Modélisation de la composante génétique des maladies humaines : Données familiales et Modèles Mixtes

Claire Dandine-Roulland

NNT : 2016SACLS259

THÈSE DE DOCTORAT
DE
L'UNIVERSITÉ PARIS-SACLAY
PRÉPARÉE À
L'UNIVERSITÉ PARIS-SUD

ÉCOLE DOCTORALE n ° 570
EDSP Santé Publique

Spécialité de doctorat : Santé Publique - Génétique Statistique

Par

Claire Dandine-Roulland

Modélisation de la composante génétique des maladies humaines : Données Familiales et Modèles Mixtes

Thèse présentée et soutenue publiquement à Villejuif, le 4 Octobre 2016.

Composition du Jury :

Mr Christophe Ambroise	Directeur de Recherche, Inserm	Président du jury
Mr Laurent Abel	Directrice de Recherche, Inserm	Rapporteur
Mme Maria Martinez	Professeur, Université d'Évry Val d'Essonne	Rapporteur
Mme Anne-Louise Leutenegger	Chargée de Recherche, Inserm	Examinatrice
Mr Bertram Müller-Myhsok	Professeur, Max Planck Institute of Psychiatry	Examinateur
Mr Philippe Broët	Professeur, Université Paris-Sud	Directeur de thèse
Mr Hervé Perdry	Maitre de Conférences, Université Paris-Sud	Co-directeur de thèse

Modélisation de la composante génétique des maladies humaines : Données Familiales et Modèles Mixtes

Thèse de doctorat par Claire Dandine-Roulland
dirigée par Philippe Broët et Hervé Perdry
au sein de l'équipe « Genostat » du CESP (UMRS 1018)

Le modèle linéaire mixte a été formalisé il y a plus de 60 ans. Celui-ci permet d'estimer un modèle avec des effets fixes équivalents à ceux du modèle linéaire classique et des effets aléatoires. Ce type de modélisation, d'abord utilisé en génétique animale, est depuis quelques années largement utilisé en génétique humaine. Les utilisations de ce modèle sont nombreuses. En effet, il peut être utilisé en étude de liaison, d'association, pour l'estimation de l'héritabilité ou encore dans la recherche d'empreinte parentale et peut s'adapter à des données familiales ou en population.

Le but de mon doctorat est d'exploiter différentes méthodes basées sur les modèles mixtes d'abord sur des données génétiques en population puis sur des données génétiques familiales.

Dans un premier temps, nous explorons dans ce manuscrit la théorie des modèles linéaires mixtes et leur utilisation en génétique. Nous adaptons aussi certaines méthodes pour les appliquer à notre recherche. Ce travail a donné lieu au développement informatique d'un package R permettant d'utiliser ces modèles dans le cadre des études génétiques.

Dans un deuxième temps, nous utilisons les modèles linéaires mixtes pour l'estimation de l'héritabilité dans une étude en population française, l'étude Trois-Cités. Nous disposons dans cette étude des génotypes des tag-SNPs habituellement utilisés dans les études d'association ainsi que des lieux de naissance et de plusieurs traits anthropométriques quantitatifs tels que la taille. L'objectif est alors d'étudier la présence et la prise en compte dans l'analyse de stratification de population dans cette étude. Dans ce manuscrit, nous analysons les coordonnées géographiques des lieux de naissance. Nos résultats mettent en évidence la difficulté pour corriger correctement la stratification de population avec les méthodes classiques dans certains cas. Nous analysons ensuite les traits anthropométriques en particulier la taille dont nous estimons l'héritabilité à 39% dans la population de l'étude Trois-Cités.

Dans la dernière partie de ce manuscrit, nous nous concentrons sur les données familiales. Nous montrons le gain d'information que peut apporter ce type de données dans la recherche des variants causaux. Puis, nous explorons l'utilisation des modèles mixtes sur des données familiales en appliquant certaines des méthodes associées dans la recherche de signaux d'association pour la Sclérose en Plaques, une maladie auto-immune, en utilisant un échantillon d'une centaine de familles nucléaires avec au moins deux germains atteints. Nous avons alors mis en évidence l'inadéquation des méthodes classiques basées sur les modèles mixtes à ce type de données. Afin de mieux comprendre ce biais de sélection et de le corriger, plus d'investigations sont nécessaires.

Modelisation of genetic risk in human diseases : Family Data and Mixed Models

Linear mixed models have been formalized 60 years ago. These models allow to estimate fixed effects, as in the linear models, and random effects. First used in animal genetics, this type of modelling have been widely used in human genetics since a few years. Mixed models can be used in many genetic analysis ; linkage and association studies, heritability estimations and Parent-of Origin effects studies for population or familial data.

My thesis' aim is to investigate mixed models based methods, for genetic data in population and, for familial genetic data.

In the first part of my thesis, we investigated the mixed model statistical theory and their multiple uses in human genetics. We also adapted methods for our own work. An R package have been created which permits to analyze genetic data in R environment with mixed models.

In a second part, we applied mixed models on Three-Cities data, a French longitudinal study, to estimate heritability of several traits. For this analysis, we have access to tag-SNPs typically used in genome-wide association studies, birthplaces and several anthropometric traits. The aim of our study is to analyze presence of population stratification and evaluate methods to correct it. In the one hand, we analyzed birthplace geographic coordinates and showed that the correction for population stratification by classical method is not sufficient in this case. In the other hand, we analyzed anthropometric traits, in particular the height for which we estimated heritability to 39% in Three-Cities study population.

In the last part, we focused on family data. In a first work, we exploited familial information in causal variant research. In a second work, we explored mixed models uses for familial data, in particular association study, on Multiple Sclerosis data. We showed that mixed model methods can not be used without taking account the ascertainment scheme : in our data, all families have at least two affected sibs. To understand and correct this phenomenon, more investigations are needed.

Articles et communications

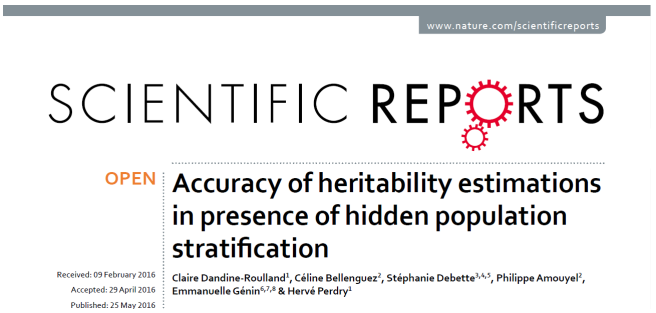
Articles


European Journal of Human Genetics (2015) 23, 1357–1363
© 2015 Macmillan Publishers Limited All rights reserved 1018-4813/15
www.nature.com/ejhg

ARTICLE

Where is the causal variant? On the advantage of the family design over the case–control design in genetic association studies

Claire Dandine-Roulland*[,1,2] and Hervé Perdry[1,2]


www.nature.com/scientificreports

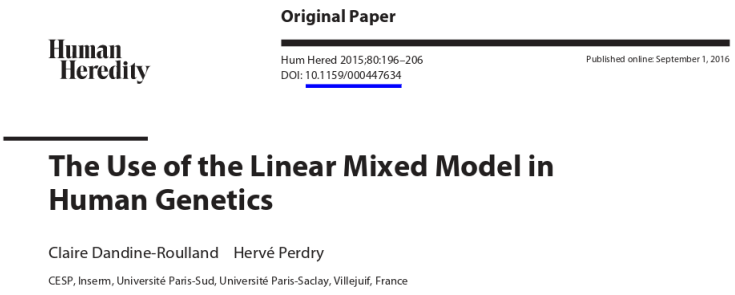
SCIENTIFIC REPORTS

OPEN

Accuracy of heritability estimations in presence of hidden population stratification

Received: 09 February 2016
Accepted: 29 April 2016
Published: 25 May 2016

Claire Dandine-Roulland[1], Céline Bellenguez[2], Stéphanie Debette[3,4,5], Philippe Amouyel[2], Emmanuelle Génin[6,7,8] & Hervé Perdry[1]


Human Heredity

Original Paper

Hum Hered 2015;80:196–206
DOI: 10.1159/000447634

Published online: September 1, 2016

The Use of the Linear Mixed Model in Human Genetics

Claire Dandine-Roulland Hervé Perdry

CESP, Inserm, Université Paris-Sud, Université Paris-Saclay, Villejuif, France

Communications

Orales

- ▷ European Mathematic Genetic Meeting (2015). The impact of population stratification on genomic heritability estimation. Abstract publié dans *Human Heredity* (**79**).

Posters

- ▷ European Mathematic Genetic Meeting (2014). Discrimination between correlated SNPs in genetic association studies : comparaison between case-control and familial studies. Abstract publié dans *Human Heredity* (**Special Edition**).

- ▷ Assises de Génétique (2016). Gaston, un package R pour les données de génome entier.
- ▷ European Mathematic Genetic Meeting (2016). Gaston, an R package for genome-wide data manipulation. Abstract publié dans *Human Heredity* (**Special Edition**).

Un grand merci !

Le meilleur ami de « merci » est « beaucoup ».

Michel Bouthot

N'étant pas des plus douée pour les discours, ces remerciements brefs mais sincères ne seront ni en vers ni en chanson et resterons très classiques...

Les premières personnes à remercier, à mon sens, sont mes deux directeurs de thèse, Hervé Perdry et Philippe Broët. Cela fait maintenant quatre années qu'ils m'ont accueillie dans leur laboratoire, d'abord en tant qu'étudiante dans le Master 2 GGS suivi par mon stage de six mois dans leur équipe puis durant les trois ans de ma thèse. En plus de m'avoir appris le métier de chercheur, ils m'ont également permise de participer et de m'investir dans différents enseignements me confortant ainsi dans mon projet professionnel.

Un merci tout particulier pour Hervé Perdry qui m'a offert l'opportunité de faire cette thèse mais aussi supportée et accompagnée pendant ces trois années et, ce, sans heurts (si si c'est vrai) !

Je tiens aussi à remercier tous les membres du PSG (mais non pas le club de football !) pour le partage de leur travaux mais aussi pour leur aide dans les réflexions liées à mon propre travail.

Je remercie également tous mes collègues de Villejuif pour leur convivialité et leur accueil. La diversité des membres du bâtiment Leriche de l'hôpital Paul Brousse a permis durant trois ans de grandes conversations à table lors du repas du midi allant de la cosmétique à la politique en passant par la culture et les sciences et, ce, parfois en un seul repas !

Il ne faut pas oublier, en dehors des heures de travail, mes proches qui m'ont soutenue et supportée durant ces trois années avec des périodes plus ou moins difficiles. J'ai une pensée toute particulière pour mon compagnon qui a su gérer mon caractère parfois explosif durant les périodes de stress malgré ces propres études qui ne sont pas toujours de tout repos.

Bonne continuation à toutes et à tous !

Avant-propos

Dans les débuts de mon doctorat, mes objectifs de travail étaient très différents de ce qui est présenté dans ce manuscrit. En effet, suite à des difficultés dans l'obtention des données génétiques prévues, j'ai eu d'abord l'occasion d'approfondir l'exploitation des données familiales en les utilisant pour la recherche de variants causaux. Puis, je me suis intéressée, avec mes directeurs de thèse, aux modèles mixtes qui prenaient une place croissante dans la littérature. Le sujet de ma thèse s'est alors construit petit à petit grâce aux différentes opportunités qui se sont présentées. Tout d'abord, grâce à la collaboration du laboratoire de Brest puis du comité scientifique de l'étude Trois-Cités, nous avons pu explorer l'utilisation des modèles linéaires mixtes et appréhender la méthodologie de ces modèles sur un problème réel qu'est l'estimation de l'héritabilité en présence de structure de population. Ce travail a demandé beaucoup de temps pour comprendre toute la complexité de cette méthodologie, créer nos outils d'analyses au travers d'un package R puis analyser de façon la plus complète possible les données de l'étude Trois-Cités. Nous avons alors obtenu depuis quelques mois les données familiales pour la Sclérose en Plaques tant attendues et nous avons donc commencé à exploiter ces données avec nos connaissances sur les modèles mixtes. Ces analyses ne sont pas encore complètes mais vont continuer après la rédaction de ce manuscrit.

J'ai essayé durant l'écriture de ce manuscrit d'être synthétique et didactique afin d'être lisible par le maximum de gens. Ce manuscrit se décompose en 6 chapitres :

1. Introduction
2. Le modèle linéaire mixte en génétique
3. Une application sur la population française : l'étude Trois-Cités
4. Le gain apporté par les données familiales, l'exemple des paires de germains atteints
5. Une application sur des données familiales, la Sclérose en Plaques
6. Perspectives

Le premier chapitre a pour vocation de donner les notions en génétique essentielles à la compréhension de ce manuscrit ainsi que d'exposer rapidement l'état actuel de la recherche en génétique. Il contient également un résumé de l'avancée des connaissances sur la Sclérose en Plaques, une maladie auto-immune qui sera étudiée dans ce manuscrit. Puis, le chapitre 2 contient la méthodologie liée aux modèles linéaires mixtes avec un résumé de son utilisation dans les études génétiques. Il est suivi d'une application aux données en population de l'étude Trois-Cités au chapitre 3. Les chapitres 4 et 5 se concentrent sur les données génétiques familiales avec un exemple de l'exploitation du gain d'information contenu dans ce type de données et

une étude de données familiales à l'aide des modèles mixtes. Enfin, au dernier chapitre, des perspectives sont proposées.

Bonne lecture !

Table des matières

Chapitre 1

Introduction

Ce premier chapitre a pour but de définir les différentes notions utilisées tout au long de ce manuscrit ainsi que de dresser brièvement le contexte scientifique dans le domaine de la génétique humaine. Nous y exposons aussi l'avancement des recherches pour la Sclérose en Plaques, une maladie auto-immune qui sera étudiée dans la suite de cette thèse.

1.1 Le génome, quelques notions essentielles

Pour commencer, définissons quelques termes spécifiques à la génétique. Cette section n'est pas exhaustive.

1.1.1 Qu'est ce que l'ADN ?

L'Acide DésoxyriboNucléique ou ADN est une molécule présente dans toute cellule vivante ainsi que dans certains virus. Elle contient la totalité de l'information génétique appelée aussi

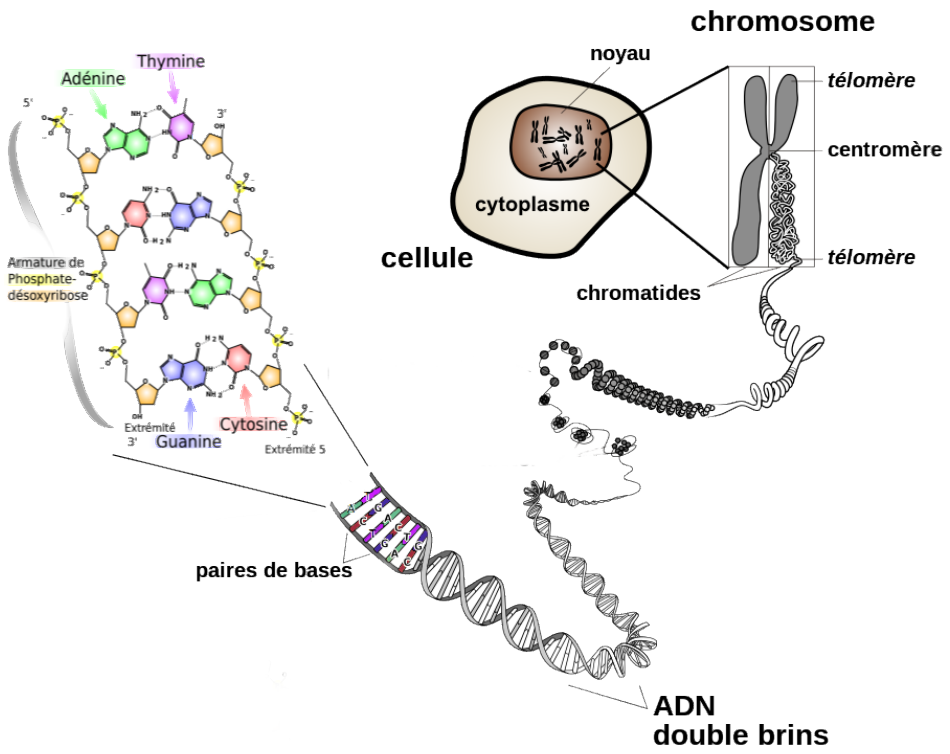


FIGURE 1.1 – L'ADN pour une cellule eucaryote.

génome. Elle est structurée en une double hélice, chacune composée de deux brins de nucléotides complémentaires. Chaque nucléotide est lui-même composé d'un groupement phosphate, d'une base azotée et d'un sucre. Les groupements phosphates et les sucres permettent aux bases azotées d'être liées ; les premiers afin de former un des brins de la double hélice et les seconds afin de relier les deux brins entre eux. L'information génétique est portée par les bases azotées. Il y en a de quatre sortes deux à deux complémentaires ; l'adénine (A) avec la thymine (T) et la cytosine (C) avec la guanine (G). Une *paire de bases* est constituée d'une base azotée et de son complémentaire sur le brin opposé.

Deux types de cellules peuvent être distingués :

- ▷ les cellules procaryotes n'ayant pas de vrai noyau et généralement un ADN circulaire unique,
- ▷ les cellules eucaryotes ayant un noyau (figure 1.1).

Dans les cellules eucaryotes, selon le stade de la cellule concernée (figure 1.2), l'hélice double brin peut être pêle-mêle dans le noyau, ou enroulée formant ainsi des *chromosomes* qui peuvent avoir une ou deux chromatides (molécule d'ADN enroulée avec d'autres protéines, figure 1.1). Dans la majorité de ce même type de cellule, il existe un autre type d'ADN, en plus de celui contenu dans le noyau, c'est l'ADN mitochondrial porté par les mitochondries ; des organites présentes dans le cytoplasme (liquide remplissant la cellule). Cet ADN a pour particularité d'être transmis presque exclusivement par la mère chez les espèces ayant une reproduction sexuée (qui nécessite deux parents génétiquement différents).

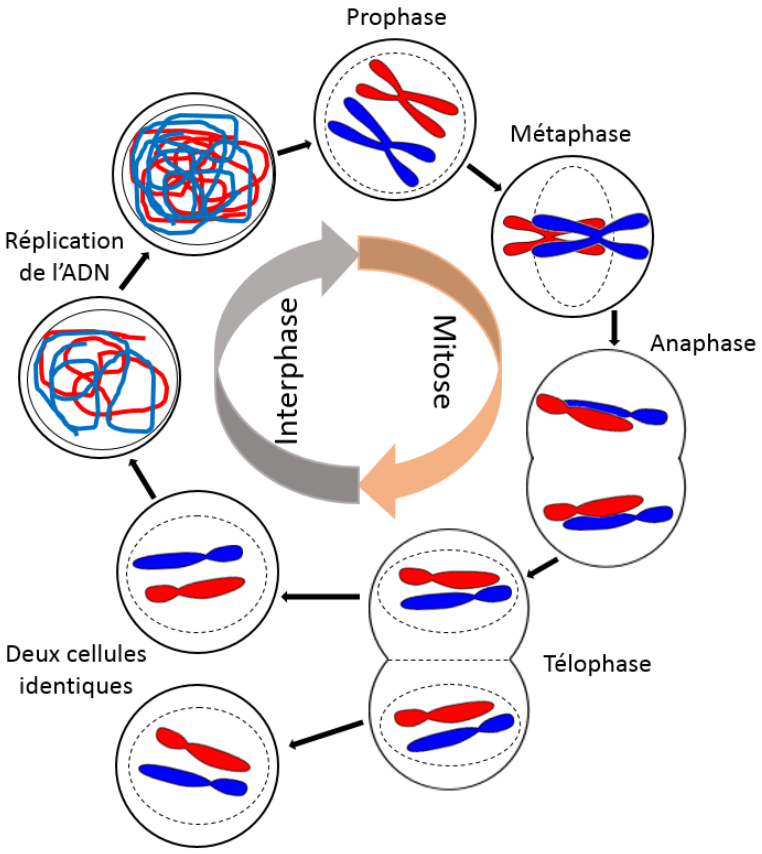


FIGURE 1.2 – Cycle d'une cellule eucaryote.

1.1.2 À quoi sert le génome ?

Le génome code les *protéines* et permet donc le fonctionnement des êtres vivants. En fait, tout le génome ne donne pas lieu à la production de protéines. Seules des portions du génome appelées *gènes* sont « lues » afin de fabriquer entre autres les protéines nécessaires à la vie. Un gène est généralement composé d'un promoteur permettant le début de lecture de celui-ci, d'introns, d'exons et d'un codon stop indiquant la fin du gène (figure 1.3).

La synthèse ou création des protéines se déroule en trois étapes (figure 1.3) :

- ▷ la transcription ; duplication de l'ADN (deux brins) en Acide Ribonucléique ou ARN messager (1 brin),
- ▷ la maturation de l'ARN messager appelée aussi épissage ; suppression des introns de l'ARN,
- ▷ la traduction ; fabrication de la protéine à partir de l'ARN mature par les ribosomes. Cette dernière étape se déroule en dehors du noyau de la cellule.

D'autres portions du génome, ne conduisant pas à la fabrication de protéines, sont aussi transcrites. Différents types d'ARN sont ainsi produits. Nous ne rentrerons cependant pas dans le détail des ARNs possibles dans ce manuscrit.

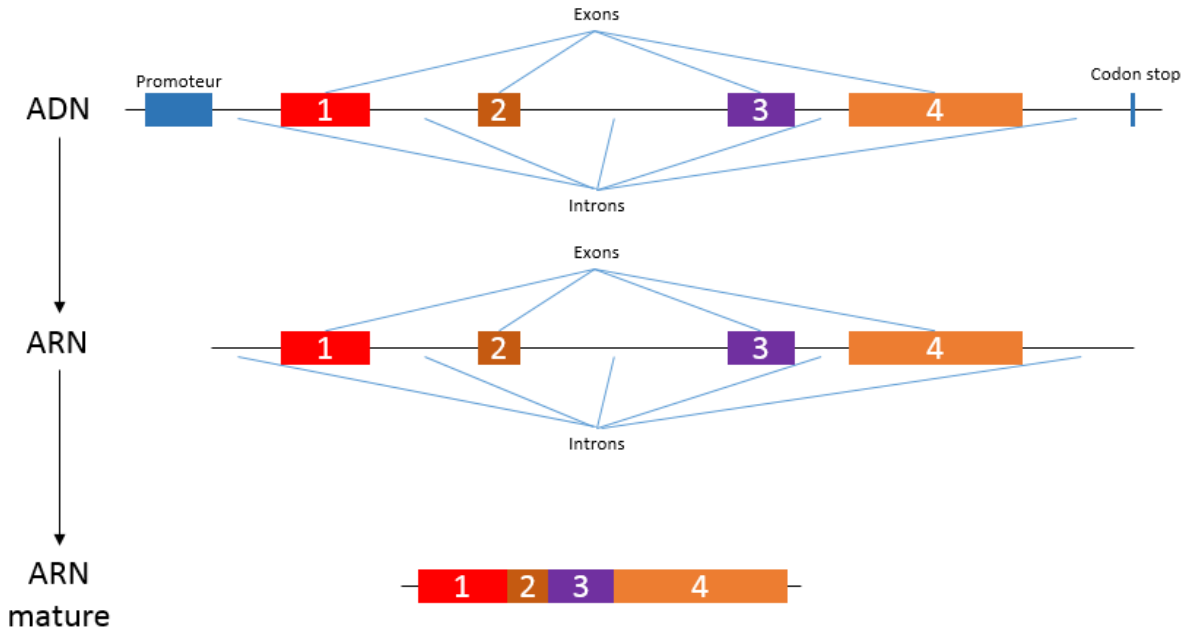


FIGURE 1.3 – Structure d'un gène et transformation en ARN mature.

1.1.3 Le génome humain

Le génome humain contient plus de 3 milliards de paires de bases réparties en 22 paires de chromosomes homologues (ayant les mêmes gènes) ou autosomes, une paire de chromosomes sexuels (XX ou XY) déterminant le sexe de l'individu et un ADN mitochondrial. Le caryotype (arrangement normal des chromosomes dans un type de cellule donné) humain est donné dans la figure 1.4. Les chercheurs estiment que le génome humain contient environ 20 000 gènes de taille

très variable et codant des protéines. La partie codante des gènes (exons) représente seulement environ 1.5% du génome entier [1]. Le reste du génome est composé entre autres des introns des gènes (supprimés lors de la maturation de l'ARN), d'ADN transcrit mais non traduit et de séquences génétiques répétées. Le rôle de beaucoup de ces portions du génome reste flou mais des avancées se font dans ce domaine [2].

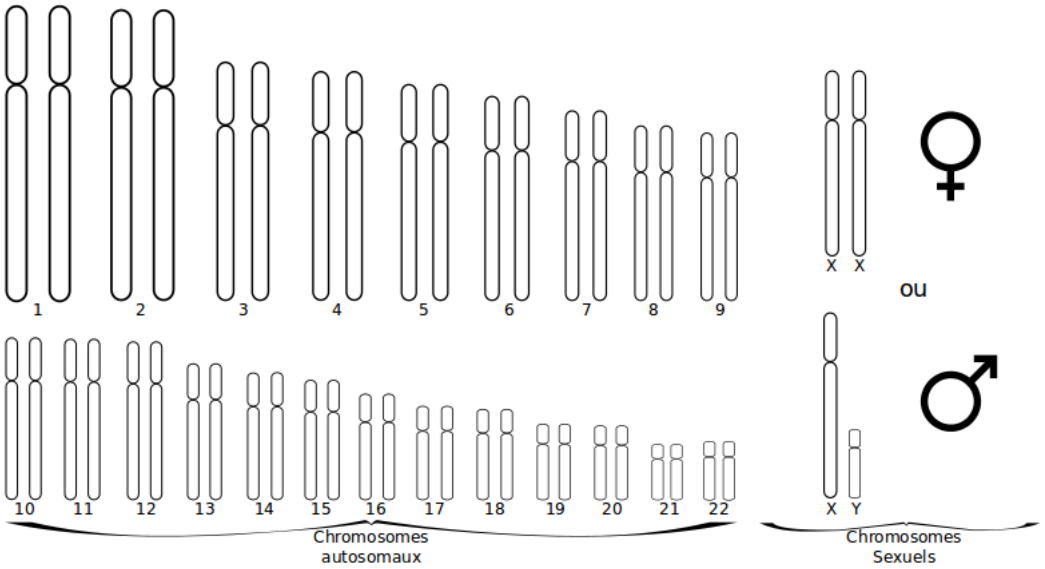


FIGURE 1.4 – Caryotype chez l'homme. La taille de chaque chromosome est proportionnelle à sa longueur réelle.

1.1.4 La variabilité du génome

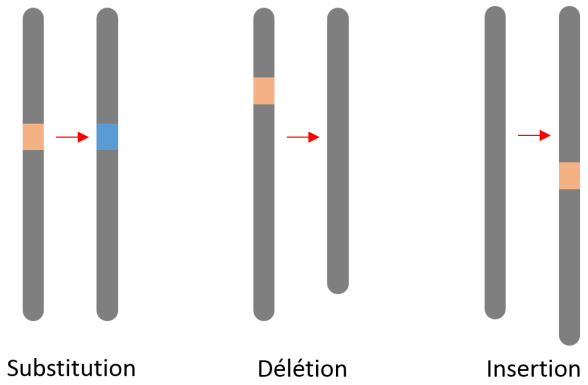
Le génome est très variable dans le monde du vivant et, ce, à plusieurs niveaux. La forme du génome ou caryotype n'est, en premier lieu, pas la même selon les espèces. Par exemple, la souris grise possède 40 chromosomes alors que l'homme en possède 46. Puis, la variabilité génétique existe aussi entre les individus d'une même espèce. En effet, à part les jumeaux monozygotes, c'est-à-dire des jumeaux ayant reçu le même patrimoine génétique, deux individus n'ont pas le même génome. La diversité des génomes présents pour une seule espèce est appelée la diversité génétique. Cette variabilité inter-individu a pour origine des modifications de l'ADN ou *mutations*. Ces modifications sont normalement « réparées » par des systèmes biologiques complexes. Cependant, il arrive que la cellule portant la mutation survive et transmette son génome aux générations suivantes créant ainsi un polymorphisme génétique. La mutation peut ensuite se diffuser dans la population.

Plusieurs types de mutations existent. Pour commencer, une mutation génétique peut avoir lieu sur :

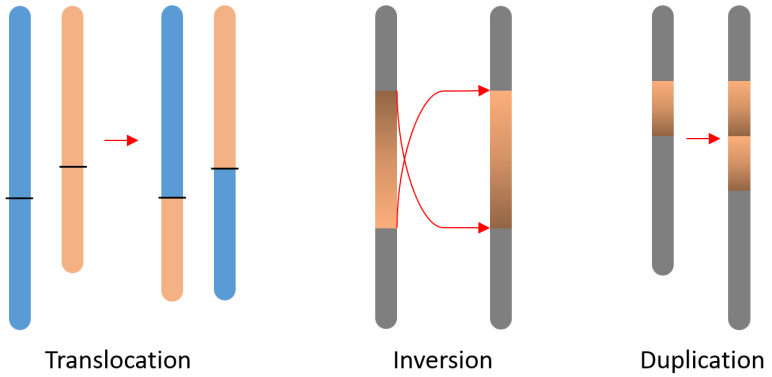
- ▷ une cellule somatique c'est-à-dire qui n'intervient pas dans la reproduction. La mutation ne pourra donc affecter que son hôte.
- ▷ une cellule germinale qui permet la formation des gamètes afin de servir à la reproduction. Dans ce cas, la mutation peut-être transmise à la descendance.

En plus de différer sur la localisation, une mutation génétique peut avoir plusieurs formes :

▷ les mutations d'une ou quelques paires de bases (substitutions, délétions ou insertions),



▷ les mutations chromosomiques qui concernent une portion importante du génome. Dans ce type de mutations, nous retrouvons les délétions et insertions évoquées précédemment pour une portion de génome beaucoup plus importante mais aussi les translocations, inversions et duplications.



En fonction de l'emplacement de la mutation génétique sur le génome, celle-ci peut n'avoir aucune conséquence ou, au contraire, avoir des répercutions très variables sur l'organisme d'un individu. Les mutations participent à la variabilité de la séquence génétique. Cette variation est alors appelée un *polymorphisme génétique*. Elle se traduit par la présence de plusieurs versions dans la population d'un même locus (emplacement sur le génome), appelées *allèles*.

Une autre source de variabilité du génome est la *recombinaison* qui a lieu durant la méiose ou la division d'une cellule souche en gamètes (figure 1.5). La méiose se décompose en deux divisions cellulaires. Avant la première division de la cellule, les chromosomes se rassemblent par paire puis se placent de chaque coté d'un plan. Durant cette étape, les bras des deux chromosomes peuvent s'entremêler. Dans ce cas, après la première division, chaque cellule fille contient un chromosome recombiné à deux chromatides. Ainsi, après la seconde division, les cellules résultantes sont des gamètes ayant des chromosomes existants déjà chez l'individu (gamètes 1 et 4 sur la figure 1.5) et d'autres ayant un « mélange » des deux chromosomes homologues de base (gamètes 2 et 3 sur la figure 1.5).

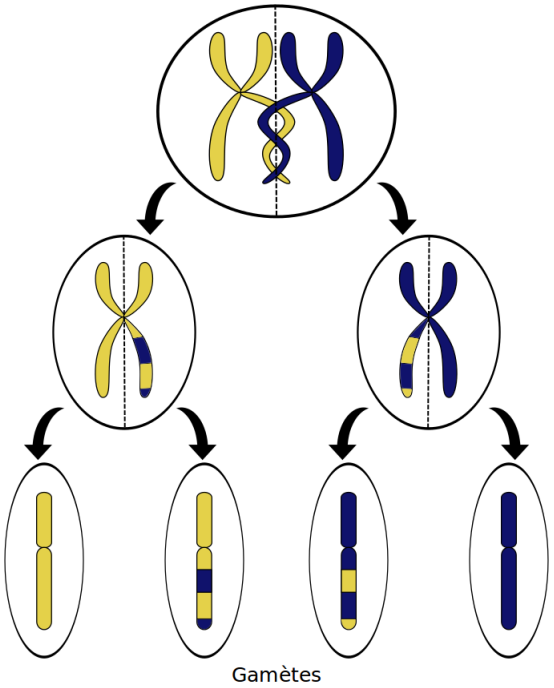


FIGURE 1.5 – Méiose d'une cellule.

Génotypes et haplotypes

Nous avons vu que l'information contenue dans le génome est représentée par les bases azotées. À partir des bases azotés d'un certain nombre de locus, nous pouvons distinguer deux types de données génétiques :

- ▷ les *génotypes* ; allèles portés par un même individu à un locus donné,
- ▷ les *haplotypes* ; un groupe d'allèles de différents locus situés sur un même chromosome.

Par exemple, sur la figure 1.6, nous pouvons lire sur le brin de référence les génotypes GA pour le locus A et TC pour le locus B et les haplotypes GT sur le premier chromosome et AC sur le second. L'information haplotypique contient, en plus de l'information génotypique, la *phase* qui indique quels allèles sont sur le même chromosome.

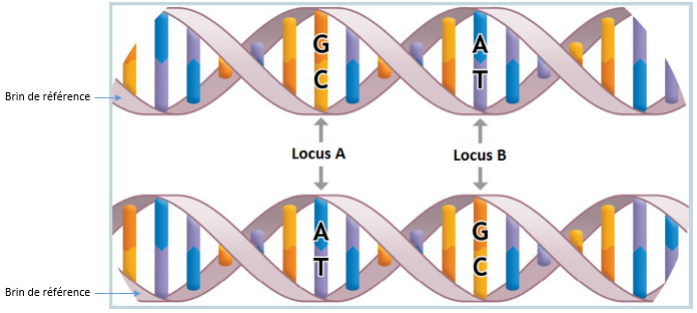


FIGURE 1.6 – Bases azotés de deux locus A et B sur les deux chromosomes.

Single Nucleotide Polymorphism (SNP)

L'un des polymorphismes génétiques possibles est le « *Single Nucleotide Polymorphism* » ou SNP. Les SNPs sont des polymorphismes d'une seule paire de bases. Par exemple, sur la figure 1.7, nous avons un SNP G/A : certains chromosomes porteront l'allèle A et d'autres l'allèle G sur le brin de référence. Si ce SNP est situé sur un chromosome autosomal (non sexuel), trois génotypes sont alors possibles : AA, AG et GG. Les SNPs représentent la majorité des polymorphismes dans le génome humain. Tout au long de ce travail, nous ne considérerons que des SNPs di-alléliques c'est-à-dire qui ont deux versions ou allèles possibles.

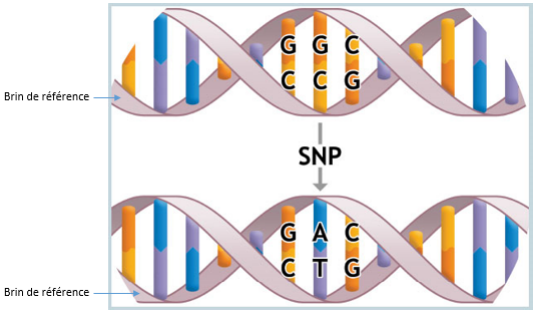


FIGURE 1.7 – Single Nucleotide Polymorphism.

Les SNPs constituent notamment de bons *marqueurs génétiques* (polymorphismes d'emplacement connu dans le génome) pour les études d'association (voir section 1.2.3). Ils nous permettent de détecter des facteurs de risques génétiques c'est-à-dire des facteurs génétiques qui favorisent l'apparition de la maladie ou affectent un trait donné.

1.1.5 Les proportions d'Hardy-Weinberg

Le principe d'Hardy-Weinberg, développé au début des années 1900, est utilisé en génétique des populations. Il s'applique à une population vérifiant certaines hypothèses :

- ▷ La population est infinie
- ▷ La composition des gamètes (cellules reproductrices) reflète fidèlement la composition allélique des individus de la population :
  - – Pas de fertilité différentielle
  - – Pas de mutations de novo (mutations apparaissant sur un gamète ou un œuf fécondé)
  - – Pas de distorsion de ségrégation méiotique c'est-à-dire pas de ségrégation durant la formation des gamètes
- ▷ Tirage au hasard des gamètes :
  - – Panmixie : formation des couples au hasard
  - – Pangamie : lors de la fécondation, les gamètes s'unissent au hasard ; pas de sélection gamétique
- ▷ La composition de la population ne change pas entre la fécondation et la reproduction :
  - – Pas de sélection

– Pas de migration

Sous ces hypothèses, si nous avons n haplotypes possibles $z_1, \ldots, z_n$ de fréquence respective $f_1, \ldots, f_n$, alors la fréquence des individus ayant les haplotypes z_i et z_j est fixe dans le temps et vaut :

$$f_{i,j} = \begin{cases} f_i^2 & \text{si } i = j \\ 2f_i f_j & \text{sinon} \end{cases}$$

Ce sont les *proportions d'Hardy-Weinberg*. Généralement, toutes les hypothèses ne sont pas vérifiées. En effet, une population vérifiant toutes les hypothèses est « idéale ». Cependant, dans beaucoup de cas et à des échelles de temps suffisamment faibles, les proportions d'Hardy-Weinberg sont une description satisfaisante de la réalité. Il est alors possible de tester si ces proportions sont respectées pour un locus donné. Le plus souvent, les locus qui s'écartent trop des proportions d'Hardy-Weinberg sont exclus des analyses car cela peut indiquer la présence d'erreurs de génotypage.

1.1.6 Le déséquilibre de liaison

Nous considérons le problème posé par le *déséquilibre de liaison* (DL) entre deux SNPs A et B proches sur le génome (figure 1.6). Nous disons que deux locus di-alléliques A/a et B/b sont en équilibre s'ils vérifient :

$$\begin{aligned} f_{AB} &= f_A f_B \\ f_{Ab} &= f_A f_b \\ f_{aB} &= f_a f_B \\ f_{ab} &= f_a f_b \end{aligned}$$

où f_x est la fréquence en population générale de l'allèle ou de l'haplotype x.

En présence de déséquilibre, ces égalités ne sont pas vérifiées. Certaines combinaisons des allèles des deux locus ou haplotypes sont préférentiellement présentes dans la population. Par exemple, si nous avons :

Allèle/Haplotype	a	b	AB	Ab	aB	ab
Fréquence	0.1	0.2	0.8	0.1	0	0.1

Nous constatons que :

$$\begin{aligned} f_{AB} &= 0.8 > f_A f_B = 0.72 \\ f_{Ab} &= 0.1 < f_A f_b = 0.18 \\ f_{aB} &= 0 < f_A f_B = 0.08 \\ f_{ab} &= 0.1 > f_a f_b = 0.02 \end{aligned}$$

Les haplotypes AB et ab sont donc plus fréquents que s'il y avait équilibre.

Le déséquilibre de liaison s'exprime à l'aide de deux mesures usuelles :

▷ $D = f_{AB} - f_A f_B = f_{ab} - f_a f_b = -(f_{aB} - f_a f_B) = -(f_{Ab} - f_A f_b)$,

▷ $r = \dfrac{D}{\sqrt{f_a f_A f_b f_B}} \in [-1, 1]$.

Le déséquilibre de liaison est une forme de corrélation entre deux locus, plus le déséquilibre est important, plus les deux locus dépendent l'un de l'autre.

Le déséquilibre de liaison est également lié à la recombinaison, plus la recombinaison est probable entre deux locus du génome, plus le déséquilibre tendra à être faible, sauf notamment en présence d'un mélange de populations. Les recombinaisons génétiques permettent de remanier les haplotypes et donc, avec du temps, de diminuer le déséquilibre de liaison.

1.2 Les études en génétique

1.2.1 Un peu d'histoire ...

Bien avant la découverte de la génétique à proprement parlé, en 1865, Gregor Mendel découvrit, à partir d'expériences, les lois de transmission de certains caractères chez les végétaux [3]. Ces lois sont des conséquences directes de l'information génétique. Il commença alors à formaliser ses observations, c'est ce que nous appelons la transmission mendélienne. Ses résultats passèrent presque inaperçus durant les 30 années suivantes puis furent découvert de nouveau en 1900 par Hugo de Vries, Carl Correns et Erich von Tschermak-Seysenegg. Entre temps, plusieurs avancées biologiques avaient été faites telles que la découverte de la substance de l'ADN en 1869 par Friederich Miescher ou la description de la division cellulaire par Walther Flemming. Au début du 20e siècle, la théorie selon laquelle les chromosomes seraient le support de l'hérédité [4] fut développée par Walter Sutton et Theodor Boveri puis confirmée par Thomas Morgan qui montra, dans les années 1910, que les chromosomes portaient les gènes eux mêmes porteurs de l'information génétique. Ce même scientifique développa également la théorie de la liaison génétique à l'aide des connaissances sur les recombinaisons chromosomiques découvertes par Frans Alfons Janssens [5] quelques années plus tôt. Ces avancées majeures dans la compréhension des mécanismes de la génétique permirent l'élaboration de la première carte génétique en 1913 [6].

Les décennies suivantes virent beaucoup d'avancées scientifiques dans le domaine de la génétique jusqu'en 1944 lorsque Oswald Avery, Colin MacLeod, et Maclyn McCarty démontrèrent que l'ADN portait l'information génétique [7]. En 1953, la structure en double hélice de l'ADN fut découverte par James Watson et Francis Crick [8] permettant ainsi la compréhension du processus d'hybridation ou le fait que deux brins complémentaires auront toujours tendance à se lier. Une autre découverte importante fut celle du mécanisme de réplication de l'ADN lors de la division cellulaire dans les années 1950 par Matthew Meselson and Franklin Stahl [9]. La compréhension de ce mécanisme permit, en 1983, à Kary Banks Mullis d'amplifier pour la première fois artificiellement la quantité d'ADN [10], c'est ce que nous appelons aujourd'hui la technique de la « Polymerase Chain Reaction » (PCR). Nous pouvons évoquer une dernière découverte majeure dans l'avancée des technologies en génétique, celle du découpage de l'ADN par une enzyme dans les années 1970 par Hamilton Smith [11] permettant la mise au point des techniques de fragmentation de l'ADN.

Les dernières découvertes évoquées sont notamment utilisées par les puces à ADN modernes apparues dans les années 1990 [12]. Elles contiennent un brin d'ADN pour chaque locus étudié.

Les principales étapes de leur utilisation sont :

- ▷ l'amplification de l'ADN afin d'augmenter sa quantité,
- ▷ la fragmentation de l'ADN avant de le placer sur la puce,
- ▷ l'hybridation de certains fragments aux séquences complémentaires contenues dans la puce,
- ▷ l'apparition de fluorescences pour les fragments bien hybridés,
- ▷ l'étude des fluorescences à l'aide d'outils informatiques pour en déduire le génotype.

Les puces à ADN fournissent donc le génotype de locus choisis au préalable uniquement. Un autre type de puces très utilisé est la puce d'expression. Elle fonctionne de la même façon, avec non plus des locus de l'ADN mais des ARNs. Les données de fluorescence représentent ainsi l'expression d'un gène donné chez un individu. Ces puces peuvent également servir à détecter la présence de fragments d'ARN qui seront ou non traduits en protéine.

Une autre technique de lecture de l'ADN est le séquençage qui fut développé pour la première fois en 1977 par Fred Sanger [13] et Walter Gilbert, et Allan Maxam [14] de façon indépendante. Le séquençage permet, contrairement aux puces à ADN, d'avoir la séquence complète de l'ADN d'un individu. Le qualité, le coût et le débit de ce processus ont été améliorés par une méthode appelée « Next Generation Sequencing » [15]. Cette avancée technologique dans le domaine de la génétique participe au problème posé par les « Big Data ». En effet, la production de données étant moins coûteuse, l'un des principaux enjeux aujourd'hui est le stockage et le traitement de ce flux de données très important, c'est ce que nous appelons les « Big Data ».

1.2.2 Les études de liaison

Développée au milieu du 20e siècle, le but de l'étude de liaison [16] est de localiser sur le génome les facteurs génétiques d'un trait ou d'une maladie. Pour cela, elles utilisent des données familiales et combinent l'information phénotypique et l'information des transmissions des allèles au locus d'intérêt. Le but est alors de tester la présence de liaison ou de corrélation entre le phénotype et le marqueur étudié ce qui indiquerait une co-transmission du locus d'intérêt avec le trait. Les études de liaison permettent notamment de détecter des régions candidates qui peuvent contenir un locus causal pour le trait étudié.

1.2.3 Les études d'association

L'étude des gènes impliqués dans une maladie ou un trait quantitatif passe par différentes étapes dont l'étude d'association. Celle-ci consiste à repérer des signaux d'association entre des marqueurs génétiques et une maladie ou un trait quantitatif. Selon le but recherché, plusieurs stratégies peuvent être appliquées afin de détecter l'association entre des variants génétiques et un trait. La première stratégie donnant lieu encore aujourd'hui au plus grand nombre de publications est l'étude d'association sur les données « *genome-wide* », en français génome entier, ou « *Genome-Wide Association Studies* » (GWAS) [17–19]. Cette méthode consiste à tester l'association de chaque SNP de façon indépendante puis de corriger le test utilisé afin d'éviter

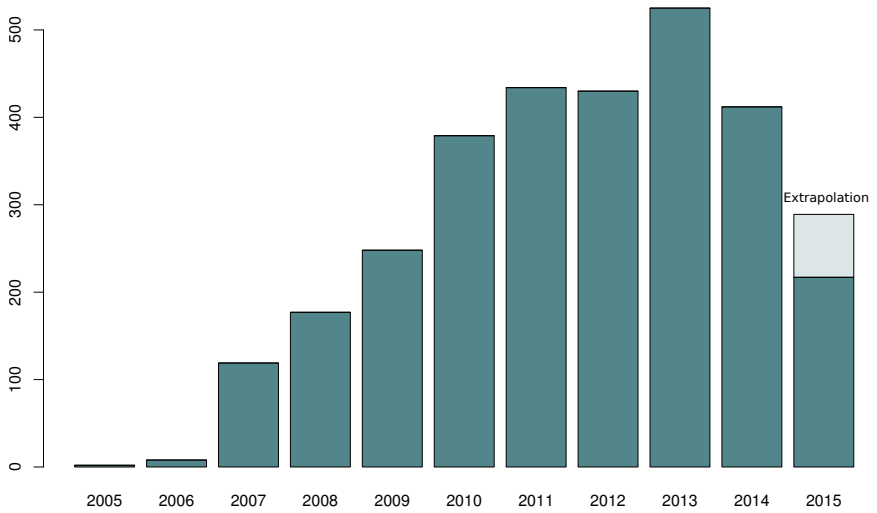


FIGURE 1.8 – Nombre de GWAS publiées entre 2005 et août 2015 d'après le "GWAS catalog" (`https://www.ebi.ac.uk/gwas/`). Pour la dernière année, une extrapolation a été appliquée.

un trop grand nombre de faux positifs. Dans le but d'améliorer la capacité des GWAS à trouver des signaux d'association avec un trait en limitant le nombre de tests, seuls les tag-SNPs (des SNPs en faible déséquilibre de liaison choisis pour résumer au mieux le génome) sont considérés. Ce type d'études a connu un essor très important depuis 2005 donnant lieu à un nombre très important de publications (figure 1.8) et la découverte de beaucoup de tag-SNPs associées aux différentes maladies complexes c'est-à-dire les maladies ayant des facteurs environnementaux et génétiques avec des effets plus ou moins faibles (figure 1.9). Cette tendance s'essouffle un peu aujourd'hui mais reste une part importante des études génétiques.

FIGURE 1.9 – Tag-SNPs trouvés comme étant associés à une ou plusieurs maladie(s) complexe(s) d'après le "GWAS catalog" (`https://www.ebi.ac.uk/gwas/`).

D'autres stratégies pour les études d'association existent. Parmi elles, nous trouvons l'étude de l'association avec non plus un variant mais une région génétique ou un ensemble de régions génétiques reliées à la même fonction (*pathway*) telle que la méthode « SKAT » [20] que nous

verrons plus tard dans ce manuscrit. Nous pouvons aussi trouver des méthodes adaptées aux variants rares [21,22] telles que les « burden tests » [23–25] ou une variante de « SKAT » [26,27].

Les études d'association ne sont pas une fin en soi. En effet, l'étude d'association n'indique pas une relation de cause à effet entre le locus et le trait d'intérêt. Elle peut être due à un autre locus en déséquilibre de liaison (corrélé) avec le locus associé ou à un autre facteur inconnu qui influence le locus et le trait (figure 1.10). Les études d'association permettent donc uniquement de donner des pistes pour des études fonctionnelles postérieures.

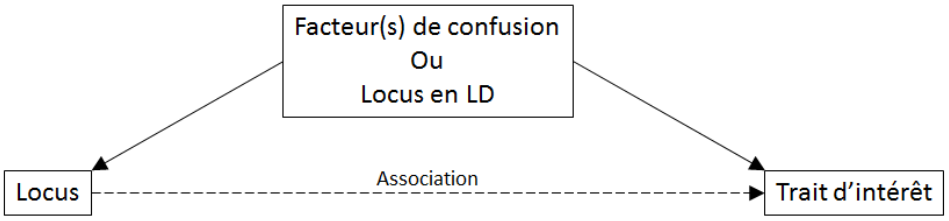


FIGURE 1.10 – Association entre les tag-SNPs et le trait d'intérêt.

1.2.4 La structure des données génétiques

Les données familiales

Les données familiales sont extraites de pedigrees ou de familles qui peuvent contenir une ou plusieurs générations. Par exemple, la figure 1.11 représente un pedigree. Dans cette figure, les ronds représentent les femmes et les carrés les hommes. Les générations vont du haut vers le bas. Les liens directs indiquent une union donnant lieu à une descendance. Dans notre exemple, nous avons également un lien double qui indique une union consanguine (entre deux individus apparentés). Pour finir, les individus barrés sont les membres décédés de la famille. L'information sur les liens familiaux de différents individus peut être résumée dans la *matrice de kinship* qui donne le coefficient d'apparentement entre tous les individus. Le coefficient d'apparentement entre deux individus est la probabilité que deux allèles d'un même gène tirés au hasard chez chacun des individus soient identiques par descendance (figure 1.12) et peut être calculé à partir de la généalogie. Quelques unes de ses valeurs sont données dans la table 1.1. La matrice de *kinship* peut aussi être estimée à partir des données de génome. En effet, la matrice

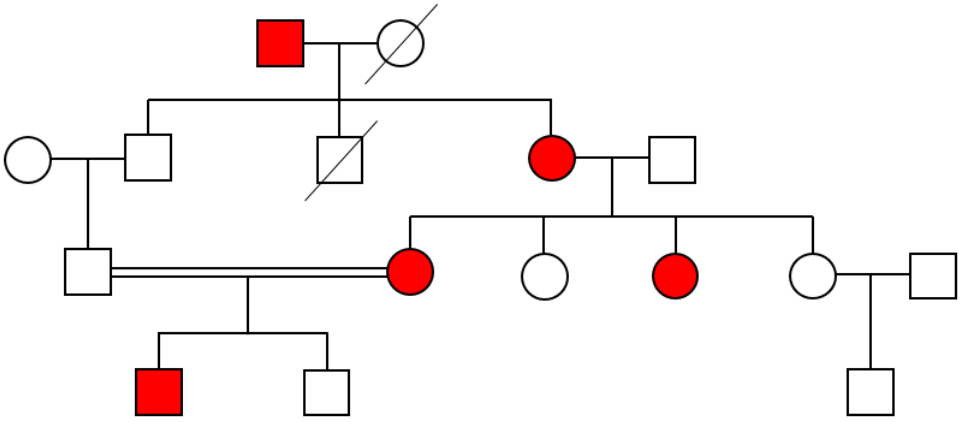


FIGURE 1.11 – Exemple de pedigree.

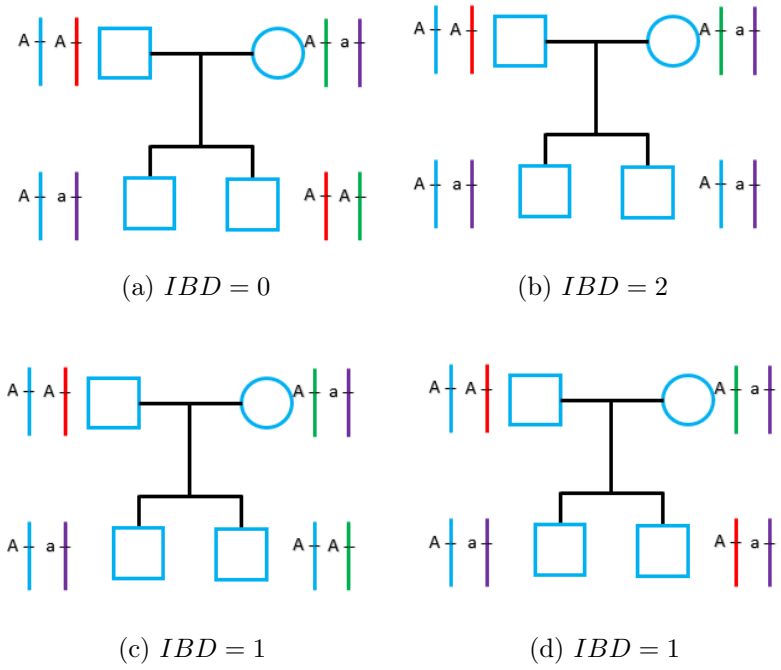
de corrélation calculée sur les génotypes standardisés (encadré 1.18) est une estimation de deux fois la matrice de *kinship*. Dans ce manuscrit, nous allons rencontrer trois types particuliers de données familiales : les paires de germains atteints, les trios et les familles nucléaires multiplex.

Lien familial	Φ
Identité	0.5
Parent/enfant	0.25
Germains	0.25
Demi-germains	0.125

TABLE 1.1 – Valeurs des coefficients d'apparentement pour quelques liens familiaux.

Les paires de germains atteints

Les données de paires de germains atteints reposent sur une fratrie avec au moins deux enfants ayant la maladie d'intérêt. Nous nous intéressons au cas où les données sont composées du génotype de l'un des germains appelé le cas index et de son état IBD (Identical By Descent) avec le second germain. L'état IBD est le nombre d'allèles (entre 0 et 2) que les deux germains ont reçu d'un même ancêtre. Attention, IBD est différent d'IBS (Identical By State) qui signifie que deux allèles sont identiques mais n'ont pas obligatoirement la même origine. Afin de mieux comprendre, nous pouvons regarder l'exemple de la figure 1.12 pour un locus di-allélique A/a avec un père homozygote AA, une mère hétérozygote Aa et un premier enfant atteint homozygote AA. Nous regardons alors toutes les possibilités pour le deuxième enfant. Dans le premier cas (figure 1.12a), le deuxième germain reçoit les chromosomes rouge et vert par le père et la mère respectivement qui sont les deux chromosomes non transmis au premier enfant ; l'état IBD est donc de 0. Les figures 1.12b et 1.12c montrent les deux possibilités pour obtenir un état IBD de 1 entre les deux germains. Nous pouvons remarquer que l'état IBS est



(a) $IBD = 0$ (b) $IBD = 2$

(c) $IBD = 1$ (d) $IBD = 1$

FIGURE 1.12 – État IBD.

différent pour ces deux possibilités ; il vaut 2 ou 1 respectivement. Le dernier cas (figure 1.12d) est celui où l'état IBD est de 2 ; les deux germains ont reçu les chromosomes bleu et violet du père et de la mère respectivement.

Les trios

Les données de trio sont issues de familles constituées des deux parents et d'un enfant atteint de la maladie étudiée (figure 1.13). Elles comportent les génotypes de ces trois membres. Des méthodes ont été développées exclusivement pour ce type de données. Les familles trio peuvent aussi permettre de créer des pseudo-témoins. C'est-à-dire que nous prenons comme témoin un individu fictif avec le génotype composé des allèles non-transmis par les parents à l'enfant malade. Par exemple, si nous regardons la famille de la figure 1.13 pour un locus di-allélique A/a, l'enfant malade a reçu le chromosome bleu avec l'allèle A de son père et le chromosome violet avec l'allèle a de sa mère donc le pseudo-témoin résultant aura les chromosomes rouge et vert et ainsi le génotype AA. Il a été montré que, si les trios sont recrutés sous la seule condition que l'enfant est atteint, les génotypes de ces pseudo-témoins sont représentatifs de la population [28].

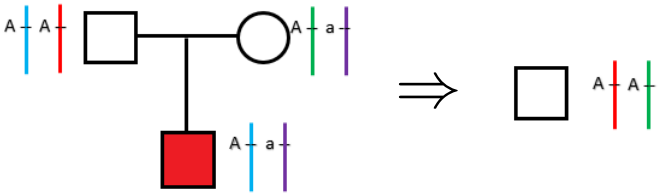


FIGURE 1.13 – Famille trio et pseudo témoin.

