UNIVERSITÉ DU QUÉBEC À MONTRÉAL

TESTS DE DÉSÉQUILIBRE DE LIAISON ET LEUR APPLICATION À UN GÈNE CANDIDAT À L'HYPERACTIVITÉ

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

DJAMILA ABED

FÉVRIER 2008

À la mémoire de mon père

À ma très chère mère

À toute ma famille

À tous mes amis

REMERCIEMENTS

J'aimerais adresser mes remerciements à des personnes qui ont beaucoup compté dans l'accomplissement de ce travail.

Tout d'abord ma directrice la professeure Sorana Froda pour son soutien moral, la rigueur de son encadrement, ses précieux conseils et les nombreuses discussions que nous avons eues tout au long de ce travail. Le professeur Fabrice Larribe pour m'avoir initié aux thématiques de la génétique et pour avoir accepté de codiriger ce travail.

Madame Gisèle Legault, analyste informatique au département de mathématiques, pour son aide technique pour la rédaction de mon mémoire. Elle a fait ça avec sa gentillesse habituelle et sa grande disponibilité. Monsieur Bertrand Fournier, analyste statisticien au département de mathématiques, pour sa précieuse aide dans la programmation SAS.

Mesdames Isabel Fortier et Louise R. Simard et Monsieur François L'Heureux de l'université McGill pour nous avoir fourni les données sur lesquelles s'est basé notre travail.

Monsieur Hassan Younes, chargé de cours au département de mathématiques, pour son soutien moral.

Les personnes qui me sont chères et qui tiennent une grande place dans mon cœur : mes parents tout d'abord, sans qui bien sûr je ne serais pas arrivée au bout de mon parcours et qui ont toujours été présents, mes frères et sœurs pour leurs encouragements dans les moments difficiles.

TABLE DES MATIÈRES

LISTE DES FIGURES vii
LISTE DES TABLEAUX viii
RÉSUMÉ xi
INTRODUCTION 1
CHAPITRE I
CONCEPTS DE GENETIQUE 4
1.1 Introduction 4
1.2 Terminologie de base 5
1.3 Arbre généalogique 9
1.4 Mode de transmission 10
1.5 Fréquences alléliques et génotypiques 11
1.6 Principe d'équilibre de Hardy-Weinberg 14
1.7 Recombinaison et liaison génétiques 16
1.7.1 Recombinaison génétique 16
1.7.2 Liaison génétique (*genetic linkage*) 17
1.8 Identification des gènes candidats 18
1.8.1 Études de liaison génétique 18
1.8.2 Études d'association 19
1.8.3 Déséquilibre de liaison (linkage disequilibrium) 23
CHAPITRE II
LE TEST DU TDT (TRANSMISSION TEST FOR LINKAGE DISEQUILIBRIUM) 26
2.1 Exemple illustratif 27
2.2 Présentation du TDT 28
2.3 Construction du test du TDT 30
2.4 Statistique du test 40

CHAPITRE III
GÉNÉTIQUE QUANTITATIVE 44
3.1 Définitions 45
3.1.1 Définition d'un QTL 45
3.1.2 Maladie complexe 45
3.3 Valeurs et moyennes d'une population 46
3.4 Influence d'un locus sur le phénotype 47
3.4.1 Moyenne d'une population 49
3.4.2 Effet moyen 50
3.5 Relation entre la valeur phénotypique et la valeur génotypique 51
3.5.1 Composantes de la valeur génotypique 51
3.5.2 Composantes de la valeur environnementale E 53
3.6 Étude des variances 54
3.6.1 Variance de la valeur phénotypique 54
3.6.2 Variance de la valeur génotypique 55
3.7 Modèle des moindres carrés pour les effets génétiques : calcul de variance 55
3.8 Ressemblance entre apparentés 58
3.8.1 Mesure de parenté 59
3.8.2 Covariance phénotypique entre les apparentés 63
CHAPITRE IV
LE TEST DU QTDT (QUANTITATIVE TRANSMISSION DISEQUILIBRIUM TEST) ... 68
4.1 Présentation du modèle 69
4.1.1 Paramètres du modèle 70
4.1.2 Calcul de l'espérance et nouvelle paramétrisation 71
4.1.3 Calcul des éléments de la matrice de variance covariance 71
4.2 Estimation des paramètres du modèle 74
4.3 Test général d'association : principes de base 75
4.4 Modèle de décomposition orthogonale des scores des génotypes 77
4.5 Modèles d'Abecasis 81
4.5.1 Cas général avec plusieurs enfants malades par famille 82
4.5.2 Cas particulier d'un enfant malade par famille 87
4.6 Tests d'association et de liaison génétique 92

4.6.1 Test de liaison génétique 92
4.6.2 Test conjoint d'association et de *linkage* 93

CHAPITRE V
ÉTUDE DE CAS : ASSOCIATION GÉNÉTIQUE CHEZ LES ENFANTS HYPERACTIFS 95

5.1 Définition et diagnostic du TDA/H 95
5.1.1 Définition du TDA/H 95
5.1.2 Symptômes du TDA/H 96
5.1.3 Outils de diagnostic du TDA/H 97
5.2 Résultats de l'enquête 99
5.2.1 Bref aperçu de l'enquête 99
5.2.2 Questionnaires 100
5.3 Analyse des données 101
5.3.1 Validation des scores 101
5.3.2 Analyse des facteurs environnementaux 106
5.3.3 Tests génétiques d'association 115

CONCLUSION 124

APPENDICE A
DISC4 126

APPENDICE B
CONNERS ENSEIGNANT 128

APPENDICE C
COTATION ENSEIGNANT 131

APPENDICE D
CONNERS PARENTS 135

APPENDICE E
COTATION PARENTS 139

APPENDICE F
ÉCHELLE DU WISC-III 143

BIBLIOGRAPHIE 145

LISTE DES FIGURES

Figure		page
Figure 1.1	Représentation d'une cellule.	6
Figure 1.2	Exemple d'un arbre généalogique.	9
Figure 1.3	Famille artificielle ayant 2 génotypes à deux loci.	10
Figure 2.1	Exemple de trente familles classées selon le génotype et selon le type d'allèles transmis	27
Figure 3.1	Valeurs arbitraires attribuées aux génotypes	48
Figure 3.2	Transmission des gènes de deux parents à leur deux enfants	59

LISTE DES TABLEAUX

Tableau		page
Tableau 1.1	Relation entre génotypes et phénotypes au gène ABO	8
Tableau 1.2	Exemple d'une population d'individus classés selon leurs génotypes.	12
Tableau 1.3	Fréquences des allèles et des génotypes.	13
Tableau 1.4	Démonstration de la loi de Hardy-Weinberg.	15
Tableau 1.5	Exemple d'illustration d'une association entre allèles en situation d'équilibre de la phase gamétique	22
Tableau 2.1	Exemple d'allèles transmis et non transmis.	28
Tableau 2.2	Combinaisons des allèles marqueurs M_1 et M_2 transmis et non transmis parmi les $2n$ parents ayant n enfants affectés.	31
Tableau 2.3	Probabilités des combinaisons des allèles marqueurs M_1 et M_2 transmis et non transmis parmi les $2n$ parents ayant n enfants affectés.	33
Tableau 3.1	Exemple de valeurs génotypiques et leur valeur moyenne dans la population	49
Tableau 3.2	Transmission des allèles B_1 et B_2 d'une mère à ses enfants.	60
Tableau 4.1	Tableau des fréquences génotypiques	69
Tableau 4.2	Moyennes et différences de l'effet additif a pour un locus biallélique à effet additif	79

Tableau 5.1	Quelques statistiques descriptives de l'échantillon des 308 enfants	100
Tableau 5.2	Comparaison des Scores pour le TDA/H	102
Tableau 5.3	Comparaison des scores pour l'hyperactivité/impulsivité	102
Tableau 5.4	Comparaison des scores pour l'inattention	103
Tableau 5.5	Résultats du test de McNemar pour les trois phénotypes	103
Tableau 5.6	Résultats du diagnostic du *DISC4* selon que l'enfant présente ou ne présente pas un des troubles	104
Tableau 5.7	Comparaison des résultats du diagnostic du *DISC4* et du score de l'échelle Conners-enseignant.	105
Tableau 5.8	Comparaison des résultats du diagnostic du *DISC4* et du score de l'échelle Conners-parent.	105
Tableau 5.9	Statistiques sur les scores obtenus par les Conners-enseignants	108
Tableau 5.10	Statistiques sur les scores obtenus par les Conners-parents	108
Tableau 5.11	Scores moyens selon le sexe Conners-enseignants	109
Tableau 5.12	Scores moyens selon le sexe et les scores du Conners-parents	110
Tableau 5.13	Résultats du diagnostic du *DISC4* selon le type du trouble	111
Tableau 5.14	Répartition des enfants selon la classe du QI et le diagnostic *DISC4*	112
Tableau 5.16	Nombre d'enfants sains et ceux ayant le TDA/H selon le risque de complication de grossesse	114
Tableau 5.17	Valeurs du rapport de cotes *RC* pour le TDA/H et des facteurs de risque	115
Tableau 5.18	Répartition des enfants selon le phénotype	117

Tableau 5.19 Test du TDT pour les enfants atteints d'un des trois phénotypes 117

Tableau 5.20 Test du TDT pour les enfants souffrant du TDA/H 118

Tableau 5.21 Test du TDT pour les enfants atteints de l'Inattention 118

Tableau 5.22 Test du TDT pour les enfants atteints d'Hyperactivité/Impulsivité 119

Tableau 5.23 P-valeurs du QTDT entre les phénotypes et le marqueur MM_1 Mesure d'association 121

Tableau 5.24 P-valeurs du QTDT entre les phénotypes et le marqueur MM_2 Mesure d'association 122

Tableau 5.25 P-valeurs du QTDT entre les phénotypes et le marqueur MM_1. Analyse de liaison 123

Tableau 5.26 P-valeurs du QTDT entre les phénotypes et le marqueur MM_2 Analyse de liaison 123

RÉSUMÉ

Ce mémoire présente deux tests génétiques permettant l'identification de gènes reliés à une maladie par l'intermédiaire des études d'association et de liaison génétique. Le premier test, le TDT (*Transmission Test for Linkage disequilibrium*), est très couramment utilisé et s'applique aux caractères binaires. Le second test, le QTDT est relativement nouveau et s'applique aux caractères mesurés sur une échelle continue. On commence par donner des notions théoriques de ces deux tests, en particulier des notions élémentaires de génétique quantitative. Par la suite, ces deux tests sont appliqués à des données réelles, des enfants atteints du TDA/H (trouble du déficit de l'attention, avec ou sans hyperactivité), et leurs résultats sont comparés. Les données comprennent des variables explicatives environnementales et génétiques. Les phénotypes (variables expliquées) sont des scores de comportement de l'enfant, qui varient sur une échelle continue. On montre que les résultats des deux tests sont similaires; le gène étudié n'est ni associé, ni lié au trouble étudié.

Mots-clés: déséquilibre de liaison, test du TDT, test du QTDT, TDA/H.

INTRODUCTION

La génétique humaine se préoccupe, entre autres, de la recherche de gènes responsables de maladies ou affections diverses. À partir de la découverte de ces gènes, on espère pouvoir proposer des traitements efficaces.

Souvent, on étudie des gènes candidats pour des maladies multifactorielles (par exemple, les cancers, les maladies cardiovasculaires, l'asthme, etc.) qui résultent des effets et interactions de nombreux facteurs, génétiques et environnementaux, et on fait appel à des techniques analytiques bien spécifiques. Parmi elles, il y a les « études d'association » basées sur des méthodes d'analyse familiale. Ces méthodes peuvent s'appliquer aussi bien à des caractères qualitatifs (ou binaires) que quantitatifs. Dans le premier cas, on regroupe les individus en classes distinctes (par exemple malade ou non malade). Par contre, dans le second cas, il est difficile de faire ce type de regroupement, c'est-à-dire classer les individus dans différentes catégories, car le caractère étudié varie sur une échelle continue. La génétique s'intéresse à l'étude de l'hérédité entre les individus. Qu'il s'agisse de différences qualitatives ou quantitatives, on suppose que plusieurs gènes (ainsi que des facteurs environnementaux) interagissent pour former ces différences. La plupart du temps, les effets de chaque facteur ne peuvent cependant pas être différenciés. Utiliser la catégorisation pour un caractère qui est de type continu, peut mener à des résultats non satisfaisants, puisqu'on n'utilise pas toute l'information disponible et il est plus difficile de tenir compte de la présence des facteurs environnementaux. Un des objectifs de notre travail est de rendre compte des résultats dans le cas de l'utilisation de ces deux types de tests.

Une méthode couramment utilisée pour l'analyse des gènes candidats (c'est-à-dire potentiellement reliés à la maladie) est le «TDT» (*Transmission Test for Linkage disequilibrium*). Celle-ci s'applique aux caractères binaires (malades, non malades), et nécessite des données sur des trios, les deux parents et leur enfant malade. Elle compare les

fréquences de transmission de l'allèle soupçonné d'être responsable de la maladie avec les fréquences de transmission de l'autre allèle. La seconde, appelée «QTDT» (*Quantitative Transmission Disequilibrium Test*), plus récente et plus puissante, mais moins connue, s'applique aux caractères quantitatifs. Le QTDT est un test qui utilise le caractère continu des phénotypes pour expliquer les différences d'ordre génétique entre les individus par rapport à ce phénotype. Il permet ainsi d'utiliser toute l'information que la dichotomisation grossière en malades-non malades ne permet pas d'utiliser. Le QTDT est basé sur le modèle de décomposition de la variance de la valeur phénotypique, ainsi que sur le test du maximum de vraisemblance. Le QTDT permet aussi de tenir compte de l'influence des facteurs environnementaux sur le phénotype.

Dans notre travail, on se propose de comparer les résultats de ces deux tests génétiques en nous basant sur des données réelles d'enfants souffrant du trouble du déficit de l'attention, avec ou sans hyperactivité, connu sous l'appellation de TDA/H. Il s'agit d'un trouble multifactoriel à forte composante génétique. Dans un premier temps, nous nous intéressons à l'étude des facteurs environnementaux impliqués dans cette affection. Suite à cela, nous nous penchons sur la composante génétique par l'étude de gènes candidats. Pour cela, dans un premier temps, nous utilisons la dichotomisation atteint/non atteint, et nous appliquons le TDT. Dans une seconde étape, nous nous intéressons à l'utilisation des scores sur une échelle continue du trouble étudié, et nous introduisons le QTDT. Nous procédons par la suite à la comparaison des résultats ainsi obtenus.

Notre travail est réparti en cinq chapitres. Dans le premier chapitre, nous introduisons les éléments de base en génétique nécessaires à la compréhension de notre travail. Ce chapitre contient les termes utilisés en génétique et quelques propriétés de la génétique des populations. Nous abordons par la suite le mode de transmission mendélien, ainsi que les méthodes utilisées pour l'identification des gènes. Ces deux dernières notions nous servent de base pour présenter le chapitre deux qui est consacré au développement du TDT. Dans ce deuxième chapitre, nous présentons le fonctionnement, les hypothèses, ainsi que la statistique du test du TDT.

Le troisième chapitre contient les notions de base de la génétique quantitative qui est une branche assez complexe de la génétique. Dans ce chapitre, on définit la notion de valeur phénotypique, qui est une notion très importante, et on donne sa décomposition en plusieurs composantes. L'idée de base de la génétique d'un caractère sur une échelle continue est d'étudier sa variation pour expliquer les différences, et l'étude de cette variation nous amène alors à la diviser en composantes attribuables chacune à une cause différente.

Le quatrième chapitre est le prolongement du chapitre trois, et il est consacré à la présentation du QTDT, en utilisant les différents concepts, définitions et formules qui sont définis au chapitre précédent.

Enfin, le chapitre cinq qui clôture notre travail est consacré à l'application de ces deux tests pour l'identification des gènes impliqués dans le TDA/H. Cette étude de la composante génétique du trouble étudié est d'abord précédée par l'étude des facteurs environnementaux qui peuvent avoir une incidence sur la maladie.

CHAPITRE I

CONCEPTS DE GENETIQUE

1.1 Introduction

Les premières lois de l'hérédité ont été découvertes en 1865 par le moine Gregor Mendel (1822-1884). Mais ce n'est que bien plus tard, soit au début du $XX^{ème}$ siècle, que l'œuvre de Mendel sera reconnue. Mendel a fait des expériences scientifiques afin de répondre à deux questions :

- comprendre pourquoi des organismes (plantes et animaux) apparentés pouvaient se ressembler;
- expliquer les ressemblances familiales.

Il a émis l'hypothèse suivante : « Les caractères héréditaires sont portés par des particules transmissibles présentes chez les parents ». Il pensait que l'observation de l'hérédité de certains caractères pourrait répondre à ces questions. Il a ainsi mené une série d'expériences portant sur le croisement entre des pois ayant différents caractères (grains lisses, ridés, verts, jaunes etc.). Ces expériences lui ont permis de découvrir l'existence d'éléments biologiques transmissibles de génération en génération qui seront dénommés, bien plus tard, gènes. Il a observé que la progéniture de certaines plantes avait des caractéristiques semblables aux caractéristiques physiques de leurs parents, ou de leurs ancêtres.

L'observation de cette transmission d'une génération à l'autre de certains caractères apparents chez les pois lui a permis de déduire les trois lois simples, les lois de Mendel, dont les deux premières sont toujours valables aujourd'hui.

Ce sont les gènes qui déterminent les caractères héritables, et l'étude de ces gènes constitue *La Génétique*. Avec ces observations, Mendel a élaboré le modèle de l'hérédité qui est devenu la base de la génétique moderne.

1.2 Terminologie de base

Tous les organismes vivants ont comme unité de base : la *cellule.* Certains organismes sont constitués d'une seule cellule, comme c'est le cas d'une bactérie. L'être humain lui, possède entre 60000 et 100000 milliards de cellules. Il existe deux types de cellules; les procaryotes, qui ne possèdent pas de noyau, et les eucaryotes, comprenant un noyau. Tous les organismes pluricellulaires sont constitués de cellules eucaryotes. La cellule est enveloppée d'une membrane composée de lipides, de protéines et de glucides. Cette membrane facilite les échanges d'aliments, d'énergie et de fluides. Le cytoplasme est une sorte de gelée qui entoure le noyau.

Le noyau de la cellule eucaryote renferme une série de paires de longs filaments qu'on appelle les ***chromosomes***. Ces derniers sont définis comme l'entité physique transmise lors du processus de reproduction par les parents à leurs enfants. L'être humain possède 23 paires de chromosomes : 22 paires sont communes aux deux sexes, et les deux chromosomes restants sont les chromosomes sexuels. Chez la femme, ils forment une paire identique, et on les appelle les chromosomes X . Chez l'homme, les chromosomes de la paires sont différents, l'un est un chromosome X , et l'autre beaucoup plus court est appelé chromosome Y . Chaque chromosome est constitué d'une seule molécule d'ADN (acide désoxyribonucléique) qui est le support moléculaire de l'information génétique (voir figure 1.1).

Les ***gènes*** sont logés le long des chromosomes, ce sont eux qui déterminent les caractères héréditaires. Ils sont des segments d'ADN. Chaque cellule humaine contient une série identique d'environ 50000 gènes et chaque individu possède deux copies de chaque gène, une transmise par le père, et l'autre par la mère.

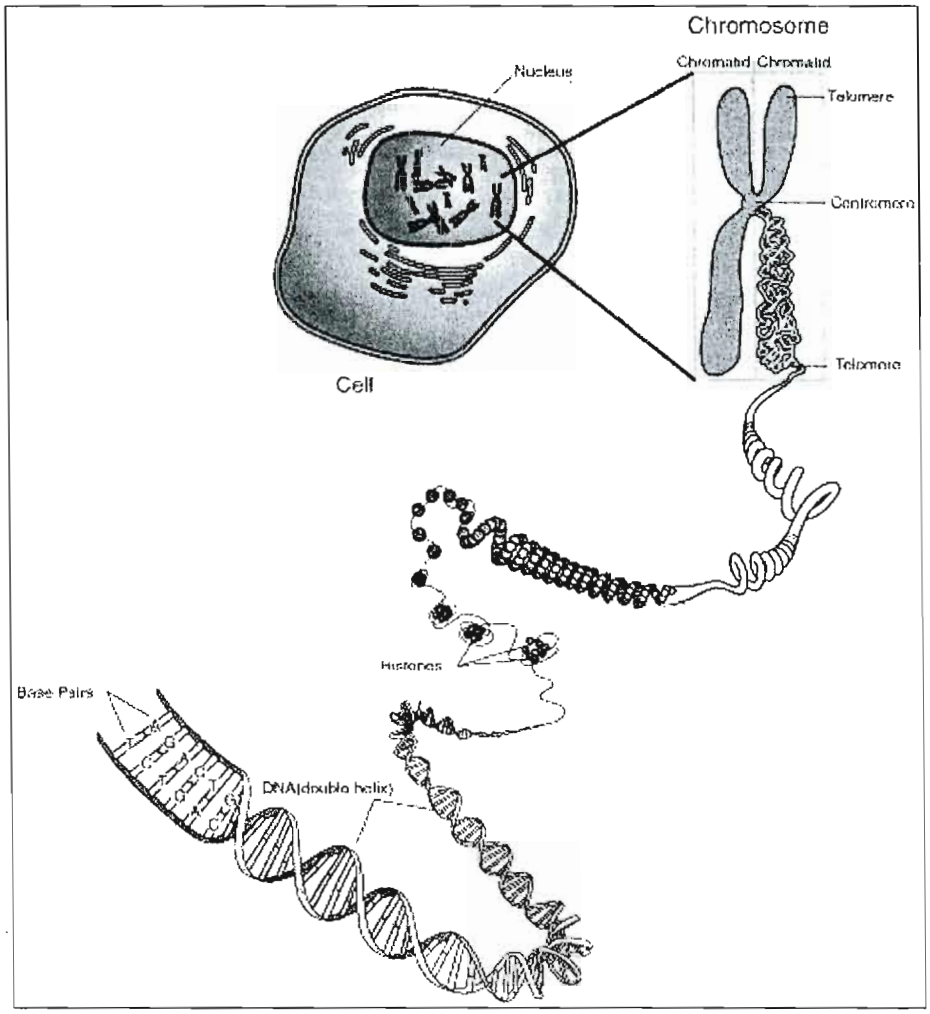

Figure 1.1 Représentation d'une cellule[1].

L'***allèle*** est une des différentes formes que peut prendre un même gène, par exemple le gène de la couleur des yeux comprend un allèle pour chaque couleur et un ***locus***[2] définit la

---

[1] Wikipedia (2007).

[2] Pour le pluriel de locus, on utilise couramment le mot *loci*.

position, ou l'emplacement précis d'un allèle ou d'un gène, le long d'un chromosome. Les allèles occupent la même position (locus) sur les chromosomes homologues.

Chez un individu, chaque gène est représenté par deux allèles situés au même locus ; un est transmis par la mère, l'autre par le père. Ces deux allèles transmis peuvent être identiques ou différents. Prenons l'exemple d'un gène au locus B ayant 2 allèles B et b, alors un individu peut être BB, bb, ou Bb. L'individu est dit ***homozygote*** pour un gène, quand il possède deux allèles identiques de ce gène (dans notre exemple BB, bb). Il est dit ***hétérozygote*** pour un gène, quand il possède deux allèles différents de ce gène (Bb dans notre exemple).

Soit un individu hétérozygote pour un gène comportant les deux allèles B et b ; si cet individu exprime le caractère correspondant à l'allèle B, cet allèle sera dit ***dominant*** par rapport, à l'allèle b qui est qualifié de ***récessif.*** Il existe aussi des allèles codominants qui s'expriment tous les deux (très fréquents pour les fleurs). La lettre majuscule est réservée pour le caractère dominant, et la lettre minuscule pour le caractère récessif.

Le ***génotype*** d'un individu désigne un ensemble de gènes. Il est constitué de plusieurs paires d'allèles. D'une façon plus simple, le génotype est défini comme un ensemble d'informations génétique contenu dans le génome, et l'expression du génotype conduit au phénotype.

Une séquence d'allèles reçue du même parent est appelée ***haplotype***. L'haplotype est donc un fragment chromosomique qui est transmis à la descendance. Pour un certain chromosome, chaque individu se caractérise ainsi par deux haplotypes, provenant l'un du père, et l'autre de la mère. Ces deux haplotypes ainsi reçus définissent le génotype de l'individu aux deux loci.

Le mot ***phénotype*** vient du grec *fainein* qui signifie paraître. C'est un mot qui désigne l'ensemble des caractères, comme la couleur des yeux, la couleur des cheveux, la taille, le poids de l'individu etc. On le définit aussi comme l'ensemble des caractères d'un individu résultant des actions conjuguées, de son *génotype* et de *l'environnement* dans lequel il évolue.

Pour illustrer ces deux concepts, ainsi que la notion de dominance et de récessivité, on se réfère à un exemple donné par Ott (1991). Celui-ci porte sur le gène *ABO* déterminant le groupe sanguin. Il considère les trois allèles majeurs du gène qui interagissent pour déterminer les divers types de groupe sanguins. Pour ce même gène du groupe sanguin, six génotypes peuvent être formés : *A/A, A/B, A/O, B/B, B/O, et O/O.* L'expression de chacun des six génotypes est ce qu'on a défini comme étant le phénotype. Pour ce gène, on retrouve quatre phénotypes, qui sont les différents groupes sanguins que peut avoir un individu : groupe A, groupe B, groupe AB, et enfin le groupe O.

La relation entre le génotype et le phénotype pour le gène *ABO* est représenté dans le tableau 1.1. Chaque cellule du tableau représente ce qu'on appelle la pénétrance, qui est définie comme une probabilité conditionnelle d'observer le phénotype sachant un génotype spécifique. Dans notre cas, la pénétrance peut prendre les valeurs 1 ou 0, selon que le phénotype apparaît ou pas. Si la pénétrance prend la valeur 1, on parle alors de pénétrance complète.

Tableau 1.1 Relation entre génotypes et phénotypes au gène *ABO*

	Phénotype			
Génotype	Type *A*	Type *B*	Type *AB*	Type *O*
A/A	1	0	0	0
A/B	0	0	1	0
A/O	1	0	0	0
B/B	0	1	0	0
B/O	0	1	0	0
O/O	0	0	0	1

Note : les chiffres du tableau indiquent la pénétrance des génotypes.

Les génotypes A/A et A/O conduisent au même phénotype de type A. Puisque l'allèle A apparaît en présence ou non de l'allèle O, on dit alors que A est dominant par rapport à l'allèle O, qui lui est considéré récessif par rapport à A, puisqu'il n'a pas d'effet en sa présence. On peut faire la même conclusion concernant les allèles B et O. L'allèle O puisqu'il est récessif, n'a d'effet que s'il est présent en deux copies (dernière cellule du tableau). Par contre, les allèles A et B sont codominants car ils s'expriment tous les deux (phénotype AB) quand ils sont présents dans le génotype.

1.3 Arbre généalogique

L'étude de la génétique humaine repose en partie sur l'analyse des arbres généalogiques. L'arbre généalogique se décrit par un schéma (voir figure 1.2) où certaines conventions sont respectées ; parmi celles-ci, on cite : les individus du sexe masculin sont désignés par des carrés, et les individus du sexe féminin le sont par des cercles. Les couples sont reliés par un trait horizontal, à partir duquel un trait vertical va les relier par le biais d'un autre trait horizontal à leur progéniture. Les générations sont numérotées à l'aide de chiffres romains, et les sujets dans chaque génération sont numérotés à l'aide de chiffres arabes. Les sujets atteints d'une maladie génétique verront leurs carrés ou cercles noircis.

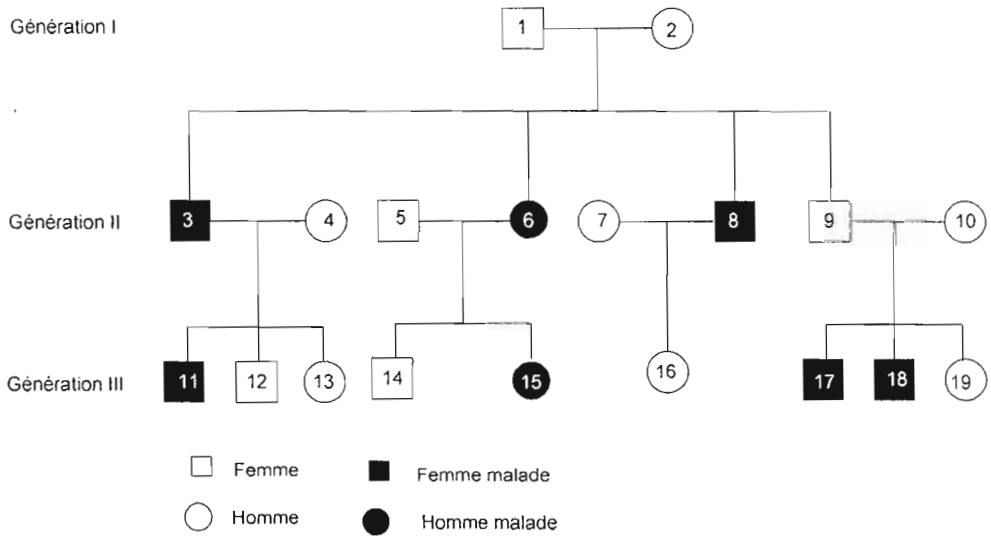

Figure 1.2 Exemple d'un arbre généalogique.

1.4 Mode de transmission

Lors de la formation des gamètes (fécondation), à un certain locus donné, chaque parent transmet un de ses allèles à son enfant avec une probabilité 1, mais un allèle particulier avec probabilité $\frac{1}{2}$; les allèles ainsi reçus par l'enfant de ses 2 parents constituent le génotype de l'enfant à ce locus.

Considérons l'arbre généalogique suivant (voir figure 1.3) ayant 3 générations I, II et III. Afin d'illustrer l'information dans un arbre généalogique, on se réfère aux individus de la génération I comme les grands-parents, ceux de la génération II comme les parents, et enfin ceux de la génération III comme les enfants.

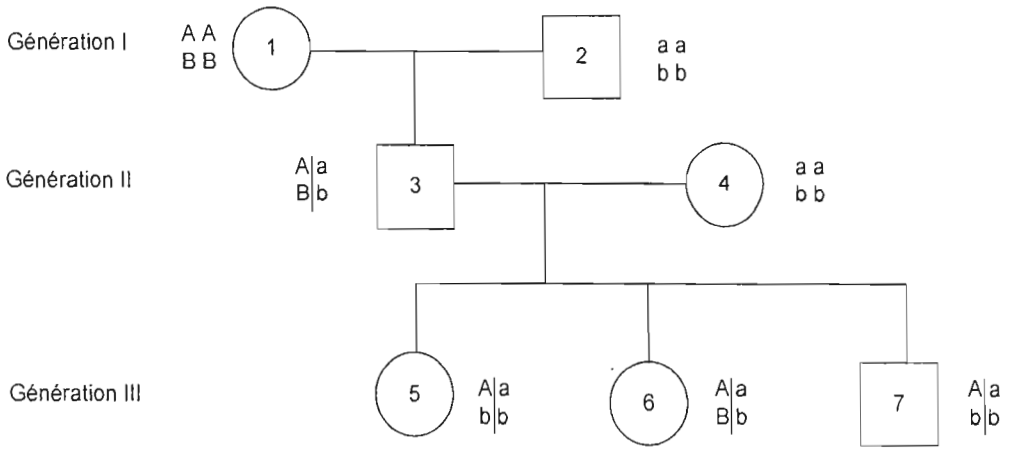

Figure 1.3 Famille artificielle ayant 2 génotypes à deux loci.

Dans la génération I, la grand mère (individu 1) et le grand père (individu 2) ont respectivement les génotypes suivants $ABAB$ et $abab$. Comme ces deux individus sont homozygotes aux deux loci considérés, cela nous permet de déduire la suite. Dans la génération II, le père (individu 3) a alors reçu les allèles (A, B) de la grand-mère, et (a, b) du grand-père. On parle alors d'haplotypes AB et ab qu'on peut définir comme des blocs d'allèles reçus *d'un même parent.*

Puisqu'on sait quels allèles ont été transmis ensemble (AB de la grand-mère, et ab du grand-père), on dit que la *phase* est connue. On indique cette information par la barre verticale dans le génotype (sur la gauche, on retrouve l'haplotype AB qui provient de la grand-mère, et sur la droite, ab qui provient du grand-père).

Quand le génotype d'un individu n'est pas marqué par la barre verticale, cela signifie que l'on ne connaît pas la phase, c'est-à-dire que l'on ne sait pas de quels parents il a reçu les génotypes, et c'est le cas des individus $1, 2,$ et 4. En principe, pour un individu doublement hétérozygote, comme dans le cas du père (individu 3), il n'est pas possible de connaître sa phase, sauf si on connaît le génotype des parents comme dans le cas élucidé ci-dessus.

1.5 Fréquences alléliques et génotypiques

Pour un locus donné, on s'intéresse à comprendre la relation qui existe entre *la fréquence génotypique (des génotypes) et la fréquence allélique (*ou *génique)*. Pour pouvoir décrire la constitution génétique d'une population, on a besoin de connaître le type et le nombre de génotypes des individus de la population.

Supposons qu'on a un locus B ayant deux allèles B_1 et B_2 présents chez les individus d'une certaine population. Trois génotypes peuvent être constitués avec ces deux allèles : B_1B_1, B_1B_2, B_2B_2. La constitution génétique de cette population sera entièrement décrite par les fréquences, ou proportions de ces trois génotypes pour ces individus. Ces fréquences sont appelées *les fréquences génotypiques*.

