UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ANALYSE EN COMPOSANTES PRINCIPALES D'HÉRITABILITÉ DANS LE CADRE DES ÉTUDES D'ASSOCIATION GÉNÉTIQUE DE GRANDE DIMENSION

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

MIFTAH HANANE

AOÛT 2016

REMERCIEMENTS

Mes remerciements vont tout d'abord à mon directeur de recherche, Dr Karim Oualkacha et ma codirectrice Dr Aurélie Labbe, qui m'ont fait confiance et m'ont donné la chance de réaliser ce projet de recherche. Merci Dr Karim pour votre gentillesse, votre disponibilité et votre aide précieuse dans le déroulement de ce mémoire. J'ai réellement apprécié la liberté d'action que vous m'avez accordée pour mener à bien ce projet de recherche, tout en gardant un œil critique et constructif. Merci de savoir si bien transmettre votre savoir et votre passion ; merci d'avoir cru en moi et de m'avoir toujours soutenu et encouragé. Merci Dr Aurélie pour vos remarques et conseils qui m'ont permis de perfectionner ce mémoire.

Je tiens à exprimer mes remerciements les plus sincères aux Dr Fabrice Larribe et Dr Geneviève Lefebvre pour avoir accepté de corriger ce mémoire, pour l'intérêt qu'ils y ont apporté et pour leurs conseils avisés.

Je voudrais aussi remercier mes professeurs de statistique pour ce qu'ils m'ont appris, particulièrement Dr Glenn Shoorock et Dr Serge Alalouf, ainsi que Gisèle Legault pour sa disponibilité à répondre à mes nombreuses questions.

Je dédie ce mémoire à la fin à ma famille et à mes amis. Ce mémoire est aussi le vôtre. Merci pour votre soutien, votre amour et la fierté que je peux lire dans vos yeux.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

PCH_K	Composante principale de l'héritabilité de Klei
ACP	Analyse en composantes principales
ANOVA	Analysis of variance
ApoE	Apolipoprotéine E
CP	Composante principale
ddl	Degré de liberté
DL	Déséquilibre de liaison
ECDF	Empirical cumulative distribution function
FDR	False discovery rate
FWER	Family wise error rate
GPD	Generalized pareto distribution
GWAS	Genome wide association study
LCQ	Locus de caractères quantitatifs
MAF	Minor allele frequency
MCD	Minimum covariance determinant
MD	Mahalanobis distance

MVE	Minimum volume ellipsoid
PCH	Principal components of heritability
PCH de Klei	Analyse en composantes principales basée sur l'héritabilité combinée à la pléiotropie
SNP	Single nucleotide polymorphism

RÉSUMÉ

Dans ce mémoire, nous présentons une nouvelle approche de réduction de la dimension dans un contexte d'études d'association génétique [1] avec présence de plusieurs variables réponses corrélées entre elles. Cette approche tente de résumer l'information utile de plusieurs variables (qualitatives ou quantitatives) corrélées dans un nombre réduit de nouvelles variables (scores) appelées composantes principales d'héritabilité. Ces nouveaux scores optimaux peuvent être utilisés dans des études d'association afin d'augmenter la puissance des tests statistiques appliqués pour détecter les gènes responsables des maladies complexes. Nous avons proposé une nouvelle statistique qui teste une telle association entre ces nouvelles composantes principales et une région génomique, et nous avons approximé sa distribution sous l'hypothèse nulle à l'aide des techniques de permutation. Nous avons également comparé notre nouvelle méthode avec une méthode existante (Klei *et al.*, 2008). À l'aide de simulation, nous avons montré que notre nouvelle méthode contrôle bien l'inflation de l'erreur de type 1 et elle a un gain important de puissance statistique comparée à la méthode existante. Enfin, nous avons illustré notre méthodologie à l'aide d'une application sur des données réelles collectées pour étudier la maladie d'Alzheimer.

Mots clés : Études d'association génétique, variables réponses corrélées, réduction de la dimension, composantes principales d'héritabilité, puissance statistique.

---

1. Les études d'association permettent d'évaluer statistiquement l'association entre une ou plusieurs covariables et un caractère relié à la maladie étudiée. Ce caractère mesuré peut être dichotomique (atteint, non atteint) ou bien continu (taux de sucre dans le contexte du diabète, par exemple).

INTRODUCTION

La susceptibilité individuelle aux maladies complexes, telles que le diabète de type 1 et 2, l'hypertension artérielle, l'asthme, certains cancers et la schizophrénie a une forte composante héréditaire. Les dernières avancées technologiques par rapport au séquençage du génome humain offrent la promesse de comprendre le mécanisme moléculaire de telles maladies et de fournir de nouvelles cibles thérapeutiques. Toutefois, ces efforts sont généralement entravés par le fait que ces nombreuses maladies complexes sont étiologiquement hétérogènes. Les caractères principaux utilisés dans les études d'association génétique sont généralement les résultats de nombreux gènes différents qui interagissent les uns avec les autres et avec des facteurs environnementaux. Les chercheurs, dans le cadre des maladies génétiques complexes, collectent le plus grand nombre possible de données, et se retrouvent à la fin avec une base de données de dimension $N \times q$ où N se compte en milliers et q se compte en millions. Cette richesse de l'information n'est pas facile à manipuler, d'où la nécessité d'utiliser des algorithmes de réduction de dimension tels que l'analyse en composantes principales (ACP). L'algorithme d'analyse en composantes principales dans les études d'association génétique n'a cependant pas apporté d'amélioration dans la découverte de nouvelles variantes génétiques (c'est-à-dire facteurs génétiques). Klei *et al.* (2008) proposent une approche alternative à l'ACP, qu'ils nomment analyse en composantes principales d'héritabilité. Contrairement à l'ACP, le principe de l'approche de Klei est de réduire les caractères à un seul caractère qui a une héritabilité plus élevée que toute autre combinaison linéaire des caractères. L'héritabilité est définie comme la

proportion de la variance phénotypique totale expliquée par les variantes génétiques. Par conséquent, l'association entre le facteur génétique et la composante principale d'héritabilité est souvent plus facile à détecter.
En adoptant une approche d'analyse individuelle des variantes génétiques, les études d'association génétique (noté GWAS, de l'anglais Genome wide association study) ont très bien réussi à identifier des variantes liées à des centaines de caractères et maladies humaines complexes. Cependant, les variantes identifiées dans des GWAS ne reflètent que quelque pourcentages de l'héritabilité estimée pour ces caractères complexes. En effet, cette approche est limitée à l'analyse d'une variante à la fois et seules celles qui atteignent le seuil de GWAS (p-valeur de 5×10^{-8}, correspond au seuil de Bonferroni pour un seuil de 5%) sont conservées pour des études ultérieures, laissant une grande partie de l'héritabilité des caractères inexpliquée.

De même, l'approche de Klei *et al.* (2008) a une puissance limitée pour identifier le rôle des variantes génétiques ayant un effet individuel faible. Il est probable que parmi ces variantes à faible effet individuel, il en existe certaines, qui conjointement, peuvent conférer un risque de prédisposition aux maladies. L'approche de Klei *et al.* (2008) laisse ainsi une grande partie de l'héritabilité des caractères inexpliquée.

Pour contourner ce problème, nous allons généraliser l'approche de Klei *et al.* au cas de plusieurs variantes génétiques afin d'identifier la composante linéaire des caractères maximisant la part de variance expliquée par l'effet joint des variantes incluses dans le modèle. Au cours de cette étude, nous avons cherché à évaluer l'intérêt que peut présenter l'analyse jointe de caractères quantitatifs corrélés, en mesurant le gain apporté par l'analyse jointe des variantes génétiques relativement aux analyses individuelles des variantes (PCH de Klei et analyse de GWAS).

Ce mémoire s'organise principalement autour de cinq chapitres : **le chapitre 1** permet d'introduire les notions de base et les fondements statistiques des caractères à effets quantitatifs. Nous présentons à cette occasion la notion de l'héritabilité de ces caractères et les différentes méthodes d'estimation correspondantes. Nous présentons également la problématique de tests statistiques causée par des comparaisons multiples ainsi que quelques procédures de contrôle correspondantes. **Le chapitre 2** présente l'analyse en composantes principales comme technique de réduction de dimension généralement utilisée, pour discuter ensuite de nouvelles approches de réduction de dimension basées sur l'héritabilité proposées par Ott et Rabinowitz (1999) et Klei *et al.* (2008) séparément. **Le chapitre 3** détaille spécifiquement l'approche que nous avons développée. Nous proposons aussi dans ce chapitre une procédure de test d'hypothèse pour tester l'association entre la composante principale d'héritabilité déduite de notre approche et les marqueurs génétiques tout en évitant une procédure de fractionnement compliquée. Nous approximons la distribution de la statistique de notre test sous l'hypothèse nulle en utilisant des techniques de permutation. **Le chapitre 4** est réservé aux simulations réalisées pour comparer les performances des approches existantes et l'approche PCH_g, une discussion des résultats est présentée ensuite. **Dans le chapitre 5** nous donnons un préliminaire de littérature sur la maladie d'Alzheimer et ensuite nous appliquons notre méthode à un jeu de données portant sur cette maladie afin de comparer les performances de notre méthode PCH_g avec celle de Klei. À la fin, nous terminons par une discussion des résultats.

CHAPITRE I

PRÉLIMINAIRES GÉNÉTIQUES ET STATISTIQUES

1.1 Terminologie génétique

L'objectif de cette section est d'introduire les fondements génétiques et statistiques nécessaires à la compréhension du travail de recherche réalisé dans le cadre de ce mémoire.

1.1.1 Acide désoxyribonucléique (ADN)

L'acide désoxyribonucléique (ADN) est une molécule, présente dans toutes les cellules vivantes, qui contient l'ensemble des informations nécessaires au développement et au fonctionnement d'un organisme. C'est aussi le support de l'hérédité, car il est transmis lors de la reproduction.

La structure standard de l'ADN est une double-hélice composée de deux brins complémentaires. Chaque brin d'ADN est constitué d'un enchaînement de nucléotides. On trouve quatre nucléotides différents dans l'ADN : Adénine, Guanine, Cytosine, Thymine, notés A, G, C et T respectivement.

1.1.2 Chromosomes, gènes, locus, allèles

Un individu possède, dans ses cellules, des **chromosomes** qui sont le support de son patrimoine génétique (voir figure 1.1). Les chromosomes présents dans les noyaux des cellules sont des structures filamenteuses décelables par teintures lors des divisions cellulaires. Les cellules sexuelles possèdent un ensemble de n paires de chromosomes, le nombre n étant caractéristique de chaque espèce : quatre chez la drosophile, sept chez le pois, dix chez le maïs, vingt-trois chez l'homme (Tjio et Levan, 1956). La série complète des chromosomes d'un individu constitue son caryotype (voir figure 1.2).

Un **gène** est une séquence d'ADN dont la transmission est héréditaire. Cette séquence est définie par sa fonction et sa structure. Les gènes se trouvent sur les chromosomes. Ils sont constitués d'une longue chaîne d'ADN. Un gène code généralement pour la formation d'un acide ribonucléique (ARN) et d'une protéine. On compte 20000 gènes chez les humains. Le **locus**[1] d'un gène sur un chromosome est l'endroit précis où se situe ce gène sur le chromosome. Les différentes formes possibles de l'information génétique portée à un locus sont nommées **allèles**.

1.1.3 Marqueurs génétiques

Avant de donner la définition de marqueur génétique, il est important de définir la notion de polymorphisme génétique. Les variations génétiques observées dans les populations constituent le polymorphisme génétique. Le polymorphisme peut être observé au niveau de l'individu entier, dans des formes variantes de protéines ou

1. Le pluriel de locus est loci.

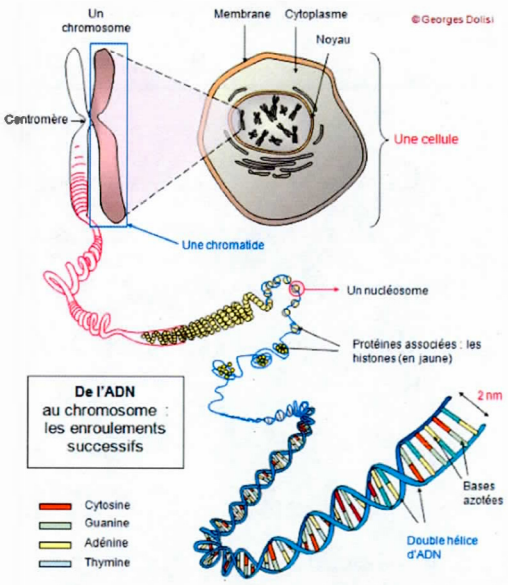


Figure 1.1: Chromosome.

Source: Goerges Dolisi (03-02-2013). *Bio-Top*. Récupéré de `http://www.associationiris.org/infos-medicales-et-traitements/les-deficits-immunitaires`.

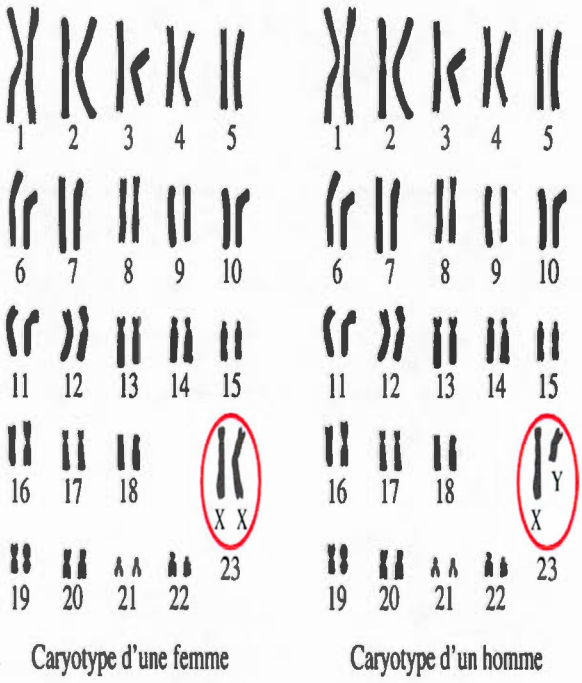


Figure 1.2: Caryotype humain.

Source: Rue des écoles. *Assistance scolaire personnalisée*. Récupéré de `http://www.assistancescolaire.com/eleve/TST2S/biologie/reviser-le-cours/chromosomes-et-caryotypes-tst2s_bio09`.

de composants des groupes sanguins, dans les caractéristiques morphologiques des chromosomes, ou au niveau de l'ADN, dans des différences nucléotidiques. Techniquement, une variation doit être présente dans au moins 1 % d'une population pour

être classée comme un polymorphisme.

Sur le plan moléculaire, on classe les polymorphismes en trois catégories : le polymorphisme de séquence, d'insertion-délétion et de nombre d'unités de répétitions dans les régions répétées (De Vienne, 1998).

Un marqueur génétique est une séquence polymorphe d'ADN aisément détectable, dont l'emplacement est connu sur un chromosome. Un marqueur génétique peut être un polymorphisme portant sur un seul nucléotide (noté SNP pour Single Nucleotide Polymorphism), et c'est ce type de polymorphisme avec lequel nous allons travailler dans ce mémoire. Un marqueur génétique peut être aussi constitué de séquences répétées de nucléotides. Dans ce cas-là on distingue les microsatellites qui sont des répétitions courtes de motifs de nucléotides, et les minisatellites qui sont des répétitions longues d'une séquence de nucléotides.

Les marqueurs génétiques représentent en quelque sorte des balises qui vont permettre la localisation des loci responsables d'une maladie (voir figure 1.3).

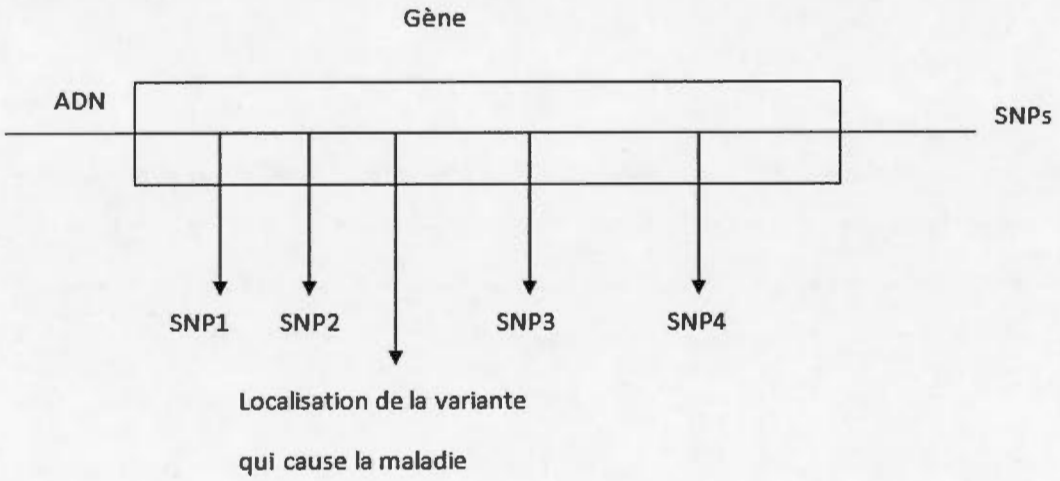


Figure 1.3: Marqueur génétique.

1.1.4 Génotype, homozygote, hétérozygote, phénotype, dominance, récessivité

Pour un locus à n allèles, la combinaison de deux allèles d'un gène qu'un individu reçoit de ses deux parents forme son **génotype**. Le génotype d'un individu pour un gène qui possède deux allèles A et a, peut donc être soit AA, Aa, ou aa. Les deux types similaires AA et aa sont dits **homozygotes**, l'autre étant dit **hétérozygote**. **Phénotype :** le caractère effectivement manifesté par un individu correspond à son phénotype, par exemple la couleur des yeux, le taux de cholestérol, l'indice de masse corporelle, etc. Le caractère observable est le produit de l'interaction de l'information génétique de l'organisme (le génotype) et du milieu dans lequel cet individu vit. **Dominance-Récessivité** : on dit que l'allèle a_i est dominant sur l'allèle a_j (ou que l'allèle a_j est récessif devant l'allèle a_i) lorsque l'action de l'allèle a_i rend inapparente celle de l'allèle a_j, quand ils sont associés à un génotype hétérozygote $(a_i a_j)$.

1.1.5 Fréquences génotypiques et fréquences alléliques

Dans une population de N individus, la fréquence génotypique d'un gène autosomique[2] est le rapport du nombre d'individus qui possèdent ce type de génotype sur le nombre total d'individus de la population. Par exemple, considérons un locus avec deux allèles A et a. Les fréquences p_{AA}, p_{Aa} et p_{aa} des génotypes AA, Aa et aa sont données dans le tableau 1.1.

La fréquence allélique est le rapport du nombre d'allèles d'une catégorie sur le nombre total d'allèles présents au locus considéré dans la population $(2 \times N)$. Les fréquences

2. Un gène autosomique est un gène situé sur un chromosome non sexuel.

Tableau 1.1: Fréquence génotypique

Génotype	Nombre d'individus	Fréquence génotypique
AA	n_1	$p_{AA} = \frac{n_1}{N}$
Aa	n_2	$p_{Aa} = \frac{n_2}{N}$
aa	n_3	$p_{aa} = \frac{n_3}{N}$
	$n_1 + n_2 + n_3 = N$	$p_{AA} + p_{Aa} + p_{aa} = 1$

des allèles peuvent être obtenues à partir des fréquences génotypiques. Supposons, par exemple, p_a la fréquence de l'allèle a et p_A la fréquence de l'allèle A. Chaque individu qui a un génotype AA possède deux exemplaires de l'allèle A et chaque individu qui a un génotype Aa en possède un seul, nous pouvons écrire :

$$\begin{aligned} p_A &= 1 \times p_{AA} + 1/2 \times p_{Aa} + 0 \times p_{aa} \\ &= P_{AA} + 1/2 \times P_{Aa}, \\ p_a &= 0 \times p_{AA} + 1/2 \times p_{Aa} + 1 \times p_{aa} \\ &= 1/2 \times p_{Aa} + p_{aa}. \end{aligned}$$

En remplaçant les fréquences génotypiques par leurs valeurs dans les fréquences alléliques, on obtient :

$$\begin{aligned} p_A &= \frac{2 \times n_1 + n_2}{2 \times N}, \\ p_a &= \frac{2 \times n_3 + n_2}{2 \times N}. \end{aligned}$$

1.1.6 Équilibre de Hardy-Weinberg

L'équilibre de Hardy-Weinberg (Lange, 2002) est un principe fondamental en génétique des populations. Il stipule que les fréquences génotypiques à un locus donné

demeurent constantes de génération en génération si les conditions suivantes sont respectées :

- la population est panmictique (c'est-à -dire les couples se forment au hasard (panmixie) et leurs gamètes[3] se rencontrent au hasard (pangamie)) ;
- la population est "infinie" (très grande : pour minimiser les variations de l'échantillonnage) ;
- il ne doit y avoir ni sélection, ni mutation, ni migration (pas de perte/gain d'allèle) ;
- les générations successives sont discrètes (pas de croisement entre générations différentes).

On dit donc de façon commune qu'une population est en équilibre de Hardy-Weinberg si la fréquence du génotype dépend uniquement des allèles. Pour un locus ayant deux allèles A et a de fréquences respectives p et q $(p+q=1)$, cet équilibre se traduit par :

$$\begin{cases} p_{AA} &= p^2, \\ p_{Aa} &= 2pq, \\ p_{aa} &= q^2, \end{cases}$$

avec

$$p^2 + q^2 + 2pq = 1.$$

L'importance de l'équilibre de Hardy-Weinberg dans une population découle du fait que la majorité des méthodes développées en statistique génétique présument un tel

3. Un gamète est une cellule à fonction reproductrice. Chez l'être humain, on distingue les gamètes mâles (spermatozoïdes) des gamètes femelles (ovules).

équilibre. Puisque plusieurs de ces méthodes ne sont pas robustes au non respect de l'équilibre de Hardy-Weinberg, il est important de tester si une population est en équilibre de Hardy-Weinberg avant de commencer une analyse. Dans l'analyse de données menée dans ce mémoire nous allons effectuer une telle analyse pour vérifier le postulat d'équilibre de Hardy-Weinberg.

Le principe du test est simple et peut être résumé en trois étapes :

1- calcul des fréquences génotypiques observées à partir des données observées;

2- calcul des fréquences génotypiques attendues selon la loi de Hardy-Weinberg;

3- comparaison des fréquences observées et attendues par un test statistique du χ^2. Le test du χ^2 teste l'hypothèse d'égalité entre la distribution observée et la distribution attendue sous l'hypothèse nulle qui stipule que le locus est en équilibre de Hardy-Weinberg.

Les fréquences observées (O_i) et attendues (E_i) sont présentés dans le tableau 1.2 où i réfère au génotype considéré.

Tableau 1.2: Fréquences observées (O_i) et attendues (E_i) des trois génotypes AA, Aa et aa pour n individus. i réfère au génotype considéré.

i	AA	Aa	aa
O_i	p_{AA}	p_{aA}	p_{aa}
E_i	p^2	$2pq$	q^2

La statistique du χ^2 correspondante est :

$$\chi^2 = \sum_i \frac{(\text{fréquences observées du génotype}_i - \text{fréquences attendues du génotype}_i)^2}{\text{fréquences attendues du génotype}_i}$$

$$= \sum_i \frac{(O_i - E_i)^2}{E_i}.$$

La valeur du χ^2 est comparée au seuil théorique d'ordre $(1-\alpha)$ de la loi de χ^2 avec un degré de liberté (noté ddl) qui est la différence entre le nombre de génotypes et le nombre d'allèles. Si le χ^2 calculé est inférieur à ce seuil théorique, on ne rejette pas l'hypothèse nulle et on conclut qu'il n'y a pas assez d'évidence pour rejeter l'équilibre de Hardy-Weinberg de la population sous étude, sinon on rejette H_0.

Exemple de l'équilibre de Hardy-Weinberg

Lors d'une étude médicale, on a déterminé le génotype de $n = 1000$ personnes (Source : Morgenthaler, 2008 p-72). Les données observées sont dans le tableau 1.3. On désire savoir si la population est en équilibre de Hardy-Weinberg :

Tableau 1.3: Exemple de données observées des trois génotypes AA, Aa et aa.

AA	Aa	aa
652	310	38

- Calcul des fréquences p et q des allèles A et a :

$$p = \left(652 + \frac{1}{2} \times 310\right)/1000 = 0.807 \text{ pour l'allèle } A,$$
$$q = \left(38 + \frac{1}{2} \times 310\right)/1000 = 0.193 \text{ pour l'allèle } a.$$

- Calcul des fréquences attendues de différentes catégories génotypiques :

$$p_{AA} = p^2 = 0.651249,$$
$$p_{Aa} = 2pq = 0.311502,$$
$$p_{aa} = q^2 = 0.037249.$$

- La statistique du test du χ^2 vaut :

$$\chi^2 = \frac{(0.652 - 0.651249)^2}{0.651249} + \frac{(0.310 - 0.311502)^2}{0.311502} + \frac{(0.038 - 0.037249)^2}{0.037249}$$
$$= 0.000023249745.$$

Puisque le seuil théorique du test est $\chi^2_{0.05,ddl=1} = 3.84$, nous avons $\chi^2 \leqslant \chi^2_{0.05,ddl=1}$, on conclut que la population est en équilibre de Hardy-Weinberg.

1.1.7 Déséquilibre de liaison

Le déséquilibre de liaison (noté DL) (Lewontin,1964) est défini comme une association non aléatoire entre les allèles de deux loci (ou plus) proches l'un de l'autre. À l'inverse, deux loci éloignés tendent à être en équilibre de liaison. Le DL joue un rôle central dans les études d'association des maladies complexes (Weiss et Clark, 2002 ; Nordborg et Tavaré, 2002) puisque les marqueurs testés sont supposés être en déséquilibre de liaison avec le locus de la maladie.

Par exemple, considérons une paire de SNPs ayant les allèles A/a et B/b. Notons p_A, p_a, p_b et p_B les fréquences alléliques correspondantes. Les haplotypes possibles sont AB, Ab, aB et ab. Soient p_{AB}, p_{Ab}, p_{aB} et p_{ab} leurs fréquences haplotypiques pour différentes combinaisons possibles des allèles. Si un individu a le génotype AA

au premier SNP et le génotype Bb au deuxième SNP, les deux haplotypes[4] possibles sont AB et Ab.

Une mesure simple de DL consiste à évaluer la différence entre la fréquence observée d'un haplotype donné et celle attendue sous l'hypothèse d'indépendance entre les loci. Cette mesure s'appelle le coefficient de déséquilibre de liaison (D) (Robbins, 1917) et sa formule est la suivante :

$$D = p_{AB} - p_A p_B = p_{ab} - p_a p_b.$$

Sous l'hypothèse d'indépendance entre les deux loci, la présence d'un allèle à un locus n'influence pas l'allèle observé dans l'autre locus. Chaque cellule du tableau 1.4 donne la fréquence de l'haplotype correspondant.

Tableau 1.4: Distribution attendue des haplotypes sous l'hypothèse d'indépendance

*Site*1 \ *Site*2	B	b	
A	$p_{AB} = p_A p_B$	$p_{Ab} = p_A p_b$	p_A
a	$p_{aB} = p_a p_B$	$p_{ab} = p_a p_b$	p_a
	p_B	p_b	1

Le tableau 1.5 donne les fréquences alléliques de la population pour deux loci sous l'hypothèse de déséquilibre de liaison. Lewontin (Lewontin, 1964) a introduit D' comme mesure de déséquilibre défini par :

4. Un haplotype est l'arrangement linéaire des allèles sur le même chromosome à deux (ou plus) loci.

Tableau 1.5: Distribution observée des haplotypes sous l'hypothèse de déséquilibre de liaison

$Site1 \setminus Site2$	B	b	
A	$p_{AB} = (p_A p_B + D)$	$p_{Ab} = (p_A p_b - D)$	p_A
a	$p_{aB} = (p_a p_B - D)$	$p_{ab} = (p_a p_b + D)$	p_a
	p_B	p_b	1

$$D' = \frac{D}{D_{max}},$$

avec :

$$D_{max} = \begin{cases} min(p_A p_b; p_a p_B) & \text{si} \quad D > 0, \\ min(p_a p_b; p_A p_B) & \text{si} \quad D < 0. \end{cases}$$

La mesure D' a la propriété de prendre des valeurs entre -1 et 1.

Une autre mesure de liaison est le coefficient de corrélation (r^2) [5] défini comme :

$$\begin{aligned} r^2 &= \frac{(p_{AB} - p_A p_B)^2}{p_A(1-p_A)p_B(1-p_B)} \\ &= \frac{D^2}{p_A p_a p_B p_b}. \end{aligned}$$

Le coefficient r^2 exprime la quantité d'information que fournit un locus sur l'autre. Des grandes/petites valeurs de D ou bien de r^2 suggèrent un fort/faible déséquilibre de liaisons, respectivement. Le déséquilibre de liaison entre les allèles d'un locus peut être aussi estimé à partir de fréquences génotypiques. En effet, dans la loi de Hardy-Weinberg, les fréquences génotypiques sont supposées être constantes de génération

5. Démonstration en annexe A.

en génération (voir sous-section 1.1.6). Le déséquilibre de liaison représente ainsi un écart par rapport à l'équilibre de Hardy-Weinberg. On écrit :

$$p_{AA} = p_A^2 + D,$$
$$p_{Aa} = 2p_A p_a - 2D,$$
$$p_{aa} = p_a^2 + D,$$

et

$$\begin{aligned} D &= p_{AA} - p_A^2 \\ &= p_{aa} - p_a^2 \\ &= 0.5(2p_A p_a - p_{Aa}). \end{aligned}$$

1.2 Caractères quantitatifs

Les caractères (phénotypes) quantitatifs sont des caractères à variation continue, par exemple l'hypertension artérielle, le taux de cholestérol, le poids, la taille, etc. La génétique quantitative suppose que des allèles à effets modérés à plusieurs loci sur le génome influencent, ensemble, la variation d'un caractère quantitatif.

1.2.1 Modèle caractère-gène

Dans cette section, nous allons présenter un modèle génétique simple (Falconer, 1975 ; Morgenthaler, 2008). Ce modèle relie le caractère Y avec un locus par :

$$Y = \mu + G + E, \tag{1.1}$$

où Y est la valeur phénotypique individuelle, μ est la moyenne de la population, G est la composante génétique qui représente l'effet du locus et E représente le terme d'erreur ou la composante environnementale. En supposant que les composantes G

et E sont indépendantes, on peut écrire la variance phénotypique de Y comme suit :

$$\sigma_Y^2 = \sigma_G^2 + \sigma_E^2.$$

L'hypothèse de non-corrélation entre E et G n'est pas toujours justifiée, mais elle est nécessaire pour simplifier le modèle.

La composante génotypique peut aussi se décomposer, de la façon suivante, en parties considérées comme indépendantes :

$$G = A + D,$$

où :

- A est l'effet additif des allèles au locus ;
- D est l'ensemble des effets de dominance entre allèles parentaux (un allèle de père et un allèle de mère).

Dans la section suivante, nous allons illustrer le modèle 1.1 ainsi que les notations d'additivité et de dominance.

1.2.2 Locus occupé par deux allèles

Certaines propriétés d'un locus à deux allèles vont être présentées dans cette sous-section, vu leur importance en génétique quantitative et leur présence dans la détermination de la variance génotypique. En effet, en conservant les notations précédentes, on peut écrire la variance génotypique comme suit :

$$\sigma_G^2 = E(G^2) - E(G)^2.$$

Cette équation fait intervenir l'espérance de la valeur génotypique, $E(G)$, que nous allons déterminer dans une première étape. Dans une deuxième étape, nous allons déterminer l'expression de la variance génotypique en termes de fréquences génotypiques.

L'espérance de valeur génotypique

Pour un locus autosomique ayant 2 formes alléliques A_1 et A_2, on note p la fréquence dans la population de l'allèle A_1 et q la fréquence dans la population de l'allèle A_2 et $p+q=1$ puisqu'il n'y a que 2 allèles. Nous supposons que la population est en équi-

Tableau 1.6: Les deux premières colonnes du tableau montrent les trois génotypes et leurs fréquences dans une population. La troisième colonne donne les valeurs génotypiques.