Les familles nucléaires multiplex

Les familles nucléaires multiplex sont des familles composées de deux générations avec aucune exigence particulière. Nous observons donc les parents et les enfants. Par exemple, la figure 1.14 représente une famille avec quatre enfants dont 3 garçons. Dans cette famille, le père et deux des fils sont atteints de la maladie étudiée et la mère est décédée. Ce type de données est surtout très utile aux études de liaison, mais elles peuvent également servir, par exemple, pour l'étude d'association ou la recherche d'empreinte parentale (section 1.2.6).

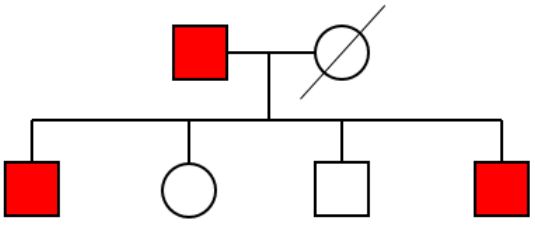


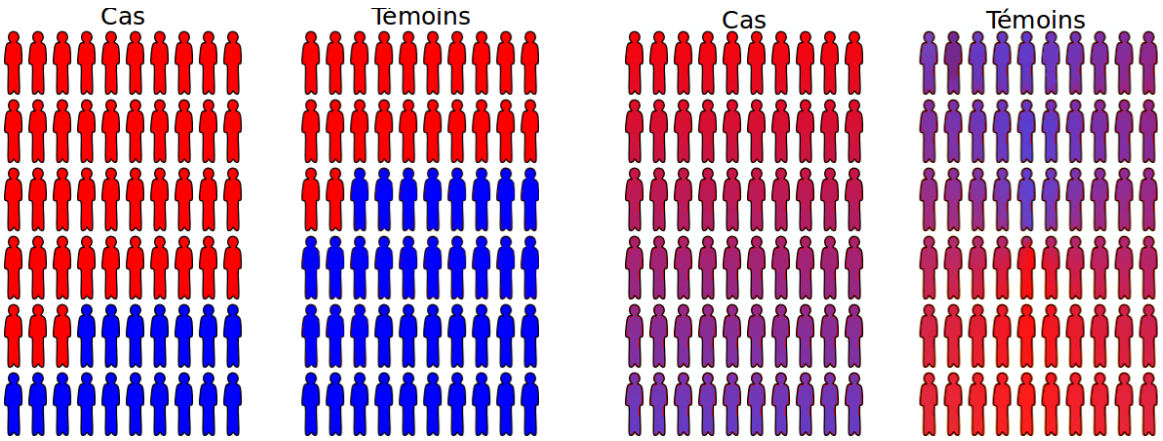
FIGURE 1.14 – Famille nucléaire.

Les données en population

Les données en population sont constituées des génotypes d'individus non-apparentés recrutés dans une ou plusieurs populations. Il n'est donc pas nécessaire dans ce type d'étude de prendre en compte la structure familiale ce qui rend leur analyse plus simple. L'avantage des analyses en population réside surtout dans la « facilité » du recrutement qui est beaucoup moins compliqué que celui de familles. Cependant, les données en population posent plusieurs problèmes :

- ▷ le manque de puissance dans certains cas comme l'étude des variants rares (il faut un effectif très important afin de détecter des signaux d'association),

- ▷ la *stratification de population* ou la présence de compositions génétiques différentes dans chaque population au travers des fréquences et du déséquilibre de liaison. Elle peut être discrète (figure 1.15a) ou continue (figure 1.15b).



(a) Stratificaton discrête. (b) Stratificaton continue.

FIGURE 1.15 – Représentations de la stratification de population.

Ainsi, en présence de stratification de population, si la distribution du trait étudié dépend de la composante génétique mais également de la population, une association fictive entre le trait et le marqueur d'intérêt peut apparaître. Dans ce cas, la stratification de population agit comme un facteur de confusion (figure 1.16). Afin d'éviter la surestimation de l'association entre la composante génétique et le trait d'intérêt, la stratification de population doit être prise en compte dans l'analyse. En pratique, cette information sur la stratification de population n'est pas disponible, il est donc nécessaire d'utiliser des moyens détournés. Les solutions les plus usitées sont l'inclusion dans l'analyse d'un certain nombre de composantes principales ou PCs (encadré 1.17) calculées sur les données génétiques ou d'information sur les origines des individus si celles-ci sont disponibles dans l'étude. Nous pouvons noter ici qu'il y a deux possibilités pour obtenir les PCs (encadré 1.17), avec la SVD de la matrice des génotypes standardisés ou la décomposition en éléments propres de la matrice de corrélation génétique ou « *Genetic Relationship Matrix* » (GRM) obtenue à partir des génotypes standardisés (encadré 1.18).

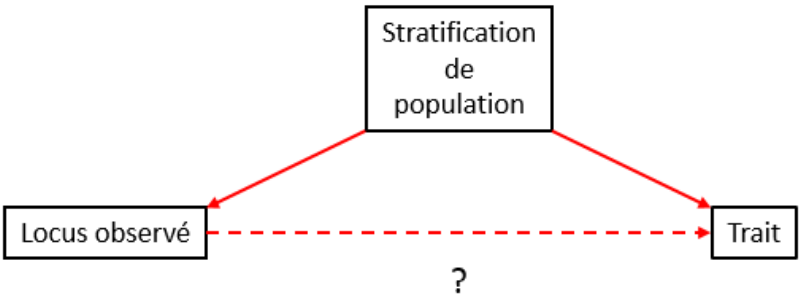


FIGURE 1.16 – La stratification de population comme facteur de confusion.

Décomposition en valeurs singulières ou « *Singular Value Decomposition* » (SVD)

La SVD permet de décomposer une matrice X de taille $(n \times m)$. Elle s'écrit

$$X = U\Sigma V'$$

avec U et V des matrices de vecteurs orthogonaux et de norme 1 de taille $(n \times n)$ et $(m \times m)$ respectivement et Σ une matrice de taille $(n \times m)$ avec les valeurs singulières σ_i sur la diagonale et des 0 partout ailleurs.
Dans ce manuscrit, nous utiliserons une version tronquée de la SVD. En effet, pour les données génétiques, $n \ll m$, nous ne gardons donc que les colonnes de la matrice V associée au bloc diagonale de Σ.

$$X = U\Sigma V_t'$$

avec U et V_t une matrice de vecteurs indépendants et de norme 1 de taille $(n \times n)$ et $(m \times n)$ respectivement et Σ la matrice diagonale des n valeurs singulières σ_i. La matrice V_t s'appelle aussi la matrice des « *loadings* », ses colonnes contiennent les contributions des variables à la définition des colonnes de U.

La SVD peut être reliée à la décomposition en éléments propres ou « *Eigen Decomposition* » qui permet de décomposer une matrice carrée en une matrice de vecteurs orthogonaux et de norme 1 et une autre diagonale,

$$XX' = U\Sigma^2 U'$$

avec U et Σ les mêmes matrices que celles définies pour la SVD.

Dans ce cas, la matrice des composantes principales ou PCs est

$$PC = U\Sigma.$$

Les PCs ont la particularité d'être orthogonales et d'expliquer une portion de l'information contenue dans X proportionnelle à la valeur singulière σ_i. Ainsi, inclure les premières PCs ou celles ayant les valeurs singulières les plus importantes permet de capturer le maximum d'informations avec un minimum de variables.

FIGURE 1.17 – Décomposition en valeurs singulières.

Matrice de corrélation génétique ou « *Genetic Relationship Matrix* » (GRM)

La matrice de corrélation génétique (GRM) de taille $(n \times n)$ est la matrice de variance-covariance de la composante génétique de n individus. Elle est calculée à partir de la matrice des génotypes centrés et réduits notée Z de taille $(n \times q)$ avec q le nombre de SNPs considérés. Si nous notons $G_{ij} = 0, 1$ ou 2 le génotype de l'individu i au SNP j codé comme le nombre d'allèles alternatifs, le même génotype standardisé vaut

$$Z_{ij} = \frac{G_{ij} - \mu_j}{\sigma_j}$$

où $\mu_j = 2p_j$ est l'espérance des génotypes au SNP j avec p_j la fréquence de l'allèle alternatif et σ_j leur variance. La variance des génotypes peut être estimée par la variance empirique ou à partir de la fréquence p_j sous l'hypothèse de l'équilibre d'Hardy-Weinberg par $\sqrt{2p_j(1-p_j)}$.
Alors, avec ces notations, la matrice de corrélation génétique est la matrice de variance-covariance empirique des génotypes centrés et réduits des individus

$$GRM = \frac{1}{q-1} ZZ'.$$

FIGURE 1.18 – Matrice de corrélation génétique.

1.2.5 L'héritabilité

Dans le cas d'un trait quantitatif, l'héritabilité, notée H^2, est définie comme la proportion de variance expliquée par les facteurs génétiques. Ce concept est apparu pour la première fois dans les travaux de recherche de Galton au cours du 19e siècle [29] mais la définition moderne de l'héritabilité a été développée par Fisher en 1918 [30]. Des travaux plus récents se concentrent sur un type particulier de facteurs génétiques ; les facteurs avec des effets additifs. Dans ce cas, nous parlons de l'*héritabilité restreinte*, notée h^2, qui est la proportion de variance d'un trait expliquée uniquement par les effets génétiques additifs.

Dans le cas d'un trait binaire, la définition de l'héritabilité est moins évidente. Il est possible de la définir en supposant l'existence d'un trait quantitatif sous-jacent appelé « liabilité » dont la valeur détermine le trait binaire en question. L'héritabilité du trait binaire est alors définie comme la proportion de variance de la liabilité expliquée par les effets génétiques.

L'estimation de l'héritabilité a longtemps été faite sur des données familiales [31]. Par exemple, beaucoup d'études de jumeaux ont été conduites car elles permettent de prendre facilement en compte l'environnement partagé dans une famille. Cependant, il reste des biais induits par la différence dans l'environnement partagé par des jumeaux monozygotes ou dizygotes [32–34].

Nous avons précédemment parlé des études d'association telles que les GWAS qui ont permis de trouver des centaines de SNPs associés à des traits complexes [17, 18] (section 1.2.3). Cependant, il a été montré que pour beaucoup de traits, ces SNPs n'expliquent pas toute la variance génétique estimée antérieurement. La proportion de variance génétique qui reste inexpliquée par ces SNPs est appelée l'héritabilité manquante (« *missing heritability* ») [35–41]. Le problème de l'héritabilité manquante a alors déclenché le développement de nouvelles méthodes utilisant les données du génome entier d'individus non-apparentés issues de GWAS pour estimer

l'héritabilité de traits complexes [41,42]. Celles-ci sont maintenant plus utilisées que les méthodes basées sur les données familiales car elles sont souvent recommandées et perçues pour donner des estimations non biaisées. La principale raison est l'absence d'environnement partagé pour les individus non-apparentés [43]. Cependant, il a été montré que la présence de stratification de population provoque des biais dans l'estimation de l'héritabilité. En effet, un facteur non génétique associé au trait étudié et qui diffère au travers des différentes sous-populations peut agir comme un effet de confusion et affecter les estimations des composantes de la variance. Le plus souvent, l'héritabilité est surestimée [44]. Il est donc nécessaire de corriger la stratification de population lorsque nous utilisons ces méthodes pour l'estimation de l'héritabilité. Comme évoqué précédemment, la solution la plus utilisée pour corriger ce biais est l'inclusion dans le modèle de quelques PCs de la matrice des génotypes standardisés, le plus souvent 10 ou 20 [45–49]. La difficulté est alors de savoir le nombre de PCs qu'il est nécessaire d'inclure afin de corriger correctement la stratification de population mais sans affecter la qualité des estimations.

Le chapitre 3 de ce manuscrit explore la variabilité des estimations de l'héritabilité génétique et évalue la correction de la stratification de population avec l'inclusion de PCs dans le modèle.

1.2.6 L'empreinte parentale

Chez un individu diploïde (ayant des paires de chromosomes homologues), deux copies de chaque gène autosomal sont présentes dans son génome. Un gène soumis à empreinte est un gène dont les différentes copies présentes dans le génome d'un même individu ne s'expriment pas de la même manière. Ce processus est dû à des facteurs épigénétiques qui influencent la fabrication des protéines à partir de l'ADN. Un cas particulier de l'empreinte génétique est l'empreinte parentale. Un gène soumis à une empreinte parentale est un gène qui a une expression différente pour chacune de ces copies en fonction de l'origine de celle-ci ; le gène transmis par la mère n'aura pas la même expression que celui transmis par le père. Un tel phénomène a été observé pour la première fois chez l'insecte puis chez les mammifères dans les années 80 [50]. Un effet d'empreinte parentale a également été trouvé pour certaines pathologies humaines telles que le diabète, certains cancers et l'obésité [1].

Afin de tester la présence d'empreinte parentale pour un trait, plusieurs méthodes utilisant des familles nucléaires ont été développées, par exemple :

- ▷ celles basées sur le test du rapport de vraisemblance pour les traits binaires ; « *Maximum Likelihood Binomial Method adapted for Imprinting* » [51] (MLB-I) pour les études de liaison et « *Parent-of-Origin likelihood Ratio Test* » [52] pour les études d'association.
- ▷ celles dérivées de la méthode du « *Transmission Disequilibrium Test* » [53,54] basée sur le compte des allèles transmis par les deux parents à leurs enfants. Cette méthode, d'abord développée pour étudier l'association de traits binaires chez des trios, a été adaptée pour pouvoir tester la présence d'empreinte parentale, analyser des traits quantitatifs et traiter des familles multiplex ou des pedigrees plus complexes. Ces analyses ont été implémentées dans le programme QTDT [2].

Il y a quelques années, une autre méthode, utilisant cette fois-ci les modèles mixtes pour tester la présence d'empreinte parentale pour un trait quantitatif, a été publiée [55]. Les auteurs ont

1. `http://www.geneimprint.com/site/home`
2. `http://csg.sph.umich.edu/abecasis/QTDT/index.html`

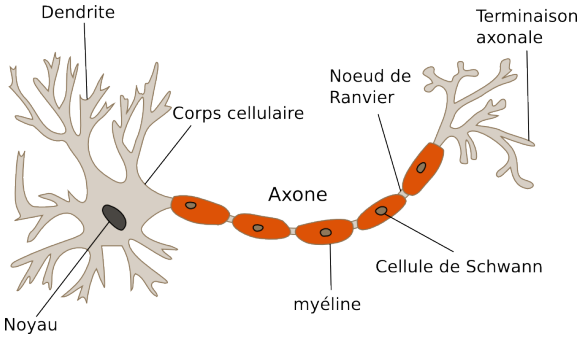
montré un gain de puissance pour leur méthode comparée à celle implémentée dans le logiciel QTDT.

1.3 Le cas de la Sclérose en Plaques

Dans les chapitres 4 et 5, les méthodes évoquées et éventuellement développées seront appliquées sur des données liées à la Sclérose en Plaques (SEP). Cette maladie touche presque 100 000 personnes en France. Dans cette section, nous allons exposer rapidement cette maladie ainsi que les facteurs environnementaux et génétiques connus comme étant associés à cette pathologie.

1.3.1 La pathologie

Le système nerveux central est constitué de la moelle épinière, du cerveau et des nerfs optiques. Ils sont, en partie, constitués de neurones. Le neurone est formé par un corps cellulaire et un axone qui transmet l'information du corps cellulaire vers les terminaisons axonales et donc vers un autre corps cellulaire (figure 1.19). Ce prolongement du neurone est entouré de myéline qui forme une gaine autour de lui. Cette substance permet de protéger l'axone et contribue à la transmission du signal nerveux.



http://www.pinville.com/reseaux-de-neurones.php

FIGURE 1.19 – Schéma d'un neurone.

La sclérose en plaques est une maladie neurologique auto-immune chronique du système nerveux central découverte par Charcot en 1868. Elle se manifeste par la dégradation de la myéline (démyélinisation) autour des fibres nerveuses par les cellules immunitaires du patient. Cette dégradation induit des interférences dans la transmission du signal nerveux. De plus, l'axone n'étant plus protégé, il se dégrade provoquant aussi une dégénérescence neurologique. La SEP se résume donc en deux processus cliniques ; un processus inflammatoire (réponse immunitaire) et une dégénérescence neurologique. Chaque processus provoque un événement distinct :

- ▷ la poussée résultant du processus inflammatoire. Elle est définie par l'apparition de nouveaux symptômes, la réapparition d'anciens symptômes ou une aggravation de ceux déjà présents. La poussée a une durée finie et une récupération variable.

- ▷ la progression résultant de la dégénérescence neurologique. Elle est définie par un aggravation continue des symptômes neurologiques sur plusieurs mois.

Plus concrètement, les symptômes de la SEP sont des troubles moteurs, sensitifs, de l'équilibre et visuels ou urinaires. La maladie est évolutive et peut aller jusqu'à la perte de la marche et finalement la mort. La forme et l'évolution de cette maladie varient beaucoup d'un patient à un autre. Trois formes principales ont été définies [56] :

- ▷ la forme rémittente-récurrente (RR) composée de poussées suivies de rémissions qui peuvent être incomplètes,
- ▷ la forme secondairement progressive (SP) composée d'une période avec des poussées (de la forme RR) suivie d'une phase de progression du handicap avec ou non des poussées,
- ▷ la forme primaire progressive (PP) définie par une progression continue du handicap à laquelle peut s'ajouter des poussées.

La communauté scientifique s'accorde à penser que la Sclérose en Plaques est une maladie complexe et que le développement de celle-ci résulte de l'action globale de plusieurs facteurs : des facteurs de susceptibilité génétique, des facteurs environnementaux et des interactions entre ces deux derniers.

1.3.2 Épidémiologie

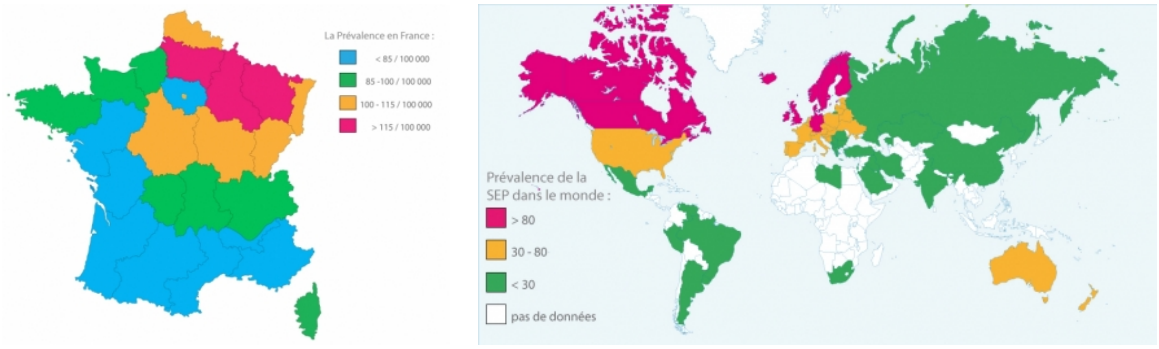
D'après le site de l'ARSEP [57], la prévalence de cette maladie dépend de la région du monde considérée. La Sclérose en Plaques concerne plus de 90 000 personnes en France avec entre 4 000 et 6 000 nouveaux cas par an.

Le premier facteur notable pour la SEP est le sexe. En effet, 3 patients atteints de la SEP sur 4 sont des femmes [57]. Il semble peu probable que cette différence soit expliquée par des gènes de susceptibilité présents sur les chromosomes sexuels mais plutôt par des particularités physiologiques telles que les hormones. De plus, ce ratio semble évoluer avec le temps [58].

Nous allons maintenant regarder les facteurs environnementaux associés à la SEP [59]. Beaucoup de facteurs ont été trouvés mais aucun n'a pu être totalement vérifié et peu d'entre eux ont une association forte.

L'un des premiers facteurs environnementaux cités est le virus d'Epstein-Barr [60–63]. Une étude récente ne trouve cependant pas d'infection active de ce virus dans le système nerveux de patients atteints de la Sclérose en Plaques [64]. Cette association reste donc pour le moment inexpliquée face à la difficulté de comprendre comment agit ce virus sur la maladie.

Un autre facteur fortement probable est la vitamine D. Plusieurs observations soutiennent cette hypothèse. La première est le lien visible entre la latitude et la prévalence de la maladie [65]. En effet, si nous regardons la figure 1.20 avec les cartes de prévalence en France et dans le monde données par l'ARSEP [57], nous observons un gradient Nord-Sud. Plus l'ensoleillement est important plus la prévalence de la SEP est faible. De plus, des études sur le mois de naissance [66] et les migrants [67] appuient cette hypothèse. Elle a cependant été remise en question dans un article datant maintenant de plusieurs années [68]. Une autre observation en faveur de l'influence de la vitamine D sur la SEP est son rôle dans la réponse immunitaire [69]. En effet, la vitamine



(a) France (b) Monde

https://www.arsep.org/fr

FIGURE 1.20 – Prévalence de la SEP en 2014

D intervient dans le développement des lymphocytes T, des cellules jouant un rôle essentiel dans la réponse immunitaire en détruisant des cellules infectées par exemple. De plus, des gènes impliqués dans la voie biologique de la vitamine D ont également été trouvés comme associés à la maladie [70].

Un dernier facteur associé à la SEP est le tabac. Plusieurs études montrent un effet du tabac sur l'apparition de la maladie mais également sur son évolution [71–75]. Le mécanisme permettant au tabac d'agir sur la maladie est pour le moment mal connu [59]. Plusieurs hypothèses ont été formulées mais aucune n'a pu être confirmée.

1.3.3 Génétique

Dès les premières découvertes sur cette maladie, la présence de facteurs génétiques est suggérée par l'observation de familles avec un grand nombre de cas de SEP [76,77]. En effet, des études ont trouvé une augmentation du risque de SEP chez les apparentés de malades. Cependant, ce risque varie beaucoup selon les études. De plus, il est impossible de différencier le risque dû à des facteurs génétiques ou à l'environnement partagé de la famille.

Une méta-analyse récente [78] a été faite sur huit études de jumeaux d'Europe et des États-Unis afin d'évaluer la portion de facteurs génétiques. Les résultats sont très variables selon les populations étudiées mais ils restent tous en faveur de l'existence de facteurs génétiques pour cette pathologie. Cette méta-analyse estime l'héritabilité de la Sclérose en Plaques à 50% avec un intervalle de confiance à 95% de [39, 61].

La présence de familles avec de nombreux cas a d'abord motivé des études de liaison qui n'ont malheureusement pas donné de résultats concluant en dehors du Complexe Majeur d'Histocompatibilité (CMH), une région du génome impliqué dans la réponse immunitaire. En effet, cette zone du génome est connue pour être liée à la SEP depuis les années 1970.

Par la suite, cette région a été retrouvée dans des études d'association familiales [80] puis dans les études GWAS. Ces dernières ont aussi trouvé comme associé avec la SEP de nombreux autres variants impliqués dans la réponse immunitaire ou avec d'autres fonctionnalités. Les associations trouvées et répliquées avant juin 2015 ont été référencées dans une revue de la littérature [79]. Les gènes associés à la SEP à travers leurs SNPs ou des SNPs intergéniques en DL avec eux sont résumés dans la table 1.2.

Fonction	Gènes	Chr	SNPs (Odd Ratio)
Réponse Immunitaire	VCAM1	1	rs11581062 (1.13), †rs7552544 (1.10)
	CD58	1	rs2300747 (1.30), rs1335532 (1.28, 1.18), rs6677309 (1.29)
	RGS1	1	†rs2760524 (1.15), †rs1323292 (1.12), †rs1359062 (1.15)
	EOMES	3	†rs170934 (1.17), †rs11129295 (1.11), †rs2371108 (1.10)
	CD86	3	rs9282641 (1.21)
	IL12A	3	rs4680534 (1.12), rs2243123 (1.09), †rs1014486 (1.11)
	IL7RA	5	rs6897932 (1.18, 1.12, 1.11), †rs6881706 (1.12)
	PTGER4	5	†rs6896969 (1.10), †rs4613763 (1.21), †rs9292777 (1.16)
	TCF7	5	†rs756699 (1.12)
	NDFIP1	5	rs1062158 (1.08)
	HLA-DRB1	6	rs3129934 (3.30, 2.34), rs3135388 (1.99, 2.75), †rs9271366 (2.78, 2.62), †rs2040406 (2.05), †rs3129889 (2.97)
	TNFAIP3, IL22RA2	6	†rs9321619 (1.12), †rs13192841 (1.10), †rs17066096 (1.14), †rs67297943 (1.11)
	IL2RA	7	rs12722489 (1.25, 1.23), rs2104286 (1.15, 1.16, 1.22), rs3118470 (1.12), †rs7090512 (1.19)
	CD6	11	rs17824933 (1.18), rs650258 (1.12), †rs34383631 (1.13)
	CXCR5	11	†rs630923 (1.12), †rs9736016 (1.10)
	TNFRSF1A	12	rs4149584 (1.58), rs180069 (1.20, 1.12, 1.14)
	CLEC16A	16	rs6498169 (1.14), rs11865121 (1.15), rs7200786 (1.15), rs12927355 (1.20)
	IRF8	16	†rs17445836 (1.25), †rs13333054 (1.12), †rs35929052 (1.15)
	STAT3	17	rs744166 (1.15, 1.13), rs9891119 (1.10), rs2293152 (1.22), rs4796791 (1.12)
	TNFSF14	19	rs1077667 (1.16)
	TYK2	19	†rs8112449 (1.10), rs34536443 (1.29)
	CD40	20	†rs6074022 (1.20, 1.15), rs2425752 (1.10), rs4810485 (1.11)
Vitamine D	CYP27B1	12	rs703842 (1.23)
	CYP24A1	20	†rs2248359 (1.12, 1.09)
Autres	MMEL1	1	†rs4648356 (1.16), rs3748817 (1.14)
	DDAH1	1	†rs233100 (1.08), rs11587876 (1.09)
	RPL5	1	rs6604026 (1.15, 1.17)
	EVI5	1	rs11810217 (1.15), rs41286801 (1.19)
	C1orf106	1	rs7522462 (1.11), rs55838263 (1.13)
	PLEK	2	†rs7595037 (1.10), †rs7595717 (1.10)
	SP140	3	rs10201872 (1.13), rs9989735 (1.16)
	ILDR1	3	rs2255214 (1.13)
	TIMMDC1	3	rs2293370 (1.13), rs1131265 (1.17)
	IQCB1	3	rs1920296 (1.12)
	RGS14	5	†rs4075958 (1.09), rs4976646 (1.12)
	BACH2	6	rs12212193 (1.09), rs72928038 (1.14)
	OLIG3	6	†rs9321619 (1.12), †rs13192841 (1.10), †rs17066096 (1.14), †rs67297943 (1.11)
	AHI1	6	rs11154801 (1.13, 1.12), †rs9321619 (1.12), †rs13192841 (1.10), †rs17066096 (1.14), †rs67297943 (1.11)
	TAGAP	6	rs1738074 (1.13), †rs212405 (1.12)
	MYC, MIR1204, MIR1205, PVT1, MIR1208, MIR3686	8	†rs4410871 (1.11, 1.12), †rs2019960 (1.12), †rs759648 (1.08)
	ZMIZ1	10	rs1250540 (1.12), rs1250550 (1.10), rs1250542 (1.15), rs1782645 (1.10)
	HHEX	10	†rs7923837 (1.10, 1.11)
	METTL1	12	rs703842 (1.23)
	AGAP2	12	rs12368653 (1.11)
	ZFP36L1	14	†rs4902647 (1.11), rs2236262 (1.08)
	GALC	14	rs2119704 (1.26), rs74796499 (1.32)
	DKKL1	19	rs2303759 (1.11), rs8107548 (1.11)
	MAPK1	22	rs2283792 (1.09, 1.10)

† = SNP intergénique (dans ce cas, les gènes donnés sont les plus proches)

Table 1.2 – Génétique de la SEP [79].

Depuis juin 2015, des réplications pour ces locus de susceptibilité ont été faites [81–83] et d'autres associations ont été découvertes :

- ▷ CD24 codant une protéine intervenant sur les surfaces des cellules immunitaires et du système nerveux central [84],
- ▷ MIR3681 (rs1534422) qui code un ARN,
- ▷ CD28 (rs6435203) intervenant sur les lymphocytes T [85],
- ▷ LPP (rs4686953) dont la protéine agit à la surface des cellules [85],
- ▷ ETS1 (rs3809006) qui intervient notamment lors de la maturation des lymphocytes T [85],
- ▷ DLEU1 (rs806349) codant un ARN [85],
- ▷ LPIN3 (rs6072343) qui intervient notamment dans la régulation de l'expression des gènes [85],
- ▷ IFNGR2 (rs9808753) intervenant dans la réponse immunitaire [85],
- ▷ ANKRD55 (rs6859219) [81, 86, 87],
- ▷ IFI30 (rs11554159) codant un enzyme réductase présente dans le lysosome, un organite qui effectue la digestion intracellulaire. Cette enzyme permet de découper les protéines en antigènes reconnus par les cellules T [81, 88],
- ▷ IL4 (rs2243250) et son récepteur IL4R (rs1801275) codant une protéine intervenant dans la différenciation des lymphocytes T [89],
- ▷ IL18 (promoteur) dont la protéine stimule l'activité des lymphocytes de l'immunité innée [90],
- ▷ NR1H3 (rs2279238) codant une protéine qui intervient lors de la transcription des gènes impliqués dans certaines réponses immunitaires [91],
- ▷ VDR intervenant dans la voie biologique de la vitamine D [92].

1.4 Les objectifs et la suite du manuscrit

Dans ce chapitre, nous avons défini les différents principes en génétique que nous utilisons tout au long de ce manuscrit. Nous avons également placé le contexte scientifique pour les types d'analyses génétiques qui nous intéressent ici ainsi que pour la Sclérose en Plaques que nous étudierons par la suite.

Dans la suite de ce manuscrit, nous allons nous focaliser principalement sur l'utilisation des modèles mixtes dans le domaine de la génétique. Pour cela, le prochain chapitre est consacré à la théorie des modèles mixtes ainsi qu'aux méthodes d'analyses des données génétiques basées sur ces modèles. Nous aborderons alors le problème pratique de la stratification de population lors de l'estimation de l'héritabilité sur des données en population à l'aide du modèle linéaire mixte. Ce chapitre repose sur les données françaises de l'étude Trois-Cités. Le chapitre suivant est une parenthèse dans l'exploration des utilisations du modèle mixte en génétique. Nous y montrons l'avantage des données familiales en les exploitant pour l'étude d'association et l'estimation du

risque allélique du variant causal sous le modèle multiplicatif même si celui-ci n'est observé qu'au travers de tag-SNPs. Pour finir, le dernier chapitre combine nos expertises faites sur les modèles mixtes d'une part et les données familiales d'autre part pour analyser avec les modèles mixtes des données familiales de la Sclérose en Plaques.

Chapitre 2

Le modèle linéaire mixte en génétique

Dans le chapitre précédent, nous avons introduit des notions de génétique et décrit succinctement le contexte scientifique actuel. Nous allons maintenant nous concentrer sur la méthodologie statistique de cette thèse. Ce chapitre est constitué tout d'abord d'un rappel non exhaustif des différents termes et notations utilisés en statistique. Puis, nous allons exposer la théorie des modèles mixtes avant de l'appliquer au cas particulier des analyses en génétique. Ce travail a notamment été valorisé avec un article (Annexe 8.3) et l'élaboration d'un package R décrit dans la dernière section de ce chapitre.

2.1 Quelques notions de statistiques et notations

2.1.1 L'aléatoire

Une *expérience aléatoire* est une expérience dont le résultat est impossible à prévoir, au contraire d'une expérience déterministe. Ce phénomène nous est en fait très familier ; tirer à pile ou face ou lancer un dé est une expérience aléatoire. Le résultat d'une expérience peut être alors décrite par une *loi de probabilité* qui à chaque événement possible associe une probabilité comprise entre 0 et 1 notée :

$$\mathbb{P} : \Omega \to [0, 1]$$

avec Ω l'ensemble des résultats ou événements possibles de l'expérience aléatoire. Dans ce manuscrit, nous utiliserons également des lois de probabilités *conditionnelles* notée $\mathbb{P}[.|A]$ avec « $|A$ » signifiant « sachant l'évènement A ».

Nous pouvons alors modéliser une expérience aléatoire par une *variable aléatoire* ou une variable qui à chaque résultat possible associe une valeur numérique. Une variable aléatoire est définie par sa loi ou sa distribution qui peut être :

- ▷ discrète, si le nombre de valeurs possibles est fini ou dénombrable. Elle est alors définie par les probabilités d'obtenir chacune des valeurs possibles.
- ▷ continue, si les valeurs sont prises dans un intervalle ou une union d'intervalles. La majorité des variables de ce type peuvent être définies par une *densité*. Une densité est une fonction qui donne un « poids » à chaque valeur de l'intervalle de définition.

D'une façon générale, une variable aléatoire sera désignée par une majuscule. Si nous notons X, une variable aléatoire, nous pouvons évoquer dans ce manuscrit :

- ▷ la moyenne ou *espérance* de X notée $\mathbb{E}[X]$,

▷ la dispersion ou *variance* de X notée $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

Si nous n'observons plus une mais deux variables aléatoires X et Y, nous pouvons également calculer la covariance, notée $\mathrm{Cov}(X, Y)$, et la corrélation, notée $\mathrm{Cor}(X, Y)$ et comprise entre -1 et 1, entre ces deux variables. Ces deux valeurs représentent le lien linéaire entre elles.

Nous allons maintenant rappeler quelques distributions que nous utiliserons dans la suite de ce manuscrit et expliciter leur notation.

Les distributions Binomiale et Multinomiale

Commençons par regarder quelques lois discrètes classiques. La loi de *Bernoulli* est une loi discrète prenant ces valeurs dans $\{0, 1\}$. Cette variable peut, par exemple, modéliser la présence ou l'absence d'un trait chez un individu. Une loi de Bernoulli est définie par la fréquence du trait dans la population p. Si la variable aléatoire X suit une loi de Bernoulli de paramètre p, nous notons $X \sim \mathcal{B}(p)$. De cette loi, nous pouvons en déduire la loi *binomiale*, notée $\mathcal{B}in(n, p)$, qui est la somme de n lois de Bernoulli de paramètre p. Cette loi modélise le nombre d'individus qui ont un trait dans une population de taille n.

Pour finir, la généralisation de cette loi s'appelle la loi *multinomiale*, $\mathcal{M}ulti(n, p_1, \ldots, p_k)$. En effet, la loi binomiale considère uniquement deux catégories ; ceux qui ont le trait et ceux qui ne l'ont pas. La loi multinomiale permet de considérer plus de deux possibilités. Cette loi nous donne les effectifs dans chaque catégorie pour une population de taille n. Par exemple, les effectifs de chaque génotype d'un variant bi-allélique A/a dans une population de taille donnée suit une loi multinomiale avec trois catégories AA, Aa et aa.

La distribution normale et la loi du χ^2

Nous utiliserons dans ce manuscrit essentiellement deux lois continues. La première est la loi normale ou gaussienne d'espérance μ et de variance σ^2. Le nombre de lois normales possibles est infini. Cependant, nous pouvons noter que la loi normale la plus utilisée est la loi normale centrée réduite ou de moyenne nulle et de variance 1, notée $\mathcal{N}(0, 1)$. La loi normale peut se généraliser à non plus une variable aléatoire mais à un vecteur de n variables aléatoires $(X_1, \ldots, X_n)$. Dans ce cas, nous parlons de loi normale multivariée notée :

$$\begin{pmatrix} X_1 \\ X_2 \\ . \\ . \\ . \\ X_n \end{pmatrix} \sim \mathcal{N}_n \left(\mu = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ . \\ . \\ . \\ \mathbb{E}[X_n] \end{pmatrix}, V = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \ldots & \mathrm{Cov}(X_2, X_n) \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \ldots & \mathrm{Var}(X_n) \end{pmatrix} \right).$$

La matrice V est la *matrice de variance-covariance* de taille $(n \times n)$. En multivarié, la loi normale centrée réduite correspond à un vecteur de variables de loi normale $\mathcal{N}(0, 1)$ indépendantes et se note $\mathcal{N}_n(\mathbb{O}_n, \mathbb{I}_n)$ avec $\mathbb{O}_n$ le vecteur de 0 de taille n et $\mathbb{I}_n$ la matrice identité de taille n c'est-à-dire la matrice de taille $(n \times n)$ avec des 1 sur la diagonale et des 0 partout ailleurs. Au cours de ce manuscrit, nous utiliserons notamment la densité de la loi normale

multivariée $\mathcal{N}_n(\mu, V)$ qui s'écrit :

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{|V|}} \exp\left(-\frac{1}{2}(x-\mu)'V^{-1}(x-\mu)\right)$$

avec $x = (x_1, \ldots, x_n)' \in \mathbb{R}^n$ un vecteur d'observations et $|V|$ le *déterminant* de la matrice V.

À partir de la loi normale centrée réduite non multivariée, nous obtenons la loi du χ^2_1 à un degré de liberté qui correspond à $\mathcal{N}(0,1)^2$. La loi χ^2_p à p degrés de liberté correspond à la somme de p χ^2_1 indépendantes.

2.1.2 Les modèles statistiques

Durant une étude, nous observons des variables aléatoires. Un *modèle statistique* est un modèle donnant une distribution à une variable à expliquer Y qui dépend d'autres variables observées appelées variables explicatives, $X_1, \ldots X_n$,

$$Y = f(X_1, \ldots X_n).$$

Celui-ci est subjectif et donc choisi par celui qui analyse les données afin de répondre au mieux à une question donnée. Les modèles statistiques les plus utilisés sont les modèles paramétriques, c'est-à-dire un modèle pour lequel le lien entre la variable à expliquer et les variables explicatives dépend uniquement de paramètres $\theta = (\theta_1, \ldots, \theta_k)$. Le but sera alors d'obtenir des *estimateurs* de ces paramètres et de tester nos hypothèses à partir du modèle statistique résultant.

Une fois le modèle choisi, nous pouvons en déduire une *vraisemblance* qui dépend des observations. La vraisemblance, notée L, est la probabilité « d'observer ce que nous observons dans notre étude ». La plupart du temps, afin de simplifier les calculs, nous regardons la *log-vraisemblance*, notée ℓ, qui est le logarithme népérien (log) de la vraisemblance. Nous allons alors maximiser cette vraisemblance afin d'obtenir le meilleur modèle possible sous les contraintes que nous avons imposées (variables observées et distribution). Dans le cadre d'un modèle paramétrique, nous obtenons ainsi les estimateurs du maximum de vraisemblance qui possèdent des propriétés très intéressantes.

2.1.3 Les tests statistiques

En statistique, nous cherchons à tester des hypothèses. Pour cela, nous définissons l'hypothèse à tester ou l'*hypothèse nulle*, notée H_0, et une *hypothèse alternative*, notée H_1. Pour tester notre hypothèse nulle, nous allons alors définir une *statistique* ou une variable aléatoire dont la loi est connue sous H_0. L'étape suivante consiste à regarder si la valeur observée de cette statistique est « trop éloignée » des valeurs attendues sous H_0. Dans ce cas, nous rejetons l'hypothèse nulle. Nous parlerons alors de *p-valeur* qui correspond à la probabilité sous l'hypothèse nulle d'obtenir une statistique de test au moins aussi éloignée que celle observée.

Dans ce manuscrit, nous regarderons également la *puissance* des tests considérés ou la probabilité de rejeter l'hypothèse nulle quand celle-ci n'est pas vérifiée.

2.2 Le modèle linéaire mixte, la théorie

Le modèle linéaire mixte est un modèle statistique rassemblant des effets fixes et des effets aléatoires. C'est dans les années 1950 que le modèle linéaire mixte tel que nous le connaissons aujourd'hui a été développé. Par la suite, il a donné lieu à de nombreux développements méthodologiques comme les modèles mixtes généralisés. Le modèle mixte suppose donc que certains des effets ne sont plus fixes mais tirés au hasard dans une loi donnée permettant ainsi de réduire le nombre de paramètres à estimer. Nous allons, dans un premier temps, nous intéresser à la théorie du modèle linaire mixte c'est-à-dire un modèle qui exprime la variable à expliquer comme une combinaison linéaire des variables explicatives. Ce modèle suppose également que les effets aléatoires et donc la variable à expliquer, suivent une loi normale. Une fois la théorie du modèle linéaire mixte mise en place, nous discuterons du cas des traits binaires puis des différentes utilisations du modèle mixte en génétique humaine.

2.2.1 Les notations

Nous allons ici exposer une méthodologie pouvant être appliquée à d'autres domaines que la génétique ; le modèle linéaire mixte. Volontairement, cette méthode sera exposée de façon très générale avant d'être appliquée à la génétique. Pour commencer, nous introduisons les notations suivantes utilisées tout au long de ce manuscrit :

- ▷ Y le vecteur d'observations à expliquer (phénotypes),
- ▷ $X \in \mathbb{R}^{n\times p}$ la matrice des covariables introduites dans le modèle,
- ▷ β les coefficients fixes pour les covariables,
- ▷ $Z_j \in \mathbb{R}^{n\times q_j}$ une matrice de variables avec des effets aléatoires pour $j = 1, \ldots, k$,
- ▷ $K_j = Z_j Z_j' \in \mathbb{R}^{n\times n}$ la matrice des corrélations entre individus estimées avec les variables incluses dans la matrice Z_j pour $j = 1, \ldots, k$.

D'autres notations seront introduites tout au long de ce chapitre.

2.2.2 Le modèle

Le modèle linéaire mixte peut s'écrire sous la forme de trois modèles équivalents qui donnent au vecteur des observations Y la même distribution gaussienne :

- ▷ le modèle avec des vecteurs d'effets aléatoires, u_j, associés à une partie des variables explicatives :

$$Y = X\beta + Z_1 u_1 + \cdots + Z_k u_k + e \tag{2.1}$$

  avec :

  - $u_j \sim \mathcal{N}_n(0, \tau_j \mathbb{I}_{q_j})$ pour $j = 1, \ldots, k$,
  - $e \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_n)$.

▷ le modèle avec des vecteurs d'effets aléatoires individuels, ω_j,

$$Y = X\beta + \omega_1 + \cdots + \omega_k + e \tag{2.2}$$

avec :

- $\omega_j \sim \mathcal{N}_n(0, \tau_j K_j)$ pour $j = 1, \ldots, k$,
- $e \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_n)$.

▷ le modèle avec un vecteur d'effets aléatoires résumant tous les termes aléatoires introduits dans le modèle :

$$Y = X\beta + \epsilon \tag{2.3}$$

avec $\epsilon \sim \mathcal{N}_n(0, V)$ où $V = \tau_1 K_1 + \cdots + \tau_k K_k + \sigma^2 \mathbb{I}_n$ est la matrice de variance-covariance.

Dans ces trois modèles, le vecteur d'observations Y suit la loi $\mathcal{N}_n(X\beta, V)$ dépendant des paramètres $\beta, \tau_1, \ldots, \tau_k, \sigma^2$. Leur différence réside dans la définition des variables latentes ou non-observées.

2.2.3 La vraisemblance

Le vecteur d'observations Y suit une loi normale multivariée de moyenne $X\beta$ et de variance $V = \tau_1 K_1 + \cdots + \tau_k K_k + \sigma^2 \mathbb{I}_n$. Donc, à une constante multiplicative près, la vraisemblance s'écrit :

$$L(\beta, \tau_1, \ldots, \tau_k, \sigma^2) = \frac{1}{\sqrt{|V|}} \exp\left(-\frac{1}{2}(Y - X\beta)' V^{-1} (Y - X\beta)\right).$$

D'où la log-vraisemblance :

$$\ell(\beta, \tau_1, \ldots, \tau_k, \sigma^2) = -\frac{1}{2} \log |V| - \frac{1}{2}(Y - X\beta)' V^{-1} (Y - X\beta).$$

Afin d'estimer les différents paramètres du modèle, la première solution envisagée est de maximiser cette vraisemblance. Cependant, les estimateurs des paramètres de la variance τ_1, $\ldots$, τ_k, σ^2 obtenus en maximisant la log-vraisemblance sont biaisés. En effet, ce biais sur les estimations des composantes de la variance est induit par le fait que la moyenne $X\beta$ est remplacée par son estimation. Un exemple simple est le modèle linéaire classique avec uniquement une variance résiduelle σ^2. Si nous reprenons les notations précédentes, nous avons $V = \sigma^2 \mathbb{I}_n$. Dans ce cas, la log-vraisemblance s'écrit :

$$\ell(\beta, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2}\sigma^2 \sum_{i=1}^{n} (Y_i - X\beta)^2,$$

et les estimateurs du maximum de vraisemblance sont :

$$\widehat{\beta} = (X'X)^{-1} X'Y$$
$$\widehat{\sigma^2} = \frac{1}{n} \sum_i \left(Y_i - X\widehat{\beta}\right)^2.$$

Cette estimation de la variance $\widehat{\sigma^2}$ est connue pour être biaisée par la présence de $X\widehat{\beta}$.

C'est pourquoi, nous utilisons la vraisemblance restreinte ou plutôt son logarithme permettant d'obtenir des estimateurs non biaisés.

2.2.4 La vraisemblance restreinte

Le biais des estimations des composantes de la variance étant provoqué par l'utilisation de $\widehat{\beta}$, l'idée est d'introduire une matrice de contrastes $C \in \mathbb{R}^{(n-p)\times n}$ telle que :

- ▷ $CX = 0$,
- ▷ $CC' = \mathbb{I}_{n-p}$.

Cette matrice nous permet de construire un nouveau modèle linéaire mixte découlant des modèles précédents :

$$\begin{aligned} CY &= CX\beta + C\omega_1 + \cdots + C\omega_k + Ce \\ Z &= \omega_1' + \cdots + \omega_k' + e' \end{aligned} \tag{2.4}$$

avec :

$$\begin{aligned} &\omega_1' \sim \mathcal{N}_n(0, \tau_1 CK_1C'), \\ &\ldots, \\ &\omega_k' \sim \mathcal{N}_n(0, \tau_k CK_kC'), \\ &e' \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_{n-p}). \end{aligned}$$

Sous ce modèle, la variable Z suit une loi normale multivariée de moyenne nulle et de variance :

$$W = \tau_1 CK_1C' + \cdots + \tau_k CK_kC' + \sigma^2 \mathbb{I}_{n-p}.$$

Nous en déduisons la log-vraisemblance restreinte suivante :

$$\begin{aligned} \ell^{\mathrm{re}}(\tau_1, \ldots, \tau_k, \sigma^2) &= -\frac{1}{2}\log|W| - \frac{1}{2}Z'W^{-1}Z \\ &= -\frac{1}{2}\log\left|CVC'\right| - \frac{1}{2}Y'C'\left(CVC'\right)^{-1}CY. \end{aligned}$$

Cette expression peut être simplifiée à l'aide de lemmes mathématiques donnés en Annexe 1, qui assurent que :

- ▷ $C'\left(CVC'\right)^{-1}C = V^{-1} - V^{-1}X\left(X'V^{-1}X\right)^{-1}X'V^{-1}$
- ▷ $\log|CVC'| = \log|V| + \log\left|X'V^{-1}X\right| + \text{constante}$

Ainsi, nous obtenons l'écriture suivante pour la log-vraisemblance restreinte :

$$\ell^{\mathrm{re}}(\tau_1, \ldots, \tau_k, \sigma^2) = -\frac{1}{2}\log|V| - \frac{1}{2}\log\left|X'V^{-1}X\right| - \frac{1}{2}Y'PY + \text{constante}$$

avec $P = V^{-1} - V^{-1}X\left(X'V^{-1}X\right)^{-1}X'V^{-1}$. Il est intéressant de noter ici que les termes importants de la log-vraisemblance ne dépendent plus de la matrice C.

À l'aide de cette log-vraisemblance restreinte, nous pouvons alors obtenir des estimations non biaisées des paramètres de la variance. Les dérivées premières et secondes, utilisées dans l'algorithme de maximisation, sont calculées en Annexe 1.

2.2.5 Les estimations des composantes de la variance

Pour maximiser ℓ^{re} et donc estimer les composantes de la variance, plusieurs solutions sont possibles. Le premier algorithme de maximisation auquel nous pouvons penser est l'algorithme Espérance-Maximisation (EM) classiquement utilisé en présence de variables latentes. Cet algorithme est composé de deux étapes itératives :

- ▷ "E-step", pour des paramètres de variance donnés, nous calculons les moyennes et les variances des effets aléatoires (ω_j et e) conditionnellement au vecteur des observations Y.
- ▷ "M-step", en utilisant les moyennes et variances calculées précédemment, nous estimons les nouveaux paramètres de variance par leur espérance.

Il peut être appliqué sur la log-vraisemblance ou la log-vraisemblance restreinte. Dans le deuxième cas, les deux étapes peuvent être résumées par les itérations suivantes :

$$\begin{aligned}
\tau_{r+1} &= \tau_r + \frac{1}{r_K}\tau_r^2\left(y'P_rKP_ry - \mathrm{Tr}\left(KP_r\right)\right) \\
&= \tau_r + \left(\frac{2\tau_r^2}{r_k}\right)\frac{\partial \ell^{\mathrm{re}}}{\partial \tau}\left(\tau_r, \sigma_r^2\right) \\
\sigma_{r+1}^2 &= \sigma_r^2 + \frac{1}{n}\left(\sigma_r^2\right)^2\left(y'P_rP_ry - \mathrm{Tr}\left(P_r\right)\right) \\
&= \sigma_r^2 + \frac{2\left(\sigma_r^2\right)^2}{n}\frac{\partial \ell^{\mathrm{re}}}{\partial \sigma^2}\left(\tau_r, \sigma_r^2\right)
\end{aligned}$$

avec r le numéro de l'itération effective.

L'algorithme EM converge avec certitude. Cependant, cette convergence peut être très longue. Nous pouvons donc regarder d'autres algorithmes de maximisation qui ont été développés . Nous avons par exemple :

- ▷ l'algorithme de Newton-Raphson :

  $$\theta^{r+1} = \theta^r + J(\theta^r)^{-1}U(\theta^r)$$

  où $U(\theta)$ est le gradient de la log-vraisemblance, $J(\theta)$ la matrice d'information observée ou l'inverse de la matrice Hessienne de la log-vraisemblance.

- ▷ le « *Fisher Scoring* » :

  $$\theta^{r+1} = \theta^r + I(\theta^r)^{-1}U(\theta^r)$$

  où $I(\theta)$ est la matrice d'information de Fisher ou l'espérance de $J(\theta)$.

avec θ le vecteur des paramètres.

Ces algorithmes demandent des calculs de matrices très lourds, notamment avec le calcul de la trace des matrices $K_jPK_{j'}P$ (la hessienne et la matrice d'information de Fisher sont en

Annexe 1). Le temps de calcul très long a justifié le développement de la méthode appelée « *Average Information Restricted Likelihood Maximisation* » (AIREML). Cette méthode utilise la même procédure que les deux algorithmes précédents mais avec une autre matrice ; la moyenne l'inverse de la hessienne et de la matrice d'information de Fisher qui vaut :

$$AI(\tau_1, \ldots, \tau_k, \sigma^2) = \begin{bmatrix} \frac{1}{2}Y'PK_1PK_1PY & \cdots & \frac{1}{2}Y'PK_1PK_kPY & \frac{1}{2}Y'PK_1PPY \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}Y'PK_kPK_1PY & \cdots & \frac{1}{2}Y'PK_kPK_kPY & \frac{1}{2}Y'PK_kPPY \\ \frac{1}{2}Y'PPK_1PY & \cdots & \frac{1}{2}Y'PPK_kPY & \frac{1}{2}Y'PPPY \end{bmatrix}$$

La trace ayant disparu, le calcul de cette matrice est plus rapide. En effet, en choisissant l'ordre de calcul des produits matriciels, nous pouvons nous limiter à des produits entre une matrice et un vecteur. Ce type de produit présente beaucoup moins d'opérations qu'un produit de deux matrices.

2.2.6 Le calcul des BLUPs

Nous avons donc une méthode efficace pour estimer les composantes de la variance de notre modèle. Une fois ces paramètres obtenus, nous pouvons estimer les vecteurs des différents effets (fixes et aléatoires). Les estimateurs d'un modèle linéaire mixte sont appelés les « *Best Linear Unbiased Predictors* » (BLUPs). Nous allons regarder séparément les effets fixes et aléatoires.

Les effets fixes

Pour les effets fixes, nous regardons le modèle sous sa troisième forme (équation (2.3)) une fois les composantes de la variance estimées :

$$Y = X\beta + \epsilon$$

avec $\epsilon \sim \mathcal{N}_n(0, \widehat{V})$ où $\widehat{V} = \widehat{\tau}_1 K_1 + \cdots + \widehat{\tau}_k K_k + \widehat{\sigma^2}\mathbb{I}_n$. Nous cherchons donc à estimer β. La log-vraisemblance restreinte ne dépendant pas de β, nous devons utiliser ici la log-vraisemblance classique. La dérivée de celle-ci par rapport à β, donnée en Annexe 1, vaut :

$$\frac{\partial \ell}{\partial \beta}\left(\beta, \widehat{\tau}_1, \cdots, \widehat{\tau}_k, \widehat{\sigma^2}\right) = X'\widehat{V}^{-1}(Y - X\beta).$$

Si nous annulons cette dérivée, nous obtenons l'équation :

$$X'\widehat{V}^{-1}X\beta = X'\widehat{V}^{-1}Y$$

qui a une unique solution :

$$\widehat{\beta} = \left(X'\widehat{V}^{-1}X\right)^{-1} X'\widehat{V}^{-1}Y. \tag{2.5}$$

Cette estimation dépend uniquement des composantes de la variance.

De plus, la variance de cet estimateur, $\widehat{\beta}$, peut être estimée par :

$$\begin{aligned}\widehat{\mathrm{Var}\left(\widehat{\beta}\right)} &= \left(X'\widehat{V}^{-1}X\right)^{-1} X'\widehat{V}^{-1}\widehat{\mathrm{Var}(Y)}\widehat{V}^{-1}X\left(X'\widehat{V}^{-1}X\right)^{-1} \\ &= \left(X'\widehat{V}^{-1}X\right)^{-1} X'\widehat{V}^{-1}\widehat{V}\widehat{V}^{-1}X\left(X'\widehat{V}^{-1}X\right)^{-1} \\ &= \left(X'\widehat{V}^{-1}X\right)^{-1}. \end{aligned} \tag{2.6}$$

Afin de faciliter les calculs, l'estimation de β peut également s'écrire sous la forme :

$$\widehat{\beta} = \left(X'X\right)^{-1} X'(Y - \widehat{V}\widehat{P}Y)$$

avec $\widehat{P} = \widehat{V}^{-1} - \widehat{V}^{-1}X(X'\widehat{V}^{-1}X)^{-1}X'\widehat{V}^{-1}$. L'intérêt de cette forme réside dans le fait que l'algorithme de maximisation nécessite de calculer le produit PY à chaque itération.

Les effets aléatoires

Nous allons maintenant regarder les estimateurs des effets aléatoires pour les différents types de modèle, les effets aléatoires associés aux variables latentes u_j puis les effets aléatoires individuels ω_j.

Les estimations des variables u_j et e, $\widehat{u_\ell}$ et $\widehat{e}$, sont obtenues en maximisant le logarithme de la densité jointe :

$$-\frac{1}{2}\left(\frac{1}{\widehat{\tau_1}}u_1'u_1 + \cdots + \frac{1}{\widehat{\tau_k}}u_k'u_k + \frac{1}{\widehat{\sigma^2}}e'e\right)$$

obtenue à partir de la densité de la loi normale multivariée et définie à une constante additive près.