Néanmoins, la génétique des populations n'est pas seulement concernée par la constitution génique des individus de la population, mais aussi par *la transmission des gènes* d'une génération à une autre. Lors de la transmission, les génotypes des parents ne se reproduisent pas et, par conséquent, ne sont pas transmis, par contre les gènes eux, restent intacts. D'une façon générale, ce qui est transmis de génération en génération dans une population ce sont les gènes et non pas les génotypes. Si on continue sur le locus présenté en exemple ci-dessus alors la fréquence allélique de B_1 est la proportion de l'allèle B_1 par

rapport à tous les tous les allèles possibles présents à ce locus. On peut vérifier que la somme des fréquences de tous les allèles présents à ce locus est égale à 1.

Pour un groupe d'individus, la connaissance des fréquences des génotypes à un locus donné conduit à la détermination des fréquences alléliques. Nous allons illustrer ces définitions par l'exemple donné par Falconer (1974, p.6).

Supposons qu'on a encore un locus B ayant 2 allèles B_1 et B_2, et supposons qu'on a une population de 100 individus classés selon leurs génotypes. On peut alors obtenir les fréquences des allèles B_1 et B_2 dans la population; les résultats sont présentés dans le tableau 1.2.

Tableau 1.2 Exemple d'une population d'individus classés selon leurs génotypes.

		B_1B_1	B_1B_2	B_2B_2	Total	
Nombre d'allèles		30	60	10	100	
Nombre d'allèles	B_1	60	60	0	120	200
	B_2	0	60	20	80	

Comme nous avons 100 individus, et que chaque individu possède 2 allèles (celui transmis par la mère, et celui transmis par le père), donc au total, il y a 200 allèles. Un individu qui possède le génotype B_1B_1 possède deux allèles B_1, par contre un individu possédant le génotype B_1B_2 contient un allèle B_1 et un allèle B_2. Au total, il y a 120 allèles B_1, et 80 allèles B_2. Les fréquences des allèles B_1 et B_2 sont alors respectivement égales à 0.6 et 0.4. D'une façon générale, on peut exprimer les fréquences alléliques en fonction des fréquences génotypiques (voir formule 1.1). Pour cela, on pose p et q les fréquences respectives des allèles B_1 et B_2, et $P, H,$ et Q les fréquences respectives des

génotypes B_1B_1, B_1B_2, et B_2B_2 (voir tableau. 1.3) . Au préalable, les conditions suivantes doivent alors être remplies :

$$p+q=1 \text{ et } P+H+Q=1.$$

Tableau 1.3 Fréquences des allèles et des génotypes.

	allèles		Génotypes		
Fréquences	B_1	B_2	B_1B_1	B_1B_2	B_2B_2
	p	q	P	H	Q

Nous allons pouvoir maintenant exprimer les fréquences des allèles en fonction de celles des génotypes. Comme chaque individu possède deux allèles, alors la proportion de l'allèle B_1 est égale à :

$$p=\frac{1}{2}(2P+H)$$

$$=P+\frac{1}{2}H.$$

Ceci nous permet alors d'écrire les fréquences respectives p et q des deux allèles B_1 et B_2 en fonction de celles des génotypes comme suit :

$$\begin{cases} p=P+\frac{1}{2}H \\ q=Q+\frac{1}{2}H. \end{cases} \qquad (1.1)$$

1.6 Principe d'équilibre de Hardy-Weinberg

Dans leur modèle, Hardy et Weinberg ont supposé les hypothèses suivantes :

- la population est de taille infinie,
- les gamètes (ovules et spermatozoïdes) s'associent au hasard c'est-à-dire que les accouplements sont aléatoires (hypothèse de panmixie),
- il n'y a pas de sélection dans la population,
- il n'y a ni mutation, ni migration dans la population.

Alors, les fréquences des allèles et des génotypes au cours des générations suivent une loi simple appelée loi de Hardy-Weinberg, qui constitue le modèle de référence en génétique des populations. Cette loi stipule que les fréquences alléliques et les fréquences génotypiques restent constantes de génération en génération. On dit alors que la population est à l'équilibre, et il existe une relation simple entre les fréquences alléliques et les fréquences génotypiques. Cette loi a été établie indépendamment en 1908 par le médecin W. Weinberg (1862-1937) et le mathématicien G.H.Hardy (1877-1947). Dans ce qui suit, nous allons expliquer les fondements de ladite loi.

Pour cela, supposons qu'on a toujours un locus donné B ayant deux allèles B_1 et B_2, on peut alors avoir six types de croisements, comme le montre le tableau 1.4 (Hartl, 1994, p.24). Quand les accouplements se font au hasard, chacun des types est produit en proportion des fréquences génotypiques dans la population. Si nous prenons le croisement $B_1B_1 \times B_1B_1$, celui-ci ne se produit que si un mâle de type B_1B_1 est croisé avec une femelle de type B_1B_1. Ce croisement se fait avec une fréquence $P \times P = P^2$ (tab. 1.3). De la même façon, le croisement $B_1B_2 \times B_2B_2$ se produit lorsqu'un mâle B_1B_2 se croise avec une femelle B_2B_2 avec une proportion HQ, ou une femelle B_1B_2 se croise avec un mâle B_2B_2 avec une proportion HQ; la proportion totale du croisement $B_1B_2 \times B_2B_2$ est donc la somme des deux proportions, soit $2HQ$ (tableau. 1.3) .

Les génotypes des enfants issus des différents croisements sont donnés dans les trois dernières colonnes du tableau 1.4. Les fréquences de ces génotypes sont obtenues à partir de la loi de ségrégation de Mendel qui stipule qu'un croisement de B_1B_1 avec B_2B_2 ne produit qu'une descendance de B_1B_2 et, un croisement de B_1B_1 avec B_1B_2 produit une descendance composée de moitié de B_1B_1 et de moitié de B_1B_2 par contre, un croisement de B_1B_2 avec B_1B_2 produit dans la descendance $\frac{1}{4}B_1B_1$, $\frac{1}{4}B_2B_2$ et $\frac{1}{2}B_1B_2$.

Ainsi, pour trouver les fréquences respectives P', H', et Q' des génotypes B_1B_1, B_1B_2 et B_2B_2 après une génération de croisement au hasard, on doit sommer pour chaque colonne, le produit entre *la fréquence du croisement* et *la fréquence du génotype des enfants*. C'est-à-dire :

Tableau 1.4 Démonstration de la loi de Hardy-Weinberg.

		Fréquences des génotypes des enfants		
Croisements	Fréquences des croisements	B_1B_1	B_1B_2	B_2B_2
$B_1B_1 \times B_1B_1$	P^2	1	0	0
$B_1B_1 \times B_1B_2$	$2PH$	1/2	1/2	0
$B_1B_1 \times B_2B_2$	$2PQ$	0	1	0
$B_1B_2 \times B_1B_2$	H^2	1/4	1/2	1/2
$B_1B_2 \times B_2B_2$	$2HQ$	0	1/2	1/2
$B_2B_2 \times B_2B_2$	Q^2	0	0	1
	Totaux (génération suivante)	P'	H'	Q'

Ces simplifications ont été obtenues en utilisant la formule (1.1). En conclusion, on obtient les fréquences génotypiques de B_1B_1, B_1B_2 et B_2B_2 qui sont respectivement de $p^2, 2pq,$ et q^2, ce qui est le résultat connu sous le nom de l'équilibre de Hardy-Weinberg.

La relation entre les fréquences alléliques et les fréquences génotypiques revêt une grande importance, du fait que beaucoup de déductions faites en génétique des populations et en génétique quantitative reposent sur celle-ci. Cette relation est la suivante : si les fréquences alléliques de deux allèles B_1 et B_2 parmi les parents sont p et q, alors les fréquences génotypiques parmi la progéniture sont :

$$f(B_1B_1) = p^2,\ f(B_1B_2) = 2pq,\ f(B_2B_2) = q^2.$$

Ce qui se résume dans le tableau ci-dessous

	Gènes des parents		Génotypes dans la progéniture		
Fréquences	B_1	B_2	B_1B_1	B_1B_2	B_2B_2
	p	q	p^2	$2pq$	q^2

Pour conclure, en supposant les hypothèses de Hardy-Weinberg, dans une génération, on peut à partir des fréquences alléliques déduire les fréquences génotypiques, et celles-ci restent stables à partir d'une première génération.

1.7 Recombinaison et liaison génétiques

1.7.1 Recombinaison génétique

Pour illustrer le concept de recombinaison, revenons à la figure 1.3, et considérons le père hétérozygote AB/ab, représenté par l'individu 3. Celui-ci peut transmettre quatre types d'haplotypes à sa descendance : AB, ab, Ab, aB (génération III).

Les haplotypes AB et ab sont les mêmes que ceux transmis par les grands-parents, par contre les haplotypes Ab et aB sont différents des haplotypes reçus par le père (donnés par les grands-parents). Ces deux derniers haplotypes comportent un allèle provenant de chaque grands-parents. Pour l'haplotype Ab, l'allèle A provient de la grand-mère, et l'allèle b du grand-père, pour l'haplotype aB, l'allèle a provient du grand-père, et B de la grand-mère.

Ainsi une recombinaison entre 2 loci est définie comme le phénomène par lequel deux grands-parents contribuent par la transmission d'un allèle chacun, pour constituer *l'haplotype de leur descendance* (génération). Le gamète ainsi transmis est un mélange des deux haplotypes du parent.

Au contraire, quand il n'y a pas de recombinaison entre deux loci, l'haplotype d'un individu ainsi produit est constitué de 2 allèles transmis par *un même parent*.

Dans notre exemple, l'enfant 6 est *non recombinant* pour l'haplotype reçu par le père, mais on ne peut pas dire, si oui ou non, il est recombinant pour l'haplotype reçu par la mère car elle est homozygote aux deux loci (elle transmet donc l'haplotype *ab*). Alors que l'enfant 7 est un *recombinant* pour le père puisqu'il a reçu l'haplotype recombiné Ab.

Il faut noter qu'une recombinaison se produit entre deux loci à un taux relié à la distance physique les séparant sur un même chromosome. Plus les loci sont proches, moins on a de recombinaisons, plus ils sont loin l'un de l'autre, plus le taux de recombinaison augmente.

1.7.2 Liaison génétique (*genetic linkage*)

La liaison génétique est définie comme la co-ségrégation de deux, ou plusieurs gènes (ou loci), au cours des générations en raison de leur proximité physique sur le génome. Le degré de liaison mesuré par le pourcentage de recombinaisons entre loci, indique une distance génétique que l'on exprime en centimorgans (voir aussi Déséquilibre de liaison section 1.8.1). Elle est mesurée par un taux (ou fraction) de recombinaison qui est défini comme étant la probabilité qu'un gamète produit par un parent soit un *recombinant*. Ce taux est généralement dénoté par θ, et par différence, $1-\theta$ représente le taux des non recombinants ; on a

$0 \leq \theta \leq 0,5$. Deux gènes parfaitement liés vont toujours être transmis ensemble lors de la *méiose* (division cellulaire), et dans ce cas $\theta = 0$, alors que deux gènes qui ne sont pas liés (s'ils sont par exemple situés sur des chromosomes différents), sont complètement indépendants, et $\theta = 0,5$. En d'autres termes, la liaison génétique est le phénomène par lequel les allèles d'un haplotype transmis par un grand parent au parent, auront tendance à être transmis encore du parent au fils. On dit alors, que les allèles des différents gènes sont génétiquement liés.

Il faut noter que dans les analyses de liaison, les valeurs de $\theta > 0,5$ sont biologiquement impossibles.

L'analyse de liaison est un outil très puissant pour détecter la présence d'un gène malade dans une région chromosomique, ce qui sera développé dans la suite.

1.8 Identification des gènes candidats

Un des grands objectifs de la recherche en génétique est de trouver les gènes causant les maladies de type génétique. Il y a principalement deux types d'études permettant d'identifier le, ou les gènes causant une maladie :

- les études de liaison,
- les études d'association.

Nous allons maintenant décrire ces deux types d'études.

1.8.1 Études de liaison génétique

Les études de liaison génétique s'appuient sur l'information génétique recueillie sur des familles. Dans ce genre d'études, pour localiser les gènes responsables d'une maladie, on essaie de trouver un ou plusieurs marqueurs[3] proches de la zone chromosomique impliquée, et dont une combinaison allélique semble presque toujours associée à la maladie, On essaie

[3] Les marqueurs génétiques sont des balises dont on connaît l'emplacement sur le chromosome.

alors d'élaborer un test de liaison dont le principe est de détecter chez l'individu malade, la présence ou non d'allèles (gènes) liés à la maladie.

Comme on a déjà vu, deux loci sont liés s'ils sont situés proches l'un de l'autre, ce qui diminue donc les chances de recombinaison $(\theta \cong 0)$ et, par le fait même, augmente les chances que leurs allèles soient transmis ensemble. Le locus de la maladie n'étant pas connu, on cherche alors des marqueurs situés dans son voisinage. Si l'allèle d'un marqueur est fréquemment transmis aux personnes malades dans les familles, on dira que ce marqueur est lié à la maladie. On peut alors concentrer la recherche du gène responsable autour de ce marqueur.

Les analyses de liaison sont considérées comme un outil très puissant pour détecter la position d'un gène causant une maladie dans une région chromosomique. Il s'agit d'analyses génétiques par lesquelles on dresse une cartographie des gènes.

1.8.2 Études d'association

On a vu que la liaison est la tendance qu'un petit segment de chromosome soit transmis par un parent sous forme intacte à ses enfants. Le résultat est que certains haplotypes dans ce court segment peuvent être préservés sur de longues générations. Cette excessive co-ségrégation allélique de certains haplotypes est appelée association allélique.

Les études d'association ont pour but d'identifier un marqueur pour lequel un allèle est plus fréquent chez les sujets malades que chez les sujets sains. On dit alors que cet allèle est associé à la maladie. Il est possible d'effectuer ces études à partir d'échantillons de cas et de témoins, mais également à partir d'échantillons de familles. Une association peut être due, soit à l'existence d'une liaison, soit à une stratification de la population, soit à d'autres facteurs qui ne seront pas considérés ici. Le cas où l'association apparaît à cause de la stratification de la population sera développé à la fin de cette section.

Les études d'association sont utilisées dans le but de confirmer l'implication d'un allèle qu'on pense être important dans l'étude d'une maladie ou pour découvrir de nouveaux gènes pouvant jouer un rôle dans la maladie. Pour illustrer, considérons le cas de deux loci A et B

ayant des allèles $A_1, A_2, ..., A_m$ et $B_1, B_2, ..., B_n$ se produisant dans la population avec des fréquences relatives $p_1, p_2, ..., p_m$, et $q_1, q_2, ..., q_n$ (Almgren et al., 2003). Au total, il y a $m \times n$ haplotypes qu'on note $A_1B_1, A_1B_2, ..., A_mB_1, A_2B_1, ..., A_mB_n$, se produisant avec des fréquences relatives $h_{11}, h_{12}, ...h_{m1}, h_{21}, ..., h_{mn}$. Si la présence de l'allèle A_i et la présence de l'allèle B_j sur l'haplotype A_iB_j sont des événements indépendants, alors la fréquence relative de la présence conjointe des allèles A_i et B_j dans le gamète sera est égale à :

$$h_{ij} = p_i q_j .$$

Si cette égalité n'est pas respectée, la rencontre entre A_i et B_j n'est donc pas due au hasard et alors on dit que les *allèles* A_i et B_j *sont associés*.

Soit θ le taux de recombinaison entre les loci A et B, et soit $h_{ij}^{(0)}$ la fréquence relative de l'haplotype A_iB_j dans la génération courante, c'est à dire la génération 0. La question est de savoir, quelle sera la fréquence du même haplotype dans la prochaine génération sous l'hypothèse des accouplements aléatoires. Dans la prochaine génération, chaque haplotype est un recombinant avec probabilité θ, ou un non recombinant avec probabilité $1-\theta$ par rapport aux loci A et B. Quand l'haplotype est un non recombinant, il a une probabilité $h_{ij}^{(0)}$ pour qu'il soit A_iB_j; quand il est un recombinant, cette probabilité est simplement $p_i q_j$, toujours sous l'hypothèse d'accouplements aléatoires. Donc la probabilité qu'un haplotype transmis à la prochaine génération soit A_iB_j est égale à :

$$h_{ij}^{(1)} = (1-\theta) h_{ij}^{(0)} + \theta\, p_i q_j . \qquad (1.2)$$

A partir de l'équation (1.2), on peut déduire l'accroissement relatif des fréquences de l'haplotype A_iB_j entre la génération 0 et la génération 1, accroissement qu'on trouve égal à :

$$h_{ij}^{(1)} - h_{ij}^{(0)} = \theta\left(p_i q_j - h_{ij}^{(0)}\right).$$

L'accroissement de la fréquence de l'haplotype dépend de l'association ; s'il n'y a pas d'association dans la génération courante, c'est-à-dire $h_{ij}^{(0)} = p_i q_i$, la fréquence de l'haplotype ne changera pas, par contre s'il y a association, l'accroissement est proportionnel à θ. S'il y a changement dans les fréquences des haplotypes entre les générations, les deux loci sont dits être *en déséquilibre de phase gamétique*. Par contre, le taux auquel la population aléatoire approche l'équilibre de *la phase gamétique* dépend du taux de recombinaison θ. On voit cela, en réécrivant (1.2) comme suit :

$$\begin{aligned} h_{ij}^{(1)} - p_i q_j &= (1-\theta) h_{ij}^{(0)} + \theta\, p_i q_j - p_i q_j \\ &= (1-\theta) h_{ij}^{(0)} + (1-\theta)\, p_i q_j \end{aligned}$$

Ce qui donne :

$$h_{ij}^{(1)} - p_i q_j = (1-\theta)\left(h_{ij}^{(0)} - p_i q_j\right).$$

Cette équation nous montre clairement qu'à la génération 1, la différence entre la fréquence de l'haplotype et sa valeur d'équilibre est diminuée par le facteur $(1-\theta)$, donc, après k générations, on aura :

$$h_{ij}^{(k)} - p_i q_j = (1-\theta)^{(k)}\left(h_{ij}^{(0)} - p_i q_j\right). \qquad (1.3)$$

Lorsque k devient grand $(k \to +\infty)$, h_{ij}^{k} se rapproche de la valeur d'équilibre $p_i q_j$. La différence entre la fréquence relative de l'haplotype et sa valeur d'équilibre est dans certains cas, utilisée comme une mesure du degré d'association entre les allèles.

Si on détecte une association entre deux loci, ce résultat suggère que ces loci sont liés. Cependant, et c'est là où émerge la difficulté, l'association peut se produire entre des loci liés ou non liés. Ce dernier cas peut se produire lorsqu'on est en présence d'une stratification de la population. Il faut alors tenir compte de ce cas dans le cadre des analyses de liaison pour ne

pas aboutir à des conclusions erronées. Avant d'illustrer cette situation par un exemple, on va d'abord définir clairement ce qu'on entend par la stratification de la population. Une *population stratifiée* est une population qui dans l'apparence est homogène, mais qui est composée de sous-populations différentes entre elles. Cette définition nous permet de comprendre l'exemple suivant tiré de Almgren et al. (2003).

Considérons trois populations qui ont chacune atteint l'équilibre de la phase gamétique par rapport aux allèles A_i et B_j (c'est-à-dire $h_{ij} = p_i q_j$).

On suppose que la taille de la population totale est N, et on suppose que les fréquences relatives des allèles A_i, B_j, A_iB_j pour chaque sous population sont telles qu'indiquées dans le tableau 1.5.

Tableau 1.5 Exemple d'illustration d'une association entre allèles en situation d'équilibre de la phase gamétique

N	A_i	B_j	A_iB_j
1000	0,300	0,500	0,150
2000	0,200	0,400	0,080
10000	0,050	0,100	0,005

Si ces trois sous populations sont fusionnées, calculons les fréquences relatives des allèles et des haplotypes avant qu'aucun croisement n'ait lieu. On a :

$$P(A_i) = (0,30\times1000 + 0,20\times2000 + 0,050\times10000)/13000 = 0,0923$$
$$P(B_i) = (0,50\times1000 + 0,40\times2000 + 0,100\times10000)/13000 = 0,1770$$
$$P(A_iB_j) = (0,15\times1000 + 0,08\times2000 + 0,005\times10000)/13000 = 0,0277.$$

La fréquence relative d'équilibre pour A_iB_j est $0,0923\times0,1770 = 0,0163$ qui

est différente de $0,0277$. Puisque $P(A_i B_j) \neq P(A_i)P(B_j)$, alors les allèles A_i et B_j sont associés dans la population fusionnée. On voit bien que dans chaque population prise à part, les allèles A_i et B_j ne sont pas associés, puisque pour chaque sous population on a $P(A_i B_j) = P(A_i)P(B_j)$; par contre en fusionnant les trois populations, on trouve une association entre ces deux allèles. On conclut que l'association trouvée entre les allèles A_i et B_j est due à la stratification de la population, et non pas à la liaison génétique entre les deux allèles.

Ce qui nous intéresse dans le cadre de notre étude est de tester et de trouver une association entre deux loci qui est due au déséquilibre de liaison et non pas à la stratification de la population (considérée comme une fausse association).

1.8.3 Déséquilibre de liaison (linkage disequilibrium)

Les études d'association sont fondées sur le principe du déséquilibre de liaison (DL). On définit le déséquilibre de liaison comme une association non aléatoire entre les allèles de deux loci. Plus deux loci sont proches l'un de l'autre, moins leurs allèles ont de chance d'être transmis de façon indépendante. Alors, on dit d'eux qu'ils sont en déséquilibre de liaison. Néanmoins, cette association doit nécessairement être causée par la liaison gamétique telle que déjà définie (et non par une stratification de la population).

S'il existe un déséquilibre de liaison, cela signifie que les gènes appartenant à des loci différents ne sont pas associés au hasard dans la population : certaines combinaisons sont moins fréquentes, d'autres plus fréquentes que ne le voudrait le hasard si on avait une association aléatoire de ces différents gènes (Henry, 2003). Cette définition nous amène à l'expression formalisée du déséquilibre de liaison. Pour cela considérons deux loci :

- le premier locus a deux allèles A_1 et A_2, avec des fréquences respectives p_{A_1} et p_{A_2}.
- le second locus a deux allèles B_1 et B_2, avec des fréquences respectives p_{B_1} et p_{B_2}.

Au lieu de s'intéresser aux génotypes des individus, on va plutôt considérer les gamètes produits par ces deux loci. Ce choix se justifie par le fait que la dépendance ou l'indépendance qui existe dans une génération résulte de celle qu'il y a avait dans la génération des gamètes qui l'ont formée.

En considérant ces deux loci, on a au total quatre combinaisons gamétiques A_1B_1, A_1B_2, A_2B_1, et A_2B_2. Soit une de ces quatre combinaisons, prenons par exemple A_1B_1. S'il y a une transmission due au hasard dans la population gamétique (indépendance statistique), donc pas de déséquilibre de liaison, on aura :

$$p_{AB} = p_A . p_B . \qquad (1.4)$$

Dans le cas contraire, c'est-à-dire s'il n'y a pas indépendance statistique, donc s'il existe un déséquilibre, on ne peut plus écrire cette égalité, puisque la fréquence de ce gamète sera soit supérieure, ou inférieure au produit des fréquences des deux gamètes de valeur $p_A p_B$. On peut alors écrire (1.4), en ajoutant à la fréquence $p_A p_B$ une quantité δ qui peut être une valeur positive, ou négative dépendamment du fait que cette fréquence (p_{AB}) est supérieure ou inférieure au produit $p_A p_B$. Ce qui permet d'écrire :

$$p_{AB} = p_A p_B + \delta \Leftrightarrow \delta = p_{AB} - p_A p_B .$$

C'est le paramètre δ qui mesure le déséquilibre de liaison dans la population pour ces deux gènes, soit l'écart par rapport à l'association aléatoire.

Le concept de déséquilibre de liaison est très utile pour la recherche en génétique. Ceci s'explique par le fait que, si on trouve une association entre un marqueur et une maladie, cela suggère qu'il existe un déséquilibre de liaison entre eux, en d'autres mots que le locus marqueur et le locus de la maladie sont liés. On devine alors que le marqueur est situé à proximité du locus de la maladie, ce qui facilite alors l'identification du gène.

Nous allons dans le chapitre suivant introduire un test dans lequel nous allons utiliser les notions que nous venons de définir, en particulier les notions d'association et de liaison génétique, il s'agit du TDT.

CHAPITRE II

LE TEST DU TDT (TRANSMISSION TEST FOR LINKAGE DISEQUILIBRIUM)

Tel que déjà mentionné, en génétique, une des problématiques est de trouver si la maladie qui fait l'objet de l'étude est d'ordre génétique, ou pas ; en d'autres mots, s'il y a un, ou plusieurs gènes responsables de la maladie. Pour vérifier si la maladie est d'ordre génétique, des études sur des jumeaux ou sur des familles sont faites afin de recenser des gènes candidats, c'est-à-dire des gènes qui pourraient causer la maladie. A partir de ce ou ces gènes candidats (potentiellement responsables de la maladie), on cherche à voir s'ils jouent un rôle dans le développement de la maladie. En termes génétiques, on cherche à trouver, s'il y a ou non, présence d'un déséquilibre de liaison. Comme les marqueurs génétiques ne sont pas toujours disponibles dans le gène candidat, on peut utiliser des marqueurs génétiques situés près du locus candidat et dont on connaît la position ; alors il s'agit de montrer qu'il y a une liaison entre ces marqueurs génétiques et le statut de la maladie.

Il existe plusieurs méthodes testant cette liaison, dont les plus connues sont celles basées sur l'étude des paires d'enfants affectés appelées « ASP » (*affected sib pairs*), ou les méthodes appelées « AFBAC » (*affected family-based con*trols) qui comparent, en utilisant le test du χ^2, le nombre d'allèles transmis aux enfants malades (cas) au nombre d'allèles transmis aux enfants non malades (témoins).

Ces méthodes ont presque toujours échoué dans leur objectif de trouver un déséquilibre de liaison (Spielman et al. 1993), car elles sont sensibles à la stratification de la population ; c'est-à-dire qu'on peut trouver une association qui est due, non pas au déséquilibre de liaison, mais à la stratification de la population. En réponse à ce problème, Spielman et al. (1993) ont

proposé le test du TDT (Transmission Test for Linkage Disequilibrium), qui est un test qui n'est nullement affecté par la présence d'une stratification de la population. Ce dernier teste la liaison en présence d'une association allélique entre le marqueur et la maladie. Un des succès de ce test se réfère au gène responsable du diabète.

Le TDT s'applique aux traits binaires, c'est-à-dire au cas où l'état de l'individu est décrit par deux valeurs : malade ou pas malade. Pour utiliser le TDT, il suffit de disposer de familles ayant au moins un enfant malade, contrairement aux tests cités ci-dessus, qui eux nécessitent des données sur des paires de frères et sœurs malades (dans le cas des ASP), ou des membres affectés et non affectés d'une même famille (dans le cas des AFBAC); ce type de données n'est pas toujours facile à collecter surtout lorsqu'il s'agit d'une maladie rare d'où l'avantage du TDT.

2.1 Exemple illustratif

Afin de mieux comprendre les notions de « transmis » et de « non transmis » utilisées dans le TDT, nous allons présenter l'exemple illustratif suivant.

Soit un ensemble de trente familles classées selon le génotype des parents, et selon les allèles transmis et non transmis à leurs enfants. Considérons un locus ayant deux allèles *B* et *b*. La figure 2.1 représente les familles classées selon le génotype des parents, et selon les types d'allèles transmis à leurs enfants.

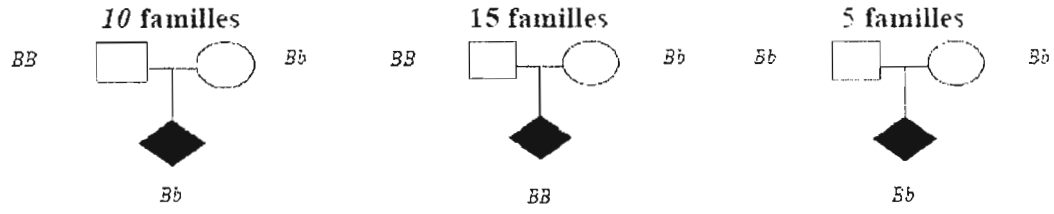

Figure 2.1 Exemple de trente familles classées selon le génotype et selon le type d'allèles transmis

On a un total de trente familles, parmi elles, il y en a dix dont un des parents est homozygote *BB* et l'autre parent est hétérozygote *Bb*. La figure montre aussi que ces parents

ont transmis à leurs enfants, les allèles B et b; l'allèle B a été transmis par les parents homozygotes (puisqu'ils ne possèdent que l'allèle B), et l'allèle b a, par conséquent, été transmis par les parents hétérozygotes. En se référant au tableau 2.1, on peut lire, qu'il y a dix allèles b transmis par les parents hétérozygotes et dix allèles B non transmis par ces mêmes parents, et il y a dix allèles B transmis par les parents homozygotes BB et dix allèles B non transmis par ces mêmes parents. On peut raisonner de la même façon pour les autres familles restantes, et trouver qu'au total il y a soixante allèles transmis et non transmis (soit 2×30).

Tableau 2.1 Exemple d'allèles transmis et non transmis

Allèles transmis	Allèles non transmis		
	B	*b*	*Total*
B	10+15	15+5	45
b	10+5	0	15
Total	40	20	60

L'idée est de comparer au niveau de l'ensemble des familles, si l'allèle qu'on suppose être responsable de la maladie (B par exemple) a été transmis plus que ne le voudrait le hasard; il s'agit donc de comparer le nombre de fois où cet allèle a été transmis (soit 45), et le nombre de fois où il n'a pas été transmis (soit 40).

2.2 Présentation du TDT

Le TDT est un test de la liaison et d'association entre une maladie et un locus marqueur. L'avantage de ce test est multiple, en ce sens que, d'une part, il n'est nullement affecté par la présence d'une stratification de population, d'autre part, comparé à d'autres tests de liaison, il ne nécessite pas de données sur des familles ayant plusieurs enfants affectés, ou de familles ayant des enfants non affectés (cas témoins). Le TDT peut fonctionner avec des familles ayant au moins un enfant malade (trios dans le cas où on a un enfant malade et ses deux parents) correspondant au plan d'expérience à un facteur. Par

contre, l'inconvénient est qu'il ne peut détecter de liaison entre un locus malade et un locus marqueur qu'en présence d'une association due à un déséquilibre de liaison, d'où son nom test de déséquilibre de liaison (liaison due à un déséquilibre de liaison).

Dans ce chapitre, on va considérer seulement le cas, où on a un seul enfant affecté par famille, car les données que nous allons utiliser pour notre étude portent sur un échantillon de familles ayant un seul enfant malade.

On considère alors un locus responsable de la maladie D (*diseased*) possédant deux allèles, D_1 et D_2, qui correspondent aux états malade et non malade ; soit aussi un locus marqueur M ayant deux allèles M_1 et M_2.

On introduit les notations suivantes :

D = *locus de la maladie*

D_1 = *allèle malade*

D_2 = *allèle normal*

M = *locus marqueur*

M_1, M_2 = *allèles au locus marqueur.*

Par hypothèse, on considère que l'allèle D_2 est dominant par rapport à D_1, c'est-à-dire que la maladie est récessive (Ott 1989). Il faut donc qu'un individu ait deux allèles D_1 pour manifester la maladie.

Soit un échantillon de n familles ayant un seul enfant affecté. Au locus M, on aura, alors, $4n$ allèles parentaux :

- $2n$ transmis,
- $2n$ non transmis.

car à un locus donné, chaque parent possède deux allèles et transmet un seul allèle à son enfant (avec une probabilité $\frac{1}{2}$), comme on a deux parents, on a donc $2n$ allèles qui sont transmis et donc $2n$ allèles non transmis.

Soit M_iM_j, $i=1,2$, $j=1,2$ le génotype du parent au locus M (on se réfère au parent par son génotype au locus M). Chaque parent de génotype M_iM_j transmet à son enfant l'allèle M_i ou M_j. L'objectif recherché est de voir si l'allèle M_1 qu'on soupçonne être associé à la maladie est lié au locus malade (D). Pour cela, on va comparer le nombre d'allèles M_1 transmis et le nombre d'allèles M_1 non transmis à l'enfant malade, en utilisant un test d'indépendance du χ^2 dont on va définir la forme. Le principe de base de ce test est de considérer les allèles transmis par les parents à leurs enfants malades comme étant les *cas*, et les allèles non transmis comme les allèles *témoins*. Les hypothèses du test sont les suivantes :

H_0 : *il n'existe pas de liaison entre le marqueur et la maladie,*
H_1 : *il existe une liaison entre le marqueur et la maladie.*

Ces hypothèses se traduisent mathématiquement par :

$$H_0 : \theta = \frac{1}{2} \quad \text{contre} \quad H_1 : \theta < \frac{1}{2},$$

où θ représente le taux de recombinaison, entre le locus marqueur et le locus malade.

2.3 Construction du test du TDT

Pour la construction du test recherché, on se concentre sur les $2n$ parents (au lieu des $4n$ gènes parentaux qui correspondent aux n familles). Le tableau 2.2 suivant décrit l'ensemble des transmissions des allèles du locus marqueur M pour les $2n$ parents (le père et la mère). Posons :

a = *est le nombre de fois où le parent* M_1M_1 *transmet* M_1 *à son enfant*,

b = *est le nombre de fois où le parent* M_1M_2 *transmet* M_1 *à son enfant*,

c = *est le nombre de fois où le parent* M_1M_2 *transmet* M_2 *à son enfant*,

d = *est le nombre de fois où le parent* M_2M_2 *transmet* M_2 *à son enfant*.

Ainsi les $2n$ parents sont divisés en quatre catégories, *a, b, c, d,* comme indiqué dans notre exemple illustratif.

Tableau 2.2 Combinaisons des allèles marqueurs M_1 et M_2 transmis et non transmis parmi les $2n$ parents ayant n enfants affectés.