Génotype	Fréquence	Valeur génotypique	Fréquence×valeur
A_1A_1	p^2	$+a$	p^2a
A_1A_2	$2pq$	d	$2pqd$
A_2A_2	q^2	$-a$	$-q^2a$

libre de Hardy-Weinberg. L'espérance génotypique de la population pour le caractère considéré est obtenue on multipliant chaque valeur génotypique par sa fréquence et en totalisant pour les trois génotypes. En utilisant le tableau 1.6, nous avons :

$$\begin{aligned}\mu = E(G) &= p^2a + 2pqd + q^2(-a)\\ &= a(p^2-q^2) + 2pqd\\ &= a(p-q)(p+q) + 2pqd\\ &= a(p-q) + 2pqd. \end{aligned} \tag{1.2}$$

La première composante de la moyenne génotypique représente la contribution d'un locus qui est due aux génotypes homozygotes, et la deuxième composante est la contribution d'un locus qui est due aux génotypes hétérozygotes.

Variance génétique

Selon le modèle (1.1) et la paramétrisation du tableau 1.6, nous pouvons écrire la

variation génétique comme la variance de la composante G, σ_G^2 [6] :

$$\begin{aligned}\sigma_G^2 &= E(G^2) - E(G)^2 \\ &= 2pq(a - d(p-q))^2 + (2pqd)^2 \\ &= 2pq\alpha^2 + (2pqd)^2 \qquad \text{avec } \alpha = a - d(p-q) \\ &= Var(A) + Var(D) \\ &= \sigma_A^2 + \sigma_D^2. \end{aligned} \tag{1.3}$$

Le premier terme dans la variance génétique est appelé variance génétique additive et le deuxième terme, variance de dominance.

L'importance relative du génotype considéré pour la valeur du caractère est exprimée par le rapport de la variance génétique à la variance du caractère (σ_Y^2) :

$$\frac{\sigma_G^2}{\sigma_Y^2}.$$

C'est ce que nous appelons l'héritabilité au sens large, et que nous allons explorer dans la section suivante.

1.3 Héritabilité : définition et estimation

1.3.1 Définition de l'héritabilité

On distingue deux types d'héritabilité (Falconer, 1975 ; Ollivier, 2002 ; Ollivier *et al.*, 1971) :

L'héritabilité au sens large, notée H^2, définie comme le rapport de la variance

6. Démonstration est donnée dans l'annexe B.

génétique à la variance phénotypique :

$$H^2 = \frac{\sigma_G^2}{\sigma_Y^2}.$$

Une héritabilité de 0.4 signifie que 40% de la variabilité entre les individus de la population concernant le caractère évalué est liée à un effet génétique.

L'héritabilité au sens strict, notée h^2, définie par le rapport de la variance génétique additive à la variance phénotypique du caractère :

$$h^2 = \frac{\sigma_A^2}{\sigma_Y^2}.$$

Le coefficient h^2 est compris entre 0 et 1. La valeur $h^2 = 1$ est atteinte s'il n'y a pas d'effet d'environnement sur le caractère, le caractère étant purement génétique. La valeur 0 correspond au cas où la totalité de la variation est d'origine environnementale, si les loci polymorphes dans la population sont neutres vis-à -vis du phénotype.

1.3.2 Estimation de l'héritabilité

L'estimation de l'héritabilité s'appuie sur une caractéristique fondamentale que présentent les phénotypes contrôlés génétiquement : la ressemblance entre apparentés. C'est la covariance entre apparentés qui va donc servir à estimer les composantes de la variance génotypique. Trois méthodes sont généralement utilisées pour l'estimation de l'héritabilité au sens strict : l'estimation de l'héritabilité basée sur la régression, l'estimation de l'héritabilité basée sur l'analyse de la variance et l'estimation de l'héritabilité par la méthode de vraisemblance maximale. À titre d'illustration, nous allons introduire les deux premières méthodes. L'estimation de l'héritabilité au sens large est basée sur l'étude comparative des jumeaux.

Méthode de l'estimation de l'héritabilité h^2 basée sur la régression

Définition : le coefficient de régression b d'une variable Y sur un autre X est définie

comme

$$b = \frac{cov(X,Y)}{Var(X)}. \tag{1.4}$$

L'estimation de l'héritabilité basée sur la régression est utilisée pour estimer l'héritabilité au sens strict et la régression "enfant-parent" et celle "enfant-parent moyen" sont les plus courantes.

Régression "enfant-parent"

La ressemblance entre un enfant et un parent est exprimée par la régression du caractère de l'enfant sur le caractère du parent. En effet, considérons Y_p la valeur phénotypique du caractère chez le parent et X_e la valeur phénotypique du caractère de l'enfant. La valeur phénotypique d'un individu dépend de sa valeur génotypique et d'un effet dû à l'environnement comme dans le modèle (1.1). On écrit alors

$$Y_p = \mu + G_p + E_p,$$
$$X_e = \mu + G_e + E_e.$$

La variabilité phénotypique commune entre l'enfant et le parent s'explique par la composante génétique commune.

Ainsi :

$$b = \frac{cov(Y_p, X_e)}{Var(Y_p)},$$
$$\begin{aligned} cov(Y_p, X_e) &= cov(G_p + E_p, G_e + E_e) \\ &= cov(G_p, G_e) \\ &= E\{(G_p - E(G))(G_e - E(G))\} \\ &= E(G_{gp} G_{gme}), \end{aligned}$$

où $G_{gp} = G_p - E(G)$ est la valeur génotypique du parent exprimée en écarts à la moyenne et $G_{gme} = G_e - E(G)$ est la valeur génotypique moyenne de l'enfant.

Tableau 1.7: Génotypes du parent, leurs fréquences, f_p, dans la population et leurs valeurs génotypiques G_{gp}. Valeur génotypique moyenne de l'enfant G_{gme} (Source : Falconer, 1975 page-119).

		Parent	Enfant
Génotype	Fréquence	G_{gp}	G_{gme}
AA	p^2	$2q(\alpha - qd)$	$q\alpha$
Aa	$2pq$	$(q-p)\alpha + 2pqd$	$1/2(q-p)\alpha$
aa	q^2	$-2p(\alpha + pd)$	$-p\alpha$

Par exemple, la valeur génotypique du parent pour le génotype AA du tableau 1.7 peut être obtenue comme suit

$$G_{gp} = a - E(G).$$

Selon l'équation (1.2) et puisque pour le génotype AA, $G_p = a$, nous pouvons écrire

$$\begin{aligned} G_{gp} &= a - \{a(p-q) + 2pqd\} \\ &= a(1-p+q) - 2pqd \\ &= 2qa - 2pqd \\ &= 2q(a + dq - pd - qd) \\ &= 2q(\alpha - qd) \\ &= 2q\alpha - 2q^2d. \end{aligned}$$

Notez que puisque le parent transmet exactement 50 % de son matériel génétique additif, alors la valeur génotypique moyenne de l'enfant, G_{gme}, est la moitié de la valeur génétique additive du parent (la moitié de la première composante de G_{gp}).

Alors à partir du tableau 1.7 :

$$\begin{aligned} cov(Y_p, X_e) &= \sum(fréquence \times G_{gp} \times G_{gme}) \\ &= pq\alpha^2 \\ &= \frac{1}{2}\sigma_A^2, \end{aligned}$$

avec $\alpha = a + d(q - p)$.

En remplaçant dans l'équation (1.4) le coefficient de régression b nous avons

$$b = \frac{\frac{1}{2}\sigma_A^2}{\sigma_{Y_p}^2} = \frac{1}{2}h^2.$$

Régression "enfant-parent moyen"

Soit Y_{pm} la valeur phénotypique moyenne des deux parents c'est-à-dire $Y_{pm} = \frac{1}{2}(Y_{mère} + Y_{père})$ où $Y_{mère}$ est la valeur phénotypique de la mère et $Y_{père}$ est la valeur phénotypique du père. X_e est la valeur phénotypique de l'enfant.

On a :

$$cov(X_e, Y_{pm}) = \frac{1}{2}\{cov(Y_{mère}, X_e) + cov(Y_{père}, X_e)\},$$

or

$$\begin{aligned} cov(Y_{mère}, X_e) &= cov(Y_{père}, X_e) \\ &= \frac{1}{2}\sigma_A^2, \end{aligned}$$

car la covariance des parents prise séparément est donnée par la régression enfant et un parent.

D'autre part :

$$Var(Y_{pm}) = \frac{1}{4}Var(Y_{mère} + Y_{père}) = \frac{1}{4}\{Var(Y_{mère}) + Var(Y_{père})\}$$

et

$$\begin{aligned} Var(Y_{mère}) &= Var(Y_{père}) = \sigma_{Y_{père}}^2, \\ Var(Y_{pm}) &= \frac{1}{2}\sigma_{Y_{père}}^2. \end{aligned}$$

Ainsi, le coefficient de régression d'un enfant sur parent moyen est :

$$b = \frac{\frac{1}{2}\sigma_A^2}{\frac{1}{2}\sigma^2_{Y_{père}}} = \frac{\sigma_A^2}{\sigma^2_{Y_{père}}} = h^2.$$

Conclusion

Le coefficient de régression d'un enfant en son parent est une mesure correcte de $\frac{1}{2}h^2$ ($b = \frac{1}{2}h^2$) et celle de l'enfant sur parents moyens est une bonne mesure de h^2 ($b = h^2$).

Remarque

L'estimation de l'héritabilité par la régression est utilisée lorsque le sujet 1 est le descendant de niveau 1 du deuxième sujet, par contre l'estimation basée sur la corrélation est utilisée lorsque les deux sujets sont de même niveau (frères et sœurs par exemple). Cette dernière mesure sera utilisée dans l'estimation de l'héritabilité au sens strict et au sens large dans ce qui suit.

Méthode d'estimation de l'héritabilité basée sur l'analyse de la variance

Rappel : analyse de la variance

L'approche traditionnelle de l'analyse de la variance (notée ANOVA) à un seul facteur est basée sur le modèle linéaire suivant :

$$y_{ij} = \mu + f_i + \varepsilon_{ij} \qquad i = 1, \ldots, N \text{ et } j = 1, \ldots, n, \tag{1.5}$$

où :

- y_{ij} est la valeur du caractère de l'individu j de la famille i. n le nombre d'individus par famille est supposé égal pour toutes les familles ;
- μ est la moyenne commune ;
- f_i est l'effet de la famille i, c'est-à -dire l'écart de la moyenne de la famille i par rapport à la moyenne commune ;
- ε_{ij} est l'erreur ou résidu, c'est-à -dire l'écart entre la valeur de l'individu j et

la moyenne de la famille i.

L'ANOVA décompose la variance totale en variance expliquée et une variance résiduelle. Les hypothèses de base du modèle linéaire par rapport à l'ANOVA sont :

$$cov(\varepsilon_{ij}, \varepsilon_{ik}) = 0 \qquad \text{et} \qquad cov(f_i, \varepsilon_{ij}) = 0. \tag{1.6}$$

Pour N familles de n frères et sœurs on a le tableau d'analyse de la variance 1.4.

Cause de variation	ddl	Somme des carrés	Moyenne des carrés	Espérances des moyennes des carrés
Interfamille	$N-1$	$SCE = n\sum_{i=1}^{N}(\overline{y}_i - \overline{y})^2$	$MCE = \frac{SCE}{N-1}$	$\sigma_\varepsilon^2 + n\sigma_f^2$
Intrafamille	$N(n-1)$	$SCR = \sum_{i=1}^{N}\sum_{j=1}^{n}(y_{ij} - \overline{y}_i)^2$	$MCR = \frac{SCR}{N(n-1)}$	σ_ε^2

Figure 1.4: Modèle d'ANOVA pour N famille de n frères et sœurs.

La variabilité interfamille est la part de la variance totale qui peut être expliquée par le modèle (aussi appelée somme des carrés expliquée, notée SCE), la variabilité intrafamille est la part de la variance totale non expliquée par le modèle (aussi appelée somme des carrés résiduelle, notée SCR).

L'estimation des composantes de la variance :

$$\sigma_f^2 = \frac{MCE - MCR}{n},$$

$$\sigma_\varepsilon^2 = MCR,$$

$$Var(y) = \sigma_y^2 = \sigma_f^2 + \sigma_\varepsilon^2.$$

Estimation de l'héritabilité basée sur ANOVA

L'estimation de l'héritabilité dans le cas de familles composées de plein-frères (de même père et mère) est basée sur la notion de corrélation définie comme :

$$t = \frac{cov(y_{ij}, y_{ik})}{\sigma_y^2}, \tag{1.7}$$

avec

$$\begin{aligned} cov(y_{ij}, y_{ik}) &= cov[(\mu + f_i + \varepsilon_{ij}), (\mu + f_i + \varepsilon_{ik})] \\ &= cov(f_i, f_i) + cov(f_i, \varepsilon_{ik}) + cov(\varepsilon_{ij}, f_i) + cov(\varepsilon_{ij}, \varepsilon_{ik}). \end{aligned}$$

Selon les hypothèses du modèle ANOVA, on peut écrire :

$$\begin{aligned} cov(y_{ij}, y_{ik}) &= cov(f_i, f_i) \\ &= \sigma_f^2. \end{aligned}$$

Par conséquent, la variance entre les effets des familles est égale à la covariance entre plein-frères (la valeur de la covariance σ_f^2 est tirée du tableau 1.9 où σ_E^2 représente la variance environnementale).

$$\sigma_f^2 = \frac{\sigma_A^2}{2} + \frac{\sigma_D^2}{4} + \sigma_E^2,$$

où :

$$\begin{aligned} \sigma_\varepsilon^2 &= \sigma_y^2 - \sigma_f^2 \\ &= \sigma_y^2 - \left(\frac{\sigma_A^2}{2} + \frac{\sigma_D^2}{4} + \sigma_E^2\right) \\ &= \sigma_A^2 + \sigma_D^2 + \sigma_E^2 - \left(\frac{\sigma_A^2}{2} + \frac{\sigma_D^2}{4} + \sigma_E^2\right) \\ &= \frac{1}{2}\sigma_A^2 + \frac{3}{4}\sigma_D^2. \end{aligned}$$

Sous l'hypothèse d'absence de variance génétique non additive et d'effet de milieu commun, l'estimation de l'héritabilité par ANOVA est déduite de l'équation de la corrélation (1.7) comme suit

$$t = \frac{\frac{1}{2}\sigma_A^2}{\sigma_f^2 + \sigma_\varepsilon^2}$$
$$= \frac{1}{2}\frac{\sigma_A^2}{\sigma_f^2 + \sigma_\varepsilon^2}$$
$$= \frac{1}{2}h^2.$$

Estimation de l'héritabilité au sens large

L'estimation de l'héritabilité au sens large est basée principalement sur l'étude comparative de jumeaux.

Étude comparative de jumeaux

Soient X et Y les valeurs phénotypiques de deux jumeaux respectivement, décomposées chacune en une valeur génotypique et une valeur due à l'environnement. Nous pouvons écrire :

$$X = G + E \quad \text{et} \quad Y = G' + E'.$$

Deux jumeaux monozygotes, c'est-à -dire issus de la division d'un œuf unique sont génétiquement identiques. Les différences entre eux résultent sûrement des circonstances de l'environnement.

$$cov(\text{vrais jumeaux}) = cov(X, Y)$$
$$= cov(E, E').$$

Par contre, les faux jumeaux ou jumeaux dizygotes issus de deux œufs différents diffèrent génétiquement autant que deux plein-frères (voir tableau 1.9) :

$$cov(\text{faux jumeaux}) = \frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2 + cov(E, E').$$

L'étude de différences génétiques entre jumeaux monozygotes (identiques) et dizygotes (non identiques) est menée dans le but de déterminer l'importance relative des facteurs génétiques en admettant que l'effet du milieu est le même chez les dizygotes

et les monozygotes.

Ainsi,

$$cov(\text{faux jumeaux}) - cov(\text{vrais jumeaux}) = \frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2.$$

Si l'on suppose qu'il n'y a pas de variance génétique non additive, la dernière équation s'écrit

$$\begin{aligned} cov(\text{faux jumeaux}) - cov(\text{vrais jumeaux}) &= \frac{1}{2}\sigma_A^2 \\ &= \frac{1}{2}\sigma_G^2. \end{aligned}$$

En termes de corrélation, soit t_v la corrélation entre vrais jumeaux et t_f corrélation entre faux jumeaux, on a :

$$t_f - t_v = \frac{1}{2}\frac{\sigma_G^2}{\sigma_y^2}.$$

Le rapport $\frac{\sigma_G^2}{\sigma_y^2}$ n'est que H^2. D'où,

$$t_f - t_v = \frac{1}{2}H^2.$$

La différence de corrélation entre les caractères des vrais jumeaux et les faux jumeaux fournit donc une estimation pour l'héritabilité au sens large si l'on suppose qu'il n'y a pas de variance génétique non additive. Cependant, puisqu'on ne peut pas raisonnablement supposer que la variance génétique non additive n'existe pas, cette différence ne peut être considérée que comme une limite supérieure de la moitié de l'héritabilité (Falconer, 1975). La tableau 1.8 donne des estimations de H^2 pour certains caractères basées sur des études comparatives des jumeaux.

Le tableau 1.9 donne la composition des covariances phénotypiques ainsi que les expressions de la régression ou de la corrélation en fonction de l'héritabilité selon le lien de parenté.

Phénotype	H^2	Phénotype	H^2
Indice de virilité	12	Poids	63
Indice de Tempérament	58	Longévité	29
Indice de sociabilité	66	Mémoire	47
Excrétion d'acides aminés	72	Taille	85
Taux de lipide sérique	44	Habilité de raisonnement numérique	76
Maximum de lactate dans le sang	34	Habilité verbale	63
Fréquence cardiaque maximale	84		

Tableau 1.8: Héritabilité au sens large (H^2), en pourcentage, basée sur des études comparatives de jumeaux. Source adaptée de "Genetics : Principales and analysis" Hartl et W.Jones, 1998 page-683.

Parenté	Covariances	Coefficient de régression (b) ou corrélation (t)
Enfant et un parent	$\frac{1}{2}\sigma_A^2$	$b = \frac{1}{2}h^2$
Enfant et parent moyen	$\frac{1}{2}\sigma_A^2$	$b = h^2$
Demi-frères	$\frac{1}{4}\sigma_A^2$	$t = \frac{1}{4}h^2$
Plein-frères	$\frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2 + \sigma_E^2$	$t > \frac{1}{2}h^2$

Tableau 1.9: Compositions des covariances phénotypiques ainsi que les expressions de la régression ou de la corrélation en fonction de l'héritabilité pour différents liens de parenté, Falconer, 1975 p-131.

La notion d'héritabilité vue dans cette section est souvent associée aux études d'association génétique. En effet, l'objectif des études d'association génétique est d'iden-

tifier les variantes génétiques causales qui expliquent la variation phénotypique, et qui, par conséquent, expliquent l'héritabilité. Une stratégie utilisée dans ces études consiste à évaluer l'effet individuel de chaque marqueur avec le caractère. Cependant, cette stratégie n'est pas toujours appropriée car elle introduit une inflation de l'erreur de type 1 à cause des comparaisons multiples (tests multiples). Plusieurs corrections ont été proposées pour faire face à ce problème (Hirschhorn et Daly, 2005 ; Balding, 2006 ; Dudoit et Van Der Laan, 2007), on cite entre autres la correction de Bonferroni et la correction de taux de faux positifs. Dans la section prochaine, nous introduisons le problème des tests multiples pour discuter ensuite de quelques procédures pour le corriger (correction Bonferroni et correction de taux des faux positifs).

1.4 Préliminaires statistiques : techniques de correction dans le cas des tests multiples

1.4.1 Tests multiples

En génétique, il arrive souvent que nous voulons tester l'association entre un caractère (Y) et plusieurs covariables (notées x_l avec $l = 1, \ldots, L$) séparément. Le modèle est exprimé pour une covariable à la fois par

$$Y_i = \mu + \beta_l x_{il} + \varepsilon_i, \quad i = 1, \ldots, n. \quad \text{et} \quad l = 1, \ldots, L, \tag{1.8}$$

avec β_l représente l'effet de covariable l sur le caractère Y. On dit que la covariable x_l (ex : SNP) est associée avec le caractère Y si le test $H_0 : \beta_l = 0$ est rejeté en faveur de $H_1 : \beta_l \neq 0$ dans le modèle (1.8).

On sait que pour chaque covariable, nous avons un risque de rejeter à tort $H_0 : \beta_l = 0$. Ce risque, noté α, est défini par :

α=Erreur de type 1=P{ rejet de H_0 sachant que H_0 est vraie }.

Lorsque nous faisons chacun des L tests au seuil α, le seuil global, à savoir la probabilité de trouver au moins un résultat significatif sera beaucoup plus grande que α. En effet, supposons que nous menons L tests d'hypothèses (synthèse des résultats est dans le tableau 1.10) donnés par $H_0^{global} = [H_0^1 \cdots H_0^L]$, et chaque test est contrôlé à un niveau α, alors nous avons

$$\alpha^{global} = P(V = \text{rejeter au moins un } H_0^l \mid \beta_1 = \beta_2 = \ldots = \beta_L = 0).$$

Pour une illustration simplifiée, supposons que ces tests sont indépendants les uns des autres, de sorte que la probabilité de rejeter un test ne dépend pas des résultats des autres tests. Alors le seuil α^{global}, appelé aussi Family Wise Error Rate (FWER), est donné par :

$$\begin{aligned}
FWER &= Pr(V \geq 1 \mid H_0^{global} \text{ Vraie}), \\
&= 1 - Pr(V = 0 \mid H_0^{global} \text{ Vraie}) \\
&= 1 - \prod_{l=1}^{m=L} Pr(\text{ne pas rejeter } H_0^l \mid H_0^l \text{ Vraie}) \\
&= 1 - \prod_{l=1}^{m=L} [1 - Pr(\text{rejeter } H_0^l \mid H_0^l \text{ Vraie})] \\
&= 1 - \prod_{l=1}^{m=L} (1 - \alpha) \\
&= 1 - (1 - \alpha)^L.
\end{aligned}$$

Ainsi, si on considère deux tests indépendants contrôlés chacun au niveau $\alpha = 0.05$, en réalité notre erreur de type 1 est égale à $1 - (1 - 0.05)^2 = 0.0975$. Dans le cas de dix tests indépendants, bien que nous contrôlions chacun des dix tests individuels au

niveau α, nous avons .

$$\begin{aligned} \text{FWER} &= 1 - (1 - 0.05)^{10} \\ &= 1 - 0.598 = 0.401. \end{aligned}$$

L'erreur de type 1 pour un seuil global donné augmente donc avec le nombre de tests effectués. Le risque doit être corrigé afin d'avoir un niveau de risque étendu à l'ensemble des tests acceptables. L'objectif est donc de trouver une méthode permettant de choisir un seuil de rejet qui contrôle le risque de première espèce. Plusieurs corrections dites de "tests multiples" ont été proposées afin de résoudre ce problème. Les notations utilisées dans le tableau 1.10 seront adoptées tout au long de cette section :

- m est le nombre d'hypothèses nulles à tester et il est connu d'avance ;
- m_0 est le nombre des hypothèses nulles vraies ;
- $m - m_0$ est le nombre d'hypothèses nulles fausses ;
- U est le nombre de vrais négatifs (ne pas rejeter H_0 sachant que H_0 est vraie) ;
- V est le nombre de faux positifs (rejeter H_0 alors que H_0 est vraie) ;
- T est le nombre de faux négatifs (ne pas rejeter H_0 alors que H_1 est vraie) ;
- S est le nombre de vrais positifs (rejeter H_0 alors que H_1 est vraie) ;
- R est le nombre total de tests rejetés.

Tableau 1.10: Synthèse des résultats de tests multiples

Décision\ Réalité	H_0 Vraie	H_1 Vraie	Total
ne pas rejeter H_0	U	T	m-R
rejeter H_0	V	S	R
Total	m_0	$m - m_0$	m

R est une variable aléatoire observable, par contre S, T, U et V sont des variables aléatoires non observables.

1.4.2 Correction de Bonferroni

L'ajustement de Bonferroni pour les comparaisons multiples est sans doute la correction la plus simple à appliquer. Il consiste simplement à utiliser $\alpha' = \frac{\alpha}{m}$ au lieu de α pour le niveau de chaque test, afin que le seuil global soit inférieur ou égal à α. Par exemple, si nous avons l'intention de procéder à $m = 10$ tests d'hypothèses et que nous voulons contrôler ces tests à un niveau global de $\alpha = 0.05$, alors le seuil de contrôle de chacun des m tests est $\alpha' = 0.05/10 = 0.005$. Le problème de cette correction est qu'elle est très conservatrice et a tendance à réduire le nombre de résultats réellement significatifs (Balding, 2006).

1.4.3 Contrôle de taux de faux positifs FDR

Le taux de faux positifs, ou « FDR de l'anglais false discovery rate», introduit par Benjamini et Hochberg (1995, noté BH), correspond à la proportion attendue de faux positifs au sein de résultats déclarés positifs. L'objectif est d'identifier le plus possible de signaux significatifs mais le moins possible de faux positifs parmi ces signaux (Storey et Tibshirani, 2003). On distingue deux cas :
Cas 1 : $H_0^{global} = [H_0^1, H_0^2, \cdots, H_0^m]$, toutes les hypothèses nulles sont vraies. On a $V = R$ (cas $S = 0$) et

$$\frac{V}{R} = \begin{cases} 0 & \text{si} \quad V = 0 \\ 1 & \text{si} \quad V \geq 1, \end{cases}$$

$$\begin{aligned} FDR &= E\left(\frac{V}{R}\right) \\ &= E\left(E\left(\frac{V}{R}\middle| R\right)\right) \\ &= E\left(\frac{V}{R}\middle| R > 0\right) P(R > 0) + E\left(\frac{V}{R}\middle| R = 0\right) P(R = 0) \\ &= E\left(\frac{V}{R}\middle| R > 0\right) P(R > 0) + 0 \times P(R = 0) \\ &= 1 \times P\left(V \geq 1 \mid H_0^{global} = [H_0^1, H_0^2, \cdots, H_0^m] \text{ sont vraies}\right) \\ &= FWER. \end{aligned}$$

Cas 2 : $H_0^{partielle} = [H_0^1, H_0^2, \cdots, H_0^{m_0}]$ sont vraies avec $m_0 < m$. On a $V < R$ et $\frac{V}{R} < 1$, avec

$$\begin{aligned}
FDR &= E\left(\frac{V}{R}\right) \\
&= E\left(E\left(\frac{V}{R}\middle| R\right)\right) \\
&= E\left(\frac{V}{R}\middle| R>0\right) P(R>0) + E\left(\frac{V}{R}\middle| R=0\right) P(R=0) \\
&= E\left(\frac{V}{R}\middle| R>0\right) P(R>0) + 0 \times P(R=0) \\
&= E\left(\frac{V}{R}\middle| V\geq 1\right) P(V\geq 1) + E\left(\frac{V}{R}\middle| V=0\right) P(V=0) \\
&= \frac{V}{R} \times P(V\geq 1) + \frac{0}{R} \times P(V=0) \\
&< P(V\geq 1) \\
&< FWER.
\end{aligned}$$

Comme le nombre de fausses hypothèses, donnés par $m - m_0$ dans le tableau 1.10, augmente, le nombre de vrais positifs, étant donné par S, auront également tendance à augmenter. À son tour, V/R sera plus petit et la différence entre FDR et FWER sera plus grande. Par conséquent, on a le résultat général que le FDR est inférieure (cas 2) ou égale à FWER (cas 1). Cela implique que toute approche qui contrôle FWER va également contrôler FDR. L'inverse, cependant, est pas vrai.

Benjamini et Hochbert décrivent une procédure pratique pour obtenir un FDR contrôlé à un niveau α, c'est-à -dire permettant de garantir un FDR inférieur à α. Cette procédure, notée BH, est décrite dans l'algorithme 1.4.3. FDR présente l'avantage d'être moins sévère sur le résultat des m tests, au prix du non rejet de quelques hypothèses nulles dont on contrôle la proportion, ce qui augmente la puissance du

test.

Algorithme de Benjamini et Hochbert FDR (1.4.3).

1 : Trier les p-valeurs des hypothèses dans l'ordre croissant,

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}.$$

2 : Trouver le nombre K défini dans l'équation suivante :

$$K = max_{l \in [1,m]} \left\{ p_{(l)} \leqslant \frac{l\alpha}{m} \right\}.$$

3 : Rejeter l'hypothèse nulle pour les tests $l \leqslant K$.

CHAPITRE II

TECHNIQUES DE RÉDUCTION DE DIMENSIONS

Motivation

Il y a un grand besoin de développer une méthodologie pour analyser et exploiter les informations contenues dans les données génétiques. Cependant, face à des données génétiques de grandes dimensions avec la redondance ou la non-pertinence des caractéristiques, la grande dimension est vue comme une des causes majeures de la complexité des données. Les algorithmes de réduction de dimensions se présentent comme une solution pour rechercher des sous-espaces de faibles dimensions tout en minimisant la perte d'information. Mais en utilisant différents algorithmes de réduction de dimensions pour analyser le même jeu de données, les conclusions peuvent être différentes. L'analyse en composante principale (ACP) et l'analyse en composante principale de l'héritabilité sont deux techniques de réduction de dimensions utiles pour l'analyse des données génétiques. Notre objectif dans ce chapitre est de présenter théoriquement les deux méthodes ainsi que de montrer l'efficacité ou la puissance des composantes principales de l'héritabilité (PCH) dans la détection des variantes génétiques par rapport à l'analyse en composantes principales classique.

2.1 Analyse en composantes principales

L'analyse en composantes principales (Jolliffe, 2002 ; Johnson *et al.*, 1992) est une technique de réduction de dimensions qui cherche des combinaisons linéaires non corrélées à partir des variables d'origine d'une base de données de grande dimension. Ces combinaisons linéaires sont appelées composantes principales.

Techniquement, la première composante principale (CP) peut être définie comme une combinaison linéaire des variables observées avec variance maximale (information maximale). La seconde composante extraite représente une deuxième combinaison des variables d'origine avec variance maximale, une variance qui n'a pas été représentée par la première composante. Les deux composantes sont non corrélées. Les autres composantes sont extraites de la même façon.

Détermination des coefficients et des composantes principales

Les composantes principales d'une matrice de données sont déterminées soit à partir de la décomposition en valeurs propres de la matrice de variances-covariances des données ou de la décomposition en valeurs singulières de la matrice des données, généralement après centrage des données pour chaque variable.

La détermination des composantes principales par rapport à notre sujet de recherche est basée sur la décomposition en valeurs propres de la matrice de variances-covariances Σ.

Décomposition en valeurs propres

Soit Y une matrice de données à n lignes et q colonnes telle que Σ est la matrice de variances-covariances des q caractères. Les lignes $i = 1, 2, \cdots, n$ décrivent les valeurs prises par l'individu i pour les q variables quantitatives. L'exposant T indique la transposée. La matrice Y peut s'écrire comme suit :

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & \cdots & y_{2q} \\ y_{31} & y_{32} & \cdots & \cdots & y_{3q} \\ .. & \cdots & \cdots & \cdots & \cdots \\ .. & \cdots & \cdots & \cdots & \cdots \\ .. & \cdots & \cdots & \cdots & \cdots \\ .. & \cdots & \cdots & \cdots & \cdots \\ y_{n1} & y_{n2} & \cdots & \cdots & y_{nq} \end{pmatrix} = \begin{pmatrix} Y_1^\top \\ Y_2^\top \\ Y_3^\top \\ .. \\ ... \\ .. \\ .. \\ Y_n^\top \end{pmatrix}.$$

La décomposition en valeurs propres de la matrice de variances-covariances est :

$$\Sigma = \Gamma \Lambda \Gamma^\top,$$

où Γ est une matrice $q \times q$ orthogonale et Λ est une matrice $q \times q$ diagonale. Notons

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_q \end{pmatrix}, \qquad \Gamma = (e_1, \cdots, e_q),$$

où les λ_i sont les valeurs propres classées par ordre décroissant de Σ et les e_i sont les vecteurs propres orthonormés de Σ, avec

$$\| e_i \| = 1 \ , \ e_i^\top e_j = 0 \ , \ i \neq j.$$

L'ACP se réalise sur plusieurs étapes, et dans chaque étape une composante principale est déterminée.