Les estimations $\widehat{u}_j$ et $\widehat{e}$ doivent également vérifier la contrainte :

$$Z_1\widehat{u_1} + \cdots + Z_k\widehat{u_k} + \widehat{e} = Y - X\widehat{\beta} = \widehat{V}\widehat{P}Y \tag{2.7}$$

En utilisant la méthode de Lagrange d'optimisation sous contrainte [93], nous obtenons les équations :

$$\begin{cases} \dfrac{1}{\widehat{\tau_1}}\widehat{u}_1 = Z_1'\lambda \\ \vdots \\ \dfrac{1}{\widehat{\tau_k}}\widehat{u}_k = Z_k'\lambda \\ \dfrac{1}{\widehat{\sigma^2}}\widehat{e} = \lambda \end{cases}$$

pour $\lambda \in \mathbb{R}^n$. Pour déterminer λ, nous revenons sur la contrainte (équation (2.7)) :

$$\begin{aligned} &Z_1\widehat{u_1} + \cdots + Z_k\widehat{u_k} + \widehat{e} = VPY \\ &\widehat{\tau_1}Z_1Z_1'\lambda + \cdots + \widehat{\tau_k}Z_kZ_k'\lambda + \widehat{\sigma^2}\lambda = \widehat{V}\widehat{P}Y \\ &\widehat{V}\lambda = \widehat{V}\widehat{P}Y \\ &\lambda = \widehat{P}Y. \end{aligned}$$

Les BLUPs des effets aléatoires sont donc :

$$\begin{aligned} \widehat{u}_1 &= \widehat{\tau}_1 Z_1' \widehat{P} Y \\ &\vdots \\ \widehat{u}_k &= \widehat{\tau}_k Z_k' \widehat{P} Y \\ \widehat{e} &= \widehat{\sigma^2} \widehat{P} Y. \end{aligned} \tag{2.8}$$

Ainsi, nous pouvons en déduire les BLUPs des effets individuels $\omega_j = Z_j u_j$:

$$\begin{aligned} \widehat{\omega}_1 &= \widehat{\tau}_1 K_1 \widehat{P} Y \\ &\vdots \\ \widehat{\omega}_k &= \widehat{\tau}_k K_k \widehat{P} Y \\ \widehat{e} &= \widehat{\sigma^2} \widehat{P} Y. \end{aligned} \tag{2.9}$$

2.2.7 Tester les paramètres du modèle

Nous pouvons distinguer ici encore deux types de paramètres ; les paramètres fixes et les composantes de la variance.

Les effets fixes

Pour les effets fixes β, nous voulons tester :

$$H_0 : \beta = \beta_0 \text{ vs } H_1 : \beta \neq \beta_0.$$

Pour cela, les tests asymptotiques classiques (figure 2.1) sont possibles :

- le test asymptotique de Wald qui regarde la différence relative entre l'estimation $\widehat{\beta}$ et β_0,

  $$T = (\widehat{\beta} - \beta_0)' \mathrm{Var}(\widehat{\beta})^{-1} (\widehat{\beta} - \beta_0) \sim \chi^2_p.$$

  La variance de $\widehat{\beta}$ est estimée en supposant les paramètres de la variance fixes. Nous avons donc un test approché. À cause de cette approximation, ce test a tendance à être non conservateur pour les petits échantillons [94].

- le test du rapport de vraisemblance, qui regarde la différence des log-vraisemblances évaluées en $\widehat{\beta}$ et β_0. Les deux vraisemblances étant calculées pour des paramètres de variances différents, les hypothèses de ce test ne sont pas vérifiées. Un ajustement de la distribution du rapport de vraisemblance a été proposé pour répondre à ce problème [95].

▷ le test du score qui regarde la pente la log-vraisemblance en β_0,

$$\begin{aligned} T &= \frac{\partial \ell}{\partial \beta}(\beta_0, \widehat{\tau}_1, \ldots, \widehat{\tau}_k, \widehat{\sigma}^2) \\ &= X'\widehat{V_0}^{-1}(Y - X\beta_0) \\ &= X'\widehat{P_0}Y \end{aligned}$$

où $\widehat{P_0}$, $\widehat{V_0}$, $\widehat{\tau}_j$ et $\widehat{\sigma}^2$ les estimations des matrices P et V et des composantes de la variance sous H_0. La variance de la variable T sous l'hypothèse nulle est estimée par :

$$\mathrm{Var}(T|H_0) = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \beta^2}(\beta_0, \widehat{\tau}_1, \ldots, \widehat{\tau}_k, \widehat{\sigma}^2)\right]$$

ou, comme le score est centré :

$$\begin{aligned} \mathrm{Var}(T|H_0) &= \mathbb{E}\left[TT'\right] \\ &= \mathbb{E}\left[X'\widehat{P_0}YY'\widehat{P_0}X\right] \\ &= X'\widehat{P_0}\mathbb{E}[YY']\widehat{P}X \\ &= X'\widehat{P_0}\left(V_0 + (X\beta_0)(X\beta_0)'\right)\widehat{P_0}X \\ &= X'\widehat{P_0}V_0\widehat{P_0}X + X'(\widehat{P_0}X\beta_0)(\widehat{P_0}X\beta_0)'X \\ &= X'\widehat{P_0}X \end{aligned}$$

car $\widehat{P_0}\widehat{V_0}\widehat{P_0} = \widehat{P_0}$ et $\widehat{P_0}X = \mathbb{O}_{n\times p}$ où $\mathbb{O}_{n\times p}$ est la matrice de 0 de taille $(n \times p)$. Il est nécessaire de préciser que cette estimation est une approximation. En effet, elle ne prend pas en compte la variance introduite par l'estimation des composantes de la variance. Ainsi, la statistique de test s'écrit :

$$T'\mathrm{Var}(T|H_0)^{-1}T \sim \chi^2_p$$

.

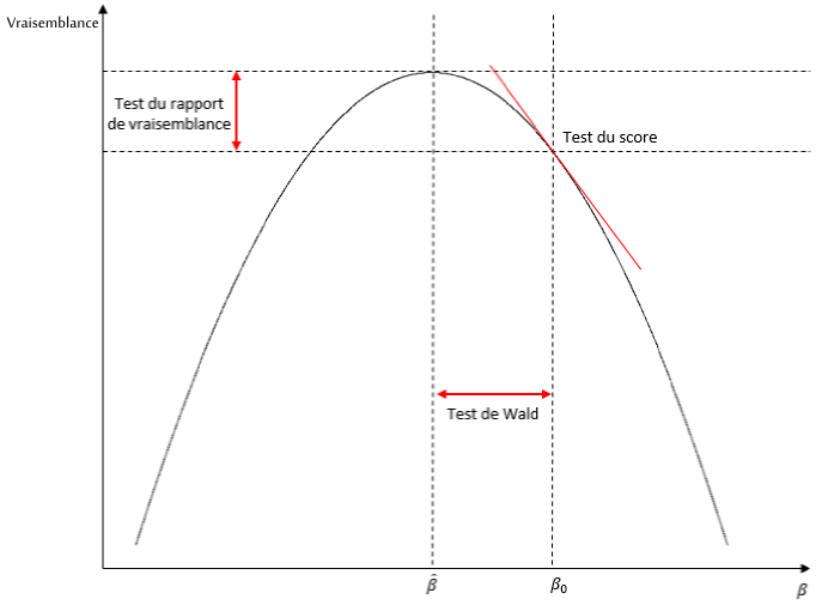


FIGURE 2.1 – Les tests classiques pour un effet fixe.

Nous pouvons noter que nous avons pris la log-vraisemblance classique pour les tests du rapport de vraisemblance et du score. En effet, la log-vraisemblance restreinte ne dépend plus de β et ne peut donc pas convenir pour tester cette hypothèse.

Les composantes de la variance

Nous allons ici nous concentrer sur les tests impliquant une seule composante de la variance :

$$H_0 : \tau_j = 0 \text{ vs } H_1 : \tau_j > 0.$$

Une généralisation est bien sûr possible mais elle est beaucoup plus complexe et non pertinente dans ce manuscrit. Pour tester cette hypothèse nulle, deux solutions sont possibles avec des distributions adaptées ; le test du score et le test du rapport de vraisemblance.

Le test du score se construit à l'aide de l'une ou l'autre des deux log-vraisemblances exposées précédemment. Si nous regardons pour commencer la vraisemblance classique, la dérivée première en τ_j (Annexe 1) vaut :

$$\frac{\partial \ell}{\partial \tau_j}(\beta, \tau_1, \ldots, \tau_k, \sigma^2) = -\frac{1}{2}\operatorname{Tr}(V^{-1}K_j) + \frac{1}{2}(Y - X\beta)'V^{-1}K_jV^{-1}(Y - X\beta).$$

Étant donné que seule la deuxième partie de cette dérivée dépend de la variable à expliquer Y, nous pouvons limiter le score à ce second terme. Le score vaut ainsi :

$$Q = (Y - X\widehat{\beta}_0)'\widehat{V}_0^{-1}K_j\widehat{V}_0^{-1}(Y - X\widehat{\beta}_0)$$

avec $\widehat{V}_0$ et $\widehat{\beta}_0$, la matrice de variance et les effets fixes estimés sous le modèle nul avec $\tau_j = 0$.

Nous savons que $Y - X\widehat{\beta}_0 = \widehat{V}_0\widehat{P}_0Y$ avec $\widehat{P}_0 = \widehat{V}_0^{-1} - \widehat{V}_0^{-1}X\left(X'\widehat{V}_0^{-1}X\right)^{-1}X'\widehat{V}_0^{-1}$. Ainsi, le score peut se réécrire :

$$\begin{aligned} Q &= Y'\widehat{P}_0\widehat{V}_0\widehat{V}_0^{-1}K_j\widehat{V}_0^{-1}\widehat{V}_0\widehat{P}_0Y \\ &= Y'\widehat{P}_0K_j\widehat{P}_0Y. \end{aligned}$$

Si nous regardons maintenant le même calcul avec la vraisemblance restreinte, nous obtenons :

$$\frac{\partial \ell^{\text{re}}}{\partial \tau_j} = -\frac{1}{2}\operatorname{Tr}(PK_j) + \frac{1}{2}Y'PK_jPY$$

Comme précédemment, seule la deuxième partie de la dérivée première de la log-vraisemblance restreinte dépend de Y. De plus, si nous remplaçons P par son estimation sous H_0, $\widehat{P}_0$, nous obtenons :

$$\frac{1}{2}Y'\widehat{P}_0K_j\widehat{P}_0Y.$$

À une constante multiplicative près, cette expression est celle du score précédent. Les deux log-vraisemblances permettent donc d'aboutir au même test.

Nous cherchons maintenant la distribution de la statistique de test Q sous H_0. Pour cela,

nous pouvons noter que :

$$Y \sim \mathcal{N}_n(X\widehat{\beta}_0, \widehat{V}_0) \qquad \Rightarrow \qquad \widehat{P}_0 Y \sim \mathcal{N}_n(\mathbb{O}_n, \widehat{P}_0\widehat{V}_0\widehat{P}_0)$$

car $\widehat{P}_0 X = \mathbb{O}_{n\times p}$. Ainsi, nous pouvons écrire $\widehat{P}_0 Y$ sous la forme $\widehat{P}_0 Y = \widehat{P}_0 \widehat{V}_0^{\frac{1}{2}} \mathcal{Z}$ avec $\mathcal{Z} \sim \mathcal{N}_n(\mathbb{O}_n, \mathbb{I}_n)$. Ce qui nous donne, pour le score, l'expression :

$$Q = \mathcal{Z}' \left(\widehat{V}_0^{\frac{1}{2}} \widehat{P}_0 K_j \widehat{P}_0 V_0^{\frac{1}{2}} \right) \mathcal{Z}.$$

Nous utilisons alors la décomposition en éléments propres (encadré 1.17) de $\widehat{V}_0^{\frac{1}{2}} \widehat{P}_0 K_j \widehat{P}_0 V_0^{\frac{1}{2}} = U'\Lambda U$ avec U la matrice orthogonale des vecteurs propres et Λ la matrice diagonale des valeurs propres λ_i. Les propriétés de la matrice U nous permettent d'en déduire que :

$$\begin{aligned} Q &= \lambda_1 \mathcal{Z}_1^2 + \ldots + \lambda_n \mathcal{Z}_n^2 \\ \Rightarrow Q &\sim \lambda_1 \chi_1^2 + \ldots + \lambda_n \chi_1^2 \end{aligned}$$

Q suit donc une combinaison linéaire de χ_1^2. Afin d'obtenir une p-valeur à partir de cette distribution, nous pouvons utiliser l'algorithme de Davies [96]. Il est nécessaire de préciser que, ici encore, la variance introduite par les estimations sous le modèle nul n'est pas prise en compte dans nos calculs. Cette distribution est donc une approximation.

La deuxième solution est le test du rapport de vraisemblance. Pour cela, la vraisemblance restreinte ou la vraisemblance classique sont possibles :

$$LRT = 2\ell\left(\widehat{\beta}, \widehat{\tau}_1, \ldots, \widehat{\tau}_k, \widehat{\sigma^2}\right) - 2\ell\left(\widehat{\beta}, \widehat{\tau}_1, \ldots, \widehat{\tau_{j-1}}, 0, \widehat{\tau_{j+1}}, \ldots, \widehat{\tau}_k, \widehat{\sigma^2}\right)$$

ou

$$RELRT = 2\ell^{re}\left(\widehat{\beta}, \widehat{\tau}_1, \ldots, \widehat{\tau}_k, \widehat{\sigma^2}\right) - 2\ell^{re}\left(\widehat{\beta}, \widehat{\tau}_1, \ldots, \widehat{\tau_{j-1}}, 0, \widehat{\tau_{j+1}}, \ldots, \widehat{\tau}_k, \widehat{\sigma^2}\right)$$

La valeur du paramètre d'intérêt τ_j ne pouvant pas être négative, notre test se situe sur le bord du domaine de définition du paramètre. La distribution du test du rapport de vraisemblance n'est donc par un χ_1^2 mais un mélange de χ^2 [97], $\frac{1}{2}\chi^2(0) : \frac{1}{2}\chi^2(1)$. D'autre part, il a été montré que le test du rapport de vraisemblance basé sur les vraisemblances restreintes respecte mieux l'erreur de type I voulue que celui basé sur les vraisemblances classiques [98]. Il est donc préférable d'utiliser les vraisemblances restreintes.

Un cas particulier plus simple est celui où le modèle contient un seul paramètre τ ($k = 1$). Alors, nous voulons tester :

$$H_0 : \tau = 0 \text{ vs } H_1 : \tau > 0.$$

Dans ce cas, nous avons $\widehat{V}_0 = \widehat{\sigma^2}\mathbb{I}_n$ ainsi l'expression du score se simplifie :

$$\begin{aligned} Q &= \frac{1}{\widehat{\sigma^4}}(Y - X\widehat{\beta}_0)' K (Y - X\widehat{\beta}_0) \\ Q &= Y'\widehat{P}_0 K \widehat{P}_0 Y. \end{aligned} \tag{2.10}$$

avec $\widehat{P}_0 = \frac{1}{\widehat{\sigma^2}}\left(\mathbb{I}_n - X\left(X'X\right)^{-1} X'\right)$.

2.3 Le partitionnement de la variance avec le modèle linéaire mixte

Nous avons montré précédemment comment estimer les composantes de la variance sous le modèle linéaire mixte. Nous allons ici regarder comment relier chacun de ces termes à la variance de notre vecteur d'observations Y.

Pour un vecteur d'observations $Y \in \mathbb{R}^n$, nous pouvons calculer la variance empirique qui vaut :

$$\mathrm{ev}(Y) = \frac{1}{n-1}\left(Y'Y - \frac{1}{n}\left(\mathbb{1}_n' Y\right)^2\right)$$

avec $\mathbb{1}_n$ le vecteur colonne de longueur n qui ne contient que des 1.

Nous définissons la forme linéaire $\Psi : \mathbb{R}^{n\times n} \to \mathbb{R}$ par :

$$\Psi(A) = \frac{1}{n-1}\left(\mathrm{Tr}(A) - \frac{1}{n}\mathbb{1}_n' A \mathbb{1}_n\right)$$

pour toutes les matrices (symétriques) $A \in \mathbb{R}^{n\times n}$. Nous avons $\mathrm{ev}(Y) = \Psi(YY')$ pour tout $Y \in \mathbb{R}^n$.

La linéarité de Ψ implique que :

$$\begin{aligned}
\mathbb{E}[\mathrm{ev}(Y)] &= \mathbb{E}[\Psi(YY')] \\
&= \Psi(\mathbb{E}[YY']) \\
&= \Psi(\mathbb{E}[Y]\mathbb{E}[Y]' + \mathrm{Var}(Y)) \\
&= \mathrm{ev}(\mathbb{E}[Y]) + \Psi(\mathrm{Var}(Y))
\end{aligned}$$

Donc dans notre cas, avec $\mathbb{E}[Y] = X\beta$ et $\mathrm{Var}(Y) = V = \tau_1 K_1 + \cdots + \tau_n K_n + \sigma^2 \mathbb{I}_n$, nous obtenons :

$$\begin{aligned}
\mathbb{E}\left[\mathrm{ev}(Y)\right] &= \mathrm{ev}(X\beta) + \Psi(V) \\
&= \mathrm{ev}(X\beta) + \tau_1\Psi(K_1) + \cdots + \tau_n\Psi(K_n) + \sigma^2. \qquad (2.11)
\end{aligned}$$

Le terme $\mathrm{ev}(X\beta)$ dans (2.11) est la variance due aux variables dans X. Le terme $\Psi(V)$ est la variance due aux différents effets aléatoires (variances génétique et résiduelle).

2.3.1 Les estimations non biaisées

La variance due aux facteurs avec des effets aléatoires, $\Psi(V)$, est estimée sans biais en substituant $\widehat{V} = \widehat{\tau}_1 K_1 + \cdots + \widehat{\tau}_k K_k + \widehat{\sigma^2}\mathbb{I}_n$ à V. À l'inverse, en substituant $\widehat{\beta}$ à β dans l'expression de $\mathrm{ev}(X\beta)$, nous obtenons une estimation biaisée. Une solution possible est d'estimer $\mathrm{ev}(X\beta)$ par :

$$\mathrm{ev}(Y) - \Psi\left(\widehat{V}\right).$$

Cependant, comme nous savons que :

$$\mathbb{E}\left(\mathrm{ev}\left(X\widehat{\beta}\right)\right) = \mathrm{ev}(X\beta) + \Psi(\mathrm{Var}(X\widehat{\beta})) = \mathrm{ev}(X\beta) + \Psi\left(X\mathrm{Var}\left(\widehat{\beta}\right)X'\right),$$

une estimation non biaisée de ev$(X\beta)$ est :

$$\text{ev}\left(X\widehat{\beta}\right) - \Psi\left(X\text{var}\left(\widehat{\beta}\right)X'\right).$$

2.4 Que faire des traits binaires ?

Depuis le début de ce chapitre, nous nous consacrons uniquement aux modèles mixtes gaussiens et donc aux traits quantitatifs. Il est possible d'étendre ce modèle aux traits binaires ($y_i = 1$ si l'individu a le trait et $y_i = 0$ sinon) avec le modèle logistique mixte. Ce modèle s'écrit :

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = X_i\beta + \omega_1^i + \ldots + \omega_k^i$$

pour $i = 1 \ldots n$, avec :

- ▷ X_i les covariables pour l'individu i et β les effets fixes,
- ▷ $w_j = (w_j^1, \ldots, w_j^n)' \sim \mathcal{N}_n(0, \tau_j K_j)$ les effets aléatoires,
- ▷ $\pi_i = \mathbb{P}[y_i = 1 | X_i, \omega_1^i, \ldots, \omega_k^i]$,
- ▷ $y_i \in \{0, 1\}$ le statut de l'individu i.

La vraisemblance obtenue pour les traits binaires ou, de façon générale, les traits extraits d'une loi de la famille exponentielle plus complexe que la loi gaussienne, dépend d'une intégrale très lourde à calculer et est donc très difficile à maximiser. Afin d'approcher la valeur de cette vraisemblance, deux approximations sont possibles, l'approximation normale de Laplace et la quadrature gaussienne [99]. Nous allons ici nous concentrer sur l'approximation de Laplace dont dépend la méthode de la « *Penalized Quasi-Likelihood* » [100]. Cette méthode utilise l'approximation normale de Laplace afin de simplifier la vraisemblance et ainsi de permettre les calculs dans un temps raisonnable. Pour le cas binaire, utiliser la *Penalized Quasi-Likelihood* revient à modéliser un vecteur de travail Z à l'aide du modèle linéaire mixte suivant :

$$Z = X\beta + \omega_1 + \ldots + \omega_k + e \tag{2.12}$$

avec les notations et les lois données précédemment ainsi que X la matrice des covariables et $e \sim \mathcal{N}_n(0, W^{-1})$ où W est la matrice diagonale des n valeurs $\pi_i(1-\pi_i)$. Les vecteurs Z et π dont dépend le modèle ne sont pas observés et doivent donc être estimés durant l'algorithme de maximisation. Celui-ci se décompose comme suit :

1. Initialiser les éléments du modèle (2.12), $\beta^{(0)}$ à 0 ou à l'estimation de β sous le modèle logistique classique et les effets génétiques à 0, $\omega_j^{(0)} = 0$ pour $j = 1 \ldots k$.
2. En déduire les valeurs de

$$\pi_i^{(0)} = \text{logit}^{-1}\left(X_i\beta^{(0)} + \omega_1^{i\,(0)} + \ldots + \omega_k^{i\,(0)}\right) = \frac{1}{1 + e^{-X_i\beta^{(0)} - \omega_1^{i\,(0)} - \ldots - \omega_k^{i\,(0)}}}$$

puis

$$Z_i^{(0)} = X_i\beta^{(0)} + \omega_1^{i\,(0)} + \ldots + \omega_k^{i\,(0)} + \frac{y_i - \pi_i^{(0)}}{\pi_i^{(0)}(1-\pi_i^{(0)})}.$$

3. Initialiser les composantes de la variance τ_j à $\mathrm{Var}(Z^{(0)})/k$ puis faire une étape de l'algorithme EM sous le modèle (2.12) pour obtenir $\tau_j^{(0)}$ avec $j = 1 \ldots k$.

4. Faire une étape de maximisation (AIREML) sous le modèle (2.12) à partir de $Z^{(0)}$ et $\pi^{(0)}$ pour mettre à jour les composantes de la variance, $\tau_1^{(1)}, \ldots, \tau_k^{(1)}$.

5. Avec ces nouveaux paramètres, $Z^{(0)}$ et $\pi^{(0)}$, en déduire $\beta^{(1)}, \omega_1^{(1)}, \ldots, \omega_k^{(1)}$.

6. Mettre à jour $\pi^{(1)}$ puis $Z^{(1)}$ avec les mêmes expressions qu'à l'étape 2.

7. Répéter les étapes 4 à 6 jusqu'à obtenir à une itération r

$$2\max\left(\frac{|\beta^{(r)} - \beta^{(r-1)}|}{|\beta^{(r)}| + |\beta^{(r-1)}|}, \frac{|\tau_1^{(r)} - \tau_1^{(r-1)}|}{|\tau_1^{(r)}| + |\tau_1^{(r-1)}|}, \ldots, \frac{|\tau_k^{(r)} - \tau_k^{(r-1)}|}{|\tau_k^{(r)}| + |\tau_k^{(r-1)}|}\right) < \text{ seuil de tolérance.}$$

Cette algorithme de maximisation est celui qu'utilisent Chen *et al.* [101] dans le contexte de l'étude d'association en population. Grâce à cette approximation par un modèle linéaire mixte, nous pouvons calculer un score afin de tester des effets fixes ou aléatoires [20, 101] (paragraphe 2.2.7) :

▷ pour les effets fixes, $H_0 : \beta = \beta_0$ vs $H_1 : \beta \neq \beta_0$, le score s'écrit :

$$\begin{aligned} T &= X'\widehat{P_0}Z \\ &= X'(Y - \widehat{\pi}_0) \end{aligned}$$

où Z, $\widehat{P_0}$ et $\widehat{\pi}_0$ sont les dernières valeurs estimées dans l'algorithme de maximisation décrit plus haut sous l'hypothèse nulle. Sous H_0, ce score a pour variance $\mathrm{Var}(T|H_0) = X'\widehat{P_0}X$ et le test s'écrit :

$$(Y - \widehat{\pi}_0)'X\left(X'\widehat{P_0}X\right)^{-1}X'(Y - \widehat{\pi}_0) \sim \chi_p^2.$$

▷ pour les effets aléatoires, $H_0 : \tau_j = 0$ vs $H_1 : \tau_j > 0$, le score s'écrit :

$$\begin{aligned} Q &= Z'\widehat{P}_0 K_j \widehat{P}_0 Z \\ Q &= (Y - \widehat{\pi}_0)K_j(Y - \widehat{\pi}_0). \end{aligned}$$

Sous H_0, $Q \sim \lambda_1\chi_1^2 + \ldots + \lambda_n\chi_1^2$ avec λ_j les valeurs propres de la matrice $\widehat{V}_0^{\frac{1}{2}}\widehat{P}_0 K_j \widehat{P}_0 V_0^{\frac{1}{2}}$ estimée sous H_0.

Cependant, il a été montré que cette approximation normale n'est pas optimale pour les traits binaires [102, 103]. Une solution plus performante est le calcul de la vraisemblance par une quadrature gaussienne lors de l'algorithme de maximisation (déjà cité précédemment). Malheureusement, cette méthode est nettement plus complexe et peut demander des temps de calcul importants. C'est pourquoi, la *Penalized Quasi-Likelihood* reste tout de même utilisée. En plus de ces deux méthodes, d'autres pratiques sont apparues dans la littérature :

▷ la plus simple, le traitement du trait binaire comme un trait quantitatif. Cette solution n'est pas la meilleure mais elle permet d'utiliser les outils optimisés pour les traits quantitatifs.

▷ le calcul du modèle logistique classique avec uniquement les covariables suivi de l'analyse de ses résidus comme un trait quantitatif (encadré 2.2).

Les résidus du modèle logistique classique

Le modèle logistique s'écrit, pour l'individu i,

$$\text{logit}(\pi_i) = \log(\pi_i/(1-\pi_i)) = X_i\beta$$

où $\pi_i = \mathbb{P}[y_i = 1|X_i]$, y_i est le statut à expliquer dans $\{0,1\}$, X_i la matrice des variables explicatives et β les effets fixes associés.
À l'aide du maximum de vraisemblance, nous estimons β par $\widehat{\beta}$. Ainsi, nous obtenons :

$$\widehat{\pi}_i = 1 \Big/ \left(1 + e^{X_i\widehat{\beta}}\right)$$

Les résidus les plus classiques de ce modèle sont, pour $i = 1 \dots n$,

▷ les résidus simples, $r_i = y_i - \widehat{\pi}_i$,

▷ les résidus de Pearson, $r_i = \frac{y_i - \widehat{\pi}_i}{\sqrt{\widehat{\pi}_i(1-\widehat{\pi}_i)}}$,

▷ les résidus de déviance, $d_i = (-1)^{1-y_i}\sqrt{|2\ln\left(\widehat{\pi}_i^{y_i}(1-\widehat{\pi}_i)^{1-y_i}\right)|}$.

FIGURE 2.2 – Les résidus du modèle logistique classique.

Aucune de ces solutions n'est totalement satisfaisante mais elles permettent tout de même l'analyse de traits binaires tout en restant prudent sur l'interprétation des résultats.

2.5 Et la génétique humaine dans tout ça ?

Nous sommes restés volontairement dans un cas très général dans l'exposé des modèles mixtes. Ces notations sont donc générales et avec un choix adapté des différents éléments, il est possible de faire plusieurs types d'analyse génétique comme le calcul d'héritabilité ou la recherche d'association. Dans le domaine de la génétique, le modèles mixte a, dans un premier temps, été utilisé pour la génétique animale. Il a été en particulier développé pour les études génétiques sur les vaches laitières dans le but de prévoir au mieux la production de lait des descendants de géniteurs donnés [104, 105]. Ce n'est que bien plus tard que le modèle mixte est apparu en génétique humaine et cela ne fait que quelques années qu'il est largement utilisé. Nous allons ici regarder différentes utilisations courantes du modèle linéaire mixte en génétique humaine. Ce travail a fait l'objet d'un article qui est en attente de publication (Annexe 8.2).

2.5.1 Les études de liaison

La première utilisation du modèle linéaire mixte en génétique humaine est dans les études de liaison (section 1.2.2). Dans l'article [106], l'auteur développe le modèle linéaire mixte suivant :

$$Y_i = X_i\beta + g_i + G_i + e_i$$

avec, pour l'individu i, Y_i le phénotype, X_i les covariables, β les coefficients des covariables, g_i l'effet du gène majeur, G_i un effet polygénique et e_i le terme d'erreur. L'une des hypothèses fortes de cette méthode est l'indépendance des différents effets introduits dans le modèle.

Par ailleurs, au locus considéré pour l'analyse de liaison dont le taux de recombinaison avec le gène majeur est noté θ, nous pouvons calculer, pour deux individus i et j, l'espérance de leur état IBD (figure 1.12), π_{ij}, ainsi que la probabilité qu'ils aient un état IBD de 2, $\triangle_{ij}$. Ainsi, sous le modèle précédent, nous pouvons exprimer les covariances entre les phénotypes de deux individus i et j comme

$$\mathrm{Cov}\,(Y_i, Y_j|\pi_{ij}, \triangle_{ij}) = \begin{cases} \sigma_a^2 + \sigma_d^2 + \sigma_G^2 + \sigma_e^2 & \text{si } i = j \\ f(\theta, \pi_{ij})\sigma_a^2 + g(\theta, \triangle_{ij})\sigma_d^2 + \Phi_{ij}\sigma_G^2 & \text{si } i \neq j \end{cases}$$

où σ_a^2 σ_d^2, σ_G^2 et σ_e^2 sont les composantes de la variance additive et dominante pour le gène majeur, polygénique et résiduelle respectivement, Φ_{ij} le coefficient de parenté entre les individus i et j (paragraphe 1.2.4) et f et g des fonctions dépendant du lien de parenté des individus i et j. Il est alors possible d'estimer les paramètres de ce modèle $(\beta, \sigma_a^2, \sigma_d^2, \sigma_G^2, \sigma_e^2, \theta)$ par maximum de vraisemblance. L'absence de liaison est ainsi équivalente à la nullité des paramètres de la variance σ_a^2 et σ_d^2 représentant l'effet additif et de dominance du gène majeur non observé respectivement qui peut être testée à l'aide d'un test du rapport de vraisemblance.

2.5.2 Les études d'association

Dans le cadre des études d'association, les modèles linéaires mixtes peuvent être utilisés de deux façons différentes.

GEMMA et Cie

Dans le premier type de modèle, le(s) marqueur(s) d'intérêt est(sont) inclus dans le modèle avec un effet fixe. Dans ce cas, les effets aléatoires sont utilisés en tant que correction. Ils représentent l'effet polygénique pour des données familiales ou la stratification de population et l'apparentement cryptique pour des données en population.

$$Y = X\beta + g\gamma + \omega + e$$

où X est la matrice des covariables, g le vecteur des génotypes du marqueur d'intérêt, β et γ des effets fixes, et $\omega \sim \mathcal{N}_n(0, \tau K)$ avec K la matrice de corrélation génétique (GRM, encadré 1.18) ou d'apparentement et $e \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_n)$ des effets aléatoires. Le but est alors de tester l'effet du marqueur d'intérêt γ,

$$H_0 : \gamma = 0 \qquad \text{vs} \qquad H_1 : \gamma \neq 0.$$

Ces modèles permettent de tester l'association marqueurs par marqueurs et peuvent donc nécessiter le calcul d'un modèle mixte pour chaque marqueur (environ 500 000 pour une étude GWAS). Cette méthode pose donc d'importants problèmes de temps de calcul. C'est pourquoi de nombreux développements ont été fais pour ce même modèle afin d'améliorer le temps d'analyse.

Nous pouvons les classer ainsi :

- ▷ les méthodes utilisant la décomposition en valeurs propres de la GRM (encadré 1.18) calculée sur le génome entier (EMMA(x) [107,108], GEMMA [109] et FaST-lmm [110]). Afin d'optimiser la correction induite pour les effets aléatoires, des auteurs ont proposé d'exclure le chromosome du marqueur d'intérêt pour calculer la GRM [111].
- ▷ les méthodes en deux étapes. La première étape consiste à calculer les composantes du modèle sans le terme $g\gamma$ puis d'en déduire des résidus et la seconde étape consiste à tester l'association de chaque marqueur avec les résidus précédemment obtenus (GRAMMAR [112,113]).
- ▷ le test du score qui ne nécessite, comme la méthode GRAMMAR, que le calcul du modèle mixte sous l'hypothèse nulle (section 2.2.7). Ce test présente l'avantage de ne pas demander le calcul de résidus. Cette solution est notamment utilisée pour les traits binaires par *Chen et al.* [101].

SKAT et Cie

Dans le second type de modèle, les marqueurs d'intérêt sont inclus dans le modèle avec des effets aléatoires :

$$Y = X\beta + Gu + e$$

où X est la matrice des covariables, G la matrice des génotypes sur la région d'intérêt de taille $(n \times q)$, β des effets fixes, et $u \sim \mathcal{N}_n(0, \tau\mathbb{I}_q)$ et $e \sim \mathcal{N}_n(0, \sigma^2\mathbb{I}_n)$ des effets aléatoires. Leur association est alors testée au travers de la composante de la variance associée. L'absence d'association entre le trait et la région d'intérêt correspond ainsi à la nullité de la composante de la variance τ,

$$H_0 : \tau = 0 \qquad \text{vs} \qquad H_1 : \tau > 0.$$

Dans ce cas, le score est utilisé afin de tester l'association entre un groupe de marqueurs et un trait (paragraphe 2.2.7 avec en particulier l'équation (2.10)). Cette méthode a été développée dans le but de regarder l'association entre un « *pathway* » (un ensemble de régions liées à un domaine précis) avec un trait dichotomique [114]. L'idée a ensuite été reprise par Wu *et al.* [20] puis d'autres méthodes dérivées ont été développées avec des types de données ou buts différents :

- ▷ l'analyse des variants rares, rSKAT [26] et SKAT-O [27],
- ▷ l'analyse des haplotypes, HKAT [115]
- ▷ un test combinant variants communs et variants rares, SKAT-A et SKAT-C [116],
- ▷ l'analyse de données familiales, ASKAT [117] et famSKAT [118]. Dans ce cas, une composante de la variance est introduite dans le modèle afin de prendre en compte les liens familiaux au travers d'un effet polygénique :

  $$Y = X\beta + Gu + \delta + e$$

  avec les mêmes notations que précédemment et $\delta \sim \mathcal{N}_n(0, \varphi 2\Phi)$ où Φ est la matrice de *kinship* (section 1.2.4).

Nous pouvons noter que la majorité de ces méthodes se concentrent sur le modèle dichotomique avec des données cas-témoins.

2.5.3 L'héritabilité restreinte

Nous rappelons que l'héritabilité restreinte d'un trait quantitatif est la proportion de variance de celui-ci expliquée par les facteurs génétiques additifs (section 1.2.5). Elle peut être estimée à l'aide du modèle linéaire mixte

$$Y = X\beta + \omega_1 + \ldots + \omega_k + e$$

avec :

- ▷ X la matrice des covariables,
- ▷ β les effets fixes associés aux covariables,
- ▷ $\omega_j \sim \mathcal{N}_n(0, \tau_j K)$ les effets aléatoires génétiques individuels où $K_j = Z_j Z_j'/(q_j - 1)$ sont les matrices de corrélation génétique entre les individus (encadré 1.18) estimées sur des parties du génome disjointes (les matrices des génotypes Z_j sont normalisées). Le plus souvent, $k = 1$ et la matrice de corrélation génétique est estimée sur tout le génome.
- ▷ $e \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_n)$ le vecteur des erreurs.

Sous ce modèle, la variance τ_j représente la variance génétique des marqueurs introduits dans Z_j et σ^2 la variance résiduelle ou environnementale. L'héritabilité est alors généralement estimée par l'expression suivante [36]

$$h^2 = \frac{\tau_1 + \ldots + \tau_k}{\tau_1 + \ldots + \tau_k + \sigma^2}.$$

Nous pouvons faire deux remarques sur cette estimation de l'héritabilité restreinte :

- ▷ Lorsque nous avons regardé la partition de la variance (section 2.3), nous avons montré que la variance expliquée par les effets génétiques est estimée par $\Psi(\tau_j K_j)$ qui dans notre cas vaut

$$\begin{aligned}\Psi(\tau_j K_j) &= \tau_j \Psi(K_j) \\ &= \frac{\tau_j}{n-1}\left(\mathrm{Tr}(K_j) - \frac{1}{n}\mathbb{1}_n' K_j \mathbb{1}_n\right).\end{aligned}$$

  Étant donné que la diagonale de la GRM K_j comporte des valeurs très proches de 1, nous pouvons dire que $\mathrm{Tr}(K_j) \simeq n$. Ainsi, nous obtenons

$$\Psi(\tau_j K_j) \simeq \tau_j - \tau_j \frac{\mathbb{1}_n' K_j \mathbb{1}_n - n}{n(n-1)}.$$

  Dans l'estimation classique de l'héritabilité, il manque donc une correction. Cependant, nous pouvons remarquer que le terme $(\mathbb{1}_n' K_j \mathbb{1}_n - n)/n(n-1)$ correspond approximativement à la moyenne des éléments de la matrice de corrélation K hors de la diagonale. Or,

cette moyenne est quasiment nulle car les corrélations entre individus sont très faibles. L'estimation de la variance génétique par τ_j est donc satisfaisante.

- ▷ La variance totale est estimée par $\tau_1 + \ldots + \tau_k + \sigma^2$ et ne prend donc pas en compte la variance expliquée par les variables incluses avec un effet fixe dans le modèle. Ce choix est assumé [36] et justifié si nous cherchons à calculer la proportion de variance expliquée par les marqueurs génétiques parmi la variance non-expliquée par les covariables.

Le modèle classique contient une seule matrice de corrélation génétique estimée sur la totalité du génome. Ce modèle revient à n'autoriser qu'une seule échelle des effets des marqueurs du génome entier. Sachant qu'il est possible de mettre plusieurs matrices de corrélation génétique, la question est de savoir comment découper le génome afin d'obtenir la meilleure estimation possible. Une première idée est de découper le génome en fonction des chromosomes, cette idée est d'ailleurs utilisée pour diagnostiquer la présence de stratification de population dans des données en population [119] (paragraphe 3.4.3). Une autre idée, proposée par Yang *et al.* [120], est de classer les marqueurs en fonction de la fréquence de leur allèle alternatif puis d'introduire une composante de la variance différente pour chaque classe. Cette méthode est justifiée par l'hypothèse selon laquelle les variants rares ont des tailles d'effets plus importantes sur les traits complexes.

L'une des limites de ce modèle est qu'il ne prend pas en compte la présence de déséquilibre de liaison entre les marqueurs dans le calcul de la matrice de corrélation génétique entre individus. Des chercheurs se sont donc intéressés à la question et ont développé une méthode de calcul de la matrice de corrélation afin de prendre en compte le DL [121, 122]. Cette nouvelle matrice peut alors être utilisée pour estimer l'héritabilité d'un trait.

Pour l'analyse des traits binaires, il est possible d'utiliser les méthodes d'estimation du modèle mixte pour les traits binaires (section 2.4) ou le « *liability threshold model* » [41, 123]. Cependant, il n'est pas judicieux d'utiliser directement le modèle linéaire mixte. En effet, les individus étant sélectionnés en fonction de leur statut, l'échantillon ne sera pas représentatif de la population et ne convient donc pas à l'estimation de l'héritabilité avec ce modèle.

2.5.4 L'origine parentale

Une autre utilisation du modèle linéaire mixte est l'étude de l'empreinte parentale (paragraphe 1.2.6) à l'aide de données de pedigree [55] :

$$Y = X\beta + g\gamma + p\delta + \omega + e$$

où Y est le vecteur des phénotypes, X est la matrice des covariables, g le vecteur des génotypes au marqueur d'intérêt, p le vecteur des différences du nombre d'allèles alternatifs transmis par la mère et le père pour le marqueur d'intérêt, β, γ et δ des effets fixes, et $\omega \sim \mathcal{N}_n(0, \tau K)$ et $e \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_n)$ des effets aléatoires. Ici, K est la matrice de *kinship* multipliée par 2 reflétant les liens familiaux ou une estimation de celle-ci (GRM) et ω est donc l'effet polygénique. Sous ce modèle, plusieurs tests sont alors possibles :

- ▷ test de l'effet propre du marqueur d'intérêt en présence ou non d'empreinte parentale,

$$H_0 : \gamma = 0 \qquad \text{vs} \qquad H_1 : \gamma \neq 0.$$

▷ test de l'empreinte parentale du marqueur d'intérêt en présence ou non d'un effet propre du marqueur,

$$H_0 : \delta = 0 \qquad \text{vs} \qquad H_1 : \delta \neq 0.$$

▷ test global de l'effet propre et de l'empreinte parentale du marqueur d'intérêt,

$$H_0 : \gamma = 0 \text{ et } \delta = 0 \qquad \text{vs} \qquad H_1 : \gamma \neq 0 \text{ ou } \delta \neq 0.$$

Dans ce modèle, le choix a été fait de modéliser l'empreinte parentale par le vecteur p qui représente la différence entre la transmission de la mère et du père (table 2.1). D'autres modélisations de l'empreinte parentale sont possibles. Par exemple, nous pourrions nous intéresser uniquement à l'empreinte maternelle en regardant uniquement quel est l'allèle transmis par la mère à son enfant.

Allèle transmis par		le père A	le père a
la mère	A	0	-1
la mère	a	1	0

TABLE 2.1 – Valeurs de p en fonction des allèles transmis par les parents à leur enfant pour un SNP di-allélique A/a.

2.6 L'origine parentale et les traits binaires

Le modèle, pour étudier l'empreinte parentale donné précédemment (paragraphe 2.5.4), a été développé pour des traits quantitatifs. Notre idée est donc de faire la même analyse en utilisant une méthode de calcul adaptée aux traits binaires. Pour cela, nous utilisons le score dérivé de la *Penalized Quasi-Likelihood* (section 2.4) permettant de tester des effets fixes dans le modèle logistique mixte suivant, pour $i = 1 \dots n$:

$$\text{logit}\left(\mathbb{P}[y_i = 1 | X_i, g_i, p_i, \omega_i]\right) = X_i\beta + g_i\gamma + p_i\delta + \omega_i$$

où y_i, X_i, g_i et p_i sont, pour l'individu i, le statut, la matrice des covariables, le génotype au marqueur d'intérêt et la différence du nombre d'allèles alternatifs transmis par la mère et le père pour le marqueur d'intérêt (table 2.1), β, γ et δ des effets fixes, et $\omega = (\omega_1, \dots, \omega_n)' \sim \mathcal{N}_n(0, \tau 2\Phi)$ des effets aléatoires avec Φ la matrice de *kinship*.

Nous voulons alors appliquer le score pour tester l'hypothèse :

$$H_0 : \gamma = 0 \text{ et } \delta = 0 \qquad \text{vs} \qquad H_1 : \gamma \neq 0 \text{ ou } \delta \neq 0.$$

Ce test a l'avantage de ne nécessiter le calcul que d'un seul modèle mixte sous l'hypothèse nulle commun à tous les variants analysés. En appliquant le score calculé dans le cas général (section 2.4), nous obtenons la statistique de test suivante :

$$Q = (y - \widehat{\pi}_0)' \begin{pmatrix} g & p \end{pmatrix} \left(\begin{pmatrix} g' \\ p' \end{pmatrix} \widehat{P_0} \begin{pmatrix} g & p \end{pmatrix} \right)^{-1} \begin{pmatrix} g' \\ p' \end{pmatrix} (y - \widehat{\pi}_0) \sim \chi^2_2$$

avec $y = (y_1, \ldots, y_n)'$, $X = (X_1, \ldots, X_n)$, $g = (g_1, \ldots, g_n)'$, $p = (p_1, \ldots, p_n)'$ et $\widehat{\pi}_0$ et $\widehat{P_0}$ définis dans la section 2.4.

Nous avons testé les performances de ce score sur des simulations. Pour cela, nous avons utilisé les haplotypes fournis dans le package R SKAT [124]. Ce jeu de données est composé de 10 000 haplotypes d'une région génétique de 200 kb (kilobase) générés avec le modèle de coalescence implémenté dans COSI [125] qui utilise la structure des ancêtres européens.

Nous avons donc simulé des familles nucléaires en suivant la procédure suivante :

1. les haplotypes des parents de chaque famille sont aléatoirement tirés dans le jeu de données des haplotypes fourni dans le package SKAT.
2. les haplotypes des enfants sont tirés parmi ceux de leurs parents en ignorant la possibilité de recombinaison. Le nombre d'enfants de chaque famille est tiré aléatoirement.
3. un SNP avec une fréquence de l'allèle mineur autour de 1% ou 20% est sélectionné aléatoirement. À partir de celui-ci, les vecteurs des génotypes g et de l'effet de l'origine parentale p sont définis pour chaque enfant.
4. le phénotype y_i de chaque individu est simulé sous le modèle

$$\text{logit}\left(\mathbb{P}[y_i = 1 | g_i, p_i, \omega_i]\right) = \beta_0 + g_i\gamma + p_i\delta + \omega_i$$

   avec $\beta_0 = \log(0.01) - \log(1 - 0.01)$ pour une prévalence du trait autour de 1%, $\omega \sim \mathcal{N}_n(0, \tau 2\Phi)$ où Φ est la matrice de *kinship* calculée en fonction des liens familiaux. Plusieurs valeurs de γ et δ ont été considérées.

Avec cette procédure, nous avons simulé 1000 réplications de 500 familles pour chaque jeu de paramètres (γ, δ) considéré. Le test du score a été appliqué sur chacun de ces jeux de données de 500 familles sans autre sélection, d'une part, en considérant le vecteur p connu et, d'autre part, en le calculant à partir des haplotypes estimés par le programme SHAPEIT version 2.12 [126] depuis les génotypes au locus d'intérêt et de 200 SNPs autour de celui-ci. Les puissances estimées pour un risque de première espèce de 5% sont données dans les figures 2.3 et 2.4 pour des SNPs d'intérêt avec une fréquence de l'allèle mineur de 1% et 20% respectivement en fonction des odds ratios génotypique (e^γ) et de l'empreinte parentale (e^δ).

Pour commencer, lorsque la fréquence de l'allèle mineur du variant d'intérêt est autour de 1% (figure 2.3), la puissance obtenue sous l'hypothèse nulle vaut 4.1% et 4.6% en considérant le vecteur p connu ou en l'estimant respectivement. Ces valeurs sont légèrement au dessous de l'erreur de type 1 attendue. Ces résultats pourraient s'expliquer par le peu de familles informatives dans nos jeux de données simulées ($\sim$ 20 familles informatives par simulation) lorsque nous considérons un variant avec une fréquence de l'allèle mineur aussi faible. En effet, le test du score utilisé reste une approximation pouvant ainsi provoquer des fluctuations dans les puissances obtenues. Ensuite, la puissance augmente avec l'un et l'autre des deux odds ratios comme attendu. Nous avons, par exemple, pour des odds ratios génotypique et de l'empreinte parentale valant 1.2 et 1.5 respectivement, une puissance estimée à 10.5% et 10.8% avec le vecteur p connu et déterminé par phasage respectivement. Les puissances les plus importantes, pour des odds ratios de 2, sont estimées à 29.9% et 28.2%. Ces puissances, relativement élevées étant donné le peu de familles informatives, peuvent s'expliquer par l'accumulation des risques génotypiques et de l'empreinte parentale.

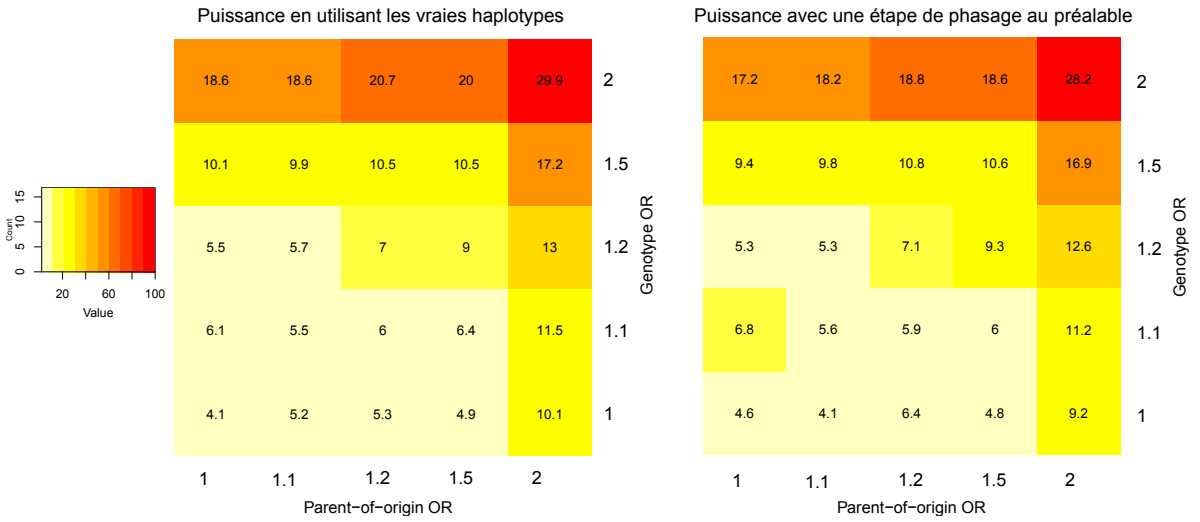


FIGURE 2.3 – Puissance du test du score avec $H_0 : \gamma = 0$ et $\delta = 0$ pour un variant dont la fréquence de l'allèle mineur vaut 1% en fonction des exponentiels des coefficients utilisés pour simuler le trait binaire sous le modèle logistique mixte pour 500 familles et 1000 réplications.

Dans le deuxième cas, lorsque la fréquence de l'allèle mineur du variant d'intérêt est proche de 20% (figure 2.4), l'erreur de type 1 est estimée à 4.7% et 6.2% en considérant le vecteur p connu ou en l'estimant respectivement. Nous constatons donc une inflation de l'erreur de type 1 avec une étape de phasage préalable qui pourrait éventuellement s'expliquer par les erreurs de phasage. En termes de puissance, nous obtenons des valeurs allant de 7.5% à 100%. Entre autre, nous estimons la puissance à 58% et 51.2% avec le vecteur p connu et déterminé par phasage respectivement pour des odds ratios génotypique et de l'empreinte parentale valant 1.2 et 1.5. Ces puissances sont très satisfaisantes et s'expliquent par un nombre plus important de familles informatives que précédemment (~ 300).

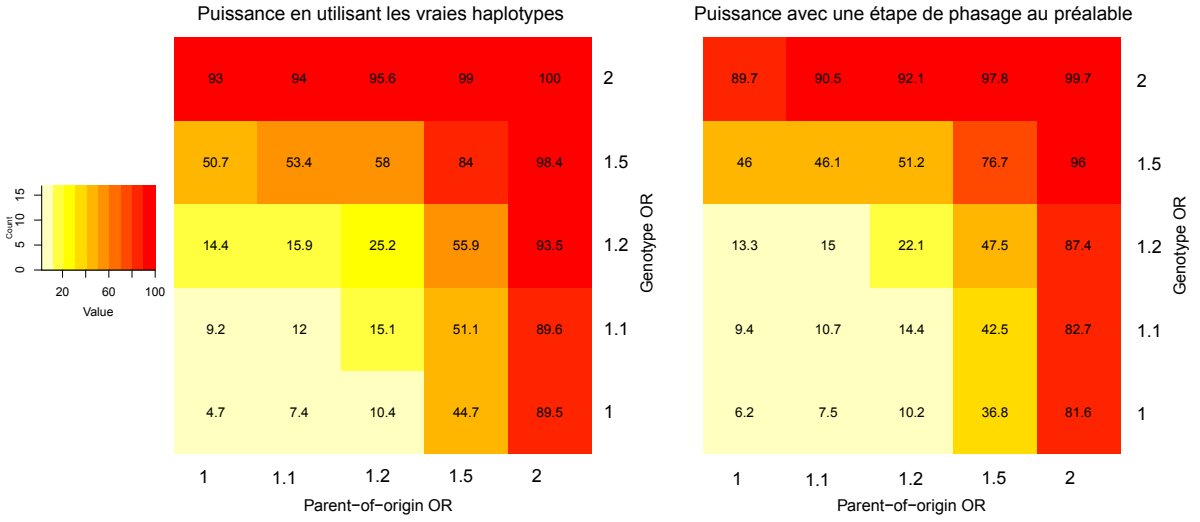


FIGURE 2.4 – Puissance du test du score avec $H_0 : \gamma = 0$ et $\delta = 0$ pour dont la fréquence de l'allèle mineur vaut 20% en fonction des exponentiels des coefficients utilisés pour simuler le trait binaire sous le modèle logistique mixte pour 500 familles et 1000 réplications.

Ces différentes résultats sont très encourageants quant aux performances de ce test. Il est cependant nécessaire de faire d'autres simulations pour appréhender totalement la qualité de la méthode que nous proposons pour tester la présence d'empreinte parentale. En particulier,

il serait intéressant d'inclure dans nos simulations un mode de sélection des familles comme la présence d'au moins un enfant atteint afin de rapprocher d'avantage de la réalité des recrutements.

2.7 Les performances pour la prédiction

Nous avons vu précédemment qu'il était possible à l'aide du modèle linéaire mixte d'estimer les BLUPs des effets aléatoires. Dans le cadre de la génétique, il est donc techniquement réalisable d'estimer les effets de chaque marqueur introduit dans le modèle avec des effets aléatoires et ainsi de faire de la prédiction. Le modèle linéaire mixte a été exploité pour la prédiction de trait binaire par Speed *et al.* [127]. L'idée de leur méthode est de découper le génome en régions définies en fonction de l'amplitude de l'association des marqueurs avec le trait étudié et d'introduire dans le modèle linéaire mixte des effets aléatoires de taille d'effets différente pour chacune de ces régions. Cette méthode permet d'obtenir une qualité de prédiction satisfaisante pour des traits ayant des gènes causaux avec un effet très fort mais obtient des résultats très mitigés pour la prédiction de traits complexes.

Nous nous sommes alors demandés si les modèles linéaires mixtes utilisés pour l'association ou le calcul de l'héritabilité permettent d'obtenir des résultats intéressants pour la prédiction. Pour cela, nous avons regardé les performances de prédiction sur des simulations de deux types de modèles :

- ▷ le modèle considérant uniquement une région du génome (à la SKAT),
- ▷ le modèle considérant la totalité du génome.

2.7.1 La prédiction avec une région du génome

Afin de simuler un trait sous le modèle linéaire mixte de la méthode SKAT (paragraphe 2.5.2), nous avons utilisé les haplotypes fournis dans le package R SKAT [124]. Nos simulations comportent plusieurs étapes :

1. une région génomique ayant la taille voulue (5kb et 20 kb) est aléatoirement sélectionnée parmi le jeu de données. Dans cette région, seuls les variants rares (avec une fréquence de l'allèle alternatif inférieure à 5%) sont conservés.
2. un certain pourcentage (5% et 10% pour une région de 20 et 5 kb respectivement) des variants rares de cette région sont aléatoirement sélectionnés pour être causaux.
3. les effets β_j de chaque variant causal sont fixés à $\beta_j = c\,|\log_{10}(maf_j)|$ avec maf_j la fréquence de l'allèle alternatif pour le variant j et $c = \log(5)/2$. Ces valeurs d'effets sont celles utilisées par les auteurs de la méthode SKAT [26] pour leur propres simulations.
4. les génotypes d'une population de taille voulue ($n = 2\,000$ ou $5\,000$), G, sont aléatoirement tirés dans le jeu de données des haplotypes fourni dans le package SKAT réduit à la région génomique sélectionnée.

5. les phénotypes de chaque individu sont générés avec le modèle :

$$Y = G^{causal}\beta + \epsilon$$

avec G^{causal} la matrice des génotypes aux variants causaux et ϵ un vecteur de bruit gaussien centré. La valeur de la variance du bruit a été choisie pour obtenir une proportion de variance expliquée par les variants causaux α fixée entre 1% et 5%.

Une fois les données simulées, nous avons estimé les effets aléatoires des variants rares u sous le modèle linéaire mixte :

$$Y = Gu + e$$

avec G la matrice des génotypes de tous les variants rares de la région considérée, $u \sim \mathcal{N}_n(0, \tau W)$ où W est la matrice diagonale des poids données à chaque variant $w_j = (1 - maf_j)^{48}$ (ces poids sont ceux suggérés par les auteurs de la méthode SKAT) et $e \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_q)$. Afin de mesurer la précision de la prédiction, deux coefficients de détermination, le plus souvent compris entre 0 et 1, peuvent être calculés :

$$R^2_{\text{gen}} = 1 - \frac{\mathbb{E}\left[(\widehat{Y} - Gu)^2\right]}{\text{Var}(Gu)} \quad \text{et} \quad R^2_{\text{tot}} = 1 - \frac{\mathbb{E}\left[(\widehat{Y} - Y)^2\right]}{\text{Var}(Y)}.$$

Nous pouvons remarquer que $R^2_{\text{tot}} = \alpha R^2_{\text{gen}}$. Donc, nous avons choisi de montrer uniquement les valeurs de R_{gen}. Afin d'obtenir des résultats représentatifs de la population, les coefficients de détermination ont été calculés sur l'ensemble des haplotypes possibles. Les résultats sont donnés dans la figure 2.5. La méthode SKAT ayant été développée pour tester l'association d'une région avec un trait, la puissance du test d'association a également été tracée dans la figure.