Allèles transmis	Allèles non transmis		Total
	$M_1 = 0$	$M_2 = 0$	
$M_1 = 1$	a	b	$a+b$
$M_2 = 1$	c	d	$c+d$
Total	$a+c$	$b+d$	$2n$

À partir de ce tableau, on peut lire que le nombre d'allèles M_1 transmis est égal à $(a+b)$, et le nombre d'allèles M_1 non transmis est égal à $(a+c)$. Comme on veut comparer la fréquence d'allèles M_1 transmis, et la fréquence d'allèles M_1 non transmis, on va alors s'intéresser à comparer $(a+b)$ et $(a+c)$.

Pour connaître quelles variables dépendent du taux de recombinaison θ, et s'assurer ainsi de prendre les valeurs qui définissent correctement les hypothèses du test, Ott (1989) a proposé d'écrire un tableau correspondant au tableau 2.2, mais en introduisant simultanément les probabilités et les paramètres suivants :

- m, la fréquence de M_1 dans la population, et donc $1-m$ est la fréquence de M_2 dans la population,
- p, la fréquence de l'allèle D_1 dans la population,
- θ, le taux de recombinaison entre le locus marqueur M et le locus du gène causant la maladie D,
- δ, le coefficient mesurant le déséquilibre de liaison, où δ est défini comme suit :

$$\delta = P(M_1D_1) - P(M_1)P(D_1) = P(M_1D_1) - mp.$$

Nous allons définir aux deux loci pris conjointement quatre types d'haplotypes : M_1D_1, M_1D_2, M_2D_1, et M_2D_2. Dans la population, les fréquences respectives des ces quatre haplotypes sont définies comme suit :

$$\begin{aligned} h_{11} &= P(M_1D_1) = mp + \delta, \\ h_{12} &= P(M_1D_2) = m(1-p) - \delta, \\ h_{21} &= P(M_2D_1) = (1-m)p - \delta, \\ h_{22} &= P(M_2D_2) = (1-m)(1-p) + \delta, \end{aligned} \qquad (2.1)$$

où h_{ij} est la fréquence de l'haplotype $M_iD_j, (\mathrm{i} = 1,2, \mathrm{j} = 1,2)$.

Ces nouveaux paramètres vont nous servir à calculer les probabilités des combinaisons de M_i et M_j, entre les transmis et les non transmis à l'enfant malade (voir tableau 2.3).

Pour faire ce calcul, on doit supposer d'abord que les accouplements sont aléatoires, et par là, il y a indépendance dans la transmission des allèles par les deux parents. On note aussi par F_1, F_2 (*father* 1 et *father* 2) les deux parents respectifs, et par C (*child*) l'enfant. Soit l'événement A : l'enfant C est malade.

Les cases du tableau 2.3 correspondent chacune à la probabilité conditionnelle par rapport à A de l'évènement T_{ij}, où T_{ij} représente l'intersection entre les deux événements

suivants : le parent F_1 possède le génotype M_iM_j et transmet l'allèle M_i à son enfant. Ainsi T_{ij} s'écrit comme :

$$T_{ij} = \{F_1 = M_iM_j, F_1 \to M_i\} \tag{2.2}$$

Tableau 2.3 Probabilités des combinaisons des allèles marqueurs M_1 et M_2 transmis et non transmis parmi les $2n$ parents ayant n enfants affectés.

Allèles	Allèles non transmis		Total
transmis	$M_1 = 0$	$M_2 = 0$	
$M_1 = 1$	$m^2 + (m\delta/p)$	$m(1-m)+[(1-\theta-m)\delta/p]$	$m+[(1-\theta)\delta/p]$
$M_2 = 1$	$m(1-m)+[(\theta-m)\delta/p]$	$(1-m)^2-\left[(1-m)\delta/p\right]$	$(1-m)-[(1-\theta)\delta/p]$
Total	$m+(\theta\delta/p)$	$(1-m)-(\theta\delta/p)$	1

Le théorème de Bayes permet d'écrire :

$$P(T_{ij} \mid A) = \frac{P(T_{ij})P(A \mid T_{ij})}{P(A)}. \tag{2.3}$$

L'équation (2.3) sera déterminée par le calcul de chaque probabilité séparément, c'est-à-dire de $P(A)$ et de $P(T_{ij})\,P(T_{ij} \mid A)$

Le calcul de la probabilité que l'enfant soit malade, c'est-à-dire $P(A)$ se fait en conditionnant par rapport au génotype D_iD_j de la maladie de l'enfant, soit :

$$P(A) = \sum_{k,l} P(A \mid D_k D_l) P(D_k D_l)$$

$$= P(A \mid D_1 D_1) P(D_1 D_1) + P(A \mid D_1 D_2) P(D_1 D_2) + P(A \mid D_2 D_2) P(D_2 D_2)$$

$$= P(A \mid D_1 D_1) p^2 + P(A \mid D_1 D_2) 2p(1-p) + P(A \mid D_2 D_2)(1-p)^2$$

Nous avons posé l'hypothèse sur la récessivité de la maladie, ce qui signifie que pour qu'un enfant soit malade (évènement A), il doit avoir reçu l'allèle malade D_1 de ses deux parents, ce qui correspond au génotype $D_1 D_1$ de l'enfant. Par conséquent, la probabilité que l'enfant soit malade s'il a reçu un seul allèle malade (soit le génotype $D_1 D_2$) est nulle (et évidemment $P(A \mid D_2 D_2) = 0$) donc les deux derniers termes de cette dernière équation sont égaux à zéro. Ainsi, la probabilité qu'un enfant soit malade sachant qu'il a reçu de ses deux parents les deux allèles malades est égale à 1, soit $P(A \mid D_1 D_1) = 1$. Donc, on obtient que :

$$P(A) = p^2. \qquad (2.4)$$

Ce résultat est en fait évident : pour qu'un enfant soit malade, il doit avoir reçu l'allèle malade D_1 de chacun de ses deux parents. Comme chaque parent transmet au hasard un de ses deux chromosomes à l'enfant, et que chaque chromosome est porteur de l'allèle D_1 avec probabilité p, il s'ensuit qu'un parent transmet l'allèle D_1 avec probabilité p, et par conséquent l'enfant est porteur de deux allèles D_1 avec probabilité p^2.

Il nous reste à trouver la dernière probabilité de l'équation (2.3), soit $P(A \mid T_{ij})$, qui va être calculée en conditionnant l'événement A par rapport au génotype de l'enfant, au locus malade *D*.

On utilise la décomposition :

$$P(A \mid T_{ij}) = \sum_{k,l} P(A \mid D_k D_l, T_{ij}) P(D_k D_l \mid T_{ij})$$
$$= P(A \mid D_1 D_1, T_{ij}) P(D_1 D_1 \mid T_{ij})$$
$$+ P(A \mid D_1 D_2, T_{ij}) P(D_1 D_2 \mid T_{ij})$$
$$+ P(A \mid D_2 D_2, T_{ij}) P(D_2 D_2 \mid T_{ij}).$$

Le même raisonnement sur la récessivité de la maladie nous permet de réduire l'équation ci-dessus de la façon suivante :

$$P(A \mid T_{ij}) = P(A \mid D_1 D_1, T_{ij}) P(D_1 D_1 \mid T_{ij})$$
$$= P(D_1 D_1 \mid T_{ij}).$$

Si l'enfant a le génotype $D_1 D_1$ cela signifie qu'il a hérité un allèle D_1 de chacun de ses deux parents, en d'autres mots, les parents F_1 et F_2 ont transmis chacun l'allèle D_1. En utilisant cette information, on peut réécrire l'équation précédente de la façon suivante :

$$P(A \mid T_{ij}) = P(F_2 \to D_1, F_1 \to D_1 \mid T_{ij}).$$

Comme la transmission d'un allèle par un parent est indépendante de la transmission de l'autre parent, alors on peut écrire cette dernière équation comme suit :

$$P(A \mid T_{ij}) = P(F_2 \to D_1) P(F_1 \to D_1 \mid T_{ij})$$
$$= pP(F_1 \to D_1 \mid T_{ij}). \qquad (2.5)$$

car la probabilité qu'un parent transmette l'allèle D_1 est p, ce qui justifie l'équation (2.5).

En utilisant les équations (2.4) et (2.5), l'équation (2.3) se réduit à la forme suivante :

$$
\begin{aligned}
P\left(T_{ij} \mid A\right) &= \frac{pP\left(F_1 \to D_1 \mid T_{ij}\right)P\left(T_{ij}\right)}{p^2} \\
&= \frac{P\left(F_1 \to D_1 \mid T_{ij}\right)P\left(T_{ij}\right)}{p} \\
&= \frac{P\left(F_1 \to D_1, T_{ij}\right)}{p} \\
&= \frac{P\left(F_1 \to D_1, F_1 \to M_i, F_1 = M_i M_j\right)}{p}.
\end{aligned} \tag{2.6}
$$

On doit donc expliciter la formule (2.6) pour les cas $i = j = 1$, $i = j = 2$ et $i \neq j$.

En considérant le locus du gène causant la maladie D, le phénotype du parent s'écrit $F_1 = M_i D_k / M_j D_l$, $(i,j) = (1,2)$ et $(k,l) = (1,2)$; Ceci nous permet d'écrire le numérateur de l'équation (2.6) sous la forme de la somme des quatre probabilités suivantes:

$$P\left(F_1 \to D_1, F_1 \to M_i, F_1 = M_i D_1 / M_j D_1\right) \tag{2.7}$$

$$P\left(F_1 \to D_1, F_1 \to M_i, F_1 = M_i D_1 / M_j D_2\right) \tag{2.8}$$

$$P\left(F_1 \to D_1, F_1 \to M_i, F_1 = M_i D_2 / M_j D_1\right) \tag{2.9}$$

$$P\left(F_1 \to D_1, F_1 \to M_i, F_1 = M_i D_2 / M_j D_2\right) \tag{2.10}$$

Pour chacune des équations $(2.7), (2.8), (2.9)$ et (2.10), on calcule (en utilisant les probabilités conditionnelles) la probabilité de l'intersection des événements suivants : le parent F_1 de génotype $M_i D_k / M_j D_l$ transmet à l'enfant l'allèle malade D_1 et l'allèle M_i. On peut voir que la quatrième probabilité (2.10) est nulle puisque le parent n'a pas d'allèle D_1, et donc la probabilité qu'il le transmette est nulle. Pour l'équation (2.7) on a :

$$
\begin{aligned}
&P\left(F_1 \to D_1, F_1 \to M_i, F_1 = M_iD_1 / M_jD_1\right) \\
&= P\left(F_1 \to D_1 \mid \left(F_1 \to M_i, F_1 = M_iD_1 / M_jD_1\right)\right) \times P\left(F_1 \to M_i, F_1 = M_iD_1 / M_jD_1\right) \\
&= 1 \times P\left(F_1 \to M_i, F_1 = M_iD_1 / M_jD_1\right) \\
&= P\left(F_1 \to M_i \mid F_1 = M_iD_1 / M_jD_1\right) \times P\left(F_1 = M_iD_1 / M_jD_1\right),
\end{aligned}
$$

$P\left(F_1 \to M_i \mid F_1 = M_iD_1 / M_jD_1\right)$ est égale à 1 quel que soit i et j comme on peut le voir ci-dessous :

Si $i = j$;

$$\left\{P\left(F_1 \to M_i \mid F_1 = M_iD_1 / M_iD_1\right) = P\left(F_1 \to M_j \mid F_1 = M_jD_1 / M_iD_1\right) = 1.\right.$$

Si $i \neq j$,

$$\left\{P\left(F_1 \to M_i \mid F_1 = M_iD_1 / M_jD_1\right) = 2 \times \frac{1}{2} = 1,\right.$$

car dans ce cas, le parent est hétérozygote M_iM_j par rapport au locus marqueur, donc M_i peut provenir soit de la grand-mère, soit du grand père, d'où le facteur 2 . On a alors,

$$P\left(F_1 \to D_1, F_1 \to M_i, F_1 = M_iD_1 / M_jD_1\right) = P\left(F_1 = M_iD_1 / M_jD_1\right).$$

On a posé que $P\left(M_iD_j\right) = h_{ij}$ et, comme les accouplements sont aléatoires, alors $P\left(M_iD_j / M_jD_i\right) = h_{ij} \times h_{ji}$, donc l'équation (2.7) est égale à $h_{i1}h_{j1}$ pour tout (i,j).

De la même façon, on écrit l'équation (2.8) :

$$
\begin{aligned}
&P\left(F_1 \to D_1, F_1 \to M_i, F_1 = M_iD_1 / M_jD_2\right) \\
&= P\left(F_1 \to D_1 \mid \left(F_1 \to M_i, F_1 = M_iD_1 / M_jD_2\right)\right) \times P\left(F_1 \to M_i, F_1 = M_iD_1 / M_jD_2\right)
\end{aligned}
$$

Le parent F_1 va transmettre l'allèle D_1, sachant qu'il est M_iD_1 / M_jD_2 et qu'il a déjà transmis l'allèle M_i, avec la probabilité $(1-\theta)$ puisqu'il n'est pas recombinant, d'où :

$$P(F_1 \to D_1 \mid (F_1 \to M_i, F_1 = M_iD_1 / M_jD_2)) = (1-\theta),$$

et par conséquent :

$$\begin{aligned} &P(F_1 \to D_1, F_1 \to M_i, F_1 = M_iD_1 / M_jD_2) \\ &= (1-\theta) \times P(F_1 \to M_i \mid F_1 = M_iD_1 / M_jD_2) \times P(F_1 = M_iD_1 / M_jD_2) \\ &= (1-\theta) h_{i1} h_{j2}, \quad \forall \ (i,j). \end{aligned}$$

Finalement l'équation (2.9) est égale à :

$$\begin{aligned} &P(F_1 \to D_1, F_1 \to M_i, F_1 = M_iD_2 / M_jD_1) \\ &= P(F_1 \to D_1 \mid (F_1 \to M_i, F_1 = M_iD_2 / M_jD_1)) \times P(F_1 \to M_i, F_1 = M_iD_2 / M_jD_1). \end{aligned}$$

Le parent F_1 va transmettre l'allèle D_1, sachant qu'il est M_iD_2 / M_jD_1 et qu'il a déjà transmis l'allèle M_i, avec la probabilité θ puisqu'il est recombinant, d'où :

$$P(F_1 \to D_1 \mid (F_1 \to M_i, F_1 = M_iD_2 / M_jD_1)) = \theta,$$

et en définitive, l'équation (2.9) devient :

$$\begin{aligned} &P(F_1 \to D_1, F_1 \to M_i, F_1 = M_iD_2 / M_jD_1) \\ &= \theta \times P(F_1 \to M_i \mid F_1 = M_iD_2 / M_jD_1) \times P(F_1 = M_iD_2 / M_jD_1), \\ &= \theta h_{i2} h_{j1}, \quad \forall \ (i,j). \end{aligned}$$

En sommant les trois probabilités calculées pour les équations (2.7), (2.8), (2.9), et après simplifications, l'équation (2.6) devient :

$$P(T_{ij} \mid A) = \frac{h_{i1}h_{j1} + h_{i1}h_{j2} + \theta\left(h_{i2}h_{j1} - h_{i1}h_{j2}\right)}{p}, \qquad \forall \ (i,j).$$

À partir de ces trois formules et des valeurs données en (2.1), nous pouvons retrouver les valeurs des différentes probabilités obtenues dans le tableau 2.3.

Pour $i = j = 1$, on a :

$$\begin{aligned} P(T_{11} \mid A) &= \frac{h_{11}h_{11} + h_{11}h_{12} + \theta\left(h_{12}h_{11} - h_{11}h_{12}\right)}{p} \\ &= \frac{h_{11}^2 + h_{11}h_{12}}{p} \\ &= \frac{(mp+\delta)^2 + (mp+\delta)\left(m(1-p) - \delta\right)}{p} \\ &= \frac{m(mp+\delta)}{p} \\ &= m^2 + m\delta/p. \end{aligned}$$

Pour $i = j = 2$

$$\begin{aligned} P(T_{22} \mid A) &= \frac{h_{21}h_{21} + h_{21}h_{22} + \theta\left(h_{22}h_{21} - h_{21}h_{22}\right)}{p} \\ &= \frac{h_{21}^2 + h_{21}h_{22}}{p} \\ &= \frac{\left(p(1-m) - \delta\right)^2 + \left(p(1-m) - \delta\right)\left((1-m)(1-p) + \delta\right)}{p} \\ &= \frac{(p - mp - \delta)(p - mp - \delta + 1 - m - p + mp + \delta)}{p} \\ &= (1-m)^2 - (1-m)\delta/p. \end{aligned}$$

Pour $i=1$ et $j=2$, on a :

$$P(T_{12}\mid A)=\frac{h_{11}h_{21}+h_{11}h_{22}+\theta(h_{12}h_{21}-h_{11}h_{22})}{p}$$
$$=\frac{(mp+\delta)(p-mp-\delta+1-m-p+mp+\delta)}{p}+$$
$$\frac{\theta\big((m-mp-\delta)(p-mp-\delta)-(mp+\delta)(1-m-p+mp+\delta)\big)}{p}$$
$$=(m+\delta/p)(1-m)-\theta\delta/p$$
$$=m(1-m)+(1-\theta-m)\delta/p.$$

Pour $i=2$ et $j=1$, on a :

$$P(T_{21}\mid A)=\frac{h_{21}h_{11}+h_{21}h_{12}+\theta(h_{22}h_{11}-h_{21}h_{12})}{p}$$
$$=\frac{(p-mp-\delta)(pm+\delta+m-mp-\delta)}{p}+$$
$$\frac{\theta\big((1-m-p+mp+\delta)(mp+\delta)-(p-mp-\delta)(m-mp-\delta)\big)}{p}$$
$$=m(1-m-\delta/p)+\theta\delta/p$$
$$=m(1-m)+(\theta-m)\delta/p.$$

On vient de voir à travers les calculs ci-dessus quelles valeurs dépendent du taux de recombinaison θ et par conséquent les valeurs qu'on va considérer dans le test. Dans ce qui suit, on va voir comment construire la statistique du test du TDT.

2.4 Statistique du test

Selon les hypothèses énoncées, notre test porte sur la valeur de θ. On peut voir très clairement dans le tableau 2.3, que seules deux valeurs dépendent de θ; il s'agit des valeurs correspondant aux nombres b et c du tableau 2.2 Cela signifie que seuls les parents hétérozygotes M_1M_2 devront être pris en considération dans le test. Ce qu'on va alors

comparer, c'est le nombre b de fois où l'allèle M_1 est transmis et le nombre c de fois où M_1 n'est pas transmis, et donc M_2 est transmis. En définitive, il s'agit de comparer b, le nombre de succès et c, le nombre d'échecs.

Maintenant qu'on a défini quelles données intégrer dans le test, il nous reste à trouver la forme du test recherché. Dans le tableau 2.3, en remplaçant θ par sa valeur sous l'hypothèse nulle, soit $\theta = \frac{1}{2}$, on a l'égalité entre les deux probabilités quelles que soient les valeurs de $m, \delta,$ et p. Ceci nous permet d'écrire que, sous H_0, $E(b) = E(c)$. Voyons maintenant quelles lois suivent b et c (sous H_0 en conditionnant sur le total $b+c$): b représente le nombre d'allèles M_1 transmis parmi les $(b+c)$ allèles parentaux. Chaque allèle a une probabilité $\frac{1}{2}$ d'être transmis, et une probabilité $\frac{1}{2}$ de ne pas l'être ; alors b se comporte comme une loi binomiale de paramètres $(b+c)$ et $\frac{1}{2}$, c'est-à-dire:

$$b \sim Bin\left(b+c, \frac{1}{2}\right),$$

où $\sim$ veut dire « de loi ». De la même façon, c est le nombre d'allèles M_2 parmi les $b+c$ allèles parentaux transmis et alors :

$$c \sim Bin\left(b+c, \frac{1}{2}\right).$$

Comme on l'a déjà mentionné ci-dessus, notre objectif est de comparer b (le nombre de fois où M_1 est transmis) et c (le nombre de fois où M_2 est transmis et donc M_1 n'est pas transmis). Pour faire ceci, on utilise le test de McNemar (voir Sprent et Smeeton 2001)qui se réduit à un test du χ^2. On part de la définition du test usuel du χ^2 défini comme suit :

$$\sum_i \frac{(O_i - E_i)^2}{E_i} \sim \chi^2 \qquad (2.11)$$

où O_i est le nombre observé, E_i est le nombre espéré, dans $i = 1,...,k$ catégories (ici $k = 2$).

Dans notre cas : $O_i \in \{b, c\}$ et $E_i = E(b) = E(c) = \frac{b+c}{2}$ (sous $H_0 : \theta = 1/2$). Cela donne :

$$\begin{aligned}
\sum_i \frac{(O_i - E_i)^2}{E_i} &= \frac{(b - E(b))^2}{E(b)} + \frac{(c - E(c))^2}{E(c)} \\
&= \frac{\left(b - \frac{b+c}{2}\right)^2}{\frac{b+c}{2}} + \frac{\left(c - \frac{b+c}{2}\right)^2}{\frac{b+c}{2}} \\
&= \frac{(b-c)^2}{2(b+c)} + \frac{(c-b)^2}{2(b+c)} \\
&= \frac{(b-c)^2}{b+c}.
\end{aligned}$$

En définitive le test recherché est de la forme :

$$\frac{(b-c)^2}{b+c} \sim \chi_1^2. \qquad (2.12)$$

On écrit la statistique (2.12) sous la forme χ_{td}^2, pour signifier que c'est la statistique du déséquilibre de transmission (*transmission/disequilibrium* χ^2) qui compare le nombre d'allèles M_1 transmis et le nombre d'allèles M_1 non transmis. On utilise le χ_{td}^2 pour tester la liaison et l'association entre un locus malade D et un locus marqueur M. Si le test est

significatif, cela voudrait dire que l'allèle M_1 est transmis d'une façon non aléatoire. Il est donc lié et associé à la maladie.

Certaines remarques peuvent être faites concernant ce test :

- le tableau 2.3 nous montre que le test χ^2_{td} dépend des deux paramètres θ et de δ qui sont respectivement, le taux de recombinaison et déséquilibre de liaison ; il n'est donc pratique qu'en présence d'un déséquilibre de liaison entre les deux loci ;

- on a l'égalité entre les valeurs b et c du tableau 2.2 ($b-c=0$), si on a $\theta = 1/2$ (il n'y a pas de liaison) ou $\delta = 0$ (n'y a pas d'association allélique). Donc la statistique du TDT (2.12) peut tester l'hypothèse de l'absence d'association $\delta = 0$ s'il existe déjà une liaison, c'est-à-dire $\theta < 1/2$, ou l'hypothèse nulle de l'absence de liaison $\theta = 1/2$, s'il y a une association $\delta \neq 0$;

- dans notre présentation, nous avons travaillé avec le cas où la maladie est récessive, tel que développé par Ott (1989). Des travaux plus récents ont montré que la statistique (2.12) peut être utilisée en présence d'une dominance d'un locus par rapport à l'autre.

On vient de voir tout au long de ce chapitre un test d'association et de liaison génétique qui s'applique aux caractères binaires. Dans le chapitre 4, nous allons présenter un test de même type, mais s'appliquant à des caractères mesurés sur une échelle continue. Ce test utilise les concepts de génétique quantitative qui sont définis dans le chapitre 3.

CHAPITRE III

GÉNÉTIQUE QUANTITATIVE

L'analyse génétique étudie les variations des caractères de nature discrète ou continue. Jusqu'à présent, on s'est intéressé surtout au cas discret, dans le cas des traits binaires par exemple. Dans ce cas, le nombre de modalités est fini, voir faible dans certains cas, comme le cas de la couleur de la peau, ou encore, le groupe sanguin, la présence ou l'absence de certains caractères comme le cas des taches de rousseur, etc. Ainsi, dans le cas discret toujours, il existe une relation entre le phénotype (le caractère observé), et le génotype (ce qui nous intéresse). La question de l'interprétation génétique est compliquée, du fait qu'on ne peut pas associer sans ambiguïtés un phénotype donné à un gène particulier, car plusieurs gènes ainsi que l'environnement peuvent interagir sur ces variations. Par ailleurs, la majorité des caractères observables chez tous les organismes vivants sont caractérisés par des variations qui sont de nature continue, comme le poids, la taille, le quotient intellectuel, etc.

Par contre, les fondements de l'hérédité des caractères à variation continue ne sont pas différents de ceux dont les variations sont discrètes. L'objectif de la *génétique quantitative* est de rendre compte de l'hérédité de ce type de caractères.

C'est aux environs de 1920 qu'on été jetées les bases de la génétique quantitative par les travaux de Fisher (1918), Wright (1921) et Haldane (1924-1932). Les années suivantes ont été plus marquées par le développement du sujet par ces généticiens et statisticiens et bien d'autres.

L'extension de la génétique mendélienne à la génétique quantitative peut être faite en deux étapes ; la première est d'introduire de nouveaux concepts liés aux propriétés génétiques

des populations, la seconde en introduisant des concepts liés à la transmission des mesures en se référant aux gènes causaux, facilement identifiables ; le développement de ces deux étapes constitue l'objectif de ce qui suit.

Avant d'introduire les nouveaux concepts de la génétique quantitative, nous allons définir certains termes que nous aurons à utiliser tout au long du chapitre.

3.1 Définitions

3.1.1 Définition d'un QTL

Un locus à effet quantitatif, QTL (quantitative trait locus), est un locus où la variation allélique est associée à la variation d'un caractère quantitatif.

La transmission d'un caractère quantitatif se rapporte à la transmission d'une caractéristique d'un phénotype qui varie en degrés, et qui peut être due à l'interaction de deux ou plusieurs gènes, ainsi que l'environnement dans lequel ils évoluent.

3.1.2 Maladie complexe

Une maladie est dite complexe lorsqu'elle fait référence à un phénotype qui n'indique pas d'hérédité mendélienne classique attribuée à un seul gène; il s'agit donc de maladies qui ne sont pas monogéniques. Plusieurs gènes peuvent interagir ensemble pour donner lieu à la maladie. Les maladies complexes se caractérisent aussi par des facteurs de risque environnementaux, comme le cas du cancer du sein, dont on a identifié jusqu'à présent deux gènes responsables : *BCR1* et *BCR2*, mais ces derniers ne sont pas seuls responsables, puisqu'il existe aussi des facteurs environnementaux en relation avec la maladie et qui constituent des facteurs de risque.

Après ces quelques définitions utiles, nous allons pouvoir introduire des concepts de génétique quantitative qui nous serviront de base pour les tests que nous développerons au chapitre 5.

3.2 Variations continues

Si notre objectif est, par exemple, de classer des individus selon leur poids, on ne peut pas se contenter d'une classification catégorielle, où l'on pourrait avoir sommairement les personnes grosses et minces, car il y a des degrés de poids et le fait de ranger les individus dans des catégories peut se révéler arbitraire. On rattache à ce genre de variations sur une échelle continue des caractères qui sont appelés des *caractères quantitatifs*.

Les principes génétiques qui véhiculent l'hérédité des caractères quantitatifs sont principalement les mêmes que ceux de la génétique de la population qu'on a présentés dans le chapitre 1. Mais du fait qu'on ne peut pas suivre d'une façon individuelle la ségrégation des gènes concernés, il y a eu besoin d'élaborer de nouvelles méthodes d'étude et d'introduire de nouveaux concepts, d'où la naissance d'une nouvelle branche de la génétique concernant les caractères mesurables qui est « la génétique quantitative », appelée encore « génétique biométrique » (Falconer 1974, p. 81).

Dans ce chapitre, on va d'abord montrer ce qui détermine la moyenne de la population, définir deux nouveaux concepts : *effets moyens* et *valeur de parenté*, qui sont utiles à la compréhension de la notion de variance génétique. On passera alors à l'étude de la variance et de ses composantes ainsi que la covariance entre les apparentés (c'est-à-dire des personnes ayant des liens de parenté).

3.3 Valeurs et moyennes d'une population

Les fréquences géniques et les fréquences des génotypes définies à la section 1.5 peuvent exprimer les propriétés génétiques d'une population. On va définir un nouveau concept, celui de *valeur*, qui va nous permettre d'établir une relation qui lie ces fréquences aux différences quantitatives exprimées par un caractère continu.

Cette *valeur* va s'exprimer dans la même unité de mesure dans laquelle le caractère est mesuré, par exemple le kilogramme, si le caractère à étudier est le poids. Pour continuer sur cet exemple, la *valeur* observée du poids sur un individu est ce qu'on appelle la *valeur phénotypique* de l'individu et, toutes les observations, que ce soit la moyenne, la variance ou covariance, doivent être fondées sur la mesure des *valeurs phénotypiques* (Falconer, 1996).

Par ailleurs, la valeur observée lorsqu'on mesure un caractère pour un individu est la *valeur phénotypique* de cet individu.

Pour pouvoir faire une analyse des propriétés génétiques d'une population, nous allons en premier lieu répartir la *valeur phénotypique* en composantes imputables, chacune à une cause différente.

3.4 Influence d'un locus sur le phénotype

Nous allons commencer par diviser la valeur phénotypique en deux composantes, l'une imputable à l'influence du génotype et l'autre imputable à l'influence de l'environnement.

On rappelle que le génotype est toute l'information génétique portée par un individu et, l'environnement, ce sont toutes les causes non génétiques qui influencent la valeur phénotypique. On associe alors deux valeurs à ces deux composantes :

- la valeur génotypique G,
- la valeur de l'écart dû à l'environnement E.

On représente alors la valeur phénotypique Y suivant le modèle suivant :

$$Y = G + E .$$

Lorsque l'on doit traiter des effets conjoints de plusieurs gènes, Fisher (1918) supposa que les effets de tous ces gènes étaient additifs (Vetta et Courgeau, 2003), c'est-à-dire qu'il n'y a ni covariance, ni interaction entre eux. D'où la dénomination de modèle additif. Il définit ensuite l'environnement comme un ensemble de « causes arbitraires externes qui sont indépendantes de l'hérédité ». Cela implique que l'environnement est indépendant des gènes et aléatoire avec un effet moyen nul et une variance inconnue. En d'autres termes, la moyenne des écarts dus à l'environnement pour la population entière est supposée égale à zéro. En terme statistique, E est une variable aléatoire d'espérance nulle, de sorte que la moyenne phénotypique soit égale à la moyenne génotypique. Le terme *moyenne de la population* se réfère alors de la même façon aux valeurs génotypique ou phénotypique.

On définit Y comme la valeur attendue d'un phénotype (étant donné un génotype) résultant de l'expression commune des gènes qui influencent le caractère. La valeur génotypique considérée ici représente un gène, ou plusieurs gènes si ceux-ci sont assez proches pour être considérés comme une entité (le cas général de plusieurs gènes ne sera pas traité dans notre étude). On s'intéresse à la contribution d'un gène sur les valeurs génotypiques. Pour cela, considérons un marqueur (L) qui ségrégatif pour deux allèles B_1 et B_2 dans une population panmictique[4] de taille infinie (Falconer 1974)

. On prend par convention :

- a la valeur génotypique de l'un des homozygotes (B_1B_1 ou B_2B_2) ,
- $-a$ la valeur génotypique de l'autre homozygote,
- d la valeur génotypique de l'hétérozygote B_1B_2 , qui mesure la dominance d'un gène par rapport à l'autre.

Par hypothèse, c'est l'allèle B_1 qui a pour effet d'augmenter la valeur génotypique. Le schéma ci-dessous représente une échelle de valeurs génotypiques, où l'origine (le point zéro) est située à mi-distance entre les valeurs des deux homozygotes.

Figure 3.1 Valeurs arbitraires attribuées aux génotypes

La valeur d de l'hétérozygote dépend du degré de dominance ce qui donne :

- s'il n'y a pas dominance, $d = 0$;
- si B_1 domine B_2, $d > 0$;

[4] C'est une population où les accouplements sont aléatoires.

- si B_2 domine B_1, $d < 0$;
- si la dominance est complète, $d = a$ ou $d = -a$;
- s'il y a surdominance, d peut être plus grand que $+a$ ou plus petit que $-a$.

3.4.1 Moyenne d'une population

Il s'agit de voir comment les fréquences géniques influencent la moyenne d'un caractère d'une population dans son ensemble. À cette fin, nous allons continuer sur l'exemple ci-dessus et définir, en plus des valeurs génotypiques, des fréquences. Soit p et q les fréquences géniques respectives de B_1 et B_2. L'ensemble de ces données est alors résumé dans le tableau 3.1 qui suit :

Tableau 3.1 Exemple de valeurs génotypiques et leur valeur moyenne dans la population

Génotype	Fréquence	Valeur	Fréquence × *Valeur*
B_1B_1	p^2	$+a$	p^2a
B_1B_2	$2pq$	d	$2pqd$
B_2B_2	q^2	$-a$	$-q^2a$
		Total	$a(p-q)+2dpq$

Par définition, la valeur moyenne de la population totale est obtenue en multipliant chaque valeur du génotype par sa fréquence et en calculant la somme totale M pour l'ensemble des génotypes. On obtient :

$$\begin{aligned} M &= p^2a + 2pqd - q^2a \\ &= a(p^2 - q^2) + 2pqd \\ &= a(p-q)(p+q) + 2pqd \\ &= a(p-q) + 2dpq, \end{aligned}$$

car $p+q=1$. Ainsi la valeur moyenne de la population est:

$$M = a(p-q)+2dpq \qquad (3.1)$$

La contribution d'un locus à la moyenne de la population est l'addition de deux termes : le premier terme, $a(p-q)$, attribué aux homozygotes et le second terme, $2dpq$, attribué aux hétérozygotes. S'il n'y a pas de dominance d'un gène sur l'autre $(d=0)$, la valeur moyenne sera égale à :

$$M = a(p-q).$$

La moyenne M représente, pour le caractère étudié, en même temps la valeur de la moyenne génotypique et la valeur de la moyenne phénotypique de la population.

3.4.2 Effet moyen

L'établissement des propriétés d'une population en relation avec sa structure familiale se fait par l'intermédiaire de la transmission d'une *valeur* d'un parent à sa progéniture, ce qui ne peut pas se faire avec les valeurs génotypiques seules, car les parents transmettent à leurs enfants des *gènes* et non pas des *génotypes* qui eux changent à chaque génération.

Une nouvelle mesure de la valeur se rapportant non pas aux génotypes, mais aux *gènes*, s'impose. Il s'agit de *la valeur additive* appelée aussi valeur d'élevage, qui est une valeur associée aux gènes que porte un individu et qu'il transmet à sa descendance. Cette grandeur est ce qu'on appelle *l'effet moyen*. C'est cet effet moyen des gènes des parents qui détermine la valeur génotypique moyenne de la progéniture. L'effet moyen dépend des valeurs a et d définies ci-dessus et des fréquences génotypiques. Pour un allèle particulier, il est défini comme l'écart moyen entre la moyenne de la population (voir formule 3.1) et la moyenne des individus qui reçoivent cet allèle d'un parent, l'autre parent ayant transmis un allèle tiré au hasard dans la population. On peut l'expliquer en supposant qu'un certain nombre de gamètes portant tous un allèle B_1 s'unissent avec des gamètes tirés au hasard dans

la population. Alors, l'écart entre la moyenne de la population et la moyenne des génotypes ainsi produits est égale à l'effet moyen de l'allèle B_i (Falconer, 1974, p. 92).

3.5 Relation entre la valeur phénotypique et la valeur génotypique

On quantifie la relation qui existe entre les valeurs génotypique et phénotypique afin de traiter un des objectifs de la génétique quantitative qu'on développera plus loin, à savoir la répartition de la variance phénotypique en composantes génétique et non génétique.

L'approche standard est de considérer la valeur phénotypique Y comme étant la somme du total des effets de tous les loci sur le caractère (valeur génotypique) et des effets environnementaux. Nous avons déjà vu au paragraphe 3.4 que le modèle portant sur la relation entre la valeur phénotypique et la valeur génotypique est de la forme :

$$Y = G + E, \qquad (3.2)$$

où Y est la valeur phénotypique, G est la valeur génotypique ou génétique qui traduit l'effet moyen d'un génotype sur un caractère et E est l'écart dû à l'environnement qui inclut toutes les valeurs non génétiques. Nous avons aussi vu l'hypothèse l'additivité des effets des gènes, et de l'indépendance entre G et E.

Les deux composantes de la valeur phénotypique Y peuvent à leur tour être subdivisées comme dans ce qui suit.

3.5.1 Composantes de la valeur génotypique

Trois composantes contribuent à la valeur génotypique G :

- Valeur génétique additive A,
- Dominance D,
- Interaction.

Nous allons définir chacune de ces composantes comme suit :

Définition 3.5.1.1 : Valeur génétique additive

La valeur génétique additive d'un individu est obtenue en sommant tous les effets moyens des gènes qu'il possède. Cette somme sera faite pour toutes les paires d'allèles à chaque locus et pour tous les loci.

Définition 3.5.1.2 : Dominance

Dans le cas où il y a un seul locus, l'écart de dominance D est l'écart entre la valeur génotypique G et la valeur génétique additive A.

$$G = A + D.$$

Celui-ci résulte de la dominance qui existe entre les allèles présents sur le locus. En l'absence de dominance, la valeur génotypique G est égale à la valeur génétique additive A.

Définition 3.5.1.3 : Interaction

Comme on vient de le mentionner ci-dessus, s'il y a juste un seul locus, la valeur génotypique ne comprend que la valeur génotypique additive et l'écart de dominance. Par contre, quand il y a plus d'un locus, la valeur génotypique peut comporter un autre écart, qui lui est dû à une combinaison non additive de gènes. En suivant l'exemple de Falconer (1974), considérons deux loci A et B. Pour un individu, soit :

- G_A la valeur génotypique pour le locus A ;
- G_B la valeur génotypique pour le locus B ;
- G la valeur génotypique pour les deux loci.

On pose alors :

$$G = G_A + G_B + I_{AB},$$

où I_{AB} est l'écart à la combinaison additive des ces valeurs génotypiques.

Dans une population, si l'interaction I pour certaines combinaisons de gènes situés sur différents locus n'est pas nulle, on dit alors que les gènes sont en interaction. On définit alors I comme l'écart d'interaction ou d'épistasie. Les loci peuvent interagir par paires ou par triplets, ce qui conduit à des interactions de plusieurs sortes. Vu la complexité de ces termes, toutes ces interactions sont rassemblées sous le terme I dans la valeur définitive du génotype. Pour l'ensemble des loci, on suppose le modèle suivant :

$$G = A + D + I. \qquad (3.3)$$

Dans le cas de plusieurs loci, A représente la somme des valeurs additives attribuées à l'ensemble des loci pris séparément et D la somme des écarts de dominance. Dans le cas où l'interaction est nulle, les gènes situés sur ce groupe de loci sont alors additifs.

3.5.2 Composantes de la valeur environnementale E

La plupart des caractères quantitatifs ne sont pas simplement déterminés par des facteurs génétiques ; les facteurs environnementaux jouent eux aussi un rôle dans la détermination des caractères. Plusieurs subdivisions du facteur environnemental sont faites par différents auteurs. Lynch et Walsh $(1998, p.108)$ proposent de diviser le facteur environnemental en effets environnementaux, généraux et spécifiques.

- Effets environnementaux généraux S : comprennent tous les facteurs influents partagés par les individus ou groupes d'individus d'une même famille. Il s'agit généralement des effets maternels, comme l'état général de santé de la mère qui influe différemment sa progéniture.

- Effets environnementaux spécifiques e : il s'agit en partie des résidus entre, la valeur attendue du phénotype et le génotype, et des effets généraux d'environnement qui influencent différemment chaque individu. Ces effets ne sont pas mesurables par l'expérimentateur.

La valeur phénotypique (3.2) peut alors s'écrire sous la forme suivante :

$$Y = G + S + e\,. \qquad (3.4)$$

En remplaçant dans (3.4) la valeur de G telle que définie dans par la formule (3.3), on obtient :

$$Y = A + D + I + S + e\,. \qquad (3.5)$$

On vient de voir comment se répartit la valeur phénotypique en différentes composantes, il s'agit maintenant de voir comment expliquer les différences quantitatives de cette valeur entre les individus. Ceci se fera par l'étude des variances, qui fera l'objet de la section suivante.

3.6 Étude des variances

L'étude génétique d'un caractère quantitatif est basée sur l'étude de sa variation. L'idée de base de l'étude de cette variation est le fait qu'on peut la subdiviser en composantes attribuées chacune à une cause différente. L'importance relative de chacune de ces composantes détermine les propriétés génétiques de la population, en particulier, le degré de ressemblance entre apparentés dont le concept sera développé par la suite. L'étude de ces variations se fera par l'étude des variances.

3.6.1 Variance de la valeur phénotypique

On a vu que l'expression d'un caractère quantitatif est influencée par des facteurs génétiques et environnementaux. Une des principales questions qui est posée est de savoir : quelle part de variation de ces caractères dans la population est due aux causes génétiques et quelle part est due à des causes environnementales.

De la même façon que nous avons décomposé la valeur phénotypique d'un caractère en deux composantes, nous décomposons aussi sa variance phénotypique en deux grandes composantes :

- composante génotypique G ,

- composante résiduelle E .

Les deux composantes de la valeur phénotypique étant supposées indépendantes entres elles (formule 3.2), la variance phénotypique est alors la somme des variances de ses composantes, soit la somme des variances génotypique et environnementale.

$$\sigma_Y^2 = \sigma_G^2 + \sigma_E^2 . \qquad (3.6)$$

3.6.2 Variance de la valeur génotypique

Comme nous avons subdivisé la valeur génotypique en composantes additive, de dominance et d'interaction (formule 3.3), on peut de la même façon répartir la variance de la valeur génotypique, en variances de ces mêmes composantes, soit :

$$\underset{(\text{génotypique})}{\sigma_G^2} = \underset{(\text{additive})}{\sigma_A^2} + \underset{(\text{dominance})}{\sigma_D^2} + \underset{(\text{interaction})}{\sigma_I^2} .$$

Parmi ces composantes, seule la variance additive revêt une importance capitale, car il s'agit de la principale cause de ressemblance entre les apparentés (point qu'on développera par la suite); c'est aussi elle qui détermine les composantes génétiques de la population. Elle a aussi pour avantage que c'est la seule valeur qui peut être estimée à partir des observations faites sur la population. En pratique, on décompose la variance totale en variance génétique additive et, un reste, qui est égal à la variance génétique non additive et celle due au milieu.

3.7 Modèle des moindres carrés pour les effets génétiques : calcul de variance

Dans cette section, on considère une procédure statistique pour la définition des effets additifs, de dominance et d'interaction des gènes. Celle-ci nous sera utile dans le calcul de la covariance génétique entre des individus ayant des liens de parenté. On prend le cas de deux loci ayant les allèles i, j au locus 1, et les allèles k, l au locus 2, qu'on pourrait aisément généraliser à plus de deux loci. On suppose aussi qu'on a des accouplements aléatoires.

Soit $G_{i\ldots}$ l'espérance conditionnelle du phénotype des individus ayant l'allèle i au 1^{er} locus sans considération de l'autre allèle relatif au second locus. Soit α_i l'effet additif de l'allèle i. Celui-ci est défini comme étant égal à l'écart entre la moyenne du génotype des

individus de la population ayant l'allèle i au 1^{er} locus $G_{i..}$ et la moyenne phénotypique de la population μ_G.

$$\alpha_i = G_{i...} - \mu_G \ . \qquad (3.7)$$

Les autres effets additifs $\alpha_j, \alpha_k, \alpha_l$ se calculent de la même façon. À chaque locus, la valeur moyenne de tous les effets est nulle.

L'effet de dominance δ_{ij} est défini en considérant $G_{ij..}$ l'espérance conditionnelle du phénotype des individus ayant les allèles i et j au 1^{er} locus, sans tenir compte du génotype au second locus. Il est obtenu ainsi :

$$\delta_{ij} = G_{ij..} - \mu_G - \alpha_i - \alpha_j \ . \qquad (3.8)$$

On procède au même raisonnement pour la définition des effets d'interaction, notés $(\alpha\alpha)$ ou $(\alpha\delta)$.

Ainsi, soit $G_{i.k.}$ la moyenne du phénotype des individus ayant l'allèle i au locus 1, et le l'allèle k au locus 2, sans considération des autres allèles. Le $ik^{ième}$ effet *additif*×*additif* s'écrit :

$$(\alpha\alpha)_{ik} = G_{i.k.} - \mu_G - \alpha_i - \alpha_k \ ; \qquad (3.9)$$

$(\alpha\alpha)_{ik}$ représente la différence entre la moyenne conditionnelle $G_{i.k}$ et les valeurs de la moyenne de la population μ_G et les effets additifs α_i et α_j.

L'effet *additif*×*dominance* mesure l'interaction entre un allèle situé sur un locus avec un génotype particulier situé sur l'autre locus. Il est défini comme la différence entre la moyenne conditionnelle $G_{i.kl}$ et les valeurs basées sur les trois effets additifs, un effet de dominance, et deux effets *additif*×*additif* impliquant les allèles, donc on a :

$$(\alpha\delta)_{ikl} = G_{i.kl} - \mu_G - \alpha_i - \alpha_k - \alpha_l - \delta_{kl} - (\alpha\alpha)_{ik} - (\alpha\alpha)_{il}. \qquad (3.10)$$

Enfin, il nous reste à définir l'effet *dominance* × *dominance* comme étant le terme :

$$\begin{aligned}(\delta\delta)_{ijkl} = G_{ijkl} &- \mu_G - \alpha_i - \alpha_j - \alpha_k - \alpha_l - \delta_{ij} - \delta_{kl} \\ &- (\alpha\alpha)_{ik} - (\alpha\alpha)_{il} - (\alpha\alpha)_{jk} - (\alpha\alpha)_{jl} \\ &- (\alpha\delta)_{ikl} - (\alpha\delta)_{jkl} - (\alpha\delta)_{ijk} - (\alpha\delta)_{ijl}\end{aligned} \qquad (3.11)$$

La composition de la valeur totale du génotype en une série d'effets peut être reconstituée en additionnant chacun des effets décrits ci-dessus, en commençant par les effets additifs (voir Lynch et Walsh, 1998). On a alors la formule qui donne la valeur d'un génotype :

$$\begin{aligned}G_{ijkl\ldots} = \mu_G &+ \left[\alpha_i + \alpha_j + \alpha_k + \alpha_l\right] + \left[\delta_{ij} + \delta_{kl}\right] \\ &+ \left[(\alpha\alpha)_{ik} + (\alpha\alpha)_{il} + (\alpha\alpha)_{jk} + (\alpha\alpha)_{jl}\right] \\ &+ \left[(\alpha\delta)_{ikl} + (\alpha\delta)_{jkl} + (\alpha\delta)_{ikj} + (\alpha\delta)_{ijl}\right] + (\delta\delta)_{ijkl} + R.\end{aligned} \qquad (3.12)$$

Le reste qui n'est pas défini désigné par R signifie qu'il peut y avoir des termes analogues provenant des interactions entre plus de deux loci (d'où l'écriture $G_{ijkl\ldots}$).

L'équation (3.12) semble peu pratique, et à première vue, il n'est pas possible de trouver la valeur du génotype sans commettre des erreurs. En supposant que les accouplements sont aléatoires, et la ségrégation des loci est indépendante, il n'y a alors pas de relation statistique (dépendance) entre les effets des gènes à l'intérieur des loci et entre les loci ; ce qui simplifie le calcul de la variance génétique, puisque celle-ci devient égale à la somme de tous les effets en annulant les termes relatifs aux covariances, soit :

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_{AA}^2 + \sigma_{AD}^2 + \sigma_{DD}^2 + \ldots,$$

où :

$$\sigma_A^2 = \sigma^2(\alpha_i) + \sigma^2(\alpha_j) + \sigma^2(\alpha_k) + \sigma^2(\alpha_l),$$
$$\sigma_D^2 = \sigma^2(\delta_{ij}) + \sigma^2(\delta_{kl}),$$
$$\sigma_{AA}^2 = \sigma^2\left[(\alpha\alpha)_{ik}\right] + \sigma^2\left[(\alpha\alpha)_{il}\right] + \sigma^2\left[(\alpha\alpha)_{jk}\right] + \sigma^2\left[(\alpha\alpha)_{jl}\right],$$
$$\sigma_{AD}^2 = \sigma^2\left[(\alpha\delta)_{ikl}\right] + \sigma^2\left[(\alpha\delta)_{jkl}\right] + \sigma^2\left[(\alpha\delta)_{ijk}\right] + \sigma^2\left[(\alpha\delta)_{ijl}\right],$$
$$\sigma_{DD}^2 = \sigma^2\left[(\delta\delta)_{ijkl}\right].$$

En termes génétiques, les différentes variances signifient :

- Variance génétique : σ_A^2
- Interaction entre les deux valeurs additives
  variance *additive* × *additive* : σ_{AA}^2
- Interaction entre valeur additive d'un
  locus et écart de dominance de l'autre : σ_{AD}^2
- Interaction entre les deux écarts de dominance
  variance *dominance* × *dominance* : σ_{DD}^2

3.8 Ressemblance entre apparentés

C'est par l'analyse statistique de la ressemblance entre apparentés que vers la fin du $XIX^{ème}$ siècle, les biométriciens ont commencé à poser les bases de l'hérédité des caractères quantitatifs.

Tout un chacun sait que les apparentés tendent à se ressembler et cette ressemblance est d'autant plus grande que la relation de parenté est proche. Le degré de ressemblance varie aussi avec le caractère ; il est élevé pour certains caractères, par exemple la couleur de la peau, et faible pour d'autres, par exemple la couleur des yeux, etc. On peut alors conclure que le degré de ressemblance dépend du lien de parenté et du caractère étudié.

Le degré de ressemblance entre apparentés est une des propriétés de la population qu'on peut facilement observer. On va dans ce qui suit introduire le concept de mesure de parenté.

3.8.1 Mesure de parenté

La mesure de parenté est basée sur le concept appelé en anglais « Identity by descent » et noté généralement *IBD*, qu'on pourrait traduire par « identique par la descendance ». On dit que deux individus partagent un allèle identique par la descendance *(IBD)*, si cet allèle provient du même ancêtre.

Définition 3.8.1.1 : Coefficient de parenté

En supposant qu'on a tiré aléatoirement deux gènes à partir de chacun de deux individus s et t, le *coefficient de parenté (coefficient of kinship)* est égal à la probabilité Θ_{st} que ces 2 gènes soient *IBD*.

Dans le cas d'une cellule familiale, deux individus partagent un allèle *IBD*, s'il y a un parent qui a transmis *le même allèle à ses deux enfants*. En d'autres mots, cet allèle provient du même grand-parent. Dans la figure 3.2, le père transmet le même allèle B_1 (trait continu) à ses deux enfants, et l'allèle B_1 est alors appelé *identique par la descendance*.

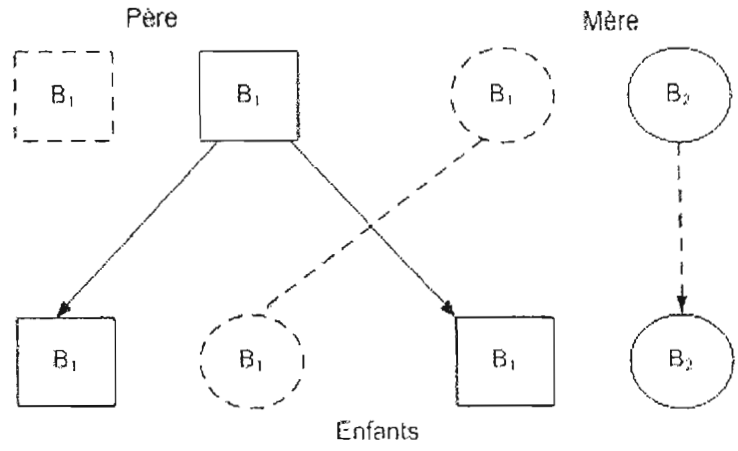

Figure 3.2 Transmission des gènes de deux parents à leur deux enfants

Considérons une famille où les deux parents n'ont pas de relation de parenté, donc sont indépendants. Supposons qu'ils ont deux enfants et que chaque enfant hérite un allèle de chacun de ses parents. La probabilité est de 0,5 pour que ces deux allèles soient *IBD*, c'est-à-dire qu'ils proviennent du grand père paternel ou de la grand-mère paternelle. De la même façon, cette probabilité est de 0,5 pour que ces deux allèles reçus proviennent du grand-père maternel, ou de la grand-mère maternelle. En effet, fixons par exemple la mère qui a deux allèles provenant des ses parents, disons l'allèle B_1 hérité de sa mère et l'allèle B_2 hérité de son père. Si on considère la transmission de ces allèles à ses deux enfants, on peut alors en dénombrer quatre possibilités qu'on peut résumer dans le tableau 3.2. Donc on voit immédiatement que dans la moitié des cas, les enfants reçoivent les allèles du même grand parent (B_1 de la grand-mère, et B_2 du grand père).

Soit N le nombre total d'allèles partagés identiques par la descendance *(IBD)* pour les deux enfants, N peut alors être vu comme le nombre de succès obtenu lors de deux expériences, où le succès représente ici la transmission par un parent de deux allèles *IBD* à une paire d'enfants.

Tableau 3.2 Transmission des allèles B_1 et B_2 d'une mère à ses enfants

Enfant 1	Enfant 2
B_1	B_1
B_1	B_2
B_2	B_1
B_2	B_2

Ainsi, N pourra être assimilé à une loi binomiale de paramètres $n=2$, où n est le nombre d'allèles qui peuvent être partagés et $p=0,5$. Ceci nous permet de calculer les probabilités des événements suivants :

$$p(N=0)=\binom{2}{0}0,5^{0}0,5^{2}=0,25$$

$$p(N=1)=\binom{2}{1}0,5^{1}0,5^{1}=0,5$$

$$p(N=2)=\binom{2}{2}0,5^{2}0,5^{0}=0,25$$

Une autre façon nous permettant d'interpréter le coefficient de parenté, et qui va nous être utile par la suite, est de considérer la distribution du nombre d'allèles partagés *IBD* pour une paire d'apparentés pour un locus donné. Afin de calculer le coefficient de parenté Θ_{st}, on considère N_{st}, le nombre d'allèles partagés *IBD* entre deux individus apparentés s et t, où N_{st} peut prendre les valeurs 0, 1 ou 2.

Soit F l'événement que deux allèles tirés au hasard, un parmi chacun des individus soient *IBD* sur ce locus, donc la probabilité de F sera égale au coefficient Θ_{st} : $P(F)=\Theta_{st}$. Pour les calculs, nous supposons que les individus appartiennent à la même fratrie.

Considérons la probabilité conditionnelle de l'événement F sachant que $\{N_{st}=i\}$, soit $P_F(i)$, $i=0,1,2$. On a en tout quatre paires d'allèles possibles formées en tirant au hasard un allèle de chaque enfant. Ces allèles sont par exemple : B_1, B_2, B_3, B_4. Dans le tableau suivant, on représente un cas typique de génotype des enfants classés selon le nombre d'allèles partagés. Nous supposons dans ce cas typique que les parents sont hétérozygotes et ne partagent aucun allèle à ce locus.

	Enfant 1	Enfant 2	Nombre d'allèles partagés IDB
Génotypes	B_1 / B_2	B_3 / B_4	0
	B_1 / B_2	B_1 / B_3	1
	B_1 / B_2	B_1 / B_2	2

Fixons une ligne, alors si on tire un allèle dans chaque colonne du tableau, on obtient selon la ligne, qu'il y a 0, 1 ou 2 choix d'allèles *IBD* (sur les quatre choix possibles). Donc on obtient :

$$P_F(0) = 0,\ P_F(1) = 1/4,\ P_F(2) = 1/2.$$

On peut remarquer que $P(F \mid N_{st}) = N_{st} / 4$, ce qui nous permet d'écrire :

$$\begin{aligned} P(F) &= \sum_i P(F \mid N_{st}) \times P(N_{st} = i) \\ &= \sum_i \frac{N_{st}}{4} \times P(N_{st} = i) \\ &= \frac{1}{4} \sum_i N_{st} \times P(N_{st} = i) \\ &= \frac{1}{4} E(N_{st}) \\ &= \frac{1}{2} E\left(\frac{N_{st}}{2}\right). \end{aligned}$$

À partir de ce résultat, le coefficient de parenté peut s'écrire sous la forme :

$$\Theta_{st} = \frac{1}{2} E\left(\frac{N_{st}}{2}\right) \Leftrightarrow 2\Theta_{st} = E\left(\frac{N_{st}}{2}\right).$$

On voit que $\frac{N_{st}}{2}$ est la proportion d'allèles partagés *IBD* par les individus s et t. Donc $E\left(\frac{N_{st}}{2}\right)$ est l'espérance de cette proportion, qu'on note π_{st}, d'où :

$$2\Theta_{st} = \pi_{st}. \qquad (3.13)$$

Cette formule se généralise à toute paire d'individus (s,t) qui sont apparentés (par exemple, cousins, oncles, neveux, etc.)

Dans la littérature, il existe plusieurs méthodes pour l'estimation des π_{st} (voir références dans Sham et al. 2000), mais celles-ci ne feront pas l'objet de notre étude.

3.8.2 Covariance phénotypique entre les apparentés

On a vu que la variance phénotypique d'un caractère quantitatif peut être répartie en deux composantes : génétique et environnementale. La question est de savoir comment estimer ces variances. L'idée vient de Fisher (1918) et Wright $(1921a)$ qui ont noté que les différentes sources génétique et environnementale de la variance contribuent différemment à la ressemblance entre les apparentés (voir Lynch et Walsh, 1998, p. 131).

Soit deux individus s et t ayant une relation de parenté quelconque (frères et sœurs, demi-frères et demi-sœurs, parents et enfants, etc.). En considérant qu'il n'y a pas d'interaction entre les génotypes et l'environnement, et en utilisant l'équation (3.4), les valeurs phénotypiques pour les deux individus s'écrivent :

$$y^s = G^s + S^s + e^s \quad \text{et} \quad y^t = G^t + S^t + e^t.$$

On peut écrire alors la covariance phénotypique entre les deux individus apparentés de la façon suivante :

$$\begin{aligned} Cov\left(y^s, y^t\right) &= Cov\left[\left(G^s + S^s + e^s\right), \left(G^t + S^t + e^t\right)\right] \\ &= Cov\left(G^s, G^t\right) + Cov\left(S^s, G^t\right) + Cov\left(e^s, G^t\right) \\ &+ Cov\left(G^s, S^t\right) + Cov\left(S^s, S^t\right) + Cov\left(e^s, S^t\right) \\ &+ Cov\left(G^s, e^t\right) + Cov\left(S^s, e^t\right) + Cov\left(e^s, e^t\right) \\ &= Cov\left(G^s, G^t\right) + Cov\left(S^s, G^t\right) + Cov\left(G^s, S^t\right) \\ &+ Cov\left(S^s, S^t\right) + Cov\left(e^s, e^t\right). \end{aligned} \tag{3.14}$$

On arrive à cette simplification, en utilisant le fait que les effets d'environnement spécifiques (e) sont des résidus aléatoires qui ne sont pas corrélés avec les autres facteurs, et qui ne le sont pas entre les individus, c'est-à-dire $Cov(e^s, G^t) = Cov(G^s, e^t) = 0$, . $Cov(e^s, S^t) = Cov(S^s, e^t) = 0$, et $Cov\left(e^s, e^t\right) = 0$,

Les termes $Cov\left(S^s, G^t\right)$ et $Cov\left(G^s, S^t\right)$ représentent chacun, la covariance entre la valeur du génotype d'un membre de la paire d'apparentés et la valeur de l'effet de l'environnement général de l'autre membre. Dans ce cas, on fait l'hypothèse que les covariances entre les valeurs génotypiques et environnementales sont nulles ; en d'autres mots, on considère que :

$$Cov\left(S^s, G^t\right) = Cov\left(G^s, S^t\right) = 0.$$

Finalement la formule (3.14) de la covariance entre deux apparentés s'écrit comme suit :

$$Cov\left(y^s, y^t\right) = Cov\left(G^s, G^t\right) + Cov\left(S^s, S^t\right). \tag{3.15}$$

La covariance génétique entre deux apparentés revêt une grande importance, car elle résulte du fait que les deux apparentés aient hérité une copie du même gène.

3.8.2.1 La covariance génétique entre les apparentés

Notre première démarche est de décomposer la covariance totale génétique G entre apparentés formule (3.15) en composantes principales décrivant les effets des différents gènes. Elle mérite donc un développement approprié. Pour la calculer, on considère une paire d'individus s et t tirés d'une population constituée de paires de personnes ayant la même relation de parenté.

D'après la formule (3.11), les valeurs génotypiques pour les 2 individus s'écrivent :

$$\begin{aligned} G_{ijkl\ldots}(s) = \mu_G &+ \left[\alpha_i^s + \alpha_j^s + \alpha_k^s + \alpha_l^s + \ldots\right] + \left[\delta_{ij}^s + \delta_{kl}^s + \ldots\right] \\ &+ \left[(\alpha\alpha)_{ik}^s + (\alpha\alpha)_{il}^s + (\alpha\alpha)_{jk}^s + (\alpha\alpha)_{jl}^s + \ldots\right] \\ &+ \left[(\alpha\delta)_{ikl}^s + (\alpha\delta)_{jkl}^s + (\alpha\delta)_{kij}^s + (\alpha\delta)_{lij}^s + \ldots\right] + (\delta\delta)_{ijkl}^s + \ldots \end{aligned}$$

$$\begin{aligned} G_{ijkl\ldots}(t) = \mu_G &+ \left[\alpha_i^t + \alpha_j^t + \alpha_k^t + \alpha_l^t + \ldots\right] + \left[\delta_{ij}^t + \delta_{kl}^t + \ldots\right] \\ &+ \left[(\alpha\alpha)_{ik}^t + (\alpha\alpha)_{il}^t + (\alpha\alpha)_{jk}^t + (\alpha\alpha)_{jl}^t + \ldots\right] \\ &+ \left[(\alpha\delta)_{ikl}^t + (\alpha\delta)_{jkl}^t + (\alpha\delta)_{kij}^t + (\alpha\delta)_{lij}^t + \ldots\right] + (\delta\delta)_{ijkl}^t + \ldots \end{aligned}$$

Les indices i, j et k, l représentent respectivement les gènes au premier et second locus, et le reste correspond au cas où l'on aurait plus de deux loci.

Afin de calculer la covariance génétique entre les deux individus, on fait référence aux travaux de Fisher (1918) qui a montré qu'étant donné que les différents effets ne sont pas corrélés pour chaque individu, ils ne le sont pas aussi entre les individus (Lynch et Walsh, 1998, p. 143). Donc l'équation de la covariance entre les deux individus s et t ne prend en compte que la covariance entre les mêmes types d'effets pour les deux individus.

$$\begin{aligned} Cov(G_s, G_t) = &Cov(A_s, A_t) + Cov(D_s, D_t) + Cov((AA)_s, (AA)_t) \\ &+ Cov((AD)_s, (AD)_t) + Cov((DD)_s, (DD)_t) + \ldots \end{aligned}$$

Dans notre cas, comme nous allons considérer que l'effet de dominance ainsi que les termes de la covariance causée par l'effet *additive*× *additive* (AA) sont nuls, la covariance génotypique se réduit alors à l'équation suivante :

$$Cov(G_s, G_t) = Cov(A_s, A_t).$$

Il reste à exprimer cette équation en termes de composantes de variance et de coefficients de parenté. On prend le cas où l'on a un seul locus 1, ce qui peut se généraliser au cas à deux loci.

$$Cov[G_s, G_t] = E[G_{ij}(s)G_{ij}(s,t)] - E[G_{ij}(s)]E[G_{ij}(t)].$$

On l'a vu dans le développement de la formule (3.12) que tous les effets moyens sont nuls et particulièrement $E(G_{ij}(s)) = E(G_{ij}(t)) = 0$. Il reste alors à calculer le terme $E(G_{ij}(s)G_{ij}(t))$:

$$\begin{aligned} E[G_{ij}(s)G_{ij}(t)] &= E[(\alpha_i^s + \alpha_j^s)(\alpha_i^t + \alpha_j^t)] \\ &= E(\alpha_i^s \alpha_i^t) + E(\alpha_i^s \alpha_j^t) + E(\alpha_j^s \alpha_i^t) + E(\alpha_j^s \alpha_j^t). \end{aligned}$$

On fait le développement pour le premier terme $E(\alpha_i^s \alpha_i^t)$; la généralisation pour les autres termes restants peut se développer de façon semblable. Rappelons que les deux gènes considérés peuvent être identiques par la descendance avec une probabilité Θ_{st} . Dans ce cas, $E(\alpha_i^s \alpha_i^t) = E[(\alpha_i)^2]$, ce qui est égal à *la moitié de la variance génétique attribuée au locus1*. Par contre, si les 2 gènes ne sont pas identiques par la descendance, ils doivent être distribués indépendamment, alors

$$E(\alpha_i^s \alpha_i^t) = [E(\alpha_i)]^2 = 0.$$

En appliquant le même raisonnement aux trois autres termes restants, la covariance génétique additive devient :

$$Cov(A_s, A_t) = 4\Theta_{st} E\left[(\alpha_i)^2\right] = 4 \times \frac{1}{2} \Theta_{st} \sigma_A^2 = 2\Theta_{st} \sigma_A^2.$$

En utilisant la formule (3.13), on obtient :

$$Cov(A_s, A_t) = \pi_{st} \sigma_A^2. \qquad (3.16)$$

Les notions de la décomposition de la valeur phénotypique en composante génétique et environnementale permettent alors l'étude des variations de chaque facteur selon un modèle basé sur la décomposition de la variance dans lequel sont introduites les informations sur les marqueurs génétiques. Nous allons utiliser dans le chapitre suivant un modèle de décomposition de variance de la valeur phénotypique en utilisant les notions introduites dans ce chapitre (décomposition de la valeur phénotypique, la notion d'apparenté, etc.) pour établir un test d'association et de déséquilibre de liaison.

CHAPITRE IV

LE TEST DU QTDT (QUANTITATIVE TRANSMISSION DISEQUILIBRIUM TEST)

Il existe plusieurs méthodes d'analyse de liaison génétique pour la localisation de gènes causant la maladie, permettant ainsi de délimiter la région du chromosome responsable ; mais celles-ci sont peu performantes quand elles sont appliquées aux maladies complexes (voir section 3.1.2) (Abecasis et al. 2000). Néanmoins, il existe des tests qui donnent de meilleurs résultats, dont le plus connu est le TDT, que l'on a présenté dans le chapitre 2. Dans le cas où les parents, plutôt que des frères et sœurs, sont utilisés comme témoins, ces tests de déséquilibre de transmission nécessitent un nombre élevé de familles afin d'assurer le même niveau de puissance, ce qui n'est pas toujours aisé à constituer.

Aussi, pour la plupart des maladies complexes, les scores obtenus à partir de phénotypes quantitatifs contiennent plus d'information que ce que fournissent les caractères binaires. Des auteurs comme Allison et Rabinowitz (1997) ont introduit un test de liaison génétique pour les caractères quantitatifs basé sur des familles. Comme le TDT, ces tests utilisent les génotypes des parents pour construire des paires de témoins. Ces tests sont des tests de déséquilibre de liaison dans des familles ayant juste un enfant. Fulker et al. (1999) ont décrit un test similaire pour des données sur des paires d'enfants où l'on n'utilise pas les génotypes des parents. Ce test est basé sur une méthode combinée pour tester l'association et la liaison génétique dans le cas d'un caractère quantitatif.

L'objectif de ce chapitre est de présenter la méthode de Fulker ainsi qu'un test encore plus général qui peut, séparément ou simultanément, modéliser l'association et la liaison génétique pour des caractères quantitatifs (Abecasis et al., 2000)

4.1 Présentation du modèle

Soit un marqueur locus candidat diallélique ayant les allèles B_1 et B_2, où :

- l'allèle B_1 apparaît avec une fréquence p dans la population,
- l'allèle B_2 apparaît avec une fréquence $q = 1 - p$ dans la population.

Trois génotypes peuvent être formés à ce locus, B_1B_1, B_1B_2, B_2B_2. Soit un ensemble de k familles, où la famille i a n_i enfants, $i=1,..,k$. On a alors un total de N enfants avec $N = \sum_{i=1}^{k} n_i$. Les individus sont identifiés selon leur appartenance à une famille, donc on parle de l'individu j de la famille i.

Dans ce chapitre, nous dénotons par A_{ij}^* la valeur de l'effet additif du génotype candidat du $j^{ème}$ enfant $j = 1,...,n_i$ dans la $i^{ème}$ famille. On considère l'effet de dominance comme nul (la valeur génotypique est égale à l'effet additif du génotype). On peut ainsi affecter aux effets additifs A_{ij}^* les valeurs génétiques arbitraires suivantes $-a$, 0 ou a correspondant respectivement aux génotypes suivants B_2B_2, B_1B_2, B_1B_1. En appliquant les fréquences génotypiques aux scores A_{ij}^*, on obtient le tableau 4.1 suivant.

Tableau 4.1 Tableau des fréquences génotypiques

Scores A_{ij}^*	$-a$	0	a
Fréquences (probabilités)	q^2	$2pq$	p^2

Soit le score phénotypique de l'individu j dans la famille i, c'est à dire la mesure du caractère d'intérêt. Selon Falconer (1996), le score des phénotypes pour la maladie étudiée est défini par l'effet d'un gène majeur, un effet relatif à la famille (qui inclut les effets du partage de l'environnement par les membres de la fratrie et la variance résiduelle

additionnelle), ainsi qu'un effet environnemental résiduel propre à chaque membre de la fratrie.

Cela se résume dans l'écriture du modèle déjà présenté au chapitre 3 équation (3.5) en considérant l'effet de dominance et d'interaction nuls :

$$Y = A^* + S + e,$$

où :

- Y représente la valeur phénotypique,
- A^* représente la valeur génétique additive (l'effet de tous les gènes impliqués),
- S représente l'effet environnemental partagé par les apparentés (l'effet familial),
- e représente l'effet environnemental non partagé.

Le modèle basé sur les composantes de la variance de la valeur phénotypique pour des frères et sœurs j et k appartenant à une famille i spécifie que, conditionnellement aux vecteurs des effets additifs A_{ij}^* et aux matrices de proportions pour les allèles *IBD* (Π_{ij}), le vecteur des valeurs phénotypiques du caractère quantitatif suit une loi normale multivariée de moyenne μ_i et de matrice de variance covariance égale à Ω_i, où les paramètres μ_i et Ω_i seront calculés aux sections 4.1.2 et 4.1.3.

4.1.1 Paramètres du modèle

Pour définir les paramètres de la loi normale du vecteur des valeurs phénotypiques, on considère :

- $Y_i = \left(y_{i1}, y_{i2}, \ldots . y_{in_i}\right)'$ le vecteur des valeurs phénotypiques du caractère quantitatif pour la famille i, c'est-à-dire la mesure du caractère d'intérêt où y_{ij} représente la valeur phénotypique du caractère pour l'individu j dans la famille i, $j = 1, \ldots, n_i$, et $i = 1, \ldots, k$;
- $A_i^* = \left(A_{i1}^*, A_{i2}^*, \ldots, A_{in_i}^*\right)$ le vecteur des valeurs génétiques additives;

- $[\Pi]_{i,jl} = \pi_{i,jl}$, la matrice des proportions d'allèles IBD partagés par les frères et sœurs j et l dans la famille *i*.

Les $\pi_{i,jl}$ sont des probabilités qui sont généralement estimées à partir des données sur des génotypes marqueurs à l'aide des moyennes, ou en utilisant une des méthodes déjà existantes et qui sont citées par Sham et al. (2000), Curtis et Sham (1994), Fulker et al. (1995), Kruglyak et Lander (1995), Almasy et Blangero (1998).

4.1.2 Calcul de l'espérance et nouvelle paramétrisation

En supposant que la moyenne de tous les effets d'environnement est nulle, la moyenne phénotypique μ_{ij} pour un enfant j dans une la famille i, ne dépend que de la valeur génotypique, et s'écrit comme suit :

$$\mu_{ij} = \mu + A_{ij}^{*} = \mu + aA_{ij},$$

où a représente l'amplitude de l'effet additif au locus candidat, A_{ij} est l'effet additif codé en $-1,0,1$, et μ est le vecteur contenant la moyenne phénotypique pour chaque sujet supposée égale pour toute la population. On a introduit ainsi une nouvelle paramétrisation par laquelle l'effet additif A_{ij} est codé en $-1, 0, 1$. Cette paramétrisation met en évidence la valeur a sur laquelle porte notre inférence.

4.1.3 Calcul des éléments de la matrice de variance covariance

Pour le calcul de la matrice de variance covariance $\Omega_{i,jl}$ pour deux enfants j et l appartenant à la même famille *i*, on se base sur le modèle général de la valeur phénotypique qu'on a présenté au chapitre 3.

$$Y_{ij} = \mu + A_{ij}^{*} + S_i + e_{ij} \qquad i = 1,\ldots,k \text{ et } j = 1,\ldots,n_i, \tag{4.1}$$

où i et j représentent respectivement la famille et l'enfant. Soit Y_{ij} et Y_{il} les valeurs phénotypiques de deux enfants j et l dans la famille *i*. Alors, on a :

$$\begin{aligned} Cov\left(Y_{ij},Y_{il}\right) &= Cov\left(\mu + A_{ij}^{*} + S_{i} + e_{ij}\,,\mu + A_{il}^{*} + S_{i} + e_{il}\right) \\ &= Cov\left(A_{ij}^{*},A_{il}^{*}\right) + Cov\left(A_{ij}^{*},S_{i}\right) + Cov\left(A_{ij}^{*},e_{il}\right) \\ &\quad + Cov\left(S_{i},A_{il}^{*}\right) + Cov\left(S_{i},S_{i}\right) \; + Cov\left(S_{i},e_{il}\right) \\ &\quad + Cov\left(e_{ij},A_{il}^{*}\right) + Cov\left(e_{ij},S_{i}\right) + Cov\left(e_{ij},e_{il}\right). \end{aligned} \qquad (4.2)$$

Comme e_{ij} et e_{il} sont des variables aléatoires indépendantes, elles ne sont pas corrélées entre elles ; de plus, elles ne le sont pas avec les autres effets, donc :

$$Cov\left(A_{ij}^{*},e_{il}\right) = Cov\left(A_{il}^{*},e_{ij}\right) = Cov\left(S_{i},e_{il}\right) = Cov\left(S_{i},e_{ij}\right) = Cov\left(e_{ij},e_{jl}\right) = 0\,.$$

La valeur génétique additive est aussi une variable aléatoire qui ne dépend que du gène lui même, ce qui nous amène à écrire :

$$Cov\left(A_{il}^{*},S_{i}\right) = 0\,.$$

L'équation (4.2) s'écrit alors :

$$\begin{aligned} Cov\left(Y_{ij},Y_{il}\right) &= Cov\left(A_{ij}^{*},A_{il}^{*}\right) + Cov\left(S_{i},S_{i}\right) \\ &= Cov\left(A_{ij}^{*},A_{il}^{*}\right) + Var\left(S_{i}\right). \end{aligned} \qquad (4.3)$$

Or, on a vu au chapitre 3 que la covariance entre deux individus apparentés est égale à :

$$Cov\left(A_{ij}^{*},A_{il}^{*}\right) = \pi_{jl}\sigma_{A}^{2} \quad \text{, où } \pi_{jl} = 2\Theta_{st}\,. \qquad (4.4)$$

En définitive (4.2) devient

$$Cov\left(Y_{ij},Y_{il}\right) = \pi_{jl}\sigma_{A^{*}}^{2} + \sigma_{S}^{2}\,.$$

Pour obtenir les éléments de la diagonale de Ω_i, donc les variances, il suffit de mettre dans (4.2) l'individu $l = j$, ce qui revient à calculer la variance phénotypique d'un individu j, d'où

$$\begin{aligned} Var\left(Y_{ij}\right) = Cov\left(Y_{ij}, Y_{ij}\right) &= Cov\left(A_{ij}^{*} + S_i + e_{ij}, A_{ij}^{*} + S_i + e_{ij}\right) \\ &= Cov\left(A_{ij}^{*}, A_{ij}^{*}\right) + Cov\left(A_{ij}^{*}, S_i\right) + Cov\left(A_{ij}^{*}, e_{ij}\right) \\ &+ \; Cov\left(S_i, A_{ij}^{*}\right) + Cov\left(S_i, S_i\right) + Cov\left(S_i, e_{ij}\right) \\ &+ Cov\left(e_{ij}, A_{ij}^{*}\right) + Cov\left(e_{ij}, S_i\right) + Cov\left(e_{ij}, e_{ij}\right). \end{aligned}$$

Après la suppression des termes nuls, on obtient le résultat suivant :

$$\begin{aligned} Var\left(Y_{ij}\right) = Cov\left(Y_{ij}, Y_{ij}\right) &= Cov\left(A_{ij}^{*}, A_{ij}^{*}\right) + Cov\left(S_i, S_i\right) + Cov\left(e_{ij}, e_{ij}\right) \\ &= Var\left(A_i^{*}\right) + Var\left(S_i\right) + Var\left(e_{ij}\right) \\ &= \sigma_{A^{*}}^2 + \sigma_S^2 + \sigma_e^2 . \end{aligned}$$

En conclusion, les éléments de la matrice de variance covariance Ω_i pour deux enfants j et l appartenant à une famille i s'écrivent comme suit :

$$\Omega_{i,jl} = \begin{cases} \sigma_a^2 + \sigma_s^2 + \sigma_e^2 & \text{si} \quad j = l \\ \pi_{jl}\sigma_a^2 + \sigma_s^2 & \text{si} \quad j \neq l \ , \end{cases} \qquad (4.5)$$

où :

- σ_a^2 représente la composante additive de la variance du phénotype,
- σ_s^2, σ_e^2 représentent respectivement les composantes des variances du milieu partagé et non partagé,
- π_{jl} représente la proportion d'allèles identiques par la descendance (*IBD*) partagés par les frères et soeurs.

4.2 Estimation des paramètres du modèle

Comme on vient de le mentionner, le vecteur des scores phénotypiques $Y_i = \left(y_{i1}, y_{i2}, \ldots y_{in_i}\right)'$ de la famille i, $i = 1,\ldots,k$, suit une loi normale multivariée de moyenne μ_i et de matrice de variance covariance Ω_i, c'est-à-dire :

$$Y_i \sim N\left(\mu_i, \Omega_i\right)$$

La densité de Y_i s'écrit comme suit :

$$f_Y\left(y_i\right) = \left(2\pi\right)^{-n_i/2} \left|\Omega_i\right|^{-1/2} \exp\left(-\frac{1}{2}\left(y_i - \mu_i\right)' \Omega_i^{-1}\left(y_i - \mu_i\right)\right) \qquad (4.6)$$

où n_i est la taille de la famille i, y_i est le vecteur des scores phénotypiques observés pour les enfants de la famille i, et μ_i est la moyenne phénotypique dans la famille i. La fonction de vraisemblance est de la forme :

$$L(\theta \mid y) = \prod_i f\left(y_i \mid \theta\right) = \prod_i \left(2\pi\right)^{-n_i/2} \left|\Omega_i\right|^{-1/2} \exp\left(-\frac{1}{2}\left(y_i - \mu_i\right)' \Omega_i^{-1}\left(y_i - \mu_i\right)\right) \qquad (4.7)$$

où $\underset{\sim}{\theta} = \left(a, \mu_i, \sigma_a^2, \sigma_s^2, \sigma_e^2\right)$ est l'ensemble des paramètres. Les estimateurs du maximum de vraisemblance s'obtiennent par la maximisation du logarithme de la fonction de vraisemblance, c'est-à-dire en maximisant :

$$\log L\left(\underset{\sim}{\theta} \mid y\right) = \frac{\left(-\sum_i^k n_i\right)}{2} \log\left(2\pi\right) - \frac{1}{2}\sum_i \left[\log\left|\Omega_i\right| + \left(y_i - \mu_i\right)' \Omega_i^{-1}\left(y_i - \mu_i\right)\right].$$

En remplaçant μ_i par sa valeur dans le modèle, on obtient :

$$\log L\left(\underset{\sim}{\theta}/y\right)=\frac{-\sum_{i}^{k} n_i}{2}\log\left(2\pi\right)-\frac{1}{2}\sum_{i}\left[\log\left|\Omega_i\right|+\left(y_i-\mu-aA_i\right)'\Omega_i^{-1}\left(y_i-\mu-aA_i\right)\right].$$

Cette fonction devient plus complexe si le nombre et la taille des familles augmentent, et la maximisation nécessite alors l'utilisation de méthodes numériques appropriées.

4.3 Test général d'association : principes de base

Une fois obtenues les estimations des paramètres par la méthode du maximum de vraisemblance, on peut établir différents tests parmi lesquels le test d'association.

Pour tester l'association entre le locus marqueur et le phénotype, on s'intéresse à la moyenne, et plus particulièrement à la valeur qui mesure l'amplitude de l'effet génétique additif, le paramètre a. Notre modèle est de la forme :

$$y_{ij}=\mu+aA_{ij}+\varepsilon_{ij}, \tag{4.8}$$

que l'on peut écrire sous forme de moyenne :

$$\mu_{ij}=\mu+aA_{ij}. \tag{4.9}$$

Ainsi, le test d'association porte sur les hypothèses suivantes :

$$\begin{cases} H_0 : a=0, \\ H_1 : a\neq 0. \end{cases}$$

À partir de (4.8), on voit que, si H_0 est vraie, il n'y a pas d'association entre le locus marqueur et le phénotype et $\mu_{ij}\equiv\mu$ pour tout (i,j), $i=1,...,k$ et $j=1,...,n_k$.

Pour faire ce type de test, on utilise la méthode du test du rapport du maximum de vraisemblance $\lambda(X)$ ayant la forme suivante :

$$\lambda(X) = \frac{\max_{\theta \in \Theta_0} L(\theta \mid x)}{\max_{\theta \in \Theta} L(\theta \mid x)}$$
$$= \frac{L\left(\hat{\hat{\theta}} \mid x\right)}{L\left(\hat{\theta} \mid x\right)} = \frac{L_0(x)}{L_1(x)},$$

où $\hat{\hat{\theta}}$ dénote l'estimateur de maximum de vraisemblance sous l'hypothèse nulle (modèle restreint) et $\hat{\theta}$ dénote l'estimateur de maximum de vraisemblance sous le modèle complet (aucune restriction imposée au paramètre). La région de rejet du test est de la forme $\{X : \lambda(X) \leq c\}$, où c est une constante à déterminer (elle dépend du seuil du test). Dans le cas d'un échantillon de grande taille (et c'est ce qu'on va supposer), on utilise le résultat asymptotique suivant :

Sous H_0 , et quand $n \to \infty$, on a :

$$-2 \log \lambda(X) \to \chi_1^2,$$

où :

$$-2 \log \lambda(X) = -2[\log L_0 - \log L_1] = 2[\log L_1 - \log L_0].$$

En définitive, on a :

$$2[\log L_1 - \log L_0] \to \chi_1^2. \qquad (4.10)$$

Par conséquent, dans notre cas, on prend la valeur a quelconque (modèle complet) ou $a = 0$ (modèle restreint), d'où :

$$2\left[\log L_1 - \log L_0\right] = 2\left[\left(\frac{\left(-\sum_i^k n_i \log(2\pi)\right)}{2} - \frac{1}{2}\sum_i\left(\log\left|\hat{\Omega}_i\right| + (y_i - \hat{\mu} - \hat{a}A_i)' \hat{\Omega}_i^{-1}(y_i - \hat{\mu} - \hat{a}A_i)\right)\right) - \left(\frac{\left(-\sum_i^k n_i \log(2\pi)\right)}{2} - \frac{1}{2}\sum_i\left(\log\left|\hat{\tilde{\Omega}}_i\right| + \left(y_i - \hat{\tilde{\mu}}\right)' \hat{\tilde{\Omega}}_i^{-1}\left(y_i - \hat{\tilde{\mu}}\right)\right)\right)\right]$$

$$= +\sum_i\left(\log\left|\hat{\tilde{\Omega}}_i\right| + \left(y_i - \hat{\tilde{\mu}}\right)' \hat{\tilde{\Omega}}_i^{-1}\left(y_i - \hat{\tilde{\mu}}\right)\right)$$
$$-\sum_i\left(\log\left|\hat{\Omega}_i\right| + (y_i - \hat{\mu} - \hat{a}A_i)' \hat{\Omega}_i^{-1}(y_i - \hat{\mu} - \hat{a}A_i)\right)$$

Dans ce cas aussi, on fera appel aux méthodes numériques appropriées pour trouver la région de rejet. Cependant, avec ce type de test, un problème se pose : l'association susceptible d'être trouvée peut être due à une stratification de la population, ce qui nous amènerait alors à tirer une fausse conclusion. Pour surmonter ce problème, Fulker et al. (1999) ont proposé un modèle qui permet de détecter, et même d'isoler, cette fausse association en décomposant le score du génotype en composantes orthogonales inter-fratrie et intra-fratrie.

4.4 Modèle de décomposition orthogonale des scores des génotypes

Dans ce qui suit, on va présenter la décomposition telle que décrite dans le modèle proposé par Fulker et al. (1999). Les éléments de leur modèle sont les suivants : soit un locus candidat biallélique ayant des allèles arbitraires B_1 et B_2, et soit $-a, 0, a$ respectivement les effets des trois génotypes, B_2B_2, B_2B_1, B_1B_1.

Dans le cas où l'on considère uniquement des paires d'enfants de mêmes parents, on peut avoir neuf combinaisons de paires de génotypes qui, sous l'hypothèse d'accouplements

aléatoires, sont caractérisés par des moyennes et des différences. Ces calculs sont présentés dans le tableau 4.2 et se basent sur la paramétrisation (4.9).

Ainsi pour une paire d'enfants d'une certaine famille i, le vecteur μ_i des moyennes peut se modéliser à partir des éléments suivants : une moyenne globale μ, une moyenne de la paire, Δ_m, ainsi que de la différence de la paire, Δ_d. On obtient pour les enfants 1 et 2 :

$$\mu_{i1} = \mu + \Delta_m + (\Delta_d/2) \quad \text{et} \quad \mu_{i2} = \mu + \Delta_m - (\Delta_d/2),$$

et on retrouve le tableau 4.2.

Pour tester le paramètre d'association a, le test du χ^2 à 1 degré de liberté a été proposé. Néanmoins, ce test peut donner lieu à une fausse association qui n'est due qu'à une stratification de la population (par exemple une population répartie en groupes de différentes ethnies).

Puisque la stratification influence la moyenne à l'intérieur des paires d'enfants, et pas leur différence, Fulker et al. 1999 ont proposé une méthode qui permet de détecter cette fausse association. Il s'agit d'une méthode dans laquelle les effets de l'association vont être répartis en composantes entre les familles b (*between*) et à l'intérieur de chaque famille w (within). C'est une méthode dans laquelle l'effet a est posé différent dans la moyenne et dans la différence des paires. Donc l'effet additif génotypique A_{ij} de l'enfant j de la famille i va se décomposer en deux composantes orthogonales telles que :

$$A_{ij} = \overline{A}_i + (A_{ij} - \overline{A}_i), \qquad (4.11)$$

où $\overline{A}_i$ représente l'effet additif entre familles et $A_{ij} - \overline{A}_i$ représente l'effet additif à l'intérieur des familles. On introduit les valeurs a_b et a_w qui sont respectivement les amplitudes des effets additifs au locus candidat sur la moyenne et sur la différence des paires.

Tableau 4.2 Moyennes et différences de l'effet additif a pour un locus biallélique à effet additif

Génotype		Effet additif			
Enfant1	Enfant2	Enfant1	Enfant2	Moyenne	Différ/2
B_1B_1	B_1B_1	a	a	a	0
B_1B_1	B_1B_2	a	0	$a/2$	$a/2$
B_1B_1	B_2B_2	a	$-a$	0	a
B_1B_2	B_1B_1	0	a	$a/2$	$-a/2$
B_1B_2	B_1B_2	0	0	0	0
B_1B_2	B_2B_2	0	$-a$	$-a/2$	$a/2$
B_2B_2	B_1B_1	$-a$	a	0	$-a$
B_2B_2	B_1B_2	$-a$	0	$-a/2$	$-a/2$
B_2B_2	B_2B_2	$-a$	$-a$	$-a$	0

On redéfinit ainsi μ_{ij} la moyenne phénotypique pour un paire d'enfants (enfant 1 et enfant 2) appartenant à la famille i qui va pouvoir s'écrire sous la forme suivante :

$$\begin{aligned}\mu_{i1} &= \mu + a_b\bar{A}_i + a_w\left(A_{i1} - \bar{A}_i\right) \\ &= \mu + a_b\bar{A}_i + a_w\left(A_{i1} - \frac{A_{i1}+A_{i2}}{2}\right) \\ &= \mu + a_b\bar{A}_i + a_w\left(\frac{A_{i1}-A_{i2}}{2}\right),\end{aligned} \qquad (4.12)$$

ainsi que, de façon similaire :

$$\mu_{i2} = \mu + a_b\bar{A}_i + a_w\left(A_{i2} - \bar{A}_i\right) = \mu + a_b\bar{A}_i - a_w\left(\frac{A_{i1}-A_{i2}}{2}\right) \qquad (4.13)$$

En remplaçant $\overline{A}_i$ par Δ_{mi} pour signifier la moyenne de l'effet additif au locus candidat pour les deux paires d'enfants et $A_{i1} - A_{i2}$ par Δ_{di} pour signifier la différence de l'effet additif au locus candidat pour les deux paires, les deux moyennes ci-dessus deviennent :

$$\mu_{i1} = \mu + a_b \Delta_{m_i} + a_w \frac{\Delta_{d_i}}{2},$$

$$\mu_{i1} = \mu + a_b \Delta_{m_i} - a_w \frac{\Delta_{d_i}}{2}.$$

En utilisant ces notations, on peut réécrire le tableau 4.2 sous la forme du tableau 4.3.

Afin de voir comment ce tableau est obtenu à partir de (4.12) et (4.13), considérons l'exemple de la deuxième ligne du tableau 4.2, où l'enfant 1 est de génotype B_1B_1 avec un effet additif égal à $+1$, et l'enfant 2 est de génotype B_1B_2 avec un effet additif de 0. Cela donne :

$$\overline{A}_i = \frac{1}{2}, \quad \text{et} \quad \begin{cases} A_{i1} - \overline{A}_i = 1 - \frac{1}{2} = + \frac{1}{2} \\ A_{i2} - \overline{A}_i = 0 - \frac{1}{2} = - \frac{1}{2} \end{cases}.$$

En notant les amplitudes des effets intra et inter familles respectivement par a_b et a_w, on réécrit les valeurs phénotypiques pour chaque enfant (voir tableau 4.3) comme suit :

$$\mu_{i1} = \mu + \frac{a_b}{2} + \frac{a_w}{2} \quad \text{et} \quad \mu_{i2} = \mu + \frac{a_b}{2} - \frac{a_w}{2}.$$

À partir de cette écriture, on peut faire des tests séparés sur a_b et a_w et éviter de conclure à une fausse association. Ces tests sont décrits à la section 4.6

Tableau 4.3 Décomposition orthogonale de l'effet additif en composantes des inter(b) et intra (w) fraterie

Génotype		Effet additif			
Enfant1	Enfant2	Enfant1	Enfant2	Moyenne	Différ/2
B_1B_1	B_1B_1	a_b	a_b	a_b	0
B_1B_1	B_1B_2	$(a_b/2)+(a_w/2)$	$(a_b/2)-(a_w/2)$	$a_b/2$	$a_w/2$
B_1B_1	B_2B_2	a_w	$-a_w$	0	a_w
B_1B_2	B_1B_1	$(a_b/2)-(a_w/2)$	$(a_b/2)+(a_w/2)$	$a_b/2$	$-a_w/2$
B_1B_2	B_1B_2	0	0	0	0
B_1B_2	B_2B_2	$(-a_b/2)+(a_w/2)$	$(-a_b/2)-(a_w/2)$	$-a_b/2$	$a_w/2$
B_2B_2	B_1B_1	$-a_w$	a_w	0	$-a_w$
B_2B_2	B_1B_2	$(-a_b/2)-(a_w/2)$	$(-a_b/2)+(a_w/2)$	$-a_b/2$	$-a_w/2$
B_2B_2	B_2B_2	$-a_b$	$-a_b$	$-a_b$	0

4.5 Modèles d'Abecasis

Le modèle de Fulker et al. (1999) s'applique au cas d'une paire de frères et sœurs et permet d'éviter de conclure à une fausse association par la décomposition des scores des génotypes A_{ij} en composantes orthogonales. Abecasis et al. (2000) ont généralisé cette décomposition de deux façons : (*i*) pour n'importe quelle taille de familles, et (*ii*) au cas des trios (les deux parents et un seul enfant malade).

Ainsi, pour chaque individu j appartenant à la famille i, l'effet additif A_{ij} de son génotype par rapport au locus candidat peut s'écrire sous la forme de deux termes : le premier terme qui est une composante inter-fratrie A_i^b et qui est représentée par l'effet moyen au sein de la fratrie, et le second terme qui est la composante intra-fratrie représentée par l'écart entre l'effet de l'enfant et l'effet moyen au sein de la fratrie, $A_{ij} - A_i^b$; cela s'écrit comme suit, (voir Sham et al., 2000) :

$$A_{ij} = A_i^b + \left(A_{ij} - A_i^b\right), \tag{4.14}$$

où $i = 1,...,k$, $j = 1,...,n_i$, et $N = \sum_{i=1}^{k} n_i$.

4.5.1 Cas général avec plusieurs enfants malades par famille

Avec plusieurs enfants par famille, la composante inter-fratrie A_i^b se calcule comme suit :

$$A_i^b = \frac{\sum_{i=1}^{n_i} A_{ij}}{n_i} = \bar{A}_i. \tag{4.15}$$

En utilisant l'équation (4.14), on peut réécrire la valeur phénotypique Y_{ij} de l'individu j de la famille i comme suit :

$$Y_{ij} = \mu + a_b A_i^b + a_w \left(A_{ij} - A_i^b\right) + \varepsilon_{ij}, \tag{4.16}$$

où a_b et a_w représentent respectivement les amplitudes des effets additifs au locus candidat entres les composantes inter-fratrie et intra-fratrie, μ représente la moyenne phénotypique au niveau de la population, donc supposée égale pour chaque individu, et ε_{ij} est un terme aléatoire.

On a donc un système à N équations et trois paramètres inconnus μ, a_b, a_w à estimer. Comme l'estimation de la constante μ ne nous intéresse pas, on peut la supprimer du système en calculant, pour chaque individu, l'écart de sa valeur phénotypique par rapport à la moyenne phénotypique globale, qui est égale à :

$$\bar{\bar{Y}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}}{N} = a_b \frac{\sum_{i=1}^{k} n_i A_i^b}{N} + a_w \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(A_{ij} - A_i^b\right)}{N} + \bar{\bar{\varepsilon}}.$$

On obtient pour un individu j de la famille i :

$$\begin{aligned} Y_{ij} - \overline{\overline{Y}} &= a_b A_i^b + a_w \left(A_{ij} - A_i^b \right) + \varepsilon_{ij} \\ &\quad - \left[a_b \frac{\sum_{i=1}^{k} n_i A_i^b}{N} + a_w \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(A_{ij} - A_i^b \right)}{N} + \overline{\overline{\varepsilon}} \right] \\ &= a_b \left(A_i^b - \overline{A}_i^b \right) + a_w \left(A_{ij} - A_i^b \right) + \left(\varepsilon_{ij} - \overline{\overline{\varepsilon}} \right). \end{aligned} \tag{4.17}$$

En effet, dans ce cas $A_i^b = \overline{A_i}$, d'où

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(A_{ij} - A_i^b \right) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(A_{ij} - \overline{A_i} \right) = \sum_{i=1}^{k} n_i \overline{A_i} - \sum_{i}^{k} n_i \overline{A_i} = 0.$$

En écriture matricielle on a le modèle :

$$\begin{bmatrix} Y_{11} - \overline{\overline{Y}} \\ \cdots \\ Y_{1n_1} - \overline{\overline{Y}} \\ \cdots \\ \cdots \\ Y_{kn_k} - \overline{\overline{Y}} \end{bmatrix} = \begin{bmatrix} A_1^b - \overline{A}_1^b & A_{11} - A_1^b \\ \cdots & \cdots \\ A_1^b - \overline{A}_1^b & A_{1n_1} - A_1^b \\ \cdots & \cdots \\ \cdots & \cdots \\ A_k^b - \overline{A}_k^b & A_{kn_k} - A_k^b \end{bmatrix} \begin{bmatrix} a_b \\ a_w \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \cdots \\ \varepsilon_{1n_1} \\ \cdots \\ \cdots \\ \varepsilon_{kn_k} \end{bmatrix}$$

On pourrait poser :

- $Z_{ij} = Y_{ij} - \overline{\overline{Y}}$,
- $X_{i1} = A_i^b - \overline{A}_i^b$,
- $X_{i2} = A_{ij} - A_i^b$,

et on a ainsi l'espérance conditionnelle :

$$E\left[Z_{ij} \mid X\right] = E\left[\left(Y_{ij} - \bar{\bar{Y}}\right) \mid X\right] = a_b\left(A_i^b - \bar{A}_i^b\right) + a_w\left(A_{ij} - A_i^b\right).$$

Afin d'estimer (a_b, a_w) par les moindres carrés, on utilise l'estimateur par le modèle de régression multiple $\hat{\beta}$ qui est égal à :

$$\hat{\beta} = \left(X'X\right)^{-1} X'Z,$$

où les matrices $X'X$ et $X'Z$ sont décrites dans ce qui suit. Notons par $\hat{\beta}_b$ et $\hat{\beta}_w$ les deux composantes du vecteur $\hat{\beta}$, soit $\hat{\beta} = \begin{pmatrix} \hat{\beta}_b \\ \hat{\beta}_w \end{pmatrix}$, et calculons :

$$X'X = \begin{bmatrix} A_1^b - \bar{A}_1^b & \cdots & A_1^b - \bar{A}_1^b & \cdots & \cdots & A_k^b - \bar{A}_k^b \\ A_{11} - A_1^b & \cdots & A_{1n_1} - A_1^b & \cdots & \cdots & A_{kn_k} - A_k^b \end{bmatrix} \times \begin{bmatrix} A_1^b - \bar{A}_1^b & A_{11} - A_1^b \\ \cdots & \cdots \\ A_1^b - \bar{A}_1^b & A_{1n_1} - A_1^b \\ \cdots & \cdots \\ \cdots & \cdots \\ A_k^b - \bar{A}_k^b & A_{kn_k} - A_k^b \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{k} n_i \left(A_i^b - \bar{A}_i^b\right)^2 & \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(A_i^b - \bar{A}_i^b\right)\left(A_{ij} - A_i^b\right) \\ \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(A_{ij} - A_i^b\right)\left(A_i^b - \bar{A}_i^b\right) & \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(A_{ij} - A_i^b\right)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{k} n_i \left(A_i^b - \bar{A}_1^b\right)^2 & 0 \\ 0 & \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(A_{ij} - A_i^b\right)^2 \end{bmatrix}$$

Les termes mixtes sont nuls car dans ce cas $A_i^b = \bar{A}_i$ et $\sum_{j=1}^{n_i} A_{ij} = n_i \bar{A}_i = n_i A_i^b$, d'où

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(A_{ij}-A_i^b\right)\left(A_i^b-\bar{A}_i^b\right) &= \sum_{i=1}^{k}\left(A_i^b-\bar{A}_i^b\right)\sum_{j=1}^{n_i}\left(A_{ij}-A_i^b\right)\\ &= \sum_{i=1}^{k}\left(A_i^b-A_i^b\right)\left(\sum_{j=1}^{n_i}A_{ij}-n_iA_i^b\right)\\ &= \sum_{i=1}^{k}\left(A_i^b-A_i^b\right)\left(n_iA_i^b-n_iA_i^b\right)\\ &= 0.\end{aligned}$$

Le calcul de $X'Z$ se fait de la même façon, et on obtient :

$$X'Z = \begin{bmatrix} A_1^b-\bar{A}_1^b & \dots & A_1^b-\bar{A}_1^b & \dots & \dots & A_k^b-\bar{A}_k^b \\ A_{11}-A_1^b & \dots & A_{1n_1}-A_1^b & \dots & \dots & A_{kn_k}-A_k^b \end{bmatrix} \times \begin{bmatrix} Y_{11}-\bar{\bar{Y}} \\ \dots \\ Y_{1n_1}-\bar{\bar{Y}} \\ \dots \\ \dots \\ Y_{kn_k}-\bar{\bar{Y}} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{k}\left(A_i^b-\bar{A}_i^b\right)\left(\sum_{j=1}^{n_i}Y_{ij}-n_i\bar{\bar{Y}}\right) \\ \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(A_{ij}-A_i^b\right)\left(Y_{ij}-\bar{\bar{Y}}\right) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{k}\sum_{j=1}^{n_i}n_i\left(A_i^b-\bar{A}_i^b\right)\left(Y_{ij}-\bar{\bar{Y}}\right) \\ \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(A_{ij}-A_i^b\right)\left(Y_{ij}-\bar{\bar{Y}}\right) \end{bmatrix}$$

On peut montrer que les estimateurs sont sans biais. En effet, d'après les équations normales de la régression multiple, on a que :

$$(X'X)\hat{\beta} = X'Z\,.$$

Si on prend l'espérance de chaque membre de l'équation ci-dessus, on obtient :

$$E\left((X'X)\hat{\beta} \mid X\right) = E(X'Z \mid X)$$
$$\Leftrightarrow (X'X)E\left(\hat{\beta} \mid X\right) = X'E(Z \mid X).$$

Par contre, dans ce cas, $E\left(\hat{\beta} \mid X\right) = E\left(\hat{\beta}\right)$.

On a

$$(X'X)E\left(\hat{\beta} \mid X\right) = \begin{bmatrix} E(\hat{a}_b)\sum_{i=1}^{k} n_i \left(A_i^b - \bar{A}_i^b\right)^2 \\ E(\hat{a}_w)\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(A_{ij} - A_i^b\right)^2 \end{bmatrix} \qquad (4.18)$$

ainsi que

$$X'E(Z \mid X) = X'E\left[\left(Y_{ij} - \bar{\bar{Y}}\right) \mid X\right]$$

$$= \begin{bmatrix} A_1^b - \bar{A}_1^b & \ldots & A_1^b - \bar{A}_1^b & \ldots & \ldots & A_k^b - \bar{A}_k^b \\ A_{11} - A_1^b & \ldots & A_{1n_1} - A_1^b & \ldots & \ldots & A_{kn_k} - A_k^b \end{bmatrix} \times \begin{bmatrix} a_b\left(A_1^b - \bar{A}_1^b\right) + a_w\left(A_{11} - A_1^b\right) \\ \ldots \\ a_b\left(A_1^b - \bar{A}_1^b\right) + a_w\left(A_{1n_1} - A_{n_1}^b\right) \\ \ldots \\ \ldots \\ a_b\left(A_k^b - \bar{A}_k^b\right) + a_w\left(A_{kn_k} - A_{n_k}^b\right) \end{bmatrix}$$

$$= \begin{bmatrix} a_b\sum_{i=1}^{n_1} n_i\left(A_i^b - \bar{A}_i^b\right)^2 + a_w\sum_{i=1}^{k}\sum_{j=1}^{n_1}\left(A_i^b - \bar{A}_i^b\right)\left(A_{ij} - A_i^b\right) \\ a_b\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(A_{ij} - A_i^b\right)\left(A_i^b - \bar{A}_i^b\right) + a_w\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(A_{ij} - A_i^b\right)^2 \end{bmatrix} \qquad (4.19)$$

$$= \begin{bmatrix} a_b\sum_{i=1}^{k} n_i\left(A_i^b - \bar{A}_i^b\right)^2 \\ a_w\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(A_{ij} - A_i^b\right)^2 \end{bmatrix},$$

car le produit mixte $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(A_{ij}-A_i^b\right)\left(A_i^b-\overline{A}_i^b\right)=0$. Enfin,

$$X'E\left(Z\mid X\right)=\begin{bmatrix} a_b\sum_{i=1}^{k} n_i\left(A_i^b-\overline{A}_i^b\right)^2 \\ a_w\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(A_{ij}-A_i^b\right)^2 \end{bmatrix}$$

Comme (4.18) est égale à (4.19), on a nécessairement :

$$\begin{pmatrix} E\left(\hat{\beta}_b\right) \\ E\left(\hat{\beta}_w\right) \end{pmatrix}=\begin{bmatrix} a_b \\ a_w \end{bmatrix},$$

ce qui montre que l'estimateur $\hat{\beta}$ est un estimateur sans biais de $\left(a_b,a_w\right)'$.

Le modèle (4.16) est équivalent au modèle initial (4.8) si $a_b=a_w=a$, qui se réalise dans le cas où il n'y a pas de stratification de la population ou d'autres causes d'association suspecte (Sham, et al. 2000).

On peut remarquer que, pour vérifier si l'association trouvée est réelle, ou si elle est due à une stratification de la population, on peut tester si $a_b=a_w$. S'il y a égalité, alors on peut conclure à une association, dans le cas contraire on peut conclure qu'une part de l'association trouvée est due à la stratification de la population. Néanmoins, dans le cas où il n'y a pas d'égalité entre les deux paramètres, mais que seule la valeur de a_w est significative, on conclut à l'existence d'une association (Fulker et al. 1999).

4.5.2 Cas particulier d'un enfant malade par famille

Une autre généralisation de la décomposition orthogonale de Fulker peut se faire au cas où on a juste un enfant malade par famille et les génotypes des parents sont connus. Alors

$$A_i^b = \frac{A_{iF} + A_{iM}}{2}.$$

Donc, pour l'enfant appartenant à la famille i, la valeur y_i de son phénotype s'écrit comme suit :

$$Y_i = \mu + a_b A_i^b + a_w \left(A_i - A_i^b \right) + \varepsilon_i, \quad i = 1,\ldots,k .$$

On procède de la même façon que pour le cas où l'on a plusieurs enfants par famille et on calcule la moyenne phénotypique pour l'ensemble des familles :

$$\overline{Y} = \frac{\sum_{i=1}^{k} y_i}{k} = \mu + a_b \overline{A}^b + a_w \left(\overline{A} - \overline{A}^b \right) + \overline{\varepsilon},$$

car dans ce cas $n_i = 1,\ i = 1,\ldots,k$ et $\sum_{i=1}^{k} n_i = k$.

On a aussi :

$$\begin{aligned} Y_i - \overline{Y} &= a_b A_i^b + a_w \left(A_i - A_i^b \right) + \varepsilon_i - \left[a_b \overline{A}^b + a_w \left(\overline{A} - \overline{A}^h \right) + \overline{\varepsilon} \right] \\ &= \left(a_b - a_w \right) \left(A_i^b - \overline{A}^b \right) + a_w \left(A_i - \overline{A} \right) + \left(\varepsilon_i - \overline{\overline{\varepsilon}} \right). \end{aligned} \tag{4.20}$$

Le calcul de l'espérance de (4.20) nous donne :

$$E\left[\left(Y_i - \overline{Y} \right) | X \right] = \left(a_b - a_w \right) \left(A_i^b - \overline{A}^b \right) + a_w \left(A_i - \overline{A} \right).$$

En écriture matricielle, cela revient à :

$$E\left[\left(Y_i-\bar{Y}\right)\mid X\right]=\begin{bmatrix}\left(a_b-a_w\right)\left(A_1^b-\bar{A}^b\right)+a_w\left(A_1-\bar{A}\right)\\ \left(a_b-a_w\right)\left(A_2^b-\bar{A}^b\right)+a_w\left(A_2-\bar{A}\right)\\ \cdots\\ \left(a_b-a_w\right)\left(A_k^b-\bar{A}^b\right)+a_w\left(A_k-\bar{A}\right)\end{bmatrix}$$

$$=\begin{bmatrix}A_1^b-\bar{A}^b & A_1-\bar{A}\\ A_2^b-\bar{A}^b & A_2-\bar{A}\\ \cdots & \cdots\\ A_k^b-\bar{A}^b & A_k-\bar{A}\end{bmatrix}\times\begin{bmatrix}a_b-a_w\\ a_w\end{bmatrix}. \qquad (4.21)$$

Soit $\hat{\beta}=\left(\hat{\beta}_b,\hat{\beta}_w\right)'$ l'estimateur de $\left(a_b-a_w,a_w\right)'$. Comme pour le cas précédent, on va montrer que l'estimateur $\hat{\beta}$ est un estimateur non biaisé pour $\left(a_b-a_w,a_w\right)'$ et en particulier $\hat{\beta}_w$ estime a sans biais, si $a_b=a_w=a$. On procède de la même façon que pour le cas où on a $j=1,\ldots,n_i$ enfants en calculant les espérances de chacun des termes de l'équation suivante :

$$\left(X'X\right)\hat{\beta}=X'Z,$$

où $Z=Y_i-\bar{Y}$. Il suffit pour cela réécrire l'équation (4.16) avec la nouvelle valeur de $X'X$

$$\left(X'X\right)E\left[\hat{\beta}\mid X\right]=\begin{bmatrix}\sum_{i=1}^{k}\left(A_i^b-\bar{A}^b\right)^2 & \sum_{i=}^{k}\left(A_i^b-\bar{A}^b\right)\left(A_i-\bar{A}\right)\\ \sum_{i}\left(A_i-\bar{A}\right)\left(A_i^b-\bar{A}^b\right) & \sum_{i=1}^{k}\left(A_i-\bar{A}\right)^2\end{bmatrix}$$

$$\times\begin{bmatrix}E\left[\hat{\beta}_b\mid X\right]\\ E\left[\hat{\beta}_w\mid X\right]\end{bmatrix} \qquad (4.22)$$

$$=\begin{bmatrix}\beta_b\sum_{i=1}^{k}\left(A_i^b-\bar{A}^b\right)^2+\beta_w\sum_{i=}^{k}\left(A_i^b-\bar{A}^b\right)\left(A_i-\bar{A}\right)\\ \beta_b\sum_{i=}^{k}\left(A_i^b-\bar{A}^b\right)\left(A_i-\bar{A}\right)+\beta_w\sum_{i=1}^{k}\left(A_i-\bar{A}\right)^2\end{bmatrix},$$

où on a posé :

$$E\left[\hat{\beta}_b\right]=\beta_b \text{ et } E\left(\hat{\beta}_w\right)=\beta_w$$

En vertu de la formule (4.21), on écrit l'équation (4.19) pour les matrices X et Z dans ce cas :

$$X'E(Z\mid X)=\begin{bmatrix} A_1^b-\bar{A}^b & A_2^b-\bar{A}^b & \ldots & A_k^b-\bar{A}^b \\ A_1-\bar{A} & A_2-\bar{A} & \ldots & A_k-\bar{A} \end{bmatrix}\begin{bmatrix} A_1^b-\bar{A}^b & A_1-\bar{A} \\ A_2^b-\bar{A}^b & A_2-\bar{A} \\ \ldots & \ldots \\ A_k^b-\bar{A}^b & A_k-\bar{A} \end{bmatrix}\begin{bmatrix} a_b-a_w \\ a_w \end{bmatrix} \tag{4.23}$$

$$=\begin{bmatrix} (a_b-a_w)\sum_{i=1}^{k}\left(A_1^b-\bar{A}^b\right)^2+a_w\sum_{i=1}^{k}\left(A_i-\bar{A}\right)\left(A_i^b-\bar{A}^b\right) \\ (a_b-a_w)\sum_{i=1}^{k}\left(A_i-\bar{A}\right)\left(A_i^b-\bar{A}^b\right)+a_w\sum_{i=1}^{k}\left(A_i-\bar{A}\right)^2 \end{bmatrix}$$

Pour montrer que les composantes de $\hat{\beta}$ sont des estimateurs non biaisés pour a_w-a_b et a_w, on part de l'égalité entre (4.22) et (4.23). Dans ce cas-ci, les facteurs ne sont plus orthogonaux et on doit s'assurer que le système d'équations ci-dessous possède une solution unique en (β_b,β_w) :

$$\begin{aligned} &\beta_b\sum_{i=1}^{k}\left(A_i^b-\bar{A}^b\right)^2+\beta_w\sum_{i=1}^{k}\left(A_i-\bar{A}\right)\left(A_i^b-\bar{A}^b\right) \\ &\quad=\left(a_b-a_w\right)\sum_{i=1}^{k}\left(A_i^b-\bar{A}^b\right)^2+a_w\sum_{i=1}^{k}\left(A_i-\bar{A}\right)\left(A_i^b-\bar{A}^b\right) \\ &\beta_b\sum_{i=1}^{k}\left(A_i-\bar{A}\right)\left(A_i^b-\bar{A}^b\right)+\beta_w\sum_{i=1}^{k}\left(A_i-\bar{A}\right)^2 \\ &\quad=\left(a_b-a_w\right)\sum_{i=1}^{k}\left(A_i-\bar{A}\right)\left(A_i^b-\bar{A}^b\right)+a_w\sum_{i=1}^{k}\left(A_i-\bar{A}\right)^2. \end{aligned} \tag{4.24}$$

Pour cela, il faut vérifier que le déterminant de la matrice $X'X$ est non nul. Cela revient vérifier que :

$$\det \neq 0 \Leftrightarrow \sum_{i=1}^{k}\left(A_i^b - \overline{A}^b\right)^2 \sum_{i=1}^{k}\left(A_i - \overline{A}\right)^2 - \left(\sum_{i=1}^{k}\left(A_i - \overline{A}\right)\left(A_i^b - \overline{A}^b\right)\right)^2 \neq 0$$

$$\Leftrightarrow \sum_{i=1}^{k}\left(A_i^b - \overline{A}^b\right)^2 \sum_{i=1}^{k}\left(A_i - \overline{A}\right)^2 \neq \left(\sum_{i=1}^{k}\left(A_i - \overline{A}\right)\left(A_i^b - \overline{A}^b\right)\right)^2 .$$

Or, on a en général :

$$\sum_i (c)^2 \sum_i (d)^2 \geq \left(\sum_i cd\right)^2 ,$$

ce qui nous donne que :

$$\sum_{i=1}^{k}\left(A_i^b - \overline{A}^b\right)^2 \sum_{i=1}^{k}\left(A_i - \overline{A}\right)^2 \geq \left(\sum_{i=1}^{k}\left(A_i - \overline{A}\right)\left(A_i^b - \overline{A}^b\right)\right)^2 ,$$

avec égalité si et seulement si :

$$A_i^b - \overline{A}^b = A_i - \overline{A}, \ \forall i ,$$

ce qui n'est pas vrai pour tout i. Notre matrice est ainsi de plein rang, et donc on a l'égalité suivante qui nous donne des estimateurs sans biais pour $(a_b - a_w)$ et a_w, car on peut vérifier que la solution de (4.24) est :

$$\beta_b = a_b - a_w, \ \beta_w = a_w .$$

Fulker et al. 1999 ont établi le modèle de base pour une approche systématique pour l'utilisation des paires de frères et sœurs puis généralisé à n'importe quelle taille de la famille par Abecassis et al., 2000. Ce modèle sert de base à une multitude de tests : association,

liaison génétique, association et liaison génétique combinées, stratification, qu'on va brièvement décrire ci-dessous.

4.6 Tests d'association et de liaison génétique

L'utilisation du modèle orthogonal de Fulker et al. (1999), nous permet d'éliminer le problème de la fausse association. Pour illustrer, nous expliquons les principes de ces tests à partir du modèle de Fulker. Dans le cadre de ces modèles et de ses généralisations, le test d'association consiste à calculer un rapport de vraisemblance de la loi du χ^2 à 1 degré de liberté, où a_w sera libre sous l'hypothèse alternative, et $a_w = 0$ sous l'hypothèse nulle ; par contre a_b sera libre sous les 2 hypothèses, ce qui nous permet d'écrire les modèles sous cette forme :

Modèle restreint : $\mu_{ij} = \mu + a_b A_i^b$

Modèle complet : $\mu_{ij} = \mu + a_b A_i^b + a_w\left(A_{ij} - A_i^b\right)$.

Les hypothèses sont alors :

$$H_0 : \quad a_w = 0$$

$$H_1 : \quad a_w \neq 0$$

On conclut à l'existence de l'association si l'hypothèse nulle est rejetée.

4.6.1 Test de liaison génétique

Lorsque notre objectif est de tester seulement la liaison génétique, celle-ci est alors modélisée à travers la structure de la covariance tout en éliminant l'effet de l'association.

Ceci nous amène à écrire les deux modèles restreint et complet sous la forme suivante :

Modèle restreint : $\mu_i = \mu$

$$\sigma^2 = \sigma_s^2 + \sigma_e^2$$

Modèle complet : $\mu_i = \mu$

$$\sigma^2 = \sigma_a^2 + \sigma_s^2 + \sigma_e^2$$

La liaison est modélisée dans la variance due à l'effet additif du locus candidat c'est-à-dire σ_a^2. Les deux hypothèses à tester sont alors les suivantes :

$$\begin{cases} H_0 : \sigma_a^2 = 0, \\ H_1 : \sigma_a^2 \neq 0. \end{cases}$$

De la même façon, on utilise le test du rapport de vraisemblance pour comparer ces deux modèles et tester si le paramètre est significatif. Si le test est significatif, c'est-à-dire que la variance due à l'effet additif du gène est différente de 0, alors on peut conclure à une liaison génétique entre le locus marqueur et le phénotype, sinon il n'y a pas de liaison.

4.6.2 Test conjoint d'association et de *linkage*

Lorsque l'hypothèse de liaison entre un locus marqueur et un phénotype est vérifiée, on peut vouloir tester si ce locus marqueur est bien le gène de la maladie. Pour cela, on utilise le test conjoint « d'association et de liaison » (Fulker et al.1999).

Le test conjoint d'association et de liaison génétique est utilisé quand on veut tester la liaison tout en modélisant l'association. Le test devrait donner comme résultat : soit le locus QTL est candidat, soit il est simplement en déséquilibre avec le locus du caractère étudié. Si on trouve que le test de liaison est significatif quand on modélise l'association, on peut conclure que ce locus n'est pas le gène fonctionnel, mais plutôt c'est un locus en déséquilibre avec le locus du caractère. On peut ainsi poser les deux modèles, restreint et complet :

Modèle restreint : $\mu_{ij} = \mu + a_b A_i^b + a_w \left(A_{ij} - A_i^b \right)$,

$$\sigma^2 = \sigma_s^2 + \sigma_e^2 .$$

Modèle complet : $\mu_{ij} = \mu + a_b A_i^b + a_w \left(A_{ij} - A_i^b \right)$,

$$\sigma^2 = \sigma_a^2 + \sigma_s^2 + \sigma_e^2 .$$

Nous ne présentons pas les détails de ces tests.

Pour conclure, dans ce chapitre, nous avons donc vu comment décomposer la valeur phénotypique en plusieurs composantes. Cette décomposition nous a permis d'introduire le test du maximum de vraisemblance sur lequel se sont basés les tests d'association et de liaison génétique.

Dans le chapitre suivant, nous présentons, en utilisant le logiciel élaboré par Abecasis et al. (2000), une application de ces tests pour un ensemble de données réelles.

CHAPITRE V

ÉTUDE DE CAS : ASSOCIATION GÉNÉTIQUE CHEZ LES ENFANTS HYPERACTIFS

Dans ce chapitre, nous allons appliquer les deux tests présentés aux chapitres 2 et 4 à des données réelles portant sur des enfants souffrant de l'hyperactivité. Mais avant de décrire les données sur lesquelles reposent les tests, il est important de donner un aperçu de la maladie d'intérêt.

L'hyperactivité est un problème de comportement. Même si le degré des symptômes varie d'une personne à l'autre, son effet est ressenti par chacun de la même façon et peu importe l'endroit où il est se trouve, à la maison, à l'école, au travail par exemple. Un enfant atteint de l'hyperactivité est généralement décrit comme un enfant impulsif (entraînant des perturbations dans sa vie quotidienne), dérangeant, qui éprouve des difficultés à se tenir en place. Il ne peut rester assis trop longtemps, il éprouve tout le temps un besoin de se défouler en courant ou en hurlant. Quand ses parents ou l'enseignant le rappellent à l'ordre, il n'arrive pas à se contrôler plus d'un instant. Le problème est neuro-développemental, l'enfant n'agit donc pas par mauvaise volonté mais par manque de contrôle. On désigne généralement ce trouble sous le nom du TDA/H (trouble du déficit de l'attention, avec ou sans hyperactivité).

5.1 Définition et diagnostic du TDA/H

5.1.1 Définition du TDA/H

Le TDA/H désigne un comportement pathologique précis qui survient chez l'enfant ; c'est l'un des troubles neuropsychiatriques le plus commun de l'enfance. On estime qu'il touche de 3 à 7 % des enfants d'âge scolaire et ce trouble peut, dans plus de la moitié des cas, persister jusqu'à l'âge adulte.

Il est caractérisé par trois principaux symptômes : un déficit de l'attention (ou inattention), impulsivité et hyperactivité. Ces trois symptômes peuvent être présents chez les enfants atteints de TDA/H mais à des degrés différents et, par conséquent, le trouble est sous-divisé en trois types : a) le type «inattention prédominante», b) le type «hyperactivité/impulsivité prédominante», c) le type mixte (TDA/H). Cette affection est considérée comme un handicap neurobiologique qui interagit avec l'habilité d'une personne à se montrer attentive par rapport à une tâche et à contrôler le comportement impulsif.

La présence du TDA/H peut causer de graves problèmes aux conséquences négatives et non négligeables pour les personnes qui en sont atteintes. De nombreux essais cliniques révèlent que les enfants souffrant du TDA/H sont beaucoup plus susceptibles que les enfants normaux d'abandonner l'école (32 à 40%), de ne pas compléter les études collégiales (à peine 5 à 10 % consentent à poursuivre des études collégiales), de se tourner vers la délinquance (40 à 50%)[5]. En outre, les enfants ayant grandi avec un TDA/H sont sujets à des troubles de comportement et de personnalité et gèrent mal leur vie d'adulte. Ils peuvent ainsi constituer un fléau pour leur entourage et la société toute entière. C'est pourquoi de nombreux chercheurs s'intéressent à l'étude des causes du TDA/H afin de trouver des traitements adéquats. Jusqu'à présent, de nombreuses études s'entendent pour dire qu'il y a une composante génétique ainsi que des facteurs environnementaux et que les deux interagissent. Plusieurs gènes candidats ont été décelés et sont toujours à l'étude.

5.1.2 Symptômes du TDA/H

Le TDA/H se manifeste généralement au début de la petite enfance, soit aux environs de 2 à 3 ans, mais ce n'est qu'au moment de la fréquentation de l'école que la maladie est décelée. Le diagnostic officiel du TDA/H repose sur les directives énoncées dans la 4e édition du *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-IV*) produit par l'Association Américaine de Psychiatrie. Les critères du diagnostic officiel établissent que le déclenchement des symptômes doit avoir lieu avant qu'une personne ait sept ans. Par contre,

---

[5] http://panda.cyberquebec.com/lesite.htm

pour les principaux chercheurs dans ce domaine, cette limite d'âge devrait être augmentée au-delà de l'âge de sept ans.

D'après le *DSM-IV,* les personnes souffrant du trouble du déficit de l'attention, avec ou sans hyperactivité (TDA/H) peuvent présenter les différents types du TDA/H selon la classification en trois catégories présentée à la sous-section 5.1.1 suivante.

5.1.3 Outils de diagnostic du TDA/H

Comme nous l'avons déjà souligné ci-dessus, le diagnostic du TDA/H est basé sur les directives du *DSM-IV* et s'établit par le biais du *DISC4.* Le *DISC4* (voir appendice A) est un outil médical qui permet de diagnostiquer tous les problèmes psychiatriques, dont le TDA/H. C'est donc l'outil diagnostic qui permet de classer chaque enfant, s'il est « *oui ou non* », inattentif, hyperactif/impulsif ou TDA/H (appendice A, catégorie 26).

Une fois la présence de la maladie établie, on veut mesurer son amplitude, mesure pour laquelle il existe une multitude d'outils disponibles, dont certains se basent sur des scores établis sur une *échelle continue*. Nous ne citerons que ceux utilisés dans notre travail et qui sont les *Conners* versions parent et enseignant (*Conners' Rating Scales*).

L'échelle d'évaluation de Conners, version parent et enseignant est un instrument sous forme d'un questionnaire (voir appendices B et D) qui est administré en même temps aux parents (version parent), et à l'enseignant de l'enfant (version enseignant), cela est fait pour des besoins de comparaison. Ce questionnaire regroupe un ensemble de questions (59 pour l'échelle de Conners-enseignant et 80 pour celle des parents) et chaque question est cotée sur une échelle de 0 à 3, dépendamment de la réponse par rapport au comportement de l'enfant, c'est-à-dire, si l'enfant présente le comportement dont il est question); on code 0 pour aucunement vrai, 1 pour un peu vrai, 2 pour plutôt vrai et 3 pour très vrai.

Ce questionnaire sert de base pour la mesure du comportement de l'enfant qui se fait à travers «la cotation de l'échelle de Conners». Celle-ci regroupe treize sous échelles (voir appendices C et E) et chaque sous-échelle correspond à un ensemble d'items (questions) associés (parmi les 59 ou 80 selon le type de Conners). Pour mesurer le degré de l'inattention, de l'hyperactivité/impulsivité, ou du TDA/H, on considère les trois sous-

échelles du *DSM-IV* (toujours dans la cotation de l'échelle) qui correspondent aux trois catégories suivantes :

- *DSM-IV Inattention*,
- *DSM-IV Hyperactivité-Impulsivité*,
- *DSM-IV Total.*

Cette classification DSM-IV se base sur dix-huit questions, neuf questions portent sur l'hyperactivité et les neuf autres questions portent sur l'inattention :

- si, sur les neuf questions sur l'inattention, six sont positives alors on diagnostique « le trouble d'inattention »,
- et si sur les neuf questions sur l'hyperactivité six sont positives, alors on diagnostique « l'hyperactivité ».
- Pour pouvoir diagnostiquer TDA/H, il faut avoir six réponses positives pour l'inattention et six réponses positives pour l'hyperactivité[6].
- Il faut aussi que les symptômes durent depuis au moins six mois et qu'ils soient apparus, à la maison comme à l'école, avant l'âge de 7 ans pour que le diagnostic soit fait.

---

[6] En Europe, en plus des douze réponses, il faut que l'enfant ait des troubles de conduite (symptômes et problèmes de comportement agressif).

5.2 Résultats de l'enquête

5.2.1 Bref aperçu de l'enquête

L'étude sur laquelle s'est basé notre travail a été entreprise par l'hôpital Sainte-Justine en collaboration avec Galileo Génomique[7]. L'objectif de l'étude est d'identifier des gènes impliqués dans le trouble de déficit de l'inattention, avec ou sans hyperactivité (TDA/H).

L'étude a été menée entre 2001 et 2003. Elle a d'abord été entreprise par l'hôpital sur 150 enfants issus de nationalités différentes et référés par leurs médecins traitants. Suite à cela, une entente entre l'hôpital et Galileo s'est faite et il y a eu recrutement par Galileo, au niveau de Montréal et en région partout au Québec, de 500 nouveaux enfants, tous des canadiens français. Dans l'échantillon, il y a donc une partie des enfants (canadiens français) qui a été évaluée à Sainte Justine (avec plus de tests et de questionnaires) et une autre partie dans les régions (avec moins de tests et de questionnaires). Comme les variables utilisées proviennent de différents fichiers que nous avons associés, la taille de l'échantillon diffère d'une analyse à l'autre. D'autre part, le problème des données manquantes ainsi que le problème de codification utilisée, qui est différente selon la source, a réduit la taille de l'échantillon. Les données sur lesquelles nous travaillons sont donc basées sur une partie de cette étude, soit 308 enfants.

L'âge des enfants de l'échantillon varie entre 5 et 12 ans avec un âge moyen de 8 ans. Parmi eux, 80 % sont des garçons et 20% sont des filles[8], soit quatre fois moins que les garçons. Le quotient intellectuel des enfants varie entre 52 et 137 avec une moyenne de 96,5. Le tableau 5.1 présente aussi des statistiques sur l'âge des deux parents au moment de l'enquête ainsi que sur le quotient intellectuel de l'enfant.

---

[7] Galileo Génomique inc. est une société dédiée à la découverte de cartes géniques, de gènes et de biomarqueurs liés aux causes des maladies communes et à la réponse aux médicaments.

[8] Résultats obtenus à partir du questionnaire sur les renseignements généraux.

Tableau 5.1 Quelques statistiques descriptives de l'échantillon des 308 enfants

Variables	Moyenne	Médiane	Mode	Écart type
Âge de l'enfant	8,1	8,0	8,0	1,5
Âge de la mère	36,9	37,0	36,0	5,1
Âge du père	39,9	39,0	39,0	6,5
QI	96,5	96,0	89,0	14,0

5.2.2 Questionnaires

L'enquête RECIH (Programme de Recherche Interdisciplinaire sur l'Hyperactivité des enfants) comporte plusieurs questionnaires; des questionnaires destinés aux enfants souffrant du trouble de comportement, aux parents et aux enseignants. Dans ce qui suit, nous allons présenter uniquement les questionnaires utilisés dans notre étude ainsi que les principales variables utilisées.

1 Questionnaire de renseignements généraux

Ce questionnaire comporte des questions sur plusieurs volets dont les plus importants sont :

- famille de l'enfant : coordonnées de l'enfant, lieu de naissance des parents et grands parents, nombre de frères et sœurs, etc.
- Grossesse et accouchement : complications durant la grossesse et l'accouchement, consommation de la mère durant la grossesse des substances telles que les cigarettes et l'alcool, etc.
- données socio-économiques : niveau d'instruction des parents, source et niveau de revenu.

2 Questionnaires sur l'échelle de Conners versions enseignant et parent (appendices B et D).

Ces questionnaires ont été décrits à la section 5.1.1.

3 *WISC-III* (appendice F).

Le *WISCIII* ou Échelle *d'Intelligence de Wechsler pour Enfants* est un test soumis à l'enfant et qui permet de mesurer son quotient intellectuel (QI).

5.3 Analyse des données

Dans notre analyse, les phénotypes utilisés sont les scores obtenus aux deux échelles de Conners.

Dans ce qui suit, nous allons commencer par comparer les résultats des différentes sources, c'est-à-dire comparer les résultats des deux échelles de mesure de Conners (enseignant et parent) en procédant à une validation des scores. Nous allons comparer aussi les résultats du diagnostic donné par le *DISC4* avec les scores fournis par les échelles de Conners. Dans une seconde étape, il s'agit d'étudier les variables environnementales appelées aussi les facteurs de risque qui peuvent influer le TDA/H. Enfin, dans la dernière étape, nous allons nous pencher sur l'étude des facteurs génétiques qui ont une incidence sur le TDA/H.

5.3.1 Validation des scores

Comme nous avons deux échelles (questionnaires parents et enseignant) qui mesurent le même phénomène, à savoir le comportement de l'enfant, on s'intéresse à voir si les résultats donnés par chacune d'entre elles sont similaires ou, au contraire, différents. Dans le premier cas de figure, nous aurons à utiliser les deux échelles pour faire les analyses, dans le second cas, nous utiliserons les deux échelles et procéderons par la suite à une comparaison des résultats. Pour cela, nous devons vérifier si les enfants sont classés de la même façon par les deux questionnaires. Nous allons procéder ainsi à une validation des scores.

i) Validation des scores des deux échelles de Conners

À partir de la valeur du score pour chaque type de trouble défini dans le *DSM-IV* de chaque échelle de cotation de Conners version, parent et enseignant, nous avons classé les enfants selon deux catégories : ceux dont la valeur du score est élevée (atteint) ou basse (non atteint). On considère comme personne atteinte, chaque enfant dont la valeur du score est supérieure à 60 (voir l'échelle en appendices C et E). Nous nous sommes arrêtés à la borne 60, car elle correspond à la limite à partir de laquelle on diagnostique un problème important, contrairement à un problème qui ne soulève qu'une faible ou basse inquiétude. On obtient ainsi un tableau où les enfants sont scindés en quatre groupes : les concordants, c'est-à-dire ceux ayant été classés de la même façon par les deux Conners, qu'ils soient atteints ou non atteints, et les non concordants, c'est-à-dire ceux ayant été classés atteints par l'un deux et non atteints par l'autre. On obtient pour chaque type de trouble les tableaux 5.2, 5.3, 5.4 suivants.

Tableau 5.2 Comparaison des Scores pour le TDA/H

Effectifs	Atteint	Non atteint	Total
Atteint	10	58	68
Non atteint	7	95	102
Total	17	153	170

Tableau 5.3 Comparaison des scores pour l'hyperactivité/impulsivité

Effectifs	Atteint	Non atteint	Total
Atteint	28	53	81
Non atteint	10	79	89
Total	38	132	170

Tableau 5.4 Comparaison des scores pour l'inattention

Effectif	Atteint	Non atteint	Total
Atteint	11	44	55
Non atteint	8	107	115
Total	19	151	170

En utilisant le test de McNemar (voir chapitre 2), on s'intéresse à vérifier s'il y a une différence significative entre les non concordants. Pour réaliser ceci, on compare le nombre d'enfants non concordants (déclarés atteints par un questionnaire mais non atteints par l'autre). Par exemple, dans le tableau 5.2, on compare 58 et 7. Ainsi, pour les trois scores obtenus par le *DSM-IV* Inattention, le *DSM-IV* Hyperactivité/Impulsivité et le *DSM-IV* total (TDA/H), les différences sont significatives, $P < 0,0001$ (tableau 5.5), ce qui nous permet de conclure que les deux questionnaires de Conners ne classifient pas tous les enfants de la même façon. Ces résultats sont confirmés par le test t de Student pour la moyenne de la différence entre les deux scores ; la comparaison de la différence des scores des trois phénotypes : Hyperactivité/Impulsivité, Inattention et le TDA/H obtenus par les Conners parents et enseignants donne des résultats significatifs $(P < 0.0001)$.

Tableau 5.5 Résultats du test de McNemar pour les trois phénotypes

Phénotypes	Statistique McNemar	ddl	P-valeur (McNemar)
Hyperactivité/impulsivité	29,35	1	< 0,0001
Inattention	24,92	1	< 0,0001
TDA/H	40,02	1	< 0,0001

Par conséquent, pour nos analyses, nous allons utiliser les phénotypes obtenus à partir de ces deux questionnaires et faire des comparaisons de résultats.

ii) Validation des scores du diagnostic du *DISC4* et l'échelle de Conners

De la même façon que nous avons comparé les résultats des échelles de Conners parent et enseignant, nous nous intéressons à savoir si le diagnostic donné par la *DISC4* correspond aux résultats donnés par l'échelle de Conners, en posant un seuil à partir duquel l'enfant est déclaré malade. Pour des besoins de comparaison, nous avons présenté le diagnostic du *DISC4* selon qu'il classe chaque enfant, soit atteint par l'un des trois types de troubles, soit non atteint comme le montre le tableau 5.6. Il y a ainsi 35 enfants qui ne présentent aucune forme de trouble et 272 qui sont classés dans au moins une des trois catégories de diagnostic.

Tableau 5.6 Résultats du diagnostic du *DISC4* selon que l'enfant présente ou ne présente pas un des troubles

Phénotypes	Nombre d'enfants	Pourcentage
Non atteints	35	11,40
Atteints	272	88,60
Total	307	100

Pour pouvoir faire la comparaison décrite ci-dessus, nous avons classé les enfants en deux catégories suivant la cotation de l'échelle de Conners parents et enseignants (appendices C et E), ceux qui sont atteints par un des trois troubles (inattention, hyperactivité/impulsivité et TDA/H) et ceux qui ne sont atteints pas aucun des troubles. On considère qu'un enfant est atteint si le score obtenu pour un de ces trois troubles est supérieur à la valeur 60[9] ; dans le cas contraire, on considère que l'enfant n'est pas atteint. Pour ce faire, nous avons créé une

9 Voir l'explication du choix de la valeur 60 à la section 5.3.1 (i).

variable notée «CONTROL» qui prend la valeur 0 (non atteint), si les trois scores de *DSM-IV* indiquent que l'enfant n'est pas malade (scores inférieurs à 60), et qui prend la valeur 1 (atteint), si au moins un des scores est supérieur à la valeur limite 60. Si on considère les résultats de la comparaison entre le diagnostic du *DISC4* et l'échelle de Conners-enseignant résumés dans le tableau 5.7, on voit qu'il y a 30 (18 %) enfants diagnostiqués par le *DISC4* comme atteints d'un des trois troubles étudiés, alors qu'ils ne présentent pas de symptômes selon l'échelle Conners, et il y a 14 (8 %) enfants diagnostiqués comme non atteints par le *DISC4*, mais ont des scores de Conners élevés (supérieurs à 60) pour au moins un des trois troubles (le tabl. 5.8 présente les résultats en nombres absolus pour la comparaison du diagnostic du *DISC4* et de la cotation de l'échelle de Conners-parent).

Tableau 5.7 Comparaison des résultats du diagnostic du *DISC4* et du score de l'échelle Conners-enseignant.

DISC4	CONTROL		
	Non atteint	Atteint	Total
Non atteint	3	14	17
Atteint	30	123	153
Total	33	137	170

Tableau 5.8 Comparaison des résultats du diagnostic du *DISC4* et du score de l'échelle Conners-parent.

DISC4	CONTROL		
	Non atteint	Atteint	Total
Non atteint	10	19	29
Atteint	6	225	231
Total	16	244	260

Le test de McNemar (voir chapitre 2) donne des résultats significatifs, nous avons $P = 0,0226$ et $P = 0,0146$ respectivement pour les Conners-enseignant et pour les Conners-parent, et donc on peut conclure que le *DISC4* et l'échelle Conners ne concordent pas toujours quant à la classification des enfants.

5.3.2 Analyse des facteurs environnementaux

Les recherches portant sur les troubles liés au TDA/H se divisent en deux tendances, celle privilégiant les facteurs génétiques et celle privilégiant les facteurs environnementaux. Dans le cadre de notre travail, nous nous intéressons à l'étude des deux types de facteurs.

L'objectif dans cette section est de chercher à expliquer la maladie de l'enfant (le TDA/H), en utilisant un modèle de régression multiple dans lequel la variable dépendante sera le score de la maladie et les variables indépendantes vont représenter les facteurs environnementaux. Comme nous avons un nombre assez important de variables, et afin de voir quelles variables intégrer dans le modèle (s'il y a lieu), nous allons faire une analyse préliminaire des données en étudiant les relations qui existent entre le score de la maladie et les différentes variables environnementales. Pour ce faire, nous avons procédé à une analyse préliminaire des données où nous avons étudié la relation entre le score de la maladie[10] obtenu par les deux échelles de Conners et chacune des variables suivantes :

- sexe de l'enfant (variable nominale),
- âge de l'enfant et des deux parents (variable continue),
- complication durant la grossesse (variable catégorielle),
- complication durant l'accouchement (variable catégorielle),
- durée de la grossesse (variable continue),
- poids à la naissance (variable continue),

[10] Nous utilisons le mot «maladie» pour désigner le trouble de comportement.

- consommation des médicaments (autres que des vitamines) par la mère durant la grossesse (variable catégorielle),
- consommation de cigarettes par la mère durant la grossesse (variable catégorielle),
- consommation d'alcool par la mère durant la grossesse (variable catégorielle),
- tranches d'âge auxquelles l'enfant a marché cinq pas ou plus (variable catégorielle),
- âge auquel l'enfant a dit une phrase d'au moins trois mots (variable catégorielle),
- score Apgar [11] (variable catégorielle),
- diagnostic d'hyperactivité dans la famille (seule la variable portant sur le diagnostic d'hyperactivité chez les frères et sœurs a été utilisée, car le nombre de réponses est négligeable pour celui des parents),
- composition de la famille (variable catégorielle),
- niveau de scolarité atteint par la mère (variable catégorielle),
- niveau de scolarité atteint par le père (variable catégorielle),
- niveau de revenu familial (variable catégorielle),
- le quotient intellectuel (variable continue).

Le choix de ces variables est justifié par la qualité des réponses et aussi par le fait que certaines d'entre elles sont citées dans des études comme étant susceptibles d'être responsables du dysfonctionnement dans le comportement de l'enfant (consommation par la mère durant la grossesse de drogues, alcool, cigarettes, etc.).

---

[11] Le score Apgar permet d'évaluer l'état de santé initial d'un nouveau né puis son évolution à quelques minutes d'intervalle. Les scores compris entre 8 et 10 indiquent une bonne santé, les scores compris entre 4 et 7 indiquent un problème et enfin les scores compris entre 0 et 3 indiquent de graves problèmes.

i) Conners-enseignants

Avant d'étudier les relations de dépendance entre ces différentes variables et les scores de la maladie étudiée, nous avons observé quelques statistiques sur les différents scores (tableaux 5.9 et 5.10) qui montrent qu'en moyenne les scores fournis par l'échelle Conners-parents sont supérieurs à ceux fournis par l'échelle de Conners-enseignants.

Nous avons aussi examiné la normalité des différents scores qui montre une distribution approximativement normale (il y a quelques valeurs extrêmes).

Tableau 5.9 Statistiques sur les scores obtenus par les Conners-enseignants

Type de trouble	Moyenne	Médiane	Mode	Écart type	Étendue
Hyperactivité/impulsivité	63,9	64,0	64,0	10,4	47,0
Inattention	65,4	66,0	74,0	10,3	49,0
TDA/H	63,9	64,0	64,0	10,4	47,0

Tableau 5.10 Statistiques sur les scores obtenus par les Conners-parents

Type de trouble	Moyenne	Médiane	Mode	Écart type	Étendue
Hyperactivité/impulsivité	71,8	72,0	90,0	12,8	49,0
Inattention	72,8	74,0	74,0	10,0	49,0
TDA/H	73,9	74,0	70,0	10,0	41,0

Comme on l'a dit en introduction, nous allons étudier les relations entre les scores de la maladie (trouble) et les variables environnementales afin de voir si on peut constituer un modèle de régression qui expliquerait le trouble de l'enfant.

L'examen de la matrice de corrélation entre chacun des trois scores : hyperactivité/impulsivité, inattention et le TDAH, avec les variables de type continu, comme le poids à la naissance, etc., n'a pas montré des relations significatives.

Pour les variables catégorielles, nous avons utilisé le test de Student dans le cas où la variable comporte deux catégories (sexe, consommation de cigarettes) ou des analyses d'ANOVA lorsque la variable indépendante comporte plus de deux catégories (revenu familial, niveau de scolarité de la mère, etc.). Ainsi, seule la variable sexe donne des résultats significatifs. Les résultats du test de Student sont résumés dans le tableau 5.11 : les scores du TDA/H sont en moyenne plus élevés chez les filles que les garçons. Par contre, les résultats ne sont pas significatifs pour l'hyperactivité/impulsivité.

Tableau 5.11 Scores moyens selon le sexe Conners-enseignants

	Filles	Garçons	t	p
Scores moyens	70	62	3,82	0,0003
Scores moyens	74	63	5,36	< 0,0001

L'ANOVA sur la variable revenu total du ménage qui est une variable classée selon huit classes de revenus, donne des résultats significatifs seulement pour le score de l'hyperactivité/impulsivité ($F = 2,04$, $P = 0,045$), tandis que l'analyse de Student-Newman-Keuls (qui permet de classer entre eux les groupes qui ont les mêmes moyennes) regroupe toutes les classes de revenus dans une seule classe. La même étude a été menée pour les variables portant sur le niveau atteint de la mère et du père et aucun résultat significatif ne se dégage.

Une autre variable qui peut avoir un effet sur le comportement de l'enfant est le type de famille dans laquelle il vit (biparentale, monoparentale, reconstituée ou autre). En utilisant le test non paramétrique de Kruskal-Wallis à un facteur, nous avons étudié l'effet de la variable type de famille sur le score de la maladie. Les résultats ne sont pas significatifs, ce

qui nous permet de conclure que cette variable (type de famille) n'a pas d'effet sur les troubles de comportement de l'enfant.

ii) Conners-parents

Les mêmes analyses pour les scores obtenus à partir de l'échelle Conners-enseignant ont été faites avec les scores obtenus à partir de l'échelle Conners-parent et les résultats sont similaires ; il y a une relation entre le TAD/H et la variable sexe de l'enfant seulement (voir tableau 1.12).

Tableau 5.12 Scores moyens selon le sexe et les scores du Conners-parents

	Filles	Garçons	*t*	*P*
Scores moyens	79	72	4,72	< 0,0001
Scores moyens	81	71	7,78	< 0,0001

Rappelons qu'on a quatre fois moins de filles que de garçons qui ont le TDA/H, par contre selon les tableaux 5.11 et 5.12, on conclut que les filles sont touchées plus sévèrement que les garçons.

On s'intéresse aussi à étudier les enfants atteints du TDA/H selon leur niveau d'intelligence mesuré par le quotient intellectuel (QI). Généralement, les enfants hyperactifs sont décrits comme étant des enfants très intelligents. Le calcul des coefficients de corrélation linéaire de Pearson, Spearman et Kendall entre les différents scores et le QI donnent effectivement des valeurs négatives pour les coefficients de corrélation ce qui signifie que la relation entre les deux variables est négative, mais les résultats ne sont pas significatifs. Nous avons refait les analyses en constituant des classes pour le quotient intellectuel d'après l'échelle de *WISC-III* et nous avons utilisé le test non paramétrique de Kruskal-Wallis à un facteur. Les résultats ne sont toujours pas significatifs, et on conclut ainsi qu'il n'y a pas de relation significative entre, les enfants souffrant d'un des symptômes du TDA/H, et le quotient intellectuel (QI).

iii) DISC4

Nous avons présenté les résultats fournis par le diagnostic du *DISC4* au tableau 5.6, où nous avons regroupé les enfants en deux catégories, selon qu'ils sont atteints par un des trois troubles ou non atteints. Dans cette section, nous allons nous intéresser à présenter les mêmes résultats, mais selon la classification exacte donné par le *DISC4*, c'est-à-dire présenter les enfants atteints selon le type de trouble qu'ils présentent (tableau. 5.13). Ainsi, selon ce diagnostic, 12 % des enfants examinés souffrent de l'hyperactivité ou impulsivité, près de 32% souffrent du déficit de l'attention et près de 45 % souffrent de type mixte TDA/H, tandis que 11 % des enfants sont déclarés comme n'ayant aucun trouble (non atteints).

Tableau 5.13 Résultats du diagnostic du *DISC4* selon le type du trouble

Phénotypes	Nombre d'enfants	Pourcentage
Non atteints	35	11,40
Inattention	98	31,92
Hyperactivité/Impulsivité	37	12,05
TDA/H	137	44,63
Total	307	100

Il s'agit maintenant d'étudier les relations entre le TDA/H et les variables environnementales. Comme la variable qu'on veut expliquer (c'est-à-dire la maladie) fournie par le *DISC4* est catégorielle, l'étude des relations qui existent entre cette variable et certaines variables environnementales sera faite en utilisant l'ANOVA ou le test d'indépendance du χ^2, selon que la variable environnementale est continue (poids) ou catégorielle (sexe). Par contre, si la variable environnementale est binaire de type «oui» ou «non» telle que la consommation de cigarettes, la consommation d'alcool, etc., on utilisera des outils propres aux études épidémiologiques, qu'on aura à définir.

Ainsi, les résultats du test d'indépendance du χ^2 ne sont significatifs que pour la variable sexe ($\chi_3^2 = 8,27, P = 0,041$), résultat que nous avons trouvé aussi avec l'échelle de Conners (avec une P-valeur plus faible).

Pour le QI, en utilisant les résultats du *DISC4* et la classification du QI selon l'échelle de *WISC-III* (tableau. 5.14), le test du χ^2 n'a pas donné de résultats significatifs. Il n'y a pas de relation entre la maladie et le quotient intellectuel de l'enfant. Le même résultat est donné par l'étude des variables continues (en utilisant l'ANOVA) telles que le poids et la taille à la naissance des enfants.

Tableau 5.14 Répartition des enfants selon la classe du QI et le diagnostic *DISC4*

Quotient intellectuel	Non malades	Inattentifs	Hyperactifs/ impulsifs	TDA/H	Total
Supérieur à la moyenne *	6	16	11	23	56
%	10,7	28,6	19,6	41,1	100
Moyenne	15	46	17	65	143
%	10,5	32,2	11,9	45,5	100
Inférieur à la moyenne *	14	35	8	46	103
%	13,6	33,9	7,8	44,7	100
Total	35	97	36	134	302**
%	11,6	32,1	11,9	44,4	100

* Comme certaines cellules avaient de petites tailles, nous avons regroupé les classes où le QI est « exceptionnel, supérieur et juste supérieur à la moyenne » dans la seule classe «supérieure» et les classes «juste inférieure à la moyenne, limite, déficience» dans la «classe inférieure à la moyenne».

** La taille de l'échantillon est égale ici à 302 à cause des données manquantes dans la table du *WISC-III*.

Enfin, nous allons terminer en étudiant un autre type de variable qu'on considère comme facteur de risque par l'utilisation d'outils d'épidémiologie. Pour ce faire, on supposera que nous avons une étude cas-témoins dans laquelle les enfants non atteints nous serviront de « témoins » et ceux atteints de « cas », et on calculera un rapport des cotes (*RC*) qui mesure l'association entre une maladie M et un facteur E d'exposition (ou de risque) comme le montre dans l'exemple théorique suivant (voir Bernard et Lapointe, 1987)

Soit la maladie M et le facteur E pour lesquels on voudrait savoir s'il y a une association entre eux ; M et E sont considérées comme des variables dichotomiques, 0 pour absent et 1 pour présent. Pour une population de taille N, on représente les effectifs selon qu'ils sont atteints ou non par la maladie et selon le fait que le facteur E est présent ou non.

En général, un tableau pour effectuer des calculs de *RC* se présente sous la forme suivante :

Tableau 5.15 Tableau permettant le calcul d'un rapport de cotes

M	E		
	1	0	Total
1	a	b	n_1
0	c c	d	n_2
Total	n_1	n_2	n

Le rapport $\frac{a}{b}$ chez les cas et le rapport $\frac{c}{d}$ chez les témoins représentent respectivement *la cote* en faveur de l'exposition et contre la non exposition au facteur E. Une cote égale à 2 indique que la chance d'avoir été exposé est deux fois plus grande que celle de ne pas l'avoir été. Si on veut comparer les cas et les témoins par rapport aux chances d'exposition, on calcule un rapport des cotes défini comme suit :

$$RC = \frac{a/b}{c/d} = \frac{ad}{bc}.$$

Si les deux groupes, cas et témoins, ont des cotes égales alors $RC = 1$ et on conclut alors qu'il n'y a pas d'association entre la maladie et le facteur d'exposition. Par contre, si la cote d'exposition est plus forte chez les cas que chez les témoins, on a $RC > 1$ et on conclut qu'il y a une association positive entre la maladie M et le facteur E.

Dans notre cas la maladie est le TDA/H, le déficit d'attention ou encore l'hyperactivité/impulsivité, et le facteur de risque est défini par l'une des variables environnementales suivantes : complication durant l'accouchement, la grossesse, consommation de médicaments, alcool, cigarettes, diagnostic d'hyperactivité dans la famille (section 5.3.1).

Si on considère l'exemple du TDA/H, nous avons alors les 137 enfants atteints comme « les cas » et les 35 enfants non atteints comme « les témoins ». Ainsi, si on considère le facteur d'exposition représentant la variable complication de grossesse, nous obtenons le tableau 5.16 suivant (les chiffres du tableau ne correspondent pas à ceux cités ci-dessus à cause des non réponses).

Tableau 5.16 Nombre d'enfants sains et ceux ayant le TDA/H selon le risque de complication de grossesse

TDA/H	Complication grossesse		Total
	1	0	
1	25	78	103
0	7	22	29
Total	32	100	132

Le rapport des cotes RC dans ce cas est égal à $\frac{25 \times 22}{78 \times 7} = 1.007 \simeq 1$, ce qui suggère qu'il n'y a pas d'association entre le TDA/H et une complication de grossesse. Le tableau 5.17 représente, pour les trois troubles étudiés, certaines associations positives avec les facteurs étudiés.

Tableau 5.17 Valeurs du rapport de cotes *RC* pour le TDA/H et des facteurs de risque

Facteurs de risques	TDA/H	Inattention	Hyperactivité/ impulsivité
Consommation cigarettes	1,3	0,6	0,6
Consommation médicaments	1,4	0,7	0,7
Complication grossesse	1,0	1,1	1,1

Nous avons testé, en utilisant le logiciel SAS, toutes ces valeurs de *RC* mais nous n'avons trouvé aucun résultat significatif : ni la consommation des cigarettes, ni la consommation de médicaments ne constituent un facteur de risque pour le TDA/H.

En conclusion, on peut dire qu'en utilisant les différentes sources de résultats concernant le TDA/H, nous n'avons eu aucun résultat significatif quant aux relations qui existent entre cette maladie et les variables environnementales étudiées, excepté la variable sexe. Nous n'avons donc pas jugé nécessaire de faire une régression multiple entre le TDA/H et les variables environnementales.

Nous allons maintenant passer à l'étude des autres facteurs qui peuvent avoir une incidence sur la maladie ; il s'agit des facteurs génétiques que nous verrons dans la section suivante.

5.3.3 Tests génétiques d'association

Nous avons étudié dans les sections précédentes les facteurs environnementaux liés au trouble du TDA/H. Dans cette section, nous allons privilégier l'étude des facteurs génétiques en utilisant les tests du TDT et du QTDT présentés aux chapitres 2 et 4.

Les données génétiques ont été obtenues par l'analyse d'ADN sanguin chez les enfants ayant été diagnostiqués positifs par rapport à l'un des troubles d'inattention, hyperactivité/impulsivité, ou le TDA/H, et leurs parents.

Comme nous travaillons sur des données génétiques, nous utilisons le terme « phénotypes » pour désigner les troubles étudiés et les scores proviennent de la cotation de l'échelle de Conners (version parent et enseignant).

i) Résultats du TDT

On a déjà présenté au chapitre 2 le test du TDT. Celui-ci consiste à analyser la liaison entre un marqueur et le locus de la maladie. Ce test se base sur le fait que le plus souvent des parents hétérozygotes transmettront à leurs enfants atteints l'allèle associé à la maladie plutôt que l'allèle non associé.

Cinq gènes candidats ont été identifiés pour les troubles d'hyperactivité. Cependant, pour notre travail, seulement les données sur un seul gène candidat ont été mises à notre disposition. Pour des raisons de confidentialité, nous nommerons ce gène par *DG*. Il s'agit d'un neurotransmetteur pour lequel nous avons deux marqueurs bialléliques que nous nommerons aussi pour les mêmes raisons de confidentialité par MM_1 et MM_2. Les allèles respectifs de ces marqueurs sont : G,A et C,T. Ce qu'on va tester avec le TDT est de savoir, si l'un, ou les deux marqueurs sont liés et associés au gène *DG* qu'on soupçonne jouer un rôle dans l'hyperactivité des enfants.

Dans ce qui suit, nous allons donner les résultats du test du TDT appliqué aux données sur les individus qui constituent notre base de données : l'enfant malade et ses deux parents (des trios). Pour classer les enfants selon qu'ils présentent les troubles d'inattention, d'hyperactivité, ou le type combiné (TDA/H), on s'est appuyé sur les résultats du diagnostic du *DISC4* (tableau 5.13).

Nous allons faire le test du TDT en plusieurs temps en combinant les phénotypes et puis en considérant chaque phénotype séparément. Ce test (le TDT) a été fait avec le module Genetics du logiciel SAS.

Tableau 5.18 Répartition des enfants selon le phénotype

Phénotypes	Nombre d'enfants	Pourcentage
TDA/H	137	50,37
Inattention	98	36,03
Hyperactivité/Impulsivité	37	13,60
Total	272	100

Dans cette étape, nous allons faire le test du TDT en considérant comme malade, chaque enfant ayant été diagnostiqué positif pour au moins un des phénotypes présentés dans le tableau 5.18, soit au total de 272 enfants malades. Comme le TDT utilise nécessairement toute l'information sur les génotypes des parents et de l'enfant, nous avons alors supprimé du test toutes les familles ayant des données manquantes (soit un total de 159 familles). Par conséquent, le test du TDT a été établi sur 113 familles ayant un enfant déclaré « positif » (malade) par rapport à l'un des trois phénotypes.

Tableau 5.19 Test du TDT pour les enfants atteints d'un des trois phénotypes

Locus	χ^2	ddl	P-valeur
MM_1	1,832	1	0,176
MM_2	0,952	1	0,329

Dans le tableau 5.19, on voit que les résultats obtenus avec ces familles pour les marqueurs MM_1 et le MM_2 ne sont pas concluants puisque les P-valeurs sont grandes.

Maintenant nous allons refaire le test séparément pour chaque groupe d'enfants ayant un des trois phénotypes étudiés. Pour les enfants atteints du TDA/H, après suppression des

familles présentant des données manquantes, l'échantillon est de 59 familles ayant un enfant atteint du TDA/H. Les résultats ainsi obtenus sont résumés dans le tableau 5.20.

Tableau 5.20 Test du TDT pour les enfants souffrant du TDA/H

Locus	χ^2	ddl	P-valeur
MM_1	1,961	1	0,161
MM_2	1,164	1	0,281

Les P-valeurs obtenues ont légèrement diminué, mais restent néanmoins élevées pour conclure à une liaison et une association entre les deux marqueurs et le gène candidat.

Les tests du TDT appliqués aux enfants atteints séparément de l'inattention et de l'hyperactivité ne sont pas interprétables. L'échantillon pour l'hyperactivité n'est que de 18 enfants malades et, pour l'inattention, il est de 51. Les résultats sont présentés aux tableaux 5.21 et 5.22.

Tableau 5.21 Test du TDT pour les enfants atteints de l'Inattention

Locus	χ^2	ddl	P-valeur
MM_1	0,089	1	1,000
MM_2	0,000	1	0,766

Tableau 5.22 Test du TDT pour les enfants atteints d'Hyperactivité/Impulsivité

Locus	χ^2	ddl	P-valeur
MM_1	0	1	1
MM_2	0	1	1

On conclut que le test du TDT ne donne aucune association ni liaison significatives entre les deux marqueurs et le TDA/H.

La valeur phénotypique étant cependant continue, il résulte une perte d'information de la caractérisation en malades/non malades des scores mesurés. Ainsi, il serait intéressant de faire une analyse similaire mais en gardant les scores continus. Ceci se fait avec le QTDT dont les résultats vont être présentés ci-dessous.

ii) Résultats du QTDT

Dans cette section, nous allons baser notre étude sur une échelle continue des phénotypes pour expliquer leurs variations. Le test utilisé est le QTDT présenté au chapitre 4. Pour faire ce test, nous avons utilisé le logiciel SAS pour mettre nos données sous forme de fichiers spécifiques et, pour les analyses, nous avons fait appel au logiciel de Abecasis (voir Center Of Statistical Genetics, 2007) proposé dans son article.

Pour chaque test du QTDT, le logiciel a besoin de deux fichiers, un pour la description des variables (extension .dat) : le marqueur, le phénotype ainsi que les covariables s'il y a lieu, et l'autre comprenant les données (extension .txt). Avant d'effectuer le test du QTDT, le logiciel s'assure lui-même de la validité des données, c'est-à-dire qu'il vérifie la conformité du génotype de chaque enfant par rapport à ses parents, par exemple dans le cas du marqueur MM_1, si le père et la mère ont respectivement les génotypes AA et GG, l'enfant doit nécessairement recevoir les allèles A et G respectivement du père et de la mère. Il vérifie aussi que les génotypes des parents sont hétérozygotes, une des conditions de la réalisation du

test. Dans le cas de la non-conformité des données, le test ne peut pas fonctionner, nous devons alors apporter les corrections nécessaires (corriger les génotypes si possible, sinon on supprime les familles posant des problèmes) et le faire fonctionner à nouveau.

Nous avons vu dans le chapitre 4 que le QTDT peut faire plusieurs tests, dont le test sur l'association et la liaison génétique. Dans ce qui suit, on se propose dans une première étape de tester l'association allèlique pour chaque marqueur avec chacun des trois phénotypes par l'utilisation du modèle sur la moyenne. À la seconde étape, nous allons tester la liaison génétique en utilisant le modèle sur les variances. Ces étapes seront faites séparément sur les scores des phénotypes fournis par les deux types de questionnaires : Conners-parents et Conners-enseignants, puis on procède à la comparaison des résultats, c'est-à-dire voir si les résultats des deux échelles concordent ou non.

Les résultats du QTDT sont basés sur un total de 306 familles [12]. Les familles dont le génotype de l'enfant ne concordent pas avec celui des parents ou celles qui ne sont pas des trios ont été supprimées du test, soit un total de 8 familles. Néanmoins, comme nous avons jumelé le fichier comprenant les données sur les génotypes avec celui comprenant les données sur les scores des phénotypes obtenus avec les Conners enseignant ou parent, nous n'avons donc pas nécessairement l'information pour les deux fichiers ; pour certaines familles l'information sur le phénotype de l'enfant existe mais pas le génotype et vice versa. L'ensemble de ces familles ainsi que celles où les parents ne sont pas hétérozygotes, sont supprimées du test.

Comme on l'a souligné dans le chapitre 4, l'association est modélisée dans le terme de la moyenne donc le modèle évalué par le logiciel est le suivant :

---

[12] La taille de l'échantillon varie en fonction de la base de données (type de questionnaire) sur laquelle nous travaillons et que nous jumelons avec d'autres pour avoir l'ensemble des variables utiles à l'étude souhaitée.

Modèle restreint :

$$\mu_i = \mu + a_b A_i^b .$$

Modèle complet :

$$\mu_i = \mu + a_b A_i^b + a_w \left(A_i - A_i^b \right).$$

où , rappelons le : μ est la moyenne, $a_b A_i^b$ est la composante de l'association inter-familles, $a_w \left(A_i - A_i^b \right)$ est la composante de l'association intra-familles.

Ainsi, chaque phénotype est testé pour l'association avec chaque allèle du marqueur considéré. Seules les familles ayant été correctement génotypées, et dont les parents sont hétérozygotes sont incluses dans le test, ce qui a réduit considérablement le nombre de familles de notre échantillon. Finalement nous avons travaillé avec des échantillons, dont la taille varie entre 33 et 45 familles, selon le phénotype et le type de Conners utilisés. Le test du QTDT a été fait en considérant toujours les mêmes phénotypes : l'hyperactivité/impulsivité, l'inattention et le TDA/H.

Pour le marqueur MM_1, nous avons mis les P-valeurs obtenues pour l'ensemble des phénotypes et pour chacun des questionnaires (voir tableau 5.23)

Tableau 5.23 P-valeurs du QTDT entre les phénotypes et le marqueur MM_1

Mesure d'association

Phénotypes	Conners-parents	Conners-enseignant
TDA/H	0,4459	0,3470
Hyperactivité-Impulsivité	0,3526	0,1126
Inattention	0,7582	0,8186

Les résultats donnent des P-valeurs très grandes qui ne nous permettent pas de conclure à une association entre le marqueur MM_1 et un des phénotypes.

Pour le marqueur MM_2, comme pour le marqueur précédent, on n'a pas obtenu de résultats significatifs par rapport à l'association avec un des phénotypes (tableau 5.24).

Tableau 5.24 P-valeurs du QTDT entre les phénotypes et le marqueur MM_2

Mesure d'association

Phénotypes	Conners-parents	Conners-enseignant
TDA/H	0,1298	0,6396
Hyperactivité-Impulsivité	0,3526	0,2996
Inattention	0,2114	0,7230

En conclusion, nous pouvons dire qu'il n'y a pas d'association entre les deux marqueurs MM_1 et MM_2 et le TDA/H .

Nous allons présenter dans cette dernière étape de notre analyse le test de liaison entre les marqueurs MM_1 et MM_2 et la maladie.

Nous avons vu que le test de liaison est modélisé par les variances, comme suit :

Modèle restreint :

$$\mu_i = \mu\,,\ \sigma^2 = \sigma_s^2 + \sigma_e^2\,.$$

Modèle complet :

$$\mu_i = \mu\,,\ \sigma^2 = \sigma_a^2 + \sigma_s^2 + \sigma_e^2\,.$$

L'analyse de liaison a donné pour les deux scores de Conners, parent et enseignant, des résultats similaires avec des P-valeurs proches de 1 (tableaux.5.25 et 5.26), qui nous amènent à conclure qu'il n'y a pas de liaison entre les deux marqueurs étudiés et la maladie.

Tableau 5.25 P-valeurs du QTDT entre les phénotypes et le marqueur MM_1.
Analyse de liaison

Phénotypes	Conners-parents	Conners-enseignant
TDA/H	1,00	1,00
Hyperactivité-Impulsivité	1,00	1,00
Inattention	1,00	1,00

Tableau 5.26 P-valeurs du QTDT entre les phénotypes et le marqueur MM_2
Analyse de liaison

Phénotypes	Conners-parents	Conners-enseignant
TDA/H	1,00	1,00
Hyperactivité-Impulsivité	1,00	1,00
Inattention	1,00	1,00

Au regard de ces résultats, on peut conclure qu'à travers l'étude du seul gène candidat étudié, les résultats des deux tests étudiés le TDT et le QTDT ne sont pas différents. Néanmoins, cette conclusion se base sur l'étude d'un seul gène, ce qui ne nous permet pas de conclure quant à la similarité des résultats des deux tests présentés.

CONCLUSION

Notre étude a porté sur la comparaison des résultats de deux tests génétiques pour l'identification de gènes responsables d'une maladie à travers d'une étude d'association et de liaison génétique. Le premier est le TDT et utilise des caractères binaires, alors que le second (QTDT) est basé sur des scores mesurés sur une échelle continue. Après avoir introduit les notions théoriques sur lesquelles reposent ces deux tests, nous avons, en utilisant des données génétiques réelles, appliqué chacun des deux tests et avons comparé les résultats obtenus.

Cette comparaison est basée sur les données d'enfants atteints du trouble d'hyperactivité connu sous l'appellation TDA/H (trouble du déficit de l'attention, avec ou sans hyperactivité). Le diagnostic donné par le *DISC4* classe les enfants selon qu'ils sont atteints par un des trois troubles classés dans le TDA/H (inattention, hyperactivité/impulsivité, ou le type mixte) ou selon qu'ils ne sont pas atteints. En parallèle, il y a deux autres outils qui, eux, mesurent sur une échelle continue l'ampleur des différents troubles. Il s'agit de l'échelle Conners versions parent et enseignant.

Avant de commencer nos analyses, nous avons d'abord cherché à savoir si les trois différentes sources de diagnostic classaient les enfants de la même manière. Pour cela, nous avons comparé les scores des deux échelles de Conners pour les trois phénotypes étudiés (l'inattention, l'hyperactivité/impulsivité et le TDA/H). Les résultats sont très significatifs. Cela veut dire que les deux échelles ne donnent pas toujours des scores similaires pour le même enfant. Nous avons alors été amenés à utiliser les scores des deux échelles tout au long de notre travail et à vérifier les résultats en se basant sur ces deux échelles.

De la même façon que nous avons comparé les résultats sur les scores des deux échelles, nous avons aussi comparé les résultats du *DISC4* avec l'échelle de Conners, et avons abouti aux mêmes résultats. Ce qui veut dire qu'il y a une différence significative entre

les différents outils de diagnostic. Il y a donc des enfants qui sont classés atteints d'un des trois troubles par le diagnostic établi par le *DISC4*, mais qui ont de faibles scores sur les deux échelles de Conners.

Nous avons poursuivi notre analyse en étudiant les facteurs environnementaux pouvant avoir une relation avec une des formes du TDA/H. Cette étude avait comme objectif de : i) déterminer les variables reliées au score du trouble étudié et établir ainsi un modèle de régression multiple pour le score étudié et ii) faire ressortir les variables environnementales à intégrer dans le test du QTDT. Cette étude a été faite sur un grand nombre de variables, tant quantitatives, que qualitatives, comme, par exemple, le sexe de l'enfant, son quotient intellectuel, la consommation de l'alcool par la mère durant la grossesse, le niveau de scolarité des parents, etc. ainsi que sur les trois outils de diagnostic (*DISC4* et les deux Conners). Les résultats n'ont fait ressortir que la variable sexe, car en moyenne le trouble touche plus sévèrement les filles que les garçons (résultat donné par les deux échelles de Conners). Comme nous n'avons trouvé qu'une seule variable explicative des scores mesurant le TDA/H, nous n'avons pas eu à déterminer un modèle de régression multiple.

Enfin, la dernière phase de nos analyses a été consacrée aux deux tests : le TDT et le QTDT, pour lesquels nous avons introduit les données génétiques des enfants et des parents ainsi que la variable «sexe de l'enfant» (pour le QTDT seulement). Afin d'effectuer les deux tests, nous avons programmé dans le module *Genetics* du logiciel SAS (pour le TDT) ainsi que dans le logiciel SAS pour mettre les données sous une forme qui soit conforme au logiciel du QTDT tel que proposé par Abecasis et al. (2000).

Les résultats des deux tests sont similaires : les deux tests n'ont indiqué ni association ni liaison génétique quant au gène candidat et la maladie. Cela confirme des études de la littérature sur le même gène qui ont fait appel à des méthodes différentes.

APPENDICE A

DISC4

LISTE DES DIAGNOSTICS DU DISC 4.0

1) 300.23 Social Phobia
2) 309.21 Separation Anxiety Disorder
3) 300.29 Specific Phobia
4) 300.01 Panic Disorder
5) Agoraphobia
6) 300.02 Generalized Anxiety Disorder
7) 313.23 Selective Mutism
8) 300.3 Obsessive-Compulsive Disorder
9) 309.81 Post Traumatic Stress Disorder
10) 307.1 Anorexia Nervosa
11) 307.51 Bulimia Nervosa
12) 307.6 Enuresis (nocturnal)
13) 307.6 Enuresis (diurnal)
14) 307.6 Encopresis (diurnal)
15) 307.6 Encopresis
16) 307.23 Tourette's Disorder
17) 307.22 Chronic Motor or Vocal Tic Disorder
18) Transient Tic Disorder
19) 307.52 Pica
20) 312.39 Trichotillomania
21) Major Depressive Episode
22) 300.4 Dysthymic Disorder
23) Manic Episode
24) Hypomanic Episode
25) Schizophrenia
26) 314.0X Attention-Deficit/Hyperactivity Disorder
27) 313.81 Oppositional Defiant Disorder
28) 312.8 Conduct Disorder
29) Alcohol Abuse
30) Alcohol Dependence
31) Nicotine Dependence
32) Marijuana Abuse
33) Marijuana Dependence
34) Substance Abuse
35) Substance Dependence

APPENDICE B

CONNERS ENSEIGNANT

N°ID: | | | | | |

DATE:_____/_____/_____
(jour/mois/année)

Nom de l'enfant : _____ **Degré scolaire :** _____

Nom de l'école : _____

Nom de l'enseignant(e) : _____

Échelle modifiée de Conners
- ENSEIGNANT -

Vous trouverez ci-dessous une liste de difficultés que les enfants peuvent avoir durant la classe. Veuillez évaluer chaque énoncé en vous référant aux comportements de l'enfant au cours du dernier mois. Pour chaque énoncé demandez-vous dans quelle mesure ce comportement peut être considéré comme un problème. Encerclez la meilleure réponse pour chacun d'eux. Si le comportement ne s'est jamais produit ou très peu souvent, encerclez le 0. Si le comportement est très problématique ou se produit très souvent, encerclez le 3. Vous encerclez le 1 ou le 2 si votre évaluation se situe entre les deux. Veuillez, s'il vous plaît, répondre à toutes les questions, en prenant soin de n'encercler qu'une seule réponse par question.

		Absolument vrai (jamais)	Un peu vrai (à l'occasion)	Plutôt vrai (souvent)	Très vrai (très souvent)
1.	Vous affronte.	0	1	2	3
2.	Agité(e) dans le sens qu'il/elle se tortille sur place.	0	1	2	3
3.	Oublie les choses qu'il/elle a déjà apprises.	0	1	2	3
4.	Semble ne pas être accepté(e) par le groupe.	0	1	2	3
5.	Se sent facilement blessé(e).	0	1	2	3
6.	Est un(e) perfectionniste.	0	1	2	3
7.	Fait des crises de colère; a des comportements explosifs, imprévisibles.	0	1	2	3
8.	Excitable, impulsif(ive).	0	1	2	3
9.	Ne porte pas assez attention aux détails ou fait des erreurs d'inattention dans ses devoirs, ses travaux ou autres activités.	0	1	2	3
10.	Insolent(e).	0	1	2	3
11.	Toujours en mouvement, semble bouger malgré lui/elle.	0	1	2	3
12.	A de la difficulté à s'engager dans des tâches qui demandent un effort mental soutenu (comme les travaux scolaires ou les devoirs à la maison), par exemple évite de s'y engager ou est réticent(e).	0	1	2	3
13.	Est l'un(e) des derniers(ères) à être choisi(e) dans les équipes ou les jeux.	0	1	2	3
14.	Est un(e) enfant émotif(ve).	0	1	2	3
15.	Tout doit être comme il/elle veut.	0	1	2	3
16.	Agité(e) ou trop actif(ve).	0	1	2	3
17.	N'arrive pas à terminer ce qu'il/elle commence.	0	1	2	3
18.	Ne semble pas écouter ce qu'on lui dit.	0	1	2	3
19.	Affronte activement les adultes et refuse de répondre à leurs demandes.	0	1	2	3
20.	Se lève de sa chaise en classe ou dans d'autres situations où il/elle devrait rester assis(e).	0	1	2	3
21.	Faible en orthographe.	0	1	2	3
22.	N'a pas d'ami(e)s.	0	1	2	3
23.	Timide, facilement effrayé(e).	0	1	2	3

N°ID: | | | | | |

		Absolument vrai (jamais)	Un peu vrai (à l'occasion)	Plutôt vrai (souvent)	Très vrai (très souvent)
24.	Toujours en train de vérifier les choses, encore et encore.	0	1	2	3
25.	Pleure souvent et facilement.	0	1	2	3
26.	Inattentif(ive), facilement distrait(e).	0	1	2	3
27.	A de la difficulté à organiser ses tâches ou ses activités.	0	1	2	3
28.	A de la difficulté à maintenir son attention dans les tâches ou les jeux.	0	1	2	3
29.	A de la difficulté à attendre son tour.	0	1	2	3
30.	Est sous la norme en lecture.	0	1	2	3
31.	Ne sait pas comment se faire des ami(e)s.	0	1	2	3
32.	Sensible à la critique.	0	1	2	3
33.	Semble faire trop attention aux détails.	0	1	2	3
34.	Ne tient pas en place.	0	1	2	3
35.	Dérange les autres enfants.	0	1	2	3
36.	Parle de façon excessive.	0	1	2	3
37.	Argumente avec les adultes.	0	1	2	3
38.	Ne peut rester tranquille.	0	1	2	3
39.	Court ou grimpe partout de façon excessive où ce n'est pas approprié.	0	1	2	3
40.	Manque d'intérêt pour ses travaux scolaires.	0	1	2	3
41.	A de faibles habiletés sociales.	0	1	2	3
42.	A de la difficulté à garder son calme dans ses jeux et ses loisirs.	0	1	2	3
43.	Aime que tout soit propre et soigné.	0	1	2	3
44.	Bouge nerveusement des mains et des pieds, se tortille sur sa chaise.	0	1	2	3
45.	Ses demandes doivent être satisfaites immédiatement – facilement frustré(e).	0	1	2	3
46.	Répond sans attendre que la question ne soit complètement formulée.	0	1	2	3
47.	Malveillant(e) ou rancunier(ère).	0	1	2	3
48.	Sa capacité d'attention est de courte durée.	0	1	2	3
49.	Perd des objets nécessaires aux tâches ou aux activités (ex.:devoirs, crayons, livres, instruments ou jouets).	0	1	2	3
50.	Porte seulement attention aux choses qui l'intéressent vraiment.	0	1	2	3
51.	Timide, en retrait.	0	1	2	3
52.	Sa distraction ou sa brève capacité d'attention est problématique.	0	1	2	3
53.	Les choses doivent êtres faites de la même façon à chaque fois.	0	1	2	3
54.	A des changements d'humeur brusques et prononcés.	0	1	2	3
55.	Interrompt ou s'impose en dérangeant les autres (ex.: se mêlent de leurs conversations ou s'introduit dans leurs jeux).	0	1	2	3
56.	Faible en arithmétique.	0	1	2	3
57.	Ne suit pas les consignes jusqu'à la fin et n'arrive pas à terminer ses travaux scolaires (pas parce qu'il/elle s'oppose ou qu'il/elle ne comprend pas les consignes).	0	1	2	3
58.	Facilement distrait(e) par ce qui se passe autour.	0	1	2	3
59.	Jamais au repos, toujours sur ses pieds.	0	1	2	3

APPENDICE C

COTATION ENSEIGNANT

COTATION ÉCHELLE DE CONNERS RÉVISÉE
VERSION ENSEIGNANTS (59 items)

Sous-échelles	Items associés
A. **Opposition** : prédisposé à enfreindre les règlements. problèmes avec l'autorité. facilement ennuyé	1, 7, 10, 19, 37, 47
B. **Problèmes cognitifs/inattention** : tendance à être inattentif. problèmes sur le plan de l'organisation, difficulté à terminer des tâches, problèmes de concentration	3, 12, 17, 21, 30, 40, 49, 56
C. **Hyperactivité** : difficulté à demeurer assis pendant de longues périodes, agité et impulsif	2, 8, 11, 20, 29, 39, 42
D. **Anxiété-Gêne** : quantité anormale de peurs et d'inquiétudes ; porté à être émotif et sensible à la critique, anxieux dans des situations non-familières, timide et isolé	5, 14, 23, 25, 32, 51
E. **Perfectionnisme** : objectifs fixés plutôt élevés. très fastidieux en ce qui concerne sa façon de faire les choses. obsessif à propos de son travail	6, 15, 24, 33, 43, 53
F. **Problèmes sociaux** : prédisposé à se percevoir comme ayant peu d'amis, faible estime de soi et manque de confiance en soi. socialement détaché de ses pairs	4, 13, 22, 31, 41
H. **Index du trouble déficitaire de l'attention/hyperactivité de Conners** : identifie les enfants et adolescents "à risque" pour le trouble déficitaire de l'attention/hyperactivité	8, 17, 26, 35, 38, 44, 48, 50, 52, 55, 57, 59
I. **Index global de Conners - Agité-Impulsif** : indique de l'agitation, de l'impulsivité et de l'inattention chez l'enfant ou l'adolescent	8, 16, 17, 26, 34, 35
J. **Index global de Conners - Labilité émotionnelle** : personnes prédisposées à avoir plus de réponses et de comportements émotionnels (pleurer. être en colère, etc.) que la normale	7, 25, 45, 54
K. **Index global de Conners total** : ce score reflète un comportement général problématique. Des scores élevés tendent à indiquer de l'hyperactivité, mais souvent les problèmes s'étendent au-delà de l'hyperactivité	7, 8, 16, 17, 25, 26, 34, 35, 45, 54

L. DSM-IV - **Inattention** : des scores élevés indiquent une correspondance au-dessus de la moyenne avec les critères du DSM-IV pour le diagnostic du trouble déficitaire de l'attention/hyperactivité, avec prédominance de l'inattention	3, 9, 12, 18, 27, 28, 49, 57, 58
M. DSM-IV - **Hyperactivité-Impulsivité** : des scores élevés indiquent une correspondance au-dessus de la moyenne avec les critères du DSM-IV pour le diagnostic du trouble déficitaire de l'attention/hyperactivité, avec prédominance de l'hyperactivité-impulsivité	11, 20, 29, 36, 39, 42, 44, 46, 55
N. DSM-IV **total** : des scores élevé indiquent une correspondance au-dessus de la moyenne avec les critères du DSM-IV pour le diagnostic du trouble déficitaire de l'attention/hyperactivité, du type combiné	3, 9, 11, 12, 18, 20, 27, 28, 29, 36, 39, 42, 44, 46, 49, 55, 57, 58

Scores bruts

Pour connaître les scores bruts de chaque sous-échelle, il suffit de faire la somme des scores obtenus aux items associés à ces sous-échelles (au total, 13 scores bruts)

Scores t

Pour connaître les scores t des sous-échelles, il faut encercler les scores bruts de ces sous-échelles dans la colonne appropriée sur la "feuille de profil" (un côté pour les garçons, l'autre pour les filles), selon la sous-échelle et selon l'âge de l'enfant (colonne 1 : 3 à 5 ans, colonne 2 : 6 à 8 ans, colonne 3 : 9 à 11 ans, colonne 4 : 12 à 14 ans, colonne 5 : 15 à 17 ans). Après avoir encerclé les scores bruts sur la feuille de profil, il faut les relier par des lignes droites afin d'avoir une bonne idée du profil de l'enfant.

Données manquantes

Si certains items n'ont pas été complétés, il faut inscrire "0" à ces items. Le score t s'en retrouvera automatiquement diminué, mais il est possible de corriger ce biais en multipliant le score brut obtenu par le nombre total d'items de l'échelle, et de le diviser ensuite par le nombre total d'items qui ont été répondus. Par exemple, si l'enfant a obtenu un score brut de 8 à une échelle particulière, mais qu'il n'a répondu qu'à 4 des 5 items de cette échelle, ce score doit être ajusté en multipliant le score obtenu (8) par le nombre total d'items de l'échelle (5), donc 8 x 5 = 40. Le résultat obtenu doit ensuite être divisé par le nombre d'items répondus (4) afin d'obtenir le score brut ajusté (40 4 = 10). Il faut tenir compte du nombre d'items non répondus lors de l'interprétation des résultats.

Interprétation des scores t

Score t	Percentiles	
plus de 70	+ de 98%	très atypique (indique un problème important)
de 66 à 70	95-98%	modérément atypique (indique un problème important)
de 61 à 65	86-94%	légèrement atypique (possibilité d'un problème important)
de 56 à 60	74-85%	peu atypique (limite : devrait soulever des inquiétudes)
de 45 à 55	27-73%	moyenne (score typique : ne devrait pas soulever d'inquiétude)
de 40 à 44	16-26%	peu atypique (les scores faibles sont bons : pas d'inquiétude)
de 35 à 39	6-15%	légèrement atypique (les scores faibles sont bons : pas d'inquiétude)
de 30 à 34	2-5%	modérément atypique (les scores faibles sont bons : pas d'inquiétude)
moins de 30	< 2%	très atypique (les scores faibles sont bons : pas d'inquiétude)

De façon générale, les scores t plus élevés sont associés à un plus grand nombre et/ou une plus grande fréquence des problèmes rapportés. Les scores t de 65 et plus indiquent habituellement un problème clinique important. Les percentiles inscrits à côté des scores t indiquent le pourcentage d'individus qui ont un score plus faible que l'individu qui a complété le questionnaire.

APPENDICE D

CONNERS PARENTS

N°ID: | | | | | |

DATE:_____/_____/_____
(jour/mois/année)

Nom de l'enfant : _____ NAM : _____

Veuillez indiquer qui complète le questionnaire :

☐ mère ☐ père ☐ tuteur légal ☐ autre (spécifier : _____)

Échelle modifiée de Conners
- PARENT -

Vous trouverez ci-dessous une liste de difficultés que les enfants peuvent avoir. Veuillez évaluer chaque énoncé en vous référant aux comportements de votre enfant au cours du dernier mois. Pour chaque énoncé, demandez-vous dans quelle mesure ce comportement a été considéré comme un problème durant le dernier mois. Encerclez la meilleure réponse pour chacun d'eux. Si le comportement ne s'est jamais produit ou très peu souvent, encerclez le 0. Si le comportement est très problématique ou se produit très souvent, encerclez le 3. Vous encerclez le 1 ou le 2 si votre évaluation se situe entre les deux. Veuillez, s'il vous plait, répondre à toutes les questions, en prenant soin de n'encercler qu'une seule réponse par question.

		Aucunement vrai (Jamais)	Un peu vrai (A l'occasion)	Plutot vrai (Souvent)	Très vrai (Très souvent)
1.	En colère et aigri(e), plein(e) de ressentiment.	0	1	2	3
2.	A de la difficulté à faire ou à terminer ses devoirs.	0	1	2	3
3.	Toujours en mouvement, semble bouger malgré lui/elle.	0	1	2	3
4.	Timide, facilement effrayé(e).	0	1	2	3
5.	Tout doit être comme il/elle veut.	0	1	2	3
6.	N'a pas d'ami(e)s.	0	1	2	3
7.	A des maux de ventre.	0	1	2	3
8.	Se bagarre.	0	1	2	3
9.	A de la difficulté à s'engager dans des tâches qui demandent un effort mental soutenu (comme les travaux scolaires ou les devoirs à la maison), par exemple évite de s'y engager ou est réticent(e).	0	1	2	3
10.	A de la difficulté à maintenir son attention dans les tâches ou les jeux.	0	1	2	3
11.	Argumente avec les adultes.	0	1	2	3
12.	N'arrive pas à finir ses tâches.	0	1	2	3
13.	Difficile à contrôler dans les centres d'achats ou les épiceries.	0	1	2	3
14.	A peur des gens.	0	1	2	3
15.	Toujours en train de vérifier les choses, encore et encore.	0	1	2	3
16.	Perd ses amis rapidement.	0	1	2	3
17.	A des maux physiques et des douleurs.	0	1	2	3
18.	Agité(e) ou trop actif(ve).	0	1	2	3
19.	A des difficultés de concentration en classe.	0	1	2	3
20.	Ne semble pas écouter ce qu'on lui dit.	0	1	2	3
21.	N'arrive pas à garder le contrôle de l'expression de ses émotions.	0	1	2	3
22.	A besoin d'être supervisé(e) de près pour terminer ses travaux.	0	1	2	3

N°ID: | | | | | |

		Aucunement vrai (Jamais)	Un peu vrai (à l'occasion)	Plutôt vrai (souvent)	Très vrai (très souvent)
23.	Court ou grimpe partout de façon excessive où ce n'est pas approprié.	0	1	2	3
24.	A peur des situations nouvelles.	0	1	2	3
25.	Préoccupé(e) par la propreté.	0	1	2	3
26.	Ne sait pas comment se faire des ami(e)s.	0	1	2	3
27.	Avant d'aller à l'école, il/elle a des douleurs physiques ou des maux, comme des maux de ventre.	0	1	2	3
28.	Excitable, impulsif(ive).	0	1	2	3
29.	Ne suit pas les consignes jusqu'à la fin et n'arrive pas à terminer ses travaux scolaires, ni ses tâches habituelles (pas parce qu'il/elle s'y oppose ou qu'il/elle ne comprend pas les consignes).	0	1	2	3
30.	A de la difficulté à organiser ses tâches ou ses activités.	0	1	2	3
31.	Irritable.	0	1	2	3
32.	Agité(e) dans le sens qu'il/elle se tortille sur place.	0	1	2	3
33.	A peur d'être seul(e).	0	1	2	3
34.	Les choses doivent être faites de la même façon à chaque fois.	0	1	2	3
35.	N'est pas invité(e) chez ses ami(e)s.	0	1	2	3
36.	A des maux de tête.	0	1	2	3
37.	N'arrive pas à terminer ce qu'il/elle a commencé.	0	1	2	3
38.	Inattentif(ive), facilement distrait(e).	0	1	2	3
39.	Parle de façon excessive.	0	1	2	3
40.	Affronte activement les adultes ou refuse de répondre à leurs demandes.	0	1	2	3
41.	Ne porte pas assez d'attention aux détails ou fait des erreurs d'inattention dans ses devoirs, ses travaux ou autres activités.	0	1	2	3
42.	A de la difficulté à attendre en ligne ou à attendre son tour dans des situations de jeu ou de groupe.	0	1	2	3
43.	A beaucoup de peurs.	0	1	2	3
44.	A des rituels qu'il/elle doit suivre jusqu'au bout.	0	1	2	3
45.	Sa distraction ou sa brève capacité d'attention est problématique.	0	1	2	3
46.	Se plaint d'être malade même lorsqu'il/elle n'a rien.	0	1	2	3
47.	Fait des crises de colère.	0	1	2	3
48.	Distrait(e) lorsqu'on lui donne des consignes pour faire quelque chose.	0	1	2	3
49.	Interrompt ou s'impose en dérangeant les autres (ex.: se mêle de leurs conversations ou s'introduit dans leurs jeux).	0	1	2	3
50.	Oublieux(se) dans les activités quotidiennes.	0	1	2	3
51.	N'arrive pas à comprendre l'arithmétique.	0	1	2	3
52.	Court partout entre deux bouchées lors des repas.	0	1	2	3
53.	A peur du noir, des animaux ou des insectes.	0	1	2	3
54.	Se fixe des objectifs très élevés.	0	1	2	3

N°ID: | | | | | |

		Aucunement vrai (jamais)	Un peu vrai (à occasions)	Plutôt vrai (souvent)	Très vrai (très souvent)
55.	Bouge nerveusement des mains et des pieds, se tortille sur sa chaise.	0	1	2	3
56.	Sa capacité d'attention est de courte durée.	0	1	2	3
57.	A les nerfs à fleur de peau ou est facilement exaspéré(e) (irrité-e) par les autres.	0	1	2	3
58.	A une écriture peu soignée.	0	1	2	3
59.	A de la difficulté à garder son calme dans ses jeux et ses loisirs.	0	1	2	3
60.	Timide, en retrait.	0	1	2	3
61.	Tient les autres comme responsables pour ses erreurs ou ses mauvaises conduites.	0	1	2	3
62.	Ne tient pas en place.	0	1	2	3
63.	Désordonné(e) ou mal organisé(e) à la maison ou à l'école.	0	1	2	3
64.	Devient dérangé(e) ou troublé(e) si quelqu'un réarrange ses choses.	0	1	2	3
65.	S'accroche à ses parents ou à d'autres adultes.	0	1	2	3
66.	Dérange les autres enfants.	0	1	2	3
67.	Fait par exprès des choses qui exaspèrent (irritent) les autres.	0	1	2	3
68.	Ses demandes doivent être satisfaites immédiatement – facilement frustré(e).	0	1	2	3
69.	S'applique seulement pour les choses qui l'intéressent beaucoup.	0	1	2	3
70.	Malveillant(e) ou rancunier(ère).	0	1	2	3
71.	Perd des objets nécessaires aux tâches ou aux activités (ex.: devoirs, crayons, livres, instruments ou jouets).	0	1	2	3
72.	Se sent inférieur(e) aux autres.	0	1	2	3
73.	Semble tout le temps fatigué(e) ou au ralenti.	0	1	2	3
74.	Faible en orthographe.	0	1	2	3
75.	Pleure souvent et facilement.	0	1	2	3
76.	Se lève de sa chaise en classe ou dans d'autres situations où il/elle devrait rester assis(e).	0	1	2	3
77.	A des changements d'humeur brusques et prononcés.	0	1	2	3
78.	Facilement frustré(e) quand il/elle fait un effort.	0	1	2	3
79.	Facilement distrait(e) par ce qui se passe autour.	0	1	2	3
80.	Répond sans attendre que la question ne soit complètement formulée.	0	1	2	3

APPENDICE E

COTATION PARENTS

COTATION ÉCHELLE DE CONNERS RÉVISÉE
VERSION PARENTS (80 items)

Sous-échelles	Items associés
A. Opposition : prédisposé à enfreindre les règlements, problèmes avec l'autorité, facilement ennuyé	1, 8, 11, 21, 31, 40, 57, 61, 67, 70
B. Problèmes cognitifs/inattention : tendance à être inattentif, problèmes sur le plan de l'organisation, difficulté à terminer des tâches, problèmes de concentration	2, 9, 12, 19, 22, 29, 41, 50, 51, 58, 71, 74
C. Hyperactivité : difficulté à demeurer assis pendant de longues périodes, agité et impulsif	3, 13, 23, 28, 32, 42, 52, 59, 80
D. Anxiété-Gêne : quantité anormale de peurs et d'inquiétudes ; porté à être émotif et sensible à la critique, anxieux dans des situations non-familières, timide et isolé	4, 14, 24, 33, 43, 53, 60, 65
E. Perfectionnisme : objectifs fixés plutôt élevés, très fastidieux en ce qui concerne sa façon de faire les choses, obsessif à propos de son travail	5, 15, 25, 34, 44, 54, 64
F. Problèmes sociaux : prédisposé à se percevoir comme ayant peu d'amis, faible estime de soi et manque de confiance en soi, socialement détaché de ses pairs	6, 16, 26, 35, 72
G. Psychosomatisation : quantité anormale de douleurs et de maux rapportés	7, 17, 27, 36, 46, 73
H. Index du trouble déficitaire de l'attention/hyperactivité de Conners : identifie les enfants et adolescents "à risque" pour le trouble déficitaire de l'attention/hyperactivité	9, 19, 29, 38, 45, 48, 55, 56, 63, 69, 76, 78
I. Index global de Conners - Agitation-Impulsivité : indique de l'agitation, de l'impulsivité et de l'inattention chez l'enfant ou l'adolescent	18, 28, 37, 38, 62, 66, 68
J. Index global de Conners - Labilité émotionnelle : personnes prédisposées à avoir plus de réponses et de comportements émotionnels (pleurer, être en colère, etc.) que la normale	47, 75, 77

K. **Index global de Conners total** : ce score reflète un comportement général problématique. Des scores élevés tendent à indiquer de l'hyperactivité, mais souvent les problèmes s'étendent au-delà de l'hyperactivité	18, 28, 37, 38, 47, 62, 66, 68, 75, 77
L. **DSM-IV - Inattention** : des scores élevés indiquent une correspondance au-dessus de la moyenne avec les critères du DSM-IV pour le diagnostic de trouble déficitaire de l'attention/hyperactivité, avec prédominance de l'inattention	9, 10, 20, 29, 30, 41, 50, 71, 79
M. **DSM-IV - Hyperactivité-Impulsivité** : des scores élevés indiquent une correspondance au-dessus de la moyenne avec les critères du DSM-IV pour le diagnostic de trouble déficitaire de l'attention/hyperactivité, avec prédominance de l'hyperactivité-impulsivité	3, 23, 39, 42, 49, 55, 59, 76, 80
N. **DSM-IV total** : des scores élevé indiquent une correspondance au-dessus de la moyenne avec les critères du DSM-IV pour le diagnostic du trouble déficitaire de l'attention/hyperactivité, du type combiné	3, 9, 10, 20, 23, 29, 30, 39, 41, 42, 49, 50, 55, 59, 71, 76, 79, 80

Scores bruts

Pour connaître les scores bruts de chaque sous-échelle, il suffit de faire la somme des scores obtenus aux items associés à ces sous-échelles (au total, 14 scores bruts)

Scores t

Pour connaître les scores t des sous-échelles, il faut encercler les scores bruts de ces sous-échelles dans la colonne appropriée sur la "feuille de profil" (un côté pour les garçons, l'autre pour les filles), selon la sous-échelle et selon l'âge de l'enfant (colonne 1 : 3 à 5 ans, colonne 2 : 6 à 8 ans, colonne 3 : 9 à 11 ans, colonne 4 : 12 à 14 ans, colonne 5 : 15 à 17 ans). Après avoir encerclé les scores bruts sur la feuille de profil, il faut les relier par des lignes droites afin d'avoir une représentation graphique du profil de l'enfant.

Données manquantes

Si certains items n'ont pas été complétés, il faut inscrire "0" à ces items. Le score t s'en retrouvera automatiquement diminué, mais il est possible de corriger ce biais en multipliant le score brut obtenu par le nombre total d'items de l'échelle, et de le diviser ensuite par le nombre total d'items qui ont été répondus. Par exemple, si l'enfant a obtenu un score brut de 8 à une échelle particulière, mais qu'il n'a répondu qu'à 4 des 5 items de cette échelle, ce score doit être ajusté en multipliant le score obtenu (8) par le nombre total d'items de l'échelle (5), donc 8 x 5 = 40. Le résultat obtenu doit ensuite être divisé par le nombre d'items répondus (4) afin d'obtenir le score brut ajusté (40÷4 = 10). Il faut tenir compte du nombre d'items non répondus lors de l'interprétation des résultats.

Interprétation des scores t

Score *t*	Percentiles	
plus de 70	+ de 98%	très atypique (indique un problème important)
de 66 à 70	95-98%	modérément atypique (indique un problème important)
de 61 à 65	86-94%	légèrement atypique (possibilité d'un problème important)
de 56 à 60	74-85%	peu atypique (limite : devrait soulever des inquiétudes)

de 45 à 55	27-73%	moyenne (score typique : ne devrait pas soulever d'inquiétude)
de 40 à 44	16-26%	peu atypique (les scores faibles sont bons : pas d'inquiétude)
de 35 à 39	6-15%	légèrement atypique (les scores faibles sont bons : pas d'inquiétude)
de 30 à 34	2-6%	modérément atypique (les scores faibles sont bons : pas d'inquiétude)
moins de 30	< 2%	très atypique (les scores faibles sont bons : pas d'inquiétude)

De façon générale, les scores t plus élevés sont associés à un plus grand nombre et/ou une plus grande fréquence des problèmes rapportés. Les scores t de 65 et plus indiquent habituellement un problème clinique important. Les percentiles inscrits à côté des scores t indiquent le pourcentage d'individus qui ont un score plus faible que l'individu qui a complété le questionnaire.

APPENDICE F

ÉCHELLE DU WISC-III

WISC-III -Abrégé
Échelle d'Intelligence de Wechsler pour Enfants
Troisième Edition - Adaptation canadienne-française

N°ID:

Cahier de notation

	Année	Mois	Jour
Date de l'évaluation			
Date de naissance			
Âge :			

Sous-tests	Scores bruts	Équ.
2. Connaissances		
4. Similitudes		
7. Blocs		
8. Vocabulaire		
(12. Séquence de chiffres)		

	Score	QI/ Indice
Échelle totale		

Équivalence du QI_WISCIII

QI ≥ 130	Exceptionnel
120≤QI≤129	Supérieur
110≤QI≤119	Moyenne supérieure
90≤QI≤109	Moyenne
80≤QI≤89	Moyenne inférieure
70≤QI≤79	Limite
QI ≤ 69	Déficience

BIBLIOGRAPHIE

Abecasis, G.R., L.R. Cardon., W.O.C. Cookson. 2000. «A general test of association for quantitative traits in nuclear families», *American Journal of Human Genetics,* vol. 66, p. 279-292.

Allison, D.B. 1997. «Transmission disequilibrium tests», *American Journal of Human Genetics,* vol. 60, p. 676-690.

Almgren, P., P-O. Bendahl, H. Bengtsson, O.Hossjer, R.Perfekt. Novembre 2003. Statistics in genetics. Lund University.

Amos, C.I. 1994. «Robust variance-components approach for assessing genetic linkage in pedigrees», *American Journal of Human Genetics*, vol.54, p.535-543.

Bernard, P.M., et C. Lapointe. 1987. *Mesures statistiques en épidémiologie.* PUQ.

Bobb, A.J., A.M. Addington, E. Sidransky, M.C. Gornick, J.P.Lerch, D.K.Greenstein, L.S. Clasen, W.S. Sharp, G. Inoff-Germain, G., F. Wavrant-De Vrièze, M. Arcos-Burgos, R. E. Straub, J.A. Hardy, F.X. Castellanos, J.L. Rapoport. 2005. « Support for association between ADHD and two candidate genes : NET1 and DRD1 ». *American Journal of Medical Genetics Part B (Neuropsychiatric Genetics)*, vol. 134 B, p. 67-72.

Falconer, D.S., et T.F.C. Mackay. 1996. *Introduction to quantitative genetics.* 4è édition England : Pearson Education Limited..

_____1974. *Introduction à la génétique quantitative.* Traduction de l'anglais par Boelle H., G. Chinardet., J. koscinski., et P. Lherminier Paris : Masson et compagnie.

Fisher, R.A. 1918. «The correlation between relatives on the supposition of Mendelian inheritance». *Transactions of the Royal society of the Edinburgh,* vol.52, p 399-433.

Fulker, D.W., et S.S. Cherny, P.C. Sham, et J.K. Hewitt. 1999. «Combined linkage and association sib-pair analysis for quantitative traits». *American Journal of Human Genetics*, vol.64, p. 259-267.

Hartl, L. 1994. *Génétique des populations.* Traduit de l'américain par Borot, N. Paris : Médecine-Sciences Flammarion.

Haseman, J.K., et R.C. Elston. 1972. «The investigation of linkage between a quantitative trait and a marker locus». *Behavior Genetics*, vol. 2, no 1, p. 3-19.

Henry, J.P., et P.H.Gouyon. 2003. *Précis de génétique des populations*. Paris : Dunod.

Jacquard, A. 1977. *Concepts en génétique des populations*. Paris, New-York, Barcelone, Milan : Masson.

Lynch, M., et B.Walsh. 1998. *Genetics and analysis of quantitative traits*. USA : Sinauer.

McCulloch, C. E., Shayle R.Searle. 2001. *Generalized, Linear, and Mixed Models*. Wiley Inter-Science.

Ott, J. 1991. *Analysis of human genetic linkage*. Édition révisée. Baltimore and London : The Johns Hopkins University Press.

Ott, J. 1989. «Statistical properties of the haplotype relative risk». *Genetic Epidemiology*, vol. 6, p. 127-130.

Rabinowitz, D. 1997. «A transmission disequilibrium test for quantitative trait loci», *American Journal of Human Genetics*, vol.47, p.342-350.

Searle, S.R. 1982. *Matrix algebra useful for statistics*. John Wiley & Sons.

Sham, P.C., S.S. Cherny, S. Purcell., et J.K. Hewitt. 2000. «Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data». *American Journal Genetics*, vol.66, p.1616-16300.

Spielman, R.S., R.E. McGinnis, et W.J. Ewens. 1993. «Transmission test for linkage disequilibrium : The insulin gene region and insulin-dependent diabetes Mellitus (IDDM) ». *American Journal of Human Genetics*, vol 52, p. 506-516.

Spielman, R.S., et W.J. Ewens. 1996. «The TDT and other family-based tests for linkage diséquilibrium», *American Journal of Human Genetics*. vol.59, p.983-989.

Sprent, P., et N.C. Smeeton. 2001. *Applied nonparametric statistical methods*. 3e édition. London New York Washington : Chapman et Hall/CRC.

Vetta, A.et Courgeau, D. 2003. «Comportements démographiques et génétique du comportement». Population/4-4,Vol. 58, p. 457-488.

Wright, S. 1921a. « Systems of mating. I. The biometric relations between parents and offspring». *Genetics*, vol 6, p.111-123.

Center Of Statistical GeneticsQTDT : **http://www.sph.umich.edu/csg/abecasis/QTDT/**.

Wikipedia 2007) : (**http://fr.wikipedia.org/wiki/Chromosome**