La première composante principale U_1 est définie par

$$U_1 = \underset{U \in \mathbb{R}^q}{\text{argmax}}\{Var(U^\top Y)\},$$

sous contrainte $U_1^\top U_1$ =1. La solution à cette maximisation est $U_1 = e_1$ où e_1 est le vecteur propre associé à la plus grande valeur propre λ_1 de Σ. Ainsi $Z_1 = e_1^\top Y$ a pour variance, $Var(Z_1) = e_1^\top \Sigma e_1$, maximale.

La deuxième composante principale U_2 est définie par

$$U_2 = \underset{U \in \mathbb{R}^q}{\text{argmax}}\{Var(U^\top Y)\},$$

sous contraintes $U_2^\top U_2 = 1$ et $U_2^\top e_1 = 0$. La solution à ce problème de maximisation est $U_2 = e_2$ où e_2 est le vecteur propre de Σ associé à λ_2 la deuxième plus grande valeur propre.

La $q^{ème}$ composante principale U_q est définie par

$$U_q = \underset{U \in \mathbb{R}^q}{\text{argmax}}\{Var(U^\top Y)\},$$

et sous contraintes $U_q^\top U_q = 1$ et $U_q^\top e_m = 0$, $m = 1, \ldots, q-1$. La solution à ce problème de maximisation est $U_q = e_q$ où e_q est le vecteur propre de Σ associé à λ_q la $q^{ème}$ plus petite valeur propre. Habituellement, la matrice de variances-covariances Σ est inconnue ; elle peut alors être estimée par $\widehat{\Sigma}$ à partir de la matrice des données Y. Soient $\widehat{\lambda}_1, \widehat{\lambda}_2, \cdots, \widehat{\lambda}_q$ les valeurs propres associées respectivement aux vecteurs propres $\widehat{e}_1, \widehat{e}_2, \cdots, \widehat{e}_q$ de la matrice $\widehat{\Sigma}$. Les composantes principales deviennent :

$$
\begin{aligned}
z_1 &= \widehat{e}_1^{\top} y = \widehat{e}_{11} y_1 + \widehat{e}_{12} y_2 + \ldots + \widehat{e}_{1q} y_q, \\
z_2 &= \widehat{e}_2^{\top} y = \widehat{e}_{21} y_1 + \widehat{e}_{22} y_2 + \ldots + \widehat{e}_{2q} y_q, \\
&\vdots \\
z_q &= \widehat{e}_q^{\top} y = \widehat{e}_{q1} y_1 + \widehat{e}_{q2} y_2 + \ldots + \widehat{e}_{qq} y_q.
\end{aligned}
$$

Avec les propriétés $Var(z_1) \geqslant Var(z_2) \geqslant \cdots \geqslant Var(z_q) \geqslant 0$ équivalent à $\widehat{\lambda}_1 \geqslant \widehat{\lambda}_2 \geqslant \cdots, \geqslant \widehat{\lambda}_q \geqslant 0$ et $\sum_{j=1}^{q} Var(z_j) = \sum_{j=1}^{q} \widehat{\lambda}_j = \sum_{j=1}^{q} \widehat{Var(y_j)}$.

2.2 Composantes principales d'héritabilité

Motivation

Pour faire face à des données de phénotypes de grande dimension, l'analyse en composantes principales (ACP) peut servir à réduire la dimension des données phénotypiques.

Les maladies telles que le cancer, les maladies mentales, le diabète ou l'Alzheimer semblent davantage se produire chez des personnes génétiquement reliées, comparées à la population générale. Malheureusement, l'ACP ne prend pas ceci en compte lors de la décomposition de la matrice de variances-covariances.

L'approche des composantes principales basée sur l'héritabilité proposée par Ott et Rabinowitz (1999) exploite l'information de l'héritabilité des caractères dans les familles en définissant les composantes principales de l'héritabilité (PCH) comme des scores avec héritabilité maximale, sous réserve que les scores soient orthogonaux (Wang *et al.*, 2007).

Méthodologie de l'approche

Dans le modèle de Ott et Rabinowitz (1999), un phénotype est décomposé en une composante spécifique à la famille (notée G) qui représente l'effet des gènes transmis par les parents, et une composante spécifique à l'individu (notée E) qui est due à l'environnement dans lequel il vit. Soit Y un vecteur, $q \times 1$, de caractères. Le modèle est défini par

$$Y = \mu + G + E. \tag{2.1}$$

Ce modèle permet de décomposer la variance phénotypique en une variance interfamille (Σ_G) et une variance intrafamille Σ_E. Ainsi, on peut écrire

$$\Sigma = \Sigma_G + \Sigma_E.$$

Au lieu de maximiser la variation totale comme dans l'analyse en composantes principales, l'approche PCH maximise la variance interfamille par rapport à la variation phénotypique. Autrement dit, le PCH est une solution à

$$\underset{W}{\text{argmax}} \quad \frac{W^\top \Sigma_G W}{W^\top \Sigma W}, \tag{2.2}$$

où W est le vecteur colonne $q \times 1$ qui contient les coefficients de la combinaison linéaire. W est un vecteur de poids qui va définir un nouveau modèle de projection des phénotypes d'un espace de grande dimension à un espace de faible dimension. Les vecteurs $W_1, W_2, \cdots, W_q$ qui résolvent le problème d'optimisation (2.2) sont aussi les solutions du système d'équations suivant :

$$\Sigma_G = \lambda \Sigma_E W.$$

Par conséquent, les scores obtenus des composantes principales d'héritabilité sont $W_1^\top Y, W_2^\top Y, \cdots, W_q^\top Y$, avec les propriétés suivantes : le premier score extrait re-

présente une variable avec héritabilité maximale, le second score extrait représentera une variable avec héritabilité maximale qui n'a pas été représentée par la première composante et qui est non corrélée avec la première. Les autres scores qui sont extraits de l'analyse ont les mêmes caractéristiques que les deux premiers (Wang *et al.*, 2007).

Dans le cas des familles composées de frères et sœurs, Ott et Rabinowitz estiment la matrice de variances-covariances de la composante spécifique à la famille et de la composante spécifique à l'individu comme suit :

$$\widehat{\Sigma}_E = \frac{1}{\sum_{i=1}^{n}(m_i - 1)} \sum_{i=1}^{n} \sum_{j=1}^{m_i} (Y_{ij} - \overline{Y}_{i.})(Y_{ij} - \overline{Y}_{i.})^\top,$$

où m_i est la taille de la $i^{ème}$ famille, et

$$\widehat{\Sigma}_G = \frac{1}{\sum_{i=1}^{n}(m_i - 1)} \sum_{i=1}^{n} \sum_{j=1}^{m_i} (Y_{ij} - \overline{Y}_{..})(Y_{ij} - \overline{Y}_{..})^\top - \widehat{\Sigma}_E.$$

Puisque la structure de corrélation est la même à l'intérieur de chaque famille, on remarque que $\widehat{\Sigma}_G$ et $\widehat{\Sigma}_E$ peuvent êtres vues comme la variance interfamilles et la variance intrafamilles respectivement dans un modèle d'analyse de la variance multivariée standard [1].

---

1. L'analyse de variance multivariée est un test statistique qui vise à déterminer si des facteurs qualitatifs ont des effets significatifs sur plusieurs variables dépendantes quantitatives prises collectivement.

2.3 Composantes principales d'héritabilité, Klei *et al.* (2008)

Motivation

Dans un organisme complexe tel que l'être humain, le nombre de caractères est dans l'ordre de millions alors que le nombre de gènes dans le génome humain est d'environ 20 000. Inévitablement, il y a au moins un gène qui affecte plusieurs caractères. Ce phénomène d'un gène (ou une mutation) qui affecte plusieurs caractères est connu sous le nom de pléiotropie.

Une approche standard d'analyse d'association génétique consiste à tester l'association entre chaque phénotype et chaque SNP. Cependant, cette approche pénalise la puissance de détection des variantes génétiques à cause des tests multiples. Le défi consiste donc à identifier un nombre réduit de phénotypes capables d'expliquer le maximum de la variabilité génétique, limitant ainsi la problématique des tests multiples.

Dans ce contexte Klei *et al.* (2008) ont proposé de capturer l'effet de la pléiotropie dans l'approche des composantes principales de l'héritabilité de Ott et Rabinowitz (1999). Ceci est fait dans le cadre de l'analyse d'association portée sur des sujets non apparentés.

Pour chaque SNP, l'approche PCH de Klei (Klei *et al.*, 2008) réduit les phénotypes à un seul caractère composé qui a une plus grande héritabilité que toute autre combinaison linéaire des phénotypes.

2.3.1 Méthodologie de l'approche

Supposons que nous avons q phénotypes observés chacun sur n sujets non apparentés. Reprenons la matrice définie précédemment, $Y_{n\times q}$. Le modèle est défini pour chaque

sujet i :

$$Y_{ij} = u_j + \beta_j x_i + \varepsilon_{ij} \quad i = 1, \ldots, n \;\; j = 1, \ldots, q, \tag{2.3}$$

où Y_{ij} est la valeur du $j^{ème}$ caractère pour l'individu i, β_j est l'effet du SNP sur le caractère j. Le SNP (noté x) a 3 niveaux (0, 1, 2). Dans le modèle additif (2.3), le SNP indique :

$$x = \begin{cases} 0 & \text{si} \quad AA \\ 1 & \text{si} \quad Aa \\ 2 & \text{si} \quad aa. \end{cases}$$

L'écart dû à l'environnement est représenté par ε_{ij}, la variable résiduelle. Le modèle peut s'écrire sous forme matricielle de la manière suivante :

$$\begin{pmatrix} Y_1^\top \\ Y_2^\top \\ Y_3^\top \\ .. \\ .. \\ .. \\ .. \\ Y_n^\top \end{pmatrix} = \begin{pmatrix} \mu_1 + \beta_1 x_1 + \varepsilon_{11} & \mu_2 + \beta_2 x_1 + \varepsilon_{12} & \cdots & \cdots & \mu_q + \beta_q x_1 + \varepsilon_{1q} \\ \mu_1 + \beta_1 x_2 + \varepsilon_{21} & \mu_2 + \beta_2 x_2 + \varepsilon_{22} & \cdots & \cdots & \mu_q + \beta_q x_2 + \varepsilon_{2q} \\ \mu_1 + \beta_1 x_3 + \varepsilon_{31} & \mu_2 + \beta_2 x_3 + \varepsilon_{32} & \cdots & \cdots & \mu_q + \beta_q x_3 + \varepsilon_{3q} \\ .. & .. & \cdots & \cdots & \cdots \\ .. & .. & \cdots & \cdots & \cdots \\ .. & .. & \cdots & \cdots & \cdots \\ .. & .. & \cdots & \cdots & \cdots \\ \mu_1 + \beta_1 x_n + \varepsilon_{n1} & \mu_2 + \beta_2 x_n + \varepsilon_{n2} & \cdots & \cdots & \mu_q + \beta_q x_n + \varepsilon_{nq}. \end{pmatrix}.$$

Cette matrice peut également s'écrire sous la forme :

$$\begin{pmatrix} Y_1^\top \\ Y_2^\top \\ Y_3^\top \\ .. \\ .. \\ .. \\ .. \\ Y_q^\top \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ .. & .. \\ .. & .. \\ .. & .. \\ .. & .. \\ 1 & x_n \end{pmatrix} \times \begin{pmatrix} \mu_1 & \mu_2 & \mu_3 & \cdots & \cdots & \mu_q \\ \beta_1 & \beta_2 & \beta_3 & \cdots & \cdots & \beta_q \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \cdots & \varepsilon_{1q} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \cdots & \varepsilon_{2q} \\ \varepsilon_{31} & \varepsilon_{32} & \cdots & \cdots & \varepsilon_{3q} \\ .. & \cdots & \cdots & \cdots & \cdots \\ .. & \cdots & \cdots & \cdots & \cdots \\ .. & \cdots & \cdots & \cdots & \cdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \cdots & \varepsilon_{nq} \end{pmatrix}.$$

On peut écrire

$$Y = X\beta + \varepsilon, \tag{2.4}$$

où $X\beta$ et ε sont l'équivalent de G et de E dans le modèle 2.1 de Ott et Rabinowitz (1999).

À partir du modèle (2.4), on déduit que la variance phénotypique est égale à la somme de la variance génotypique et de la variance résiduelle. Nous avons

$$\begin{aligned} Var(Y) &= Var(X\beta) + Var(\varepsilon) \\ &= \Sigma_G + \Sigma_\varepsilon, \end{aligned}$$

avec

$$\Sigma_G = \beta\beta^\top 2p(1-p),$$

où p désigne la fréquence de l'allèle le moins représenté ou le moins fréquent sur le SNP (aussi appelé allèle mineur, noté MAF). La part de variation due à d'autres effets génétiques non pris en considération dans la variance génotypique est incluse dans la variation résiduelle.

L'objectif de l'approche de Klei est de réduire les caractères corrélés en un nouveau

caractère identifié par le vecteur de poids W tel que l'héritabilité attribuable au SNP considéré est maximisé par la combinaison linéaire $Y_W = W^\top Y = w_1 y_1 + w_2 y_2 + \ldots + w_q y_q$. L'héritabilité est définie à partir des paramétrisations du modèle de Klei par

$$h^2 = \underset{W}{\text{argmax}} \frac{W^\top \Sigma_G W}{W^\top \Sigma W}.$$

Le vecteur de poids W qui maximise cette quantité peut être obtenu en analysant la décomposition de Choleski et la décomposition en valeur propre de Σ_ε et Σ_G respectivement. Cependant, les matrices de variances-covariances Σ_G et Σ_ε ne sont pas connues en pratique. Klei *et al.* (2008) décrivent un algorithme à suivre pour calculer les PCH avec un maximum d'héritabilité :

Algorithme de l'approche PCH de Klei (2.3.1).

— Définir la matrice des phénotypes $Y_{n\times q}$ et la matrice $X_{n\times 2}$.

— Estimer les paramètres du modèle $\widehat{\beta}$ par la méthode de moindres carrés ordinaires (voir Draper *et al.*, 1966 ; Johnson *et al.*, 1992).

$$\widehat{\beta} = (X^\top X)^{-1} X^\top Y.$$

— Estimer $\widehat{Y}$ et $\widehat{\varepsilon}$.

$$\hat{Y} = X\hat{\beta},$$

$$\hat{\varepsilon} = Y - \hat{Y}.$$

— Estimer la variance génétique :

$$\begin{aligned}\widehat{\Sigma_G} &= \widehat{\beta}^\top (X^\top X)^{-1} \widehat{\beta} \\ &= 2p(1-p)\widehat{\beta}\widehat{\beta}^\top.\end{aligned}$$

— Estimer la variance résiduelle :

$$\widehat{\sigma_\varepsilon^2} = \frac{1}{n-2}(Y - \widehat{Y})^\top (Y - \widehat{Y}).$$

— Faire la décomposition de Choleski de la variance résiduelle :

$$\widehat{\sigma_\varepsilon^2} = LL^\top.$$

— Faire la décomposition en valeurs propres de la matrice :

$$L^{-1}\sigma_G^2(L^{-1})^\top \equiv PDP^\top.$$

— Il y a une seule valeur propre λ_1 de la matrice D différente de zéro.

— Calculer le vecteur des poids W :

$$W = (L^{-1})^\top V,$$

avec V le vecteur propre associé à la valeur propre λ_1.

— Calculer le nouveau score $Y_W = W^\top Y$ qui a une héritabilité maximale (le PCH sera noté PCH_K).

L'héritabilité associée à $W^\top Y$ est donnée par $\frac{\lambda_1}{1+\lambda_1}$. Une fois que la composante principale d'héritabilité est déterminée, le test d'association qui relie le marqueur avec la composante principale d'héritabilité est effectué. Une révision du test d'association est donnée dans la section prochaine.

Notons que la procédure de Klei peut être appliquée de façon séparée pour chaque SNP $l = 1, \ldots, L$.

2.3.2 Modèle de PCH de Klei : test d'association

L'approche PCH de Klei estime l'effet marginal de chaque marqueur (association pas-à-pas) sur la composante principale de l'héritabilité correspondante, indépendamment des autres marqueurs. Dans la suite, nous nous concentrons sur un SNP l en particulier, mais notons qu'en pratique le test doit être effectué pour chaque SNP. L'association linéaire entre le PCH_K (noté Y_W) et le SNP peut être modélisée comme

suit :

$$Y_W = W^\top Y = \mu_W + \beta_W x + \varepsilon_i,$$

avec :

- H_0 : Le SNP n'est pas associé au PCH_K correspondant,
- H_1 : Le SNP est associé au PCH_K correspondant.

Formellement, ceci se traduit par :

$$H_0 : \beta_W = 0,$$

$$H_1 : \beta_W \neq 0.$$

On définit sous H_0 la statistique de test T :

$$T = \frac{\widehat{\beta_W}}{\sqrt{V(\widehat{\beta_W})}},$$

où

$$\widehat{\beta_W} = \left(\frac{W^\top \widehat{\Sigma} W h_W^2}{2p(1-p)} \right)^{\frac{1}{2}}.$$

On souligne que p dans l'expression de $\widehat{\beta}$ désigne la fréquence du marqueur. La variance $V(\widehat{\beta_W})$ est estimée par l'estimateur sans biais $S^2(\widehat{\beta_W})$:

$$S^2(\widehat{\beta_1}) = \left(\frac{W^\top \widehat{\Sigma} W (1 - h_W^2)}{2np(1-p)} \right).$$

La statistique de test est $T \sim \mathcal{N}(\delta, 1)$ de paramètre de non-centralité

$$\delta = \left(\frac{n h_W^2}{1 - h_W^2} \right)^{\frac{1}{2}}.$$

Remarque : à partir de ces calculs, on note que la puissance du test augmente si la taille de l'échantillon augmente ou bien l'héritabilité augmente.

D'une façon générale, pour chaque marqueur, on attribue une valeur de la statistique considérée. En fonction de cette valeur par rapport à un seuil déterminé (en général 5%), on décide si l'on considère ou non le marqueur comme étant statistiquement associé au PCH_K correspondant. Cependant, le fait d'utiliser le même jeu de données pour estimer le PCH_K et pour estimer β_W l'effet du SNP sur le score $Y_W = W^\top Y$, peut induire une surestimation de β_W et introduire ainsi de l'inflation de l'erreur de type 1 dans le test d'association (Mei *et al.*, 2010). Pour remédier à ce problème, Klei *et al.* (2008) ont proposé de diviser l'échantillon en deux sous-ensembles disjoints : observations n_0 pour estimer les W, et les observations restantes sont utilisées pour tester l'association entre Y_W et le marqueur, avec n_0/n étant une petite fraction, par exemple 0.2. Comme les résultats dépendent de manière significative de la façon dont l'échantillon est divisé, le processus est répété plusieurs fois en utilisant des techniques de Bootstrap et la p-valeur résultante pour le test d'association est dérivée de la moyenne. Bien sûr, cette méthode rend le calcul trop complexe. Nous présentons dans le chapitre prochain les tests de permutation que nous proposons pour contrôler l'inflation de l'erreur de type 1 dans l'approche de Klei ainsi que dans notre méthode qui généralise la méthode de Klei à l'analyse de plusieurs SNPs à la fois. Cette technique de permutation est bien meilleure au niveau du temps de calcul, si nous comparons avec l'approche Bootstrap de Klei.

CHAPITRE III

COMPOSANTES PRINCIPALES D'HÉRITABILITÉ GÉNÉRALISÉES

Motivation

Les études d'association pangénomiques dans les populations humaines ont découvert des centaines de SNPs associés de façon significative avec les caractères complexes (Hindorff *et al.*, 2009), mais ces polymorphismes n'expliquent qu'une petite fraction de la variation génétique des caractères complexes.

Des échantillons de plusieurs dizaines de milliers d'individus sont nécessaires pour identifier des variantes ayant de faibles effets. Ainsi, plus les effets sont faibles, plus les effectifs nécessaires à leur détection doivent être élevés. Néanmoins, il est difficile de prévoir l'effectif nécessaire à l'explication de la variation génétique d'un caractère complexe, et le coût des analyses à grande échelle de données de séquence reste prohibitif pour l'étude des maladies complexes.

Pour la taille par exemple, Yang *et al.* (2010) ont montré que si l'on considère l'ensemble des SNPs, la fraction de la variance expliquée passait de 5% à 45%. Yang *et al.* (2010) expliquent que la grande partie d'héritabilité n'est pas manquante, mais n'a pas été détectée parce que les effets individuels sont trop petits pour passer les

tests de signification du GWAS ($\sim 10^{-8}$, correspond au seuil de Bonferroni pour un seuil 5%). Ainsi, une grande partie de l'héritabilité de la taille est probablement due à des variantes ayant des effets trop faibles pour être détectés.

Plusieurs auteurs s'accordent désormais sur le fait que de multiples variantes communes ayant un faible effet contribuent dans l'architecture des maladies complexes. L'approche PCH de Klei (Klei *et al.*, 2008) peut s'avérer peu puissant pour détecter des variantes génétiques à faibles effets étant donné que l'approche est basée sur une analyse individuelle des marqueurs. Pour accéder à l'ensemble de la variabilité des variantes génétiques et maximiser ainsi l'information apportée par les marqueurs, nous proposons une extension de l'approche de Klei qui réduit les caractères corrélés en un seul caractère avec un maximum d'héritabilité et qui permet d'analyser, contrairement au modèle de Klei, plusieurs marqueurs conjointement.

L'utilisation d'un modèle multi-marqueurs dans l'approche de PCH peut optimiser la variabilité génétique et augmenter par conséquent la puissance de détection d'un locus de caractères quantitatifs (noté LCQ) et la précision de sa localisation ; c'est notre hypothèse de départ basée sur les résultats de Yang *et al.* (2010).

3.1 Analyse en composantes principales d'héritabilité avec plusieurs variantes génétiques

L'idée de l'analyse en composantes principales d'héritabilité généralisée (notée PCH_g) est de projeter les phénotypes d'un espace de grande dimension à un espace de petite dimension, tout en tenant compte de l'association entre les phénotypes et plusieurs SNPs simultanément. Cette approche est considérée comme une extension de celle définie par Klei (eq 2.4), en analysant plusieurs SNPs à la fois. Ainsi, pour un phénotype donné j, Y_{ij}, le modèle pour un individu i peut s'écrire de la façon suivante :

$$Y_{ij} = \mu_j + \sum_{l=1}^{L} \beta_{jl} x_{il} + \varepsilon_{ij},$$

où les valeurs du SNP l x_{il} sont dans $\{0, 1, 2\}$ pour $l = 1, \ldots, L$, μ_j est la moyenne générale des individus pour le phénotype j, β_{jl} est l'effet du marqueur l sur le caractère j et ε_{ij} est un bruit que l'on suppose gaussien. Sous forme matricielle, le modèle peut s'écrire comme :

$$Y_{n\times q} = X_{n\times(L+1)}\beta_{(L+1)\times q} + \varepsilon_{n\times q}, \tag{3.1}$$

où X est la matrice de design décrivant les marqueurs avec l'ordonnée à l'origine. La nouvelle approche PCH_g, comme celle de Klei, cherche à réduire la dimension de la matrice des phénotypes Y par la spécification d'un nouveau caractère y^* avec maximum d'héritabilité. L'héritabilité s'exprime ainsi par

$$h^2 = \underset{W}{\text{argmax}} \frac{W\Sigma_G W}{W\Sigma W}, \tag{3.2}$$

où

$$\Sigma_G = Var(\beta^\top X) = \beta^\top \Sigma_X \beta,$$

avec Σ_G est la matrice de variances-covariances des L SNPs. On redéfinit l'algorithme de l'analyse en composantes principales d'héritabilité pour inclure les effets joints des marqueurs comme suit

Algorithme de l'approche PCH_g (3.1).

— Définir la matrice des phénotypes $Y_{n\times p}$ et la matrice $X_{n\times(1+L)}$.

— Estimer les paramètres du modèle $\widehat{\beta}$.

— Estimer $\widehat{Y}$ et $\widehat{\varepsilon}$:

$$\widehat{Y} = X\widehat{\beta},$$

$$\widehat{\varepsilon} = Y - \widehat{Y}.$$

— Estimer la variance génétique à partir de la matrice X centrée (notée X_c) :

$$\widehat{\Sigma}_G = \widehat{\beta}^\top (X_c^\top X_c)\widehat{\beta}.$$

— Estimer la variance résiduelle :

$$\widehat{\Sigma}_\varepsilon = \frac{1}{n-q-1}(Y - \widehat{Y})^\top (Y - \widehat{Y}).$$

— Faire la décomposition de Choleski de la variance résiduelle :

$$\widehat{\Sigma}_\varepsilon = LL^\top.$$

— Faire la décomposition en valeurs propres de la matrice :

$$L^{-1}\widehat{\Sigma}_G (L^{-1})^\top \equiv PDP^\top.$$

— Calculer le vecteur des poids W :

$$W = (L^{-1})^\top V,$$

avec V le vecteur propre de $L^{-1}\widehat{\Sigma}_G(L^{-1})^\top$ associé à la valeur propre λ_1.

— Calculer le nouveau score $y^* = W^\top Y$ qui a une héritabilité maximale.

L'héritabilité associée à $y^* = W^\top Y$ est donnée par $\dfrac{\lambda_1}{1+\lambda_1}$.

L'application de l'algorithme de PCH_g résumera l'information génétique des L SNPs dans y^*. Par la suite, nous allons ajuster un modèle de régression pour le nouveau caractère y^* versus les L SNPs comme suit

$$y_i^* = \beta_0^* + \beta_1^* x_{i1} + \beta_2^* x_{i2} + \cdots + \beta_L^* x_{iL} + \varepsilon_i, \qquad i = 1, \cdots, n,$$

où y_i^* représente la $i^{ème}$ valeur du vecteur y^*, et x_{il} est le nombre d'allèles mineurs que possède l'individu i pour le SNP l, $l = 1, \ldots, L$. ε_i est le terme d'erreur. Sous

forme matricielle, le modèle de régression linéaire multiple s'écrit :

$$y^* = X\beta^* + \varepsilon,$$

avec

$$y^* = [y_1^* \cdots y_n^*] \quad \text{et} \quad \beta^* = [\beta_0^* \beta_1^* \cdots \beta_L^*]^\top.$$

Pour tester la signification globale de la régression, on définit les hypothèses :

H_0 : aucune des variables n'a d'effet sur y^*

$$\forall l \in \{1, \ldots, L\}, \quad \beta_l = 0.$$

H_1 : au moins une des variables à un effet sur y^*

$$\exists l \in \{1, \ldots, L\}, \text{tel que } \beta_l \neq 0.$$

Pour conclure s'il y a une association significative entre les marqueurs et la composante principale d'héritabilité, nous nous basons sur la statistique F de Fisher :

$$F_W^{obs} = \frac{n - L - 1}{L} \frac{SCE}{SCR}, \tag{3.3}$$

où

$$SCE = \sum_{i=1}^{n} (\widehat{y}_i^* - \overline{y})^2,$$

avec

$$\widehat{y}^* = X\widehat{\beta}^* \text{ et } \overline{y} = \frac{\sum_{i=1}^{n} y_i^*}{n}$$

$$SCR = \sum_{i=1}^{n} (y_i^* - \widehat{y}_i^*)^2.$$

Note : SCE est la somme des carrés expliquée par le modèle, SCR est la somme des carrés résiduelle et n le nombre d'observations.

Généralement, lors du test F, on suppose que les données (ou les erreurs aléatoires) suivent une distribution normale. On propose ici un test de permutation, car on utilise les données deux fois : une fois pour estimer W et une deuxième fois pour calculer le F dans le modèle du PCH_g. Ainsi, F_W^{obs} ne suit plus une distribution de Fisher

$F_{L,n-L-1}$. Cette procédure de test contrôle l'erreur globale de type 1, tout en évitant d'une part l'utilisation d'une méthode de fractionnement de l'échantillon compliquée et d'autre part un temps de calcul long. Plus de détails sur cette technique sont donnés dans la section suivante.

3.2 Test de permutation

Le test d'association entre le marqueur et le phénotype porte sur le paramètre de régression β ($\beta = 0$ vs $\beta \neq 0$). Sous l'hypothèse nulle H_0 ($\beta = 0$), il n'y a pas d'association entre le marqueur et le phénotype et $Y_{ij} = \mu_j$. Pour faire ce type de test, on a besoin de déterminer la distribution de la statistique de test sous H_0. Cependant, la distribution de cette statistique sous l'hypothèse nulle peut s'avérer compliquée. Ainsi, le calcul de la puissance et de l'erreur de type 1 s'avère difficile. Pour remédier à un tel problème, nous pouvons faire appel à des techniques de rééchantillonage, comme par exemple les techniques de Bootstrap et de permutation. Dans ce qui suit, nous allons introduire des techniques de permutation pour estimer la distribution d'une statistique de test sous l'hypothèse nulle. Une telle technique est utilisée pour le calcul de la puissance de notre statistique, F_W^{obs}, proposée dans ce chapitre. Les tests de permutation que nous présentons ici sont des procédures basées sur des réarrangements des étiquettes d'un ensemble de données (Edgington, 1980) et qui peuvent être utilisées pour approximer la distribution de la statistique du test sous l'hypothèse nulle. Nous illustrons la méthodologie de cette approche dans un cadre général. Dans un test de permutation, la statistique observée du test, Q_{obs}, est obtenue à partir des données observées, et comparée à des statistiques Q_b obtenues à partir de B jeux de données permutées. Autrement dit, on effectue B permutations pour les données originales et on calcule B statistiques du test de

la même façon que Q_{obs}. La permutation des étiquettes des données nous assure que les statistiques calculées suivent la distribution sous l'hypothèse nulle. Ainsi, la proportion des valeurs de Q_b, $b = 1, 2, \ldots, B$ qui sont supérieures ou égales à Q_{obs} estime la p-valeur théorique de notre test. On la note p-valeur et on écrit

$$\text{p-valeur} = \frac{\sum_{b=1}^{B} I(Q_b \geq Q_{obs})}{B},$$

où I est la fonction indicatrice. On rejette H_0 lorsque la p-valeur est inférieure ou égale à un seuil fixé d'avance, par exemple 5%.

Les tests de permutation présentent l'avantage de créer un seuil directement à partir des données, mais, malheureusement, ils sont intensifs en calcul, car il nécessite le calcul d'un grand nombre de permutations possibles pour atteindre une bonne estimation de niveau de signification. Pour surmonter ce problème de calcul intensif, Knijnenburg *et al.* (2009) ont proposé une approche pour estimer la p-valeur avec moins de permutations ($N \ll B$), où N représente le nombre nécessaire de permutations. Cette méthode consiste à modéliser la queue de la distribution des valeurs permutées par la distribution de Pareto généralisée (Gumbel, 1958), notée GPD. La fonction de distribution cumulative empirique de Pareto généralisée s'écrit sous la forme :

$$F_{\zeta,\delta}(z) = \begin{cases} 1 - (1 + \zeta\frac{z}{\delta})^{-\zeta^{-1}} & \text{si } \zeta \neq 0, \\ 1 - exp\left(\frac{-z}{\delta}\right) & \text{si } \zeta = 0, \end{cases}$$

où δ est le paramètre d'échelle et ζ est le paramètre de forme, et

$$\begin{cases} z \in (0, \infty) & \zeta \leqslant 0, \\ z \in \left(0, \frac{\delta}{\zeta}\right) & \zeta > 0. \end{cases}$$

La distribution de Pareto généralisée sera ajustée pour les valeurs $Z = z_1^*, \ldots, z_{N_{exc}}^*$ telles que $z_i^* = Q_i^* - t, \forall i : Q_i^* > t$, où Q^* sont les statistiques ordonnées obtenues par permutation des données $(Q_1^* > Q_2^*)$, t est un seuil d'excédent et N_{exc} est le nombre de statistiques Q_i^* qui satisfont $Q_i^* > t$.
Les p-valeurs des tests de permutation selon l'approximation de la distribution $F_{\zeta,\delta}(z)$ de Pareto sont calculées comme suit :

$$P_{gpd} = \frac{N_{exc}}{N}(1 - F_{\widehat{\zeta},\widehat{\delta}}(Q_{obs} - t)), \tag{3.4}$$

où $\widehat{\zeta}$ et $\widehat{\delta}$ sont les estimateurs de maximum de vraisemblance de ζ et δ obtenus en ajustant la distribution de Pareto généralisée au jeu de données $\{z_1^*, \ldots, z_{N_{exc}}^*\}$. Des précisions importantes qui pourraient intéresser les lecteurs et qui justifient le choix de la distribution de Pareto généralisée pour modéliser la queue de la distribution se trouvent entre autres dans Raggad (2009).