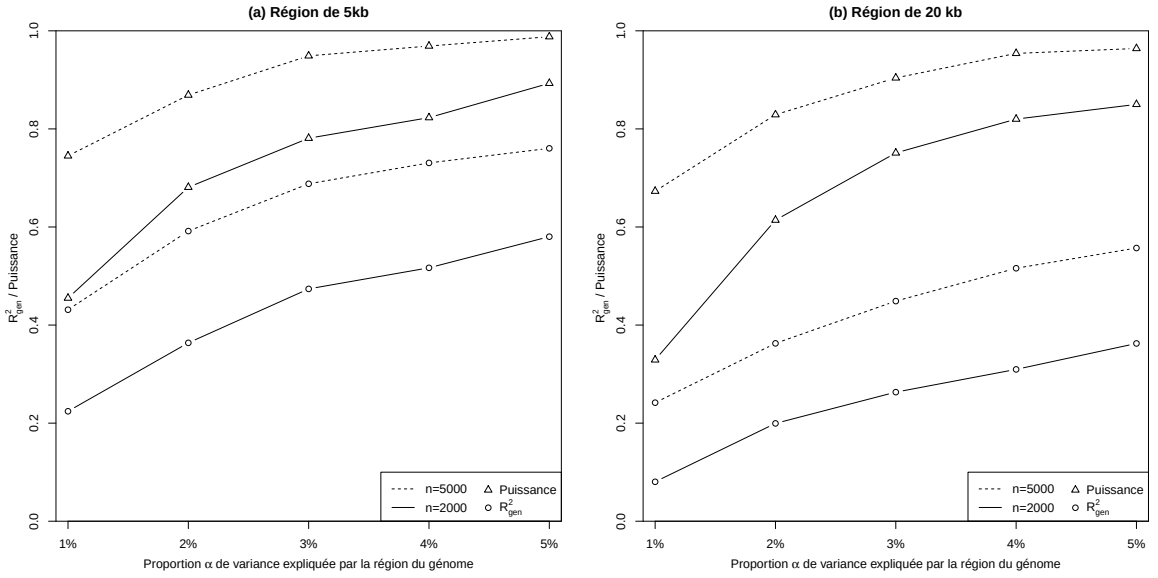


FIGURE 2.5 – Performances du modèle SKAT pour l'association et la prédiction.

Pour commencer, la puissance du test d'association est bonne. Elle varie de 33% à 99%. Ces résultats sont cohérents avec les performances du test d'association SKAT annoncées [26]. La puissance et la qualité de prédiction évoluent dans le même sens. Comme attendu [128], elles

augmentent avec le nombre d'individus utilisés pour estimer les effets des variants et avec la proportion de variance du trait expliquée par les variants causaux. De plus, les performances de ce modèle diminuent avec la taille de la région considérée. Ce résultat peut être expliqué par le fait que l'information sur les effets des variants causaux dans la région est plus « diluée » lorsque la région génomique considérée est plus étendue. La valeur du R^2_{gen} varie entre 20% et 76%. Par exemple, pour une région génomique de 20kb avec $n = 2\,000$ et $\alpha = 2\%$, la puissance du modèle SKAT est de 60% et le coefficient de détermination pour la prédiction est de seulement 20%. Ces résultats ne sont pas très convaincants. De plus, les simulations sont faites sous le modèle linéaire mixte proposé par SKAT [26]. Nous pouvons donc penser qu'ils sont optimistes pour une situation réelle.

2.7.2 La prédiction avec le génome entier

Nous allons ici tester la qualité de prédiction dans un échantillon de taille n d'un trait quantitatif défini par le modèle :

$$Y = Zu + \epsilon = \omega + \epsilon$$

avec Z la matrice des q génotypes normalisés, $u \sim \mathcal{N}_n(0, \frac{\tau}{q-1}\mathbb{I}_n)$ les effets aléatoires des marqueurs ou $\omega \sim \mathcal{N}_n(0, \tau K)$ ($K = ZZ'/(q-1)$ la matrice de corrélation génétique) les effets aléatoires individuels et $\epsilon \sim \mathcal{N}_n(0, \sigma^2\mathbb{I}_n)$.

Pour cela, nous divisons notre échantillon en deux sous-échantillons, un échantillon d'apprentissage de taille n_1 et un échantillon test de taille n_2. Nous notons alors Z_1 et Z_2 les matrices des génotypes normalisés des deux sous-échantillons de taille $n_1 \times q$ et $n_2 \times q$. Dans ce cas, la matrice de corrélation génétique K sur l'ensemble de l'échantillon peut se décomposer de la façon suivante :

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}$$

avec $K_{ij} = Z_i Z_j'/(q-1)$.

À partir des composantes du modèle τ et σ^2, nous pouvons calculer les BLUPs des effets des marqueurs $\widehat{u}$ puis estimer les phénotypes de l'échantillon test Y_2 par :

$$\begin{aligned}
\widehat{Y_2} &= Z_2\widehat{u} \\
&= \frac{\tau}{q} Z_2 Z_1 P Y \\
&= K_{21}\left(K_{11} + \frac{\sigma^2}{\tau}\mathbb{I}_{n_1}\right)^{-1} Y_1 \\
&= K_{21}\left(K_{11} + \frac{1-h^2}{h^2}\mathbb{I}_{n_1}\right)^{-1} Y_1
\end{aligned}$$

avec $h^2 = \dfrac{\tau}{\tau + \sigma^2}$.

Afin dévaluer la qualité de la prédiction sur notre échantillon test, nous définissons les

mêmes coefficients de détermination que précédemment :

$$R^2_{\text{gen}} = 1 - \frac{\mathbb{E}\left[(\widehat{Y}_2 - Z_2 u)^2\right]}{\text{Var}(Z_2 u)} \quad \text{et} \quad R^2_{\text{tot}} = 1 - \frac{\mathbb{E}\left[(\widehat{Y}_2 - Y_2)^2\right]}{\text{Var}(Y_2)} = \frac{\text{Var}(Y_2) - \mathbb{E}\left[(\widehat{Y}_2 - Y_2)^2\right]}{\text{Var}(Y_2)}.$$

et, comme précédemment, nous pouvons remarquer que $R^2_{\text{tot}} = h^2 R^2_{\text{gen}}$.

Étant donné que les phénotypes Y_2 et $\widehat{Y}_2$ sont d'espérance nulle, le coefficient de détermination R^2_{gen} peut être estimé par :

$$R^2_{\text{gen}} = \frac{\tau + \sigma^2}{\tau} \times \frac{\text{Var}(Y_2) - \text{Var}(Y_2) + 2\text{Cov}(Y_2, \widehat{Y}_2) - \text{Var}(\widehat{Y}_2)}{\tau + \sigma^2} = \frac{2Y_2'\widehat{Y_2}/n_2 - \widehat{Y}_2'\widehat{Y_2}/n_2}{\tau}.$$

Si nous n'observons pas les phénotypes de l'échantillon test Y_2, nous pouvons prendre l'espérance du coefficient de détermination. Pour cela, nous devons calculer :

$$\begin{aligned}
\mathbb{E}\left(Y_2'\widehat{Y_2}\right) &= \mathbb{E}\left(\text{Tr}\left(Y_2' K_{21}\left(K_{11} + \frac{1-h^2}{h^2}\mathbb{I}_{n_1}\right)^{-1} Y_1\right)\right) \\
&= \mathbb{E}\left(\text{Tr}\left(K_{21}\left(K_{11} + \frac{1-h^2}{h^2}\mathbb{I}_{n_1}\right)^{-1} Y_1 Y_2'\right)\right) \\
&= \text{Tr}\left(K_{21}\left(K_{11} + \frac{1-h^2}{h^2}\mathbb{I}_{n_1}\right)^{-1} \mathbb{E}\left(Y_1 Y_2'\right)\right) \\
&= \text{Tr}\left(K_{21}\left(K_{11} + \frac{1-h^2}{h^2}\mathbb{I}_{n_1}\right)^{-1} \tau K_{12}\right) \\
&= \tau \text{Tr}\left(K_{21}\left(K_{11} + \frac{1-h^2}{h^2}\mathbb{I}_{n_1}\right)^{-1} K_{12}\right)
\end{aligned}$$

et

$$\begin{aligned}
\mathbb{E}\left(\widehat{Y_2}'\widehat{Y_2}\right) &= \text{Tr}\left(\mathbb{E}\left(\widehat{Y_2}\widehat{Y_2}'\right)\right) \\
&= \text{Tr}\left(\text{Var}(\widehat{Y}_2)\right) \\
&= \text{Tr}\left(K_{21}\left(K_{11} + \frac{1-h^2}{h^2}\mathbb{I}_{n_1}\right)^{-1} \text{Var}(Y_1)\left(K_{11} + \frac{1-h^2}{h^2}\mathbb{I}_{n_1}\right)^{-1} K_{12}\right) \\
&= \tau \text{Tr}\left(K_{21}\left(K_{11} + \frac{1-h^2}{h^2}\mathbb{I}_{n_1}\right)^{-1} K_{12}\right).
\end{aligned}$$

Ainsi, nous obtenons :

$$\mathbb{E}\left(R^2_{\text{gen}}\right) = \frac{1}{n_2}\text{Tr}\left(K_{21}\left(K_{11} + \frac{1-h^2}{h^2}\mathbb{I}_{n_1}\right)^{-1} K_{12}\right)$$

Il est important de noter que ce calcul a été fait en supposant les composantes de la variance connues. Grâce à cette expression, il est possible de calculer l'espérance de R^2_{gen} avec uniquement la GRM K et l'héritabilité h^2.

Il est difficile de simuler des matrices de corrélation génétique reflétant une population non stratifiée ou avec une stratification modérée. Pour regarder la qualité des prédictions sous ce modèle, nous avons donc fait le choix d'utiliser la matrice de corrélation génétique des données de l'étude Trois-Citées après contrôle qualité que nous étudierons plus loin dans ce manuscrit (chapitre 3). Nous disposons ainsi d'une GRM pour 5 793 individus non apparentés. Dans cet échantillon, nous sélectionnons un échantillon d'apprentissage de taille variable n_1 puis le coefficient de détermination est estimé sur tous les individus restants. Les résultats pour différentes valeurs d'héritabilité sont donnés dans la figure 2.6. Malgré le calcul de l'espérance de R^2_{gen} fait sous le modèle linéaire mixte, les valeurs des coefficients de détermination sont vraiment très bas et ne dépassent pas 7.3% même avec une héritabilité de 1. Ces résultats semblent confirmer que les performances du modèle linéaire mixte en termes de prédiction sont très faibles pour des données en population avec des individus non apparentés.

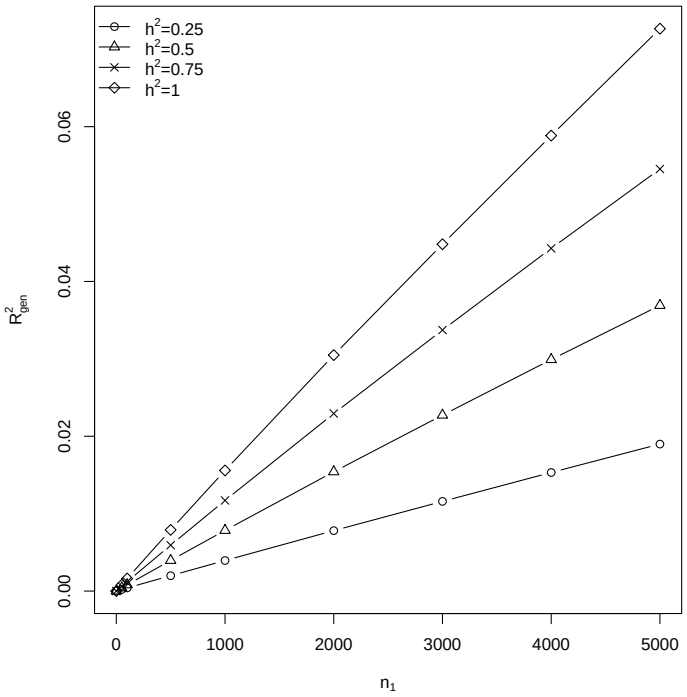


FIGURE 2.6 – Performances du modèle linéaire mixte pour la prédiction.

2.8 Le développement informatique

Dans le but d'avoir des outils permettant la manipulation de données du génome entier et face à la difficulté de trouver un logiciel unique permettant de faire le contrôle qualité, calculer la GRM (encadré 1.18) et analyser, à l'aide de modèles mixtes, des données génétiques, nous avons développé un package intitulé « `gaston` » [129] pour le logiciel de statistique R [1]. Ce package R offre des fonctions manipulant efficacement des grandes matrices de génotypes (SNPs) et permettant ainsi de faire des analyses descriptives et des contrôles qualité. Il contient également l'implémentation d'algorithmes pour les modèles linéaires mixtes couramment utilisés pour estimer l'héritabilité ou réaliser des tests d'association. Grâce aux packages `Rcpp`, `RcppParallel`, `RcppEigen`, les fonctions implémentées dans `gaston` sont principalement écrites en langage C++ permettant un gain de temps et de mémoire. Beaucoup de ces fonctions

1. `https://cran.r-project.org/`

sont également parallélisées pour une efficacité accrue. Le package `gaston` est disponible sur le CRAN avec son code source, un manuel détaillé ainsi qu'une vignette [2]. Dans cette section, nous allons exposer rapidement les possibilités offertes par le package `gaston`.

2.8.1 Le package R `gaston`

Matrices de génotypes

Dans le package, une classe S4 appelée `bed.matrix` a été définie pour les matrices de génotypes. Chaque ligne correspond à un individu et chaque colonne à un SNP. `gaston` peut charger, dans un objet `bed.matrix`, des données génétiques à partir de fichiers :

- ▷ VCF,
- ▷ BED, BIM et FAM (fichiers au standard de PLINK [130]).

Dans le but d'optimiser la manipulation de jeux de données importants, les données sont stockées de façon compacte dans la mémoire (chaque génotype est codé sur 2 bits).

En première approche, un objet `bed.matrix` peut être vu comme une matrice classique contenant uniquement des 0, 1, 2, et NA. En particulier, il est possible de sélectionner une sous-matrice de génotypes en utilisant la syntaxe classique pour les matrices dans l'environnement du logiciel R. Si besoin, un objet `bed.matrix` peut être aussi :

- ▷ converti en une matrice numérique
- ▷ multiplié par un vecteur ou une matrice.

Ces propriétés peuvent être utilisées, par exemple, pour simuler un phénotype quantitatif.

Statistiques descriptives et contrôle qualité

Le package `gaston` permet de calculer des statistiques descriptives des individus et des variants (callrate, fréquence de l'allèle mineur, etc). Il permet également de tester les proportions d'Hardy-Weinberg avec un test du χ^2 ou un test exact [131]. Toutes ces statistiques peuvent notamment servir au contrôle qualité ou permettre de sélectionner un sous-échantillon.

Matrices des génotypes standardisés

Pour les différentes analyses, il est souvent nécessaire de centrer et réduire la matrice des génotypes ; chaque génotype G_{ij} (i est l'index de l'individu et j celui du SNP) est remplacé par :

$$\frac{G_{ij} - \mu_j}{\sigma_j}$$

où $\mu_j = 2p_j$ est la moyenne des génotypes codés 0, 1 et 2 avec p_j la fréquence de l'allèle alternatif et σ_j est la variance empirique des génotypes ou son espérance sous les proportions

2. `cran.r-project.org/web/packages/gaston`

d'Hardy-Weinberg, $\sqrt{2p_j(1-p_j)}$. Le package `gaston` permet d'indiquer si nous voulons ou non centrer et réduire la matrice des génotypes ainsi que les valeurs de l'espérance et de la variance à utiliser.

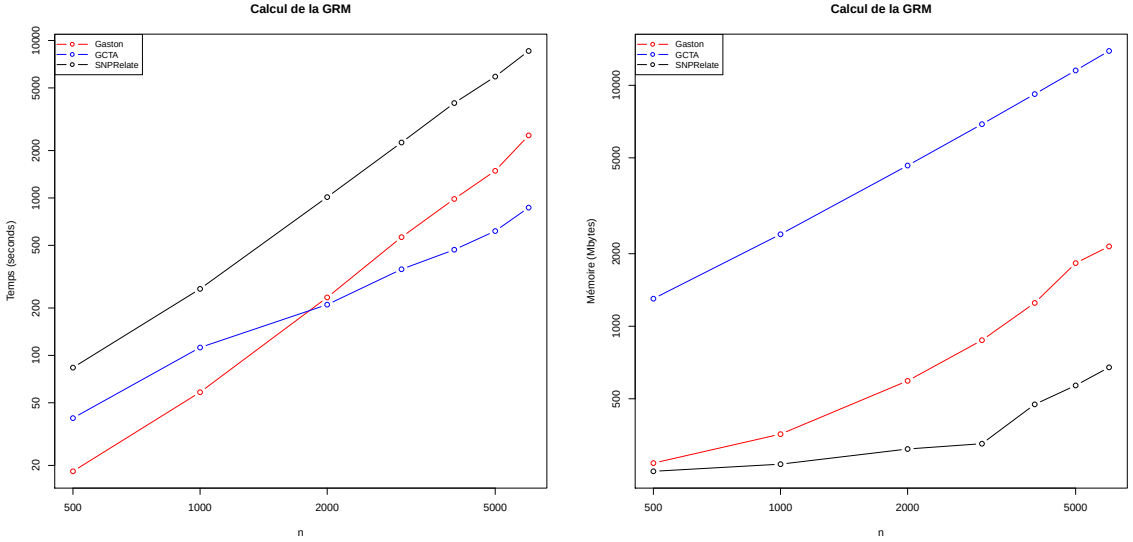
Matrice de corrélation génétique (GRM)

Si Z est une matrice de génotypes centrés réduits $(n \times q)$ (avec q et n le nombre de SNPs et d'individus respectivement), nous avons vu (encadré 1.18) que la matrice de corrélation génétique des individus peut être calculée comme :

$$GRM = ZZ'/(q-1).$$

Le package `gaston` calcule cette matrice. Il est alors possible de calculer les composantes principales (PCs) de la matrice des génotypes et de les utiliser pour détecter des individus aberrants (« *outliers* ») ou pour corriger la structure de population dans les tests d'association. Il est également possible de récupérer les *loadings* associés aux PCs.

Le calcul de la GRM est également fait par un logiciel indépendant GCTA [132] et le package Bioconductor pour le logiciel R `SNPRelate` version [1.6.2] [133]. Nous avons donc comparé les performances de notre package avec ces deux autres programmes en termes de temps de calcul et de mémoire utilisée. Pour les trois programmes, quatre cœurs ont été utilisés et les temps de calcul ont été mesurés sur un CPU Intel Xeon E5. Les performances ont été mesurées pour un jeu de données avec environ 600 000 SNPs et un nombre d'individus variant entre 500 et 6 000. Les résultats sont donnés dans la figure 2.7. Dans cette figure, les temps de calcul et la mémoire utilisée sont représentés avec une échelle logarithmique afin de faciliter la lecture. En termes de temps de calcul, le package R `gaston` obtient les meilleures performances pour des échantillons avec moins de 2 000 sujets. Par la suite, c'est le programme GCTA qui obtient les meilleurs temps. Ceux-ci restent tout de même proches. Avec le package R `SNPRelate`, les



(a) Temps de calcul. (b) Mémoire utilisée.

FIGURE 2.7 – Comparaison des programmes existants pour le calcul de la GRM.

temps de calcul sont nettement plus importants autour de 1 000 secondes par exemple pour un jeu de données de 2 000 individus. En termes de mémoire utilisée, les meilleurs programmes sont, dans l'ordre, le package `SNPRelate`, `gaston` puis pour finir GCTA. Le package `gaston` est alors un bon compromis et le meilleur en termes de temps dans l'environnement du logiciel R.

Déséquilibre de liaison

Le package `gaston` propose aussi d'estimer le déséquilibre de liaison par le produit croisé de la transposée de la matrice des génotypes standardisés :

$$DL = Z'Z/(n-1).$$

Ce calcul peut être fait sur une petite partie du génome pour tracer un graphique, par exemple, ou être utilisé afin d'extraire un jeu de SNPs en faible déséquilibre de liaison (recommandé dans certains cas avant de le calcul la GRM).

Estimations sous le modèle mixte

Le package `gaston` permet d'estimer les modèles linéaires mixtes pour un trait quantitatif comme :

$$Y = X\beta + \omega_1 + \ldots + \omega_k + \varepsilon$$

où X est la matrice de covariables de taille $(n \times p)$, $\omega_j \sim \mathcal{N}_n(0, \tau_j K_j)$ les effets aléatoires génétiques avec $K_j = \frac{1}{q_j} Z_j Z_j'$ la GRM calculée pour la matrice de génotypes Z_j de taille $(n \times q)$ et $\varepsilon \sim N(0, \sigma^2 \mathbb{I}_n)$. Le package `gaston` estime les paramètres du modèle, $\beta, \tau_1, \ldots, \tau_k$ et σ^2, avec l'algorithme AIREML [134] et donne également les BLUPs des effets fixes et aléatoires.

Pour un modèle linéaire mixte avec une seule composante de la variance ($k = 1$), notre package propose une autre méthode pour estimer les paramètres du modèle. Celle-ci utilise la diagonalisation qui est une transformation des données permettant de calculer efficacement la vraisemblance restreinte. Elle nécessite cependant le calcul préalable de la décomposition en vecteurs propres de la matrice de corrélation génétique K, qui peut être long. Les détails de cette méthode sont donnés en Annexe 2.

Les performances du package `gaston` pour l'estimation des paramètres ont été comparées à celles de GCTA [132] et celles des fonctions `VarComp`, `reml` et `lmekin` des packages R `VarComp`, `NAM` et `coxme` respectivement. Les mêmes jeux de données et le même matériel informatique que précédemment, ont été utilisés pour estimer les temps de calcul de ces différents programmes. Les résultats sont donnés dans la figure 2.8 en fonction de la taille d'échantillon avec une échelle logarithmique. Les couleurs rouge, bleue et noire correspondent au package `gaston`, GCTA et les autres packages R respectivement. Pour commencer, les temps de calcul pour les packages R `VarComp`, `NAM` et `coxme` sont plus importants que ceux obtenus avec notre package. Le package `gaston` est aussi légèrement meilleur que GCTA pour l'utilisation de l'algorithme AIREML. La méthode basée sur la diagonalisation tourne très rapidement mais le temps de calcul de la décomposition en vecteurs propres de la GRM est plus important ou équivalent à celui de l'estimation d'un modèle linéaire mixte avec l'algorithme AIREML. L'astuce de la diagonalisation est donc judicieuse uniquement lorsque que nous considérons plusieurs modèles

avec la même GRM. Notre package est donc le meilleur en termes de temps de calcul et propose également la méthode basée sur la diagonalisation pour estimer plusieurs modèles rapidement pour une même GRM.

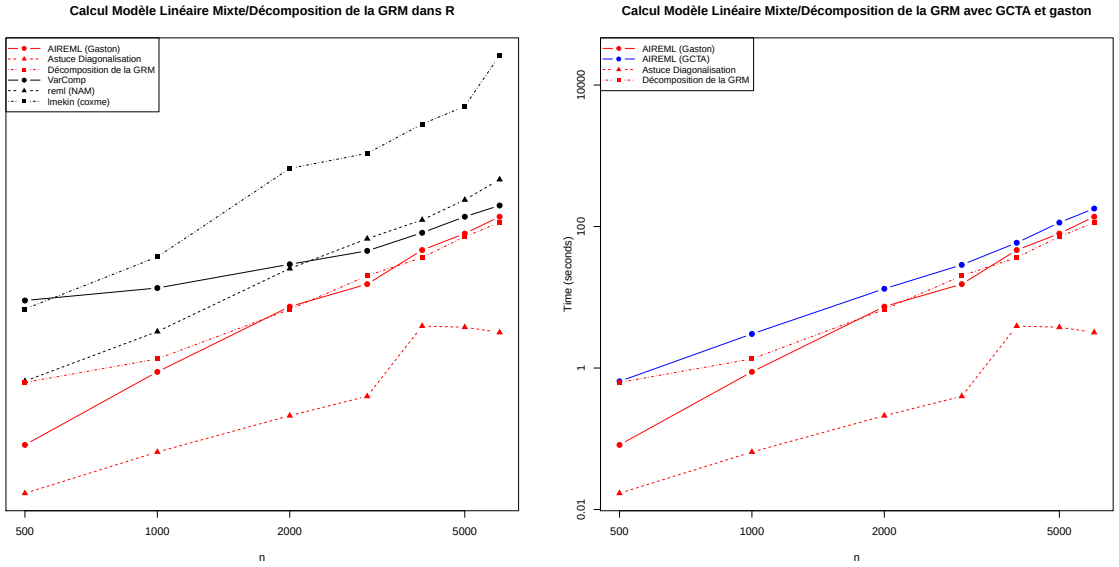


FIGURE 2.8 – Comparaison des programmes existants pour le calcul du modèle linéaire mixte.

Le package R `gaston` permet également de modéliser un modèle logistique mixte avec une ou plusieurs matrices de corrélation génétique :

$$\text{logit}\left(\mathbb{P}[y_i = 1 | X_i, \omega_1^i, \ldots, \omega_k^i]\right) = X_i\beta + \omega_1^i + \ldots + \omega_k^i$$

avec les mêmes notations que précédemment et y_i le statut de l'individu i (0 ou 1). Les paramètres du modèle $\beta, \tau_1, \ldots, \tau_k$ sont estimés par l'algorithme AIREML appliqué à la *Penalized Quasi-Likelihood* (section 2.4). Nous avons vérifié notre programme avec la fonction `glmmPQL` incluse dans la package R `MASS` dans des cas plus simples que celui de la génétique humaine. En termes de temps de calcul, pour un échantillon de 500 individus, notre programme a besoin de 0.284 seconde avec 4 cœurs alors que la fonction `glmmPQL` a besoin de 276.296 secondes avec 64 cœurs (cette fonction utilise obligatoirement la totalité des cœurs disponibles). Nous n'avons donc pas poussé plus loin la comparaison.

Analyses

Grâce à l'estimation des composantes du modèle mixte, `gaston` permet plusieurs types d'analyse pour des phénotypes quantitatifs ou binaires :

- ▷ l'étude des composantes de la variance et l'estimation de l'héritabilité restreinte d'un trait quantitatif,
$$h^2 = \tau/(\tau + \sigma^2).$$
- ▷ la prédiction, à partir des estimations des BLUPs de ω, il est possible d'en déduire les BLUPs des effets aléatoires des SNPs et de s'en servir pour la prédiction.

- ▷ l'analyse GWAS avec une correction de la structure de population ou de l'effet polygénique à la façon de GEMMA (paragraphe 2.5.2). Pour un trait quantitatif, les tests de Wald, du rapport de vraisemblance et du score sont disponibles. Pour les traits binaires, seul le test du score est implémenté.
- ▷ le test du score pour regarder la significativité d'une composante de la variance pour des traits quantitatifs et binaires.
- ▷ le test du score pour tester la nullité de certains effets fixes pour des traits quantitatifs et binaires.

2.8.2 Conclusion et futures extensions

Gaston offre un ensemble de fonctions pour une manipulation efficace de larges matrices de génotypes avec toute la souplesse de R mais également un certain nombre d'analyses basées sur les modèles mixtes. Ses performances en termes de temps de calcul sont également compétitives. Dans de futures versions, nous allons chercher à encore améliorer son efficacité mais aussi inclure d'autres fonctions pour diversifier les analyses génétique possibles.

Chapitre 3

Une application à la population française : l'étude Trois-Cités

Dans le chapitre précédent, nous avons développé la méthodologie du modèle linéaire mixte ainsi que son utilisation dans le domaine de la génétique. Nous allons maintenant appliquer ce modèle sur les données en population de l'étude Trois-Cités [135].

3.1 Les objectifs de notre étude

Dans ce chapitre, nous allons nous concentrer sur l'estimation de l'héritabilité d'un trait quantitatif en appliquant le modèle linéaire mixte suivant à des données en population (paragraphe 2.5.3) :

$$Y = X\beta + \omega + e \tag{3.1}$$

avec :

- ▷ X la matrice des covariables qui peut éventuellement contenir des PCs obtenues à partir des données génétiques,
- ▷ β les effets fixes associés aux covariables,
- ▷ $\omega \sim \mathcal{N}_n(0, \tau K)$ les effets aléatoires génétiques individuels où $K = GG'/(q-1)$ est la matrice de corrélation génétique entre les individus estimée avec les données génétiques sur tout le génome (la matrice des génotypes G est standardisée),
- ▷ $e \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_n)$ le vecteur des erreurs.

Nous avons vu précédemment (paragraphe 1.2.4) que les analyses sur des données en population sont affectées par la stratification de population. Le but de ce chapitre est d'explorer l'effet de l'inclusion des premières PCs avec des effets fixes dans le modèle et de tenter de juger la qualité de la correction de la stratification de population en fonction du nombre de PCs incluses dans le modèle. Pour cela, nous allons utiliser les données de l'étude des Trois-Cités (3C) dont les participants viennent de toute la France. Il paraît alors très intéressant d'explorer la stratification de population à l'échelle d'un pays comme la France qui ne présente aucune stratification a priori exploitable. Une partie de ces résultats a fait l'objet d'une publication [136] (Annexe 8.3).

3.2 Les données

Ce travail est appliqué aux données de l'étude 3C [135] ; une étude de cohorte longitudinale qui a débuté en 1999. Le recrutement s'est effectué entre 1999 et 2001 à partir des listes électorales de trois grandes villes de France ; Dijon, Montpellier et Bordeaux. Les participants ont été sélectionnés aléatoirement sous certaines conditions d'éligibilité :

- ▷ un âge au moins égal à 65 ans,
- ▷ résidence dans l'une des villes de recrutement,
- ▷ aucun soin de longue durée institutionnalisé.

Ces données ont ainsi l'avantage de représenter, du moins en partie, la population française d'une tranche d'âge donnée. Le protocole de recrutement est décrit avec plus de détails dans l'article [135]. De cette façon, 9 294 individus ont été inclus dans l'étude 3C. Le but de cette étude est de suivre les participants pendant au moins 4 ans et de voir s'ils développent une maladie liée à la démence comme la maladie d'Alzheimer. Notre accès aux données se limite aux participants pour lesquels un échantillon de sang a pu être prélevé puis envoyé au Centre National de Génotypage afin d'être génotypé avec une puce Illumina Human610-Quad BeadChip. Nous avons ainsi les données génétiques pour 6 748 individus. Dans notre analyse, les données des pathologies ne sont pas disponibles. Nous ne disposons que de quelques traits anthropométriques et du lieu de naissance. Les variables disponibles seront décrites plus en détails dans la suite de cette section.

3.2.1 Le contrôle qualité des données génétiques

Un contrôle qualité a été fait sur les individus et sur les SNPs en utilisant PLINK [130]. Dans notre analyse, nous ne considérons que les autosomes et les SNPs diallélíques. Un contrôle qualité sur les individus a alors été fait. Tout d'abord, nous n'avons conservé que les individus avec un taux d'hétérozygotie compris entre le taux d'hétérozygotie moyen dans l'échantillon plus ou moins trois fois son écart-type. Nous avons également imposé un pourcentage de valeurs manquantes par individu de moins de 5%. Une vérification de la cohérence entre les chromosomes sexuels et le sexe annoncé a été faite. Les fortes corrélations génétiques ont aussi été regardées. Les paires avec une corrélation génétique estimée sur le génome (paragraphe 2.8.1) supérieure à 0.98 ont été interprétées comme un duplicat et donc supprimées. Pour les individus apparentés (corrélation supérieure à 0.2), un individu par paire tiré au hasard a été enlevé de l'analyse. Les corrélations génétiques ont été calculées à partir de SNPs sélectionnés pour être en faible déséquilibre de liaison. Afin de vérifier l'origine européenne des individus, ceux-ci ont été projetés sur les 10 premières composantes principales obtenues par l'analyse en composantes principales des données Hapmap avec les SNPs déjà sélectionnés. Les individus trop éloignés des populations européennes de Hapmap ont été enlevés de l'analyse. Après concertation avec nos collaborateurs ayant des données plus précises sur les origines des sujets, les individus n'ayant pas le français pour langue maternelle ou n'étant pas nés en France métropolitaine (Corse non comprise) ont été exclus.

Une fois toutes ces étapes effectuées, il nous reste 6 214 individus. Cependant, nous avons ici utilisé un seuil classique de corrélation génétique. Or, il a été conseillé afin d'éviter une

inflation des estimations de l'héritabilité, de garder uniquement les individus avec une corrélation inférieure à 0.025 [119]. Si nous regardons les valeurs des corrélations génétiques entre nos individus (figure 3.1), nous remarquons des valeurs assez importantes.

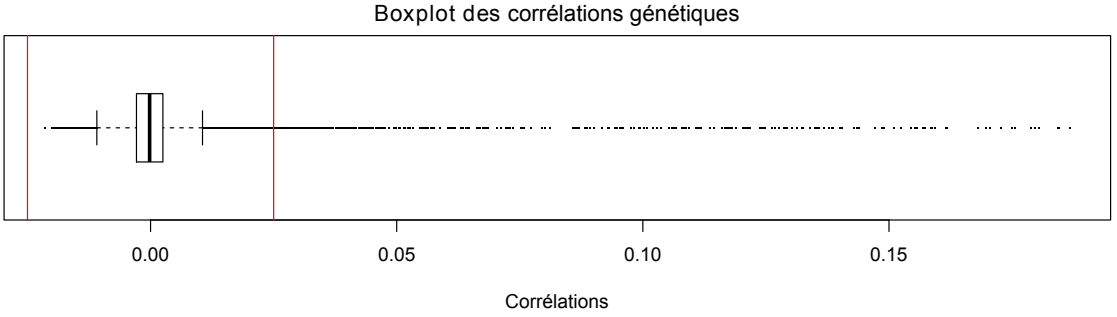


FIGURE 3.1 – Valeurs de la matrice de corrélation génétique.

Nous avons donc exclu certains individus afin de garder des corrélations entre individus assez faibles (inférieures à 0.025) . Au final, 5 793 individus ont été conservés pour l'analyse dont 1 499, 3 676 et 618 ont été recrutés à Bordeaux, Dijon et Montpellier respectivement.

Un contrôle qualité a aussi été appliqué aux SNPs. Nous avons uniquement gardé les SNPs ayant un pourcentage de valeurs manquantes inférieur à 1%. Nous avons testé les proportions d'Hardy-Weinberg et exclu les SNPs ayant obtenu une p-valeur inférieur à 10^{-8}. 509 931 SNPs ont alors été sélectionnés.

Dans la suite de notre analyse, nous avons besoin de calculer des composantes principales afin de corriger la stratification de population en les incorporant dans le modèle avec des effets fixes. Le but de ces PCs est de résumer au mieux le génome. Elles ne doivent donc pas être déterminées principalement par une ou quelques régions du génome. Afin de voir la contribution de chaque SNP aux différentes PCs, nous avons calculé les *loadings* (figure Annexe 3.1 en Annexe 3). Nous remarquons que les premières PCs sont principalement définies par des parties précises du génome comme HLA sur le chromosome 6. C'est pourquoi, pour améliorer la qualité des PCs, nous avons fait un « *pruning* » ou élagué les données avant le calcul des PCs. Pour cela, nous avons choisi d'exclure les SNPs avec une fréquence de l'allèle mineur inférieure à 5%, une p-valeur pour les proportions d'Hardy-Weinberg inférieure à 10^{-5} et un déséquilibre de liaison supérieur à 0.1. Nous avons aussi ôté les SNPs des régions de déséquilibre de liaison étendu connues pour les populations européennes [137]. En effet, ces régions, au nombre de 24, sont trop longues pour être détectées lors de l'élagage des données. Il est donc nécessaire de les exclure séparément. Ainsi, pour les données élaguées, 49 277 SNPs ont été conservés. Les nouvelles composantes principales obtenues ont des *loadings* plus satisfaisants (figure Annexe 3.2 en Annexe 3). En effet, aucune région du génome ne prédomine. Nous avons donc choisi d'utiliser dans nos analyses les PCs calculées à partir des données élaguées [138]. Les résultats obtenus avec les PCs calculées sur les données génétiques non élaguées sont donnés dans l'annexe 4 à titre de comparaison. Pour l'estimation de la composante de la variance, nous avons choisi de garder la totalité des génotypes ayant passé le contrôle qualité dans le but de capturer le maximum d'informations.

3.2.2 Les phénotypes simulés

Afin de tester les performances du modèle linéaire mixte, nous avons simulé des phénotypes à partir des génotypes de l'étude 3C sous deux modèles différents de la forme :

$$Y = X\beta + Zu + e$$

avec :

- ▷ β des effets fixes,
- ▷ Z la matrice de la totalité des génotypes normalisés,
- ▷ $u \sim \mathcal{N}_n(0, \tau \mathbb{I}_q)$,
- ▷ $e \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_n)$,
- ▷ $\tau = 1$ et $\sigma^2 = 1$ (les proportions de variance génétique et résiduelle sont les mêmes).

et X la matrice des covariables qui dépend du modèle :

- ▷ dans le premier modèle, les covariables X sont les 10 premières composantes principales calculées sur les données génétiques élaguées,
- ▷ dans le deuxième modèle, nous avons introduit une seule covariable, la latitude.

Dans les deux cas, les coefficients β ont été choisis pour obtenir une variance expliquée par les effets fixes d'environ 20%. Les variances génétique et résiduelle expliquent donc 40% de la variance totale chacune.

3.2.3 Les phénotypes réels

Plusieurs phénotypes anthropologiques ainsi que les lieux de naissance nous ont été fournis. Nous avons donc analysé les phénotypes suivants :

- ▷ la latitude et la longitude du lieu de naissance (reflétant en partie la stratification de population),
- ▷ la stature connue pour être fortement héritable,
- ▷ le poids, le BMI, la circonférence du crâne et le rapport de la circonférence de la taille sur celle des hanches.

Les statistiques descriptives de chacune de ces variables sont données dans la table 3.1.

Nous pouvons remarqué que l'âge est peu dispersé, l'étude ayant été construite pour recruter uniquement des participants de plus de 65 ans.

Le premier type de variables analysées regroupe la latitude et la longitude du lieu de naissance. Celles-ci ont été choisies car elles reflètent, du moins en partie, la stratification de la population française. En effet, les participants sont nés dans les années 1920 ou 1930 en moyenne.

	Hommes ($N = 2298$)			Femmes ($N = 3495$)			Tous ($N = 5793$)		
Phénotype	Moy	Sd	n	Moy	Sd	n	Moy	Sd	n
Âge	74.15	5.56	2 298	74.39	5.49	3 495	74.30	5.52	5 793
Latitude	46.78	1.73	2 090	46.76	1.63	3 171	46.77	1.67	5 261
Longitude	3.32	2.40	2 090	3.35	2.45	3 171	3.34	2.43	5 261
Stature (cm)	169.58	6.35	2 290	156.60	6.17	3 461	161.77	8.91	5 751
Poids (kg)	75.58	11.27	2 292	62.58	11.32	3 485	67.74	12.97	5 777
BMI	26.27	3.53	2 288	25.52	4.36	3 457	25.82	4.06	5 745
Circonférence crânienne (cm)	57.75	2.05	2 243	55.37	2.07	3 414	56.32	2.37	5 657
Ratio taille sur hanches	0.95	0.07	2 117	0.84	0.07	3 180	0.88	0.09	5 297

TABLE 3.1 – Statistiques descriptives des données 3C (Moy=Moyenne et Sd = Écart-type).

À cette époque, les gens étaient beaucoup moins mobiles qu'aujourd'hui. Nous pouvons donc supposer que le lieu de naissance reflète plutôt bien l'origine des participants. Nous pouvons aussi noter que pour environ 500 individus le lieu de naissance n'est pas renseigné. Le nombre de valeurs disponibles reste néanmoins largement suffisant pour avoir de bonnes estimations. La figure 3.2 représente les lieux de naissance disponibles sur la carte de France. Ces lieux recouvrent une bonne partie de la France.

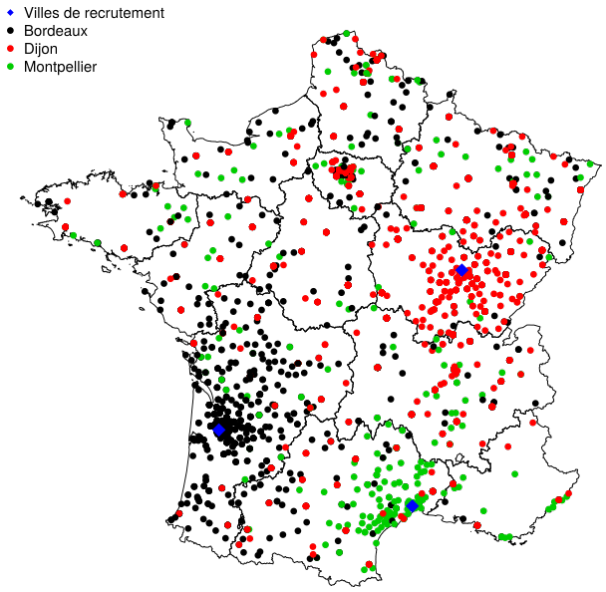


FIGURE 3.2 – Carte de France avec les lieux de naissance. La carte de France a été faite à l'aide du package R `rgdal` avec les coordonnées des limites régionales issues de ©OpenStreetMap contributors.

Le second type de variables analysées est composé de traits anthropométriques. Chacun de ces traits a déjà fait l'objet d'études génétiques et notamment d'estimations d'héritabilité résumées dans la table 3.2. La stature est l'exemple historique connu pour être fortement héritable c'est-à-dire qui est transmis par les parents à leurs enfants. Ce trait a été pour le première fois étudié par Galton et Fisher [30, 139] puis a fait l'objet de beaucoup de GWAS ayant pour but de trouver des variants associés à ce trait qui pourrait expliquer sa forte héritabilité [140–142]. Ensuite, les traits liés à l'adiposité ont aussi été beaucoup étudiés. En effet, des variants ont été trouvés comme associés à différentes mesures de l'adiposité chez l'homme telles que le BMI [143–148]. Pour finir, les études sur les caractéristiques du squelette sont moins courantes mais il semble qu'il y ait tout de même une composante génétique [149].

Pour chaque phénotype, les individus présentant des données manquantes ont été exclus lors de l'analyse.

	Données familiales	Études de jumeaux	Données en population
Stature	0.92 [150]	0.68 to 0.94 [149, 151–153]	0.44 to 0.62 [42, 119–121, 153–155]
Poids	-	0.37 [153]	0.19 [119], 0.26 [153]
BMI	0.24 to 0.81 (mean 0.46) [156]	0.47 to 0.90 (mean 0.75) [156]	0.16 to 0.27 [119, 120, 153, 155]
		0.28 [153]	
		0.45 to 0.84 [157]	
Circonférence des hanches	-	0.15 [153]	0.16 [153], 0.17 (men or women)† [155]
Circonférence de la taille	-	-	0.23 (men)†, 0.19 (women)† [155]
Ratio taille sur hanches	-	-	0.16 (men)†, 0.18 (women)† [155]
Circonférence crânienne	0.66 [158]	0.75 [157]	-
Traits du squelette			
(dont la circonférence crânienne)	-	0.59 [149]	-

† Résultats avec ajustement sur le BMI et stratifié sur le sexe. La référence [156] est une méta-analyse sur 88 études de jumeaux et 27 études de familles.

Table 3.2 – Estimations de l'héritabilité dans la littérature.

3.3 Les phénotypes simulés

Pour commencer, nous regardons le comportement des estimations à l'aide de phénotypes simulés. Pour cela, 100 réplications ont été faites. Les résultats sont donnés sous forme de graphiques représentant les proportions de variance expliquées par les différentes composantes du modèle en fonction du nombre de PCs incluses dans le modèle avec des effets fixes (section 2.3). Les proportions de variance expliquées par les effets fixes, la composante génétique et les résidus (environnement) sont représentées en gris foncé, gris clair et blanc respectivement. Dans un souci de lisibilité, une échelle logarithmique a été choisie pour l'axe des abscisses (nombre de PCs). Ce graphique est commun à toutes les analyses faites dans ce chapitre. Pour les simulations, en particulier, les valeurs tracées correspondent aux moyennes sur la totalité des simulations. Les lignes supplémentaires en pointillé correspondent aux écart-types des estimations.

Nous commençons par regarder les performances du modèle linéaire mixte pour des phénotypes simulés sous le premier modèle qui contient les 10 premières PCs en covariable. Les résultats sont donnés dans la figure 3.3. En l'absence de correction de la stratification de population, la proportion de variance génétique est estimée en moyenne à 100%. Nous avons une forte inflation de l'estimation de la variance génétique et donc de l'héritabilité. Avec l'inclusion des premières PCs avec des effets fixes dans le modèle, cette inflation diminue jusqu'à atteindre les bonnes valeurs avec au moins 10 PCs, le nombre de PCs utilisées pour les simulations. Nous avons inclus jusqu'à 2 000 PCs dans le modèle afin de vérifier la stabilité des estimations des composantes de la variance. Nous constatons que la variance des estimations n'augmente pas beaucoup même avec un nombre important de PCs en effets fixes.

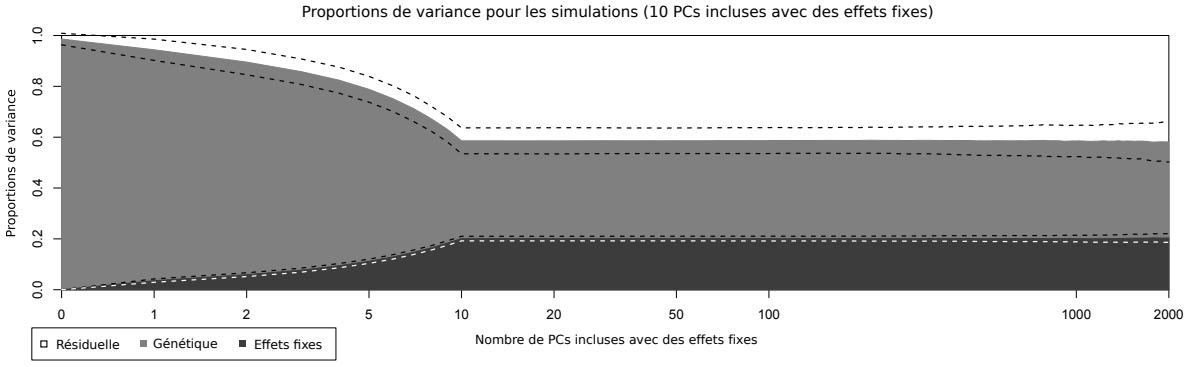


FIGURE 3.3 – Proportions de variance estimées pour des phénotypes simulés en fonction du nombre de PCs incluses dans le modèle (échelle logarithmique). Le blanc, le gris clair et le gris foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

Regardons maintenant les résultats obtenus pour les simulations faites sous le deuxième modèle avec la latitude en covariable (figure 3.4). Sans correction de la stratification de population, la proportion de variance expliquée par la composante génétique est estimée autour de 82%, ce qui est le double de la vraie valeur. L'inclusion des deux premières PC explique 7% de la variance du trait et permet ainsi de baisser l'estimation de la proportion de variance génétique à 49% ce qui reste au dessus de la valeur attendue. Les 9% en excès peuvent être expliqués par l'utilisation de la latitude pour la simulation du trait. En effet, comme nous le constaterons plus loin, la composante génétique explique 44% de la variance de la latitude lorsque les deux premières PCs sont incluses dans le modèle avec des effets fixes. La latitude

contribuant à un cinquième du phénotype simulé, il n'est donc pas surprenant que cet excès de 9% corresponde approximativement à un cinquième de 44%. Ce constat est rassurant quant au comportement des estimations du modèle linéaire mixte. De plus, nous pouvons aussi conclure que les premières PCs ne suffisent pas à retrouver toute la variance expliquée par la latitude. La variance due aux effets fixes est donc sous-estimée. Les estimations restent ici aussi stables même avec un grand nombre de PCs incluses dans le modèle.

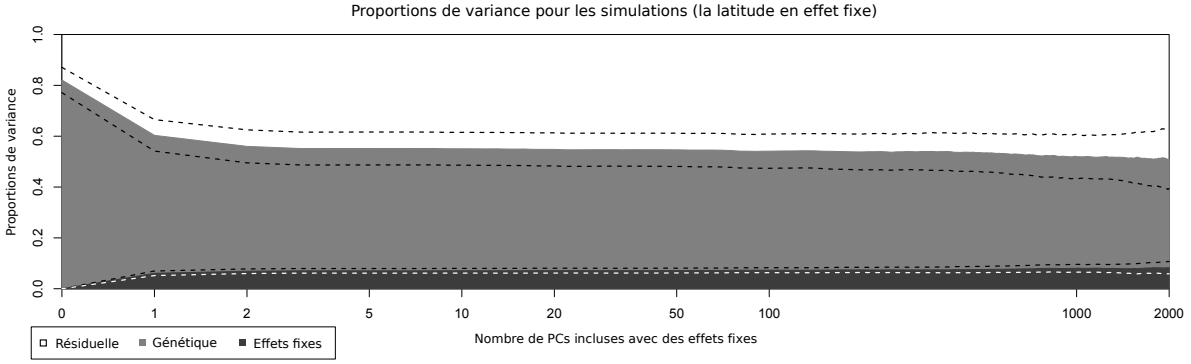


FIGURE 3.4 – Proportions de variance estimées pour des phénotypes simulés en fonction du nombre de PCs incluses dans le modèle (échelle logarithmique). Le blanc, le gris clair et le gris foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

3.4 Les coordonnées géographiques

Dans cette section, nous allons maintenant analyser les coordonnées géographiques des lieux de naissance des participants de l'étude 3C. Le but ici est de voir comment évoluent les estimations des variances et de l'héritabilité de ces traits liés à la stratification de population.

3.4.1 La latitude

Nous commençons par regarder les résultats obtenus pour la latitude des lieux de naissance donnés dans la figure 3.5. En l'absence de correction de la stratification de population, la variance est estimée comme étant entièrement génétique. Par la suite, la variance génétique diminue au profit de la variance environnementale et des effets fixes. Avec les premières PCs, la

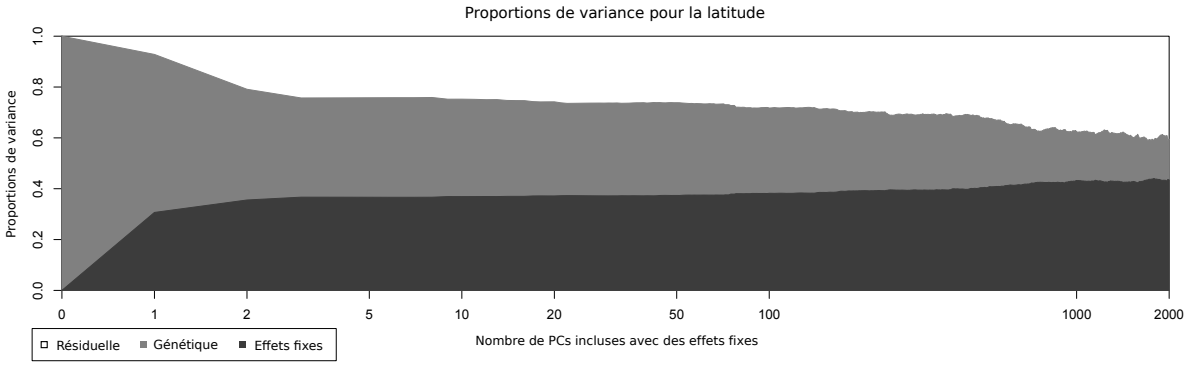


FIGURE 3.5 – Proportions de variance estimées pour la latitude en fonction du nombre de PCs incluses dans le modèle (échelle logarithmique). Le blanc, le gris clair et le gris foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

variance génétique diminue de façon importante (de 100% à environ 40%). En effet, la première PC explique 31% de la variance de la latitude. Avec l'inclusion des PCs suivantes, l'estimation de la variance génétique continue de diminuer lentement ; elle atteint environ 36% pour 50 PCs et 16% pour 2 000 PCs.

Plus de détails sur les estimations des variances ainsi que leurs erreur-types sont donnés dans la table 3.3. Cette table contient toutes les estimations des composantes de la variance ainsi que l'estimation de l'héritabilité accompagnées de leur erreur-type. Elle contient également les résultats du test de rapport de vraisemblance pour $H_0 : h^2 = 0$. L'héritabilité calculée ici est celle utilisée par Visscher (section 2.5.3) et ne prend donc pas en compte la variance expliquée par les effets fixes. Nous remarquons que cette héritabilité reste significative même avec 2 000 PCs incluses dans le modèle avec des effets fixes. Ce résultat est assez troublant et semble indiquer que les PCs ne sont pas suffisantes pour corriger la stratification de population d'un trait comme la latitude.

	LRT	p-valeur	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma^2_P}$ (se)	$\widehat{h^2}$ (se)
0 PC	1854.15	**<1e-40**	2.05 (0.040)	0.00014 (0.039)	2.05 (0.040)	1.000
1 PC	295.37	**<1e-40**	1.60 (0.035)	0.18 (0.031)	1.78 (0.035)	0.897 (0.054)
2 PCs	126.24	**1.4e-29**	1.14 (0.033)	0.55 (0.031)	1.69 (0.033)	0.676 (0.061)
3 PCs	98.87	**1.3e-23**	1.03 (0.033)	0.64 (0.031)	1.66 (0.032)	0.616 (0.063)
4 PCs	99.13	**1.2e-23**	1.03 (0.033)	0.64 (0.031)	1.67 (0.032)	0.617 (0.063)
5 PCs	99.37	**1.0e-23**	1.03 (0.033)	0.64 (0.031)	1.67 (0.032)	0.618 (0.063)
10 PCs	95.50	**7.4e-23**	1.00 (0.033)	0.65 (0.031)	1.66 (0.032)	0.608 (0.063)
20 PCs	88.35	**2.7e-21**	0.98 (0.033)	0.68 (0.031)	1.66 (0.032)	0.589 (0.063)
50 PCs	83.73	**2.8e-20**	0.96 (0.033)	0.69 (0.031)	1.65 (0.032)	0.582 (0.064)
100 PCs	68.81	**5.4e-17**	0.89 (0.033)	0.75 (0.031)	1.64 (0.032)	0.543 (0.066)
500 PCs	38.96	**2.2e-10**	0.73 (0.034)	0.86 (0.031)	1.59 (0.033)	0.459 (0.074)
1 000 PCs	15.26	**4.7e-5**	0.52 (0.035)	1.02 (0.032)	1.54 (0.033)	0.338 (0.086)
2 000 PCs	5.22	**0.0112**	0.44 (0.042)	1.11 (0.037)	1.55 (0.038)	0.284 (0.122)

TABLE 3.3 – Estimations des paramètres du modèle pour la latitude et leur erreur-type (se) en fonction du nombre de PCs incluses dans le modèle : test du rapport de vraisemblance (LRT) et la p-valeur associée, $\widehat{\tau}$, $\widehat{\sigma^2}$ et $\widehat{\sigma^2_P} = \widehat{\tau} + \widehat{\sigma^2}$ l'estimation des variances génétique, résiduelle et totale respectivement et $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ l'héritabilité estimée.

3.4.2 La longitude

Les mêmes analyses ont été faites pour la longitude. Les résultats obtenus sont résumés dans la figure 3.6. Ici aussi, sans stratification de population, la proportion de variance génétique est estimée à 100%. L'ajout des deux premières PCs provoque une forte baisse de la variance génétique estimée à 68% au profit uniquement des effets fixes. Les deux premières PCs expliquent donc 32% de la variance de la longitude. Puis, la variance génétique continue de diminuer doucement au profit, cette fois-ci, des variances environnementale et des effets fixes. Avec 50 et 2 000 PCs incluses dans le modèle, la variance génétique est estimée à 54% et 26% respectivement.

Comme précédemment, plus de détails sur les estimations sont donnés dans la table 3.6. La première chose que nous pouvons noter est que les héritabilités sont estimées à 1 lorsque aucune, une ou deux PCs sont incluses dans le modèle. En effet, comme nous l'avons déjà

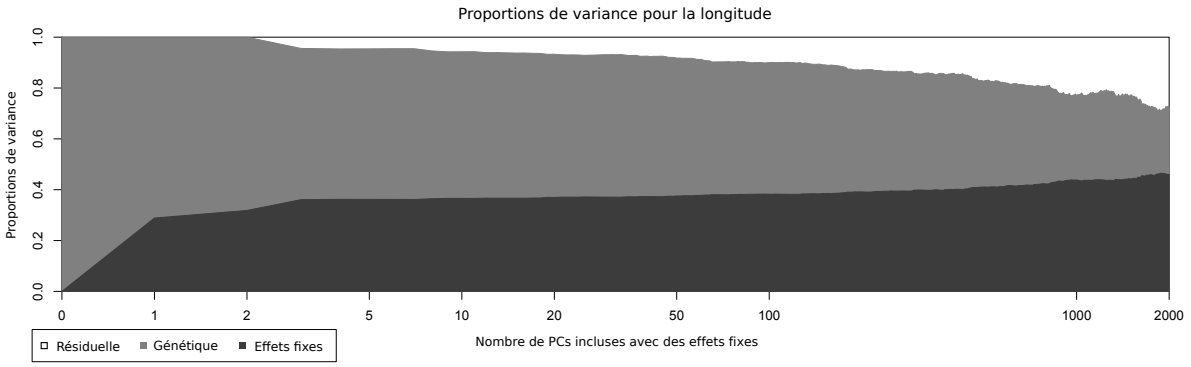


FIGURE 3.6 – Proportions de variance estimées pour la longitude en fonction du nombre de PCs incluses dans le modèle (échelle logarithmique). Le blanc, le gris clair et le gris foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

noté plus haut, l'ajout des deux premières PCs diminue la variance génétique uniquement au profit de la variance des effets fixes, ce qui n'affecte pas les estimations de l'héritabilité car celles-ci ne tiennent pas compte de la proportion de variance expliquée par les effets fixes. La totalité des héritabilités calculées est significativement non nulle. Même avec 2 000 PCs incluses dans le modèle, l'héritabilité est estimée à 48% avec une erreur-type de 12% et est donc très significative. Ces résultats sont similaires à ceux obtenus avec la latitude et semblent ici aussi indiquer que les PCs ne sont pas suffisantes pour corriger la stratification de population de la longitude.

	LRT	p-valeur	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma^2_P}$ (se)	$\widehat{h^2}$ (se)
0 PC	1892.63	**<1e-40**	4.34 (0.085)	0.00029 (0.089)	4.34 (0.084)	1.000
1 PC	563.65	**<1e-40**	3.83 (0.075)	0.00025 (0.069)	3.83 (0.075)	1.000
2 PCs	405.08	**<1e-40**	3.71 (0.072)	0.00025 (0.065)	3.71 (0.072)	1.000
3 PCs	221.81	**<1e-40**	3.28 (0.069)	0.24 (0.061)	3.53 (0.069)	0.931 (0.060)
4 PCs	219.54	**<1e-40**	3.27 (0.069)	0.26 (0.061)	3.52 (0.069)	0.928 (0.060)
5 PCs	219.56	**<1e-40**	3.27 (0.069)	0.25 (0.061)	3.52 (0.069)	0.928 (0.060)
10 PCs	204.70	**<1e-40**	3.19 (0.069)	0.32 (0.061)	3.50 (0.068)	0.910 (0.061)
20 PCs	193.57	**<1e-40**	3.12 (0.069)	0.37 (0.061)	3.49 (0.068)	0.894 (0.062)
50 PCs	177.83	**<1e-40**	3.00 (0.068)	0.46 (0.061)	3.46 (0.068)	0.868 (0.063)
100 PCs	157.09	**2.4e-36**	2.87 (0.068)	0.56 (0.061)	3.43 (0.067)	0.838 (0.065)
500 PCs	87.59	**4.0e-21**	2.35 (0.070)	0.97 (0.063)	3.32 (0.068)	0.707 (0.073)
1 000 PCs	43.24	**2.4e-11**	1.90 (0.072)	1.30 (0.065)	3.20 (0.069)	0.594 (0.087)
2 000 PCs	14.00	**9.1e-5**	1.51 (0.083)	1.61 (0.072)	3.12 (0.077)	0.485 (0.124)

TABLE 3.4 – Estimations des paramètres du modèle pour la longitude et leur erreur-type (se) en fonction du nombre de PCs incluses dans le modèle : test du rapport de vraisemblance (LRT) et la p-valeur associée, $\widehat{\tau}$, $\widehat{\sigma^2}$ et $\widehat{\sigma^2_P} = \widehat{\tau} + \widehat{\sigma^2}$ l'estimation des variances génétique, résiduelle et totale respectivement et $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ l'héritabilité estimée.

3.4.3 Comment interpréter ces résultats ?

Il ne paraît pas si étonnant que, sans correction de la stratification de population, la variance soit estimée comme étant entièrement génétique. En effet, il a été montré sur des données de population de l'Europe que nous pouvons retrouver la géographie de celle-ci avec les premières PCs de la matrice des génotypes [159]. Les données génétiques peuvent donc expliquer en partie

les coordonnées géographiques. Le résultat plus étonnant se situe après l'inclusion des deux premières PCs ; la proportion de variance expliquée par la génétique diminue très lentement avec l'ajout des PCs suivantes dans le modèle. Si nous regardons les corrélations des premières PCs avec les coordonnées géographiques (figure 3.7), nous remarquons que les corrélations avec les trois premières PCs sont fortes contrairement aux autres. Cela explique en partie la forme des courbes obtenues précédemment. Nous allons maintenant faire des analyses complémentaires afin de mieux comprendre nos résultats.

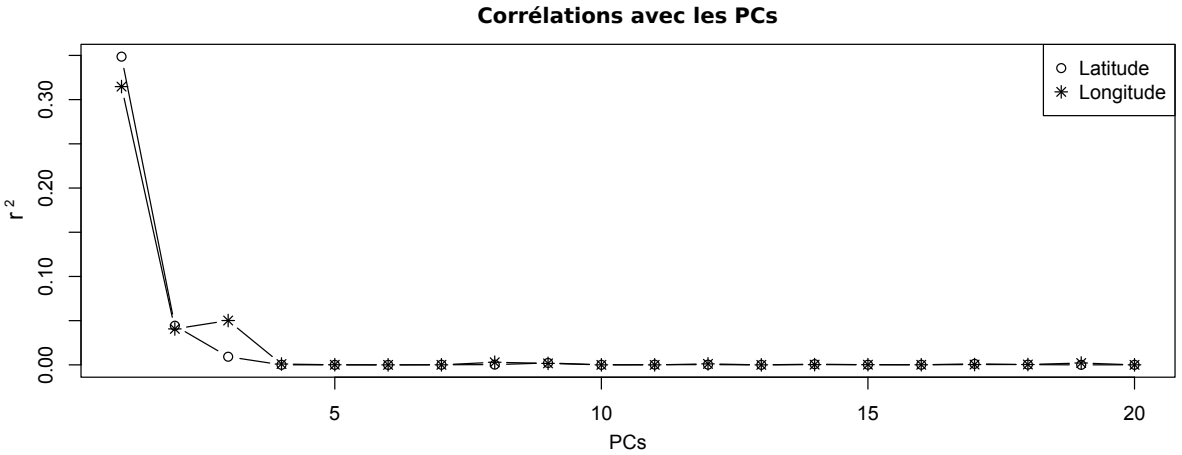


FIGURE 3.7 – Corrélations entre les 20 premières PCs et les coordonnées géographiques.

Les coordonnées géographiques permutées

La première vérification que nous avons effectuée est la même analyse après permutations des valeurs des coordonnées géographiques. Les résultats sont donnés dans la table 3.5. Nous constatons que les proportions de variance génétique et des effets fixes sont estimées approximativement à 0, ce qui nous conforte dans le fait que les résultats obtenus ne sont pas dus au hasard. En effet, cela nous permet d'exclure la possibilité que les résultats de nos analyses soient expliqués par la distribution des coordonnées géographiques.

Latitude				Longitude			
	Proportions de variance				Proportions de variance		
PCs	Effets Fixes	Génétique	Résiduelle	PCs	Effets Fixes	Génétique	Résiduelle
0	5.1e-11	6.6e-5	1.00	0	0	6.6e-5	1.00
5	0	6.6e-5	1.00	5	0	6.6e-5	1.00
10	0	6.6e-5	1.00	10	7.0e-5	6.6e-5	1.00
20	0	6.6e-5	1.00	20	1.1e-3	6.6e-5	1.00
100	0	6.6e-5	1.00	100	1.3e-4	6.6e-5	1.00
500	0	0.017	0.98	500	0	6.6e-5	1.00
1000	0	6.7e-5	1.00	1000	0	6.7e-5	1.00
2 000	3.2e-3	0.021	0.98	2 000	3.2e-3	6.8e-5	1.00

TABLE 3.5 – Proportions de variance estimées pour les coordonnées géographiques permutées.

La détection de la stratification de population

Il y a quelques temps, Yang *et al.* ont proposé une méthode pour détecter la structure de population [119]. Pour cela, ils utilisent le modèle linéaire mixte avec une matrice de corrélation génétique par chromosome autosomal. Nous avons donc des magnitudes d'effets aléatoires différents pour chaque chromosome :

$$Y = X\beta + \omega_1 + \cdots + \omega_{22} + e$$

avec

$$\omega_1 \sim \mathcal{N}_n(0, \tau_1 K_1),\ \ldots,\ \omega_{22} \sim \mathcal{N}_n(0, \tau_{22} K_{22}),\ e \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_n).$$

où K_i est la matrice de corrélation génétique pour la $i^{\text{ième}}$ paire de chromosomes, n le nombre d'individus, Y le vecteur des phénotypes et X la matrice des covariables contenant éventuellement des PCs de la matrice de corrélation génétique estimée sur les données élaguées de la totalité du génome afin de corriger la stratification de population. Sous ce modèle, l'héritabilité restreinte pour le chromosome j est estimée par :

$$h^2_{tot,j} = \frac{\tau_j}{\tau_1 + \ldots + \tau_{22} + \sigma^2}.$$

La méthode proposée dans l'article consiste à calculer les variances génétiques sous le modèle ci-dessus mais aussi sous le modèle avec une seule des 22 matrices de corrélation génétique dans le modèle :

$$Y = X\beta + \omega_j + e$$

dont nous déduisons une autre estimation de l'héritabilité pour le chromosome j :

$$h^2_j = \frac{\tau_j}{\tau_j + \sigma^2}.$$

Nous avons au final 23 modèles donnant chacun des estimations des composantes de la variance et donc de l'héritabilité, $h^2_{tot,1}, \ldots, h^2_{tot,22}$ et $h^2_1, \ldots, h^2_{22}$. Afin de détecter la présence de stratification de population, les auteurs calculent les différences entre l'héritabilité estimée avec le modèle simple et celle estimée avec le modèle complet, $h^2_j - h^2_{tot,j}$ puis font la régression des valeurs obtenues sur la taille des chromosomes. Cette régression permet alors de diagnostiquer la présence de structure de population. Tout d'abord, les auteurs ont montré sur leurs données que l'élimination des individus apparentés fait tendre l'intercept de la régression vers 0 et en ont déduit qu'un intercept significatif peut indiquer la présence d'apparentement cryptique. De plus, si la pente de cette régression est significative, cela indique la présence de stratification de population. En effet, sous l'hypothèse que les variants dont la distribution diffère au travers des populations, sont répartis aléatoirement sur le génome, un chromosome plus long contiendra un plus grand nombre de ces variants informatifs. Ainsi, si ces variants affectent les estimations de l'héritabilité, celle-ci augmentera avec la taille du chromosome.

Nous avons appliqué cette méthode aux coordonnées géographiques de l'étude 3C. Les résultats sont présentés dans la table 3.6. Pour la latitude, nous donnons les paramètres estimés de la régression lorsque aucune, 5, 10, 20 et 100 PCs sont incluses dans le modèle avec des effets fixes. En l'absence de correction de la stratification de population, la pente estimée est très significative. Avec l'ajout des premières PCs dans le modèle, la pente diminue jusqu'à obtenir une p-valeur supérieure à 5% pour 20 PCs incluses en covariable. Pour la longitude, nous

regardons les mêmes modèles qu'avec la latitude. La pente estimée est très forte en l'absence de correction puis chute avec l'inclusion des premières PCs. Par la suite, la pente diminue lentement puis passe en dessous du seuil de significativité lorsque le modèle linéaire mixte contient entre 50 et 100 PCs en covariable. Selon la méthode proposée dans l'article [119], 20 et 100 PCs incluses dans le modèle avec des effets fixes semblent suffire à corriger la stratification de population pour la latitude et la longitude respectivement. Cependant, d'après nos résultats précédents, l'héritabilité estimée des coordonnées géographiques est encore très importante pour ces modèles. Nous constatons également un intercept significatif lorsque aucune correction de stratification de population n'est appliquée. D'après les auteurs de l'article [119], cela indique la présence d'apparentements cryptiques (ou éloignés) dans ce modèle.

Trait		Intercept	p-valeur	Pente	p-valeur
Latitude	0 PC	0.185	**3.2e-13**	8.9e-4	**2.8e-10**
	5 PCs	4.1e-3	0.14	4.5e-5	**0.025**
	10 PCs	4.3e-3	0.11	3.9e-5	**0.041**
	20 PCs	4.7e-3	0.086	2.9e-5	0.12
	50 PCs	4.5e-3	0.097	2.5e-5	0.17
	100 PCs	3.1e-3	0.17	2.3e-5	0.15
Longitude	0 PC	0.192	**8.0e-13**	1.0e-3	**1.4e-10**
	5 PCs	2.5e-3	0.53	9.5e-5	**2.6e-3**
	10 PCs	1.6e-3	0.69	8.6e-5	**6.5e-3**
	20 PCs	1.3e-3	0.72	7.9e-5	**5.5e-3**
	50 PCs	1.8e-3	0.62	6.8e-5	**0.013**
	100 PCs	2.7e-3	0.49	4.8e-5	0.086

TABLE 3.6 – Résultats de la régression de la différence des héritabilités estimées sur un seul chromosome dans un modèle avec un seul chromosome et celui avec tous les chromosomes en même temps sur la longueur des chromosomes (en mega-base). Un intercept significatif est interprété comme une indication de la présence d'apparentement cryptique. Une pente significative est interprétée comme une indication de la présence de stratification de population.

Les interactions entre les PCs

Une autre idée est d'introduire les interactions d'ordre 2 des premières PCs dans le modèle avec des effets fixes en plus des effets propres de celles-ci. Cette analyse complémentaire nous permet de voir si les interactions entre les PCs permettent d'expliquer une plus grande part de la variance des coordonnées géographiques. Nous avons introduit jusqu'à 50 PCs avec leurs carrés et leurs interactions deux à deux dans le modèle linéaire mixte. Les résultats sont donnés dans la figure 3.8. Pour la première PC, nous constatons que rajouter son carré dans le modèle permet d'expliquer davantage la variance de la latitude (34% au lieu de 31%) mais n'apporte rien pour la longitude (la proportion de variance expliquée par les effets fixes reste à 29%). Les dix premières PCs ainsi que leurs interactions d'ordre 2 n'expliquent que 39% et 41% de la variance de la latitude et la longitude respectivement. Au final, avec 50 PCs incluses dans le modèle avec leurs interactions deux à deux, 35% et 45% de la variance reste expliquée par la composante génétique pour la latitude et la longitude respectivement. L'introduction des interactions d'ordre 2 ne permet donc pas d'expliquer l'importante proportion de variance attribuée à la composante génétique et ne fournit donc pas une explication satisfaisante.

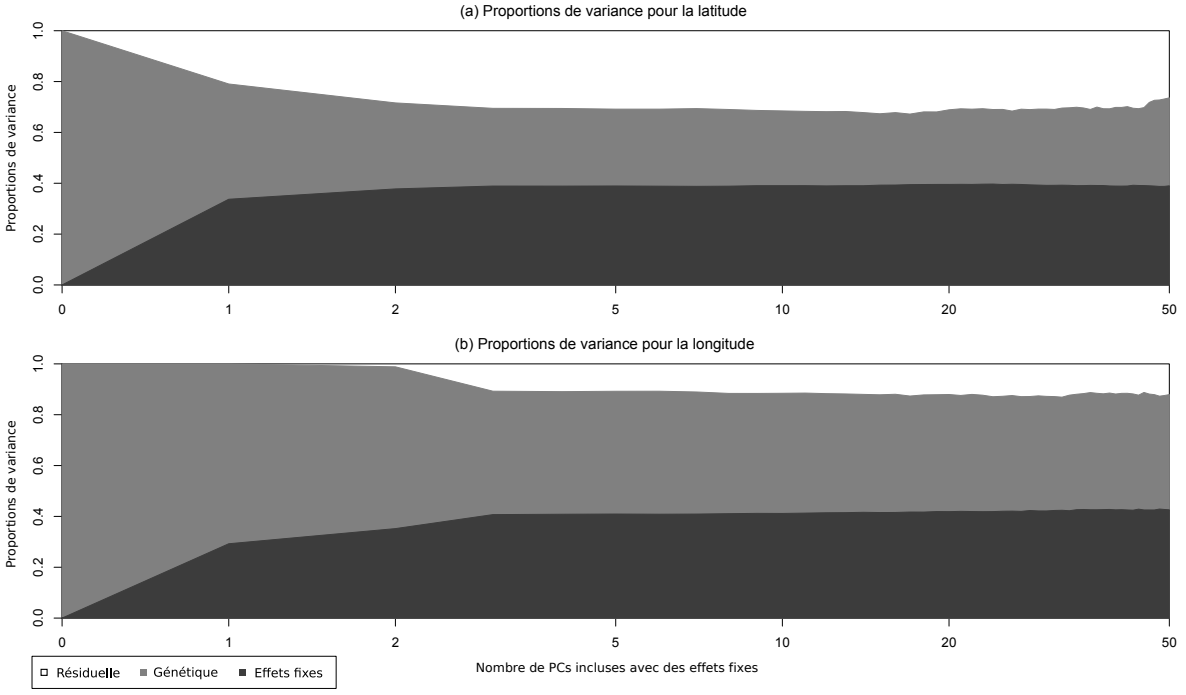
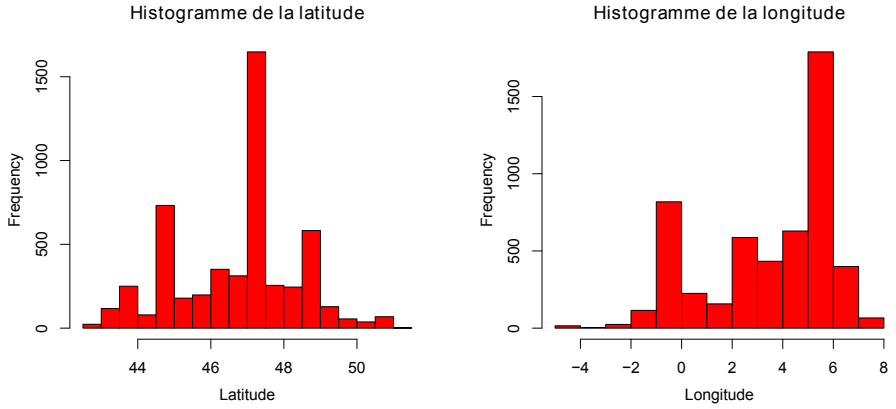


FIGURE 3.8 – Proportions de variance estimées pour les coordonnées géographiques en fonction du nombre de PCs incluses dans le modèle avec leurs interactions (échelle logarithmique). Le blanc, le gris clair et le gris foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

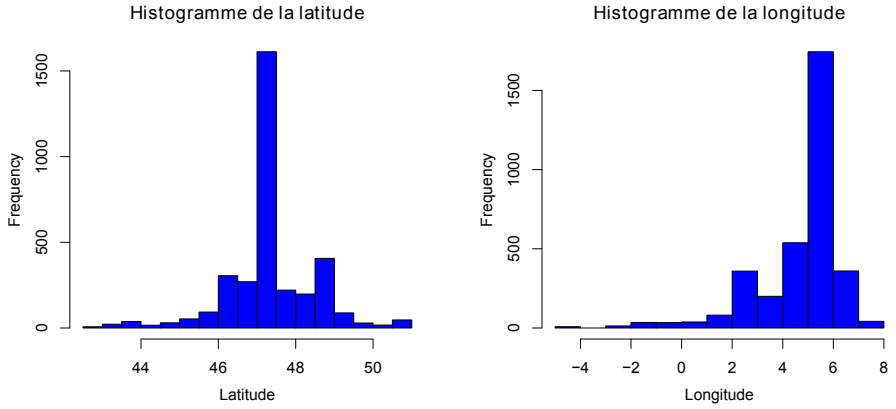
L'analyse d'un seul centre

Nous avons noté lors de nos explorations que les coordonnées géographiques sur l'ensemble des participants ont une distribution avec trois modes (figure 3.9a). Nous avons donc soupçonné que la forme de ces deux distributions pouvait influencer les estimations faites sous le modèle linéaire mixte qui suppose une distribution gaussienne du trait. Afin de vérifier si cette explication est correcte, nous avons analysé un seul centre, Dijon, qui a recruté le plus grand nombre de participants (3 676). En effet, si nous nous limitons au centre de recrutement de Dijon, les distributions des coordonnées géographiques n'ont plus qu'un seul mode et se rapprochent donc d'une loi gaussienne (figure 3.9b).

Les analyses ont ainsi été refaites sur ce sous-échantillon de l'étude 3C, les résultats obtenus sont donnés dans la figure 3.10. Pour la latitude, l'inclusion des deux premières PCs fait baisser la proportion de variance expliquée par la composante génétique autour de 54% au profit des variances résiduelle et des effets fixes. Pour la longitude maintenant, l'inclusion des deux mêmes premières PCs diminue la proportion de variance génétique autour de 68% au profit de la variance des effets fixes estimant ainsi pour ces trois modèles une héritabilité de 1. Par la suite la proportion de variance génétique diminue lentement jusqu'à 16% et 27% pour la latitude et la longitude respectivement. Ces résultats sont vraiment très proches de ceux obtenus sur la population totale. Ce n'est donc pas cette forme de distribution en trois modes qui pose problème.



(a) Ensemble de la population 3C



(b) Participants recrutés à Dijon

FIGURE 3.9 – Histogrammes des coordonnées géographiques.

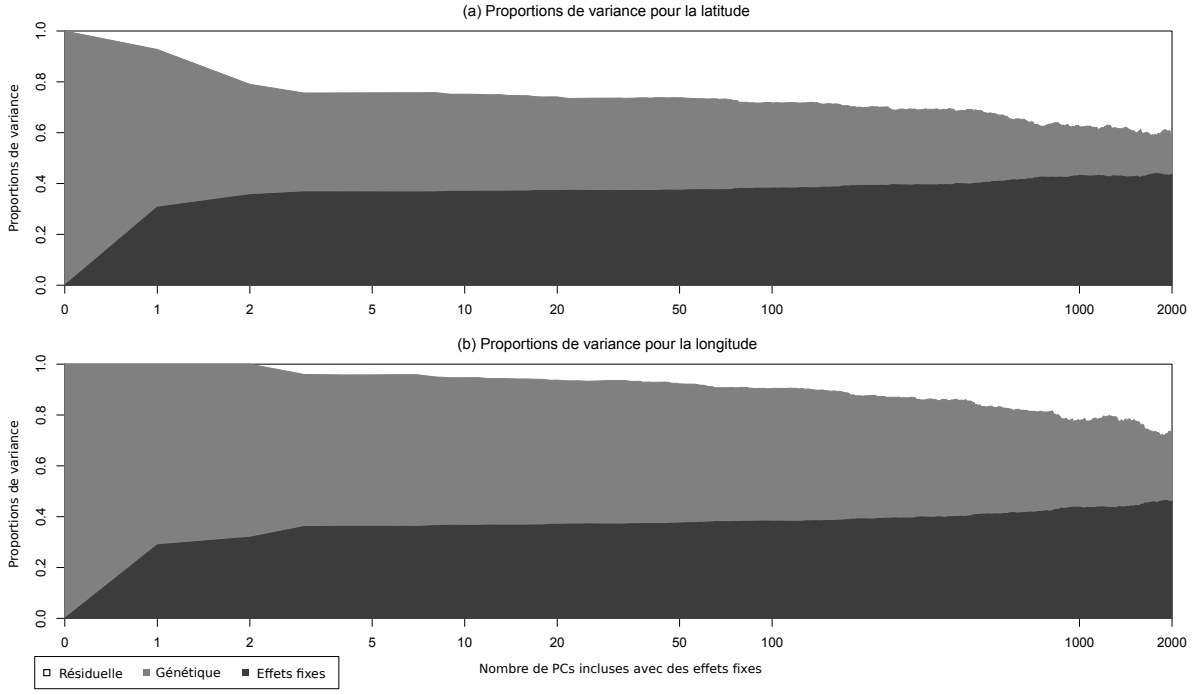


FIGURE 3.10 – Proportions de variance estimées pour les coordonnées géographiques pour le centre Dijon en fonction du nombre de PCs incluses dans le modèle (échelle logarithmique). Le blanc, le gris clair et le gris foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

Human Genome Diversity Project (HGDP)

Dans le but de mieux comprendre nos résultats, nous les avons comparés avec ceux obtenus pour un jeu de données contenant une stratification de population franche ; les données HGDP incluant des individus de différentes populations du monde. Les données HGDP sont des données publiques accessibles sur le site de l'université de Stanford [1] pour les génotypes et sur le site du CEPH pour les données individuelles [2]. Elles contiennent les génotypes de 1043 individus issus de 57 populations différentes pour 666 918 SNPs dont 644 258 autosomaux. Pour ces données, aucun contrôle qualité n'a été fait sur les marqueurs au préalable et seuls les individus avec un pourcentage de valeur manquante inférieure à 1.5% sont disponibles. Pour notre analyse complémentaire, nous avons gardé uniquement les 939 individus connus pour être non apparentés et dont l'origine géographique est disponible (table 3.7). Nous avons alors, sur ce sous-échantillon, appliqué un contrôle qualité sur les marqueurs. Seuls les SNPs ayant un pourcentage de valeurs manquantes inférieur à 1% et une p-valeur pour le test des proportions d'Hardy-Weinberg supérieure à 10^{-8} ont été conservés dans l'analyse (578 791 SNPs). Pour les données élaguées utilisées dans le calcul des PCs, nous avons appliqué un seuil de 10^{-5} pour le test des proportions d'Hardy-Weinberg, de 5% pour la fréquence de l'allèle mineur et de 0.1 pour le déséquilibre de liaison (63 432 SNPs). Les nuages de points des premières PCs calculées à l'aide des données élaguées sont tracés dans la figure 3.11 en fonction des régions du monde. Celles-ci différencient bien les différentes régions de recrutement des données HGDP.

Région du monde	Effectifs
Amérique	64
Asie	428
Europe	157
Moyen-Orient	133
Afrique du Nord	29
Afrique Subsaharienne	101
Océanie	27

Table 3.7 – Régions du monde du sous-échantillon des données HGDP utilisé dans l'analyse.

Les résultats obtenus avec le modèle linéaire mixte pour les coordonnées géographiques des lieux d'habitation de la population d'origine des individus de HGDP sont résumés dans la figure 3.12. Pour les deux types de coordonnée géographique, sans correction de la stratification de population, la proportion de variance expliquée par la composante génétique est estimée à 100%. Pour la latitude, la première PC explique 37% de la variance du trait. Après l'inclusion des 10 premières PCs, la proportion de variance génétique chute à 7.5% puis à 6% avec les 20 premières PCs. Si nous regardons maintenant la longitude, la première PC explique plus de la moitié de la variance (51%). La proportion de variance expliquée par la composante génétique est de 3% après l'inclusion des 10 premières PCs dans le modèle avec des effets fixes puis continue à diminuer très lentement jusqu'à la nullité. Malgré la proportion de variance très faible attribuée à la composante génétique, l'héritabilité de la latitude est estimée à 1 car la proportion de variance résiduelle est estimée à 0. Pour la longitude, la petite proportion de variance qui n'est pas expliquée par les PCs en covariable est d'abord attribuée à la composante génétique puis à la composante résiduelle. L'estimation de l'héritabilité de la longitude passe donc de 1 à 0

1. http://www.hagsc.org/hgdp/files.html
2. ftp://ftp.cephb.fr/hgdp_v3/

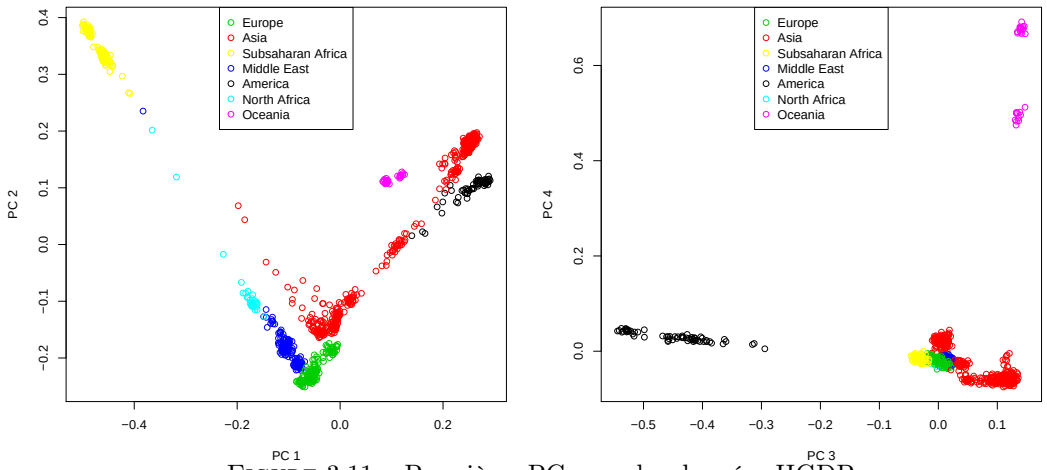


FIGURE 3.11 – Premières PC pour les données HGDP.

entre 50 et 100 PCs incluses avec des effets fixes. Ces résultats nous montrent que, dans le cas de données avec une très forte stratification de population (au niveau du monde), l'inclusion des premières PCs en effets fixes dans le modèle permet d'expliquer la quasi-totalité de la variance des coordonnées géographiques. Ces résultats correspondent davantage à ce que nous attendions même si l'héritabilité est toujours estimée à 1 pour la latitude. Il ne faut cependant pas oublier que ces données sont assez particulières. En effet, les populations représentées dans ces données viennent du monde entier et la précision des coordonnées géographiques n'est pas individuelle. La latitude et la longitude sont donc discrètes dans cet échantillon. De plus, la pertinence de la longitude peut être ici discutable car sa définition, relative au méridien de Greenwich, a pour conséquence que les continents les plus éloignés sont l'Asie et l'Amérique malgré leur proximité physique.

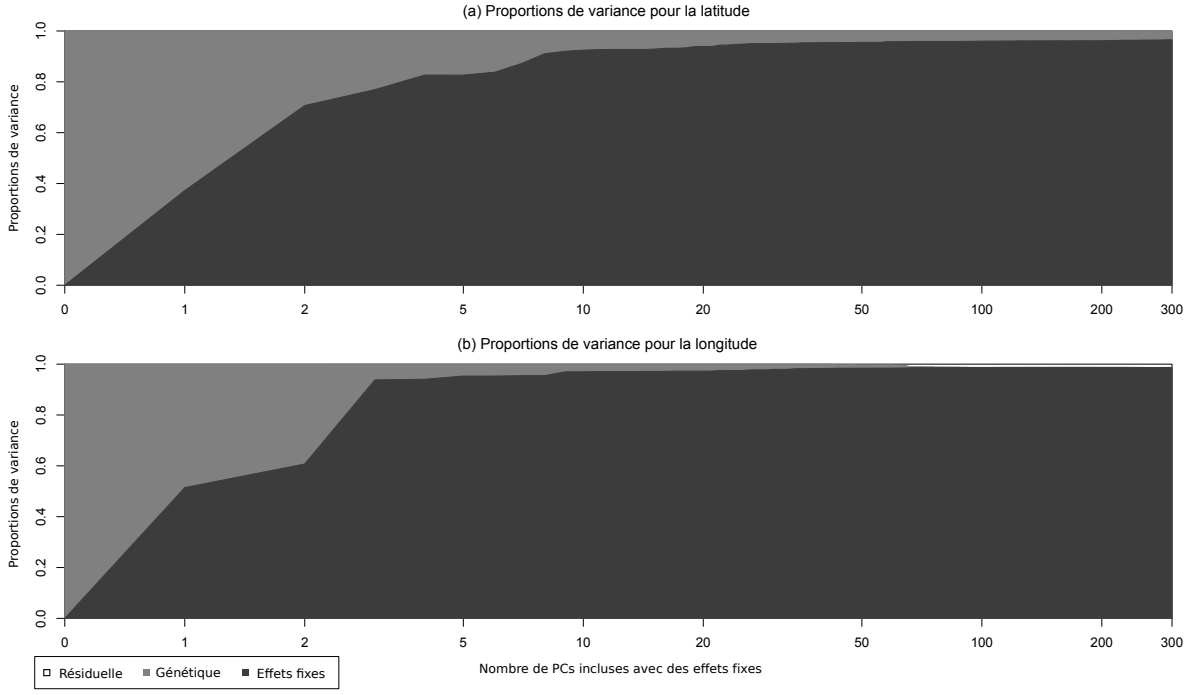


FIGURE 3.12 – Proportions de variance estimées pour les coordonnées géographiques pour les données HGDP en fonction du nombre de PCs incluses dans le modèle (échelle logarithmique). Le blanc, le gris clair et le gris foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

3.4.4 Conclusion

Les résultats obtenus pour les coordonnées géographiques suggèrent que l'inclusion des premières PCs dans le modèle linaire mixte avec des effets fixes dans le but de corriger la stratification de population, n'est pas suffisante si la stratification de population est « continue », aussi bien pour la génétique que pour la géographie, et à l'échelle d'un pays comme la France. Nous n'avons, pour le moment, pas d'explication totalement satisfaisante pour ces résultats. Cependant, plusieurs possibilités ont pu être écartées. Il paraît donc judicieux d'introduire dans le modèle les coordonnées géographiques des lieux de naissance si celles-ci sont disponibles afin de corriger au mieux la stratification de population dans une étude en population.

3.5 La stature

Dans cette nouvelle section, nous allons analyser, à l'aide du modèle linéaire mixte, la stature connue pour sa forte héritabilité. En effet, celle-ci est estimée entre 44% et 94% selon le type d'étude (table 3.2). Pour cela, plusieurs analyses vont être faites afin de comprendre au mieux nos résultats. Dans chaque analyse, le sexe et l'âge sont inclus en covariable dans les différents modèles en plus des éventuelles PCs.

Les résultats obtenus avec le modèle linéaire mixte de base donné dans l'équation (3.1) sont résumés dans la figure 3.13. La première chose que nous pouvons noter est que l'âge et le sexe expliquent plus de la moitié (52%) de la variance totale de la stature. Sans correction de la stratification de population, la proportion de variance expliquée par la composante génétique est estimée à 22%. Nous remarquons une légère baisse de la variance génétique avec l'inclusion de la première PC avec un effet fixe dans le modèle autour de 19%. Par la suite, les estimations restent stables malgré quelques fluctuations.

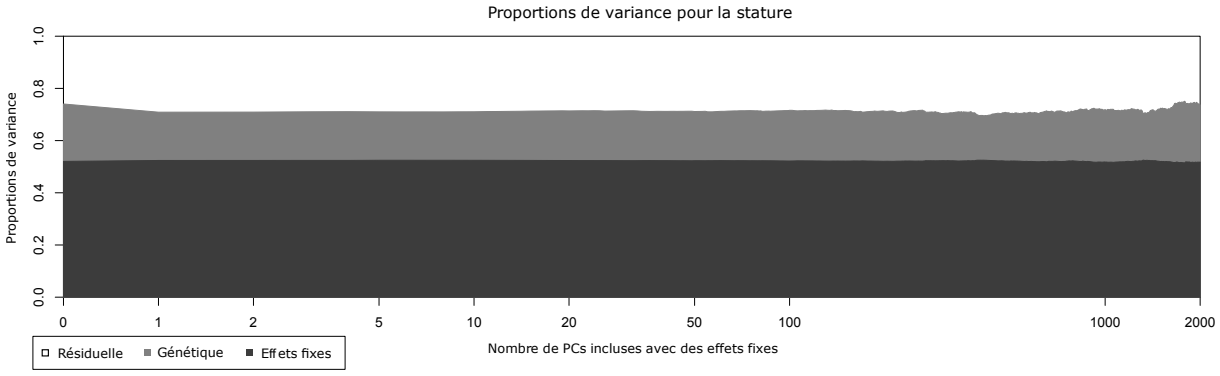


FIGURE 3.13 – Proportions de variance estimées pour la stature en fonction du nombre de PCs incluses dans le modèle avec le sexe et l'âge (échelle logarithmique). Le blanc, le gris clair et le gris foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

Nous pouvons maintenant regarder avec plus de détails les estimations des composantes de la variance données dans la table 3.8. En terme d'héritabilité, sans correction de la stratification de population, l'héritabilité est estimée à 46%. L'inclusion de la première PC fait baisser l'héritabilité à 39% avec une erreur-type de 7%. Par la suite, les estimations ne varient presque plus. Les estimations de l'héritabilité sont donc significativement différentes de 0 dans tous les modèles considérés et sont cohérentes avec les estimations faites dans la littérature. Ces

résultats nous montrent que la stature est affectée par la stratification de population en France et que l'inclusion de la première PC semble suffire à corriger celle-ci.

	LRT	p-valeur	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma^2_P}$ (se)	$\widehat{h^2}$ (se)
0 PC	68.80	**5.6e-17**	17.48 (0.719)	20.64 (0.689)	38.12 (0.687)	0.459 (0.061)
1 PC	36.32	**8.4e-10**	14.70 (0.712)	23.12 (0.690)	37.83 (0.685)	0.389 (0.066)
5 PCs	36.57	**7.4e-10**	14.73 (0.712)	23.03 (0.690)	37.76 (0.685)	0.390 (0.066)
10 PCs	36.60	**7.2e-10**	14.77 (0.712)	23.00 (0.690)	37.78 (0.686)	0.391 (0.066)
20 PCs	37.87	**3.8e-10**	15.10 (0.715)	22.75 (0.691)	37.85 (0.686)	0.399 (0.066)

Table 3.8 – Estimations des paramètres du modèle pour la stature et leur erreur-type (se) en fonction du nombre de PCs incluses dans le modèle : test du rapport de vraisemblance (LRT) et la p-valeur associée, $\widehat{\tau}$, $\widehat{\sigma^2}$ et $\widehat{\sigma^2_P} = \widehat{\tau} + \widehat{\sigma^2}$ l'estimation des variances génétique, résiduelle et totale respectivement et $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ l'héritabilité estimée.

3.5.1 L'ajout de la longitude et la latitude en covariable

Nous allons maintenant essayer de confirmer que l'effet de la première PC sur la stature indique bien la présence d'une stratification de la population pour ce trait. Pour cela, la première chose que nous pouvons regarder est l'effet de l'inclusion des coordonnées géographiques dans le modèle sur les estimations. En effet, nous pouvons nous demander si l'effet de la première PC sera toujours présent. Les résultats sont donnés dans la figure 3.14 et la table 3.9.

La première chose à noter est que les covariables (sexe, âge, latitude et longitude) expliquent 53% de la variance de la stature. Cette valeur est très proche de celle obtenue précédemment. Ensuite, l'effet de la première PC disparaît. La proportion de variance expliquée par la composante génétique est estimée à 17% ce qui est plus faible que celle obtenue avec les modèles sans les coordonnées géographiques (22% ou 19%). Ces résultats suggèrent que l'inclusion de la première PC n'est en fait pas suffisante pour corriger la stratification de population liée à l'origine géographique.

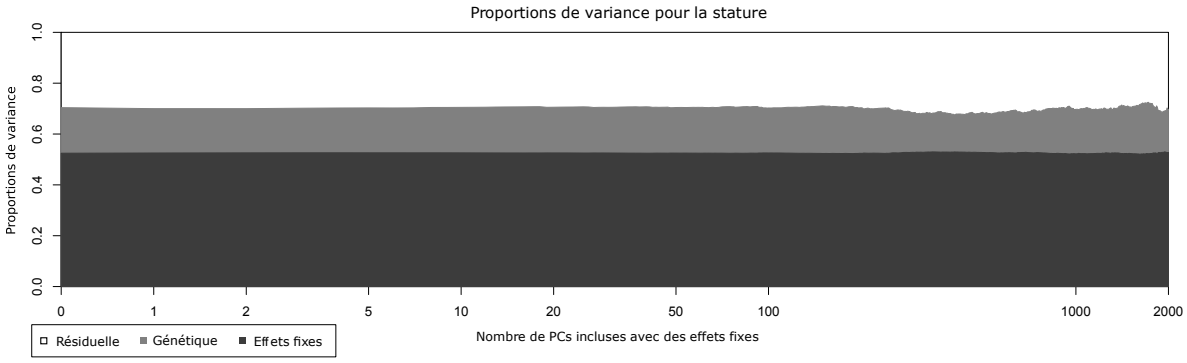


Figure 3.14 – Estimations des proportions de variance pour la stature avec les coordonnées géographiques en covariable en fonction du nombre de PCs incluses dans le modèle avec des effets fixes (échelle logarithmique). Le blanc et les gris clair et foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

Si nous regardons maintenant plus en détails les valeurs des estimations ainsi que l'héritabilité pour quelques modèles, nous constatons que la diminution dans l'estimation des proportions de variance génétique se ressent dans l'estimation de l'héritabilité. Nous obtenons une hérita-

bilité estimée de 37% avec une erreur-type autour de 7% au lieu de 46% ou 39% sans ou avec correction de la stratification de population respectivement. L'héritabilité reste tout de même très significative.

	LRT	p-valeur	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma^2_P}$ (se)	$\widehat{h^2}$ (se)
0 PC	30.05	**2.1e-8**	14.09 (0.737)	23.29 (0.712)	37.38 (0.715)	0.377 (0.070)
1 PC	27.98	**6.1e-8**	13.72 (0.736)	23.61 (0.715)	37.33 (0.714)	0.368 (0.071)
5 PCs	28.59	**4.5e-8**	13.89 (0.734)	23.39 (0.712)	37.28 (0.706)	0.373 (0.071)
10 PCs	29.34	**3.0e-8**	14.08 (0.736)	23.22 (0.715)	37.30 (0.712)	0.378 (0.071)

TABLE 3.9 – Estimations des paramètres du modèle pour la stature et leur erreur-type lorsque les coordonnées géographiques sont incluses en covariable en plus du sexe, de l'âge et d'un nombre variable de PCs. La table contient aussi le test du rapport de vraisemblance (LRT) testant la significativité de l'héritabilité et la p-valeur associée. $\widehat{\tau}$ est l'estimation de la variance génétique, $\widehat{\sigma^2}$ l'estimation de la variance résiduelle, $\widehat{\sigma^2_P} = \widehat{\tau} + \widehat{\sigma^2}$ l'estimation de la variance totale et $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ l'héritabilité estimée.

3.5.2 Stratification de population ou effet centre ?

Une autre chose que nous pouvons vérifier est la présence d'un effet centre qui pourrait expliquer l'effet de la première PC. En effet, la première PC semble être différente selon le centre de recrutement (figure 3.15) et les traits étudiés peuvent être affectés par les méthodes et les conditions de mesure des centres de recrutement.

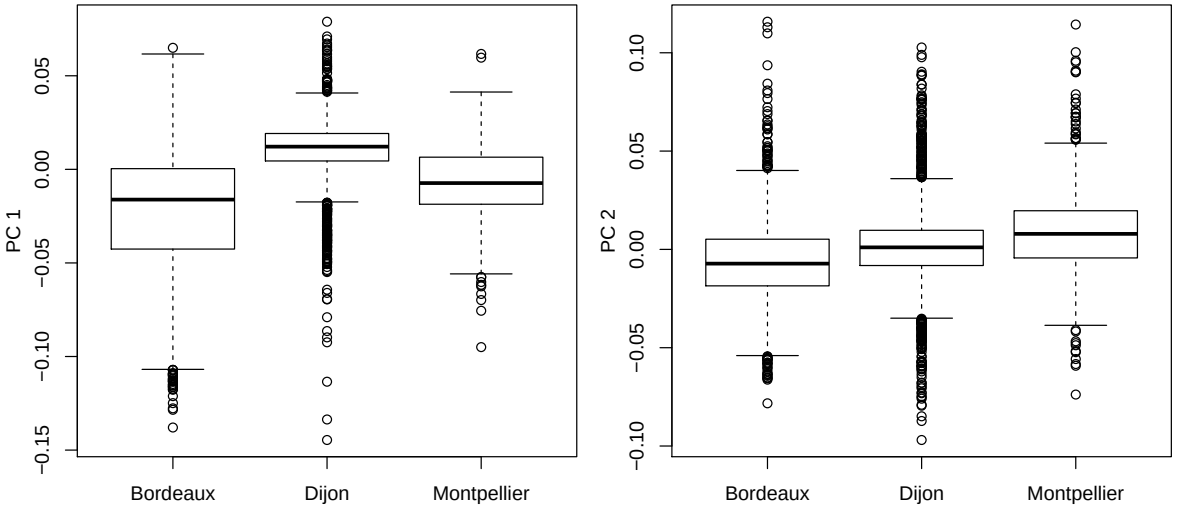


FIGURE 3.15 – Boxplots des valeurs des deux premières PC en fonction du centre de recrutement.

Afin de vérifier la présence éventuelle d'un effet centre, nous avons dans un premier temps ajouté les centres de recrutement en covariable dans le modèle sous forme de deux indicatrices. Cette méthode est classique en épidémiologie pour corriger un éventuel effet centre. Les résultats sont donnés dans la figure 3.16 et la table 3.10. La proportion de variance expliquée par le sexe, l'âge et le centre est estimée à 53%. Le centre n'a donc pas un effet fort sur la proportion de variance expliquée par les covariables. L'effet de la première PC est ici aussi très amoindri. En effet, avec l'ajout de la première PC, la part de variance attribuée à la composante génétique

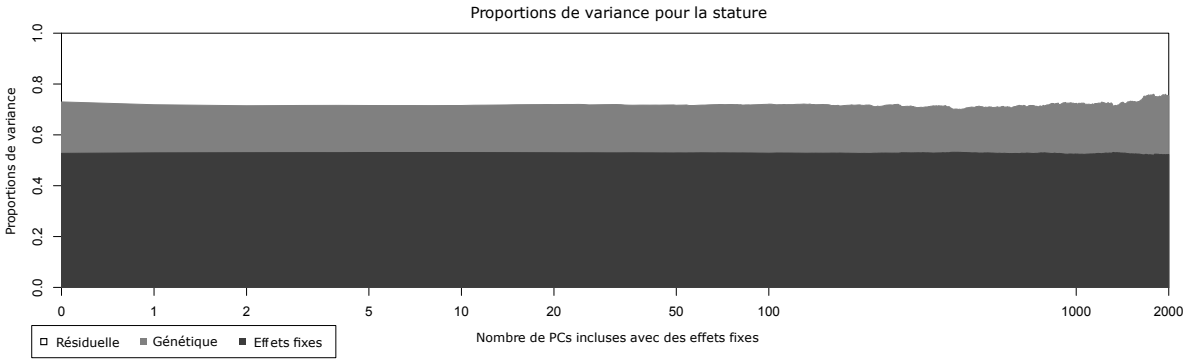


FIGURE 3.16 – Estimations des proportions de variance pour la stature avec le centre inclus en covariable et en fonction du nombre de PCs incluses dans le modèle avec des effets fixes (échelle logarithmique). Le blanc et les gris clair et foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

passe de 20% à 19%. Les estimations sont donc légèrement plus faibles que celles obtenues dans les analyses principales après correction de la stratification de population. De la même façon, les estimations de l'héritabilité (table 3.10) sont très similaires à celles obtenues précédemment. La différence notable est l'effet très faible de la première PC qui diminue l'héritabilité estimée de 43% à 40% pour la stature. Ces résultats semblent indiquer que l'effet de la première PC est dû à un effet centre. Cependant, si nous revenons sur la carte de France avec la distribution des coordonnées géographiques ou la distribution des deux premières PCs en fonction du centre de recrutement (figures 3.2 et 3.15), nous pouvons voir que les centres ne sont pas indépendants des coordonnées géographiques ou de la première PCs. Cela se confirme si nous comparons les distributions des coordonnées géographiques pour tous les centres ou pour Dijon uniquement (figures 3.9). Nous pouvons donc supposer que le centre et la stratification de population sont liés.

	LRT	p-valeur	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma^2_P}$ (se)	$\widehat{h^2}$ (se)
0 PC	49.40	**1.0e-12**	16.06 (0.707)	21.47 (0.682)	37.53 (0.676)	0.428 (0.063)
1 PC	39.82	**1.4e-10**	15.05 (0.705)	22.37 (0.682)	37.42 (0.679)	0.402 (0.065)
5 PCs	37.45	**4.7e-10**	14.69 (0.702)	22.60 (0.681)	37.29 (0.676)	0.394 (0.065)
10 PCs	37.36	**4.9e-10**	14.71 (0.703)	22.60 (0.681)	37.31 (0.677)	0.394 (0.066)

TABLE 3.10 – Estimations des paramètres du modèle pour la stature et leur erreur-type lorsque le centre est inclus comme covariable. Le sexe, l'âge et un nombre variable de PCs sont inclus dans le modèle. La table contient aussi le test du rapport de vraisemblance (LRT) testant la significativité de l'héritabilité et la p-valeur associée. $\widehat{\tau}$ est l'estimation de la variance génétique, $\widehat{\sigma^2}$ l'estimation de la variance résiduelle, $\widehat{\sigma^2_P} = \widehat{\tau} + \widehat{\sigma^2}$ l'estimation de la variance totale et $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ l'héritabilité estimée.

Afin de vérifier nos suppositions, nous avons refait l'analyse uniquement sur les individus recrutés à Dijon. Pour cette analyse complémentaire, nous avons choisi de garder les PCs estimées sur la totalité de la population qui représentent la France entière. Les résultats sont donnés dans la figure 3.17 et la table 3.11. La variance attribuée au sexe et à l'âge est la même que dans l'analyse principale (52%). Nous constatons de nouveau un effet notable de la première PC. En effet, avec l'inclusion de la première PC avec un effet fixe dans le modèle, la proportion de variance génétique estimée passe de 22% à 19%. L'ajout de la première PC se répercute également sur l'estimation de l'héritabilité qui passe de 46% à 39%. Ces résultats indiquent que l'effet de la première PC n'est pas dû à un effet centre et que les résultats précédents

s'expliquent par le fait que le centre de recrutement n'est pas indépendant des coordonnées géographiques et donc des premières PCs.

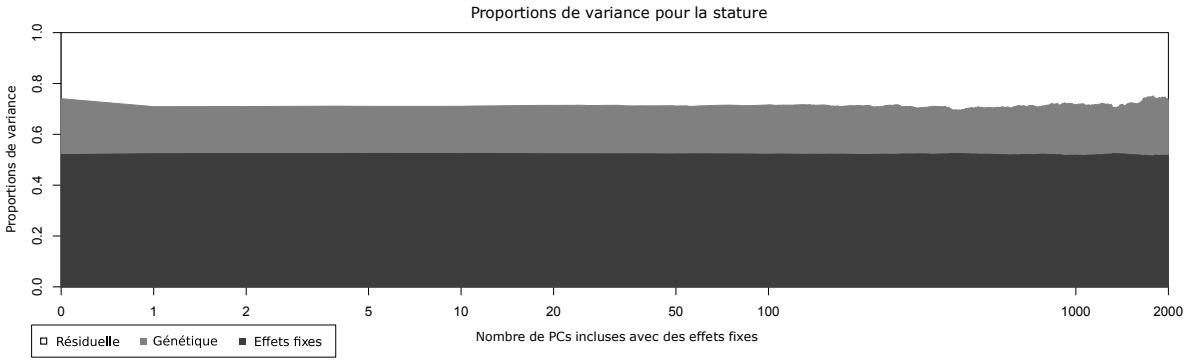


FIGURE 3.17 – Estimations des proportions de variance pour la stature lorsque seuls les individus recrutés à Dijon sont analysés en fonction du nombre de PCs incluses dans le modèle avec des effets fixes (échelle logarithmique). Le blanc et les gris clair et foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

	LRT	p-valeur	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma^2_P}$ (se)	$\widehat{h^2}$ (se)
0 PC	68.80	**5.6e-17**	17.48 (0.718)	20.64 (0.690)	38.13 (0.687)	0.459 (0.061)
1 PC	36.32	**8.4e-10**	14.70 (0.713)	23.13 (0.691)	37.83 (0.686)	0.389 (0.066)
5 PCs	36.56	**7.4e-10**	14.74 (0.711)	23.03 (0.689)	37.77 (0.685)	0.390 (0.066)
10 PCs	36.60	**7.3e-10**	14.77 (0.713)	23.01 (0.690)	37.78 (0.686)	0.391 (0.066)

TABLE 3.11 – Estimations des paramètres du modèle pour la stature et leur erreur-type lorsque seuls les individus recrutés à Dijon sont analysés. Le sexe, l'âge et un nombre variable de PCs sont inclus dans le modèle. La table contient aussi le test du rapport de vraisemblance (LRT) testant la significativité de l'héritabilité et la p-valeur associée. $\widehat{\tau}$ est l'estimation de la variance génétique, $\widehat{\sigma^2}$ l'estimation de la variance résiduelle, $\widehat{\sigma^2_P} = \widehat{\tau} + \widehat{\sigma^2}$ l'estimation de la variance totale et $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ l'héritabilité estimée.

3.5.3 La détection de la stratification de population

Comme pour les coordonnées géographiques nous avons appliqué à la stature la méthode proposée par Yang *et al.* [119] (paragraphe 3.4.3) pour détecter la présence de stratification (table 3.12). Dans tous les modèles regardés, les intercepts estimés ne sont pas significatifs indiquant que les estimations ne sont pas sujettes à la présence d'apparentement cryptique. Pour la stature, les résultats nous montrent que l'introduction d'une seule PC dans le modèle avec un effet fixe suffit à corriger la stratification de population (les pentes ne sont plus significatives

	Intercept	p-valeur	Pente	p-valeur
0 PC	1.4e-3	0.46	6.8e-5	**3.9e-5**
1 PC	-1.3e-5	0.99	6.9e-6	0.44
2 PCs	-6.1e-5	0.96	6.7e-6	0.43

TABLE 3.12 – Résultats de la régression de la différence des héritabilités estimées sur un seul chromosome dans un modèle avec un seul chromosome et celui avec tous les chromosomes en même temps sur la longueur des chromosomes (en mega-base). Un intercept significatif est interprété comme une indication de la présence d'apparentement cryptique. Une pente significative est interprétée comme une indication de la présence de stratification de population.

pour les modèles avec au moins deux PCs en covariable). Cette analyse confirme nos conclusions pour la présence de stratification de population corrigée avec l'ajout de la première PC en covariable.

3.5.4 Conclusion

Dans cette section, les héritabilités estimées pour la stature sont en adéquation avec les valeurs données dans la littérature. Nos analyses complémentaires semblent appuyer l'hypothèse selon laquelle l'effet fort de la première PC serait dû à la présence de stratification de population. Il n'est pas exclu ici que celui-ci s'explique par l'effet de l'environnement, puisque l'effet de la première PC peut être expliqué soit par un gradient (nord-sud ou est-ouest) allélique de certains gènes impliqués dans la détermination de la stature, soit par une caractéristique environnementale liée à la géographie. Il est, dans le cadre de cette analyse, impossible de conclure avec certitude pour l'une ou l'autre des explications.

3.6 Les autres traits anthropométriques

Afin d'alléger la lecture, nous avons fait le choix d'approfondir l'analyse de la stature dans la section précédente. Nous allons maintenant regarder les résultats pour les autres traits anthropométriques mais de façon moins détaillée. Les analyses complémentaires sont données en Annexe 5.4.

3.6.1 Le poids et le BMI

Nous allons maintenant regarder les résultats obtenus pour deux traits liés à l'adiposité, le poids et le BMI. Les proportions de variance génétique, résiduelle et des effets fixes sont données dans la figure 3.18. Le sexe et l'âge expliquent 26% de la variance du poids. Pour le BMI, ces deux covariables ont un effet très faible (1.4%). Dans les deux cas, l'inclusion des premières PCs n'a pas d'effet notable sur les estimations, la proportion de variance génétique est estimée à 16% et 19% respectivement.

Nos observations sont confirmées par la table 3.13 contenant les estimations des variances et de l'héritabilité avec leur erreur-type. Nous constatons une légère fluctuation des estimations des variances mais celle-ci reste minime. L'héritabilité est estimée à 22% et 19% pour le poids et le BMI respectivement. Dans les deux cas, nous obtenons une héritabilité significativement différente de 0 compatible avec les résultats déjà obtenus dans la littérature (table 3.2). Ces résultats semblent indiquer que ces deux traits ne sont pas affectés par la stratification de population. L'absence d'effet de la stratification de population est confirmée avec la méthode proposée par Yang *et al.* [119] (Annexe 5.4). Étant donné l'absence d'évolution dans les estimations avec l'ajout de la première PC en covariable, ces deux traits ne font pas l'objet d'autres analyses complémentaires dans ce manuscrit.

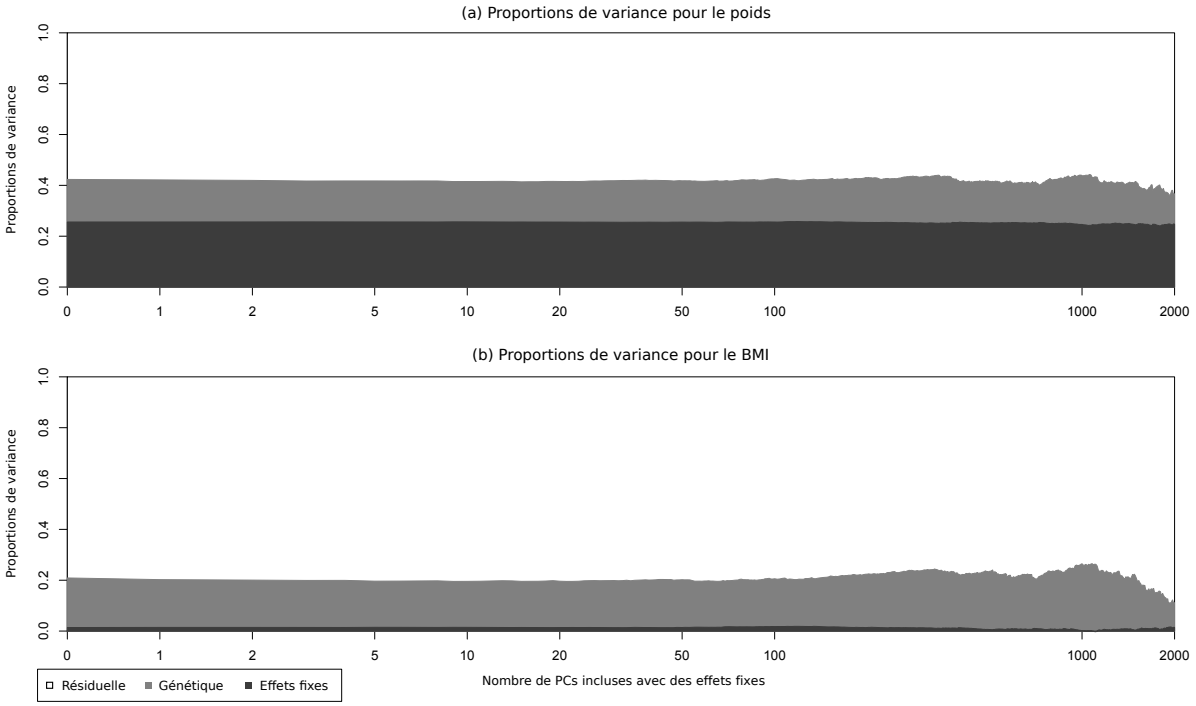


FIGURE 3.18 – Proportions de variance estimées pour le poids et le BMI en fonction du nombre de PCs incluses dans le modèle (échelle logarithmique). Le blanc, le gris clair et le gris foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

Trait		LRT	p-valeur	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma_P^2}$ (se)	$\widehat{h^2}$ (se)
Poids	0 PC	13.92	**9.6e-5**	28.24 (2.32)	97.23 (2.32)	125.47 (2.11)	0.225 (0.062)
	1 PC	13.34	**1.3e-4**	27.91 (2.32)	97.53 (2.32)	125.44 (2.12)	0.222 (0.062)
	10 PCs	12.07	**2.6e-4**	26.77 (2.32)	98.59 (2.32)	125.36 (2.13)	0.214 (0.062)
	20 PCs	12.10	**2.5e-4**	26.91 (2.32)	98.56 (2.32)	125.47 (2.13)	0.214 (0.063)
BMI	0 PC	10.44	**6.2e-4**	3.23 (0.302)	13.09 (0.302)	16.31 (0.304)	0.198 (0.063)
	1 PC	9.26	**1.2e-3**	3.11 (0.302)	13.19 (0.302)	16.30 (0.304)	0.191 (0.064)
	10 PCs	8.47	**1.8e-3**	2.99 (0.302)	13.30 (0.303)	16.30 (0.304)	0.184 (0.064)
	20 PCs	8.52	**1.8e-3**	3.01 (0.302)	13.29 (0.303)	16.31 (0.305)	0.185 (0.064)

TABLE 3.13 – Estimations des paramètres du modèle pour le poids et le BMI et leur erreur-type (se) en fonction du nombre de PCs incluses dans le modèle : test du rapport de vraisemblance (LRT) et la p-valeur associée, $\widehat{\tau}$, $\widehat{\sigma^2}$ et $\widehat{\sigma_P^2} = \widehat{\tau} + \widehat{\sigma^2}$ l'estimation des variances génétique, résiduelle et totale respectivement et $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ l'héritabilité estimée.

3.6.2 La circonférence du crâne

Le trait analysé est maintenant la circonférence du crâne. Les résultats sont tracés dans la figure 3.19. Les effets des covariables (sexe et âge) expliquent 24% de la variance de la circonférence du crâne. Sans correction de la stratification de population, la proportion de variance génétique est estimée à 13%. Puis, la valeur de son estimation chute à 8% avec l'inclusion de la première PC au profit de la variance résiduelle. Par la suite, les estimations se stabilisent avec de légères fluctuations pour un nombre important de PCs incluses dans le modèle avec des effets fixes.

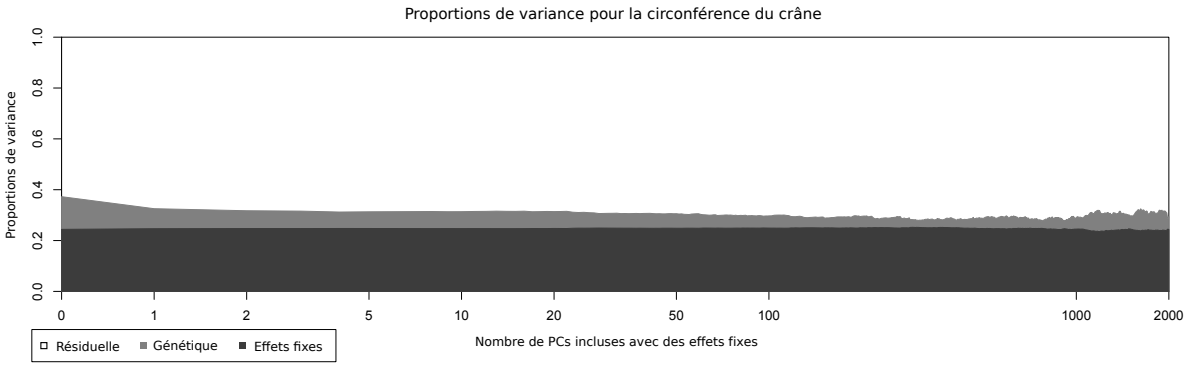


FIGURE 3.19 – Proportions de variance estimées pour la circonférence du crâne en fonction du nombre de PCs incluses dans le modèle (échelle logarithmique). Le blanc, le gris clair et le gris foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

Si nous regardons les estimations de l'héritabilité au sens de Visscher (table 3.14), nous obtenons 17% sans correction de la stratification de population. Dans ce premier modèle, l'héritabilité est significativement non nulle. Avec l'inclusion de la première PC, l'estimation de l'héritabilité baisse à 10% et est à la limite de la significativité. L'inclusion de la deuxième PC fait basculer la valeur de l'estimation de l'héritabilité sous la valeur minimale pour être significative. Nous avons donc finalement une estimation de l'héritabilité non significative de 9% avec une erreur-type de 6%. L'estimation de l'héritabilité obtenue est très inférieure à la valeur trouvée dans la méta-analyse [149]. Cependant, dans cette étude, tous les traits liés au squelette ont été réunis. Il est donc difficile de faire une comparaison rigoureuse.

	LRT	p-valeur	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma^2_P}$ (se)	$\widehat{h^2}$ (se)
0 PC	9.85	**8.5e-4**	0.722 (0.078)	3.52 (0.079)	4.24 (0.080)	0.170 (0.057)
1 PC	2.79	**0.047**	0.442 (0.078)	3.78 (0.079)	4.23 (0.079)	0.104 (0.063)
2 PCs	2.16	0.071	0.393 (0.078)	3.83 (0.079)	4.22 (0.079)	0.093 (0.064)
5 PCs	1.88	0.085	0.368 (0.078)	3.85 (0.079)	4.22 (0.079)	0.087 (0.064)
10 PCs	1.93	0.082	0.373 (0.078)	3.85 (0.079)	4.22 (0.079)	0.088 (0.064)
20 PCs	1.92	0.083	0.372 (0.078)	3.85 (0.079)	4.22 (0.080)	0.088 (0.064)

TABLE 3.14 – Estimations des paramètres du modèle pour la circonférence crânienne et leur erreur-type (se) en fonction du nombre de PCs incluses dans le modèle : test du rapport de vraisemblance (LRT) et la p-valeur associée, $\widehat{\tau}$, $\widehat{\sigma^2}$ et $\widehat{\sigma^2_P} = \widehat{\tau} + \widehat{\sigma^2}$ l'estimation des variances génétique, résiduelle et totale respectivement et $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ l'héritabilité estimée.

3.6.3 Le rapport de la circonférence de la taille sur celle des hanches

Le dernier trait que nous avons analysé est le rapport de la circonférence de la taille sur celle des hanches. Les résultats sont données dans la figure 3.20. La proportion de variance expliquée par les covariables est estimée à 41%. Sans correction de la stratification de population, la part de génétique dans la variance du trait est estimée à 12%. Après l'inclusion d'au moins une PC, celle-ci baisse à 8% au profit de la proportion de variance résiduelle puis se stabilise.

Si nous regardons les estimations de l'héritabilité estimée (table 3.15), nous constatons que celle-ci est estimée à 21% avec une erreur-type de 6% en l'absence de stratification de population puis baisse entre 13 et 14% lorsqu'au moins une PC est incluse dans le modèle avec

des effets fixes. Ces estimations sont toutes significativement non nulles même après l'inclusion de la première PC qui fait monter la p-valeur autour de 2%. L'héritabilité estimée sur notre échantillon est nettement plus faible que celle obtenue sur des données familiales ce qui est aussi observé pour beaucoup d'autres traits (table 3.2). De plus, nous pouvons noter que l'estimation de l'héritabilité de ce trait se situe entre les estimations de l'héritabilité du BMI (19%) et de la circonférence crânienne qui fait partie des caractéristiques du squelette (9%). Or, ces deux traits peuvent être liés au rapport de la circonférence de la taille sur celle des hanches.

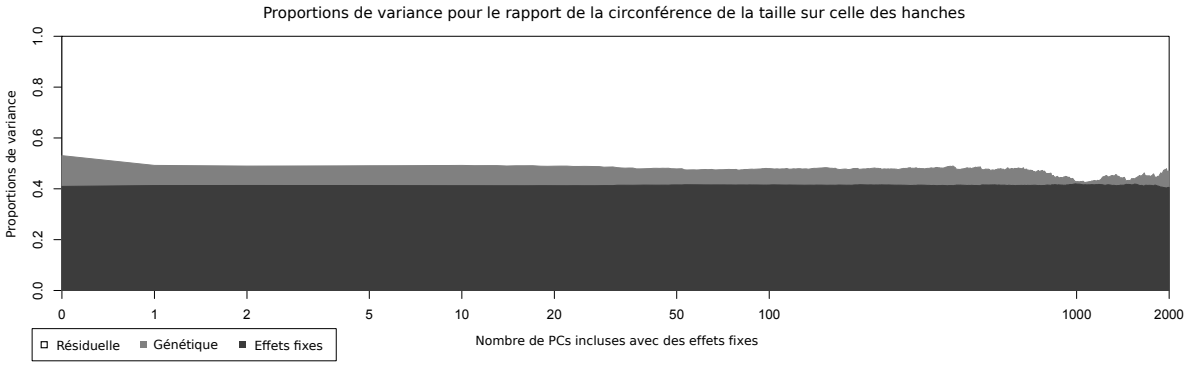


FIGURE 3.20 – Proportions de variance estimées pour le rapport de la circonférence de la taille sur celle des hanches en fonction de nombre de PCs incluses dans le modèle (échelle logarithmique). Le blanc, le gris clair et le gris foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

	LRT	p-valeur	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma^2_P}$ (se)	$\widehat{h^2}$ (se)
0 PC	13.09	**1.5e-4**	9.2e-4 (8.6e-5)	3.6e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.205 (0.061)
1 PC	4.29	**0.019**	6.0e-4 (8.6e-5)	3.9e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.135 (0.067)
5 PCs	4.11	**0.021**	5.9e-4 (8.6e-5)	3.9e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.133 (0.067)
10 PCs	4.28	**0.019**	6.1e-4 (8.6e-5)	3.9e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.136 (0.067)
20 PCs	3.94	**0.024**	5.8e-4 (8.6e-5)	3.9e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.131 (0.066)

TABLE 3.15 – Estimations des paramètres du modèle pour le rapport de la circonférence de la taille sur celle des hanches et leur erreur-type (se) en fonction du nombre de PCs incluses dans le modèle : test du rapport de vraisemblance (LRT) et la p-valeur associée, $\widehat{\tau}$, $\widehat{\sigma^2}$ et $\widehat{\sigma^2_P} = \widehat{\tau} + \widehat{\sigma^2}$ l'estimation des variances génétique, résiduelle et totale respectivement et $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ l'héritabilité estimée.

3.7 Petit résumé et discussion

Nous avons commencé par des simulations afin d'explorer le comportement des estimations des composantes de la variance sous le modèle linéaire mixte. Nous avons alors montré qu'il est possible d'introduire un nombre important de PCs sans affecter la qualité des estimations. Bien sûr, la qualité des estimations dépend également de la taille de l'échantillon. Par exemple, pour un échantillon avec environ 6 000 individus, inclure 100 ou 500 PCs ne diminue pas significativement la précision des estimations. Il est donc possible d'utiliser un grand nombre de PCs pour corriger la stratification de population.

Nous avons ensuite analysé les coordonnées géographiques des lieux de naissance dans la cohorte 3C. Nous avons d'abord été surpris par les estimations de l'héritabilité à 100% pour la latitude et de la longitude en l'absence de correction de la stratification de population. En effet,

nous attendions une héritabilité estimée forte mais pas au point d'atteindre 100%. De plus, nous pensions que l'inclusion des premières PCs fortement corrélées aux coordonnées géographiques avec des effets fixes dans le modèle suffirait à réduire les estimations de l'héritabilité à 0 ou du moins à une valeur proche de 0. Or, les héritabilités estimées restent très importantes après l'inclusion des premières PCs (autour de 68% pour la latitude) puis continuent à diminuer très doucement. Pour tous les modèles essayés, les héritabilités estimées restent très significatives pour la latitude et la longitude. Au travers d'analyses complémentaires, nous avons cherché à comprendre l'origine de ces résultats. Cependant, nos conclusions se sont révélées robustes aux différents tests que nous avons pu réaliser. Browning *et al.* [44] ont obtenu des résultats similaires sur des données cas-témoins simulées à partir de l'étude cas-témoins WTCCC avec un déséquilibre très important dans la répartition des cas dans les trois populations considérées (90% des cas venaient de la population écossaise ou galloise, les 10% restant venant de l'Angleterre). Cependant, ce scénario est un cas extrême [43]. Il semble plus réaliste d'imaginer l'influence d'un facteur environnemental dépendant des coordonnées géographiques sur un trait quantitatif qui provoquerait la surestimation de son héritabilité. Pour finir, nous avons aussi appliqué une méthode développée pour diagnostiquer la présence de stratification de population ou d'apparentement cryptique [119].Pour les coordonnées géographiques, elle ne détecte plus de stratification de population lorsque 20 et 100 PCs sont incluses dans le modèle avec des effets fixes pour la latitude et la longitude respectivement. Ces résultats suggèrent que cette méthode est efficace mais a des difficultés pour détecter la stratification de population pour un trait dépendant directement de la géographie comme, par exemple, l'ensoleillement.

Nous avons ensuite analysé les traits anthropométriques. Les héritabilités estimées obtenues sont globalement cohérentes avec les résultats présents dans la littérature. Le BMI et le poids ne semblent pas affectés par la présence d'une stratification de population. Au contraire, la stature, le rapport de la circonférence de la taille sur celle des hanches et la circonférence de crâne semblent être sujets à la stratification de population car les estimations des proportions de variance expliquée par les différents éléments du modèle changent significativement avec l'inclusion des deux premières PCs avec des effets fixes. Le trait le plus affecté est la circonférence du crâne pour laquelle l'héritabilité estimée n'est plus significativement différente de 0 après l'inclusion des premières PCs. Par la suite, avec l'inclusion des PCs suivantes, les estimations ne fluctuent plus beaucoup. Pour la stature, nous avons ensuite cherché à savoir si inclure les coordonnées géographiques dans le modèle avec des effets fixes donnait des résultats différents. Dans ce cas, l'effet des premières PCs disparait et les héritabilités estimées sont plus faibles. Ces résultats suggèrent que les premières PCs n'avaient pas correctement corrigé la stratification de population. L'inclusion des coordonnées géographiques du lieu de naissance serait a priori plus efficace. Nous nous sommes aussi posés la question de la présence d'un effet centre. Nos différentes analyses complémentaires tentent à prouver que l'effet de la première PC est bien due à une stratification de population et non à un effet centre. Cette stratification peut refléter la présence d'un facteur environnemental variant avec la géographie ou un gradient allélique de facteurs génétiques associés à la stature. Dans le deuxième cas, la correction de stratification de population n'est plus nécessaire et provoque même un biais dans nos estimations. Cependant, il est impossible ici de savoir quelle explication reflète la réalité. Pour les deux autres traits anthropométriques affectés par la stratification de population, la circonférence du crâne et le rapport de la circonférence de la taille sur celle des hanches, les mêmes analyses complémentaires ont été faites et donnent les mêmes conclusions [136] (Annexe 5).

Il a été fait le choix de prendre les données génétiques élaguées afin de calculer les PCs utilisées pour corriger la stratification de population. Nous nous sommes alors demandés si

utiliser les PCs calculées sur la totalité des données génétiques influence nos résultats. Nous avons donc refait les analyses principales avec ces nouvelles PCs. Les résultats sont donnés en Annexe 4 et nous constatons qu'ils sont très similaires mais surtout ne contredisent pas nos différentes conclusions.

Chapitre 4

Le gain apporté par les données familiales, l'exemple des paires de germains atteints

Dans le chapitre précédent, nous nous sommes intéressés à l'estimation de l'héritabilité avec le modèle linéaire mixte sur des données en population. Dans les deux prochains chapitres, nous allons nous concentrer sur les données familiales. Ce premier chapitre sur les données familiales a pour but de montrer ce que peut apporter l'utilisation d'un type particulier de données, les paires de germains atteints (paragraphe 1.2.4). Pour cela, nous allons momentanément laisser les modèles mixtes de côté et nous placer dans le cadre de la recherche des facteurs génétiques d'une maladie. Dans ce chapitre, nous allons explorer deux types d'analyses spécifiques aux données de paires de germains atteints pour lesquelles nous disposons uniquement des données génétiques de l'un des germains (germain index) et de son état IBD avec le second germain (figure 1.12). Il est intéressant de noter que le calcul de l'état IBD entre les deux germains ne nécessite pas la totalité des données génétiques des parents et du second germain ; celui-ci peut être estimé à partir de données moins denses, en particulier, issues d'analyses de liaison pré-existantes.

La première application est le test d'association proposé par Perdry *et al.* [160], permettant de tester si un variant est associé à un trait binaire.

Dans le cadre des données cas témoins, afin de gagner en puissance en limitant le nombre de tests, le test d'association est uniquement fait sur des tag-SNPs, c'est-à-dire des SNPs choisis pour résumer la majorité du génome. Lorsqu'une association est trouvée, nous ne savons donc pas si le variant trouvé est le variant causal c'est-à-dire qui agit réellement sur le phénotype, ou s'il est corrélé ou en déséquilibre de liaison avec l'hypothétique SNP causal. Aussi, les données cas-témoins ne permettent pas d'aller au-delà dans l'analyse. Nous allons montrer dans une deuxième application que les paires de germains atteints permettent de faire de l'inférence sur les propriétés de l'hypothétique variant causal en déséquilibre de liaison avec le variant associé au trait d'intérêt. En effet, ce type de données contient de l'information de liaison notamment utilisée pour les études du même nom. Cette information permet alors d'obtenir de l'information sur les variants en DL avec les variants observés. Grâce à cela, nous pouvons faire de l'inférence et acquérir de l'information sur le possible variant causal. Ce type de résultats peut être utilisé, par exemple, pour diriger des études fonctionnelles postérieures.

Ces deux méthodes ont également été appliquées à des données familiales pour la Sclérose en Plaques, une maladie auto-immune.

Nous finirons par quelques résultats à propos des performances de la méthode si le modèle génétique supposé n'est pas le modèle réel.

Ce travail a permis la publication d'un article [161] (Annexe 8.1) et d'un package R intitulé

ASPBay [162].

4.1 Première application : un locus di-allélique observé, test du score

Comme première analyse utilisant des paires de germains atteints, nous allons présenter rapidement le test du score développé par Perdry *et al.* [160] qui permet de tester l'association entre un SNP bi-allélique A/a et une maladie à partir des génotypes et de l'état *IBD* de n cas index et des génotypes de m témoins. Pour cette analyse, nous observons les effectifs suivants :

Génotype	$IBD = 0$	$IBD = 1$	$IBD = 2$	Total
AA ou $G = 0$	n_{00}	n_{01}	n_{02}	$n_{0.}$
Aa ou $G = 1$	n_{10}	n_{11}	n_{12}	$n_{1.}$
aa ou $G = 2$	n_{20}	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.0}$	$n_{.1}$	$n_{.2}$	n

et

Génotype	Effectif
$G = 0$	m_0
$G = 1$	m_1
$G = 2$	m_2
Total	m

où G est le génotype au SNP A codé comme le nombre d'allèles alternatifs a et *IBD* l'état IBD entre les deux germains d'une paire. Afin de tester l'association entre le SNP A et la maladie, nous nous plaçons sous un modèle génétique multiplicatif avec l'hypothèse que les risques relatifs chez le second germain atteint sont les mêmes que ceux du cas index. Nous avons donc les risques génotypiques suivants :

Génotype	$G = 0$	$G = 1$	$G = 2$
Risque ou $\mathbb{P}[Att_1 \mid G_1]$	ψ_0	$\psi_0\psi$	$\psi_0\psi^2$
Risque ou $\mathbb{P}[Att_2 \mid ASP, G_2]$	ψ'_0	$\psi'_0\psi$	$\psi'_0\psi^2$

avec :

- ▷ Att_1 et Att_2 les événements « *germain index atteint* » et « *deuxième germain atteint* »,
- ▷ G_1 et G_2 les génotypes du cas index et du second germain respectivement,
- ▷ ψ_0 et ψ'_0 les risques de base chez les cas index et les germains respectivement,
- ▷ ψ le risque associé à l'allèle a.

Nous supposons aussi que ce SNP respecte les proportions d'Hardy-Weinberg chez les témoins issus de la population générale (paragraphe 1.1.5). Le modèle multiplicatif choisi implique alors que les génotypes chez les cas sont également en proportions d'Hardy-Weinberg. Ainsi, si nous notons f_a et f_A les fréquences des allèles a et A du SNP A, nous avons les fréquences génotypiques suivantes :

Génotype	$G = 0$	$G = 1$	$G = 2$
Fréquence en population ou $\mathbb{P}[G]$	f_A^2	$2f_Af_a$	f_a^2
$P[G \mid \text{Cas}]$	$\dfrac{f_A^2}{(f_A + f_a\psi)^2}$	$\dfrac{2\psi f_Af_a}{(f_A + f_a\psi)^2}$	$\dfrac{\psi^2 f_a^2}{(f_A + f_a\psi)^2}$

Sous cette hypothèse, nous pouvons remarquer que le rapport de cotes ou « *odds ratio* » allélique et le risque relatif ψ sont égaux. En effet, par définition l'odds ratio allélique vaut :

$$\begin{aligned}
\text{OR} &= \frac{\mathbb{P}[\text{Cas}|a]}{\mathbb{P}[\text{Témoins}|a]} \Big/ \frac{\mathbb{P}[\text{Cas}|A]}{\mathbb{P}[\text{Témoins}|A]} \\
&= \frac{\mathbb{P}[a|\text{Cas}]}{\mathbb{P}[a|\text{Témoins}]} \Big/ \frac{\mathbb{P}[A|\text{Cas}]}{\mathbb{P}[A|\text{Témoins}]} \\
&= \frac{\frac{f_a\psi}{f_a+f_a\psi} f_A}{\frac{f_A}{f_A+f_x\psi} f_a} \\
&= \psi
\end{aligned}$$

Nous utiliserons donc par la suite le terme de « OR » pour le paramètre ψ.

Sous ce modèle génétique, la log-vraisemblance s'écrit :

$$\ell(\psi, f_a) = \sum_{i,k} n_{ki} \ln P_{k.i} + \sum_k m_k \ln \mathbb{P}[G = k]$$

avec $P_{k.i} = \mathbb{P}[G_1 = k, IBD = i|ASP]$ les fréquences des cas index ayant le génotype k et un état IBD de i avec leur germains Les valeurs de ces fréquences sont calculées en détail dans l'Annexe 6 en fonction des paramètres f_a et ψ.

Nous voulons alors tester l'association du SNP A avec la maladie :

$$H_0 : \psi = 1 \quad \text{vs} \quad H_1 : \psi \neq 1.$$

D'après l'article [160], le score obtenu à partir de la log-vraisemblance précédente s'écrit :

$$U = U_1 \widehat{f_a} + U_0$$

avec :

- ▷ $U_1 = \sum_{k,i}(2+i)n_{ki}$,
- ▷ $U_0 = \frac{1}{2}\sum_{k,i}(2+i)kn_{ki}$,
- ▷ l'estimateur de la fréquence de a, $\widehat{f_a} = \dfrac{\sum_i n_{1i} + 2\sum_i n_{2i} + m_1 + 2m_2}{2(m+n)}$.

La statistique de score vaut alors :

$$T = \frac{U}{\sqrt{\widehat{\text{Var}(U)}}}$$

avec $\widehat{\text{Var}(U)} = \frac{1}{4}\dfrac{(1-\widehat{f_a})\widehat{f_a}n(19m+n-1)}{n+m}$. Sous l'hypothèse nulle, nous avons $T \sim \mathcal{N}_n(0,1)$, donc, pour faire notre test, il faut comparer ce score au quantile de la loi normale centrée réduite adaptée au seuil de signification voulu. Perdry *et al.* ont démontré que l'information familiale exploitée par ce score apporte un gain de puissance en comparaison du test d'Armitage [163] avec des effectifs égaux qui est le test d'association classique pour des données cas-témoins.

4.2 Deuxième application : l'inférence sur un variant causal non observé

Le score exposé dans la première application s'applique à un seul SNP et, dans le but de gagner de la puissance, il sera typiquement appliqué sur des SNPs choisis pour être en faible déséquilibre de liaison. Nous ne savons donc pas si le SNP associé est causal ou en déséquilibre de liaison avec le SNP causal. Nous allons maintenant nous intéresser au cas où nous n'observons pas le locus causal mais uniquement un locus en déséquilibre de liaison avec lui, le but étant d'inférer les différents paramètres du modèle sous-jacent.

4.2.1 Le modèle génétique

Nous supposons maintenant que nous observons uniquement un locus bi-allélique B en déséquilibre de liaison avec le variant causal A. Les paramètres de ce nouveau modèle génétique sont :

- ▷ les fréquences en population générale des allèles alternatifs aux deux locus, f_a et f_b, et le déséquilibre de liaison (d ou r^2, le coefficient de corrélation). Si nous notons $f_A = 1 - f_a$ et $f_B = 1 - f_b$, nous pouvons en déduire les fréquences haplotypiques :

$$\begin{aligned} f_{AB} &= f_A f_B + d \\ f_{aB} &= f_a f_B - d \\ f_{Ab} &= f_A f_b - d \\ f_{ab} &= f_a f_b + d \end{aligned}$$

- ▷ le risque relatif associé à l'allèle a du locus causal A, ψ, avec un modèle multiplicatif :

Génotype	AA	Aa	aa
Risque	ψ_0	$\psi_0\psi$	$\psi_0\psi^2$

  Nous pouvons noter ici la présence du paramètre ψ_0 représentant le risque de base. Cependant, nous verrons plus loin que ce paramètre disparait de la vraisemblance calculée à partir de ce modèle.

Sous ce modèle, nous obtenons les fréquences génotypiques suivantes au locus observé B chez les cas index :

$$\begin{aligned} P'_{k.i} &= \mathbb{P}[G_{B.1} = k, IBD = i | ASP] \\ &= \sum_{k'} \mathbb{P}[G_{A.1} = k', G_{B.1} = k, IBD = i | ASP] \\ &= \sum_{k'} \mathbb{P}[G_{B.1} = k | G_{A.1} = k', IBD = i, ASP] \times \mathbb{P}[G_{A.1} = k', IBD = i | ASP] \\ &= \sum_{k'} \mathbb{P}[G_{B.1} = k | G_{A.1} = k'] \mathbb{P}[G_{A.1} = k', IBD = i | ASP] \\ &= \sum_{k'} \frac{\mathbb{P}[G_{B.1} = k, G_{A.1} = k']}{\mathbb{P}[G_{A.1} = k']} P_{k.i} \end{aligned} \tag{4.1}$$

avec :

- ▷ $G_{A.1}$ et $G_{B.1} \in \{0, 1, 2\}$ les génotypes du premier germain, le cas index, aux locus A et B,
- ▷ $G_{A.2}$ et $G_{B.2} \in \{0, 1, 2\}$ les génotypes du second germain aux locus A et B,
- ▷ $IBD \in \{0, 1, 2\}$ le nombre d'allèles IBD chez les deux germains.

Dans ce calcul, nous supposons que l'état IBD est le même aux deux locus A et B ce qui est équivalent à négliger le taux de recombinaison entre les locus A et B. Cette hypothèse est légitime car les deux locus considérés A et B sont proches dans le génome (dans le même gène). Les différents termes de cette expression ont déjà été définis ou calculés. En effet, celle-ci dépend des fréquences génotypiques aux locus A et B, $\mathbb{P}[G_{B.1} = k, G_{A.1} = k']$ et $\mathbb{P}[G_{A.1} = k']$, qui s'écrivent en fonction des fréquences haplotypiques données précédemment et donc en fonction des fréquences alléliques f_a et f_b et le déséquilibre de liaison d (tableau 4.1). Le dernier terme $P_{k.i}$ a été calculé précédemment (Annexe 6).

	BB	Bb	bb
AA	f_{AB}^2	$2f_{AB}f_{Ab}$	f_{Ab}^2
Aa	$2f_{AB}f_{aB}$	$2f_{AB}f_{ab} + 2f_{aB}f_{Ab}$	$2f_{Ab}f_{ab}$
aa	f_{aB}^2	$2f_{aB}f_{ab}$	f_{ab}^2

TABLE 4.1 – Fréquences génétiques aux locus A et B.

Pour cette application, nous observons les effectifs suivants :

Génotype	$IBD = 0$	$IBD = 1$	$IBD = 2$	Total
BB	n_{00}	n_{01}	n_{02}	$n_{0.}$
Bb	n_{10}	n_{11}	n_{12}	$n_{1.}$
bb	n_{20}	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.0}$	$n_{.1}$	$n_{.2}$	n

et

Génotype	Effectif
BB	m_0
Bb	m_1
bb	m_2
Total	m

Ainsi, la vraisemblance en fonction des paramètres (ψ, f_a, f_b, d) pour des observations données est :

$$\ell(\psi, f_a, f_b, d) = \sum_{i,k} n_{ki} \ln P'_{k.i} + \sum_k m_k \ln Q_k \tag{4.2}$$

où $Q_k = \mathbb{P}[G_B = k]$ est la probabilité du génotype k pour le locus B en population générale.

4.2.2 L'inférence Bayésienne sur le variant causal

Nous proposons d'utiliser les SNPs associés à la maladie pour récupérer de l'information sur les variants causaux dans la région. Supposons qu'un variant B en déséquilibre de liaison avec le variant causal A est observé. Dans ce cas, il est possible de faire de l'inférence sur le variant A. En particulier, il est possible d'estimer le déséquilibre de liaison entre A et B ainsi que le risque allélique de A. L'inférence est réalisable grâce à l'information de liaison contenue

dans les données familiales, plus précisément dans l'information IBD, et n'est donc pas possible avec des données cas-témoins.

Dans le but de faire cette inférence, nous avons utilisé l'algorithme de Metropolis-Hastings [164] avec la vraisemblance calculée dans la section précédente (équation (4.2)). Nous allons ici regarder l'identifiabilité du modèle pour confirmer la possibilité de faire de l'inférence. Puis, nous exposerons l'algorithme de Metropolis-Hastings ainsi que ses applications.

4.2.3 L'identifiabilité du modèle

Avant de faire de l'inférence en utilisant la vraisemblance écrite dans l'équation (4.2), nous devons vérifier l'identifiabilité des paramètres.

Les paramètres sont identifiables uniquement sous certaines conditions. Nous supposons que $\psi > 0$. Cette hypothèse résulte de la définition de ψ qui est un odds ratio. Nous supposons aussi que $\psi \neq 1$ et $d \neq 0$, le locus A est le vrai locus maladie, et il est en déséquilibre de liaison avec B. Pour notre objectif, cette hypothèse est inoffensive car nous appliquons la méthode uniquement dans le cas d'une association significative entre la maladie et le locus B. Évidemment, si $\psi = 1$ ou $d = 0$, nous ne pouvons pas obtenir de l'information sur le locus A à travers le locus B. En particulier, le paramètre f_a n'est pas identifiable dans ce cas.

Nous remarquons aussi que, même si nous supposons que $\psi \neq 1$ et $d \neq 0$, changer la convention sur les noms des deux allèles du locus non observé A produit un jeu de paramètres équivalent avec $f'_a = 1 - f_a$, $\psi' = 1/\psi$ et $d' = -d$. Ainsi, nous montrons par la suite que, pour un ensemble de paramètres donné (ψ, f_a, f_b, d), les valeurs associées de P'_{ki} et Q_k sont atteintes uniquement pour (ψ, f_a, f_b, d) et $\left(\frac{1}{\psi}, 1 - f_a, f_b, -d\right)$.

Nous pouvons calculer f_b à partir des fréquences observées des génotypes chez les témoins. Donc, il nous reste à calculer trois paramètres f_a, ψ et d. Nous prenons :

$$\begin{aligned}
x_1 &= \frac{2P'_{20}}{2P'_{20} + P'_{10}} - f_b = \frac{(\psi - 1)d}{\psi f_a + (1 - f_a)} \\
x_2 &= \frac{2P'_{22}}{2P'_{22} + P'_{12}} - f_b = \frac{(\psi + 1)(\psi - 1)d}{\psi^2 f_a + (1 - f_a)} \\
x_3 &= \frac{2\sum_i P'_{2i}}{2\sum_i P'_{2i} + \sum_i P'_{1i}} - f_b = \frac{((\psi - 1)f_a + \psi + 2)(\psi - 1)d}{2 + (\psi - 1)f_a((\psi - 1)f_a + \psi + 3)}
\end{aligned}$$

Notons que les paramètres présents dans ces expressions sont uniquement d, f_a et ψ. De plus, les seuls paramètres apparaissant dans les expressions de $\dfrac{x_1}{x_2}$ et $\dfrac{x_1}{x_3}$ sont f_a et ψ. Éliminant f_a entre les deux équations, nous obtenons :

$$\left(\frac{x_1}{x_2} - 1\right)\left(\frac{x_1}{x_2} - \frac{x_1}{x_3}\right)\psi^2 + \left(\left(\frac{x_1}{x_2} - 1\right)\left(\frac{x_1}{x_2} - \frac{x_1}{x_3}\right) + \left(\frac{x_1}{x_3} - 1\right)\right)\psi + \left(\frac{x_1}{x_2} - 1\right)\left(\frac{x_1}{x_2} - \frac{x_1}{x_3}\right) = 0.$$

Les coefficients d'ordre 0 et 2 dans l'équation sont égaux, l'une de ses deux solutions est l'inverse de l'autre, ψ et $\frac{1}{\psi}$. Avec l'expression de $\dfrac{x_1}{x_2}$, nous avons aussi :

$$f_a = (\psi + 1)\frac{x_1}{x_2} - \frac{1}{(\psi + 1)}(\psi - 1)\left(\frac{x_1}{x_2} - 1\right).$$

Les deux solutions ψ et $\frac{1}{\psi}$ correspondent à f_a et $1 - f_a$ respectivement. Finalement, nous avons, pour $\psi \neq 1$,

$$d = \frac{\psi f_a + (1 - f_a)}{(\psi - 1)} x_1,$$

donnant les deux solutions d et $-d$.

4.2.4 L'algorithme de Metropolis-Hastings

Pour faire de l'inférence sur le SNP causal, nous utilisons la marche aléatoire de Metropolis-Hastings. Pour cela, nous avons besoin de distributions instrumentales pour les paramètres afin de les tirer aléatoirement dans leur domaine de définition. Pour commencer, nous utilisons une reparamétrisation de la vraisemblance avec les fréquences haplotypiques. Les trois paramètres f_a, f_b, d sont remplacés par :

$$\begin{aligned}
f_1 &= f_{ab} = f_a f_b + d \\
f_2 &= f_{aB} = f_a(1 - f_b) - d \\
f_3 &= f_{Ab} = (1 - f_a) f_b - d \\
f_4 &= f_{AB} = 1 - f_{ab} - f_{aB} - f_{Ab}
\end{aligned}$$

De façon similaire, nous utilisons $\varphi = \log\psi$ pour le paramètre de risque. Nous avons alors un nouveau vecteur de paramètres $\theta = (\varphi, f_1, f_2, f_3, f_4)$.

Nous utilisons une distribution a priori gaussienne centrée de précision ϵ pour φ dans $\mathbb{R}$, $g(\varphi)$ (si $\epsilon = 0$ la distribution a priori est impropre), et une distribution de Dirichlet a priori

$$\Delta(f_1, f_2, f_3, f_4) \propto \prod_i f_i^{-\frac{1}{2}}$$

pour les fréquences haplotypiques. L'algorithme pour échantillonner N valeurs $\theta^{(1)}, \ldots, \theta^{(N)}$ dans la distribution a posteriori

$$H(\theta) \propto L(e^{\varphi}, f_1, f_2, f_3, f_4) \times g(\varphi) \times \Delta(f_1, f_2, f_3, f_4)$$

se déroule comme suit.

Initialiser $\theta^{(1)}$ avec $\varphi^{(1)} = 0$ et $(f_1^{(1)}, f_2^{(1)}, f_3^{(1)}, f_4^{(1)})$ tiré dans la distribution uniforme sur le simplexe de dimension 3, et répéter les étapes suivantes pour $c = 1, \ldots, N$:

1. Générer un point candidat $\theta^* = (\varphi^*, f_1^*, f_2^*, f_3^*, f_4^*)$ comme suit :
   - ▷ $\varphi^* = \varphi^{(c)} + U$ avec $U \sim \mathcal{N}_n(0, \sigma_\varphi)$
   - ▷ Pour $i = 1, \ldots, 4$, $f_i^* = \dfrac{f_i^{(c)} e^{Z_i}}{\sum_{j=1}^4 f_j^{(c)} e^{Z_j}}$ avec les variables indépendantes $Z_i \sim \mathcal{N}_n(0, \sigma_f)$.
2. Alors
$$\rho = \min\left\{1, \frac{H(\theta^*)}{H\left(\theta^{(c)}\right)} \times \frac{\prod_i f_i^*}{\prod_i f_i^{(c)}}\right\}$$
   et $\theta^{(c+1)} = \theta^*$ avec une probabilité égale à ρ, $\theta^{(c+1)} = \theta^{(c)}$ avec une probabilité égale à $1 - \rho$.

Pour résoudre le problème de l'équivalence entre les deux ensembles de paramètres vue précédemment, nous appliquons la transformation $(\psi, f_a, f_b, d) \mapsto (1/\psi, 1 - f_a, f_b, -d)$ pour tout ensemble de paramètres avec $\psi < 1$ à la fin de l'algorithme d'échantillonnage.

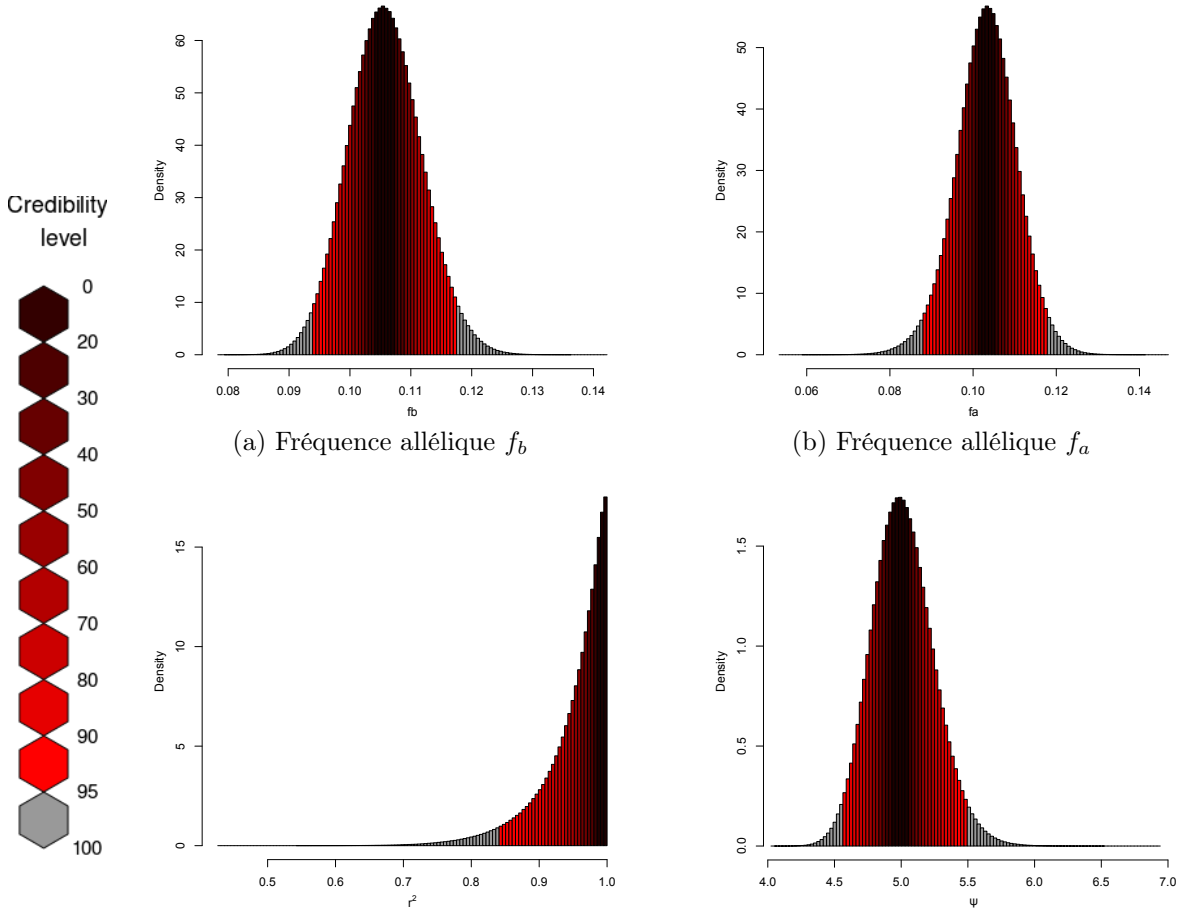
Après avoir généré notre échantillon, plusieurs choses sont possibles afin de contrôler la qualité des données :

- ▷ le "burn-in" qui consiste à ôter les premiers points de l'échantillon afin de considérer uniquement les points déjà dans la distribution limite,
- ▷ le "thinning" qui consiste à prendre uniquement un point toutes les t observations pour réduire la corrélation entre les observations.

4.2.5 Les simulations

Nous avons simulé plusieurs jeux de données composés de 1000 paires de germains atteints et 1000 contrôles afin de tester le comportement de notre méthode. La distribution utilisée pour ces simulations est celle décrite dans la section 4.2.1 (équation (4.1)).

Le premier jeu de données a été simulé sous le modèle ayant deux locus en déséquilibre de liaison complet ($r^2 = 1$), de fréquence de l'allèle alternatif égale à 0.1 et un risque allélique de 5



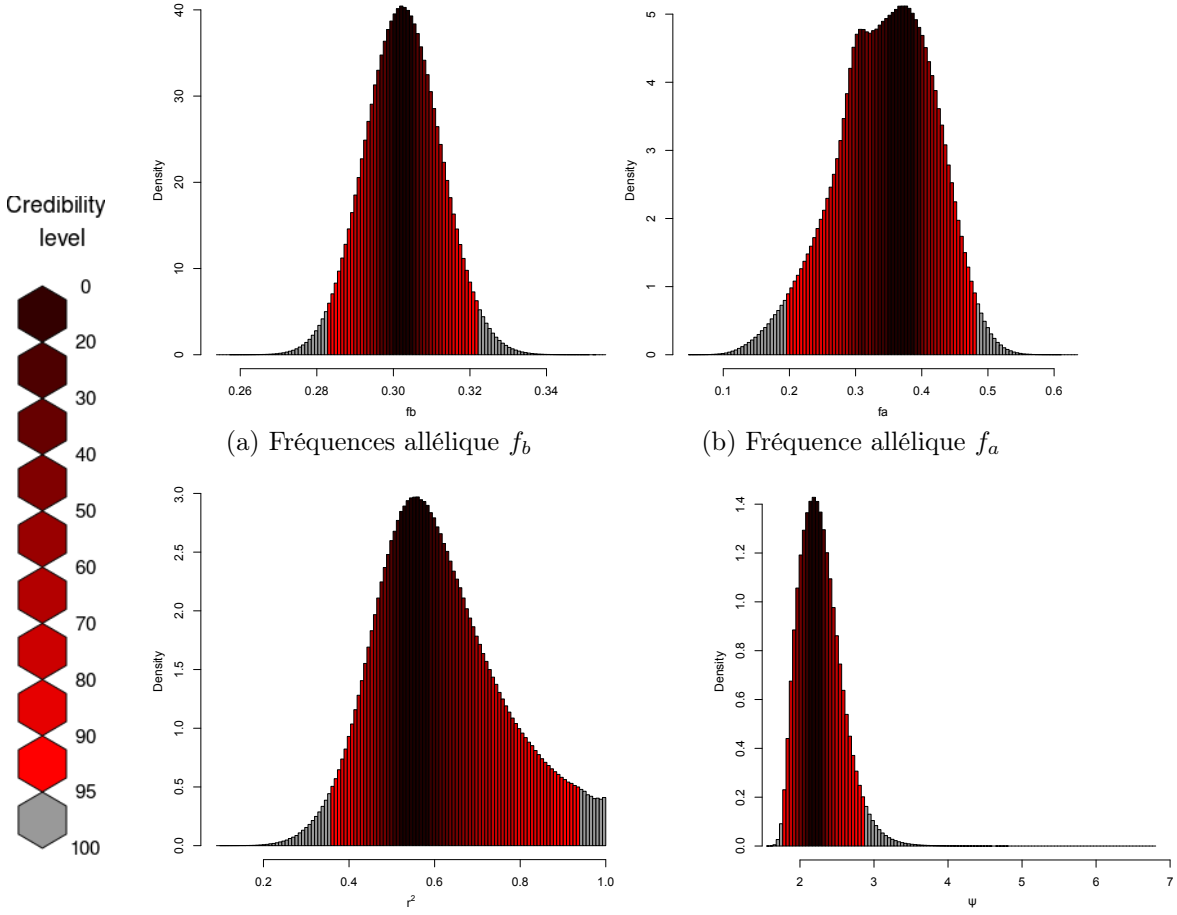
(a) Fréquence allélique f_b (b) Fréquence allélique f_a

(c) Déséquilibre de liaison r^2 (d) Risque allélique ψ

FIGURE 4.1 – Histogrammes pour des simulations avec $r^2 = 1$, $\psi = 5$, $f_a = f_b = 0.1$.

pour le locus causal. Dans ce scénario, rien ne permet de différencier le locus causal du locus observé. Les distributions a posteriori des différents paramètres du modèle f_a, f_b, ψ et r^2 obtenues à partir de nos données simulées sont représentées dans la figure 4.1 sous forme d'histogrammes. Chaque nuance de rouge représente la région de crédibilité pour un seuil donné. Par exemple, un intervalle de crédibilité à 95% est un intervalle qui contient 95% des observations. Nous avons choisi de sélectionner les valeurs les plus observées pour définir les intervalles de crédibilité. Le gris représente la totalité de l'échantillon. Les intervalles de crédibilité à 95% sont $[0.094; 0.118]$ et $[0.088; 0.118]$ pour les fréquences alléliques f_b et f_a, $[0.84; 1]$ pour le déséquilibre et $[4.6; 5.5]$ pour l'odds ratio. Les vraies valeurs sont contenues dans ces intervalles et correspondent aux modes des distributions a posteriori. Il est également visible que la distribution de f_a est plus étalée que celle de f_b ce qui est attendu car le locus B est observé contrairement au locus A. Nous pouvons également remarquer que les distributions a posteriori des fréquences alléliques et du risque allélique semblent approximativement gaussiennes. La distribution du déséquilibre de liaison r^2 est quant à elle asymétrique.

Un second échantillon simulé a été généré sous le modèle génétique ayant deux locus A et B en déséquilibre de liaison avec $r^2 = 0.8$, de fréquence de l'allèle alternatif égale à 0.33 et 0.3 respectivement et un risque allélique de 2 pour le locus A. Les résultats sont ceux de la figure 4.2. Les intervalles de confiance à 95% sont $[0.283; 0.322]$ et $[0.196; 0.482]$ pour les fréquences alléliques f_b et f_a, $[0.358; 0.939]$ pour le déséquilibre et $[1.8; 2.9]$ pour l'odds ratio. Les vraies valeurs sont là aussi contenues dans ces intervalles et très proches des modes des distributions



(a) Fréquences allélique f_b

(b) Fréquence allélique f_a

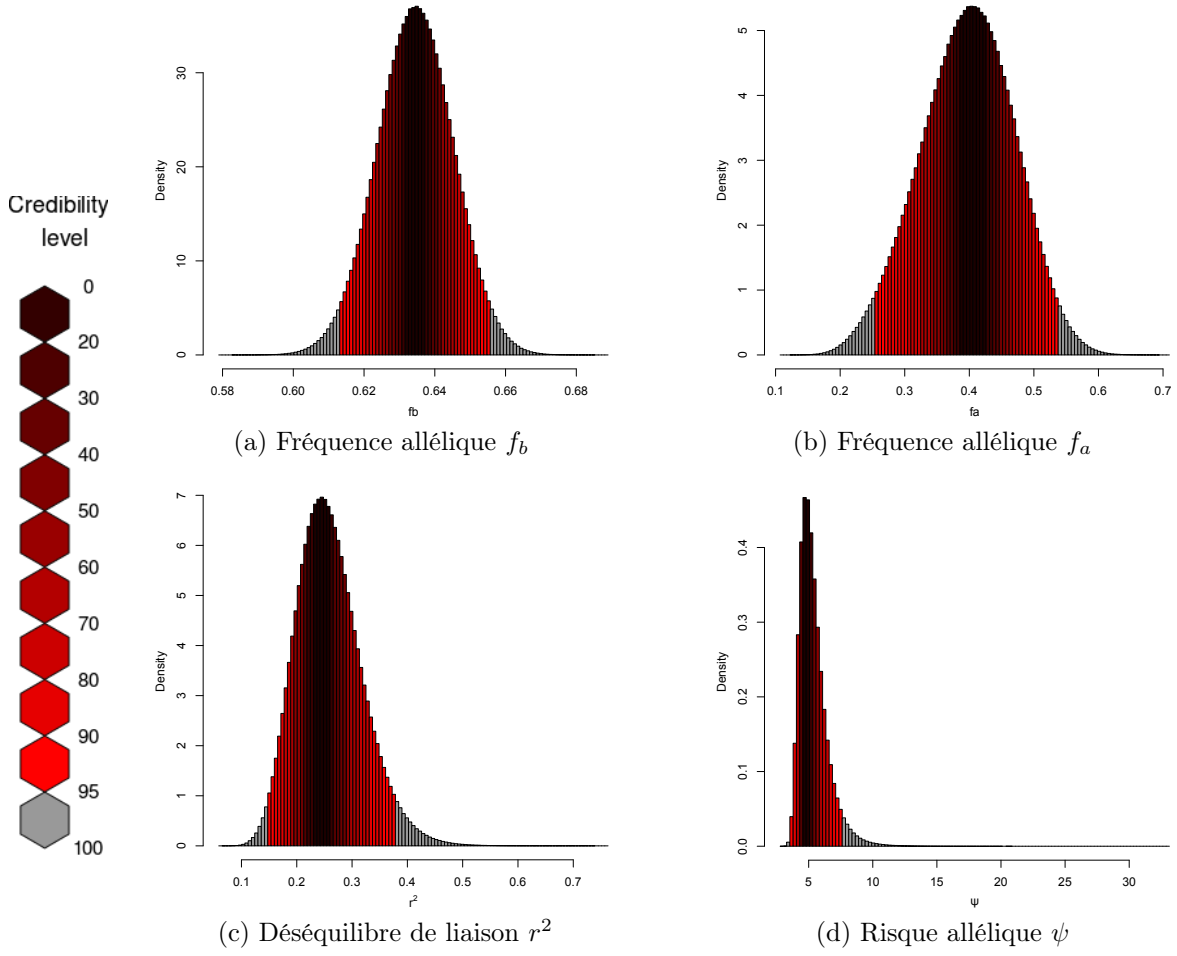
(c) Déséquilibre de liaison r^2

(d) Risque allélique ψ

FIGURE 4.2 – Histogrammes pour des simulations avec $r^2 = 0.8$, $\psi = 2$, $f_a = 0.33$ et $f_b = 0.3$.

a posteriori. La fréquence allélique du locus B est mieux estimée que celle du locus A. Si nous comparons ces intervalles à ceux du premier jeu de données simulées, nous remarquons que l'intervalle de crédibilité à 95% est plus grand pour la fréquence allélique du locus A lorsque le déséquilibre de liaison est moins important. La distribution a posteriori de r^2 est légèrement asymétrique.

Nous pouvons regarder un dernier échantillon simulé sous le modèle avec deux locus en faible déséquilibre de liaison ($r^2 = 0.2$), de fréquence de l'allèle alternatif égale à 0.34 et 0.64 respectivement et un risque allélique de 5 pour le locus A (figure 4.3). Pour ces simulations, les intervalles de confiance à 95% sont, graphiquement, $[0.613; 0.656]$ et $[0.254; 0.538]$ pour les fréquences alléliques f_b et f_a, $[0.148; 0.378]$ pour le déséquilibre et $[3.5; 7.6]$ pour l'odds ratio. Les vraies valeurs sont là aussi contenues dans ces intervalles.



(a) Fréquence allélique f_b

(b) Fréquence allélique f_a

(c) Déséquilibre de liaison r^2

(d) Risque allélique ψ

FIGURE 4.3 – Histogrammes pour des simulations avec $r^2 = 0.2$, $\psi = 5$, $f_a = 0.34$ et $f_b = 0.64$.

4.3 Application sur des données de la Sclérose en Plaques

Nous allons maintenant appliquer le score et l'algorithme de Metropolis-Hastings sur des données de paires de germains atteints par la Sclérose en Plaques (section 1.3).

4.3.1 Les données

Nous allons maintenant tester nos méthodes sur une partie des données récoltées pour une précédente étude [165]. Ces données sont issues de familles françaises avec, au moins, un enfant atteint de la Sclérose en Plaques. Pour cet échantillon, nous disposons de 26 tag-SNPs du gène *IL2RA* situé sur le chromosome 10 et connu comme associé avec la SEP. Le premier à suggérer cette association est Matesanz *et al.* qui a comparé des cas et des témoins pour 4 SNPs de ce gène [166]. Puis, elle a été clairement établie par une étude d'association pangénomique (Genome-Wide Association Study, GWAS) sur des patients anglais et américains [167] et répliquée sur des populations d'origine caucasienne [166–170]. Plusieurs études [171–175] ont trouvé que cette association était, en partie, due au SNP rs2104286 dont la fonction a également été regardée. Une autre étude a montré, à l'aide de données familiales, une association avec une combinaison de SNPs r2256774 et r3118470 [165]. Ici, en utilisant une partie des données de Babron *et al.* [165], nous allons tester l'association de 26 tag-SNPs de ce gène (table 4.2). Puis, nous modéliserons le risque allélique des SNPs associés à la SEP.

#	SNP	Position	Allèles	#	SNP	Position	Allèles
1	rs12359875	6091113	T/C	14	rs4749924	6122402	C/A
2	rs12722605	6093169	T/A	15	rs11598648	6124031	A/G
3	rs12244380	6093380	G/A	16	rs11256497	6127800	A/G
4	rs9663421	6095610	T/C	17	rs791587	6128705	G/A
5	rs12722596	6096300	G/A	18	rs791589	6129577	G/A
6	rs2386841	6097738	A/C	19	rs791590	6130328	T/A
7	rs12722588	6100439	A/G	20	rs10905669	6132099	T/C
8	rs2076846	6103259	G/A	21	rs2476491	6135416	A/T
9	rs12722561	6109899	A/G	22	rs2256774	6137171	G/A
10	rs6602392	6118085	A/C	23	rs2104286	6139051	G/A
11	rs7072398	6119852	A/G	24	rs3118470	6141719	C/T
12	rs11256456	6120718	C/T	25	rs12722489	6142018	A/G
13	rs11256457	6120800	G/C	26	rs12722486	6143768	A/G

Table 4.2 – Noms et positions des SNPs.

Nous avons 522 familles trio (avec un seul enfant malade) et 101 familles avec plus d'un enfant malade. Chaque patient a été vu par un neurologue et diagnostiqué à l'aide du critère de Poser [176]. Ces familles trio ont été utilisées pour créer des pseudo-témoins (figure 1.13). Puis, nous utilisons les familles multiplex pour créer des paires de germains atteints. Pour cela, nous tirons au sort deux germains atteints par famille.

L'état IBD a été calculé à l'aide du logiciel Merlin [177]. Ce logiciel calcule la probabilité de chaque état IBD possible. Nous avons gardé, avec l'état le plus probable, les paires de germains atteints qui avaient une probabilité moyenne sur les 26 SNPs supérieure à 0.8 pour l'un des états IBD.

Après ces différentes étapes, nous obtenons :

- ▹ 522 génotypes témoins,
- ▹ 82 paires de germains atteints avec le génotype du cas index et l'état IBD.

4.3.2 Les résultats

Pour commencer, regardons le déséquilibre de liaison entre les 24 SNPs considérés. La figure 4.4 représente les valeurs du déséquilibre r^2 entre les 24 différents SNPs deux à deux. Outre quelques valeurs de déséquilibre telles que celle entre les SNPs 1 et 4, nous remarquons que les valeurs du déséquilibre entre les différents SNPs sont plutôt faibles. Il faut prendre en compte ce graphique dans notre interprétation des résultats pour l'inférence.

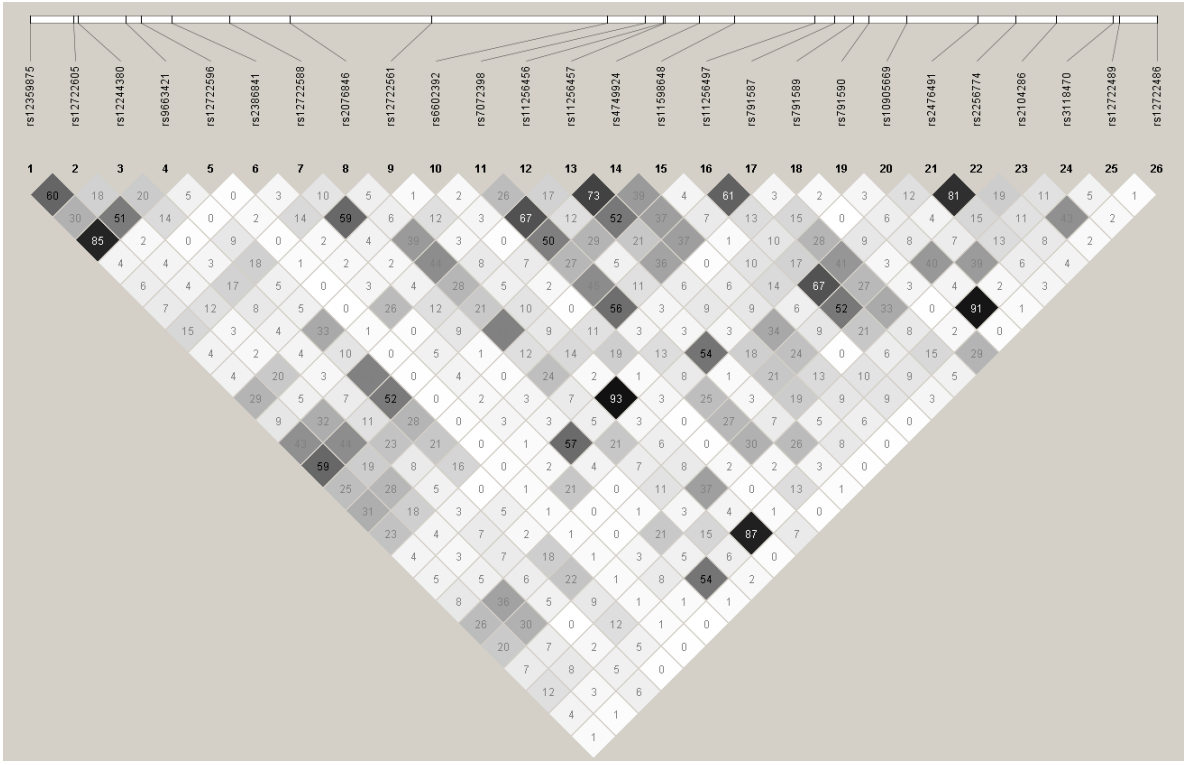


FIGURE 4.4 – Déséquilibre de liaison (r^2).

Nous avons ensuite appliqué le score sur nos données françaises afin de tester l'association entre les 26 SNPs génotypés et la Sclérose en Plaques (table 4.3). Avec une correction de Bonfer-

	Score	p-valeur	$\widehat{\psi}$
SNP 1	9.63	**0.0019**	0.64
SNP 2	4.88	0.027	0.68
SNP 3	3.11	0.078	1.18
SNP 4	7.62	0.0058	0.69
SNP 5	1.70	0.19	1.23
SNP 6	0.01	0.93	0.97
SNP 7	0.64	0.42	0.86
SNP 8	0.21	0.65	1.02
SNP 9	0.27	0.60	1.08
SNP 10	0.26	0.61	0.89
SNP 11	0.53	0.47	1.06
SNP 12	0.002	0.96	0.97
SNP 13	0.17	0.68	0.93

	Score	p-valeur	$\widehat{\psi}$
SNP 14	2.11	0.15	0.84
SNP 15	0.31	0.58	1.02
SNP 16	0.13	0.71	0.85
SNP 17	3.05	0.081	1.20
SNP 18	0.16	0.69	1.02
SNP 19	0.61	0.44	1.11
SNP 20	2.39	0.12	1.20
SNP 21	4.13	0.042	0.76
SNP 22	1.48	0.22	0.87
SNP 23	0.35	0.55	0.93
SNP 24	15.02	**0.00011**	1.52
SNP 25	0.44	0.51	1.09
SNP 26	0.004	0.95	1.01

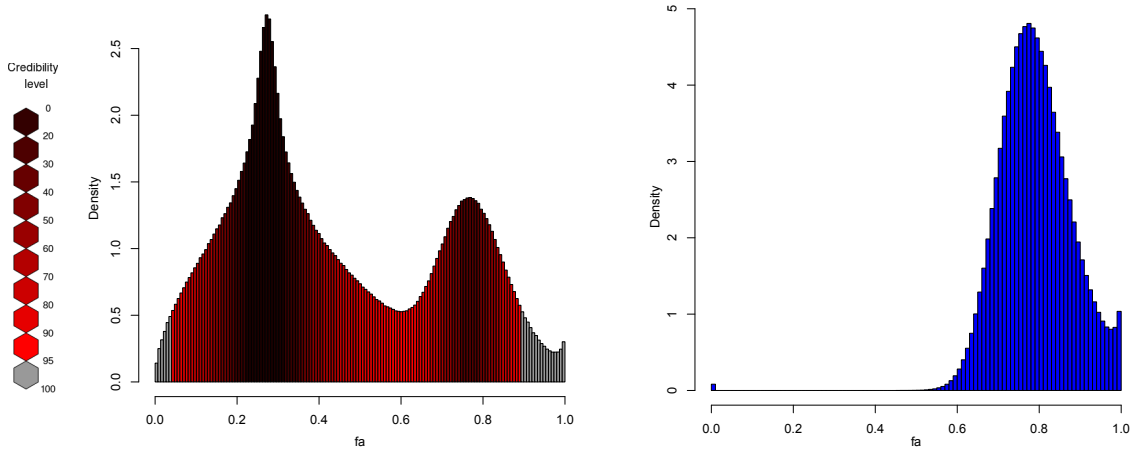
TABLE 4.3 – Valeurs des statistiques et p-valeurs pour les données familiales.

roni qui consiste à prendre comme seuil de significativité $5\%/nombre\ tests = 0.0019$, les SNPs 24 (rs3118470) et 1 (rs12359875) sont les seuls SNPs significativement associés à la Sclérose en Plaques avec un odds ratio pour l'allèle mineur estimé à 1.52 et 0.64 respectivement. Le SNP 4 (rs9663421) obtient une p-valeur proche de la significativité. Nous constatons également que le SNP 23, rs2104286, connu dans la littérature ne ressort pas ici.

Nous allons maintenant appliquer l'algorithme de Metropolis-Hastings sur le SNP 24 que nous avons désigné comme étant le SNP le plus associé parmi les SNPs observés dans notre échantillon ainsi que sur le SNP 1 qui a un p-valeur en-dessous du seuil choisi. Nous allons aussi appliquer l'algorithme de Metropolis-Hastings sur le SNP 4 qui n'est pas très loin de la significativité et en fort DL avec le SNP 1 dans notre échantillon. Il parait en effet intéressant de regarder comment notre méthode se comporte pour ces deux SNPs.

Le SNP 24

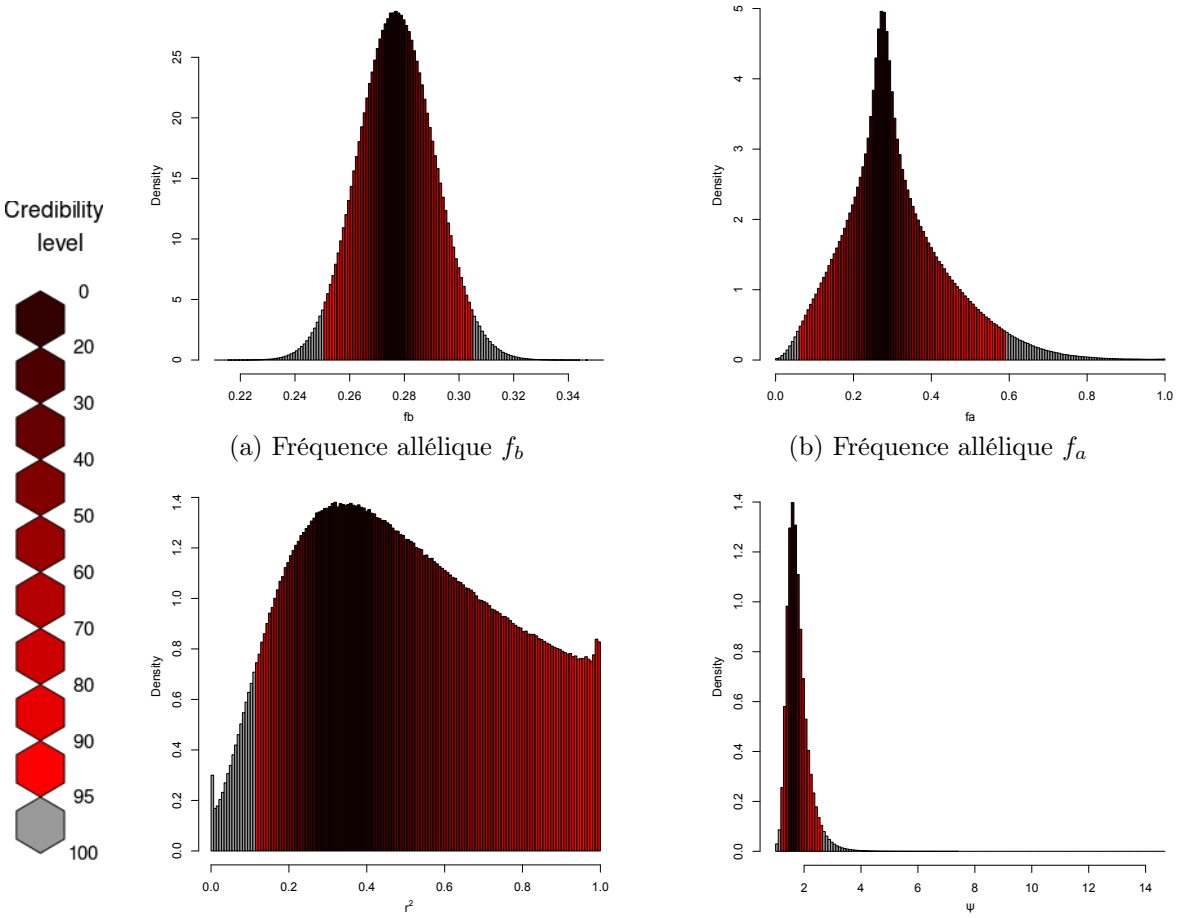
Dans un premier temps, nous avons appliqué l'algorithme de Metroplis-Hastings sans a priori sur la distribution du logarithme de risque, $\varphi = \log(\psi)$. En regardant la distribution de la fréquence de l'allèle délétère du variant causal non observé, nous remarquons deux modes (figure 4.5a). Nous remarquons également de fortes valeurs pour le risque. En allant plus loin, nous constatons que le deuxième mode observé autour de 0.8 est associé à des risques supérieurs à 10 (figure 4.5a). En effet, l'algorithme d'échantillonnage se « perd » dans l'espace de très fortes valeurs de ψ. Nous avons donc refait les analyses avec une distribution a priori gaussienne centrée de précision $\epsilon = 3$ (de variance $1/3$) pour le logarithme du risque φ (nous appliquerons la même distribution a priori pour les autres SNPs étudiés).



(a) Toutes (b) Pour $\psi > 10$

FIGURE 4.5 – Fréquences alléliques pour le SNP non observé.

La nouvelle distribution a posteriori obtenue en appliquant l'algorithme de Metroplis-Hastings sur le SNP 24 est représentée dans la figure 4.6. Les intervalles de crédibilité à 95% sont $[0.251; 0.305]$ et $[0.060; 0.591]$ pour les fréquences alléliques f_b et f_a, $[0.114; 1]$ pour le déséquilibre et $[1.2; 2.7]$ pour l'odds ratio. Les valeurs semblent tout de même très étalées.



(a) Fréquence allélique f_b

(b) Fréquence allélique f_a

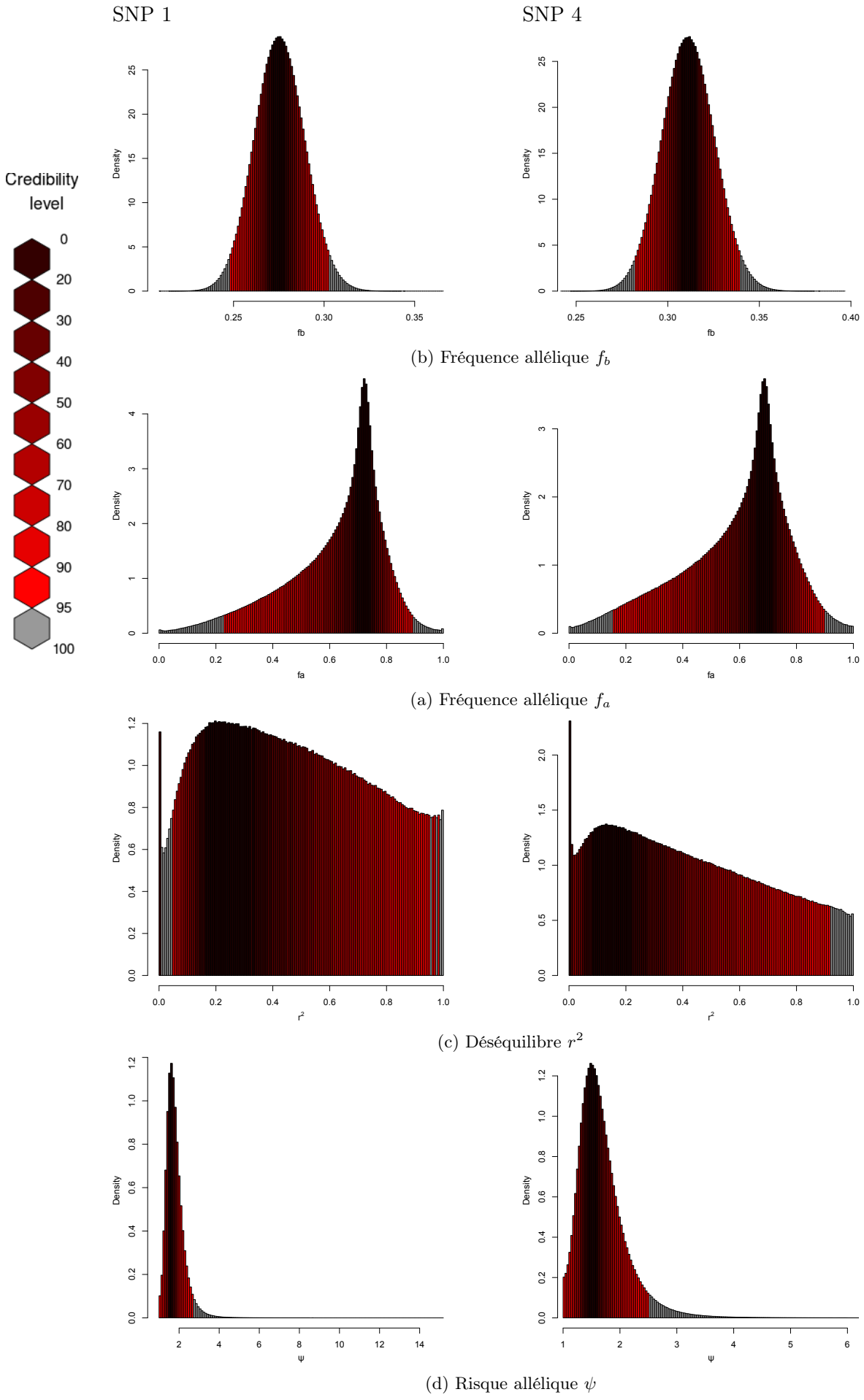
(c) Déséquilibre r^2

(d) Risque allélique ψ

FIGURE 4.6 – Metropolis-Hastings pour le SNP 24.

Les SNPs 1 et 4

Nous allons maintenant regarder conjointement les SNPs 1 et 4 qui sont en fort déséquilibre de liaison ($r^2 = 0.85$). Les résultats sont donnés dans la figure 4.7. Les distributions a posteriori des fréquences alléliques donnent les intervalles de crédibilité à 95% de $[0.248; 0.303]$ et $[0.282; 0.339]$ pour les SNPs 1 et 4 respectivement. Pour la fréquence du locus causal putatif, nous trouvons des distributions a posteriori très semblables pour les deux SNPs avec des intervalles de crédibilité à 95% de $[0.228; 0.893]$ et $[0.154; 0.899]$ et des modes autour de 70%. Pour le déséquilibre de liaison, les distributions a posteriori estimée sur les SNPs 1 et 4 se ressemblent également. Les intervalles de crédibilité à 95% pour r^2 sont $[0.047; 0.953]$ et $[0.154; 0.899]$ respectivement. Ces intervalles sont très larges. Nous avons tout de même un mode autour de 0.3 pour le SNP 1 et autour de 0.2 pour le SNP 4. Cette différence peut s'expliquer par la légère différence des fréquences alléliques des deux SNPs. Nous constatons aussi un « pic » en 0 qui semble difficile à expliquer. Pour finir, le risque allélique du locus causal putatif obtient des intervalles de crédibilité à 95% de $[1; 2.7]$ et $[1; 2.5]$. Ces deux intervalles sont très proches et les modes des deux distributions a posteriori sont tous deux autour de 1.5.



(b) Fréquence allélique f_b

(a) Fréquence allélique f_a

(c) Déséquilibre r^2

(d) Risque allélique ψ

FIGURE 4.7 – Metropolis-Hastings pour les SNPs 1 et 4.

4.4 Conclusions et discussions

Ces analyses reflètent bien l'avantage des données familiales. En effet, grâce à l'information de liaison contenue par l'état IBD dans les paires de germains atteints, nous avons d'abord gagné en puissance pour tester l'association entre un variant et un trait puis nous avons pu inférer les paramètres d'un variant bi-allélique causal non observé dans nos données ce qui est impossible avec des données cas-témoins. Dans les études cas-témoins, afin de détecter le variant causal, l'une des solutions est de faire de l'imputation pour obtenir une couverture du génome assez importante avant de faire des études plus poussées. Cependant, l'imputation repose sur des données en population de référence et n'est donc pas toujours optimale. Au contraire, les données familiales, grâce à l'information de liaison qu'elles contiennent, nous permettent directement de capturer de l'information sur les variants causaux non observés sans nécessiter l'utilisation d'un panel de référence extérieur.

Sur nos données de la Sclérose en Plaques, nous détectons deux signaux d'association avec les SNPs 24 (rs3118470) et 1 (rs12359875) (table 4.3). Nous avons alors pu appliquer l'algorithme de Metropolis-Hastings sur ces deux SNPs. Nous trouvons que ces deux marqueurs refléteraient en fait la présence de deux autres variants non observés causaux en déséquilibre de liaison avec eux. Cette conclusion ne peut cependant se faire que sous l'hypothèse d'un variant causal unique ayant un risque multiplicatif ce qui semble discutable. En effet, si nous regardons la forme des distributions a posteriori pour la fréquence de l'allèle délétère du variant causal putatif, nous constatons que celles-ci sont asymétriques contrairement à ce que nous avons obtenu sur nos simulations. Ces résultats suggèrent que le modèle supposé ne reflète pas la réalité.

De nos jours, la recherche sur les maladies complexes se concentre sur les designs cas-témoins avec énormément de sujets. Cependant, l'utilisation jointe de l'information de liaison et d'association contenue dans les données familiales permet de créer des designs efficaces dans l'étude des maladies complexes. Utiliser l'information de liaison dans les études d'association ne permet pas seulement d'obtenir un gain de puissance [160] mais aussi d'augmenter la capacité à estimer le risque allélique associé aux variants comme illustré dans les papiers ultérieurs pour l'arthrite rhumatoïde [178, 179] et pour la Sclérose en Plaques [165]. Un autre exemple est la méthode MASC [180] qui a été développée dans le but d'exploiter toute l'information contenue dans les données familiales. Le test d'association de Perdry et al. [160] a été construit dans la même idée.

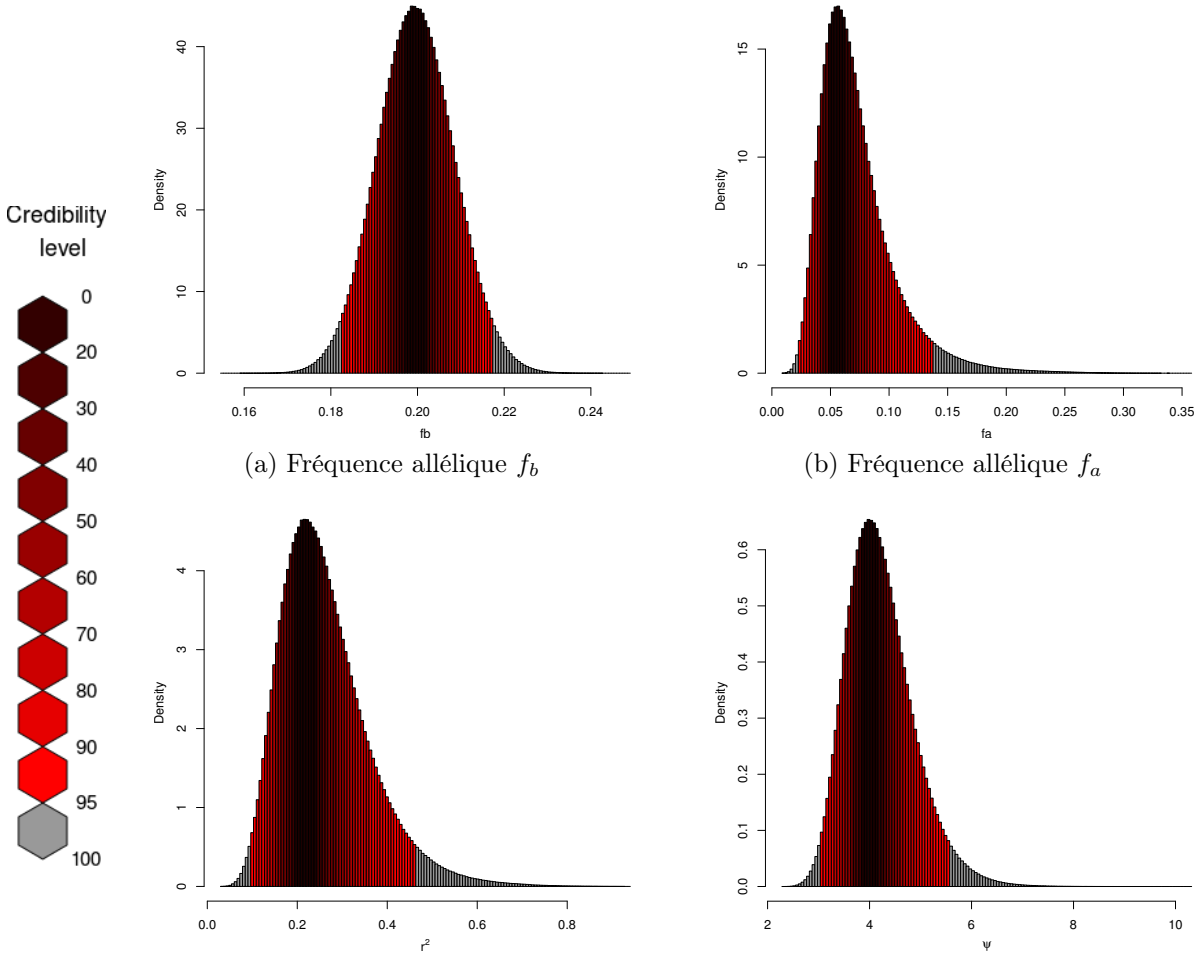
4.5 Pour aller plus loin ...

Nous avons montré que les distributions a posteriori sont satisfaisantes avec des modes très proches des vraies valeurs lorsque nous appliquons l'algorithme de Metropolis-Hastings sur des jeux de données générés sous le modèle supposé par notre méthode. Nous pouvons donc nous demander comment se comporte la méthode lorsque le modèle supposé n'est pas le vrai modèle. Pour cela, nous avons aussi fait des simulations sous un modèle plus complexe avec trois variants ; deux causaux non observés A et C et un observé B en déséquilibre de liaison avec les deux premiers.

Les paramètres pour cet échantillon simulés sont :

- ▷ les fréquences de l'allèle alternatif $f_a = 0.17$, $f_b = 0.2$ et $f_c = 0.12$,
- ▷ les déséquilibres de liaison entre les locus causaux et le locus observé $r^2_{AB} = 0.8$ et $r^2_{BC} = 0.5$,
- ▷ les odds ratio alléliques des variants causaux $\psi_a = 0.5$ et $\psi_c = 5$.

Les distributions a posteriori obtenues sont données dans la figure 4.8. Nous obtenons des intervalles de crédibilité à 95% de $[0.182; 0.217]$ et $[0.023; 0.138]$ pour les fréquences alléliques f_b et f_a, de $[3.0, 5.6]$ pour le risque allélique et de $[0.097; 0.463]$ pour le déséquilibre de liaison. Pour la fréquence allélique du locus observé, nous obtenons un mode très proche de la vraie valeur et un intervalle de crédibilité très petit. Si nous regardons les résultats pour les paramètres du variant causal putatif non observé, nous remarquons que les modes de la fréquence allélique et du déséquilibre de liaison semblent correspondre à la moyenne des fréquences des haplotypes des locus A et C les plus à risque $f_{Ac} = 0.030$ et $f_{ac} = 0.090$ et au déséquilibre de liaison calculé entre le locus B et « l'haplotype moyen » (0.234). Si nous regardons le mode de la distribution du risque allélique donnée par l'algorithme de Metropolis-Hastings, nous constatons qu'il est également proche de la moyenne des risques associés aux deux haplotypes déjà cités (5 pour Ac et 2.5 pour ac). Le modèle n'arrive donc pas à distinguer les deux variants causaux tagués par le même SNP.



(a) Fréquence allélique f_b (b) Fréquence allélique f_a

(c) Déséquilibre de liaison r^2 (d) Risque allélique ψ

FIGURE 4.8 – Metropolis-Hastings un jeux de données simulées avec deux locus causaux.

Chapitre 5

Une application à des données familiales, la Sclérose en Plaques

Nous avons précédemment exploré, d'une part, la méthodologie des modèles mixtes ainsi que leur utilité pour des données en population et, d'autre part, le gain d'information apporté par des données familiales au travers de l'information de liaison. Nous allons maintenant, dans ce chapitre, exploiter l'information familiale à l'aide du modèle mixte. Pour cela, nous avons les génotypes de familles nucléaires françaises pour lesquelles au moins deux enfants sont atteints de la Sclérose en Plaques. Nous allons appliquer plusieurs analyses utilisant les modèles mixtes sur ces données afin d'estimer l'héritabilité de la gravité de la maladie ainsi que rechercher des variants associés à la Sclérose en Plaques.

5.1 Les données

5.1.1 Le contrôle qualité des données génétiques

Nos données sont constituées de 438 individus répartis en 110 familles nucléaires multiplex avec au moins deux enfants atteints de la Sclérose en Plaques issues de l'étude du REFGENSEP. Les individus atteints de la Sclérose en Plaques ont été vus par un neurologue et diagnostiqués à l'aide du critère de Poser [176]. Pour chaque individu, 964 058 variants ont été génotypés à l'aide d'une puce exomique « *Illumina Human Omni Express Exome Chip* » [1]. Cette puce contient les tag-SNPs classiques utilisés pour des GWAS mais aussi environ 270 000 marqueurs fonctionnels situés sur des exons. Elle nous permet donc de capturer plus d'informations sur les variants fonctionnels dont certains sont rares et ne peuvent donc pas être vus au travers des tag-SNPs.

Nous avons fait un contrôle qualité sur les individus et sur les SNPs en utilisant le package R nommé `gaston` [129] que nous avons développé. Pour commencer, nous avons exclu les variants avec un callrate inférieur à 99% ou monomorphes dans notre échantillon. Nous avons également vérifié les proportions d'Hardy-Weinberg chez les fondateurs de notre échantillon (les parents dans notre cas). Les variants autosomaux avec une p-valeur pour les proportions d'Hardy-Weinberg inférieure à 10^{-8} ont été exclus. Nous avons alors fait un contrôle qualité des individus en ne conservant que les individus avec un callrate supérieur à 95% et avec un taux d'hétérozygosité compris entre le taux d'hétérozygosité moyen estimé sur les fondateurs plus ou moins trois fois son erreur-type. Nous avons également vérifié le sexe ainsi que la présence de duplicat dans notre échantillon. Pour finir, nous avons de nouveau exclu les SNPs avec un callrate inférieur à 99% ou monomorphes. Au final, nous conservons 426 individus répartis en

1. `http://www.illumina.com/products/infinium_humanomniexpressexome_beadchip_kits.ilmn`

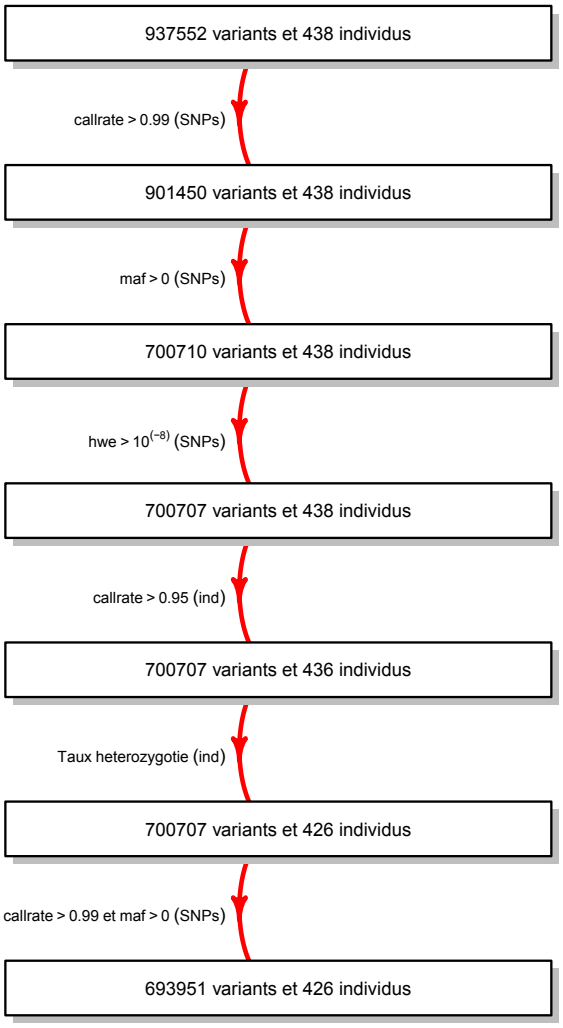


FIGURE 5.1 – Contrôle qualité des données de la Sclérose en Plaques.

104 familles et 712 816 SNPs dont 18 865 sur les chromosomes sexuels. Ces différentes étapes sont résumées dans la figure 5.1. Il a été également vérifié que tous les individus sont d'ascendance européenne en projetant leurs génotypes sur les deux premières composantes principales des Européens de 1000 Genomes [2] (Annexe 6.3).

5.1.2 La matrice de *kinship* et les GRMs

Dans le cas de données familiales, plusieurs matrices de corrélation génétique des individus K sont possibles. La première possibilité est de prendre la matrice 2Φ avec Φ la matrice de *kinship* ou d'apparentement déterminée à partir des liens familiaux. La matrice K peut aussi être estimée à partir des génotypes standardisés. Dans ce cas, K est la matrice de corrélation génétique ou GRM (figure 1.18) que nous avons précédemment calculée pour des données en population. Étant donné que nous avons besoin de cette matrice K dans les différents modèles mixtes possibles afin de corriger l'effet polygénique, nous pouvons nous demander quelle matrice est la plus judicieuse.

La matrice de *kinship* Φ rassemble les coefficients d'apparentement entre les individus de l'échantillon (paragraphe 1.2.4). Nous rappelons que le coefficient d'apparentement entre deux

2. `http://www.1000genomes.org/`

individus est la probabilité que deux allèles d'un même gène tirés au hasard chez chacun des individus soient identiques par descendance (figure 1.12 et table 1.1). Cette première possibilité repose uniquement sur les liens familiaux annoncés qui peuvent malheureusement contenir des erreurs. De plus, dans certains cas, l'information est incomplète et il n'est pas possible de définir le coefficient d'apparentement entre deux individus. Par exemple, pour une famille nucléaire, si l'un des parents n'est pas renseigné, il est impossible de déterminer si les enfants sont des germains ou des demi-germains.

L'autre possibilité est de prendre la matrice de corrélation génétique estimée sur les variants autosomaux. Pour cela, nous pouvons utiliser la totalité de l'information disponible ayant passé le contrôle qualité, ou seulement une partie de cette information en élaguant d'abord les données. En effet, l'utilisation de variants corrélés ou de variants rares peut affecter très fortement l'estimation de la matrice de *kinship*. Nous avons donc comparé les matrices de corrélation génétique estimées à partir de trois jeux de données différents :

- ▷ le premier avec tous les variants,
- ▷ le second contenant uniquement les variants fréquents (ayant une fréquence de l'allèle mineure supérieure à 5%),
- ▷ le troisième avec des données génétiques élaguées. Pour ce dernier jeu de données, nous avons exclu les SNPs avec une fréquence de l'allèle mineur inférieure à 5% ou une p-valeur pour les proportions d'Hardy-Weinberg inférieure à 10^{-5}. Puis, nous avons éliminé des SNPs de façon à ce que le déséquilibre de liaison entre les SNPs restants soit toujours inférieur à 0.1. Ces différentes statistiques descriptives ont été estimées uniquement sur les fondateurs.

Les différentes façons de calculer la GRM que nous venons d'exposer peuvent donner des valeurs très variables [181]. Afin d'explorer cette variabilité, nous avons comparé les valeurs obtenues pour ces différentes matrices de corrélation génétique à la matrice de *kinship* déterminée à partir des liens familiaux.

La première chose à regarder est la différence entre la matrice 2Φ calculée à partir des liens familiaux et celle calculée à partir des données génétiques élaguées. Pour cela, nous regardons la distribution des corrélations génétiques entre chaque paire d'individus en fonction du double de leur coefficient d'apparentement (figure 5.2). Nous constatons alors quelques paires d'individus pour lesquelles la corrélation génétique calculée sur les données génétiques élaguées est nettement différente du coefficient d'apparentement (points entourés en rouge dans la figure) :

- ▷ dans la famille « 30-002 » constituée de deux parents et de deux enfants, nous constatons que la corrélation génétique entre le père et l'un des enfants est quasiment nulle et que celle entre cet enfant et le second germain est autour de 0.25. Ces résultats suggèrent une fausse paternité.
- ▷ dans la famille « MBOB-004 » constituée d'une mère et de ses trois enfants, deux des enfants ont une corrélation génétique très proche de 1. Après vérification, il s'agit en réalité de jumeaux monozygotes.
- ▷ dans la famille « 22-89 » constituée d'une mère et de ses quatre enfants, deux des enfants ont des corrélations génétiques autour de 0.25 avec les deux autres germains. Ces valeurs suggèrent que les quatre germains sont issus de deux pères différents.

La matrice de corrélation génétique estimée sur les données élaguées nous permet ainsi de détecter des liens familiaux mal renseignés et de corriger certaines valeurs de la matrice de *kinship*. Nous remarquons également quelques valeurs assez importantes pour des paires non apparentées, allant jusqu'à 0.1.

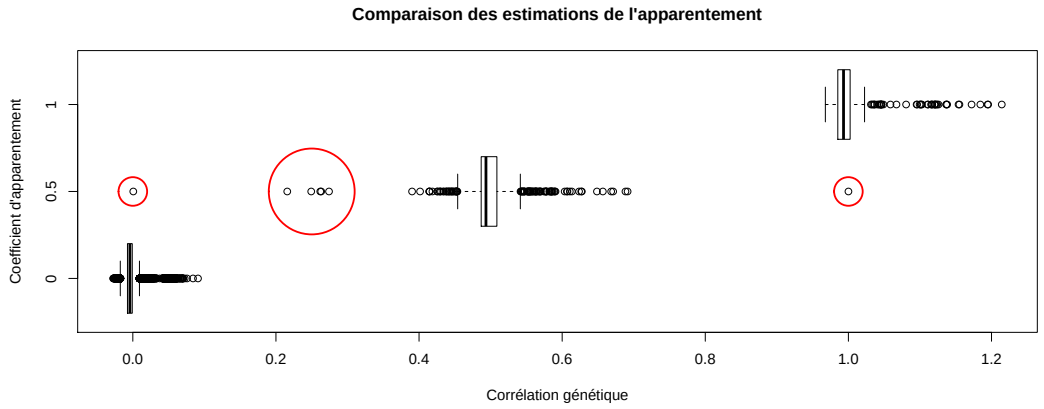


FIGURE 5.2 – Comparaison des estimations des corrélations génétiques sur les données génétiques élaguées avec le double des coefficients d'apparentement.

Une fois la matrice de *kinship* Φ corrigée, nous avons refait la comparaison entre les valeurs de 2Φ et celle des trois GRMs proposées (figure 5.4). Pour commencer, les valeurs des corrélations génétiques correspondent en moyenne, dans les trois cas, au double des coefficients d'apparentement correspondants. Nous constatons aussi que les valeurs des corrélations génétiques sont nettement plus dispersées lorsque nous conservons la totalité des variants pour calculer la GRM. En effet, nous constatons que certaines paires d'individus avec un coefficient d'apparentement à 0.5 ont des corrélations génétiques très proches de 1 lorsque nous utilisons le jeu de données entier. Nous avons également des individus dont les corrélations génétiques avec eux-mêmes atteignent presque la valeur 2. De plus, les valeurs obtenues avec les données élaguées ou les variants fréquents sont très proches. Ce qui suggère qu'exclure les variants rares suffit à obtenir des estimations plus stables. Nous pouvons aussi comparer à l'aide d'un nuage de points les corrélations génétiques. Les deux premières matrices ayant des valeurs très

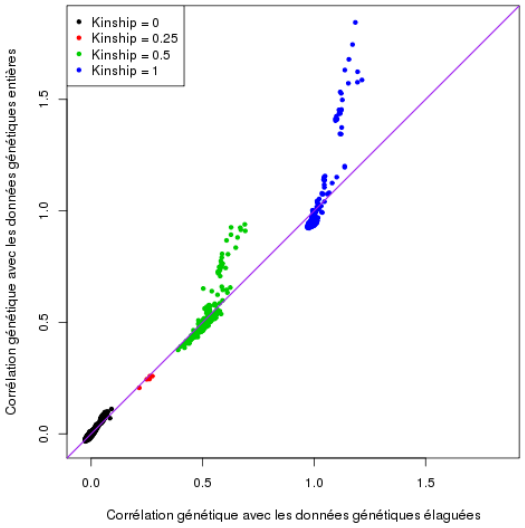


FIGURE 5.3 – Comparaison des deux GRM calculées sur les données génétiques élaguées ou non.

proches, nous avons choisi de comparer les matrices de corrélation génétique utilisant la totalité du génome ou les données élaguées afin de vérifier si les paires d'individus avec des corrélations importantes sont les mêmes (figure 5.3). Nous constatons alors que les valeurs de la GRM obtenue avec les données génétiques entières sont légèrement plus faibles lorsque celles-ci sont proches du coefficient d'apparentement correspondant. À l'inverse, les valeurs des corrélations génétiques déjà nettement plus importantes que le double des coefficients d'apparentement pour les données élaguées le sont d'autant plus pour les données du génome entier. Ce phénomène peut s'expliquer notamment par la présence des variants rares.

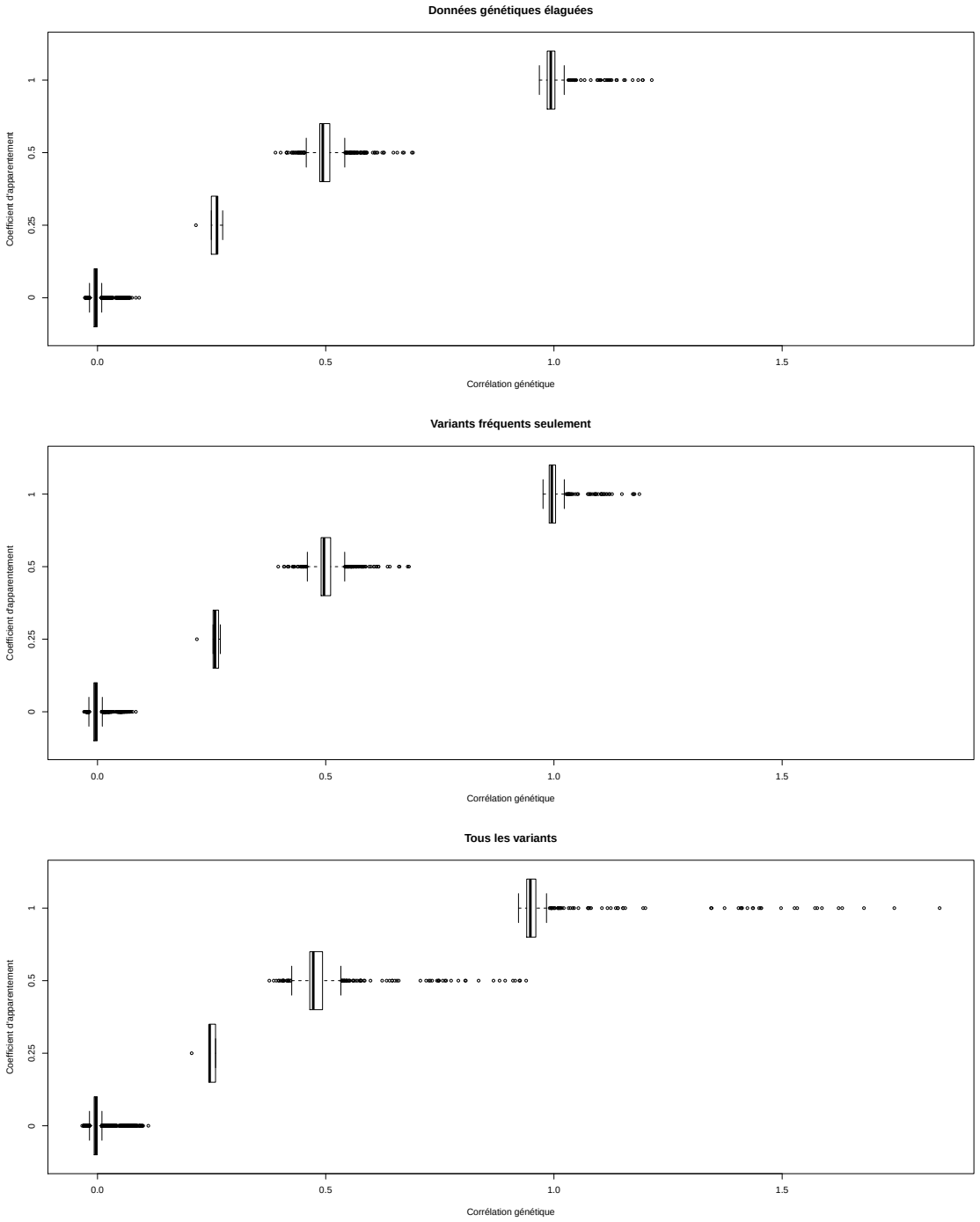


FIGURE 5.4 – Comparaison des corrélations génétiques avec les coefficients d'apparentement calculées à partir des trois jeux de données génétiques.

Dans la suite de ce chapitre, afin de distinguer plus facilement les différentes matrices de corrélation possibles, nous notons :

- ▷ 2Φ la matrice de corrélation calculée à partir des liens parentaux avec Φ la matrice de *kinship*,
- ▷ K_{all} la matrice de corrélation génétique estimée sur les données génétiques entières,
- ▷ K_{elag} la matrice de corrélation génétique estimée sur les données génétiques élaguées.

5.1.3 Les phénotypes

Outre les données génétiques et les liens familiaux, nous disposons de plusieurs données cliniques. La première est le statut de chaque individu pour la Sclérose en Plaques. Dans notre échantillon, après le contrôle qualité, nous avons 215 individus sains et 211 malades. Pour chaque individu, nous avons également quatre autres variables :

- ▷ le sexe,

Sexe	Femme	Homme
Effectif	269	157

- ▷ le centre de recrutement répartis dans toute la France, les deux centres principaux étant Paris et Toulouse,
- ▷ la date de naissance,
- ▷ le nombre de copies des haplotypes du gène HLA-DRB1 reliés au sérotype HLA-DR15 connu comme étant fortement associé à la Sclérose en Plaques [182].

HLA-DR15	X/X	15/x	15/15	NA
Effectif	200	168	15	45

Pour les 211 individus atteints de la Sclérose en Plaques, nous avons également des précisions sur la maladie au travers de trois autres variables (table 5.1) :

- ▷ la date du diagnostique nous permettant d'en déduire l'âge de début de maladie,
- ▷ la forme de la Sclérose en Plaques (section 1.3) codée « RR », « SP » et « PP » pour les formes rémittente-récurrente, secondairement progressive et primaire progressive respectivement,
- ▷ le score de gravité de la Sclérose en Plaques ou « *Multiple Sclerosis Severity Score* » (MSSS), une valeur permettant de quantifier la progression de la maladie [183]. Ce score est déterminé à partir du niveau de handicap de l'individu et de la durée de la maladie.

Forme	RR ($n = 107$)			SP ($n = 35$)			PP ($n = 10$)			Tous ($n = 211$)		
Variables	Moy	Sd	n	Moy	Sd	n	Moy	Sd	n	Moy	Sd	n
Âge début	30.02	8.94	106	28.16	8.29	34	41.92	8.66	10	30.39	9.28	209
MSSS	3.07	2.57	97	6.45	3.38	33	7.93	1.51	9	4.13	2.98	139

Moy = Moyenne et Sd = Écart-type

TABLE 5.1 – Statistiques descriptives pour les individus atteints de la Sclérose en Plaques.

Dans la suite de ce chapitre, nous avons fait le choix de regarder deux traits différents :

- ▷ le score de gravité de la Sclérose en Plaques qui est un trait quantitatif,
- ▷ le statut de la maladie qui est un trait binaire.

Pour le premier trait, nous avons d'abord estimé l'héritabilité. Puis, pour les deux traits, nous avons étudié l'association avec les SNPs disponibles dans notre étude. Afin d'explorer l'effet de la structure familiale ainsi que la qualité de la correction, nous avons comparé les résultats obtenus sous le modèle mixte (section 2.5.2) avec ceux obtenus à l'aide d'un test d'association ajusté que nous proposons pour les données familiales dans la prochaine section.

5.2 Le « *Variance Adjusted Association Test* »

Dans cette section, nous proposons un test d'association simple permettant un ajustement sur la structure familiale. Ce test est proposé comme comparaison suite à des résultats exposés dans la suite de ce chapitre. Pour cela, nous notons y un vecteur de phénotypes quantitatifs ou binaires et G un vecteur de génotypes. Nous considérons ici que y est fixé et G est une variable aléatoire. Nous supposons que y est centré, et que les génotypes G sont centrés et réduits. Nous proposons alors de regarder la statistique de test :

$$S = y'G = \sum_i y_i G_i.$$

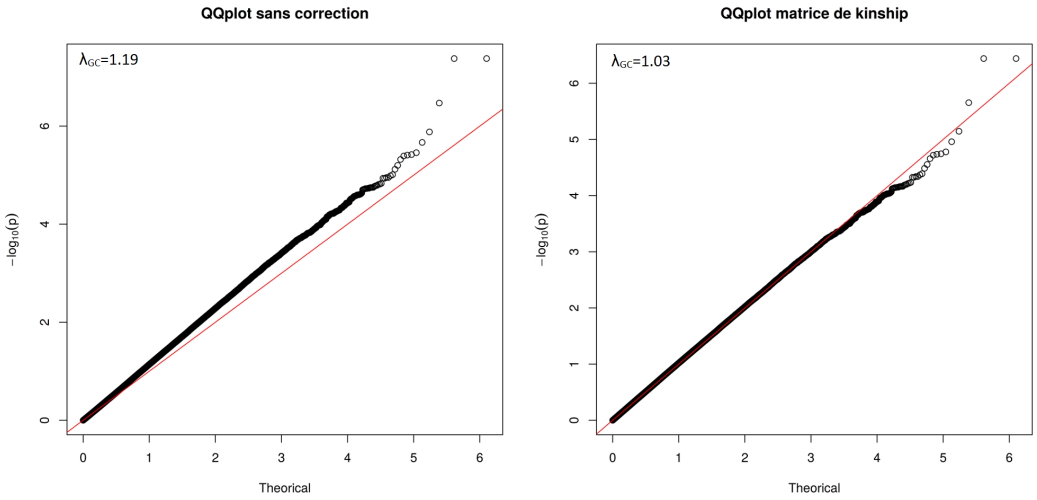
Nous pouvons approximer la loi de cette statistique par une loi normale, d'espérance 0 et de variance $y'Ky$ où $K = \mathrm{Var}(G) = 2\Phi$. Nous obtenons ainsi la statistique approchée :

$$T_v = \frac{(y'G)^2}{y'Ky} \sim \chi^2_1.$$

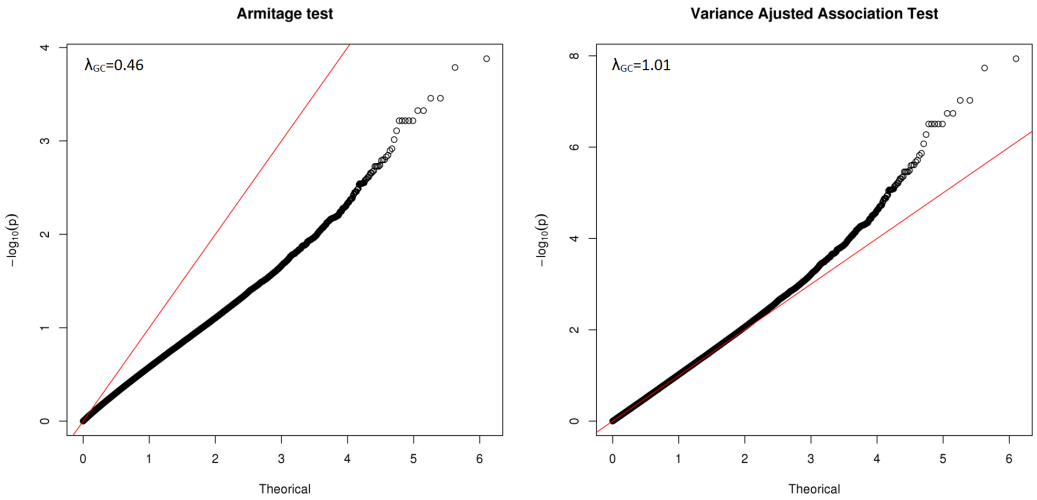
Nous pouvons noter que si les individus ne sont pas apparentés, la matrice de variance-covariance K est la matrice identité. Dans ce cas, la statistique T_v coïncide avec le test d'Armitage [163] pour un trait binaire, et avec un test d'association linéaire pour un trait quantitatif. Dans la suite de ce manuscrit, nous appellerons ce test *Variance Adjusted Association Test* ou VAAT.

Nous avons appliqué notre statistique aux données de Sclérose en Plaques, d'une part, en ignorant la structure familiale ($K = \mathbb{I}_n$) et, d'autre part, en prenant en compte les apparentements ($K = 2\Phi$) afin de vérifier rapidement la pertinence de l'ajustement. Pour cela, nous

avons appliqué le VAAT sur un trait quantitatif, le score de gravité et un trait binaire, le statut de la maladie. Pour chaque phénotype, nous avons d'abord calculé les résidus du modèle linéaire et du modèle logistique pour le score de gravité et le statut respectivement afin de prendre en compte les covariables (le sexe et la forme de la maladie pour le score de gravité et le sexe uniquement pour le statut de la maladie). Le vecteur y utilisé est alors le vecteur des résidus standardisés (pour le trait binaire, nous avons pris les résidus simples, encadré 2.2). Les diagrammes quantile-quantile (QQplot) permettant de comparer les p-valeurs du VAAT aux p-valeurs théoriques pour une loi du χ^2_1, la loi de notre statistique de test sous l'hypothèse nulle, sont donnés dans la figure 5.5. Nous constatons que, lorsque la structure familiale n'est pas prise en compte, la statistique de test ne suit pas une loi du χ^2_1 en particulier pour le trait binaire. Pour le score de gravité quantitatif, la variance de la statistique est plus grande que celle attendue pour une loi de χ^2_1. Pour le statut de la maladie, nous observons l'inverse. Lorsque nous prenons en compte les corrélations entre les génotypes de nos individus dues à la structure familiale, notre statistique de test semble bien suivre une loi du χ^2_1 pour les va-



(a) MSSS (quantitatif)



(b) Statut (binaire)

FIGURE 5.5 – QQ-plot des p-valeurs du test du score sous le modèle mixte pour l'association et du *Variance Adjusted Association Test* pour les comparer à une distribution du χ^2_1. $\lambda_{GC} = \text{médiane}(T_v)/0.456$ est le facteur d'inflation génomique.

riants non associés aux phénotypes. Ces conclusions sont appuyées par le facteur d'inflation génomique dont la valeur s'écarte de 1 en présence de biais.

5.3 Un trait quantitatif, le score de gravité de la Sclérose en Plaques

Dans cette section, nous allons analyser le niveau de gravité de la maladie disponible pour 139 individus malades de notre échantillon Ce score a pour but de quantifier la gravité de la Sclérose en Plaques. Avant de commencer les analyses, nous avons d'abord regardé si des covariables étaient corrélées avec ce trait. Le sexe et la forme de la maladie sont les seules variables qui semblent affecter le score de gravité. Nous avons donc introduit ces covariables dans nos modèles sous forme d'indicatrices. Plus précisément, pour la forme, nous avons décidé de rassembler les formes progressives primaires et secondaires et de les opposer aux formes rémittentes-récurrentes. Ce choix a été motivé par le peu d'individus présentant une forme progressive primaire (7) et la proximité des scores de gravité pour les deux types de Sclérose en Plaques progressives (figure 5.6). Pour le score de gravité de la Sclérose en Plaques, nous avons dans un premier temps estimé son héritabilité restreinte puis nous avons cherché des signaux d'association.

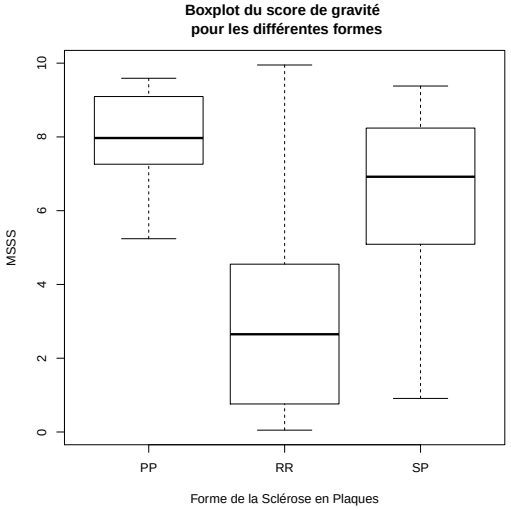


FIGURE 5.6 – Boxplot des valeurs du score de gravité de la Sclérose en Plaques en fonction de la forme de la maladie.

5.3.1 L'estimation de l'héritabilité

Nous commençons par estimer l'héritabilité restreinte du score de sévérité. Pour cela, nous considérons le modèle suivant :

$$MSSS = \beta_0 + S\beta_1 + F\beta_2 + \omega + e$$

avec :

- ▷ $MSSS$ le vecteur des scores de gravité de la Sclérose en Plaques,

- ▷ S le vecteur des sexes valant 0 si l'individu est une femme et 1 sinon,
- ▷ F le vecteur indicatrice valant 1 si la forme de la Sclérose en Plaques est « RR » et 0 sinon,
- ▷ β_0, β_1 et β_2 les effets fixes,
- ▷ $\omega \sim \mathcal{N}_n(0, \tau K)$ avec K le double de la matrice de *kinship* ou la matrice de corrélation génétique estimée sur tout le génome ou sur des données élaguées,
- ▷ $e \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_n)$ le vecteur des erreurs.

Nous avons estimé les composantes pour plusieurs variantes de ce modèle linéaire mixte, incluant ou non la forme, F, en covariable et prenant alternativement les trois matrices de corrélation K possibles citées précédemment. Les résultats sont donnés dans la table 5.2 qui contient les proportions de variance expliquées par les effets fixes, la composante génétique et les résidus, l'estimation de l'héritabilité, h^2, (paragraphe 2.5.3) avec son erreur standard et la p-valeur du test du rapport de vraisemblance pour l'hypothèse nulle $H_0 : h^2 = 0$. Pour commencer, le sexe explique entre 9% et 10% de la variance du score de gravité, la maladie est plus sévère chez les hommes. Lorsque la forme de la maladie est ajoutée dans le modèle, la variance expliquée par les effets fixes monte à un peu plus de 33% majoritairement au détriment de la proportion de variance résiduelle. Dans les modèles ne contenant pas la forme en covariable, pour les différentes matrices K, la proportion de variance génétique est estimée à 47%, 46% et 48% respectivement. Lorsque la forme est introduite dans le modèle, ces proportions augmentent légèrement à un peu plus 48%, 52% et 50% respectivement. Pour finir, l'héritabilité est estimée entre 51% et 52% avec une erreur standard proche de 25% sous les modèles ne contenant pas la forme en covariable et augmente à 73%, 79% et 76% pour chacune des matrices K possibles avec une erreur standard autour de 22% avec l'ajout de la forme en covariable. Les valeurs obtenues lorsque nous changeons de matrice K sont très proches (l'écart entre deux valeurs est toujours inférieur à 30% de l'erreur-type). Dans tous les cas, le test du rapport de vraisemblance est significatif malgré des erreurs standards très importantes dues au faible effectif.

Matrice de corrélation	Forme incluse dans le modèle	Proportions de variance			h^2 (se)	p-valeur
		Effets fixes	Génétique	Résiduelle		
2Φ	Non	0.097	0.467	0.436	0.517 (0.245)	**0.021**
K_{all}	Non	0.094	0.463	0.444	0.511 (0.246)	**0.024**
K_{elag}	Non	0.097	0.475	0.429	0.526 (0.243)	**0.018**
2Φ	Oui	0.336	0.483	0.181	0.727 (0.226)	**0.0018**
K_{all}	Oui	0.333	0.524	0.143	0.786 (0.219)	**0.0009**
K_{elag}	Oui	0.336	0.501	0.163	0.755 (0.223)	**0.0012**

TABLE 5.2 – Estimations des proportions de variance des effets fixes, génétique et résiduelle, de l'héritabilité h^2 avec son erreur-type (se) et LRT, la p-valeur associée au test du rapport de vraisemblance (LRT) pour $H_0 : h^2 = 0$.

Dans cette analyse, nous utilisons des individus apparentés qui ont certainement un environnement commun. Or, il est connu que la présence d'environnement partagé peut biaiser les estimations de l'héritabilité restreinte. Nous pouvons donc soupçonner une surestimation de nos estimations. Il est cependant difficile de quantifier l'environnement partagé et donc d'éventuellement corriger nos estimations.

Afin d'essayer de confirmer nos estimations, nous avons refait l'analyse en sélectionnant aléatoirement un individu par famille dans le but d'avoir uniquement des individus non apparentés. Cette sélection nous donne 78 individus pour notre nouvelle analyse. Ici, la matrice K ne peut être qu'une matrice de corrélation génétique estimée sur les données génétiques élaguées ou non. Le résultat dépendant du tirage, nous avons effectué 100 tirages aléatoires d'individus non apparentés. Les histogrammes des proportions de variance des effets fixes, génétiques et résiduelles estimées pour chacune des réplications sont donnés dans la figure 5.7. Nous constatons que les estimations sont très instables. Cela peut s'expliquer par le très faible effectif utilisé dans cette analyse. De plus, étant donné que les familles ont été recrutées sous la condition qu'elles contiennent au moins deux germains atteints, il est également possible que le mode de recrutement biaise l'analyse. Nous remarquons tout de même que, pour la majorité des tirages, la proportion de variance génétique est estimée autour de 0.7 pour les deux GRMs possibles. Cette analyse complémentaire ne nous rassure que partiellement.

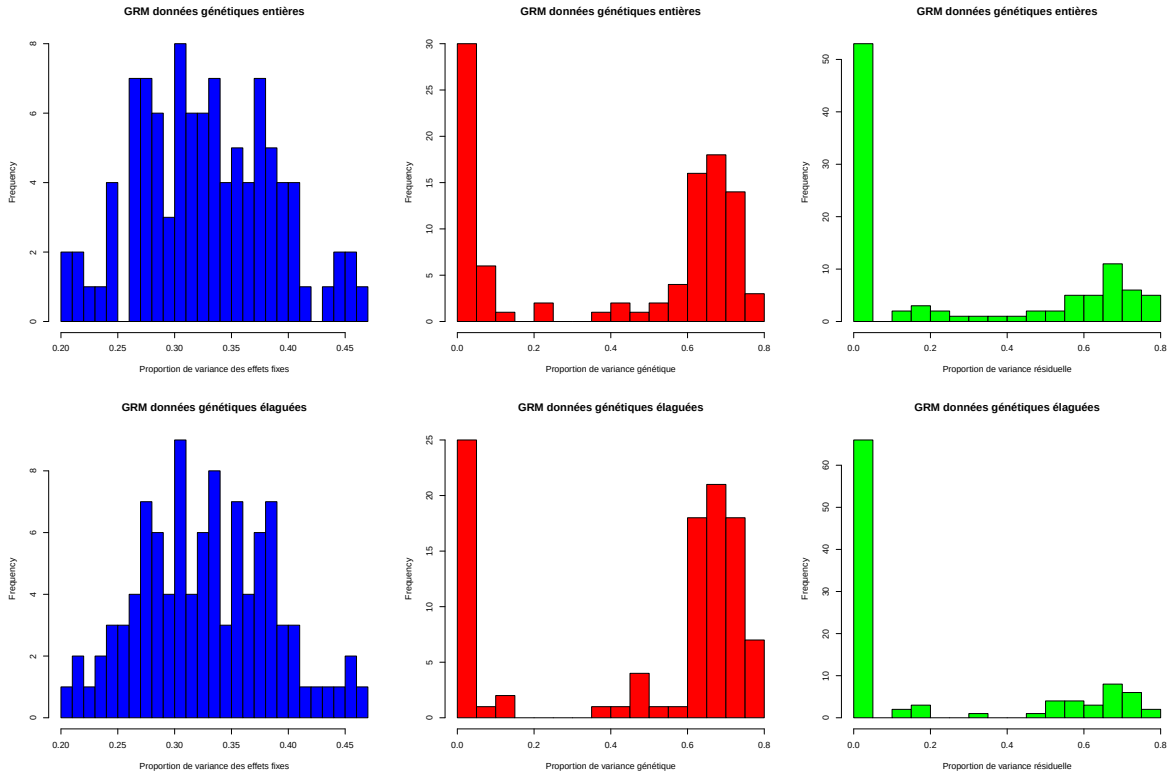


FIGURE 5.7 – Histogrammes des proportions de variance des effets fixes, génétiques et résiduelles estimées pour les 100 tirages.

5.3.2 L'association

Il semble donc que la sévérité de la Sclérose en Plaques soit en partie due à des facteurs génétiques. Nous allons maintenant analyser l'association entre ce trait et les SNPs ayant une fréquence de leur allèle mineur supérieur à 1% disponibles dans nos données. Pour cela, nous avons opté pour l'analyse point par point avec un test du score (section 2.5.2). Nous considérons donc le modèle suivant :

$$MSSS = \beta_0 + S\beta_1 + F\beta_2 + g\gamma + \omega + e$$

avec les mêmes notations que précédemment ainsi que g le vecteur des génotypes au locus d'intérêt et γ l'effet fixe associé. Nous testons alors, à l'aide du score sous le modèle linéaire mixte (paragraphe 2.2.7), l'hypothèse nulle :

$$H_0 : \gamma = 0 \qquad \text{vs} \qquad H_1 : \gamma \neq 0.$$

Sous ce modèle, le terme aléatoire permet de corriger la structure familiale. Les résultats obtenus pour l'estimation de l'héritabilité étant très proches d'une matrice K à l'autre, nous avons choisi, pour l'étude d'association, de nous concentrer sur le modèle avec les effets aléatoires génétiques $\omega \sim \mathcal{N}_n(0, \tau 2\Phi)$ où Φ est la matrice de *kinship*. En effet, les résultats obtenus avec les autres matrices K possibles (GRM) sont très similaires et sont donnés en Annexe 7. Les p-valeurs obtenues avec le test du score sont tracées dans la figure 5.8. Nous constatons qu'aucun variant n'atteint ni n'approche le seuil classique dans les études GWAS de 5.10^{-8}. Avec une p-valeur égale à 7.23×10^{-6}, rs11058793 est le SNP avec la statistique du score la plus importante. Ce SNP du chromosome 12 est sur les gènes LINC00943 et LINC00944 codant non pas des protéines mais des ARNs.

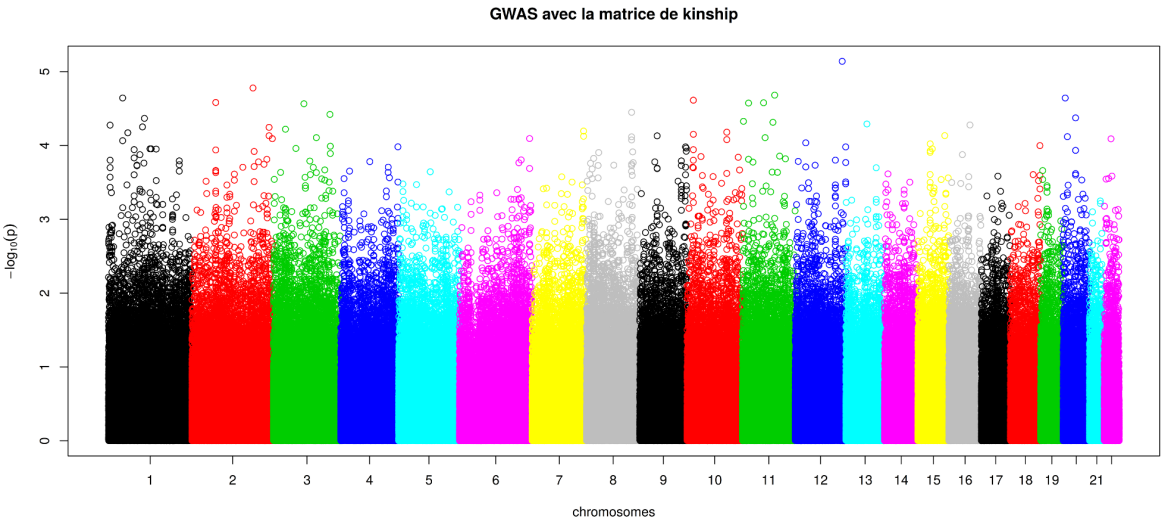


FIGURE 5.8 – P-valeurs du test du score sous le modèle mixte pour l'association. Les couleurs représentent les différents chromosomes.

Afin de valider nos résultats, nous avons regardé le QQplot permettant de comparer nos p-valeurs aux p-valeurs théoriques pour une loi du χ^2_1, la loi du score du modèle linéaire mixte sous l'hypothèse nulle (figure 5.9). Outre les p-valeurs très faibles, la distribution du score est satisfaisante. Ces résultats nous confirment que la correction de la structure familiale est correcte.

Nous avons ensuite refait la même analyse avec le *Variance Adjusted Association Test* (section 5.3) en ajustant sur le sexe et la forme. La distribution de la statistique de ce test, déjà vérifiée précédemment, est satisfaisante (figure 5.5). Les résultats de cette analyse sont donnés dans la figure 5.10. Avec ce nouveau test d'association, nous n'atteignons pas non plus le seuil de significativité classique des GWAS à 5.10^{-8}. Nous distinguons tout de même quelques variants avec des p-valeurs intéressantes inférieures à 10^{-5}. La table 5.3 donne les marqueurs avec une p-valeur inférieure à 10^{-5} pour au moins l'un des deux tests d'association que nous avons appliqués. Pour commencer, le SNP rs11058793 précédemment trouvé avec cette fois-ci une p

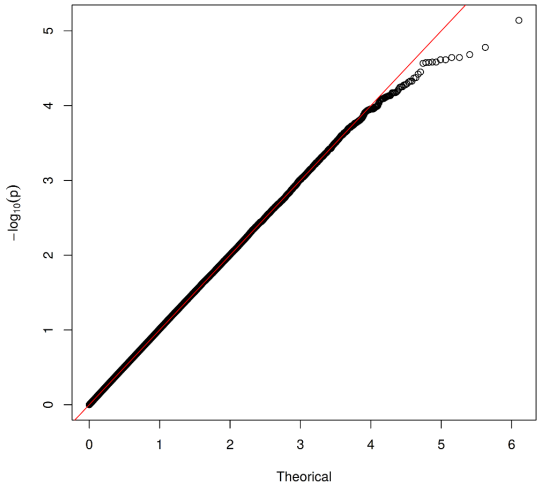


FIGURE 5.9 – QQ-plot des p-valeurs du test du score sous le modèle mixte pour l'association pour les comparer à une distribution du χ^2_1.

valeur un peu plus grande que 10^{-5}. Nous trouvons également un p-valeur intéressante pour un SNP inter-génique, rs12650036, sur le chromosome 4, et trois SNPs sur le chromosome 11 dont deux appartiennent au gène TSPAN32 suppresseur de tumeur et le dernier au gène CNTN5 qui participe au développement du système nerveux, ce qui en fait un candidat intéressant.

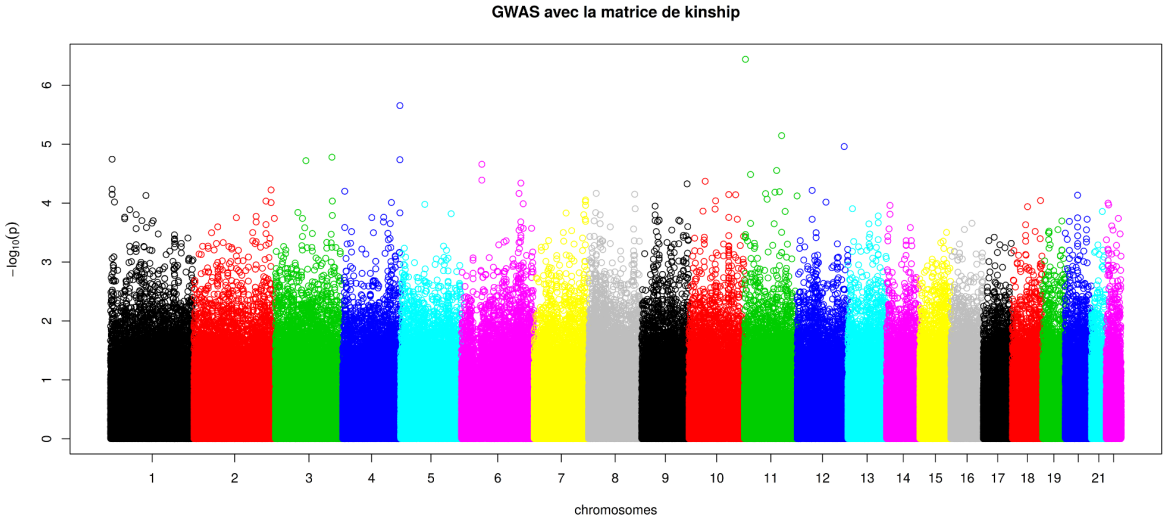


FIGURE 5.10 – P-valeurs du *Variance Adjusted Association Test* pour l'association. Les couleurs représentent les différents chromosomes.

SNP	Chromosome	Gène	p-valeur (lmm)	p-valeur (VAAT)
rs12650036	4	-	1.04e-4	2.21e-6
rs11022157	11	TSPAN32	4.71e-5	3.63e-7
exm878270	11	TSPAN32	4.71e-5	3.63e-7
rs1145408	11	CNTN5	1.39e-4	7.16e-6
rs11058793	12	LINC00943, LINC00944	7.23e-6	1.10e-5

TABLE 5.3 – SNP ayant une p-valeur inférieure à 10^{-5} pour l'un ou l'autre des deux tests d'association.

5.4 Un trait binaire, le statut de la Sclérose en Plaques

Nous allons maintenant étudier l'association avec le statut de la Sclérose en Plaques, valant 1 si l'individu est atteint et 0 sinon, disponible pour tous les individus inclus dans notre étude. Pour cette analyse, nous avons choisi d'introduire le sexe en covariable et d'utiliser une méthode d'étude d'association point par point. Pour cela, nous considérons le modèle :

$$\text{logit}\left(\mathbb{P}[MS_i = 1|S_i, \omega_i]\right) = \beta_0 + S_i\beta_1 + g_i\gamma + \omega_i \tag{5.1}$$

pour $i = 1 \dots n$ avec :

- ▷ MS_i le statut de la Sclérose en Plaques pour l'individu i,
- ▷ S_i le sexe de l'individu i valant 0 si l'individu est une femme et 1 sinon,
- ▷ β_0 et β_1 les effets fixes,
- ▷ g_i le génotype de l'individu i au locus étudié et γ l'effet fixe associé,
- ▷ $\omega = (\omega_1, \dots, \omega_n)' \sim \mathcal{N}_n(0, \tau K)$ avec $K = 2\Phi$ le double de la matrice de *kinship*.

Nous testons alors, à l'aide du score sous le modèle logistique mixte (paragraphe 2.4), l'hypothèse nulle :

$$H_0 : \gamma = 0 \qquad \text{vs} \qquad H_1 : \gamma \neq 0.$$

Le graphique des p-valeurs obtenues avec le test du score sous ce modèle est donné en Annexe 7 ainsi que les résultats obtenus avec les autres matrices de variance-covariance possibles pour les effets aléatoires ω (GRM). Nous constatons uniquement un pic sur le chromosome 6 mais surtout qu'aucun variant n'obtient une p-valeur plus petite que 10^{-4} ce qui est très inférieur au seuil classique utilisé pour les GWAS. Nous avons donc regardé la distribution des p-valeurs obtenues avec cette analyse afin de les comparer à celle d'une loi du χ^2_1. Le QQplot résultant sous le modèle (5.1) est donné dans la figure 5.11. Nous constatons alors que le test du score

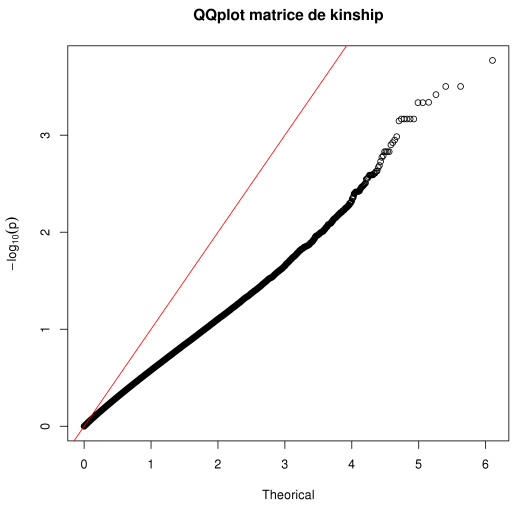


FIGURE 5.11 – QQ-plot des p-valeurs du test du score sous le modèle mixte pour l'association pour les comparer à une distribution du χ^2_1.

donne des valeurs très inférieures à celles attendues pour une loi du χ^2_1 (comme le faisait le test d'Armitage, figure 5.5) ce qui explique les p-valeurs très hautes que nous avons obtenues. Les conclusions sont les mêmes lorsque nous changeons la matrice de corrélation K (Annexe 7).

Si nous regardons plus en détail les estimations obtenues sous l'hypothèse nulle avec le modèle logistique mixte, nous constatons que la composante de la variance associée à la structure familiale est estimée à 0. Le test du score est donc calculé en ignorant la structure familiale, ce qui explique nos résultats.

Afin de mieux comprendre le comportement des estimations sous le modèle mixte, nous avons simplifié le problème en considérant le statut de la maladie comme un trait quantitatif sous l'hypothèse nulle :

$$MS_i = \beta_0 + S_i\beta_1 + \omega_i + e_i$$

avec les mêmes notations que précédemment et $e = (e_1, \ldots, e_n)' \sim \mathcal{N}_n(0, \sigma^2\mathbb{I}_n)$. Ceci nous a permis de regarder les valeurs de la log-vraisemblance en fonction de la valeur de l'héritabilité h^2 ou des composantes de la variance τ et σ^2 (figure 5.12). Nous constatons alors que la log-vraisemblance ne présente pas de réel maximum. Elle augmente lorsque l'héritabilité diminue et, ce, même pour des valeurs négatives jusqu'à ne plus être définie lorsque le déterminant de l'estimation de la matrice de variance du trait devient négatif. Ces résultats expliquent l'échec du modèle logistique mixte dans notre analyse.

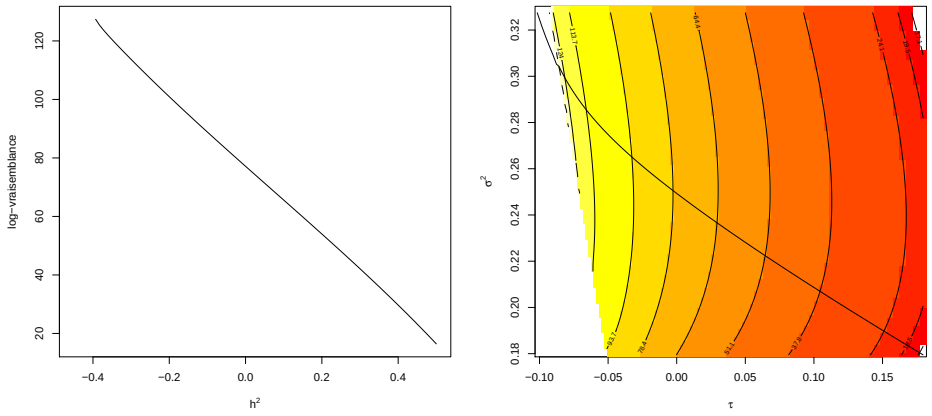


FIGURE 5.12 – Log-vraisemblance du modèle linéaire mixte pour le statut de la Sclérose en Plaques en fonction de l'héritabilité ou des paramètres de la variance.

Nous avons donc appliqué le *Variance Adjusted Association Test* afin d'étudier l'association du statut de la maladie avec les SNPs observés dans notre étude. En effet, nous avons vu, dans la section 5.3, que les faibles statistiques de test obtenues en estimant la matrice de variance-covariance des génotypes par 2Φ semblent bien suivre une loi du χ^2_1 (figure 5.5). Les p-valeurs pour ce test sont données dans la figure 5.13. Un pic sur le chromosome 6 est visible. Il correspond à la région du gène HLA impliqué dans la réponse immunitaire et connu pour être associé à la Sclérose en Plaques. Les SNPs atteignant le seuil classique de significativité pour les GWAS, 5.10^{-8}, sont donnés dans la table 5.4. Ils sont au nombre de 2 et situés dans la région HLA sur le chromosome 6, l'un sur le gène HLA-DRA et l'autre sur le gène HLA-DQB1.

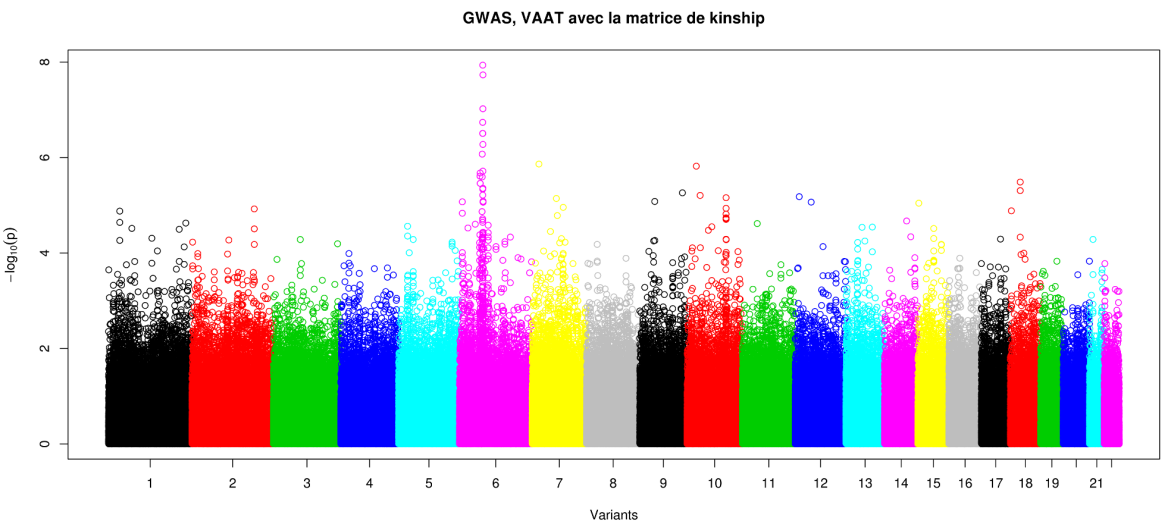


FIGURE 5.13 – P-valeurs du *Variance Adjusted Association Test* pour l'association. Les couleurs représentent les différents chromosomes.

SNP	Chromosome	Gène	p-valeur
rs7197	6	HLA-DRA	1.16e-8
exm-rs9274407	6	HLA-DQB1	1.85e-8

TABLE 5.4 – SNP ayant une p-valeur inférieure à 5.10^{-8}.

Les résultats obtenus en estimant la matrice de variance-covariance de génotypes par une des deux GRMs possibles sont donnés en Annexe 7. Les p-valeurs obtenues et donc nos conclusions sont très similaires.

5.5 Résumé et discussion

Dans ce chapitre, nous avons analysé des données pour la Sclérose en Plaques de familles françaises à l'aide des modèles mixtes. Avant toute chose, nous avons regardé les différences entre les matrices de corrélation possibles pour prendre en compte la structure familiale dans nos modèles mixtes. Nous avons, d'une part, le double de la matrice de *kinship* calculée à partir des liens familiaux et, d'autre part, les matrices de corrélation génétique (GRM) estimées à partir des données génétiques entières ou élaguées. Nous avons observé que les matrices de corrélation génétique nous permettent de confirmer ou d'infirmer les liens familiaux annoncés lors du recrutement. Nous avons notamment pu détecter un enfant n'ayant pas pour père biologique le père de famille. Nous avons également constaté la variabilité des valeurs obtenues lorsque nous calculons une GRM. En effet, celles-ci varient beaucoup d'un individu à un autre mais aussi en fonction des données génétiques utilisées. La difficulté est alors de savoir quelle matrice est la plus adaptée dans l'analyse avec les modèles mixtes. Nous avons ici fait les analyses avec chacune des matrices proposées et avons obtenu des résultats similaires. Dans notre cas, le choix de la matrice de corrélation ne semble pas affecter de façon importante les estimations du modèle mixte.

Nous avons également proposé un test d'association point par point simple mais qui permet

de prendre en compte la structure familiale que nous avons appelé VAAT. Ce test nous a servi, dans ce chapitre, de comparatif avec le modèle mixte pour l'étude d'association variant par variant.

Nous nous sommes alors concentrés sur l'analyse d'un premier trait quantitatif, le score de gravité de la Sclérose en Plaques. Nous avons, dans un premier temps, cherché à estimer son héritabilité. Nous obtenons une héritabilité significativement différente de 0 autour de 75%. Nous avons alors voulu confirmer cette estimation en prenant un individu par famille afin de refaire l'analyse sur des individus non apparentés. Cependant, l'instabilité des estimations au travers des différents tirages ne permet pas de conclure avec certitude. En effet, l'estimation de l'héritabilité peut être discutable car la matrice de corrélation utilisée dans le modèle mixte reflète la ressemblance génétique mais peut aussi refléter un environnement partagé par les familles. Il n'y a malheureusement pas de solution évidente afin de corriger l'estimation de l'héritabilité sous ce modèle.

Pour ce même trait, nous avons ensuite effectué une étude d'association point par point. Le test du score et le VAAT donnent des résultats similaires. Leurs différences résident dans les p-valeurs les plus basses qui paraissent légèrement surestimées pour le test du score sous le modèle mixte ce qui n'est pas le cas pour le VAAT. Suite à ces analyses, cinq SNPs obtiennent une p-valeur inférieure à 10^{-5}. Aucun de ces cinq SNPs n'est sur un gène qui a déjà été trouvé comme associé à la Sclérose en Plaques ce qui fait penser à des faux positifs. De plus, nous avons trouvé dans la littérature deux études d'association avec la sévérité de la Sclérose en Plaques [184, 185] ; aucune de ces deux études ne trouve de SNP significatif au seuil 5.10^{-8}, et aucun de nos 5 SNPs n'est dans les régions les plus associées à la sévérité dans ces études. Le gène CNTN5, qui code une protéine d'adhésion neuronale, reste cependant un candidat intéressant.

Nous nous sommes ensuite intéressés au statut de la Sclérose en Plaques représenté par un trait binaire. Pour ce trait, nous avons effectué une analyse d'association point par point à l'aide du modèle mixte et avons constaté un problème de convergence de celui-ci. Le modèle mixte estime que la structure familiale n'a pas d'effet sur le trait, biaisant, ainsi, les tests d'association. Ce phénomène pourrait s'expliquer par le biais de recrutement dans notre étude. En effet, les familles ont été recrutées sous la condition qu'au moins deux germains soient atteints de la Sclérose en Plaques. De ce fait, le plus souvent les enfants sont atteints (65% des familles n'incluent que des enfants atteints) alors que les parents sont sains (seuls deux parents sont atteints) ce qui induit pour la composante polygénique de la maladie une matrice de variance différente de la matrice de kinship, qui serait valable pour des familles choisies en population générale. Le modèle mixte prenant comme matrice de corrélation la matrice de *kinship* ne parvient pas à modéliser correctement ces données particulières. Nous avons également appliqué, pour ce trait, le test d'association que nous avons proposé dans ce chapitre. Les nouveaux résultats semblent plus fiables et nous permettent de trouver deux SNPs associés à la Sclérose en Plaques dans la région HLA. Étant donné la taille de notre échantillon, nous pouvons supposer que ces résultats sont dus à un manque de puissance.

Une autre solution serait d'adapter la matrice de corrélation K considérée dans le modèle linéaire mixte. Ionita *et al.* [186] ont proposé de prendre pour matrice K la matrice des corrélations des génotypes centrés sur leurs espérances conditionnelles aux génotypes des parents. Cette analyse nécessite de considérer uniquement les enfants. Cependant dans notre échantillon, la plupart des enfants recrutés sont atteints, à l'exception de quelques enfants sains qui ont été

recrutés dans les familles où l'un des parents n'est pas observé. La composition de notre jeu de données fait presque exclusivement d'enfants atteints rend cette méthode inutilisable en l'état.

Nous avions également comme projet de chercher des traces d'origine parentale pour la Sclérose en Plaques en appliquant la méthode que nous avons proposée pour le trait binaire dans le chapitre 2 (section 2.6). Cependant, cette méthode repose sur le modèle logistique mixte déjà utilisé dans l'étude d'association qui a échoué à modéliser correctement les données. Nous avons donc décidé de ne pas analyser la présence d'origine parentale à l'aide du test du score sous le modèle logistique mixte.

Chapitre 6

Perspectives

Afin de faciliter la lecture du présent manuscrit, chaque chapitre contient sa propre conclusion et une courte discussion. Nous allons ici nous concentrer sur les perspectives qu'amènent nos résultats et les futurs travaux possibles qui pourraient nous permettre d'approfondir certains aspects ou répondre à des questions encore sans réponse.

6.1 L'exploitation de l'information de liaison des données familiales

Dans le cadre des études familiales, nous avons d'abord exposé un test du score pour l'association puis une méthode bayésienne, l'algorithme de Metropolis-Hastings permettant de faire de l'inférence sur un variant causal non observé mais représenté par un tag-SNP dans notre échantillon. En effet, nous avons montré que l'information de liaison contenue dans ce type de données permet de capturer de l'information également sur les variants en déséquilibre de liaison avec le SNP observé. Dans ce travail, nous avons utilisé un type particulier de données familiales, les paires de germains atteints pour lesquelles nous ne disposons que des génotypes de l'un des germains ainsi que de son état IBD avec le second germain. Pour ces deux approches, nous pouvons penser à plusieurs extensions. La première est d'augmenter le niveau de complexité du modèle génétique considéré. En effet, nous avons vu dans ce manuscrit que notre méthode ne peut détecter si le vrai modèle génétique contient plus d'un variant causal. D'autres idées peuvent être évoquées pour augmenter la complexité du modèle comme :

- ▷ considérer non plus des variants isolés mais des haplotypes,
- ▷ s'affranchir du modèle génétique multiplicatif,
- ▷ considérer un plus grand nombre de variants causaux possibles.

Augmenter la complexité du modèle peut permettre de s'approcher davantage de la réalité. Cependant, afin de conserver l'identifiablité du modèle et une précision correcte des estimations, la complexité du modèle choisi doit prendre en compte la quantité d'informations disponible.

Une deuxième extension possible est de considérer des données familiales plus complexes afin d'apporter plus d'informations et d'améliorer la précision de l'inférence. Nous pouvons, par exemple, continuer à considérer des paires de germains atteints mais prendre en compte, dans l'analyse, les génotypes des deux germains et des parents. Une autre possibilité serait de considérer des familles plus complexes comme des familles nucléaires multiplexes.

6.2 Les modèles mixtes en génétique

Dans le dernier chapitre de ce manuscrit, nous avons mis en évidence une limite du modèle mixte appliqué aux données familiales. En effet, nous avons montré que le modèle mixte échoue dans la modélisation de données issues de familles sélectionnées à partir du trait étudié (dans notre cas, des familles avec au moins deux enfants atteints). Dans ce cas, la sélection de l'échantillon peut biaiser les estimations de façon très importante rendant invalides les tests d'association basés sur un modèle mixte utilisant une matrice de corrélation qui ne tient pas compte de cette procédure de recrutement. Afin de dépasser cette difficulté, plusieurs solutions peuvent être évoquées :

- ▷ se tourner vers des méthodes classiques construites spécialement pour étudier l'association sur des données familiales telles que FBAT [187–189].
- ▷ se tourner vers des solutions conceptuellement plus simples telles que le VAAT que nous avons utilisé dans le dernier chapitre de ce manuscrit. Afin de valider cette solution, il est cependant nécessaire d'évaluer son comportement par des simulations sous divers modèles. Nous pouvons également réfléchir à des perfectionnements pour ce test d'association.
- ▷ continuer à travailler sous le modèle mixte en modélisant la structure de variance afin qu'elle prenne en compte le mode de recrutement dans la même idée que la méthode proposée dans l'article [186]. Cependant, cette solution paraît complexe à mettre en œuvre.

Dans le deuxième chapitre, nous avons également proposé une méthode permettant de tester la présence d'empreinte parentale avec les modèles mixtes pour un trait binaire (section 2.6). Nous avons alors fait des simulations simples et calculé des puissances estimées qui sont très encourageantes. Il est cependant nécessaire d'aller plus loin dans l'évaluation de cette méthode. Nous pouvons par exemple regarder :

- ▷ l'effet des données manquantes sur le test du score,
- ▷ l'influence sur la puissance d'individus manquants dans certaines familles.

De plus, les problèmes rencontrés pour l'application du modèle mixte à des données issues de familles sélectionnées sur le trait étudié, affecteront probablement de la même façon notre méthode pour la recherche d'empreinte parentale. Nous pouvons, dans un premier temps, évaluer le comportement de la méthode sur nos données ou sur des simulations. Nous pouvons également réfléchir à des stratégies alternatives telles que celles que nous avons évoquées plus haut pour le test d'association.

6.3 Les données de la Sclérose en Plaques

Que ce soit pour la gravité de la Sclérose en Plaques ou le statut de la maladie, les études d'association point par point de nos données n'ont pas permis de trouver des variants significativement associés aux traits étudiés mis à part la région HLA pour le statut. Ces résultats étaient anticipables étant donné la petite taille de notre échantillon. Il n'est pas envisageable,

avec un échantillon de cette taille, de découvrir des variants causaux pour la SEP, qui n'auraient pas été découverts dans les grandes études GWAS qui ont été réalisées. Afin d'exploiter au mieux ces données, nous pouvons notamment :

- ▷ nous tourner vers une stratégie de gènes candidats (en ciblant par exemple, les régions où une association a déjà été découverte, ou encore des *pathways* pertinents pour l'étiologie de la SEP),
- ▷ étudier des phénotypes secondaires qui ont été moins étudiés que la maladie elle-même : la sévérité (pour laquelle nous estimons une héritabilité importante, même si ce résultat doit être pris avec précaution) ou l'âge aux premiers symptômes,
- ▷ exploiter l'information propre aux données familiales, pour la recherche d'empreinte parentale ou pour des méthodes de modélisation basées sur l'utilisation conjointe de l'état IBD et des génotypes.

N'ayant reçu ces données que dernièrement, le temps nous a manqué pour compléter ces analyses. J'espère néanmoins pouvoir explorer un maximum de ces possibilités dans les derniers mois de ma thèse...

Bibliographie

[1] International Human Genome Sequencing Consortium et al. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011) :931–945, 2004.

[2] ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414) :57–74, 2012.

[3] Gregor Mendel. Experiments in plant hybridization. 1865.

[4] Walter S Sutton. The chromosomes in heredity. *The Biological Bulletin*, 4(5) :231–250, 1903.

[5] Fr A Janssens. *La Théorie de la Chiasmatypie...* 1909.

[6] Alfred Henry Sturtevant. The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. *Journal of experimental zoology*, 14(1) :43–59, 1913.

[7] Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *The Journal of experimental medicine*, 79(2) :137–158, 1944.

[8] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356) :737–738, 1953.

[9] Matthew Meselson and Franklin W Stahl. The replication of DNA in escherichia coli. *Proceedings of the national academy of sciences*, 44(7) :671–682, 1958.

[10] KBFF Mullis, Fred Faloona, Stephen Scharf, RK Saiki, GT Horn, and H Erlich. Specific enzymatic amplification of DNA in vitro : the polymerase chain reaction. In *Cold Spring Harbor symposia on quantitative biology*, volume 51, pages 263–273. Cold Spring Harbor Laboratory Press, 1986.

[11] Hamilton O Smith and KW Welcox. A restriction enzyme from haemophilus influenzae : I. Purification and general properties. *Journal of molecular biology*, 51(2) :379–391, 1970.

[12] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235) :467–470, 1995.

[13] Frederick Sanger, Steven Nicklen, and Alan R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12) :5463–5467, 1977.

[14] Allan M Maxam and Walter Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, 74(2) :560–564, 1977.

[15] Eric E Schadt, Steve Turner, and Andrew Kasarskis. A window into third-generation sequencing. *Human molecular genetics*, page ddq416, 2010.

[16] Newton E Morton. Sequential tests for the detection of linkage. *American journal of human genetics*, 7(3) :277, 1955.

[17] W Gregory Feero, Alan E Guttmacher, and Teri A Manolio. Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine*, 363(2) :166–176, 2010.

[18] Peter Donnelly. Progress and challenges in genome-wide association studies in Humans. *Nature*, 456(7223) :728–731, 2008.

[19] TA Manolio. Genomewide association studies and assessment of the risk of disease. *New England Journal of Medecine*, 363(2), 2010.

[20] Michael C Wu, Peter Kraft, Michael P Epstein, Deanne M Taylor, Stephen J Chanock, David J Hunter, and Xihong Lin. Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6) :929–942, 2010.

[21] Andrew P Morris and Eleftheria Zeggini. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic epidemiology*, 34(2) :188–193, 2010.

[22] Jennifer Asimit and Eleftheria Zeggini. Rare variant association analysis methods for complex traits. *Annual review of genetics*, 44 :293–308, 2010.

[23] Stephan Morgenthaler and William G Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases : a cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1) :28–56, 2007.

[24] Fang Han and Wei Pan. A data-adaptive sum test for disease association with multiple common or rare variants. *Human heredity*, 70(1) :42–54, 2010.

[25] Bo Eskerod Madsen and Sharon R Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*, 5(2) :e1000384, 2009.

[26] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89(1) :82–93, 2011.

[27] Seunggeun Lee, Michael C Wu, and Xihong Lin. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4) :762–775, 2012.

[28] Glenys Thomson. Mapping disease genes : family-based association studies. *American Journal of Human Genetics*, 57(2) :487, 1995.

[29] Francis Galton. *Hereditary genius*. Macmillan and Company, 1869.

[30] RA Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(02) :399–433, 1918.

[31] Douglas Scott Falconer. *Introduction to quantitative genetics.* DS Falconer, 1960.

[32] Oscar Kempthorne and Richard H Osborne. The interpretation of twin data. *American Journal of Human Genetics*, 13(3) :320, 1961.

[33] Sandra Scarr. Environmental bias in twin studies. *Eugenics Quarterly*, 15(1) :34–40, 1968.

[34] Sandra Scarr and Louise Carter-Saltzman. Twin method : Defense of a critical assumption. *Behavior genetics*, 9(6) :527–542, 1979.

[35] Brendan Maher. Personal genomes : The case of the missing heritability. *Nature News*, 456(7218) :18–21, 2008.

[36] Peter M Visscher, William G Hill, and Naomi R Wray. Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics*, 9(4) :255–266, 2008.

[37] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability : Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4) :1193–1198, 2012.

[38] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265) :747–753, 2009.

[39] Adrián Blanco-Gómez, Sonia Castillo-Lluva, María del Mar Sáez-Freire, Lourdes Hontecillas-Prieto, Jian Hua Mao, Andrés Castellanos-Martín, and Jesus Pérez-Losada. Missing heritability of complex diseases : Enlightenment by genetic variants from intermediate phenotypes. *BioEssays*, 2016.

[40] Alexander Gusev, Gaurav Bhatia, Noah Zaitlen, Bjarni J Vilhjalmsson, Dorothée Diogo, Eli A Stahl, Peter K Gregersen, Jane Worthington, Lars Klareskog, Soumya Raychaudhuri, et al. Quantifying missing heritability at known GWAS loci. *PLoS Genetics*, 9(12) :e1003993, 2013.

[41] Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88(3) :294–305, 2011.

[42] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7) :565–569, 2010.

[43] Michael Goddard, Hong Lee, Jian Yang, Naomi Wray, and Peter Visscher. Response to Browning and Browning. *American Journal of Human Genetics*, 89(1) :193–195, 2011.

[44] Sharon R Browning and Brian L Browning. Population structure can inflate SNP-based heritability estimates. *American Journal of Human Genetics*, 89(1) :191–193, 2011.

[45] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8) :904–909, 2006.

[46] Yiwei Zhang and Wei Pan. Principal component regression and linear mixed model in association analysis of structured samples : competitors or complements ? *Genetic Epidemiology*, 39(3) :149–155, 2015.

[47] Luc Janss, Gustavo de Los Campos, Nuala Sheehan, and Daniel Sorensen. Inferences from genomic models in stratified populations. *Genetics*, 192(2) :693–704, 2012.

[48] Vincent Segura, Bjarni J Vilhjálmsson, Alexander Platt, Arthur Korte, Ümit Seren, Quan Long, and Magnus Nordborg. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics*, 44(7) :825–830, 2012.

[49] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7) :459–463, 2010.

[50] Denise P Barlow. Gametic imprinting in mammals. *Science*, 270(5242) :1610–1613, 1995.

[51] Quentin Vincent, Alexandre Alcaïs, Andrea Alter, Erwin Schurr, and Laurent Abel. Quantifying genomic imprinting in the presence of linkage. *Biometrics*, 62(4) :1071–1080, 2006.

[52] Clarice R Weinberg. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *The American Journal of Human Genetics*, 65(1) :229–235, 1999.

[53] Gonçalo R Abecasis, William OC Cookson, and Lon R Cardon. Pedigree tests of transmission disequilibrium. *European Journal of Human Genetics*, 8(7) :545–551, 2000.

[54] GR Abecasis, LR Cardon, and WOC Cookson. A general test of association for quantitative traits in nuclear families. *The American Journal of Human Genetics*, 66(1) :279–292, 2000.

[55] Nadezhda M Belonogova, Tatiana I Axenovich, and Yurii S Aulchenko. A powerful genome-wide feasible approach to detect parent-of-origin effects in studies of quantitative traits. *European Journal of Human Genetics*, 18(3) :379–384, 2010.

[56] Fred D Lublin, Stephen C Reingold, et al. Defining the clinical course of multiple sclerosis results of an international survey. *Neurology*, 46(4) :907–911, 1996.

[57] ARSEP. `www.arsep.org`.

[58] Sarah-Michelle Orton, Blanca M Herrera, Irene M Yee, William Valdar, Sreeram V Ramagopalan, A Dessa Sadovnick, George C Ebers, Canadian Collaborative Study Group, et al. Sex ratio of multiple sclerosis in Canada : a longitudinal study. *The Lancet Neurology*, 5(11) :932–936, 2006.

[59] Lazaros Belbasis, Vanesa Bellou, Evangelos Evangelou, John PA Ioannidis, and Ioanna Tzoulaki. Environmental risk factors and multiple sclerosis : an umbrella review of systematic reviews and meta-analyses. *The Lancet Neurology*, 14(3) :263–273, 2015.

[60] J Pakpoor, G Disanto, JE Gerber, R Dobson, UC Meier, G Giovannoni, and SV Ramagopalan. The risk of developing multiple sclerosis in individuals seronegative for Epstein-Barr virus : a meta-analysis. *Multiple Sclerosis Journal*, 19(2) :162, 2013.

[61] Yahya H Almohmeed, Alison Avenell, Lorna Aucott, and Mark A Vickers. Systematic review and meta-analysis of the sero-epidemiological association between Epstein Barr virus and multiple sclerosis. *PloS one*, 8(4) :e61110, 2013.

[62] O Santiago, J Gutierrez, A Sorlozano, J de Dios Luna, E Villegas, and O Fernandez. Relation between Epstein-Barr virus and multiple sclerosis : analytic study of scientific production. *European journal of clinical microbiology & infectious diseases*, 29(7) :857–866, 2010.

[63] Ciro Valent Sumaya, Lawrence W Myers, and George W Ellison. Epstein-Barr virus antibodies in multiple sclerosis. *Archives of neurology*, 37(2) :94, 1980.

[64] SA Sargsyan, AJ Shearer, AM Ritchie, MP Burgoon, S Anderson, B Hemmer, C Stadelmann, S Gattenlöhner, GP Owens, D Gilden, et al. Absence of Epstein-Barr virus in the brain and CSF of patients with multiple sclerosis. *Neurology*, 74(14) :1127–1135, 2010.

[65] Steve Simpson, Leigh Blizzard, Petr Otahal, Ingrid Van der Mei, and Bruce Taylor. Latitude is significantly associated with the prevalence of multiple sclerosis : a meta-analysis. *Journal of Neurology, Neurosurgery & Psychiatry*, 82(10) :1132–1141, 2011.

[66] Ruth Dobson, Gavin Giovannoni, and Sreeram Ramagopalan. The month of birth effect in multiple sclerosis : systematic review, meta-analysis and effect of latitude. *Journal of Neurology, Neurosurgery & Psychiatry*, pages jnnp–2012, 2012.

[67] P Cabre, Al Signate, S Olindo, H Merle, D Caparros-Lefebvre, O Bera, and D Smadja. Role of return migration in the emergence of multiple sclerosis in the French West Indies. *Brain*, 128(12) :2899–2910, 2005.

[68] Nils Koch-Henriksen and Per Soelberg Sorensen. Why does the north–south gradient of incidence of multiple sclerosis seem to have disappeared on the Northern hemisphere ? *Journal of the Neurological Sciences*, 311(1) :58–63, 2011.

[69] GC DeLuca, SM Kimball, J Kolasinski, SV Ramagopalan, and GC Ebers. Review : the role of vitamin D in nervous system health and disease. *Neuropathology and Applied Neurobiology*, 39(5) :458–484, 2013.

[70] Lauren E Mokry, Stephanie Ross, Omar S Ahmad, Vincenzo Forgetta, George Davey Smith, Aaron Leong, Celia MT Greenwood, George Thanassoulis, and J Brent Richards. Vitamin D and risk of multiple sclerosis : a mendelian randomization study. *PLoS Medicine*, 12(8) :e1001866, 2015.

[71] Peng Zhang, Rui Wang, Zhijun Li, Yuhan Wang, Chunshi Gao, Xin Lv, Yuanyuan Song, and Bo Li. The risk of smoking on multiple sclerosis : a meta-analysis based on 20,626 cases from case-control and cohort studies. *PeerJ*, 4 :e1797, 2016.

[72] Fotini Pittas, Anne-Louise Ponsonby, Ingrid AF van der Mei, Bruce V Taylor, Leigh Blizzard, Patricia Groom, Obioha C Ukoumunne, and Terry Dwyer. Smoking is associated with progressive disease course and increased progression in clinical disability in a prospective cohort of people with multiple sclerosis. *Journal of Neurology*, 256(4) :577–585, 2009.

[73] R Zivadinov, B Weinstock-Guttman, K Hashmi, N Abdelrahman, M Stosic, M Dwyer, Hussein S, J Durfee, and M Ramanathan. Smoking is associated with increased lesion volumes and brain atrophy in multiple sclerosis. *Neurology*, 73(7) :504–510, 2009.

[74] Peter Sundström, Lennarth Nyström, and Göran Hallmans. Smoke exposure increases the risk for multiple sclerosis. *European journal of neurology*, 15(6) :579–583, 2008.

[75] Anna Karin Hedström. Tobacco and multiple sclerosis susceptibility. 2016.

[76] DA Dyment, MZ Cader, CJ Willer, N Risch, AD Sadovnick, and GC Ebers. A multigenerational family with multiple sclerosis. *Brain*, 125(7) :1474–1482, 2002.

[77] Stefanie Binzer, K Imrell, M Binzer, S Vang, B Rogvi-Hansen, J Hillert, and E Stenager. Multiple sclerosis in a family on the Faroe islands. *Acta Neurologica Scandinavica*, 121(1) :16–19, 2010.

[78] Corrado Fagnani, Michael C Neale, Lorenza Nisticò, Maria A Stazi, Vito A Ricigliano, Maria C Buscarinu, Marco Salvetti, and Giovanni Ristori. Twin studies in multiple sclerosis : A meta-estimation of heritability and environmentality. *Multiple Sclerosis Journal*, 21(11) :1404–1413, 2015.

[79] VV Bashinskaya, OG Kulakova, AN Boyko, AV Favorov, and OO Favorova. A review of genome-wide association studies for multiple sclerosis : classical and hypothesis-driven approaches. *Human Genetics*, 134(11-12) :1143–1162, 2015.

[80] F Clerget-Darpoux, A Govaerts, and N Feingold. HLA and susceptibility to multiple sclerosis. *Tissue Antigens*, 24(3) :160–169, 1984.

[81] Theresa Dankowski, Dorothea Buck, Till FM Andlauer, Gisela Antony, Antonios Bayas, Lukas Bechmann, Achim Berthele, Thomas Bettecken, Andrew Chan, Andre Franke, et al. Successful replication of GWAS hits for multiple sclerosis in 10,000 Germans using the exome array. *Genetic Epidemiology*, 39(8) :601–608, 2015.

[82] Song Wu, Qian Liu, Ji-Min Zhu, Ming-Rui Wang, Jing Li, and Mei-Guo Sun. Association between the IL7R T244I polymorphism and multiple sclerosis risk : a meta analysis. *Neurological Sciences*, pages 1–8, 2016.

[83] Pouya Khankhanian, Pierre-Antoine Gourraud, Antoine Lizee, and Douglas S Goodin. Haplotype-based approach to known MS-associated regions increases the amount of explained risk. *Journal of medical genetics*, pages jmedgenet–2015, 2015.

[84] Georgia G Braliou, Katerina G Pantavou, Panagiota I Kontou, and Pantelis G Bagos. Polymorphisms of the CD24 gene are associated with risk of multiple sclerosis : A meta-analysis. *International Journal of Molecular Sciences*, 16(6) :12368–12381, 2015.

[85] Christina M Lill, Felix Luessi, Antonio Alcina, Ekaterina A Sokolova, Nerea Ugidos, Belén de la Hera, Léna Guillot-Noël, Sunny Malhotra, Eva Reinthaler, Brit-Maren M Schjeide, et al. Genome-wide significant association with seven novel multiple sclerosis risk loci. *Journal of Medical Genetics*, 52(12) :848–855, 2015.

[86] I Alloza, D Otaegui, A Lopez de Lapuente, A Antigüedad, J Varadé, C Núñez, R Arroyo, E Urcelay, O Fernandez, L Leyva, et al. ANKRD55 and DHCR7 are novel multiple sclerosis risk loci. *Genes and Immunity*, 13(3) :253–257, 2012.

[87] Christina M Lill, Brit-Maren M Schjeide, Christiane Graetz, Tian Liu, Vincent Damotte, Denis A Akkad, Paul Blaschke, Lisa-Ann Gerdes, Antje Kroner, Felix Luessi, et al. Genome-wide significant association of ANKRD55 rs6859219 and multiple sclerosis risk. *Journal of Medical Genetics*, 50(3) :140–143, 2013.

[88] International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium 2, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359) :214–219, 2011.

[89] Zhaoqiang Zhang, Lei Wang, Xiao Sun, Li Zhang, and Lianyuan Lu. Association of IL4 and IL4R polymorphisms with multiple sclerosis susceptibility in Caucasian population : A meta-analysis. *Journal of the Neurological Sciences*, 363 :107–113, 2016.

[90] Gürdal Orhan, Esra Eruyar, Semra Öztürk Mungan, Fikri Ak, and Bensu Karahalil. The association of IL-18 gene promoter polymorphisms and the levels of serum IL-18 on the risk of multiple sclerosis. *Clinical Neurology and Neurosurgery*, 146 :96–101, 2016.

[91] Zhe Wang, A Dessa Sadovnick, Anthony L Traboulsee, Jay P Ross, