinaison consiste en les premier et quatrième haplotypes de la figure 5.1 alors que la deuxième combinaison consiste en les deuxièmes et troisièmes haplotypes. Seuls les individus possédant la deuxième combinaison d'haplotypes possèdent l'haplotype dysfonctionnel. Si une telle situation se produit, à quel genre de résultat doit-on s'attendre avec une procédure comme la régression logique ou la réduction multifactorielle de dimensionalité ? Réussiront-elles à détecter l'haplotype sous forme d'une interaction ?

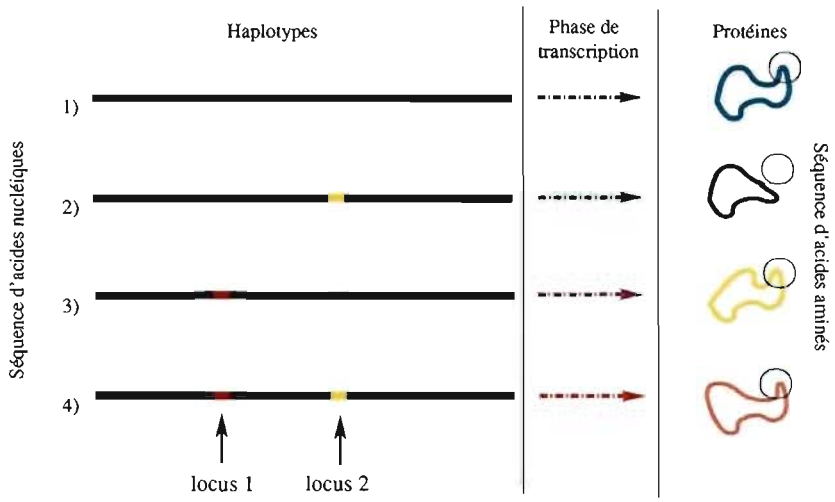

FIG. 5.1. Les 4 haplotypes possibles pour le génotype (bleu, rouge) et (bleu, jaune) et leur produit protéiné.

Premièrement, remarquons que la situation telle que décrite précédemment est comparable, mais pas en tout point identique, à une situation de deux marqueurs en déséquilibre de liaison puisqu'étant sur le même gène et donc à proximité, la probabilité de déséquilibre de liaison est plus grande et en interaction particulière. Pour montrer ceci, présentons sous forme de tableau, les génotypes d'haplotypes phasés et les génotypes non phasés qu'ils impliquent (voir le tableau 5.2). Notons qu'un génotype d'haplotype phasé sera présenté sous la forme (1,2)+(3,4), c'est-à-dire le rassemblement d'un haplotype (1,2) avec un haplotype (3,4) pour former le génotype ((1,3),(1,4)), écrit sous la forme habituelle. On voit que le génotype non phasé doublement hétérozygote est le seul où l'on ne peut pas inférer le génotype phasé. Mais alors, si la situation est telle que seuls les individus possédant le deuxième haplotype ont un risque accru d'être atteints de la maladie, ceci correspond aux individus possédant les génotypes d'haplotypes phasés 2 et 3. Pour les génotypes non phasés, ceci correspond aux cas où l'individu possède dans son génotype au moins une copie de bleu et au moins une copie de vert en n'incluant pas une partie inconnue d'individus doublement hétérozygote. Donc, ne pas effectuer de distinction entre génotype non phasé et haplotypes phasés revient à

TAB. 5.2. Génotypes d'haplotypes et leur génotype non phasé

Num. de génotype	Géno. d'haplotypes	Géno. non-phasé
1	(bleu,vert) + (bleu,vert)	((bleu,bleu),(vert,vert)
2	(bleu,vert) + (bleu,jaune)	((bleu, bleu),(jaune,vert)
3	(bleu,vert) + (rouge,vert)	((bleu,rouge),(vert,vert))
4	(bleu,vert) + (rouge,jaune)	((bleu,rouge),(vert,jaune))
5	(bleu,jaune) + (bleu,jaune)	((bleu,bleu),(jaune,jaune))
6	(bleu,jaune) + (rouge,vert)	((bleu,rouge),(vert,jaune))
7	(bleu,jaune) + (rouge,jaune)	((bleu,rouge),(jaune,jaune))
8	(rouge,vert) + (rouge,vert)	((rouge,rouge),(vert,vert))
9	(rouge,vert) + (rouge,jaune)	((rouge,rouge),(jaune,vert))
10	(rouge,jaune) + (rouge,jaune)	((rouge,rouge),(jaune,jaune))

masquer légèrement l'effet de l'haplotype. Par contre, plus il y a de marqueurs impliqués dans l'haplotype plus l'effet de cet haplotype se masque.

On conclura de cette section que l'absence d'effet de marqueurs en déséquilibre de liaison résulte en l'absence attendue de détection de ces marqueurs dans les modèles sélectionnés. Qui plus est, il est rare que des interactions entre gènes qui sont en déséquilibre de liaison soient détectées par chance : celles-ci devront être investiguées. Notons qu'elles peuvent signifier une interaction entre ces marqueurs, mais également la présence d'un seul locus de susceptibilité qui soit en déséquilibre de liaison avec ces deux premiers marqueurs. Finalement, il est possible que l'effet d'un haplotype important ne soit pas détecté dû à l'absence d'information sur la phase des haplotypes.

5.2. LA MALADIE

Le trouble du déficit de l'attention et hyperactivité (TDAH) appartient aux plus communes maladies surgissant tôt dans l'enfance. Dans le monde, on estime que la maladie touche 1 enfant sur 15 environ. Cette maladie touche principalement les hommes avec environ 4 fois plus de ces derniers que les femmes. Les

symptômes qu'exprime la personne malade varient entre des problèmes d'attention, de concentration, d'impulsivité et d'irritabilité. L'étiologie de la maladie n'est pas très connue mais plusieurs études indiquent une composante génétique non négligeable avec une héritabilité (ou proportion de variation phénotypique explicable par la génétique) estimée à 77% (Biederman et Faraone (2005)). Entre autres, il semble qu'un allèle du gène de transporteur de dopamine, connue dans la littérature comme la répétition 10 du gène DAT, soit répétitivement associé à la maladie dans les études d'association. Par ailleurs, le principal médicament contre le trouble du déficit de l'attention, nommément le méthylphénidate ou MPH, agit comme inhibiteur du transporteur de dopamine ce qui soutient d'autant plus l'hypothèse de l'implication du gène DAT dans la maladie. Une méta-analyse des études d'association du gène DAT rapporte un effet faible mais significatif du gène DAT avec la maladie (Yang et al. (2000)). Depuis, plusieurs autres gènes ont été associés à la maladie mais les résultats ne se répètent pas toujours. Comme nous l'avons expliqué plus tôt dans le mémoire, ne pas considérer d'interaction entre gènes peut causer une apparence de manque de réplicabilité dans de telles études. Nous appliquerons la méthode développée dans ce mémoire afin d'explorer l'interaction entre gènes pour le TDAH d'un jeu de données que nous décrivons ici.

5.3. Les données

Les enfants et adolescents atteints de TDAH ont été sélectionnés par le programme des troubles graves du comportement de l'Institut Douglas ainsi que sur l'ensemble des patients s'étant présentés dans des buts diagnostics à l'Hôpital Douglas. Ont été exclus de l'étude les enfants ou adolescents avec un quotient intellectuel plus petit que 70 ou qui exposaient d'autres affections particulières (syndrome de Tourette, psychoses, autres troubles du développement). Pour tous ces cas, on a collecté les données génétiques sur un ensemble de 55 marqueurs sur 9 gènes différents. Le tableau 5.5 regroupe l'ensemble des marqueurs et des gènes qui seront à l'étude dans ce chapitre. On a également recensé les données génétiques des parents de chaque cas. On obtient ici un jeu de données avec à chaque

ligne l'identificateur de la personne, l'identificateur de son père, l'identificateur de sa mère et pour chaque marqueur la valeur de son génotype à ce marqueur ainsi qu'une dernière variable pour indiquer si la personne est atteinte de la maladie.

Deux sous-échantillons ont été formés à partir d'un échantillon de 338 cas indépendants. Ainsi, le premier sous-échantillon est composé de 169 cas indépendants ainsi que de 52 pseudocontrôles qui ont été soutirés de ces cas (les autres cas avaient des génotypes incomplets pour les pères principalement) sur 54 marqueurs. Le marqueur MAOA a été enlevé des analyses dues à des erreurs récurrentes de génotypage. Le deuxième sous-échantillon regroupe des données de 231 cas indépendants et de 84 pseudo contrôles de ces 231 cas sur 21 marqueurs sélectionnés afin de maximiser le nombre de pseudocontrôles disponibles. En effet, on a été obligé de soustraire de l'échantillon les données provenant d'un individu qui n'a pas toute son information génétique pour tous les marqueurs étudiés. Donc, le fait d'enlever certains des marqueurs à l'étude a permis de garder plus de trios complets pour le deuxième échantillon. Les marqueurs étudiés pour ce deuxième sous-échantillon sont présentés dans le tableau 5.6. A été exclu tout membre de familles où l'on a décelé une erreur mendélienne (génotypes non possibles des parents ou des enfants) pour un ou plusieurs des marqueurs étudiés. Par exemple, on aurait enlevé des données le trio suivant : la mère d'un enfant a génotype "a-a" à un marqueur, le père a un génotype "a-b" à ce marqueur et l'enfant possède le génotype "b-b" au marqueur. Cette impossibilité peut être dû à une mulitude de raisons, expliquant la raison du retrait de ces familles.

Les données regroupent des marqueurs qui possèdent seulement deux allèles ainsi que des marqueurs qui en possèdent plus de deux. Les marqueurs de deux allèles peuvent être représentés par deux variables tel que dans les trois modèles simulés et présentés dans le quatrième chapitre. Par exemple, le marqueur NET1A possède deux allèles, soit les allèles A et G. On utilise deux variables dichotomiques, par exemple X et Y, pour représenter le génotype d'un individu à ce marqueur. La variable X pourrait représenter l'événement "l'individu possède au moins un A dans son génotype au marqueur NET1A" et Y représenterait "l'individu possède deux A dans son génotype au marqueur NET1A". Dans ce

cas, les variables X et Y prennent la valeur 1 si l'événement qu'ils représentent se réalise et 0 sinon. On a représenté chaque marqueur qui n'est pas bi-allélique (DAT, DINT8, DRD4, D4120) par quatre variables. Pour chacun de ces marqueurs, on choisit l'un des deux allèles les plus fréquents chez les parents. On catégorise également les allèles moins fréquents (tous les allèles sauf les deux plus fréquents) en tant que "allèles rares". Les quatre variables associées à ce marqueur sont dichotomiques de valeur 1 lorsque l'événement qu'ils représentent se réalise et 0 sinon. Les deux premières variables représentent respectivement les événements "Au moins une copie de l'allèle de référence est présent au marqueur" et "Deux copies de l'allèle de référence sont présents au marqueur". Les deux dernières variables représentent les mêmes événements pour les "allèles rares". Par exemple, soit le marqueur DAT. Ce marqueur n'est pas bi-allélique puisque les allèles possibles à ce marqueurs sont les allèles "3", "6", "7", "8", "9", "10", "11" et "15" avec les allèles "9" et "10" beaucoup plus fréquents. On fixe l'allèle "9" comme allèle de référence (on aurait aussi pu prendre l'allèle "10"). Alors, un individu de génotype ("9","11") au marqueur DAT aura comme variables associées X1, X2, X3 et X4 de valeurs respectives 1, 0, 1 et 0 respectivement. Un total de 116 variables sont nécessaires pour décrire le génotype d'un individu aux 54 marqueurs étudiés pour le premier échantillon. Un total de 44 variables sont nécessaires pour décrire les 21 marqueurs étudiés pour le deuxième échantillon. Ces variables dichotomiques serviront de variables prédictives pour les régressions logiques. Chaque sous-échantillon sera étudié par la régression logique avec exactement les mêmes paramètres que lors des simulations (revoir la section d'application de la régression logique Monte Carlo). Le nom de ces variables associées à chaque marqueur est présenté dans les tableaux 5.3 et 5.4. Ces tableaux nous seront utiles afin d'interpréter les résultats de la régression logique Monte Carlo qui va identifier certaines variables comme faisant partie de la structure du modèle de maladie. Par exemple, si pour le premier échantillon la procédure de la régression logique Monte Carlo identifie les variables V5 et V17 très régulièrement dans le même arbre logique au cours de la chaîne de Markov, on pourra suspecter une interaction entre les marqueurs DINT8 et HT1B1.

TAB. 5.3. Les marqueurs et les variables associées pour le premier échantillon

Marqueur	Variables	Marqueur	Variables
DAT	V1 V2 V3 V4	NET14	V63 V64
DINT8	V5 V6 V7 V8	NET15	V65 V66
DRD4	V9 V10 V11 V12	NET16	V67 V68
DEX9	V13 V14	NET17	V69 V70
DINT9	V15 V16	NET18	V71 V72
HT1B1	V17 V18	NET19	V73 V74
D4521	V19 V20	NET20	V75 V76
COMT	V21 V22	NET21	V77 V78
SHTT	V23 V24	NET22	V79 V80
D4120	V25 V26 V27 V28	NET23	V81 V82
NET1A	V29 V30	NET24	V83 V84
NET1B	V31 V32	NET25	V85 V86
NET1C	V33 V34	NET26	V87 V88
NET2A	V35 V36	NET27	V89 V90
NET2B	V37 V38	NET28	V91 V92
NET2C	V39 V40	NET29	V93 V94
NET3	V41 V42	TPH2.1	V95 V96
NET4	V43 V44	TPH2.2	V97 V98
NET5	V45 V46	TPH2.5	V99 V100
NET6	V47 V48	TPH2.6	V101 V102
NET7	V49 V50	TPH2.7	V103 V104
NET8	V51 V52	TPH2.8	V105 V106
NET9	V53 V54	TPH2.9	V107 V108
NET10	V55 V56	SNAP25.1	V109 V110
NET11	V57 V58	SNAP25.2	V111 V112
NET12	V59 V60	SNAP25.3	V113 V114
NET13	V61 V62	SNAP25.4	V115 V116

TAB. 5.4. Les marqueurs et les variables associées pour le deuxième échantillon

Marqueur	Variables	Marqueur	Variables
DAT	V1 V2 V3 V4	NET18	V25 V26
DINT9	V5 V6	NET20	V27 V28
SHTT	V7 V8	NET22	V29 V30
NET1A	V9 V10	NET23	V31 V32
NET1B	V11 V12	NET24	V33 V34
NET1C	V13 V14	NET27	V35 V36
NET6	V15 V16	TPH2.1	V37 V38
NET9	V17 V18	TPH2.2	V39 V40
NET10	V19 V20	TPH2.6	V41 V42
NET14	V21 V22	SNAP25.2	V43 V44
NET16	V23 V24	-	-

TAB. 5.5. Tous les gènes et les marqueurs sur ces gènes

Gène/Protéine	Nom des marqueurs
Transporteur de dopamine	DINT8, DINT9, DEX9, DAT
Récepteur de dopamine D4	DRD4, D4521, D4120
Monoamine Oxydase A	MAOA
5-hydroxytryptamine (sérotonine) récepteur	HT1B
Catéchol-O-méthyle transférase	COMT
Transporteur de Sérotonine	5HTT
Transporteur Norépinéphrine	NET1A,1B,1C,2A,2B,2C,3 à 29
Tryptophan hydroxylase 2	TPH2.1,.2,.5,.6,.7,.8,.9
synaptosome 25	SNAP25.1,.2,.3,.4

TAB. 5.6. Les gènes et les marqueurs inclus dans le deuxième échantillon sur ces gènes

Gène/Protéine	Marqueurs inclus dans le 2e sous-éch.
Transporteur de dopamine	DAT, DINT9
Récepteur de dopamine D4	-
Monoamine Oxydase A	-
5-hydroxytryptamine (sérotonine) récepteur	-
Catéchol-O-méthyle transférase	-
Transporteur de Sérotonine	5HTT
Transporteur Norépinéphrine	NET1A à C,6,9,10,14,16,18,20,22 à 24,27
Tryptophan hydroxylase 2	TPH2.1,.2,.6
synaptosome 25	SNAP25.2

5.4. RÉSULTATS

Comme on a vu au chapitre trois, les résultats de la régression logique Monte Carlo sont principalement les proportions de présence de variables seules, couples de variables et trios de variables sur l'ensemble des modèles visités par la chaîne de Markov. Une représentation graphique d'afficher de tels résultats a été présenté dans ce même chapitre. Les résultats seront présentés de manière similaire dans cette section pour le premier ainsi que le deuxième échantillon. On ajoutera également pour chaque échantillon les 5 variables qui sont apparus le plus régulièrement en singleton, les 5 couples de variables qui sont apparus le plus régulièrement en interaction double ainsi que les 5 triplets de variables qui sont apparus le plus régulièrement en interaction triple au cours de la régression logique Monte Carlo. Dans les analyses, on se donnera un seuil de 15%, qui équivaut à l'un des deux seuils utilisés dans Kooperberg et Ruczinski (2005). Puisque de petites tailles échantillonnales sont étudiées, le danger de détection de faux positifs est élevé, notamment pour de très petits seuils de détection. Il faudra être particulièrement prudent dans les analyses.

Pour le premier échantillon comprenant tous les marqueurs sauf le marqueur MAOA, les résultats sont présentés dans les figures 5.2 à 5.4 ainsi qu'aux tableaux 5.7 à 5.9. La première des figures (figure 5.2) nous donne une idée de la taille potable du modèle qui représente le mieux les données. Cette figure nous indique également si un quelconque signal a été détecté par la régression logique Monte Carlo. Si la taille du modèle varie beaucoup, ceci pourrait être indicateur d'une absence de signal dans le modèle car dans ce cas, ajouter ou enlever une variable au modèle ne modifie pas les vraisemblances du modèle. Dans le cas étudié, la taille du modèle semble varier principalement entre 1 et 2, indiquant un signal et que peu de variables semblent impliquées dans le modèle de la maladie. Les figures 5.3 et 5.4 nous donnent une indication de la distribution des proportions de variables et couples sélectionnés. Une absence de signal se traduirait par des distributions presque uniformes, la chaîne de Markov impliquant de manière aléatoire une variable puis une autre, sans préférence. On voit des figures 5.3 et 5.4 qu'il semble y avoir une variable plus fréquemment visitée, soit la variable V55,

et possiblement un couple de variable, ce qui est représenté par une case blanche dans la figure 5.4 indiquant une fréquence relativement plus élevée.

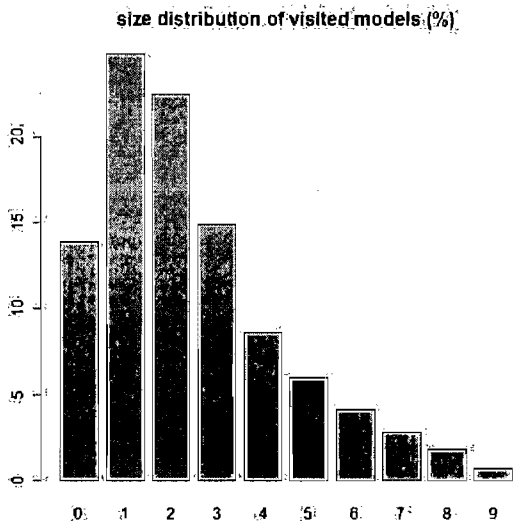

FIG. 5.2. Taille des modèles visités en nombre de variables - Échantillon 1

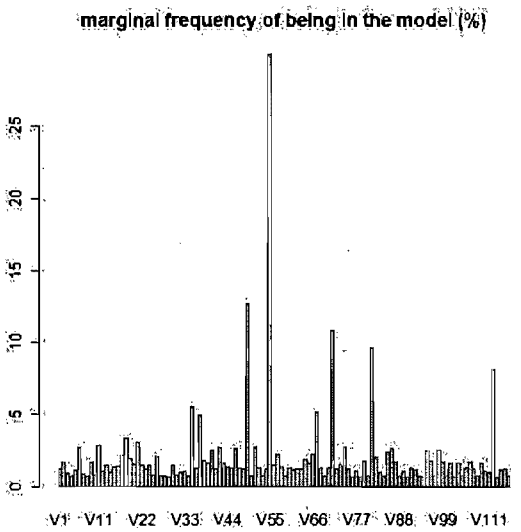

FIG. 5.3. Fréquences marginales d'inclusion dans les modèles pour chaque variables - Échantillon 1

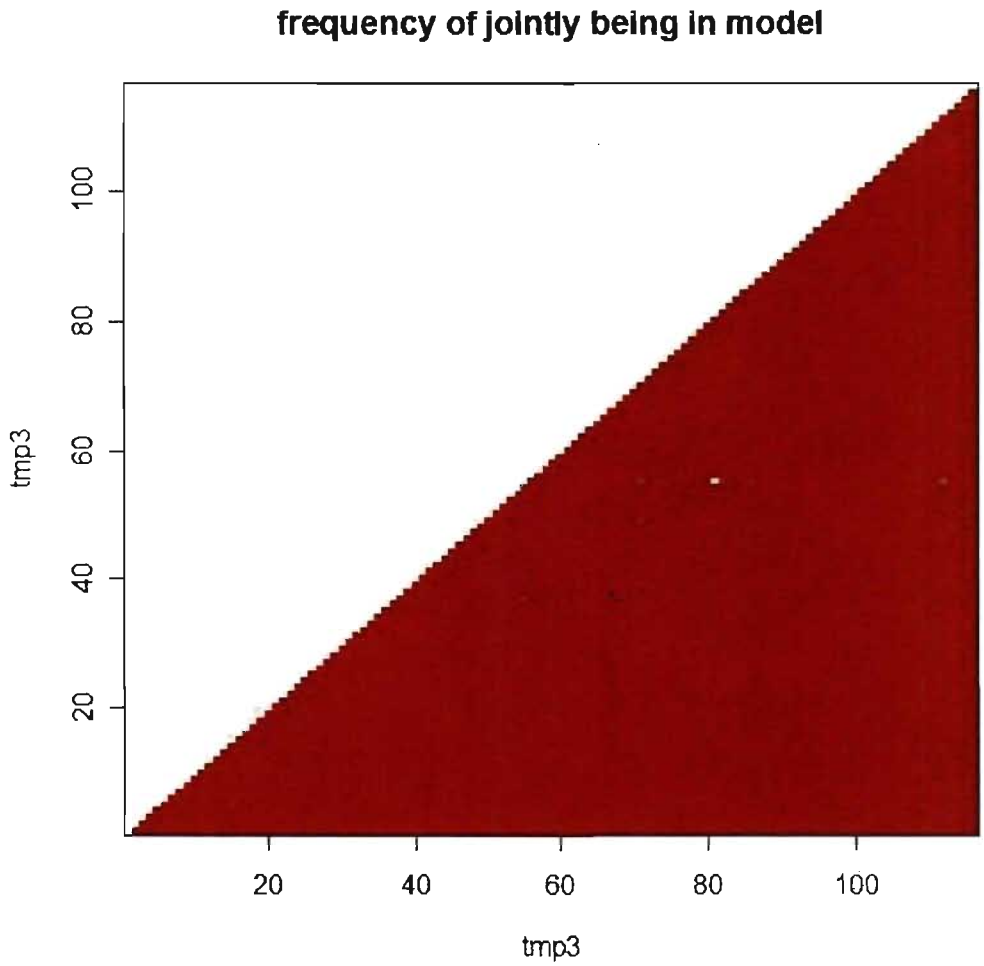

FIG. 5.4. Représentations des fréquences d'inclusions de couples de variables - Échantillon 1

Le tableau 5.7 confirme que la variable V55 est le plus souvent impliquée dans les modèles visités. Sa proportion d'inclusion dépasse le seuil que nous nous sommes fixé avec une proportion d'inclusion d'autour de 29.94 %. Par contre, les tableaux 5.8 et 5.9 montrent des proportions d'inclusion bien en-deça du seuil de détection pour les interactions doubles et triples. La prudence nous indiquerait ainsi que la variable V55 est le seul signal détecté par la procédure de la régression logique Monte Carlo pour le premier échantillon. Le modèle de régression pour la maladie est représenté par la figure 5.5.

TAB. 5.7. Les cinq variables qui reviennent le plus souvent - Échantillon 1

Variable	Proportion de présence (en %)
V55	29.94066
V49	12.65714
V71	10.83492
V81	9.57260
V112	8.09824

TAB. 5.8. Les cinq couples de variables qui reviennent le plus souvent - Échantillon 1

Couple de variables	Proportion de présence dans le même arbre (en %)
V55 V81	6.06626
V55 V112	2.29904
V55 V71	1.92556
V37 V55	1.75994
V55 V86	1.45688

TAB. 5.9. Les cinq triplets de variables qui reviennent le plus souvent - Échantillon 1

Triplets de variables	Proportion de présence dans le même arbre (en %)
V55 V81 V87	0.22312
V37 V55 V71	0.21278
V55 V81 V112	0.20714
V35 V55 V71	0.18214
V2 V55 V71	0.18156

La variable 55 représente la présence d'au moins un "C" au génotype de Net10, soit le marqueur de nomenclature scientifique rs192303. On interprète que le modèle sélectionné conclut en une augmentation de la susceptibilité face à la maladie

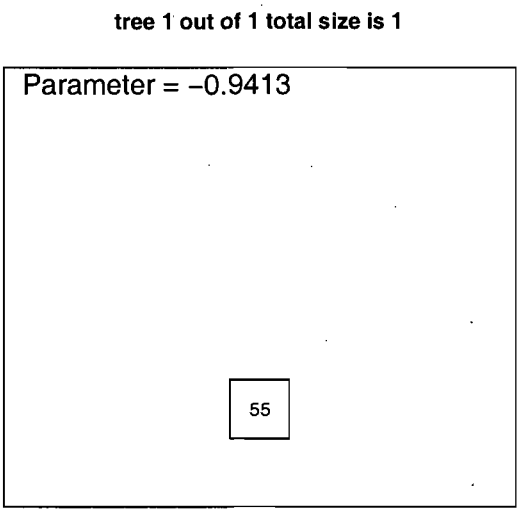

FIG. 5.5. Modèle pour le premier échantillon

associée à la possession du génotype "GG" au marqueur rs192303. On a calculé la proportion de cas possédant ce génotype dans le premier échantillon qui vaut environ 53.25% alors que la proportion de contrôle est de 30.77%. Notons également que la variable sélectionnée a un effet modéré (on rappelle que l'exponentiel de 0.9413 équivaut presque au risque relatif) et est en déséquilibre de liaison avec tous les autres marqueurs NETs. Vu la taille de l'échantillon et spécialement le nombre de pseudocontrôles, on concluera qu'une des variables NETs est potablement responsable directement ou indirectement d'une partie de la susceptibilité face à la maladie. Dans tous les cas, il serait fort aventureux de conclure en une quelconque interaction avec de tels résultats.

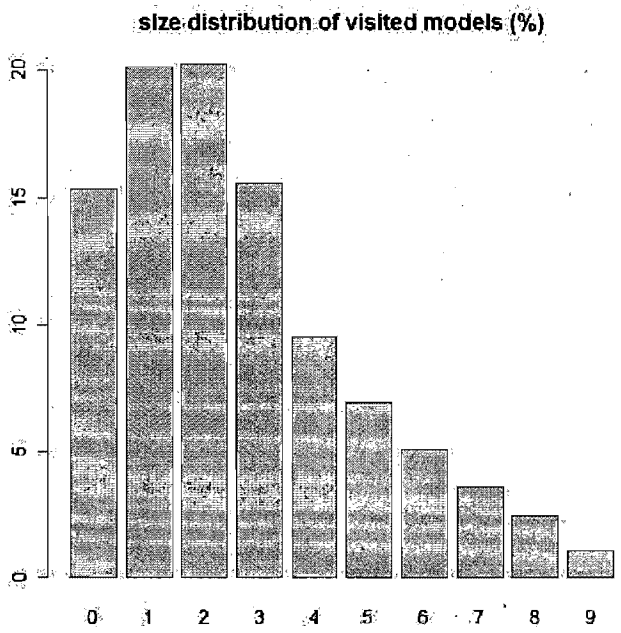

FIG. 5.6. Taille des modèles visités en nombre de variables - Échantillon 2

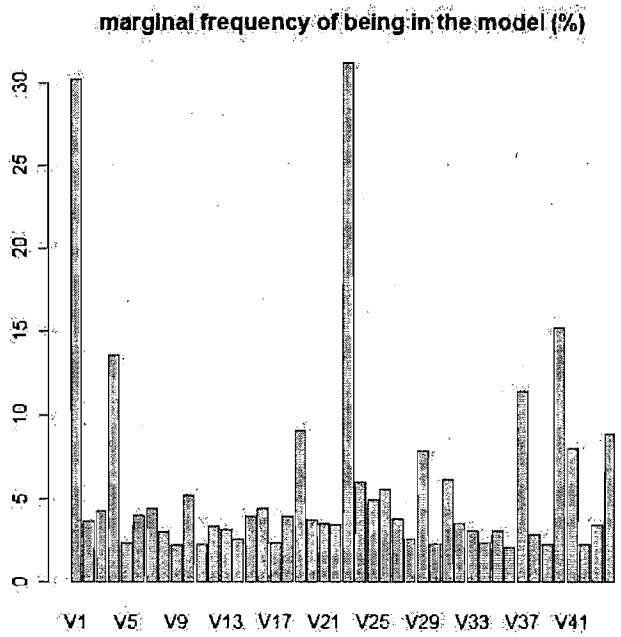

FIG. 5.7. Fréquences marginales d'inclusion dans les modèles pour chaque variables - Échantillon 2

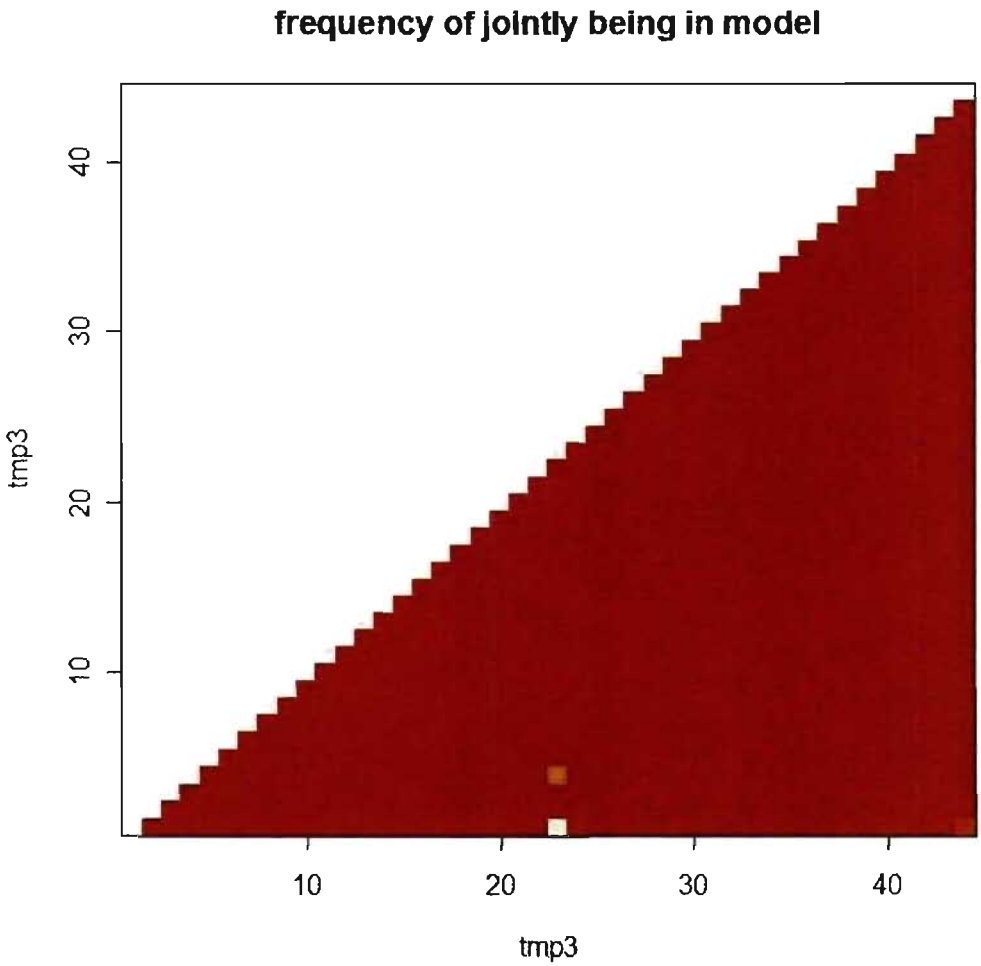

FIG. 5.8. Représentations des fréquences d'inclusions de couples de variables - Échantillon 2

Le deuxième échantillon est de plus grande taille (surtout quant au nombre de pseudocontrôles) et les résultats sont présentés dans les figures 5.6 à 5.8 ainsi que les tableaux 5.10 à 5.12. On interprète de la figure 5.6 qu'un signal semble présent avec environs 1 ou 2 variables impliquées. Les figures 5.7 et 5.8 nous montrent l'influence marginale qui se démarque pour deux variables ainsi que pour un ou bien deux couples de variables.

Le tableau 5.10 nous indique que ce sont les variables V1, V23 ainsi que V40 (plus faiblement) qui sont impliqués le plus fréquemment dans les modèles sélectionnés de maladie. Ces trois proportions dépassent également le seuil fixé de

TAB. 5.10. Les cinq variables qui reviennent le plus souvent - Échantillon 2

Variable	Proportion de présence (en %)
V23	31.16274
V1	30.22094
V40	15.26746
V4	13.52688
V37	11.39296

TAB. 5.11. Les cinq couples de variables qui reviennent le plus souvent - Échantillon 2

Couple de variables	Proportion de présence dans le même arbre (en %)
V1 V23	14.92844
V4 V23	6.27054
V1 V44	2.69466
V23 V29	2.23980
V1 V29	1.94148

TAB. 5.12. Les cinq triplets de variables qui reviennent le plus souvent - Échantillon 2

Triplets de variables	Proportion de présence dans le même arbre (en %)
V1 V23 V29	1.44880
V1 V23 V38	0.67524
V1 V20 V23	0.62352
V1 V23 V24	0.57432
V4 V23 V29	0.51144

15% pour un effet d'une variable simplement considéreé. Le tableau 5.11 est très intéressant puisqu'il démontre qu'un couple de variable se démarque et apparaît une proportion de fois plus élevé que le seuil de détection pour un couple en

interaction. Ainsi, le couple de variables V1 et V23 a été détecté en interaction dans le modèle de la maladie. Aucun effet triple n'a été détecté (tableau 5.12).

Les variables V1, V23 ainsi que V40 correspondant aux marqueurs DAT, NET16 (de nomenclature scientifique rs36021) et TPH2.2. Notons que la variable V4 correspondant au marqueur DAT est présente dans presque 15% des modèles également. Le modèle comprenant seulement l'interaction double prend la forme de la figure 5.9 où les variables indiquent une augmentation particulière-

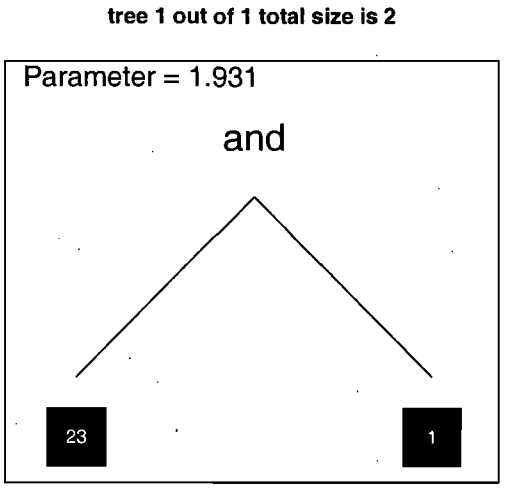

FIG. 5.9. Modèle pour le deuxième échantillon

ment grande de la susceptibilité pour la maladie chez les individus ne possédant pas d'allèle "9" au marqueur DAT et possédant deux allèles "T" au marqueur NET16. Il est intéressant de noter que la deuxième interaction en ordre d'importance correspond à l'interaction suivante : une plus grande susceptibilité pour les individus possédant deux allèles "10" au marqueur DAT et possédant deux allèles "T" au marqueur NET16. Ces deux interactions sont pratiquement identiques puisque lorsque les allèles autres que "9" et "10" sont rares dans la population et l'échantillon (moins de 2 %). La proportion de cas possédant l'interaction est de 20.35 % (47 sur 231) alors que la proportion de pseudocontrôles possédant cette interaction est de 3.57 % (3 sur 84).

Seuls des résultats se répétant d'une expérience à l'autre assurerait que nous avons bel et bien obtenu un réel effet d'interaction entre les deux gènes aux marqueurs DAT et NET16. Par contre, quelques faits supportent nos résultats. Premièrement, il avait été noté que la taille de l'échantillon (surtout le nombre de contrôle) est particulièrement peu élevé de sorte que seuls les génotypes à très grands effets possédent une probabilité intéressante d'être détectés dans la procédure. Or, l'effet du génotype est justement particulièrement élevé avec un coefficient d'interaction de 1.931 (voir la figure 5.9) qui se rapporte sensiblement au risque relatif de la maladie. Ceci est énorme en comparaison aux risques relatifs de gènes détectés seuls. De plus, il est intéressant de noter qu'un génotype répétitivement associé à la maladie dans la littérature (le génotype "10-10" du marqueur DAT) a été sélectionné dans l'interaction. Finalement, il faut noter que l'interaction est pertinente du point de vue théorique puisque les gènes de transport de la dopamine (DAT) et du transport de norépinéphrine sont tous deux impliqués dans le transport de la dopamine, malgré que le rôle de ce dernier gène dans le processus biologique est peu connu (voir Chun-Hyung et al. (2006)).

5.5. Possibilités d'applications immédiates

Nous avons vu l'application de la méthode sur de données récemment collectées. Aujourd'hui, beaucoup d'études ont été effectuées par des échantillons de cas et de leurs parents, notamment dus aux avantages de robustesse par rapport à la stratification de population, concept qui ne peut pas être directement mesuré dans une population humaine. Or, il existe présentement certaines méthodes applicables aux triades qui peuvent évaluer successivement les interactions doubles entre gènes. Des problèmes d'interprétation des résultats, de possibilité d'interactions de plus grand ordre et de multiplicité de tests surviennent. Or, à titre exploratoire, il serait intéressant de revisiter des jeux de données cas-parents antérieurs par régression logique ou réduction multifactorielle de dimensionalité. L'information soutirée ouvrirait de nouvelles pistes à la fois expérimentales et théoriques et pourrait expliquer le manque de réplicabilité de certains résultats.

En ce sens, l'application proposée dans ce mémoire s'avère un atout très intéressant.

De plus, en ce qui concerne toute méta-analyse, on a vu plus tôt que de ne pas considérer d'interaction alors que celle-ci est présente peut impliquer des résultats différents d'une population à l'autre pour l'effet d'un gène. Il est alors possible que des résultats importants soient rejetés pour des raisons de stratification. De là l'importance de la robustesse face à celle-ci notamment lors de méta-analyses. Une recommandation fort utile serait donc de ne pas négliger l'interaction entre gènes lors de méta-analyse. Les méthodes proposées dans ce mémoire sont donc particulièrement importantes pour toute méta-analyse d'échantillons composés de triades cas-parents.

Qui plus est, la taille échantillonnale nécessaire à la détection de l'interaction est relativement plus grande par rapport à la détection d'effets de gènes seuls (la fréquence d'une interaction, dans la population, est souvent plus petite que la fréquence de génotype à un seul locus). Comme le recueil d'échantillons cas-contrôles comporte d'autant plus de risque de stratification de population que la population est grande, tout échantillon de taille suffisante à la détection d'interaction est susceptible d'inclure des individus provenant de sous-populations différentes. L'examen des données cas-contrôles risque donc d'inclure plusieurs faux positifs et de mener à tort un bon nombre de recherches coûteuses. Ainsi, si le recueil d'un très grand nombre de triades est envisageable, il est préférable de procéder par ceux-ci plutôt que par le recueil de cas et de contrôles.

5.6. Travaux futurs

En plus des quelques suggestions de tests et simulations à effectuer pour rendre compte des situations non considérées dans ce mémoire, la présente section résume plusieurs idées d'adaptation de la méthode afin de la rendre plus exacte ou plus puissante, de même que d'autres approches possiblement intéressantes, qu'il serait pertinent d'explorer.

Nous avons soulevé la question suivante : est-il possible d'utiliser une vraisemblance exacte dans le processus de sélection de modèle ? Cette vraisemblance

exacte, on l'a vu dans le chapitre 3, équation (3.13) (Cordell et Clayton (2002)). La vraisemblance (3.13) fait intervenir, au lieu d'un seul pseudocontrôle par cas, 3 pseudocontrôles que sont les 3 autres combinaisons d'haplotypes qui auraient pu être transmis des parents à l'enfant. On note que ces haplotypes doivent être complètement déterminés : la phase ou la séquence d'allèles sur chacune des deux paires de chromosomes doit être connue. Or, pour l'instant, la technologie et les coûts ne permettent pas d'avoir de tels haplotypes "phasés" : on doit se contenter de ne pas connaître la phase ou du moins de l'inférer. Par contre, plus le nombre de loci est grand, plus il devient difficile de retracer l'haplotype des chromosomes des parents : il est fort possible qu'aucun trio cas-parents ne puisse être complètement phasé. Par contre, il est proposé dans l'article Cordell et Clayton (2002) une vraisemblance quelque peu différente, qui prend la forme suivante :

$$L = \frac{exp(X_i'\beta)}{exp(X_*\beta)} \tag{5.1}$$

où X_* représente le génotype non-transmis du parent à l'enfant. En note à part, il est important d'observer que la prévalence dans les sous-populations peut varier selon la sous-population sans influencer la vraisemblance modifiée (5.1) ; le terme s'annule au numérateur et au dénominateur, pour la contribution de chaque famille. Par contre, pour que la vraisemblance modifiée soit valable, l'effet associé aux différents génotypes (nommément les paramètres β) doit être équivalent pour toutes les sous-populations. Cette vraisemblance possède ceci de différent avec la vraisemblance utilisée dans la régression : elle apparie le cas et le pseudocontrôle.

L'impact d'une telle vraisemblance devrait être étudié si les restrictions computationnelles sont respectables. Si on ne considère pas la vraisemblance modifiée, il faudrait étudier par d'autres simulations l'effet de la dépendance entre cas et pseudocontrôles sur le nombre de faux positifs ou bien sur la fréquence de détection des interactions.

Notons qu'au lieu d'investiguer une méthode spécialisée dans l'étude d'interaction entre gènes et de tenter d'adapter celle-ci aux données provenant de cas et de leurs parents, on aurait également pu effectuer l'adaptation inverse c'est-à-dire comme première étape, inspecter les méthodes effectuant des tests d'association génétique pour ce type de données. En fait, par la lecture de nombreux articles,

quelques idées ont surgi par rapport à trois des méthodes d'association génétique sur des données familiales. Celles-ci sont les plus rencontrées le plus régulièrement dans la littérature : ce sont les méthodes implantées dans les logiciels UNPHASED (UNPHASED pour "unphased genotype data", Dudbridge (2003)) et FBAT (FBAT pour "family-based association test", Lunetta et al. (2000)) ainsi que la méthode discutée au deuxième chapitre élaborée par Wilcox et Weinberg (Wilcox et al. (1998)).

Par rapport à cette dernière, on aurait pu étendre la catégorisation du tableau 3.1 afin de considérer plus d'effets dans le modèle : le modèle testé inclut des paramètres d'interaction. En fait, cette idée est explorée dans l'article Wilcox et al. (1998) pour l'effet d'interaction entre un génotype et un facteur environnemental, nommément un effet du génotype maternel pour le développement interutérin de l'enfant. Un inconvénient surgit du fait que d'ajouter de l'interaction entre gènes ajoute au nombre de catégories à considérer de même qu'au nombre de paramètres à évaluer. C'est un inconvénient par le fait que pour toute régression de Poisson lorsque certaines catégories ont un compte de zéro, l'estimation des paramètres n'est pas fiable. Donc, si certains génotypes sont plus rares, il est nécessaire d'introduire un biais afin de pallier à cet inconvénient. Ajoutons à cela des problèmes de multiplicité de tests (il faut considérer toute interaction une à une) et d'interprétation si on considère l'interaction d'un point de vue statistique, cette approche ne semble pas intéressante de prime abord.

Il est fort probablement plus envisageable d'adapter la deuxième méthode, celle implantée par Dudbridge (Dudbridge (2003)). En effet, celle-ci utilise des vraisemblances explicitées dans ce chapitre et qui sont décrites dans Cordell et Clayton (2002). Tout test effectué quant à l'effet de marqueur ou allèle est un test de rapport de vraisemblance et évite du même coup toute difficulté reliée aux génotypes rares. En fait, le logiciel UNPHASED inclut la possibilité d'investiguer toute interaction en regardant la vraisemblance des données à partir des modèles de maladie sous la forme suivante :

$$f(E(Y_{ij}|G_{ij})) = \beta_0 + \beta_1 X(G_{ij}) + \beta_2 Z(G_{ij}) + \beta_3 X(G_{ij}) Z(G_{ij}) \qquad (5.2)$$

où X et Z codent pour deux loci différents. Par exemple, X pourrait prendre les valeurs 1 si le génotype est homozygote pour un allèle à un premier locus et zéro sinon et la même idée peut être appliquée pour la variable Z et un deuxième locus. On remarque immédiatement que le type d'interaction considéré est statistique, la méthode ne peut pas explorer les interactions de plus de deux marqueurs et fait face à des problèmes de multiplicité de tests. Par contre, il est tout à fait envisageable de considérer des interactions sous forme d'expressions logiques dans la modélisation de la maladie afin d'effectuer des tests. Il ne resterait plus qu'à se doter d'une méthode puissante de gestion de la multiplicité des tests (comme celle proposée dans Benjamini et Hochberg (1995)) afin d'obtenir une méthode similaire à la régression logique pour triades cas-parents. Ce qui rendrait cette adaptation intéressante est qu'elle pourrait profiter des fonctionnalités permettant d'obtenir de l'information quant à la phase des haplotypes ainsi que l'utilisation de triades incomplètes par algorithme "Expectation-Maximization", c'est-à-dire lorsqu'il manque le génotype d'un individu (ce qui rend inutilisable ces données puisqu'on ne peut y soutirer de pseudocontrôle). Il serait donc probablement intéressant d'envisager ce type d'adaptation ou d'inclure les avantages de cette méthode aux travaux effectués ici dans de travaux futurs.

La troisième et dernière méthode est celle du logiciel FBAT. La méthode étudie encore cette fois-ci des marqueurs individuels mais peut également étudier des haplotypes. L'idée développée est d'utiliser une statistique qui prend la forme suivante :

$$\sum_i (Y_i - \mu)(X(g_i) - E(X(g_i)|P_i)) \tag{5.3}$$

où, cette fois-ci, on utilise l'indice i pour le ième cas, Y_i représente son phénotype (peut être dichotomique ou quantitatif), $X(g_i)$ représente un code pour le génotype de cet individu (par exemple, si on étudie un seul allèle à la fois, $X(g_i)$ peut représenter le nombre de cet allèle dans le génotype du cas) et finalement P_i représente le génotype des parents. Les auteurs notent également que μ représente une valeur moyenne de Y_i qui est souvent inconnue mais sa valeur ne biaise pas le test.

Le test de la statistique se fait depuis des hypothèses nulles qui imposent une transmission équivalente de chaque allèle des parents à l'enfant. Plus d'information sur le test et des approximations sont disponibles dans l'article de Lunetta et al. (2000). La méthode est très développée relativement aux autres, avec notamment des méthodes pour utiliser les familles incomplètes (il manque des données des parents), des familles étendues (inclusion des cousins, par exemple), des covariables (l'effet d'une exposition environnementale), etc. Le principal avantage de la méthode est qu'elle réussit à évaluer de façon tout à fait ingénieuse la puissance de chaque test avant même d'effectuer celui-ci (voir Steen et al. (2005)). Ainsi, une sélection de tests est possible permettant d'éviter la question de multiplicité de tests. Une avenue de recherche possible pourrait être d'utiliser les mesures de puissance telles que décrites dans cet article pour la sélection d'interaction. On a vu du fonctionnement de la régression logique Monte Carlo que le processus de sélection de modèle passe par les fonctions de vraisemblance a priori et a posteriori qui permettent d'évaluer les probabilités d'acceptation d'opérations sur les arbres logiques (voir le chapitre 3). Or, une modification possible à la méthode serait d'inclure des mesures de puissance telles que décrites dans Steen et al. (2005) en plus des fonctions de vraisemblance afin de déterminer les probabilités d'acceptation de la chaîne de Markov. Sinon, serait-il possible d'utiliser la statistique FBAT dans l'étude de l'interaction gène-gène et ainsi profiter des autres avantages de la méthode ? Pour l'instant, l'équipe du Harvard School of Public Health menée par les Nan M. Laird et Christoph Lange travaille sur des moyens de tester ce type d'interaction et croit obtenir des résultats intéressants sous peu. Il sera très intéressant de voir tout développement de leur part.

Finalement, notons que concomitant avec le présent projet, d'autres travaux portant sur des méthodes d'exploration de l'interaction gène-gène ont été publiés. Il est clair que beaucoup de nouvelles méthodes surgissent et que de constantes améliorations seront apportées dans le futur. Un de ces travaux est décrit dans l'article Zhang et Liu (2007) et s'avère une piste intéressante à regarder. Notamment, les auteurs comparent la méthode qu'ils ont développé avec plusieurs autres méthodes explorant l'interaction dont entre autres la régression logique et

la réduction multifactorielle de dimensionalité. Les résultats semblent prometteurs mais la méthode ne s'applique encore qu'aux échantillons cas-contrôles seulement. Pourra-t-on utiliser les échantillons composés de triades avec cette nouvelle approche ?

CONCLUSION

Depuis peu dans la littérature génétique, on perçoit l'urgence du développement de méthodes permettant de détecter l'interaction entre gènes et leur influence sur les susceptibilités des individus. On a vu que ce premier problème peut être associé à un manque de réplicabilité des études d'association, à des puissances de détection amoindries de génotypes importants ainsi qu'à l'incompréhension de phénomènes biologiques sous-jacents aux maladies complexes. En réponse à cette demande, plusieurs outils informatiques se sont développés et permettent la détection d'interaction entre gènes par sélection de modèles. La régression logique et la réduction multifactorielle de dimensionalité font partie de ces méthodes et ont été décrites dans ce mémoire. Par contre, on soupçonne ces méthodes d'être assujetties au problème de stratification allélique de la population. À ce sujet, nous avons démontré que l'influence d'une telle hétérogénéité est réduite par l'usage de trios cas-parents. D'où finalement la motivation de ce mémoire : nous prémunir d'une méthode permettant d'étudier l'interaction entre gènes par l'usage de trios cas-parents qui soit facile d'interprétation, permet l'étude d'interaction de plus de deux gènes et qui permet d'éviter la question de multiplicité de tests.

La solution proposée est de généraliser au génotype entier le rassemblement des cas à des pseudo-contrôles caractérisés des haplotypes non transmis des parents au cas malade. On a par ailleurs développé un théorème permettant de comprendre le rapprochement des échantillons cas-pseudo-contrôles aux échantillons cas-contrôles. Ce rapprochement suggère le bon fonctionnement des méthodes de la régression logique et de la réduction multifactorielle de dimensionalité appliquées à des échantillons composés de triades. Afin d'assurer de tels propos, trois modèles ont été simulés et a suivi l'étude de la détection des interactions. Les

résultats suggèrent que l'utilisation de triades dans l'exploration de l'interaction est faisable et même souhaitable si on soupçonne de la stratification dans la population étudiée. En discussion, on a également vu que la procédure proposée dans ce mémoire peut s'appliquer naturellement à d'autres situations dont notamment lorsqu'il y a déséquilibre de liaison entre marqueurs et lorsque les haplotypes sont importants. Par contre, les méthodes proposées sont encore embryonnaires et il y a place à amélioration.

Finalement, on a appliqué la méthode à des données génétiques de parents et de leur enfant atteint du trouble de déficit de l'attention et d'hyperactivité. L'échantillon de plus grande taille a permis de soulever une interaction possiblement intéressante, celle entre le gène du transporteur de dopamine (DAT) ainsi que le gène du transporteur de norépinéphrine (NET). Plus précisément, les résultats suggèrent que les individus possédant un génotype "10-10" au marqueur DATALL ainsi que "TT" au génotype NET16 (rs36021) possèdent une grande susceptibilité face au TDAH. Aucun test n'a été effectué puisque les modèles ne sont pas emboîtés et que les tests par permutation ne sont pas trivialement applicables.

Dans un proche avenir, on espère que l'interaction entre gènes sera étudiée plus systématiquement. Également, il serait souhaitable que la méthode proposée dans ce mémoire soit utilisée sur de nombreux échantillons cas-parents. Peut-être cela permettra-t-il de soulever le voile sur le fonctionnement des gènes, leur influence sur les maladies et peut-être même d'effectuer un pas de plus pour le développement de médication efficace.

Bibliographie

Agresti A. (1990). *Categorical Data Analysis*, 734 pages.

Benjamini, Yoav et Hochberg, Yosef (1995). Controlling the False Discovery Rate : A Pratical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society (Methodology Series)*, vol. **57**, pages 289-300.

Biederman, J. et Faraone, S. (2005). Attention-deficit hyperactivity disorder, *The Lancet*, vol. **366**, pages 237-248.

Carlborg, Örjan et Haley, Chris S. (2004). Epistasis : too often neglected in complex trait studies ?, *Nature Reviews - Genetics*, vol. **5**, pages 618-625.

Kim, Chun-Hyung, Hahn, Maureen K., Joung, Yoosook, Anderson, Susan L., Steele, Angela H., Mazei-Robinson, Michelle S., Gizer, Ian, Teicher, Martin H., Cohen, Bruce M., Robertson, David, Waldman, Irwin D., Blakely, Randy D. et Kim, Kwang-Soo (2006). A polymorphism in the norepinephrine transporter gene alters promoter activity and is associated with attention-deficit hyperactivity disorder, *Proceedings of the National Academy of Sciences of the United States of America*, vol. **103**, pages 19164-19169.

Cordell, Heather J. (2002). Epistasis : what it means, what it doesn't mean, and statistical methods to detect it in humans, *Human Molecular Genetics*, vol. **11**, pages 2463-2468.

Cordell, Heather J. et Clayton, David G. (2002). A Unified Stepwise Regression Procedure for Evaluating the Relative Effects of Polymorphisms within a Gene Using Case/Control or Family Data : Application to HLA in Type 1 Diabetes, *American Journal of Human Genetics*, vol. **70**, pages 124-141.

Dudbridge, Frank (2003). Pedigree Disequilibrium Tests for Multilocus Haplotypes, *Genetic Epidemiology*, vol. **25**, pages 115-121.

Green, Peter J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination, *Biometrika*, vol. **82**, pages 711-732.

Hosmer, David W. et Lemeshow, Stanley (2000). *Applied Logistic Regression*, 392 pages.

Jakulin, Aleks et Bratko, Ivan (2003). Analyzing Attribute Dependencies, *Lecture Notes in Computer Science*, vol. **2838**, pages 229-240.

Kooperberg, Charles et Ruczinski, Ingo (2005). Identifying Interacting SNPs Using Monte Carlo Logic Regression, *Genetic Epidemiology*, vol. **28**, pages 151-179.

Lange, Kenneth (2002). *Mathematical and statistical methods for genetic analysis*, 265 pages.

Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S., et Hirschhorn J. N. (2003). Meta-analysis of a genetic association studies supports a contribution of common variants to susceptibility of common disease, *Nature Genetics*, vol. **33**, pages 117-182.

Lunetta, Kathryn L., Faraone, Stephen V., Biederman, Joseph et Laird, Nan M. (2000). Family-Based Tests of Association and Linkage That Use Unaffected Sibs, Covariates, and Interactions, *American Journal of Human Genetics*, vol. **66**, pages 605-614.

Moore, Jason H., Gilbert, Joshua C., Tsai, Chia-Ti, Chiang, Fu-Tien, Holden, Todd, Barney, Nate et White, Bill C. (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility, *Journal of Theoretical Biology*, vol. **241**, pages 252-261.

Ott, Jurg (1999). *Analysis of Human Genetic Linkage*, 416 pages.

Ripley, B.D. (1996). *Pattern recognition and neural networks*, 415 pages.

Ritchie, Marylyn D., Hahn, Lance W., Roodi, Nady, Bailey, L. Renee, Dupont, William D., Parl, Fritz F. et Moore, Jason H. (2001). Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer, *American Journal of Human Genetics*, vol. **69**, pages 138-147.

Ruczinski, Ingo (2003). *Logic Regression*, Thèse de Doctorat, Bloomberg School of Public Health.

Ruczinski, Ingo, Kooperberg, Charles et Leblanc, Michael (2003). Logic Regression, *Journal of Computational and Graphical Statistics*, vol. **12**, pages 475-511.

Self, Steven G., Longton, Gary, Kopecky, Kenneth J. et Liang, Kung-Yee (1991). On Estimating HLA/Disease Association with Application to a Study of Aplastic Anemia, *Biometrics*, vol. **47**, pages 53-61.

Spielman, Richard S., McGinnis, Ralph E. et Ewens, Warren J. (1993). Transmission Test for Linkage Disequilibrium : The Insulin Gene Region and Insulin-dependent Diabetes Mellitus (IDDM), *American Journal of Human Genetics*, vol. **52**, pages 506-516.

Steen, Kristel Van, McQueen, Matthew B., Herbert, Alan, Raby, Benjamin, Lyon, Helen, DeMeo, Dawn L., Murphy, Amy, Su, Jessica, Datta, Soma, Rosenow, Carsten, Christman, Michael, Silverman, Edwin K., Laird, Nan M., Weiss, Scott T. et Lange, Christoph (2005), Genomic screening and replication using the same data set in family-based association testing. *Nature Genetics*, vol. **37**, pages 683-691.

Terwilliger, Joseph Douglas, et Ott, Jurg (1994). *Handbook of Human Genetic Linkage*, 320 pages.

Wilcox, Allen J., Weinberg, Clarice R. et Lie, Rolv Terje (1998). Distinguishing the Effects of Maternal and Offspring Genes through Studies of "Case-Parent Triads", *American Journal of Epidemiology*, vol. **148**, pages 893-901.

Wilson, S.R. (2001). Epistasis and its possible effects on transmission disequilibrium tests, *Annals of Human Genetics*, vol. **62**, pages 565-575.

Yang, B., Chan, R.C., Chan, R.C., Jing, J., Li, T., Sham, P., Chen, R.Y. (2000). A meta-analysis of association studies between the 10-repeat allele of a VNTR polymorphism in the 3'-UTR of dopamine transporter gene and attention deficit hyperactivity disorder, *American Journal of Medical Genetics B Neuropsychiatric Genetics*, vol. **163**, pages 1-33.

Zhang, Yu et Liu, Jun S. (2007). Bayesian inference of epistatic interactions in case-control studies, *Nature Genetics*, vol. **39**, pages 1167-1173.

Zondervan, Krina T. et Cardon, Lon R. (2004). The complex interplay among factors that influence allelic association. *Nature Reviews - Genetics*, vol. **5**, pages 89-100.