La procédure de choix du nombre d'excédents

On note $Q_1^* > \cdots > Q_N^*$ les statistiques d'ordre[1] relatives à $Q_1, \cdots, Q_N$ et on fixe $N_{exc} = 250$. Knijnenburg *et al.*, (2009) justifient le choix de nombres d'excédent par le fait que l'approximation de la queue par la distribution de Pareto généralisée est souvent utilisée pour l'extrapolation[2]. On choisit ensuite t comme

---

1. Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires supposées i.i.d. On appelle statistique d'ordre le vecteur X^* obtenu en ordonnant dans l'ordre croissant l'échantillon :
$X_1 \leqslant X_2 \leqslant \ldots \leqslant X_n$.
On pose $X_{(1)}$= min$(X_1, X_2, \ldots, X_n)$ et $X_{(n)}$=max$(X_1, X_2, \ldots, X_n)$.

2. En mathématiques, l'extrapolation est le calcul d'un point d'une courbe dont on ne dispose pas d'équation, à partir d'autres points, lorsque l'abscisse du point à calculer est au-dessus du

$$t = \frac{Q^*_{N_{exc}} + Q^*_{N_{exc}+1}}{2}.$$

On ajuste la distribution de Pareto généralisée pour les 250 statistiques $z_i^* = Q_i^* - t$ suivi d'un test d'ajustement pour s'assurer que la distribution Pareto généralisée est un bon modèle pour la queue de la distribution des statistiques permutées Q_i^*. Si le test d'ajustement ne rejette pas l'hypothèse nulle, H_0 $z_i^* \sim Pareto$, pour un seuil nominal $\alpha = 0.05$, on retient $N_{exc} = 250$ et on calcule P_{gpd} comme dans l'équation (3.4). Sinon, nous réduisons de 10 le nombre d'excédent ($N_{exc} = 240$) jusqu'à ce que la distribution de Pareto généralisée s'ajuste bien aux $\{z_1^*, \ldots, z_{N_{exc}}^*\}$.

Algorithme pour le choix du nombre d'excédents (3.2).

— **Initialisation** On pose $N_{exc} = 250$.

— **Étape 1**

On teste l'hypothèse H_0 contre l'hypothèse alternative H_1 :

H_0 : La distribution de Pareto généralisée est un bon modèle pour les Z_i^* ;

H_1 : La distribution de Pareto généralisée n'est pas un bon modèle pour les Z_i^*.

— **Étape 2**

Si l'hypothèse H_0 est n'est pas rejetée, alors $N_{exc} = 250$ et la procédure se termine. Si l'hypothèse H_1 est rejetée, on reprend le calcul à la première étape avec $N_{exc} = N_{exc} - 10$.

— **Étape 3**

Reprendre les étapes 1 et 2 jusqu'au non rejet de H_0 ou jusqu'à ce que $N_{exc} = 10$.

---

maximum ou en dessous du minimum des points connus. Source Wikipédia

L'algorithme proposé par Knijnenburg *et al.* (2009) pour estimer les p-valeurs dans les tests de permutation

Une approche standard pour approximer la p-valeur permutationnelle notée "p-valeur", sans s'intéresser principalement à la queue de la distribution des valeurs permutées, est basée sur la fonction de distribution cumulative empirique donnée comme suit

$$\text{p-valeur} = P_{ecdf} = \frac{\sum_{b=1}^{N} I(Q_b \geq Q_{obs})}{N}. \tag{3.5}$$

De plus, pour montrer l'utilité de l'estimation de la p-valeur par P_{gpd} au lieu de P_{ecdf} Knijnenburg *et al.* (2009) proposent d'utiliser $T^* = \sum_{b=1}^{N} I(Q_b \geq Q_{obs})$ comme critère permettant de choisir entre les deux. Knijnenburg *et al.* (2009) ont montré que pour $T^* \geqslant 10$, le P_{ecdf} peut être un bon estimateur de la p-valeur théorique. Ainsi, ils ont proposé l'algorithme suivant :

Algorithme de Knijnenburg *et al.* (2009) (3.2).

Requiert : F_b : ensemble de statistiques de Fisher obtenues après chaque permutation, F_{obs} : statistique de Fisher calculée des données non permutées et N_{exc} est le nombre d'excédent fixé au début à 250.

1 : Donner le nombre de permutations à utiliser N.

2 : Calculer les statistiques permutées.

3 : Appeler l'algorithme de choix du nombre d'excédents.

4 : Calculer $T^* = \sum_{b=1}^{N} I(F_b \geqslant F_{obs})$; si $T^* \geq 10$ alors p-valeur $= P_{ecdf}$, sinon p-valeur$= P_{gpd}$.

Notez que nous avons adopté l'algorithme ci-dessus pour notre méthode avec plusieurs caractères et plusieurs SNPs à la fois. L'adaptation de cet algorithme à l'approche de Klei se fait de façon similaire puisque cette dernière est un cas particulier de notre méthode. Pour illustrer l'approximation $ECDF$ et GPD, 5000 valeurs permutées ont été générées aléatoirement à partir de la distribution de Fisher (voir figure 3.1). Dans la figure 3.2, de gauche, les valeurs permutées qui excèdent Q_{obs}=5 vont définir l'excédent et sont modélisées à l'aide de GPD. Q_{obs} est la statistique de test obtenu à partir de données non permutées. L'approximation GPD de la queue (mise à l'échelle à l'intervalle $\left[\left(1-\frac{N_{exc}}{N}\right),1\right]$) est représentée aux côtés de la fonction de distribution cumulative théorique dans la figure 3.2 de droite. La p-valeur théorique, qui est dérivée de la fonction de distribution cumulative de la distribution théorique Fisher (notée P_f) est comparée avec l'approximation de ECDF (notée P_{ecdf}) et l'approximation GPD (notée P_{gpd}) pour des valeurs qui excèdent $Q_{obs} = 5$.

Adaptation de l'algorithme de Knijnenburg *et al.* (2009) à notre approche PCH_g

On rappelle que pour tester la signification globale du modèle, nous nous basons sur la statistique F de Fisher (3.3). F_{obs} est la statistique d'origine ou de référence déduite de notre approche (sans permutation). Ensuite, on calcule la statistique de test F_b $(b = 1, \ldots, N)$ sur les échantillons permutés. La p-valeur permutationnelle sera déduite de l'algorithme décrit précédemment en considérant F_{obs} et F_b $(b = 1, \ldots, N)$ comme paramètres d'entrée de l'algorithme. On rejette H_0 lorsque la p-valeur est inférieure ou égale à un seuil nominal fixé d'avance.

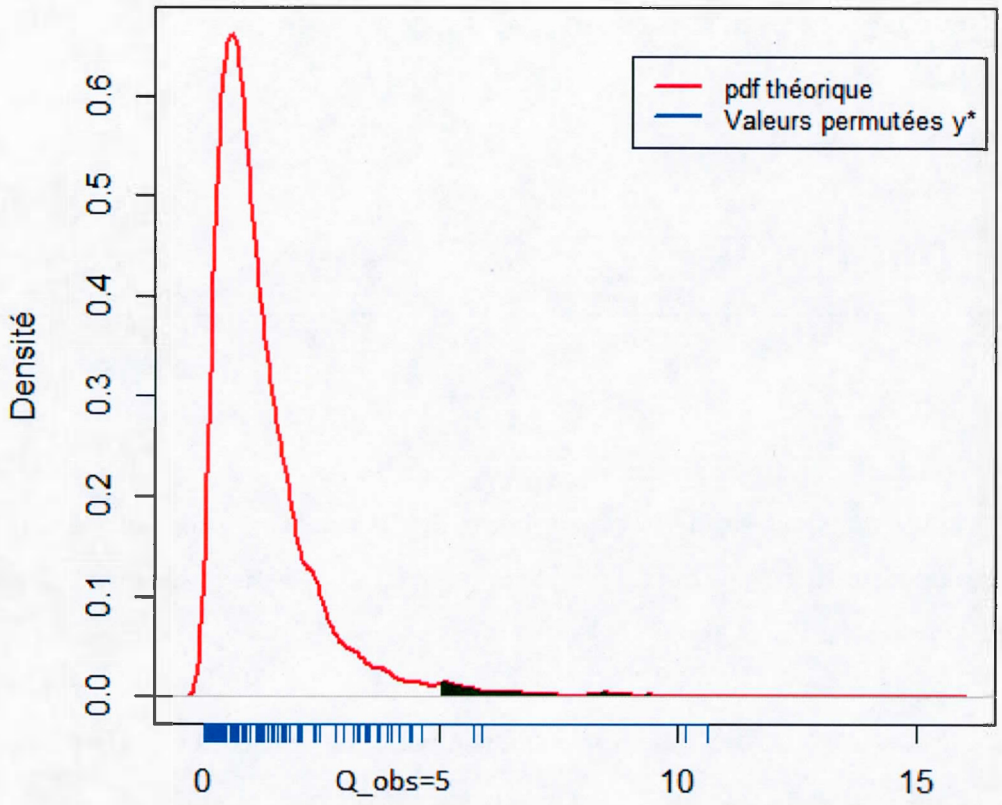


Figure 3.1: 5000 valeurs permutées ont été générées aléatoirement à partir de la distribution de Fisher à (5,10) degrés de liberté, la zone verte représente les valeurs permutées qui excédent $Q_{obs} = 5$ où Q_{obs} est la statistique de test obtenue à partir de données non permutées.

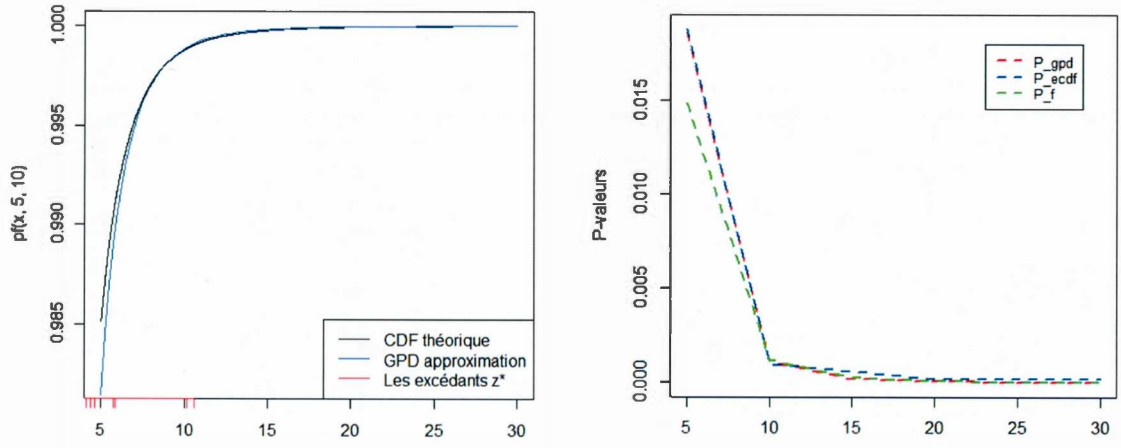


Figure 3.2: L'approximation de la queue d'une distribution de Fisher par GPD et ECDF. La p-valeur théorique, qui est dérivée de la fonction de distribution cumulative de la distribution de Fisher (notée P_f) est comparée avec l'approximation de ECDF (notée P_{ecdf}) et l'approximation GPD (notée P_{gpd}) pour des valeurs qui excédent $Q_{obs} = 5$. Q_{obs} est la statistique de test obtenue à partir de données non permutées.

3.3 PCH : effet de dépendance entre les marqueurs

La colinéarité des variables explicatives, dans un modèle de régression, exprime une situation où ces variables sont hautement corrélées les unes aux autres. Le terme multicolinéarité est également utilisé pour désigner ce genre de situation. La colinéarité se manifeste par :

— des valeurs propres nulles de la matrice de corrélation issue de la matrice X ;

— de petites variations dans les colonnes de X ou les Y peuvent produire de

grandes variations dans les estimations des coefficients de pente ;
— des estimations de pentes dont la variance est grande ;
— en outre, quand les les variables indépendantes sont fortement corrélés, le R^2 d'une régression peut être grand alors que les estimations de pentes individuelles ne sont pas statistiquement significatives (Neter *et al.*, 1996).

En génétique, la colinéarité s'exprime par le déséquilibre de liaison élevé entre les marqueurs. Dans notre approche, un tel déséquilibre de liaison entre les SNPs peut causer la colinéarité dans la matrice X de notre modèle (3.1).

3.3.1 L'effet du déséquilibre de liaison sur l'héritabilité

Si la corrélation entre un SNP et un ou plusieurs SNPs de son voisinage est forte ($r^2 > 0.8$) alors son signal peut être reproduit, et cela peut conduire à une surestimation de sa contribution à l'héritabilité du caractère sous étude (Speed *et al.*, 2012). L'approche que nous avons proposée utilise toute l'information génotypique dans la matrice des marqueurs. Ainsi, la possibilité d'avoir un déséquilibre de liaison est fort possible. Ce déséquilibre de liaison peut se traduire par une multicolinéarité entre les marqueurs et par conséquent par la singularité de la matrice X. Dans les études d'association génétique, une façon pour corriger ce problème consiste à supprimer un des deux marqueurs qui présentent un fort déséquilibre de liaison ($r^2 > 0.80$) car si la corrélation entre un SNP et un ou plusieurs SNPs de son voisinage est forte, l'information apportée par tous les SNPs sera la même que celle apportée par un seul de ces SNPs (Johnson *et al.*, 2001 ; Patil *et al.*, 2001) : on parle dans ce cas de tagSNP. L'idée est de minimiser le nombre de SNPs tout en maintenant un bon niveau de couverture de la variabilité. Ainsi, le nombre de SNPs peut être considé-

rablement réduit sans beaucoup de perte de puissance pour les études d'association (Zhang *et al.*, 2002). Généralement, le seuil de corrélation de tagSNP est supérieure à 0.8 ($r^2 > 0.8$).

Dans notre approche, une solution pour corriger le problème de multicolinéarité entre les marqueurs qui ont une corrélation forte mais inférieure à 0.8 serait d'utiliser l'ACP pour la matrice X et régresser les premières composantes principales expliquant une proportion importante de la variation dans X (détail 3.3.2). Le déséquilibre de liaison peut aisément être visualisé à l'aide de la carte triangulaire dite "Carte thermique" (voir figure 3.3).

3.3.2 PCH_g : l'apport de l'ACP au problème de multicolinéarité

Pour limiter les difficultés dues à la colinéarité qui peut exister entre les marqueurs, d'autres estimateurs des coefficients de régression existent, parmi eux la régression orthogonale ou régression sur composantes principales. La régression orthogonale est basée d'abord sur la décomposition préliminaire de la matrice X en valeurs singulières et ensuite sur la régression par moindre carrés ordinaires sur les composantes principales. L'idée générale de chaque étape est donnée comme suit

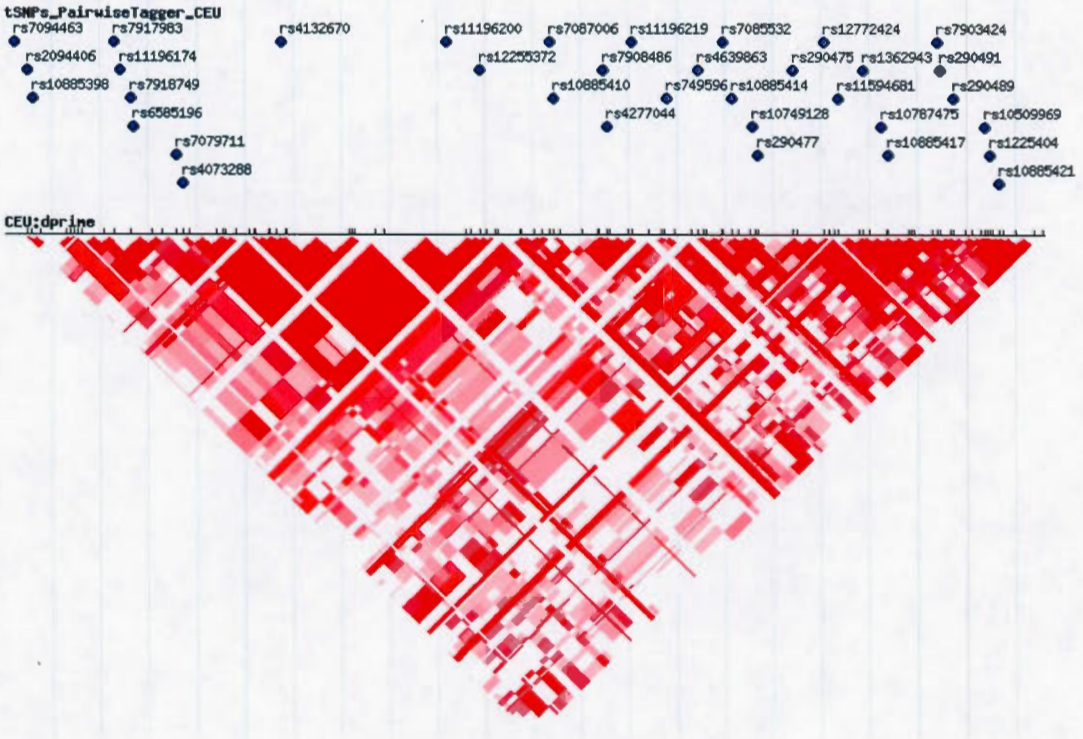
Étape 1 : décomposition en valeurs singulières de X.

Puisque la matrice $X^\top X$ est une matrice symétrique, nous pouvons écrire

$$X^\top X = P\Lambda P^\top, \tag{3.6}$$

où P est la matrice des vecteurs propres normalisés de $X^\top X$, c'est-à -dire que P est une matrice orthogonale ($P^\top P = PP^\top = I$) et $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ est la matrice diagonale des valeurs propres classées par ordre décroissant.

Figure 3.3: Carte thermique («Heat map») du déséquilibre de liaison du gène TCF7L2 (Chromosome 10) basée sur la mesure D'. Le déséquilibre de liaison est évalué à l'aide de la matrice triangulaire des dépendances (mesurées par D' pour chaque paire de SNP). Pour une paire de SNP donnée, plus la couleur est sombre, plus les SNPs sont corrélés. Les tagSNPs identifiés pour ce gène sont donnés aussi dans la figure. Source : International HapMap Project, source de données : HapMap Data Rel 27 PhaseII+III, Feb09, on NCBI B36 assembly, dbSNP b126.



En remplaçant X par $XPP^\top$ nous avons

$$Y = XPP^\top\beta + \varepsilon.$$

Étape 2 : régression par moindre carrés ordinaires sur les composantes principales.

Le modèle de régression peut être réécrit de la façon suivante :

$$Y = X^*\beta^* + \varepsilon, \tag{3.7}$$

où $\beta^* = P^\top \beta$ et $X^* = XP$. Les colonnes de la matrice X^* sont les composantes principales avec de meilleures propriétés (orthogonalité et variance maximale expliquée). Cela permet d'éviter par conséquent le problème d'inversion de la matrice.

L'estimateur des moindres carrés vaut alors

$$\widehat{\beta^*} = (X^{*\top} X^*)^{-1} X^{*\top} Y. \tag{3.8}$$

L'analyse en composante principale de l'héritabilité généralisée va utiliser la nouvelle matrice X^* comme matrice des marqueurs. Cependant, rien ne garantit que ces composantes soient pertinentes pour expliquer la matrice des phénotypes.

CHAPITRE IV

ÉTUDES DE SIMULATION

4.1 Scénarios de simulation

Pour détecter les SNPs associés aux caractères d'intérêt, nous avons précédemment formulé l'hypothèse que l'analyse conjointe des SNPs peut donner une puissance statistique plus élevée que l'analyse d'un SNP à la fois. Pour illustrer la méthodologie proposée, nous allons faire une comparaison de la puissance de notre nouvelle méthode introduite au chapitre 3 et des méthodes concurrentes, entre autres la méthode de Klei et la méthode univariée. Nous allons aussi mener d'autres études de simulation : une étude va porter sur l'estimation des p-valeurs par l'algorithme de Knijnenburg *et al.* (2009) dans l'approche PCH_g, une autre va porter sur le cas du déséquilibre de liaison entre les SNPs, afin de visualiser l'impact du déséquilibre de liaison sur la puissance de notre méthode PCH_g. Plusieurs scénarios de simulation ont été conduits, selon différentes valeurs des paramètres suivants :

— nombre de sujets $n = 500$;

— nombre de marqueurs génétiques inclus dans le modèle $l = 10$;

— nombre de caractères $q = 5$;

— le déséquilibre de liaison entre les marqueurs $r^2 \in \{0, 0.2, 0.4, 0.6, 0.8\}$;

— les caractères simulés présentent des héritabilités inférieures à 0.017 ($h^2 \in [0.00021, 0.017]$).

Pour tous les scénarios, le modèle statistique utilisé pour générer les caractères est donné par

$$Y = X\beta + \varepsilon, \tag{4.1}$$

où la matrice X inclut l'ordonnée à l'origine ainsi que plusieurs SNPs codés tels que $x_l \in \{0, 1, 2\}$ et le terme d'erreur ε est la réalisation d'une loi normale multivariée de dimension q de moyenne nulle et de matrice variances-covariances Σ_ε (noté $\varepsilon \sim MVN_q(0, \Sigma_\varepsilon)$). Les propriétés statistiques évaluées (erreur de type 1 et puissance) ont été calculées en se basant sur 1000 réplications :

(a) l'erreur de type I empirique est définie comme la moyenne du nombre de réplications révélant une association significative avec la composante principale d'héritabilité correspondante lorsque l'hypothèse nulle est vraie ;

(b) la puissance a été calculée comme étant la moyenne du nombre de réplications, parmi les 1000, révélant une association significative avec la composante principale d'héritabilité correspondante, lorsque l'hypothèse nulle est fausse.

4.1.1 Simulation des données à partir du modèle PCH

Dans cette section, nous allons expliquer comment les données génotypiques et les caractères vont être simulés pour être utilisés ensuite dans l'évaluation des performances de notre approche par rapport à l'approche de Klei et aux analyses univariées.

Considérons le scénario de simulation suivant : nous avons considéré cinq caractères ($j = 1, \ldots, 5$) corrélés positivement entre eux ($cov(y_j, y_{j'}) > 0$). Dix variantes communes corrélées vont être générées selon une distribution normale multivariée pour

un échantillon de 500 observations :

$$x_i \sim MVN_{10}(0, \Sigma_x) \quad \text{pour} \quad i = 1, \ldots, n,$$

où Σ_x est la matrice de variances-covariances de x. Notez que le choix de simulation de la corrélation entre les SNPs par la loi normale multivariée est choisie seulement en raison, d'une part de sa simplicité et d'autre part parce qu'on peut manipuler la corrélation entre les SNPs, chose qui n'est pas évidente avec la distribution multinomiale par exemple, mais d'autres techniques existent (voir Ferrari et Barbiero, 2012). Ensuite, les colonnes de la matrice $x = (x_1, \ldots, x_n)^\top$ vont être converties en génotypes codés en $\{0, 1, 2\}$ par une procédure de discrétisation basée sur les probabilités 90% et 95%. Ces quantiles sont sélectionnés pour avoir la fréquence de l'allèle mineur (p_l) des 10 SNPs qui varie entre 0.055 et 0.088. La procédure est résumée dans l'algorithme suivant avec $i = 1, \ldots, 500$ l'indice des sujets, $l = 1, \ldots, L = 10$ l'indice des marqueurs dans l'étude et q_1 et q_2 sont les quantiles 90% et 95% de x_j respectivement.

Procédure de discrétisation de la matrice x

- Générer 10 marqueurs selon la distribution normale multivariée $x_i \sim MVN(0, \Sigma_x)$ $(i = 1, \ldots, n)$. Poser $x = (x_1, \ldots, x_n)^\top$.
- Pour chaque marqueur $l = 1, \ldots, L$, calculer q_1 et q_2 pour chaque colonne de x.
- Si $x_{il} \leq q_1$ alors $x_{il} = 0$.
- Si $q_1 < x_{il} \leq q_2$ alors $x_{il} = 1$.
- Si $x_{il} > q_2$ alors $x_{il} = 2$.

Cet algorithme de discrétisation sera utilisé chaque fois que les SNPs sont générés à partir d'une distribution normale multivariée. Une fois que la matrice x est générée, l'effet de chaque variante génétique ($l = 1, \ldots, 10$) sur chaque caractère, représenté

par β_{jl}, est calculé selon la procédure suivante :

- à partir de l'équation (3.2) l'héritabilité d'un seul caractère j est

$$h^2_{jl} = \frac{Var(X\beta)}{Var(Y)} = \frac{2p_l(1-p_l)\beta^2_{jl}}{2p_l(1-p_l)\beta^2_{jl} + \sigma^2_{\varepsilon_{jl}}},$$

où β_{jl} est l'effet du SNP l sur le caractère j et p_l est la fréquence de l'allèle mineure du SNP l.

- on résout cette équation pour β_{jl}, on trouve l'effet de la variante l sur le caractère j égal à :

$$\beta^2_{jl} = \frac{h^2_{jl}}{(1-h^2_{jl})2p_l(1-p_l)}.$$

Un exemple de valeurs considérées pour h^2 est donné dans la matrice h^2 en bas et est schématisé dans la figure 4.1.

$$h^2 = \begin{pmatrix} 0.001 & 0.002 & 0 & 0 & 0 \\ 0.0006 & 0.001 & 0 & 0 & 0 \\ 0.0004 & 0.00021 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0005 & 0.01 & 0.003 \\ 0 & 0 & 0.0015 & 0.0027 & 0.008 \\ 0 & 0 & 0.0025 & 0.013 & 0.001 \\ 0 & 0 & 0.001 & 0.0026 & 0.002 \\ 0 & 0 & 0.004 & 0.0015 & 0.016 \\ 0 & 0 & 0.0023 & 0.01 & 0.017 \end{pmatrix}.$$

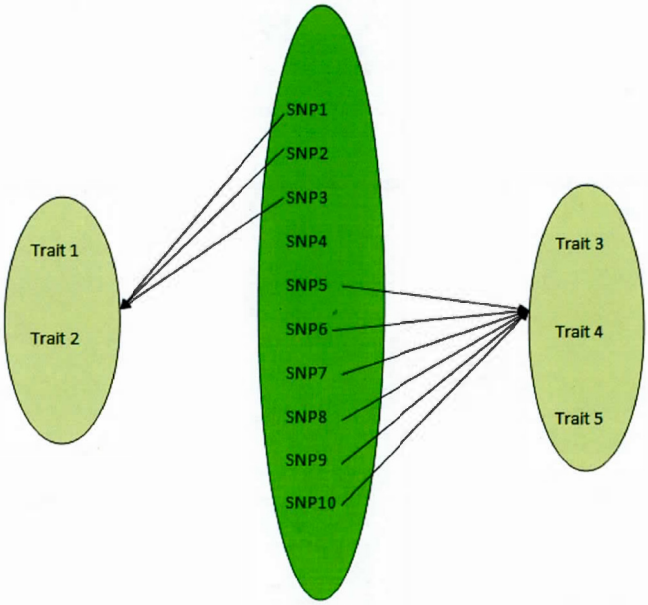


Figure 4.1: La figure présente les effets pléiotropiques introduits dans le scénario de simulation. L'ellipse à gauche inclut les caractères 1 et 2 qui partagent trois SNPs en commun (1, 2 et 3). L'ellipse au milieu présente les 10 SNPs considérés dans la simulation. L'ellipse à droite montre les caractères 3, 4 et 5 qui partagent en communs six SNPs (5, 6, 7, 8, 9 et 10). Le SNP4 n'est associé à aucun caractère.

La matrice de caractères correspondante est alors générée selon le modèle

$$Y_{500\times5} = X_{500\times11}\beta_{11\times5} + \varepsilon_{500\times5},$$

avec

$$\varepsilon_{.j} \sim MVN_5(0, \Sigma_\varepsilon),\ \varepsilon_{.j} \text{ est la } j^{ème} \text{ colonne de } \varepsilon_{500\times5},$$

et

$$\Sigma_\varepsilon = \begin{pmatrix} 0.25 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.25 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.75 & 0.55 \\ 0.00 & 0.00 & 0.75 & 1.00 & 0.63 \\ 0.00 & 0.00 & 0.55 & 0.63 & 1.50 \end{pmatrix}.$$

4.1.2 Méthodes : analyse univariée, PCH de Klei et PCH_g

Pour évaluer l'inflation de l'erreur de type 1 et la puissance des analyses univariées, de l'approche PCH de Klei et de notre nouvelle approche PCH_g pour le modèle 4.1.1, 1000 réplications vont être générées. Nous allons comparer d'abord l'analyse univariée et l'approche de Klei, et l'approche qui a une puissance plus élevée sera ensuite comparée à notre approche PCH_g.

L'analyse univariée

L'analyse univariée est une approche simple souvent utilisée dans le cadre des analyses d'association. Elle consiste à tester l'association de chaque caractère avec un SNP à la fois. Cependant, cela implique la réalisation d'un grand nombre de tests, et par conséquent cela peut engendrer un grand taux de faux positifs et une perte de puissance du test. Le nombre total des tests effectués lors de ce scénario de simulation est 50. La correction de Bonferroni citée dans l'article de Klei *et al.* (2008) est utilisée pour contrôler le seuil global du test.

L'approche PCH de Klei

Pour chaque réplication, on applique l'algorithme de PCH de Klei pour obtenir la p-valeur pour chaque SNP considéré, où la p-valeur est estimée par la fonction de distribution cumulative empirique, P_{ecdf} décrite dans la section 3.2 équation (3.5).

On rappelle que l'approche PCH de Klei est susceptible à l'inflation de l'erreur du type 1 parce que les mêmes données sont utilisées deux fois, une fois pour maximiser l'héritabilité estimée (le PCH) et une deuxième fois pour tester l'association entre le PCH et le SNP. On utilise les tests de permutation décrits en section 3.2 pour corriger cette inflation, et pour contrôler le seuil global du test, on utilise la correction de Bonferroni.

L'approche PCH_g

L'approche PCH_g que nous avons développée est susceptible comme celle de Klei à l'inflation de l'erreur de type 1 parce que les mêmes données sont utilisées deux fois, une fois pour maximiser l'héritabilité estimée et une deuxième fois pour tester l'association entre le PCH et les SNPs. On utilise les tests de permutation décrits en section 3.2 pour corriger cette inflation. L'approche PCH_g a l'avantage, à l'inverse des deux méthodes précédentes, de ne pas avoir le problème de tests multiples, alors on n'aura pas besoin d'utiliser la correction Bonferroni pour corriger le seuil global du test.

4.2 Résultats

4.2.1 Comparaison de l'approche PCH de Klei versus analyse univariée

Nous présentons ici les résultats des simulations effectuées en utilisant l'approche de Klei et l'analyse univariée pour dix SNPs générés selon le modèle décrit en 4.1.1 et en fixant le niveau de corrélation entre les SNPs à 0.2. On se contente pour illustrer les deux approches précédentes de montrer les résultats pour deux SNPs sélectionnés, le SNP7 et 9. Le résultat de l'application de deux approches sous l'hypothèse nulle est

donné dans la figure 4.2. Les tests de permutation contrôlent bien l'erreur de type 1, les courbes de détection pour le SNP7 et le SNP9 pour les deux approches sont très proches de la droite $y = x$ coupant l'axe des ordonnées y à 0 et dont la pente est égale à 1. La puissance de l'approche de PCH Klei versus celles des tests univariés est

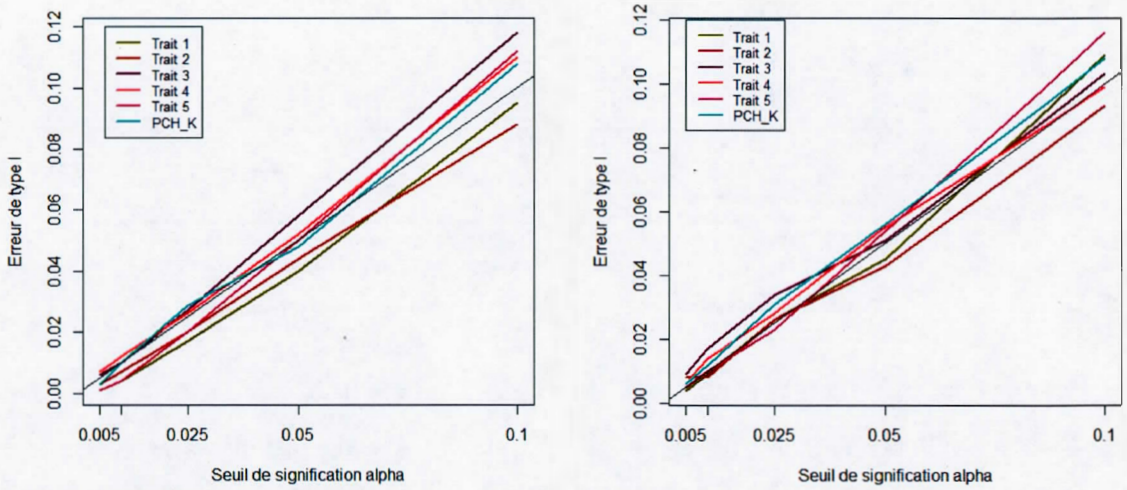


Figure 4.2: La figure à gauche présente le résultat de l'erreur de type 1 du PCH de Klei versus tests univariés pour le SNP7, et la figure à droite présente celle du SNP9. L'axe des abscisses représente le seuil de signification $\alpha \in \{0.005, 0.01, 0.025, 0.05, 0.1\}$, l'axe des ordonnées présente les probabilités empiriques sous H_0 dites probabilités d'erreur du type 1. La courbe PCH_K désigne les p-valeurs de la composante de l'héritabilité déduites selon l'algorithme de Klei sous H_0 pour un seuil α. Les 5 autres courbes représentent les résultats de l'analyse univariée par rapport à chaque caractère. La ligne de référence avec l'ordonnée à l'origine égale à 0 et la pente égale à 1 est également représentée.

donnée dans la figure 4.3. Un gain de puissance de détection des variantes causales

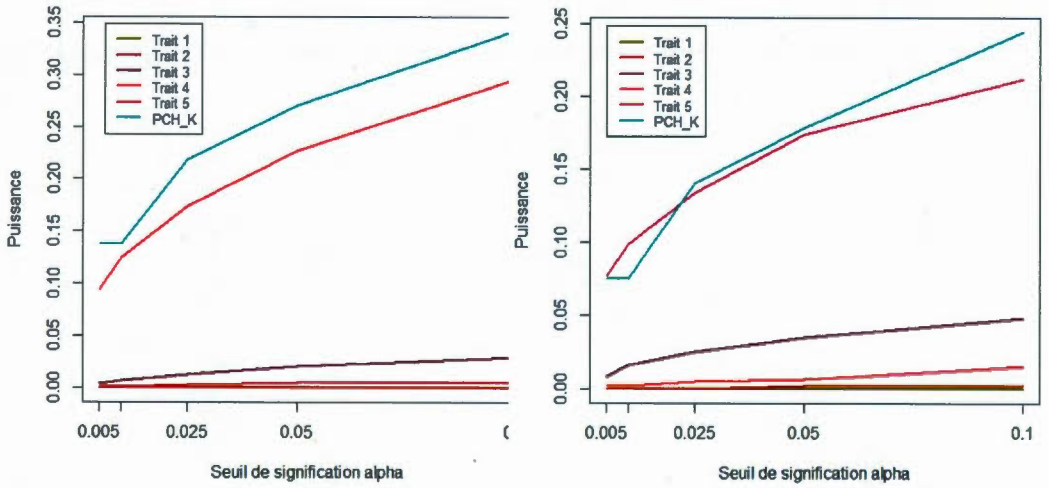


Figure 4.3: Puissance de l'approche PCH de Klei versus tests univariés pour les SNPs 7 (figure à gauche) et 9 (figure à droite). L'axe des abscisses représente le seuil de signification $\alpha \in \{0.005, 0.01, 0.025, 0.05, 0.1\}$, l'axe des ordonnées présente la puissance sous H_1. La courbe PCH_K désigne les p-valeurs de la composante de l'héritabilité déduites selon l'algorithme de Klei sous H_1 pour un seuil α. Les 5 autres courbes représentent les résultats de l'analyse univariée par rapport à chaque caractère.

est observé en utilisant l'approche PCH de Klei par rapport à l'analyse individuelle des SNPs, mais cette puissance reste petite parce que nous avons choisi des valeurs petites pour h^2.

L'identification des SNPs qui contribuent à des caractères communs fournira des informations de diagnostics très précieuses car elle facilitera le diagnostic précoce, et

permettront des moyennes thérapeutiques plus efficaces. Malheureusement, cet objectif est difficile à atteindre avec l'analyse univariée qui ne prend pas en considération la corrélation entre les caractères. L'approche PCH de Klei peut répondre à cette requête. Dans le prochain scénario de simulation, nous allons évaluer l'algorithme Knijnenburg *et al.* (2009) dans l'approche PCH_g.

4.2.2 PCH_g et l'algorithme de Knijnenburg *et al.* (2009)

On applique l'algorithme de PCH_g dans le modèle généré comme décrit dans 4.1.1 où le niveau de corrélation fixé à 0.2 avec des estimations des p-valeurs d'une part par ECDF, et d'autre part par l'algorithme de Knijnenburg *et al.* (2009) (voir chapitre 3 Section 3.2).

Les résultats sont donnés dans la figure 4.4. Les différences entre les deux courbes représentent le choix de l'algorithme d'utiliser la distribution de Pareto généralisée au lieu de la fonction de distribution cumulative empirique (ECDF) pour estimer les p-valeurs lorsque le nombre de valeurs permutées qui excèdent la statistique observée du test est inférieur à 10. Cette implémentation de l'algorithme de Knijnenburg *et al.* (2009) a l'avantage de bien contrôler l'erreur de type 1. Dorénavant, l'algorithme Knijnenburg *et al.* (2009) sera utilisé dans les deux approches dans ce qui suit.

4.2.3 Comparaison de l'approche PCH_g avec l'approche de Klei *et al.* (2008)

Afin de valider le modèle d'analyse simultanée de plusieurs SNPs, nous supposons premièrement un modèle d'analyse avec des SNPs non corrélés (noté M1), et deuxièmement un modèle d'analyse avec des SNPs corrélés (noté M2). Un échantillon de

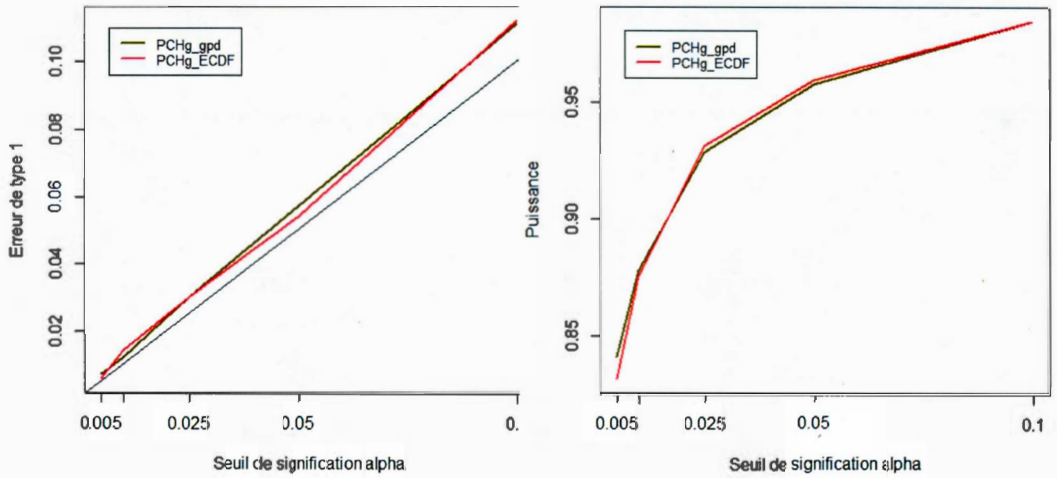


Figure 4.4: Estimation des p-valeurs de l'approche PCH_g avec la fonction de distribution empirique cumulative (ECDF) d'une part et d'autre part par l'algorithme de Knijnenburg *et al.* (2009) sous l'hypothèse nulle (figure à gauche) et sous l'hypothèse alternative (figure à droite).

500 observations a été considéré et 1000 réplications ont été générées pour évaluer l'erreur du type 1 ainsi que la puissance de notre approche PCH_g et celle de Klei (PCH de Klei). Les corrélations considérées pour les deux modèles (M1 et M2) est 0 et 0.2 respectivement. La matrice des SNPs et la matrice de caractères sont générées selon le modèle décrit en 4.1.1 en tenant compte de niveau de corrélation fixé pour chaque modèle (M1 et M2). La figure 4.5 et la figure 4.6 montrent les résultats de l'analyse de simulation à l'aide de notre approche PCH_g comparée avec l'approche PCH de Klei en termes d'erreur de type 1 et en termes de puissance. L'erreur de type 1 pour les deux modèles (M1 et M2) est bien contrôlée pour les deux scénarios simu-

lés, puis, à partir du graphique à droite de la figure 4.6, on remarque que la puissance de l'approche de Klei est affectée par la non-prise en considération des effets joints des SNPs. On peut donc conclure que les puissances obtenues avec l'approche de Klei sont faibles comparativement à la puissance de l'analyse simultanée des SNPs de notre approche. Ceci pourrait être expliqué par le fait que l'effet principal de chaque SNP ainsi que l'effet joint des SNPs est capturé par notre approche.

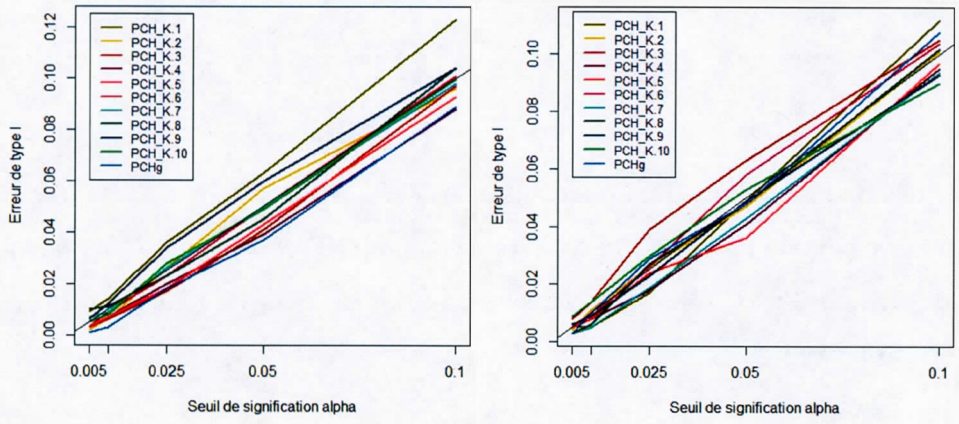


Figure 4.5: Résultat de PCH de Klei et PCH_g pour des SNPs non corrélés (figure à gauche) et des SNPs corrélés (figure à droite) sous H_0. Abréviation : PCH_K pour composante principale de l'héritabilité relative à chaque SNP ($l = \{1, \ldots, 10\}$) selon l'approche de Klei, et PCH_g représente la composante principale de l'héritabilité de notre nouvelle approche. La droite de référence avec l'ordonnée à l'origine égale à 0 et la pente égale à 1 est également représentée.

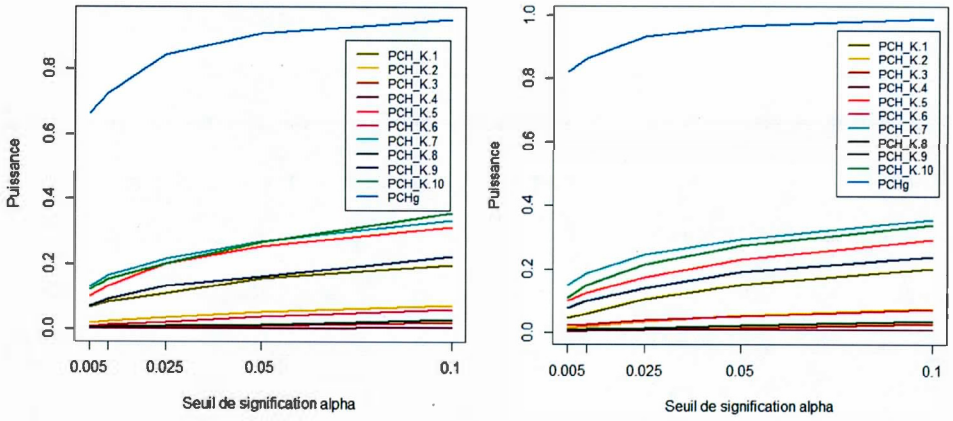


Figure 4.6: Résultat de PCH de Klei et PCH_g pour des SNPs non corrélés (figure à gauche) et des SNPs corrélés (figure à droite) sous H_1. Abréviation : PCH_K pour composante principale de l'héritabilité relative à chaque SNP ($l = \{1, \ldots, 10\}$) selon l'approche de Klei, et PCH_g représente la composante principale de l'héritabilité de notre nouvelle approche.

4.2.4 Les performances de PCH_g en considérant différents scénarios de corrélation entre les marqueurs

L'objectif de cette analyse de simulation est d'étudier l'impact de la variation de déséquilibre de liaison sur la puissance de l'approche PCH_g. Les corrélations testées entre les SNPs sont : 0, 0.2, 0.4, 0.6 et 0.8. Ces niveaux de corrélations vont être utilisés pour modéliser 5 modèles générés chacun selon le modèle décrit en 4.1.1 pour un seuil de signification fixé à 5 %.

Le résultat de comparaison de puissance pour différentes valeurs de corrélation entre les SNPs selon l'algorithme de PCH_g est donné dans la figure 4.7. On remarque que

plus la corrélation entre les 10 marqueurs augmente, plus la puissance de détection augmente.

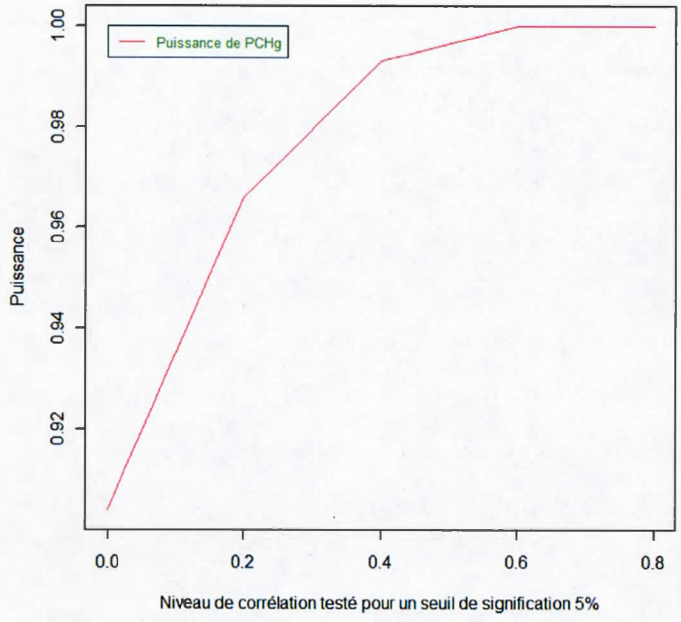


Figure 4.7: Comparaison de puissance de l'approche PCH_g pour différentes valeurs de corrélation entre les SNPs, $r^2 \in \{0, 0.2, 0.4, 0.6, 0.8\}$, pour un seuil de signification fixé à 5%. Abréviation : PCH_g pour composante principale de l'héritabilité de notre nouvelle approche relative à chaque valeur de corrélation.

4.3 Discussion

Les résultats des études de simulations montrent que :

- en présence d'une corrélation entre les caractères, l'analyse d'un SNP à la fois par l'approche de Klei engendre toujours une perte de puissance pour détecter les SNPs associés aux caractères, comparativement à notre méthode qui permet d'étudier simultanément les SNPs. Ainsi, l'approche de Klei ne capture donc pas l'effet joint des SNPs, d'où l'avantage d'utiliser notre méthode pour détecter d'autres associations non découvertes par l'étude d'un SNP à la fois ;

- la variabilité génétique expliquée par notre approche est supérieure à celle de Klei puisque PCH_g est toujours plus puissante que PCH de Klei dans tous les scénarios réalisés. Ainsi, on peut avancer que PCH_g explique une partie de l'héritabilité manquante dans l'approche de Klei ;

- en se basant seulement sur les puissances estimées, nous pouvons déduire qu'en présence de corrélation entre les caractères et des effets joints des SNPs, l'analyse univariée est la moins efficace car elle repose sur une sélection des SNPs selon leurs effets individuels. Par contre, PCH de Klei performe mieux que l'analyse univariée, mais le fait qu'elle analyse séparément les SNPs pourra mener à un résultat non concluant sur leurs effets joints.

Sur les données simulées, nous avons démontré l'avantage de l'analyse simultanée de plusieurs marqueurs (PCH_g) par rapport à une analyse individuelle des SNPs

(PCH de Klei). Ces résultats sont encourageants et montrent que l'analyse jointe de plusieurs marqueurs apporte une meilleure information sur la variabilité génétique. Un tel gain d'information peut élargir ainsi la compréhension du fonctionnement des gènes en étudiant plusieurs caractères et plusieurs marqueurs simultanément. Ces derniers résultats rejoignent la conclusion de Yang *et al.* (2010) sur l'utilité des analyses jointes de marqueurs.

Une illustration de la méthodologie proposée dans ce mémoire, à l'aide des données réelles, sera présentée dans le prochain chapitre.

CHAPITRE V

APPLICATION

Afin de tester et valider l'approche que nous avons développée, nous utiliserons un jeu de données réelles portant sur la maladie d'Alzheimer.

5.1 L'Alzheimer

L'Alzheimer est une maladie neurologique qui entraîne une altération progressive des fonctions mentales. La forme familiale de la maladie d'Alzheimer est attribuable à des modifications ou des altérations dans des gènes spécifiques qui sont transmis des parents aux enfants, tandis que la forme non familiale ou sporadique est due à un ensemble complexe d'éléments d'ordre génétique et environnemental.
La maladie d'Alzheimer est caractérisée par l'accumulation dans le cortex cérébral de deux protéines : les plaques amyloïdes et les protéines Tau (de l'anglais Tubulin associated unit) hyperphosphorylés. Une grande concentration des plaques amyloïdes entre les neurones et des protéines Tau hyperphosphorylés à l'intérieur des neurones endommage ces dernières et provoque éventuellement leur mort. Cela entraînera des déséquilibres au niveau de la communication neuronale, ce qui mène aux premiers déficits cognitifs des années plus tard.

5.1.1 L'aspect génétique

La forme précoce de la maladie d'Alzheimer (qui est essentiellement familiale) se caractérise par des mutations génétiques identifiées dans trois chromosomes : le chromosome 1, chromosome 14 et chromosome 21. Des études récentes à grande échelle (Karch et Goate, 2015 ; Hollingworth *et al.*, 2011) ont identifié plusieurs gènes de susceptibilité plus modestes, y compris CR1, CLU et PICALM, associé principalement à la forme sporadique (non familiale) de l'Alzheimer (c'est celle que nous allons étudier ici).

5.1.2 ApoE : gène candidat pour l'Alzheimer

Le gène candidat [1], "ApoE" [2], situé sur le chromosome 19, a été identifié aussi comme significativement lié à une augmentation de risque de la forme sporadique de la maladie d'Alzheimer, ainsi qu'à une forme familiale tardive (Bertram *et al.*, 2010 ; Hollingworth *et al.*, 2011).

Le gène ApoE présente un polymorphisme représenté principalement par les allèles $\varepsilon 2$ (=ApoE2), $\varepsilon 3$ (=ApoE3) et $\varepsilon 4$ (=ApoE4). Le premier se rencontre dans $\sim 7-8\%$ de la population, le deuxième se rencontre dans $\sim 75-80\%$ de la population et le troisième se rencontre dans $\sim 14-15\%$ de la population (Cedazo-Mínguez et Cowburn (2001) et Zannis *et al.* (1993)).

---

1. Un gène « candidat» est un gène qui a été identifié dans d'autres études comme étant à l'origine du caractère ou parce que sa fonction présumée pourrait l'impliquer dans le phénotype étudié.

2. l'apolipoprotéine E (ApoE) est une protéine ubiquitaire, jouant un rôle dans le monde transport du cholestérol et des phospholipides. L'ApoE est un des composants des lipoprotéines.

De nombreuses équipes ont confirmé que l'allèle $\varepsilon 4$ de l'ApoE est d'environ 2 à 4 fois plus fréquent chez les patients d'Alzheimer que dans la population générale, tandis que l'allèle $\varepsilon 2$ semble avoir un effet protecteur (order,EH.CORDER,EH. *et al.*,1994). D'autres études ont montré que la présence de l'ApoE $\varepsilon 4$ sous la forme hétérozygote augmente par deux le risque de la maladie d'Alzheimer ; la présence de l'ApoE $\varepsilon 4$ sous la forme homozygote augmente par 11 le risque de la maladie.

L'allèle $\varepsilon 4$ de l'ApoE est donc un facteur de risque important dans la maladie d'Alzheimer. Cependant, certains sujets porteurs de cet allèle ne sont pas atteints. Chee Seng ku et Chia (2010) suggèrent que des variantes génétiques (SNPs) à faibles effets, à l'intérieur du locus de l'ApoE, peuvent également être impliqués dans le risque associé à ce gène. Les approches qui augmentent la puissance de ces analyses sont potentiellement d'une grande valeur. L'approche de Klei a des points forts, mais elle limite la puissance avec les analyses individuelles des variantes ce qui peut pénaliser les SNPs à faibles effets individuels. Ainsi, l'analyse en composante principale de l'héritabilité généralisée peut être utilisée dans ce contexte, afin d'identifier les variantes génétiques à faibles effets.

Dans la section 5.2, nous allons analyser un jeu de données portant sur la maladie d'Alzheimer. Notre analyse consiste à identifier les associations potentielles entre l'accumulation de la protéine "amyloïde" et les génotypes de milliers de SNPs sur le chromosome 19.

5.2 Données utilisées dans la recherche

Les données utilisées dans l'évaluation de notre approche proviennent de la base de données "Alzheimer's Disease Neuroimaging Initiative (ADNI)".[3] Elles sont composées de 27 sujets malades (AD pour Alzheimer's disease), de 264 sujets présentant une déficience cognitive légère (notée MCI) et de 124 sujets de contrôle normaux (notés CN). Les phénotypes sont les mesures de la protéine amyloïdes β prises sur 96 régions du cerveau de chaque sujet à l'aide de techniques d'imagerie cérébrale. Au total, nous avons donc 96 phénotypes observés sur 416 sujets ($Y_{416\times 96}$ matrice de 416×96). Notez que les plaques amyloïdes (variable réponse / phénotypes) ont tendance à s'étendre/concentrer aux différentes régions du cerveau qui sont impliquées dans le fonctionnement normal de ce dernier. Les régions impliquées dans le langage ainsi que dans la perception du corps parmi les objets qui l'entourent ont été entre autres identifiées comme étant des zones cibles d'accumulation de la protéine amyloïde.

5.2.1 Caractéristiques démographiques des sujets

L'évaluation de la fonction cognitive globale des sujets impliqués dans l'étude est faite sur la base du test "Mini Mental State Examination", noté MMSE, de Folstein *et al.* (1975). Le MMSE consiste en une série de 30 questions réparties en six catégories :

---

3. ADNI est un effort global de recherche qui soutient activement la recherche et le développement de traitements qui ralentissent ou arrêtent la progression de la maladie de l'Alzheimer. Il contient des bases de données et des ressources relatives à la maladie. Il permet aux scientifiques de mener des recherches de cohésion et de partager des données compatibles avec d'autres chercheurs dans le monde entier. L'étude ADNI comporte trois phases : ADNI1, ADNI GO et ADNI2. `Source:http://adni.loni.usc.edu/study-design/`

l'orientation dans le temps et dans l'espace, l'apprentissage, l'attention et le calcul, le rappel libre, le langage et les praxies constructives.

L'évaluation de la charge amyloïde est mise en évidence sur les images du cerveau par un biomarqueur [4] av45-surv utilisé dans les recherches cliniques pour différencier les sujets atteints d'Alzheimer et les sujets ayant une déficience cognitive légère (MCI) des sujets de contrôle (Camus *et al.*, 2012). Le biomarqueur se fixe sur les plaques amyloïdes et permet ainsi de quantifier la charge amyloïde cérébrale. Un descriptif détaillé des sujets participant à l'étude (416 sujets) selon leur statut d'inclusion (AD, MCI ou bien CN) est donné dans les tableaux 5.1, 5.2, 5.3 et 5.4.

L'âge des sujets diffère en moyenne entre le groupe de contrôle normal et les groupes AD et MCI. Cependant, il n'y a pas de différence en moyenne entre les trois groupes par rapport au niveau d'éducation. Le tableau 5.3 montre que la moyenne de av45-surv-global (valeur globale sur les 96 régions) est plus élevée chez les sujets atteints de l'Alzheimer que chez les sujets de contrôle dans toutes les régions de cerveau. Les sujets MCI ont également montré une moyenne plus élevée par rapport aux sujets de contrôle. Le tableau 5.4 montre que la moyenne de MMSE est plus faible chez les sujets atteints de l'Alzheimer que chez les sujets de contrôle dans toutes les régions du cerveau. Les sujets MCI ont également montré une moyenne faible par rapport aux sujets de contrôle mais plus élevée que celle des sujets atteints.

Répartition des sujets selon la pathologie et le sexe

Cette étude comporte un échantillon qui regroupe 225 femmes et 191 hommes. La

---

4. Un biomarqueur se définit comme une entité mesurable et quantifiable dont la présence ou l'activité est associée à la maladie. Le biomarqueur idéal doit être très spécifique, lié au processus pathologique, doit permettre de prédire l'évolution de la maladie, et doit refléter le niveau de réponse à un éventuel traitement étiologique.

Tableau 5.1: Évaluation de l'âge des sujets de l'étude.

Sujets	Obs	Moyenne ± Écat-type	Min	Max
CN	124	74.26 ± 6.077	56	90
MCI	264	71.46 ± 7.674	55	91
AD	27	76.44 ± 8.697	56	91

Tableau 5.2: Évaluation de l'éducation des sujets.

Moyenne ± Écat-type	Min	Max
16.49 ± 2.555	12	20
16.05 ± 2.587	9	20
15.96 ± 2.457	12	20

Tableau 5.3: Évaluation de av45-surv-global des sujets de l'étude.

Sujets	Obs	Moyenne ± Écat-type	Min	Max
CN	124	1.252 ± 0.217	0.97	1.91
MCI	264	1.313 ± 0.241	0.86	2.03
AD	27	1.475 ± 0.239	0.98	1.87

Tableau 5.4: Évaluation du test MMSE des sujets de l'étude.

Moyenne ± Écat-type	Min	Max
28.85 ± 1.53	21	30
28.24 ± 1.69	20	30
23.19 ± 2.42	19	29

Note : **L'education est défini par le nombre d'années d'études terminées de la personne. L'âge au diagnostic de la personne est exprimé en années.**

répartition des sujets selon la pathologie est donnée dans la figure 5.1. Dans notre étude, on trouve une prédominance de développer la maladie (MCI et AD) chez le sexe féminin par rapport au sexe masculin et cette prédominance est plus marquée dans le groupe MCI que le groupe AD (voir figure 5.1).

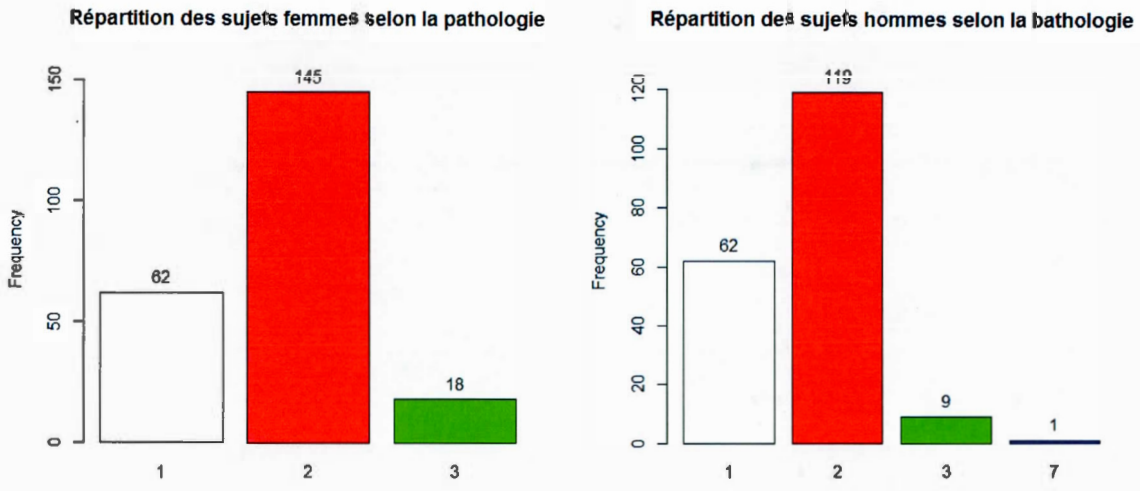


Figure 5.1: Répartition des sujets (femmes ou hommes) selon la pathologie où 1 indique le groupe de contrôle (CN), 2 le groupe ayant une déficience légère (MCI) et 3 pour le groupe identifié comme ayant l'Alzheimer (AD).

5.2.2 Évaluation de l'effet des covariables sur l'état d'inclusion des sujets (le diagnostic)

Pour évaluer la significativité des covariables (l'âge, le sexe et l'éducation) par rapport au diagnostic des sujets (CN, MCI, AD), nous allons utiliser la régression logistique multinomiale ou polytomique. Ce choix de régression se justifie par le fait que la variable dépendante est une variable catégorique à plus que 2 catégories à l'inverse de la régression logistique standard (Variable dépendante binaire). On sélectionne l'état d'inclusion "contrôle" comme catégorie de référence dans le modèle. La variable Y dans le modèle désigne l'état d'inclusion où le diagnostic codé 1, 2, 3 pour CN,

MCI, AD respectivement.

Nous souhaitons savoir si le coefficient de chaque covariable est significatif et nous fixons le risque de première espèce à 5%. L'hypothèse nulle du test s'écrit

$$H_0 : \beta_{k,j} = 0, \qquad \forall k$$

où j est la covariable, et k l'indice de catégorie (AD, MCI, CN).

Nous utilisons le test de rapport de vraisemblance. Nous calculons la déviance du modèle réduit et nous la comparons à celle du modèle complet. L'ensemble des variables explicatives spécifiées dans le modèle complet sont l'âge, l'éducation et le sexe. La statistique du test suit une loi de χ^2 à *ddl* degrés de liberté (c'est la différence entre le nombre de paramètres dans le modèle complet et le nombre de paramètres dans le modèle réduit). Le calcul est fait avec le logiciel R avec **package "nnet"**. Avec un

Analyse du modèle	Déviance résiduel	Écart de déviance	p-valeurs
Modèle complet	647.8685		
Modèle Réduit (Sexe)	651.7697	3.9012	0.1421887
Modèle Réduit (Education)	652.9918	5.1233	0.07717
Modèle Réduit (Age)	666.3147	18.4462	$9.873214e^{-5}$

$\chi^2(ddl = 1)$, nous avons une p-valeur=1-*pchisq*(Écart de déviance, *ddl*). Nous rejetons l'hypothèse nulle pour la covariable Age, le coefficient n'est pas simultanément nul dans l'ensemble des logit. Par contre, les covariables *Sexe* et *Education* ne sont pas significatives dans l'ensemble des logit.

5.3 Contrôle qualité de la base de données (matrice du génotype)

Les données génotypiques dont nous disposons se présentent sous la forme d'une matrice 'sujets × SNPs'. L'échantillon est composé de 416 sujets génotypés sur un

ensemble de 12361 marqueurs. Sur le chromosome 19, les données des génotypes ne sont pas utilisables telles quelles : un contrôle qualité de ces données est nécessaire avant de les analyser pour minimiser l'impact des erreurs de génotypages[5] sur les analyses d'association.

Pour le contrôle de qualité des SNPs, nous avons appliqué les seuils suivants :

- les marqueurs ayant des données manquantes doivent être exclus de l'étude ;
- les marqueurs ayant une fréquence allélique inférieure à 1% doivent être supprimés, parce que la puissance de détecter cette association est faible ;
- l'équilibre d'Hardy-Weinberg par marqueur doit être respecté. Lors du test d'adéquation à la loi d'Hardy-Weinberg, tous les marqueurs qui ont une p-valeur inférieure à $1 \times E^{-}6$ (seuil de correction de Bonferroni) doivent être écartés ;
- les marqueurs en déséquilibre de liaison ($r^2 > 0.8$) sont exclus, où r^2 désigne la corrélation entre une paire de SNPs.

Au total, nous avons supprimé 2371 SNPs à cause des valeurs manquantes et 2672 à cause du déséquilibre de liaison. Tous les génotypes ont été trouvés en équilibre de Hardy-Weinberg et évalués en utilisant un test Khi-deux standard. Ce contrôle de qualité a été réalisé à l'aide du logiciel R. Seuls les sujets qui ont des informations complètes sur les marqueurs et les phénotypes sont inclus dans notre analyse. Ainsi, toutes les analyses des sections suivantes ont été basées sur 7318 SNPs et 416 sujets qui ont des données complètes pour les variantes et les 96 phénotypes.

---

5. Erreurs de génotypages sont les erreurs de classification qui affectent les génotypes (Guedj, 2007 page-32).

5.4 Résultats et discussions

5.4.1 Analyse de corrélation entre les phénotypes (charges amyloïdes)

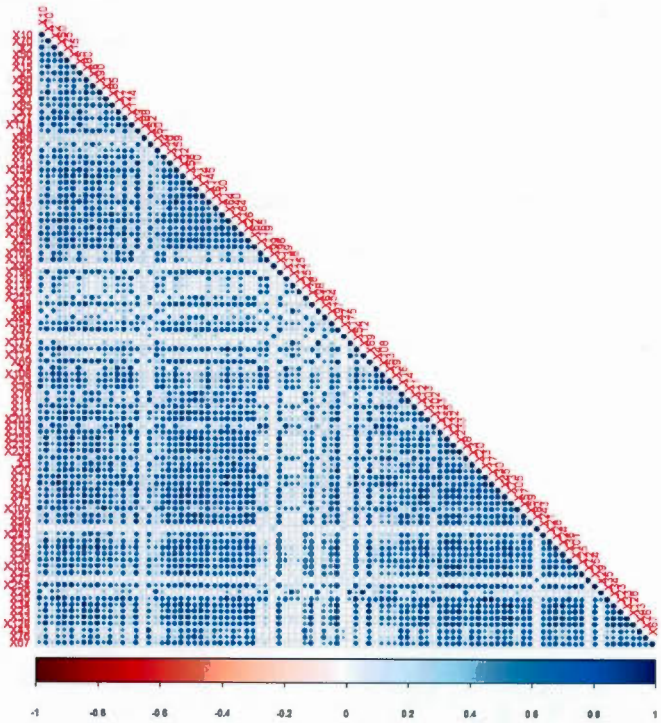
Pour établir l'utilité de PCH de Klei et PCH_g dans l'étude des données ADNI, une analyse de corrélation entre les phénotypes est réalisée. Selon le test de corrélation de Pearson, nous avons trouvé que 92.23% des caractères sont corrélés à un seuil de 5 %. Les résultats sont donnés dans les figures 5.2 et 5.3 en terme de corrélation et en terme de p-valeur. Dans la section suivante, une vérification de l'existence de données aberrantes au niveau des 96 phénotypes est effectuée.

5.4.2 Analyse des valeurs aberrantes

Grubbs (1950) définit une valeur aberrante comme étant une observation qui semble dévier de façon marquée par rapport à l'ensemble des autres membres de l'échantillon dans lequel elle apparaît. Carletti G (1988) définit une valeur aberrante comme une valeur qui paraît suspecte parce qu'elle s'écarte d'une façon importante des autres valeurs de la variable étudiée ou ne semble pas respecter une norme ou une relation bien définie.

La détection des observations aberrantes (ou bien atypiques) est d'une importance capitale dans l'analyse statistique des données multivariées parce que les moments (la moyenne et la variance) de la base de données peuvent être contaminés par des observations aberrantes, ce qui entraîne par conséquent des estimateurs biaisés et des fausses interprétations. Si la détection des observations atypiques est possible dans une base de données de dimension deux, elle est très compliquée pour les données

Figure 5.2: Carte thermique du r^2 pour les paires de caractères. Pour une paire de caractères donnés, plus la couleur est sombre, plus les caractères sont corrélés.



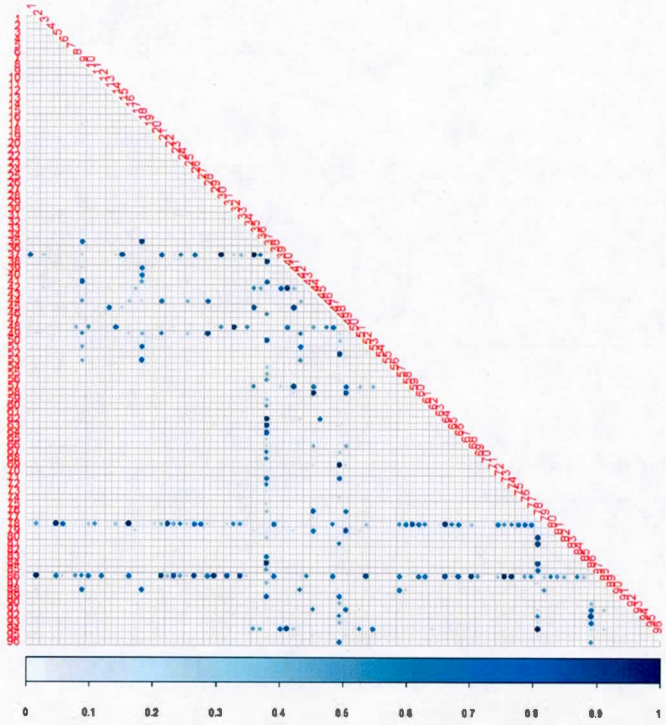
multivariées à cause de la dimension et de la redondance de l'information. Une analyse préliminaire des boîtes à moustaches [6] relatives aux phénotypes de l'étude (voir figures 5.4) montre l'existence des observations aberrantes univariées.

Divers tests sont usuellement utilisés aussi pour détecter les valeurs aberrantes, on cite entre autres le test de Grubbs dont :

---

6. Une boîte à moustaches est un moyen pratique et compact pour visualiser la distribution d'une variable.

Figure 5.3: Carte thermique des p-valeurs pour les paires de caractères. Pour une paire de caractères donnés, plus la couleur est claire, plus les p-valeurs sont petites.



H_0 : "L'observation n'est pas aberrante" ;

H_1 : "L'observation est aberrante".

L'application de ce test sur notre base de données réelle détecte 72 valeurs aberrantes pour l'ensemble des phénotypes. Cette approche fonctionne bien si une seule valeur aberrante est présente, mais elle peut être affectée par l'effet de masquage s'il y a plus d'un point aberrant parce qu'une valeur aberrante lointaine peut faire de toutes autres valeurs aberrantes des valeurs normales. Certains raffinements tels que la

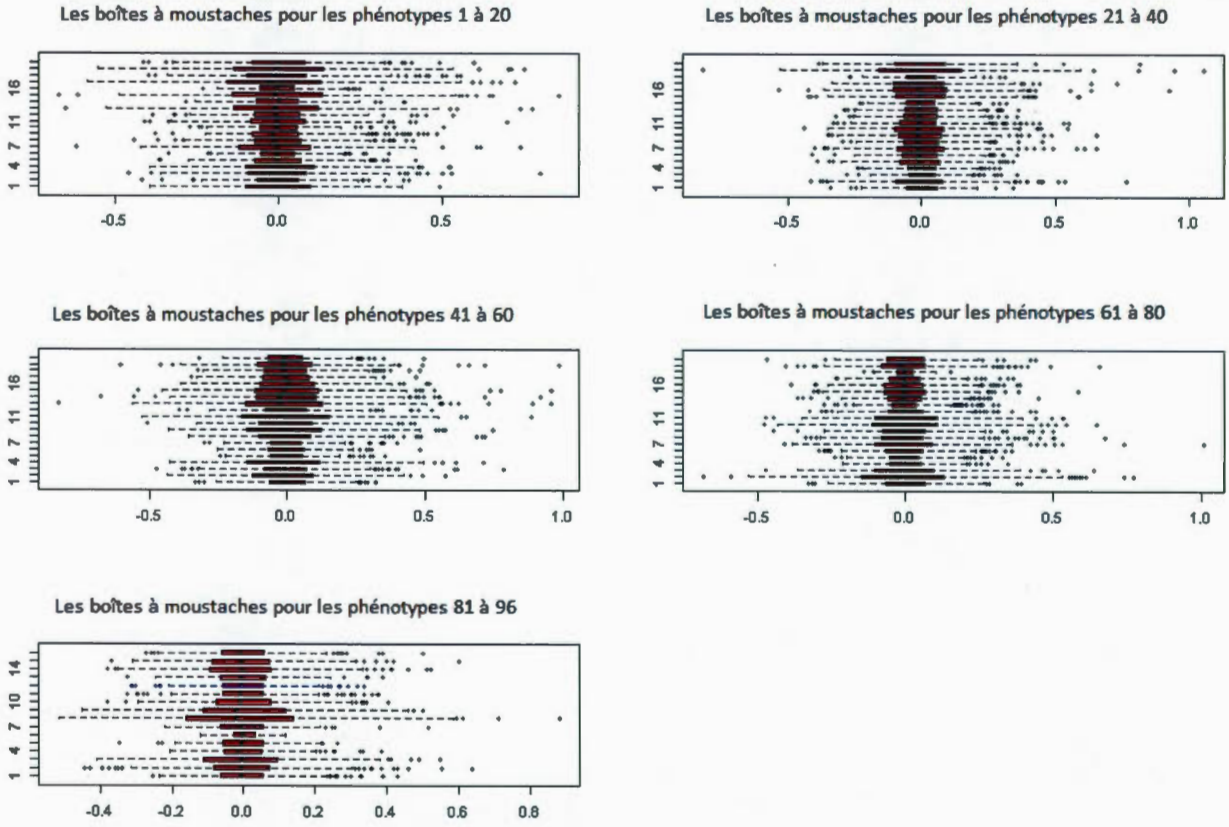


Figure 5.4: Boîtes à moustaches des 96 phénotypes.

suppression itérative des valeurs aberrantes ont été proposés, cependant les problèmes liés à l'approche existent encore dans ces méthodes.

Détection des observations aberrantes par mesure de distance

Les mesures de distance sont souvent utilisées dans la détection des valeurs aberrantes. Ainsi, pour obtenir une mesure de distance fiable pour la détection des valeurs aberrantes dans les données multivariées, il est préférable de savoir comment

la moyenne et la matrice de variances-covariances sont estimées.

Trois méthodes sont souvent utilisées pour la détection des valeurs aberrantes, premièrement la distance de Mahalanobis et deux méthodes robustes, l'une basée sur le déterminant minimale de la covariance, l'autre technique a pour objectif de minimiser le volume de l'ellipse qui engendre le nuage de points.

Distance de Mahalanobis

La distance de Mahalanobis Crettaz de Roten et Helbling (1996) est définie comme suit :

$$D_i = \sqrt{(x_i - T(X))' \widehat{\Sigma}^{-1} (x_i - T(X))}, \tag{5.1}$$

où $T(X)$ représente la moyenne et $\widehat{\Sigma}$ la matrice de covariance empirique. La distance de Mahalanobis (noté MD) prend en considération la dépendance entre les variables par la matrice de covariance empirique $\widehat{\Sigma}$.

La distance de Mahalanobis utilise l'espace des composantes principales pour calculer la distance de chaque sujet par rapport au centre du nuage de points dans cet espace. La projection du nuage de points des individus sur les composantes principales, considère les sujets avec une grande distance de Mahalanobis, par rapport à l'origine comme observations aberrantes. Autrement dit, la méthode suggère des observations aberrantes qui ont une distance considérable par rapport au centre du nuage de points (voir figure 5.5). Mais une difficulté majeure de l'ACP provient de sa sensibilité aux valeurs aberrantes. En effet l'estimateur $\widehat{\Sigma}$ est efficace lorsque les données sont distribuées suivant une loi normale multivariée. Cependant, il est sensible face à des points aberrants, et dans ce cas, les conclusions faites sur la base des composantes principales peuvent être totalement erronées. Afin de réduire cette sensibilité, d'autres techniques d'ACP insensibles à ces valeurs aberrantes sont pro-

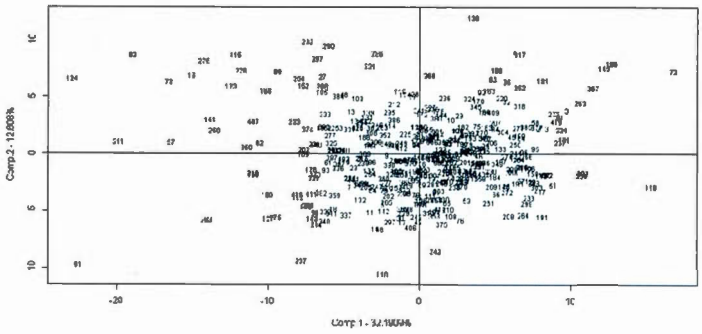


Figure 5.5: Projection des individus sur les deux axes qui représentent les deux composantes principales basées sur l'analyse en composantes principales. La distance entre chaque observation et le centre des nuages de points définit la distance de Mahalanobis.

posées dans la littérature.

En effet, les méthodes dites d'ACP robustes aux valeurs aberrantes tentent de remplacer les estimateurs classiques de la moyenne et de la matrice de covariance par des estimateurs robustes. Rousseeuw et *al.* définissent deux estimateurs robustes pour remplacer les estimateurs classiques. Il s'agit de l'estimateur "Minimum covariance déterminant MCD" et "Ellipsoïde de volume minimal, noté MVE" que nous allons voir.

Minimum covariance déterminant

Rousseeuw (1984) propose l'estimateur robuste, MCD, pour estimer la moyenne et la covariance. Cette méthode consiste à trouver un sous-ensemble de h observations ($h \leq n$, n le nombre d'observations dans le jeu de données) dont la matrice de covariance a le plus petit déterminant.

L'estimateur MCD est déterminé pour l'échantillon $\{x_1, \ldots, x_n\}$ en sélectionnant le

sous-échantillon $\{x_{i_1}, \ldots, x_{i_h}\}$ de taille h $(1 \leq h \leq n)$, qui minimise le déterminant de la matrice de covariance empirique parmi tous les sous-échantillons possibles de taille h. L'estimateur MCD de position est alors défini comme

$$T_n = \frac{1}{h} \sum_{k=1}^{h} x_{i_k},$$

et l'estimateur MCD de la matrice de covariance empirique est défini par

$$\widehat{\Sigma} = c_1 \frac{1}{h} \sum_{k=1}^{h} (x_{i_k} - T_n)(x_{i_k} - T_n)^\top,$$

où $c_1 = \dfrac{1-\alpha}{F_{\chi^2_{p+2}}(q_\alpha)}$ ($F_{\chi^2_{p+2}}$ étant la fonction de répartition de la loi de χ^2 à $p+2$ degrés de liberté), et $q_\alpha = \chi^2_{p,1-\alpha}$ est défini comme le quantile $(1-\alpha)$ de la distribution χ^2. Généralement, h prend la valeur $h = n(1-\alpha)$, avec $\alpha = 0.50$.

La détection des données aberrantes basée sur MCD se résume dans l'algorithme 5.4.2. La plus grande valeur possible pour h est donnée par :

$$h = \frac{n+p+1}{2}.$$

Algorithme de MCD 5.4.2

1 : Tous les sous-ensembles de cardinal h sont formés ;

2 : La matrice de covariance pour chacun de ces sous-ensembles est calculée ;

3 : Le sous-ensemble dont le déterminant de la matrice de covariance est le plus petit est sélectionné ;

4 : La moyenne et la covariance de ce sous-ensemble sont utilisées pour calculer la distance de chaque observation par rapport à cette moyenne ;

5 : Pour une observation i, si D_i est supérieure à un certain seuil($\chi^2_{p,0.975}$), elle est définie comme aberrante, sinon elle est considérée comme observation normale.

Ellipsoïde de volume minimal (MVE)

Rousseeuw (1985) a introduit la méthode robuste MVE pour la détection des valeurs aberrantes dans les données multidimensionnelles. C'est une extension multivariée de la méthode de la médiane (MCD) qui prend en considération de multiples observations aberrantes. L'idée de base de cette méthode (Abou-Moustafa et Ferrie, 2007 ; Chen *et al.*, 2008) est que la distance de Mahalanobis entre deux points x et y s'écrit $D^2(x,y) = (x-y)^\top \Sigma^{-1}(x-y)$. Or la distance de x par rapport à l'origine est égale à $D^2(x,0) = x^\top \Sigma^{-1} x = c^2$, qui est l'équation d'un ellipsoïde centré à l'origine, avec des axes principaux alignés sur les axes des cordonnées. Ainsi, les observations aberrantes sont des points essentiellement situés sur la limite du volume minimal couvrant l'ellipsoïde. La procédure de détection des observations aberrantes par l'approche MVE est donnée dans l'algorithme 5.4.2.

Algorithme de MVE 5.4.2

1 : Pour une matrice de données X avec p variables et n observations, tirer un sous-échantillon de $p+1$ observations différentes, indexé par $J = (j_1, ..., j_{p+1})$, et calculer la moyenne arithmétique $T_J = \dfrac{1}{p+1} \sum_{j \in J} x_j$ et la matrice covariance correspondante $C_J = \dfrac{1}{p} \sum_{j \in J} (x_j - T_J)^\top (x_j - T_J)$ où C_J est non singulière ;

2 : Calculer $m_J^2 = [(x_j - T_J) C_J^{-1} (x_j - T_J)^\top]_{h:n}$ où $h = \dfrac{n+p+1}{2}$.

3 : Calculer $P_J = (det(m_J^2 C_J))^{\frac{1}{2}}$;

4 : Répétez la procédure ci-dessus pour un grand nombre de sous-échantillons J, et conserver celui avec la plus faible P_J ;

5 : Pour ce sous-échantillon retenu J, calculer $T(X) = T_J$ et $C(X) = c^2(n,p)(\chi^2_{p,0.50})^{-1} m_J C_J$, où c^2 est un terme de correction calculé comme $[1 + \frac{15}{(n-p)}]^2$ et $\chi^2_{p,0.50}$ est la médiane de la distribution de χ^2 avec p degrés de liberté.

La $T(X)$ et $C(X)$ calculée à l'étape (5) de l'algorithme de MVE 5.4.2 sont les estimateurs de la moyenne et de la matrice de variances-covariances déduites de l'approche de MVE. Ainsi, la statistique suivante, peut être calculée à partir de ces estimateurs :

$$D^2 = (x_j - T(X))C^{-1}(X)(x_j - T(X))^\top.$$

Pour une observation x_j, si $D_j^2 > \chi^2_{p,0.975}$ cette observation est considérée comme aberrante, sinon elle est considérée normale.

Le calcul de cette distance est fait avec le logiciel R (Package MASS) et le résultat est donné dans la figure 5.7.

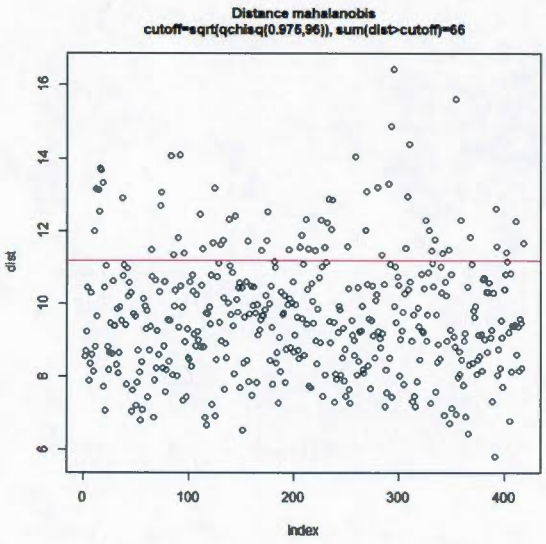


Figure 5.6: Détection des valeurs aberrantes en utilisant la mesure MCD.

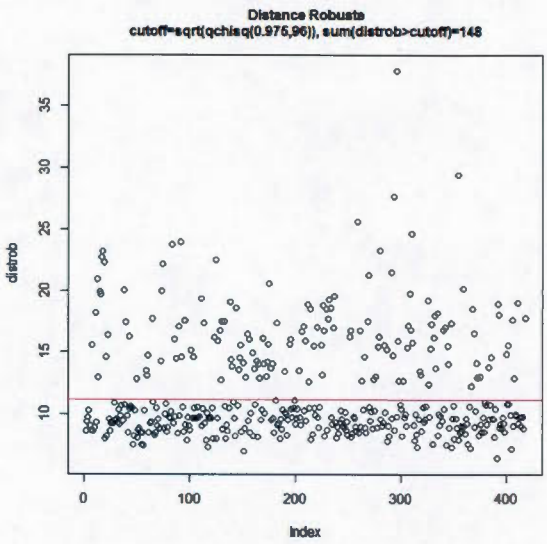


Figure 5.7: Détection des valeurs aberrantes en utilisant la mesure MVE.

Les tests d'hypothèses utilisés dans les deux approches MCD et MVE sont fondés sur un test de khi-deux qui exige la normalité des phénotypes, or des tests de normalité portés sur notre base de données ont montré que nos 96 phénotypes n'ont pas une distribution normale. Ceci peut expliquer le grand nombre de valeurs aberrantes détectées par les deux approches dans notre jeu de données. Ainsi, on s'est limité à supprimer seulement trois observations qui ont des grandes distances par rapport aux autres (voir figure 5.8). Dans la section prochaine, nous allons évaluer l'effet des

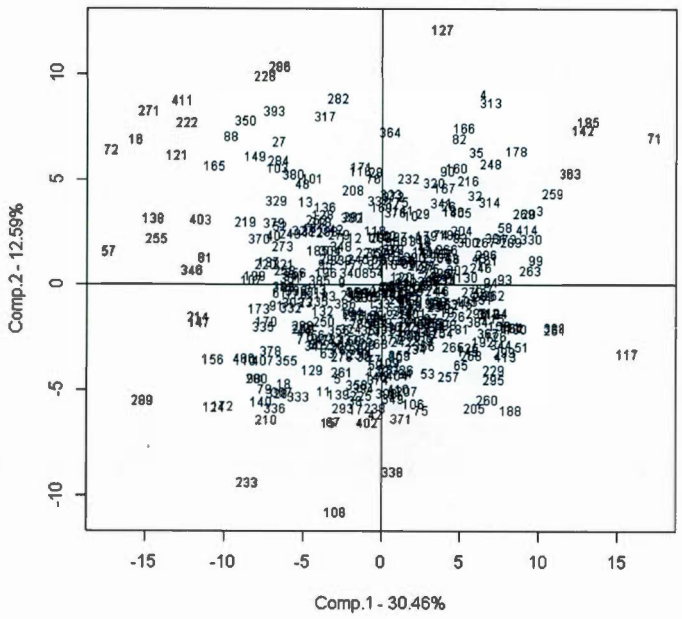


Figure 5.8: Projection des individus sur les deux axes qui représentent les deux composantes principales basées sur l'analyse en composantes principales après suppression de 3 valeurs aberrantes.

covariables sur les phénotypes considérés.

5.4.3 Effet des caractéristiques des sujets sur les phénotypes

Les caractéristiques démographiques des sujets de l'étude peuvent avoir des effets sur les phénotypes. Généralement, ces caractéristiques sont considérées comme des covariables et elles doivent alors être prises en compte lors des analyses, car si une covariable a un grand effet sur le phénotype, son inclusion dans les analyses réduira la variation résiduelle et augmentera notre capacité de détecter l'association phénotypes-génotype. Dans ce cadre, nous allons évaluer l'effet des covariables (l'âge au diagnostic, le sexe et l'éducation) sur les phénotypes. Nous allons analyser leurs significations sur les phénotypes bruts dans une étape, et ensuite extraire l'effet des covariables significatives des phénotypes, que nous appellerons "phénotypes ajustés", et enfin analyser l'association des génotypes par rapport aux phénotypes ajustés.
Pour tenir compte des effets de ces covariables, chaque phénotype a été ajusté au moyen d'une régression linéaire multiple :

$$y_i = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Education}_i + \beta_3 \text{Sexe}_i + \varepsilon_i. \tag{5.2}$$

Ensuite, nous avons calculé par rapport à toutes les 96 variables, le nombre de fois où la covariable a été significative pour un seuil de 5%. Cette procédure est répétée pour chaque covariable avec correction de Bonferroni pour le seuil global de signification. Les covariables sélectionnées par le modèle de régression dans la première étape sont : le sexe et l'âge où la première covariable est significative dans 15.26% des cas, l'âge est significatif dans 11.45% des cas. Le sexe et l'âge dans notre analyse semblent avoir un effet sur la concentration des plaques amyloïdes. Ce résultat concorde avec la plupart des données de la littérature. Selon plusieurs études, les femmes ont un risque plus élevé de développer la maladie d'Alzheimer que les hommes (Fratiglioni *et al.*, 1999). Dans une étude publiée en 1999, l'incidence de la maladie d'Alzheimer était, avant 80 ans, plus élevée chez les hommes que chez les femmes, alors que c'est

l'inverse après 80 ans (Letenneur *et al.*, 1999). Cette différence pourrait s'expliquer par l'espérance de vie plus élevée chez les femmes que chez les hommes.
Le niveau d'éducation n'est pas significatif dans tous les cas testés pour évaluer son effet par rapport a chaque phénotype. L'éducation ne semble pas être liée à la charge amyloïde ce qui concorde avec certaines études portantes sur l'implication de l'éducation dans l'Alzheimer (Cobb *et al.*, 1995 ; Letenneur *et al.*, 1999). Étant donné que les deux covariables sélectionnées sont reliées à la charge amyloide, les analyses vont être conduites par conséquent selon la procédure déclarée précédemment, les algorithmes des deux approches (PCH de Klei et PCH_g) vont être effectués directement sur les phénotypes ajustés par rapport à l'âge et le sexe. Cependant, nous aimerions souligner que la méthode de Bonferroni largement utilisée pour la correction des tests multiples n'est pas idéale pour cette application génétique, car elle suppose que tous les SNPs sont statistiquement indépendants les uns des autres, alors que ce n'est pas le cas ici (dépendance via déséquilibre de liaison). On va utiliser la technique de FDR discutée dans le chapitre 1 comme technique de correction des tests multiples. Les résultats des analyses sont donnés dans la section suivante.

5.4.4 Résultat d'application de l'approche PCH

Pour identifier des variantes à risque de la charge amyloïde, nous avons conçu une étude d'association basée sur l'approche PCH de Klei et notre nouvelle approche PCH_g.

Approche PCH_g

Dans PCH_g, on cherche à estimer les effets combinés des SNPs. Puisque nous n'avons pas un choix optimal pour le nombre de SNPs à analyser avec notre approche, nous

avons opté pour une façon simple qui consiste à analyser 20 SNPs à la fois, sous forme de fenêtres glissantes, ces fenêtres sont chevauchantes (voir figure 5.9). Cette approche va permettre d'analyser des régions chromosomiques d'intérêt plutôt que des SNPs isolés.

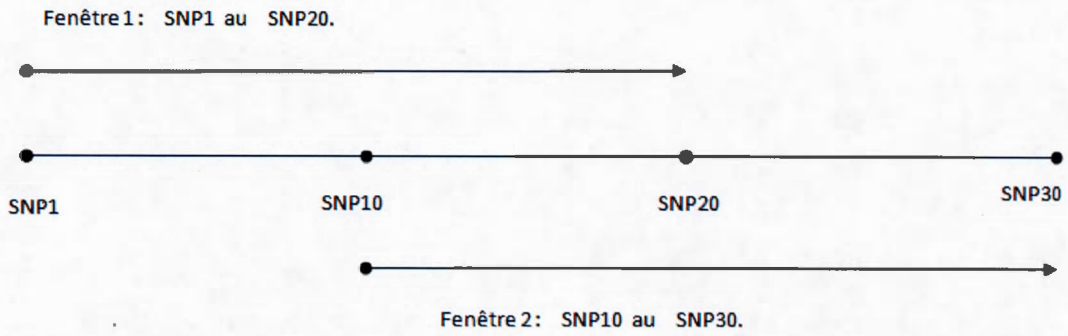


Figure 5.9: Illustration de l'idée des fenêtres glissantes. La première fenêtre contient le SNP1 jusqu'au SNP20. Le deuxième fenêtre contient le SNP10 jusqu'au SNP30.

Pour identifier des régions à risque de l'Alzheimer, nous avons conçu une stratégie d'association en deux étapes :

- l'analyse en composante principale combinée à l'héritabilité généralisée, PCH_g, est appliquée dans chaque fenêtre ;
- les fenêtres déclarées, associées significativement avec le PCH, après correction FDR, vont être analysées individuellement pour déterminer le (ou les) marqueur (s) qui sont identifiés en commun avec l'approche de PCH de Klei.

L'application de l'algorithme de PCH_g sur les phénotypes ajustés a montré 7 fenêtres avec des p-valeurs très significatives. Elles sont données dans le tableau 5.9 avec les

p-valeurs correspondantes.

Tableau 5.9: Résultats significatifs de l'approche PCH_g, à un seuil de 5%, après correction de FDR.

Fenêtre significative	p-valeurs
78	$1.412089e^{-04}$
265	$1.998835e^{-04}$
266	$8.749569e^{-06}$
436	$7.954393e^{-05}$
481	$4.058467e^{-04}$
499	$3.966272e^{-15}$
632	$1.345825e^{-07}$

Le graphique type manhattan de PCH_g est donné dans la figure 5.10. Les détails des fenêtres identifiées dans PCH_g incluant l'identifiant du marqueur sont donnés dans l'annexe C. Les fréquences alléliques des fenêtres significatives sélectionnées par l'approche PCH_g sont présentées dans la figure 5.11 qui donne la fréquence minimale et maximale correspondant à chaque fenêtre indiquée dans l'axe des abscisses, et les détails sont donnés dans l'annexe D.

Pour voir si les fenêtres ont des signaux similaires ou bien différents, nous avons décidé de vérifier le déséquilibre de liaison entre ces fenêtres. Les corrélations trouvées entre les fenêtres 265, 266 sont dues simplement au fait que les fenêtres glissantes sont chevauchantes (les fenêtres glissantes avancent de 20 marqueurs dans la première étape et reculent de 10 dans son deuxième choix d'analyse). Par conséquent, nous pouvons avancer que la part partagée entre ces fenêtres contient de l'information

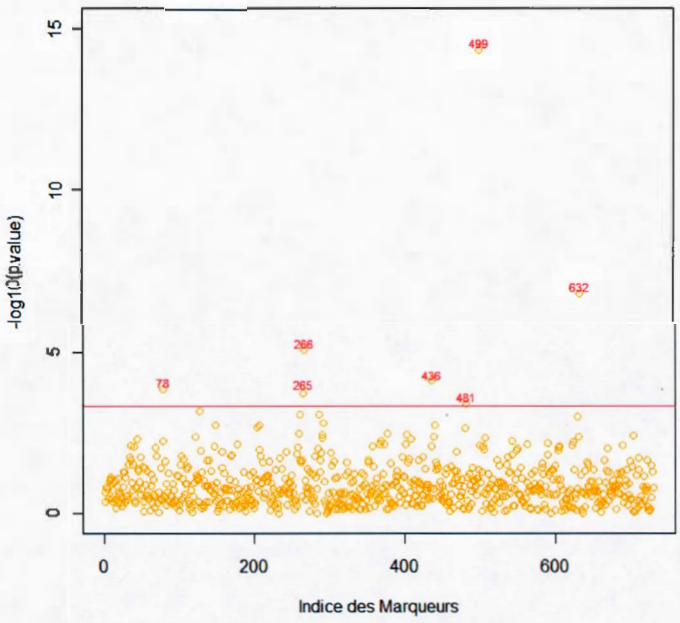
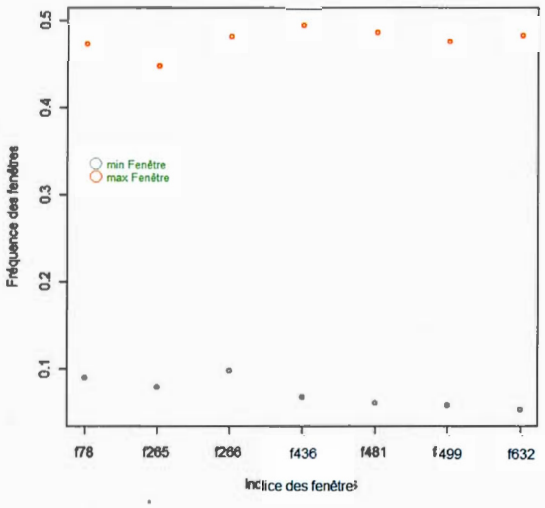


Figure 5.10: Le graphique type manhattan de l'approche PCH_g appliquée a l'analyse d'association charge amyloïde-SNPs. L'axe des ordonnées y représente -$log10$(p-valeur) et l'axe des abscisses x désigne les indices des marqueurs selon l'ordre de la matrice des génotypes. La ligne rouge indique un seuil commun déduit de la procédure FDR ($-log10\,(7 \times 0.05/730)$) pour l'ensemble des fenêtres ($m = 730$). m est le nombre d'hypothèses testées.

génétique utile.

Au total, nous avons trouvé 19 fenêtres pour lesquelles la matrice X était singulière. Pour contourner ce problème, nous avons utilisé la décomposition en valeurs singulières de la matrice X, et nous avons réappliqué l'algorithme de PCH_g qui n'a pas

Figure 5.11: Le minimum et le maximum des fréquences des SNPs sélectionnées par l'approche PCH_g correspondant à chaque fenêtre.



donné des résultats significatifs pour aucune des 19 fenêtres.

Approche PCH de Klei

Nous avons utilisé l'algorithme de PCH de Klei combiné avec l'algorithme d'ajustement correspondant décrit dans la section 3.2 pour réaliser les 7318 tests d'association. L'analyse d'association entre les phénotypes ajustés et les SNPs a révélé un marqueur hautement significatif "$rs769449\_A$" (p-valeur=$1.139189e-6$, MAF=0.1778846). Le résultat de l'algorithme de PCH de Klei est donné dans la figure 5.12 à l'aide d'un graphique type manhattan.

Une dernière comparaison pour compléter notre analyse sera faite sur les capacités de détection de l'analyse univariée dans la sous-section prochaine 5.4.5 et sur sa per-

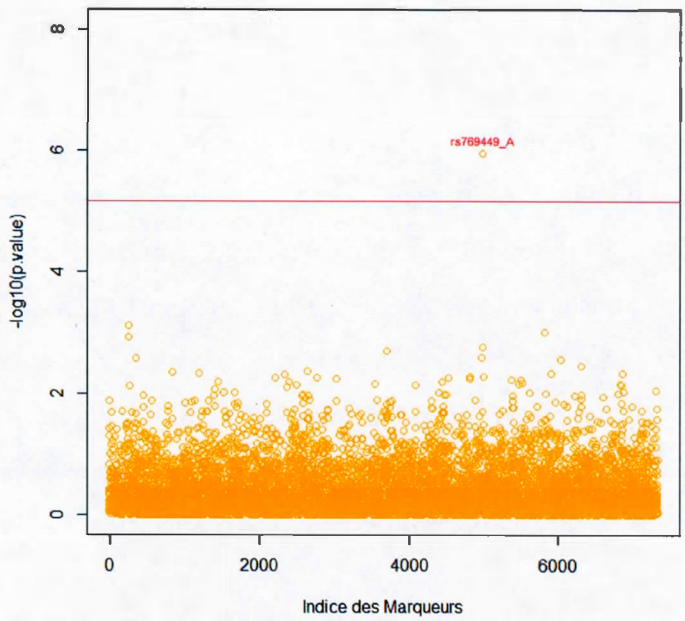


Figure 5.12: Graphique de type manhattan du résultat de l'application de l'algorithme de PCH de Klei. L'axe des ordonnées y représente -log10(p-valeur) et l'axe des abscisses x désigne les indices des marqueurs selon l'ordre de la matrice des génotypes. La ligne rouge indique un seuil commun déduit de la procédure FDR pour l'ensemble des marqueurs ($1 \times 0.05/7318$).

formance de détection. On se demande si l'analyse univariée sera capable de détecter les variantes causales (une ou plusieurs) identifiées dans les deux approches PCH_g et PCH de Klei.

5.4.5 Analyse univariée des données ADNI

Normalement il y a 702 525 analyses univariées à effectuer (96×7318), ce qui représente un grand temps de calcul. Une approche alternative est d'utiliser un pseudo phénotype, qui est une mesure globale de l'amyloïde bêta dans le cerveau d'un sujet, puis effectuer une analyse univariée. Dans la base de données réelle, on récupère la variable SURV45 qui désigne le pseudo phénotype. Une analyse préliminaire de la signification des covariables (l'âge, le sexe et l'éducation) par rapport au pseudo phénotype n'a montré des résultats significatifs pour aucune covariable. Les résultats de l'analyse univariée du pseudo phénotype et les 7318 SNPs sont donnés dans la figure 5.13. L'analyse univariée a détecté seulement trois variantes causales, ce qui confirme la puissance de notre approche sur l'analyse univariée.

5.4.6 Comparaison du résultat de PCH_g et PCH de Klei

L'approche PCH de Klei a réussi à montrer une association solide du marqueur $rs769449$ appartenant au gène ApoE mais il n'a pas réussi à détecter des associations pangénomiques supplémentaires significatives. Le marqueur hautement significatif détecté par l'approche PCH de Klei a été identifié aussi par l'approche PCH_g dans la fenêtre 499, ce qui prouve la puissance de notre approche. De plus, un ensemble de SNP avec une forte évidence d'association a été identifié par l'approche PCH_g a été déjà identifié dans la littérature comme significativement associé à la charge Amyloïde que PCH de Klei n'a pas réussi à identifier comme significatifs (voir tableau 5.10).

L'approche PCH de Klei ne met en valeur qu'un seul SNP alors que l'approche PCH_g pointe des régions importantes non identifiées par PCH de Klei.

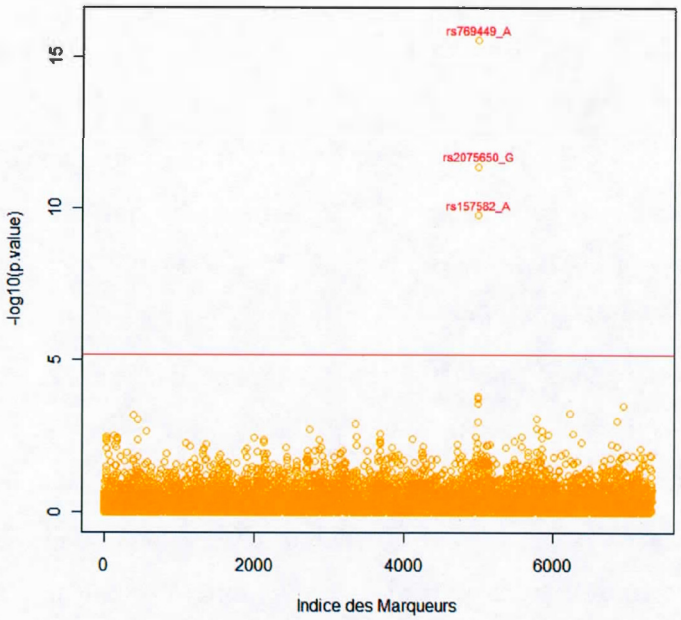


Figure 5.13: Graphique de type manhattan des résultats des analyses univariées des 7318 SNPs par rapport au pseudo phénotype. L'axe des abscisses désigne les marqueurs et l'axe des ordonnées les -log10(p-valeurs) de l'analyse univariée de pseudo phénotype par rapport à chaque SNP désigné dans l'axe d'abscisses.

On souligne à l'occasion que la différence de nombre de SNPs identifiés par l'approche univariée et l'approche PCH de Klei est due possiblement au fait que le phénotype utilisé (SURV45) dans l'analyse univariée est une moyenne globale de la charge amyloïde alors que les phénotypes utilisés dans l'approche de Klei sont des observations spécifiques aux régions et il est possible qu'il y a de la multicolinéarité entre les phénotypes, ce qui a biaisé les résultats. En résumé, nous apportons comme conclusion déduite de l'approche PCH_g plusieurs constats :

- notre étude constitue une preuve de plus d'une association entre les marqueurs associés à ApoE et l'Alzheimer ;
- en dehors de la région du gène ApoE, nous avons détecté des associations significatives avec la région du gène TOMM40. Plusieurs SNPs dans cette région ont montré des p-valeurs hautement significatives dans plusieurs études génétiques portant sur l'Alzheimer, ce qui suggère que les signaux que nous avons trouvés dans notre approche dans la fenêtre 499 sont des signaux réels et non des artefacts de l'analyse ou erreur du type 1 ;
- ces résultats suggèrent que les SNPs non ApoE associés sont enrichis avec des marqueurs qui sont vraiment associés à la charge ;
- l'utilisation de PCH_g nous a permis d'identifier des variantes de risque qui ne sont pas identifiées par GWAS, peut-être parce que ses variantes ne passent pas la correction de tests multiples stricts appliqués dans le GWAS ;
- toutefois, il est encourageant de constater que plusieurs SNPs correspondent à des régions proches des pics connus et /ou à d'excellents gènes candidats biologiques.

Ensemble, ces résultats confirment la spécificité de nos résultats et que PCH_g peut être utilisé pour identifier les variantes génétiques qui influencent les différentes facettes de la maladie d'Alzheimer. Ainsi, une confirmation dans d'autres populations (autre que celle incluse dans notre échantillon) est nécessaire pour déterminer la généralisation de la contribution de chaque SNP au risque d'Alzheimer et la possibilité de variantes étiologiques de population spécifique bien que l'effet de ApoE a été étudié en détail dans plusieurs populations. Il reste aussi un défi de conversion de ces résultats en des interprétations biologiquement significatives. Nous avons fait des recherches à partir du site `http://www.ncbi.nlm.nih.gov/`, National Center for Biotechnology Information, pour identifier les gènes impliqués dans chaque fenêtre

détectée par l'approche PCH_g. Le résultat est donné dans l'annexe E. Nous avons regardé les articles reliés à chaque gène dans le site. Et lors de cette recherche, il y avait une fenêtre qui a attiré notre attention. C'est la fenêtre 78 qui semble répondre à certaines questions souvent posées dans le cadre de la maladie d'Alzheimer et que nous allons discuter dans la prochaine section.

5.5 Gènes candidats détectés dans la fenêtre 78

Notre étude est la première à notre connaissance qui implique le gène DPP9 (dipeptidyl-peptidase 9) dans la charge amyloïde de la maladie d'Alzheimer. Nous avons mené une investigation dans la littérature sur ce gène et nous avons trouvé des études qui portent sur son rôle potentiel dans les réponses inflammatoires, l'apoptose et la neurogénèse. Et pour établir le lien entre DPP9 et l'Alzheimer, nous nous sommes intéressés aux caractéristiques de ses activités dans la maladie d'Alzheimer.
Dans ce but, nous allons d'abord présenter les caractéristiques des réponses immunitaires dans le cas de la maladie d'Alzheimer et l'effet de DPP9 sur ce système, ensuite nous parlons de l'apoptose et l'implication de DPP9 dans ce comportement cellulaire, et enfin nous discutons de la neurogénèse dans le cas d'Alzheimer, pour voir l'effet de DPP9 sur cette propriété cellulaire.

5.5.1 Le rôle de l'inflammation dans l'Alzheimer

Les réponses immunes contre les pathogènes se déroulent d'une façon particulière dans le système nerveux central (le cerveau, le cervelet et la moelle épinière). Enfermé dans une coquille rigide (crâne, colonne vertébrale), il supporte difficilement les

oedèmes accompagnant les réactions inflammatoires[7]. La protection immunitaire du système nerveux central (noté SNC) est assurée principalement par une population de cellules résidantes : les microglies (Smith *et al.*, 2012). Lors d'une inflammation, les microglies deviennent hyper ramifiées, prolifèrent, migrent vers le site de la lésion, et deviennent une source importante de cytokines pro-inflammatoires ou de facteurs cytotoxiques (IL1, $TNF\alpha$ et IL6). Un niveau élevé de ces cytokines implique un effet toxique néfaste sur la survie des neurones (Li *et al.*, 2014).

Wisniewski *et al.* (1989, 1992) et Weitz et Town (2012) ont observé, par analyse au microscope électronique, qu'au moins 80% des plaques amyloïdes colocalisent avec les microglies activées dans le cerveau de patients atteints de l'Alzheimer. Des expériences in vivo et in vitro[8] suggèrent que les microglies, recrutées dans le site des plaques amyloïdes, sont en mesure d'entourer et de phagocyter les peptides $A\beta$. Ceci suggère une déficience du système immunitaire dans la maladie d'Alzheimer.

Plusieurs cytokines[9] pro-inflammatoires ont été trouvés dérégulées chez les patients atteints d'Alzheimer (Akiyama *et al.*, 2000). IL1, IL6 et TNFα qui sont des cytokines pro-inflammatoires, ont été trouvées régulées à la hausse chez les patients d'Alzhei-

---

7. `http://www.larousse.fr/encyclopedie/medical`

8. in vivo sont des recherches ou des examens pratiqués sur un organisme vivant, par opposition à in vitro qui sont des recherches effectuées sur des organes, des tissus, des cellules, des composants de la cellule, des protéines, ou des biomolécules. source Wikipédia.

9. Protéine ou polypeptide sécrété par une cellule et agissant sur la cellule elle-même ou sur les cellules du même tissu en vue de produire un effet spécifique de la cytokine et de la cellule qui le reçoit.

mer (Bauer *et al.*, 1991), et ont été vues comme impliquées dans l'accumulation des dépôts amyloïdes. Selon les résultats de certaines études, l'IL-1 favorise la synthèse (Goldgaber *et al.*, 1989 ; Mackenzie, 2000) et le processus (Buxbaum *et al.*, 1992) de l'APP et peut donc promouvoir la production d'amyloïde et le dépôt des plaques, l'IL6 peut moduler la synthèse APP (Vandenabeele et Fiers, 1991), et l'augmentation dans la transcription et l'expression de l'APP (Ringheim *et al.*, 1998) et $TNF\alpha$ induit la mort des neurones (D'Souza *et al.*, 1996 ; Good *et al.*, 1996 ; McKee *et al.*, 1998 ; Mogi *et al.*, 1994).

5.5.2 Rôle de DPP9 dans le système nerveux central

Le gène DPP9 humain est localisé sur le chromosome 19 dans le locus 19p13.3. Olsen et Wagtmann (2002) ont rapporté que DPP9 est exprimée de manière ubiquitaire, avec des niveaux d'expression très élevés dans le muscle squelettique, le cœur et le foie, et le plus bas dans le cerveau. Des preuves suggèrent d'importantes contributions de DPP9 à divers processus biologiques, y compris le comportement cellulaire, la biologie du cancer, de la pathogenèse de la maladie, et les réponses immunitaires.

Dans le cadre d'une étude sur l'implication de DPP9 dans la maladie athéroscléreuse humains où DDP9 s'est trouvé très abondant dans les régions riches en macrophages, Zhang *et al.* (2013) ont mené des expériences et ont trouvé que DPP9 à un effet anti-inflammatoire lié à ses activités enzymatiques (Matheeussen *et al.*, 2013). L'inhibition de DPP9 diminue les cytokines pro-inflammatoires IL6 et $TNF\alpha$. Une surexpression de DPP9 induit une surproduction de cytokines (IL1, $TNF\alpha$, IL6), d'où le potentiel de son rôle dans l'Alzheimer par rapport à la variante découverte dans le cadre notre approche PCH_g. Les niveaux élevés des cytokines observées dans

l'Alzheimer peuvent être expliqués par une surexpression de DPP9 qui sort du niveau habituellement observé dans le cerveau.

Une activité similaire a été observée dans l'asthme[10]. Selon une étude récemment publiée par Pen et *al.* (2014) portant sur l'asthme des adultes et la démence (L'Alzheimer constitue 70 % à 80 % des cas de la démence), un total de 12771 patients souffrant d'asthme nouvellement diagnostiqués entre 2001-2003 ont été évalués avec 51 084 personnes sans asthme. Cette étude de cohorte a abouti à la conclusion que le risque de développement de la démence est significativement accru chez les patients souffrant d'asthme par rapport à ceux de la population générale. En outre, les augmentations de risque de démence sont plus élevés chez les personnes ayant des fréquences élevées d'hospitalisations.

Une activité significative de DPP9 liée à l'asthme a été signalée, et il semblerait que les activités enzymatique de DPP9 jouent un rôle important durant les réactions allergiques (Schade *et al.*, 2008).

5.5.3 Rôle de DPP9 dans l'apoptose dans le cas de la maladie d'Alzheimer

L'apoptose est définie comme une autodestruction cellulaire « programmée » qui peut survenir de façon naturelle dans le cadre d'un renouvellement cellulaire normal. Lorsque l'apoptose ne fonctionne pas correctement, les cellules qui devraient être éliminées peuvent persister et devenir immortelles (exemple : le cancer et la leucémie). Lorsque l'apoptose fonctionne trop bien, il tue trop de cellules et inflige des

10. L'asthme est une maladie inflammatoire chronique des voies respiratoires, qui se caractérise par une production élevée des cytokines.

dommages aux tissus (exemple la maladie de Parkinson). Des recherches récentes ont montré chez les personnes atteintes d'Alzheimer (Hochstrasser et *al.*) une diminution significative des niveaux de facteur de croissance épidermique[11] (EGF) et un mécanisme d'apoptose dérégulé qui conduit à une vaste destruction neuronale.

Dans une étude portant sur le rôle de DPP9 dans la voie de signalisation de facteur de croissance épidermique, Yu *et al.* (2006) ont trouvé qu'une surexpression de DPP9 peut provoquer une autodestruction spontanée des cellules en absence d'un stimulateur externe, suggérant un rôle de DPP9 dans la régulation de l'apoptose.
En effet, dans des conditions saines, à la suite de la liaison d'un facteur de la famille des facteurs de croissance épidermique avec un récepteur spécifique, une dimérisation du récepteur est observée. Une activation de l'activité tyrosine kinase est induite après la dimérisation, ce qui se traduit par la phosphorylation de résidus tyrosines présents sur les récepteurs, puis par le recrutement de protéines possédant des motifs de reconnaissance pour ces phosypholyrosines. Ces événements vont entraîner l'activation d'une voie de signalisation.
La voie de signalisation PI3K/AKT/mTOR, activée par la liaison du facteur de croissance EGF avec le récepteur EGFR, est importante dans la régulation du cycle cellulaire. La phosphorylation de PI3K active AKT qui peut alors phosphoryler et

11. L'EGF est un facteur de croissance aux multiples actions faisant partie d'une grande famille protéique. Son rayon d'action s'étend sur l'ensemble des tissus. L'action des facteurs de croissance de la famille de l'EGF est médiée par une famille de quatre récepteurs membranaires ubiquitaires appelés ErbB.

activer mTOR[12] et CREB. CREB régule la transcription de gènes anti-apoptotiques par exemple MCL-1, Bcl2 ou c-Jun. L'activation de cette voie de signalisation provoque la diminution de l'entrée des cellules en apoptose et permet donc la survie des cellules.

Selon les résultats de certaines études (Yao *et al.*, 2011 ; Zhang *et al.*, 2013), une surexpression de DPP9 atténue le facteur de croissance épidermique en inhibant l'activité de la voie de signalisation PI3T/AKT. Cet effet est spécifique au EGF (facteur de croissance épidermique). DPP9 inhibe la phosphorylation de AKT, ce qui entraîne une augmentation de l'apoptose spontanée et la suppression de la prolifération cellulaire. Cet effet inhibiteur sur la phospholyration de AKT dépend largement de l'activité enzymatique de DPP9 et il est médié par caspase 3/caspase 9 qui ont été signalés significativement élevés dans le cas d'une surexpression de DPP9.

5.5.4 Rôle de DPP9 dans la neurogénèse

La neurogénèse est un ensemble de processus qui permet la création d'un neurone fonctionnel au sein du système nerveux central. La neurogénèse permet la formation et le développement du cerveau. Elle peut aussi répondre à un traumatisme et permettra la réparation ou la cicatrisation de cellules endommagées du cerveau (Altman, 1969). La neurogénèse adulte (Alvarez-buylla et García-verdugo, 2002) est un processus séquentiel, durant lequel une cellule souche neuronale quiescente peut être activée et entamer les étapes de prolifération, différenciation, maturation et d'inté-

12. mTOR (de l'anglais mammalian target of rapamycin, en français cible de la rapamycine chez les mammifères) est une enzyme de la famille des sérine/thréonine kinase qui régule la prolifération cellulaire, la croissance cellulaire, la mobilité cellulaire, la survie cellulaire, la biosynthèse des protéines et la transcription.

gration fonctionnelle. De nombreuses études ont confirmé que certaines parties du cerveau des primates[13], y compris l'être humain, maintiendraient leur capacité de produire de nouveaux neurones durant toute la vie adulte (Kempermann et Gage, 1999 ; Kornack et Rakic, 1999 ; Gould *et al.*, 1997). En 1998, Eriksson *et al.* publient une étude où ils démontrent que de nouveaux neurones sont générés dans le gyrus dentelé de personnes ayant jusqu'à 72 ans.

Ce système est déficient dans le cas de la maladie d'Alzheimer[14], puisque plusieurs études longitudinales sur la maladie d'Alzheimer ont trouvé que l'accumulation des plaques amyloïdes qui entraînent successivement la mort des neurones commence de 10 à 15 ans avant l'apparition des symptômes à l'âge d'environ 65 ans (Buée *et al.*, 2010) et la neurogénèse n'intervient pas pour remplacer les neurones endommagés dans cette phase où ces derniers sont limités, ce qui soulève des interrogations.

Des études ont montré que des facteurs de croissance tels que EGF jouaient un rôle important dans la neurogénèse (Kempermann, 2006). Ce dernier serait capable de stimuler les cellules neuronales progénitrices quiescentes dans une étude sur des souris âgées par l'activation de la signalisation mTOR. Une explication potentielle est qu'à travers son action sur AKT (Yao *et al.*, 2011), DPP9 inhibe la neurogénèse.

Un autre gène TNFAIP8L1 identifié dans la même fenêtre semble avoir lui ici une implication dans l'apoptose, mais faute de temps et de ressources littéraires à libre accès, nos recherches ont été limitées.

---

13. Les primates correspondent à un ordre de mammifères, regroupant entre autres les singes, les lémuriens, les loris, les tarsiers ou l'homme. Source : `http://www.futura-sciences.com/`.

14. `http://www.cnrs.fr/insb/6.recherche/parutions2/articles2015/c-rampon.html`

Tableau 5.10: Liste des SNPs identifiés dans la fenêtre 499 et qui ont été déclarés significativement impliqués dans l'Alzheimer dans la littérature.

Gène	SNP	p-valeur	Source
PVRL2	rs6859	$5.39\ 10^{-7}$	ogue MWLOGUE MW *et al.* (2011)
		$7.78\ 10^{-28}$	Pérez-Palma *et al.* (2014)
Tomm40	rs157580	2.7710^{-29}	ogue MWLOGUE MW *et al.* (2011)
		9.6010^{-35}	Pérez-Palma *et al.* (2014)
Tomm40	rs157582	2.5410^{-91}	ogue MWLOGUE MW *et al.* (2011)
Tomm40	rs8106922	2.0610^{-28}	ogue MWLOGUE MW *et al.* (2011)
		1.1710^{-25}	Pérez-Palma *et al.* (2014)
Tomm40	rs1160985	2.7810^{-33}	ogue MWLOGUE MW *et al.* (2011)
Tomm40	rs2075650	1.7010^{-94}	ogue MWLOGUE MW *et al.* (2011)
		8.5410^{-116}	Pérez-Palma *et al.* (2014)
APOC1	rs445925	3.0210^{-4}	ogue MWLOGUE MW *et al.* (2011)
-	rs439401	8.8210^{-29}	Pérez-Palma *et al.* (2014)
ApoE	rs405509	2.2910^{-27}	Pérez-Palma *et al.* (2014)
ApoE	rs769449	4.710^{-19}	Zhang et Pierce (2014)
PVRL2	rs8104483	-	Meng *et al.* (2009)

CONCLUSION

La mise à disposition de quantités sans cesse croissantes de données issues du séquençage génomique a ouvert la voie à de nouvelles études d'association génétique. La plupart de ces études en vigueur ont été fondées sur l'analyse individuelle des marqueurs. Cependant, ces analyses pénalisent la puissance de détection des gènes ayant des effets pléiotropiques. Dans certains cas, les gènes ayant des effets pléiotropiques peuvent être trouvés en examinant séparément chaque caractère. Cependant, deux problèmes majeurs ne rendent pas toujours cette stratégie appropriée. Tout d'abord, les effets pléiotropiques pour chaque caractère peuvent être trop faibles pour être identifiés. Deuxièmement, plusieurs problèmes de tests peuvent soit réduire la puissance soit introduire de l'inflation à l'erreur du type 1 des tests d'association. Dans ce cadre, nous avons introduit une nouvelle approche permettant de tirer parti de ces idées : composantes principales d'héritabilité dans le cas des données génétiques de grandes dimensions (notée PCH_g).

Nous avons présenté dans ce mémoire comment il était possible avec PCH_g de rechercher des marqueurs à faibles effets en considérant le problème de recherche des signaux significatifs sous l'angle d'un algorithme qui cherche à résumer l'information utile de plusieurs variables (qualitatives ou quantitatives) corrélées dans un nombre réduit de nouvelles variables (scores) appelées composantes principales d'héritabilité. Ces nouveaux scores optimaux peuvent être utilisés dans des études d'association afin d'augmenter la puissance des tests statistiques pour détecter les gènes responsables

des maladies complexes. Nous avons proposé une nouvelle statistique qui teste une telle association entre ces nouvelles composantes principales et une région génomique, et nous avons approximé sa distribution sous l'hypothèse nulle à l'aide des techniques de permutation. Les études de simulation indiquent que cette approche est efficace pour différents types de corrélations entre les marqueurs, et qu'elle peut augmenter considérablement la puissance des études d'association. Lors de l'évaluation de l'atrophie associée à des fonctions cognitives spécifiques à l'Azheimer, PCH_g s'est montré très efficace en identifiant des régions d'intérêt très significatives. Nous avons aussi montré les limites et avantages des approches actuelles, en particulier l'approche des composantes principales d'héritabilité similaire à notre approche mais basée sur une analyse individuelle de marqueur, suggérée par Klei (Klei *et al.*, 2008). Nous avons également discuté d'un autre type d'information à prendre en compte dans la recherche de signaux : la notion de fenêtre glissante.

Une des principales limites à laquelle nous sommes soumis, concerne la taille de la fenêtre glissante à utiliser pour estimer les petites p-valeurs. La fenêtre glissante utilisée pour balayer les séquences présente le désavantage de fixer à priori sa taille. Et malgré cette limitation inhérente à la modélisation du problème, la recherche de signaux a permis d'avancer dans la connaissance des marqueurs du diagnostic de la maladie d'Alzheimer. Une deuxième limite qui affecte la qualité de notre approche est liée au nombre de permutations nécessaires pour tester l'association entre la composante principale d'héritabilité et la région génomique. Nous avons été obligés, dans la base de données réelle, d'augmenter le nombre de permutations pour chercher la stabilité des résultats dans l'estimation de p-valeur dans l'algorithme PCH_g ce qui a engendré des calculs intensifs. Une solution consiste à chercher la distribution exacte de la statistique de test. Une autre limite engendrée principalement par le jeu de

données est liée à la taille modeste de l'échantillon comparée au grand nombre de marqueurs à tester. Dès lors, certaines valeurs possibles de combinaisons de marqueurs ne vont pas être observées et la qualité de l'estimation des paramètres d'un modèle peut en être affectée.

ANNEXE A

DÉMONSTRATION DE LA FORMULE DU COEFFICIENT DE CORRÉLATION COMME MESURE DE DÉSÉQUILIBRE DE LIAISON.

Le coefficient r^2 est basé sur la statistique de χ^2 (voir A.1) de Pearson pour un tableau de contingence de non association entre les lignes et les colonnes du tableau (hypothèse nulle). Spécifiquement, nous pouvons écrire

$$r^2 = \frac{\chi^2}{N}, \tag{A.1}$$

avec

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

où $i = 1, 2$ lignes, $j = 1, 2$ colonnes et $(O_{ij} - E_{ij})^2 = (ND)^2$. Le nombre de degrés de liberté ddl=1. Au seuil de signification α, on rejette l'hypothèse nulle si $\chi^2 > \chi^2_{\alpha,ddl}$. Ainsi, nous avons

$$
\begin{aligned}
\chi^2 &= \sum_{ij} \frac{(ND)^2}{E_{ij}} \\
&= (ND)^2 \sum_{ij} \frac{1}{E_{ij}} \\
&= (ND)^2 \left(\frac{1}{Np_Ap_B} + \frac{1}{Np_Ap_b} + \frac{1}{Np_ap_B} + \frac{1}{Np_ap_b}\right) \\
&= ND^2 \frac{p_ap_b + p_ap_B + p_Ap_b + p_Ap_B}{p_Ap_ap_Bp_b} \\
&= ND^2 \frac{p_a(p_b + p_B) + p_A(p_b + p_B)}{p_Ap_ap_Bp_b} \\
&= ND^2 \frac{p_a + p_A}{p_Ap_ap_Bp_b} \\
&= ND^2 \frac{1}{p_Ap_ap_Bp_b}.
\end{aligned}
\tag{A.2}
$$

Par suite, nous avons

$$
r^2 = \frac{D^2}{p_Ap_ap_Bp_b}.
$$

ANNEXE B

LA VARIANCE GÉNOTYPIQUE

La formule de la variance génétique a été donnée dans 1.2.2, ici on démontre le résultat obtenu :

$$\sigma_G^2 = p^2(a-\mu)^2 + 2pq(d-\mu)^2 + q^2(-a-\mu)^2.$$

On développe les expressions entre parenthèses :

$$\begin{aligned}
a-\mu &= a - a(p-q) - 2pqd \\
&= a - ap + aq - 2pqd \\
&= a(1-p) + aq - 2pqd \\
&= aq + aq - 2pqd \\
&= 2aq - 2pqd \\
&= 2q(a-pd) \\
&= 2q(a-pd+dq-dq) \\
&= 2q(a-d(p-q)-dq).
\end{aligned}$$

On pose

$$\alpha = a - d(p - q),$$

alors

$$a - \mu = 2q(\alpha - dq).$$

Ainsi, nous avons

$$\begin{aligned}
-a - \mu &= -a - (p - q)a - 2pqd \\
&= -a(1 - q) - pa - 2pqd \\
&= -ap - pa - 2pqd \\
&= -2ap - 2pqd \\
&= -2p(a + qd) \\
&= -2p(a + qd + pd - pd) \\
&= -2p(a + d(q - p) + pd) \\
&= -2p(\alpha + pd) \\
&= -2p\alpha - 2p^2 d.
\end{aligned}$$

Nous avons aussi

$$\begin{aligned}
d-\mu &= d-(p-q)a-2pqd \\
&= d+(q-p)a-2pqd-2pqd+2pqd \\
&= d+(q-p)a-4pqd+2pqd \\
&= (q-p)a+d-4pqd+2pqd \\
&= (q-p)a+(p+q)d-4pqd+2pqd \\
&= (q-p)a+pd+qd-4pqd+2pqd \\
&= (q-p)a-4pqd+pd+qd+2pqd \\
&= (q-p)a-2pqd-2pqd+pd+qd+2pqd \\
&= (q-p)a-pd(2q-1)+qd(1-2p)+2pqd \\
&= (q-p)a+pd(1-2q)+qd(1-2p)+2pqd \\
&= (q-p)a+pd(1-q-q)+qd(1-p-p)+2pqd \\
&= (q-p)a-pd(q-p)+qd(q-p)+2pqd \\
&= (q-p)(a-pd+qd)+2pqd \\
&= (q-p)(a+d(q-p))+2pqd \\
&= (q-p)\alpha+2pqd.
\end{aligned}$$

En remplaçant dans la variance génétique, nous pouvons écrire

$$\begin{aligned}
\sigma_G^2 &= p^2(a-\mu)^2+2pq(d-\mu)^2+q^2(-a-\mu)^2 \\
&= p^2(2q(\alpha-dq))^2+2pq((q-p)\alpha+2pqd)^2+q^2(-2p\alpha-2p^2d)^2 \\
&= p^2(4q^2\alpha^2+4q^4d^2-4q^3\alpha d)+2pq((q-p)^2\alpha^2+4p^2q^2d^2 \\
&+ 4pqd(q-p)\alpha)+q^2(4p^2\alpha^2+4p^4d^2+8p^3\alpha d).
\end{aligned}$$

On isole les coefficients de $\alpha, \alpha d, d^2$ séparément

$$\begin{aligned}[4p^2 + 2pq(q-p)^2 + 4p^2q^2]\alpha^2 &= [8p^2q^2 + 2pq^3 - 4p^2q^2 + 2p^3q]\alpha^2 \\ &= 2pq(2pq + p^2 + q^2)\alpha^2 \\ &= 2pq\alpha^2.\end{aligned}$$

Puisque

$$\begin{aligned}[-8p^2q^3 + 8p^2q^2(q-p) + 8p^3q^2]\alpha d &= [-8p^2q^2(q-p) + 8p^2q^2(q-p)]\alpha d \\ &= 0,\end{aligned}$$

et

$$\begin{aligned}[4p^2q^4 + 8p^3q^3 + 4p^4q^2]d^2 &= 4p^2q^2[q^2 + p^2 + 2pq]d^2 \\ &= (2pqd)^2,\end{aligned}$$

nous pouvons conclure que

$$\sigma_G^2 = 2pq\alpha^2 + (2pqd)^2.$$

ANNEXE C

LISTE DES SNPS IDENTIFIÉS DANS LES FENÊTRES SIGNIFICATIVES DÉTECTÉ PAR L'APPROCHE PCH_G

Tableau C.1: Liste des SNPs identifiés dans les fenêtres significatives détecté par l'approche PCH_g.

fenêtre 78	fenêtre 265	fenêtre 266
rs17434614_C	rs1019445_T	rs1059369_T
rs12979155_G	rs17724992_G	rs7258465_C
rs4530284_G	rs3746183_T	rs888669_T
rs7250560_A	rs7226_A	rs4808814_T
rs12611262_T	rs17725099_A	rs10401741_A
rs10410702_G	rs888663_G	rs8109364_G
rs17363184_T	rs8101804_T	rs10409392_A
rs11880374_C	rs1059369_T	rs7258589_T
rs9917028_A	rs7258465_C	rs3170474_T
rs10417957_C	rs888669_T	rs2238647_A
rs7255247_C	rs3195944_G	rs731945_T
rs1865084_A	rs749451_T	rs17211392_T
rs892162_C	rs1043063_T	rs7251067_A
rs8110359_A	rs731945_T	rs10419977_T
rs10413885_A	rs17211392_T	rs709679_C
rs8108253_C	rs7251067_A	rs1043192_A
rs11148_G	rs4808814_T	rs12459017_T
rs8105807_G	rs10401741_A	rs10402779_G
rs10415651_C	rs8109364_G	rs11668617_A
rs11671605_G	rs10409392_A	rs3764628_A

Tableau C.2: Liste des SNPs identifiés dans les fenêtres significatives détecté par l'approche PCH_g (suite).

fenêtre 436	fenêtre 481	fenêtre 499	fenêtre 632
rs12976749_C	rs4802189_A	rs8104483_G	rs2965262_A
rs9304576_A	rs4251938_G	rs11879589_A	rs2965261_T
rs12052046_T	rs4760_C	rs395908_T	rs2914357_T
rs979971_A	rs2302524_C	rs2075642_A	rs17207094_T
rs12984247_T	rs4251864_C	rs387976_G	rs16984712_C
rs11083473_G	rs2239372_A	rs6859_A	rs2914354_T
rs749702_T	rs2239373_G	rs283814_T	rs2965260_A
rs4802747_T	rs4251842_G	rs157580_G	rs10407424_T
rs2287728_A	rs2239374_T	rs2075650_G	rs10416473_G
rs4491650_G	rs2286960_T	rs157582_A	rs3960965_A
rs8111466_T	rs344783_T	rs8106922_G	rs8109165_G
rs1966961_A	rs344776_T	rs1160985_T	rs4801974_C
rs12980315_G	rs344770_A	rs405509_C	rs11084235_C
rs10424200_T	rs10775541_C	rs769449_A	rs10406904_T
rs4802887_T	rs4493171_T	rs439401_T	rs2052858_C
rs1035525_C	rs436681_G	rs445925_T	rs7256991_A
rs2075588_G	rs1994417_T	rs1064725_G	rs12983823_A
rs2229259_A	rs346047_A	rs5157_T	rs4803083_G
rs17850798_T	rs11083718_A	rs5167_G	rs8105396_C
rs889327_T	rs346044_T	rs10413089_C	rs8104523_A

ANNEXE D

DÉTAIL DES FRÉQUENCES DES SNPS INCLUS DANS LES FENÊTRES SÉLECTIONNÉES PAR L'APPROCHE PCH_G

Les fréquences alléliques des fenêtres significatives sélectionnées par l'approche PCH_g sont donnés dans les tableaux suivants :

Tableau D.1: Détail de MAF des SNPs inclus dans les fenêtres 78, 265 et 266 sélectionnées par l'approche PCH_g.

Fenêtre 78		Fenêtre 265		Fenêtre 266	
SNPs	Fréquences	SNPs	Fréquences	SNPs	Fréquences
rs17434614_C	0.1973	rs1019445_T	0.2252	rs1059369_T	0.2458
rs12979155_G	0.4479	rs17724992_G	0.3208	rs7258465_C	0.2736
rs4530284_G	0.3281	rs3746183_T	0.2034	rs888669_T	0.1695
rs7255247_C	0.4746	rs3195944_G	0.3099	rs731945_T	0.4153
rs1865084_A	0.1998	rs749451_T	0.1065	rs17211392_T	0.2433
rs892162_C	0.4237	rs1043063_T	0.1186	rs7251067_A	0.4964
rs7250560_A	0.454	rs7226_A	0.322	rs4808814_T	0.1404
rs12611262_T	0.3656	rs17725099_A	0.2264	rs10401741_A	0.2203
rs10410702_G	0.1477	rs888663_G	0.4249	rs8109364_G	0.0666
rs17363184_T	0.0981	rs8101804_T	0.4492	rs10409392_A	0.0884
rs11880374_C	0.3293	rs1059369_T	0.2373	rs7258589_T	0.2385
rs9917028_A	0.3898	rs7258465_C	0.4007	rs3170474_T	0.4734
rs10417957_C	0.2869	rs888669_T	0.1998	rs2238647_A	0.1017
rs8110359_A	0.0896	rs731945_T	0.1429	rs10419977_T	0.2966
rs10413885_A	0.1901	rs17211392_T	0.4831	rs709679_C	0.4564
rs8108253_C	0.2748	rs7251067_A	0.4153	rs1043192_A	0.2954
rs11148_G	0.1211	rs4808814_T	0.27	rs12459017_T	0.46
rs8105807_G	0.3414	rs10401741_A	0.0969	rs10402779_G	0.0981
rs10415651_C	0.1114	rs8109364_G	0.3039	rs11668617_A	0.1404
rs11671605_G	0.1998	rs10409392_A	0.1453	rs3764628_A	0.092

Tableau D.2: Détail de MAF des SNPs inclus dans les fenêtres 436, 481 et 499 sélectionnées par l'approche PCH_g.

Fenêtre 436		Fenêtre 481		Fenêtre 499	
SNPs	Fréquences	SNPs	Fréquences	SNPs	Fréquences
rs12976749_C	0.3656	rs4802189_A	0.1973	rs8104483_G	0.0787
rs9304576_A	0.2893	rs4251938_G	0.4479	rs11879589_A	0.2772
rs12052046_T	0.3148	rs4760_C	0.3281	rs395908_T	0.1598
rs979971_A	0.0521	rs2302524_C	0.4746	rs2075642_A	0.0981
rs12984247_T	0.1804	rs4251864_C	0.1998	rs387976_G	0.4455
rs11083473_G	0.4177	rs2239372_A	0.4237	rs6859_A	0.368
rs749702_T	0.0981	rs2239373_G	0.454	rs283814_T	0.2288
rs4802747_T	0.201	rs4251842_G	0.3656	rs157580_G	0.2663
rs2287728_A	0.0799	rs2239374_T	0.1477	rs2075650_G	0.2046
rs4491650_G	0.2821	rs2286960_T	0.0981	rs8106922_G	0.2252
rs8111466_T	0.46	rs344783_T	0.3293	rs1160985_T	0.3208
rs1966961_A	0.4625	rs344776_T	0.3898	rs405509_C	0.2034
rs12980315_G	0.0714	rs344770_A	0.2869	rs769449_A	0.3099
rs10424200_T	0.3426	rs10775541_C	0.0896	rs439401_T	0.1065
rs4802887_T	0.4843	rs4493171_T	0.1901	rs445925_T	0.1186
rs1035525_C	0.0605	rs436681_G	0.2748	rs1064725_G	0.322
rs2075588_G	0.4588	rs1994417_T	0.1211	rs5157_T	0.2264
rs2229259_A	0.3051	rs346047_A	0.3414	rs157582_A	0.3184
rs17850798_T	0.1235	rs11083718_A	0.1114	rs5167_G	0.4249
rs889327_T	0.414	rs346044_T	0.1998	rs10413089_C	0.4492

Tableau D.3: Détail de MAF des SNPs inclus dans la fenêtre 632 sélectionnée par l'approche PCH_g.

SNPs	Fréquences
rs2965262_A	0.2252
rs2965261_T	0.3208
rs2914357_T	0.2034
rs17207094_T	0.3099
rs16984712_C	0.1065
rs2914354_T	0.1186
rs2965260_A	0.322
rs10407424_T	0.2264
rs10416473_G	0.4249
rs3960965_A	0.4492
rs8109165_G	0.2373
rs4801974_C	0.4007
rs11084235_C	0.1998
rs10406904_T	0.1429
rs2052858_C	0.4831
rs7256991_A	0.4153
rs12983823_A	0.27
rs4803083_G	0.0969
rs8105396_C	0.3039
rs8104523_A	0.1453

ANNEXE E

LISTE DES GÈNES IDENTIFIÉS PAR RAPPORT AUX FENÊTRES SIGNIFICATIVES DÉTECTÉS PAR PCH_G

Fenêtre 78		Fenêtre 265		Fenêtre 266	
SNPs	gènes	SNPs	gènes	SNPs	gènes
rs892162_C		rs1019445	PGPEP1	rs1059369	GDF15
rs10417957_C	TNFAIP8L1	rs17724992	PGPEP1	rs7258465	SSBP4
rs11671605_G	DPP9	rs3746183	PGPEP1	rs888669	SSBP4
rs9917028_A	TNFAIP8L1	rs3195944	PGPEP1	rs731945	ELL
rs8105807_G	DPP9-AS1	rs749451	PGPEP1	rs17211392	ELL
rs10415651_C	DPP9	rs1043063	PGPEP1	rs7251067	ELL
rs10410702_G		rs7226	PGPEP1	rs4808814	
rs4530284_G		rs17725099		rs10401741	
rs7255247_C		rs888663		rs8109364	KXD1
rs7250560_A		rs8101804	GDF15	rs10409392	C19orf60
rs8108253_C		rs1059369	GDF15	rs7258589	C19orf60
rs8110359_A	TNFAIP8L1	rs7258465	SSBP4	rs3170474	C19orf60
rs12611262_T		rs888669	SSBP4	rs2238647	CRLF1
rs11148_G	C19orf10	rs731945	ELL	rs10419977	CRLF1
rs1865084_A		rs17211392	ELL	rs709679	TMEM59L
17434614_C		rs7251067	ELL	rs1043192	TMEM59L
rs12979155_G		rs4808814		rs12459017	
rs11880374_C		rs10401741		rs10402779	
rs17363184_T		rs8109364	KXD1	rs11668617	KLHL26
rs10413885_A		rs10409392	C19orf60	rs3764628	KLHL26

Fenêtre 436		Fenêtre 481		Fenêtre 499	
SNPs	gènes	SNPs	gènes	SNPs	gènes
rs12976749	EIF3K	rs4802189		rs8104483	PVRL2
rs9304576		rs4251938	PLAUR	rs11879589	PVRL2
rs12052046		rs4760	PLAUR	rs395908	PVRL2
rs979971	ACTN4	rs2302524	PLAUR	rs2075642	PVRL2
rs12984247	ACTN4	rs4251864	PLAUR	rs387976	PVRL2
rs11083473	ACTN4	rs2239372	PLAUR	rs6859	PVRL2
rs749702	ACTN4	rs2239373	PLAUR	rs283814	PVRL2
rs4802747	ACTN4	rs4251842	PLAUR	rs157580	TOMM40
rs2287728	ACTN4	rs2239374	PLAUR	rs2075650	TOMM40
rs4491650		rs2286960	PLAUR	rs157582	TOMM40
rs8111466		rs344783	PLAUR	rs8106922	TOMM40
rs1966961		rs344776		rs1160985	TOMM40
rs12980315		rs344770		rs405509	APOE
rs10424200		rs10775541		rs769449	APOE
rs4802887	LGALS4	rs4493171		rs439401	
rs1035525	LGALS4	rs436681		rs445925	APOC1
rs2075588	ECH1	rs1994417		rs1064725	APOC1
rs2229259	ECH1	rs346047		rs5157	APOC4-APOC2
rs17850798	ECH1	rs11083718		rs5167	APOC2
rs889327	ECH1	rs346044		rs10413089	LOC105372418

Fenêtre 632	
SNPs	gènes
rs2965262	
rs2965261	Suspecté
rs2914357	
rs17207094	
rs16984712	
rs2914354	
rs2965260	
rs10407424	
rs10416473	FAM90A27P
rs3960965	
rs8109165	BIRC8
rs4801974	
rs11084235	
rs10406904	
rs2052858	
rs7256991	
rs12983823	
rs4803083	
rs8105396	
rs8104523	

BIBLIOGRAPHIE

Abou-Moustafa, K. et Ferrie, F. (2007). The minimum volume ellipsoid metric. In F. Hamprecht, C. Schnörr, et B. Jähne (dir.), *Pattern Recognition*, volume 4713 de *Lecture Notes in Computer Science* 335–344. Springer Berlin Heidelberg.

Akiyama, H., Barger, S., Barnum, S., Bradt, B., Bauer, J., Cole, G. M., Cooper, N. R., Eikelenboom, P., Emmerling, M., Fiebich, B. L. *et al.* (2000). Inflammation and alzheimer's disease. *Neurobiology of aging*, *21*(3), 383–421.

Altman, J. (1969). Autoradiographic and histological studies of postnatal neurogenesis. iv. cell proliferation and migration in the anterior forebrain, with special reference to persisting neurogenesis in the olfactory bulb. *The Journal of Comparative Neurology*, *137*(4), 433–457.

Alvarez-buylla, A. et García-verdugo, J. M. (2002). Neurogenesis in adult subventricular zone. *J. Neurosci*, 629–634.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, *7*(10), 781–791.

Bauer, J., Strauss, S., Schreiter-Gasser, U., Ganter, U., Schlegel, P., Witt, I., Yolk, B. et Berger, M. (1991). Interleukin 6 and α 2 macroglobulin indicate an acute phase state in alzheimer's disease cortices. *FEBS Letters*, *285*(1), 111 – 114.

Benjamini, Y. et Hochberg, Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Bertram, L., Lill, C. M. et Tanzi, R. E. (2010). The genetics of alzheimer disease : back to the future. *Neuron*, *68*(2), 270–281.

Buée, L., Blum, D., Bombois, S., Buée-Scherrer, V., Caillet-Boudin, M.-L., Colin, M., Deramecourt, V., Dhaenens, C.-M., Galas, M.-C., Hamdane, M., Humez, S., Maurage, C.-A., Pasquier, F., Sablonnière, B., Schraen-Maschke, S. et Sergeant, N. (2010). Comment les acteurs moléculaires de la pathologie alzheimer permettent

de comprendre la démence ? quelles conséquences diagnostiques et thérapeutiques ? *Therapie*, *65*(5), 401–407.

Buxbaum, J. D., Oishi, M., Chen, H. I., Pinkas-Kramarski, R., Jaffe, E. A., Gandy, S. E. et Greengard, P. (1992). Cholinergic agonists and interleukin 1 regulate processing and secretion of the alzheimer beta a4 amyloid protein precursor. *Proceedings of the National Academy of Sciences*, *89*(21), 10075–10078.

Camus, V., Payoux, P., Barré, L., Desgranges, B., Voisin, T., Tauber, C., La Joie, R., Tafani, M., Hommet, C., Chételat, G. *et al.* (2012). Using pet with 18f-av-45 (florbetapir) to quantify brain amyloid load in a clinical environment. *European journal of nuclear medicine and molecular imaging*, *39*(4), 621–631.

Carletti, G. (1988). Comparaison empirique de méthodes statistiques de détection de valeurs anormales à une et à plusieurs dimensions.

Cedazo-Mínguez, A. et Cowburn, R. (2001). Apolipoprotein e : a major piece in the alzheimer's disease puzzle. *Journal of cellular and molecular medicine*, *5*(3), 254–266.

Chee Seng ku, En Yun Loy, Y. P. et Chia, K. S. (2010). The poursuit of genom wide association studies : where are we now ? *Human Genetics*, *55*(4), 195–206.

Chen, Y., Chen, X. et Xu, L. (2008). Developing a size indicator for fish populations. *Scientia Marina*, *72*(2), 221–229.

Cobb, J., Wolf, P. A., Au, R., White, R. et D'agostino, R. (1995). The effect of education on the incidence of dementia and alzheimer's disease in the framingham study. *Neurology*, *45*(9), 1707–1712.

Corder,EH., Saunders,AM., Risch,NJ., Strittmatter, WJ., Gaskell,PC., S., Rimmler,JB., Locke,PA., Conneally,PM., Schmader,KE., Roses,AD., S., Haines, JL. et Pericak-Vance,MA. (1994). Protective effect of apolipoprotein e type 2 allele for late onset alzheimer disease. *Nature Genetics*, *7*(2), 180–184.

Crettaz de Roten, F. et Helbling, J.-M. (1996). Données manquantes et aberrantes : Le quotidien du statisticien analyste de données. *Revue de Statistique Appliquée*, *44*(2), 105–115.

De Vienne, D. (1998). *Les marqueurs moléculaires en génétique et biotechnologies végétales*. Editions Quae.

Draper, N. R., Smith, H. et Pownell, E. (1966). *Applied regression analysis.* Wiley New York.

D'Souza, S. D., Bonetti, B., Balasingam, V., Cashman, N. R., Barker, P. A., Troutt, A. B., Raine, C. S. et Antel, J. P. (1996). Multiple sclerosis : Fas signaling in oligodendrocyte cell death. *The Journal of Experimental Medicine*, *184*(6), 2361–2370.

Dudoit, S. et Van Der Laan, M. J. (2007). *Multiple testing procedures with applications to genomics.* Springer Science & Business Media.

Edgington, E. (1980). *Randomization tests.* Marcel Dekker,Inc.

Eriksson, P. S., Perfilieva, E., Björk-Eriksson, T., Alborn, A.-M., Nordborg, C., Peterson, D. A. et Gage, F. H. (1998). Neurogenesis in the adult human hippocampus. *Nature medicine*, *4*(11), 1313–1317.

Falconer. (1975). *Introduction à la génétique quantitative.* Dunod.

Ferrari, P. A. et Barbiero, A. (2012). Simulating ordinal data. *Multivariate Behavioral Research*, *47*(4), 566–589.

Folstein, M. F., Folstein, S. E. et McHugh, P. R. (1975). Mini-mental state : a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, *12*(3), 189–198.

Fratiglioni, L., Launer, L., Andersen, K., Breteler, M., Copeland, J., Dartigues, J., Lobo, A., Martinez-Lage, J., Soininen, H. et Hofman, A. (1999). Incidence of dementia and major subtypes in europe : A collaborative study of population-based cohorts. neurologic diseases in the elderly research group. *Neurology*, *54*(11 Suppl 5), S10–5.

Goldgaber, D., Harris, H. W., Hla, T., Maciag, T., Donnelly, R. J., Jacobsen, J. S., Vitek, M. P. et Gajdusek, D. C. (1989). Interleukin 1 regulates synthesis of amyloid beta-protein precursor mrna in human endothelial cells. *Proceedings of the National Academy of Sciences*, *86*(19), 7606–7610.

Good, P. F., Werner, P., Hsu, A., Olanow, C. W. et Perl, D. P. (1996). Evidence of neuronal oxidative damage in alzheimer's disease. *The American journal of pathology*, *149*(1), 21.

Gould, E., Mcewen, B. S., Tanapat, P., Galea, L. A. M. et Fuchs, E. (1997). Neurogenesis in the dentate gyrus of the adult tree shrew is regulated by psychosocial stress and nmda receptor activation. *Journal of Neuroscience*, (17), 2492–2498.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, 27–58.

Guedj, M. (2007). *Méthodes Statistiques pour l'analyse des données génétiques d'association à grande échelle.* (Statistique génétique). Université d'Évry-val d'Éssone.

Gumbel, E. J. (1958). *Statistics of extremes.* Columbia University Press,New York.

Hartl, D. L. et W.Jones, E. (1998). *Genetics : Principles and Analysis.4th ed.* Jones and Bartlett Publishers.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. et Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences, 106*(23), 9362–9367.

Hirschhorn, J. N. et Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics, 6*(2), 95–108.

Hollingworth, P., Harold, D., Jones, L., Owen, M. J. et Williams, J. (2011). Alzheimer's disease genetics : current knowledge and future challenges. *International journal of geriatric psychiatry, 26*(8), 793–802.

Johnson, G., Esposito, L., Barratt, BJ., Smith, AN., Heward, J., Di Genova, G., Ueda, H., Cordell, HJ., Eaves, IA. et Dudbridge, F., o. (2001). Haplotype tagging for the identification of common disease genes. *Nat. Genet., 29*(2), 233–237.

Johnson, R. A., Wichern, D. W. *et al.* (1992). *Applied multivariate statistical analysis.* Prentice hall Englewood Cliffs, NJ.

Jolliffe, I. (2002). *Principal component analysis.* Wiley Online Library.

Karch, C. M. et Goate, A. M. (2015). Alzheimer's disease risk genes and mechanisms of disease pathogenesis. *Biological Psychiatry, 77*(1), 43 – 51.

Kempermann, G. (2006). *Adult neurogenesis : stem cells and neuronal development in the adult brain.* Oxford University Press.

Kempermann, G. et Gage, F. H. (1999). New nerve cells for the adult brain. *Scientific american-american edition*, *280*, 48–67.

Klei, L., Luca, D., Devlin, B. et Roeder, K. (2008). Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic epidemiology*, *32*(1), 9–19.

Knijnenburg, T. A., Wessels, L. F., Reinders, M. J. et Shmulevich, I. (2009). Fewer permutations, more accurate p-values. *Bioinformatics*, *25*(12), i161–i168.

Kornack, D. R. et Rakic, P. (1999). Continuation of neurogenesis in the hippocampus of the adult macaque monkey. *Proceedings of the National Academy of Sciences*, *96*(10), 5768–5773.

Lange, K. (2002). *Mathematical and statistical methods for genetic analysis*. Springer Science & Business Media.

Letenneur, L., Gilleron, V., Commenges, D., Helmer, C., Orgogozo, J. M. et Dartigues, J. F. (1999). Are sex and educational level independent predictors of dementia and alzheimer's disease? incidence data from the paquid project. *Journal of Neurology, Neurosurgery & Psychiatry*, *66*(2), 177–183.

Lewontin, R. (1964). The interaction of selection and linkage. i. general considerations; heterotic models. *Genetics*, *49*(1), 49.

Li, Y., Tan, M.-S., Jiang, T. et Tan, L. (2014). Microglia in alzheimer's disease. *BioMed Research International*, ((2014) : 437483).

Logue MW, Schu M, Vardarajan BN *et al.* (2011). A comprehensive genetic association study of alzheimer disease in african americans. *Archives of Neurology*, *68*(12), 1569–1579.

Mackenzie, I. R. (2000). Anti-inflammatory drugs and alzheimer-type pathology in aging. *Neurology*, *54*(3), 732–732.

Matheeussen, V., Waumans, Y., Martinet, W., Van Goethem, S., Van der Veken, P., Scharpé, S., Augustyns, K., De Meyer, G. et De Meester, I. (2013). Dipeptidyl peptidases in atherosclerosis : expression and role in macrophage differentiation, activation and apoptosis. *Basic Research in Cardiology*, *108*(3).

McKee, A. C., Kawall, N. W., Schumacher, J. S. et Beal, M. F. (1998). The neurotoxicity of amyloid beta protein in aged primates. *Amyloid*, *5*(1), 1–9.

Mei, H., Chen, W., Dellinger, A., He, J., Wang, M., Yau, C., Srinivasan, S. et Berenson, G. (2010). Principal-component-based multivariate regression for genetic association studies of metabolic syndrome components. *BMC Genetics*, *11*(1), 100.

Meng, Y. A., Yu, Y., Cupples, L. A. et Farrer, Lindsay A ANDLunetta, K. L. (2009). Performance of random forest when snps are in linkage disequilibrium. *BMC bioinformatics.*, (10 :78).

Mogi, M., Harada, M., Riederer, P., Narabayashi, H., Fujita, K. et Nagatsu, T. (1994). Tumor necrosis factor- α (tnf- α) increases both in the brain and in the cerebrospinal fluid from parkinsonian patients. *Neuroscience Letters*, *165*(1–2), 208 – 210.

Morgenthaler, S. (2008). *Génétique statistique.* Springer Science & Business Media.

Neter, J., Kutner, M. H., Nachtsheim, C. J. et Wasserman, W. (1996). *Applied linear statistical models*, volume 4. Irwin Chicago.

Nordborg, M. et Tavaré, S. (2002). Linkage disequilibrium : What history has to tell us. *Trends Genet.*, *18*(2), 83–90.

Ollivier, L. (2002). *Eléments de génétique quantitative.* Editions Quae.

Ollivier, L. *et al.* (1971). L'héritabilité et sa mesure. *Bulletins et Mémoires de la Société d'anthropologie de Paris*, *7*(2), 159–167.

Olsen, C. et Wagtmann, N. (2002). Identification and characterization of human dpp9, a novel homologue of dipeptidyl peptidase iv. *Gene*, *299*(1), 185–193.

Ott, J. r. et Rabinowitz, D. (1999). A principal-components approach based on heritability for combining phenotype information. *Human heredity*, *49*(2), 106–111.

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P. *et al.* (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, *294*(5547), 1719–1723.

Pérez-Palma, E., Bustos, B. I., Villamán, C. F., Alarcón, M. A., Avila, M. E., Ugarte, G. D., Reyes, A. E., Opazo, C., De Ferrari, G. V. et the Alzheimer's Disease Neuroimaging Initiative, t. N.-L. F. S. G. (2014). Overrepresentation of glutamate signaling in alzheimer's disease : Network-based pathway enrichment using meta-analysis of genome-wide association studies. *PLos one*, *9*(4), e95413.

Raggad, B. (2009). Fondements de la théorie des valeurs extrêmes, ses principales applications et son apport à la gestion des risques du marché pétrolier. *Mathématiques et sciences humaines. Mathematics and social sciences*, (186), 29–63.

Ringheim, G. E., Szczepanik, A. M., Petko, W., Burgher, K. L., Zhu, S. Z. et Chao, C. C. (1998). Enhançement of beta-amyloid precursor protein transcription and expression by the soluble interleukin-6 receptor/interleukin-6 complex. *Molecular Brain Research*, *55*(1), 35 – 44.

Robbins, R. B. (1917). Some applications of mathematics to breeding problems. *Genetics*, *2*(5), 489.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American statistical association*, *79*(388), 871–880.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, *8*, 283–297.

Schade, J., Stephan, M., Schmiedl, A., Wagner, L., Niestroj, A. J., Demuth, H.-U., Frerker, N., Klemann, C., Raber, K. A., Pabst, R. et Hörsten, S. v. (2008). Regulation of expression and function of dipeptidyl peptidase 4 (dp4), dp8/9, and dp10 in allergic responses of the lung in rats. *Journal of Histochemistry & Cytochemistry*, *56*(2), 147–155.

Smith, J. A., Das, A., Ray, S. K. et Banik, N. L. (2012). Role of pro-inflammatory cytokines released from microglia in neurodegenerative diseases. *Brain Research Bulletin*, *87*(1), 10 – 20.

Speed, D., Hemani, G., Johnson, M. et Balding, D. (2012). Improved heritability estimation from genome-wide snps. *American journal of human genetics*, *91*(6), 1011–1021.

Storey, J. D. et Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, *100*(16), 9440–9445.

Tjio, J. H. et Levan, A. (1956). The chromosome number of man. *Hereditas*, *42*(1-2), 1–6.

Vandenabeele, P. et Fiers, W. (1991). Is amyloidogenesis during alzheimer's disease due to an il-1-/il-6-mediated 'acute phase response' in the brain ? *Immunology Today*, *12*(7), 217 – 219.

Wang, Y., Fang, Y. et Wang, S. (2007). Clustering and principal-components approach based on heritability for mapping multiple gene expressions. Dans *BMC proceedings*.

Weiss, K. et Clark, A. (2002). Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.*, *18*(1), 19–24.

Weitz, T. M. et Town, T. (2012). Microglia in alzheimer's disease : it's all about context. *International journal of Alzheimer's disease*, *2012*.

Wisniewski, H., Wegiel, J., Wang, K., Kujawa, M. et Lach, B. (1989). Ultrastructural studies of the cells forming amyloid fibers in classical plaques. *The Canadian journal of neurological sciences. Le journal canadien des sciences neurologiques*, *16*(4 Suppl), 535–542.

Wisniewski, H., Wegiel, J., Wang, K. et Lach, B. (1992). Ultrastructural studies of the cells forming amyloid in the cortical vessel wall in alzheimer's disease. *Acta neuropathologica*, *84*(2), 117–127.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E. et Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, *42*(7), 565–569.

Yao, T.-W., Kim, W.-S., Yu, D. M., Sharbeen, G., McCaughan, G. W., Choi, K.-Y., Xia, P. et Gorrell, M. D. (2011). A novel role of dipeptidyl peptidase 9 in epidermal growth factor signaling. *Molecular Cancer Research*, *9*(7), 948–959.

Yu, D. M. T., Wang, X. M., McCaughan, G. W. et Gorrell, M. D. (2006). Extraenzymatic functions of the dipeptidyl peptidase iv-related proteins dp8 and dp9 in cell adhesion, migration and apoptosis. *FEBS Journal*, *273*(11), 2447–2460.

Zannis, V. I., Kardassis, D. et Zanni, E. E. (1993). Genetic mutations affecting human lipoproteins, their receptors, and their enzymes. In *Advances in Human Genetics 21* 145–319.

Zhang, C. et Pierce, B. L. (2014). Genetic susceptibility to accelerated cognitive decline in the {US} health and retirement study. *Neurobiology of Aging*, *35*(6), 1512.e11 – 1512.e18.

Zhang, H., Chen, Y., Keane, F. M. et Gorrell, M. D. (2013). Advances in understanding the expression and function of dipeptidyl peptidase 8 and 9. *Molecular Cancer Research*, *11*(12), 1487–1496.

Zhang, K., Calabrese, P., Nordborg, M. et Sun, F. (2002). Haplotype block structure and its applications to association studies : power and study designs. *The American Journal of Human Genetics*, *71*(6), 1386–1394.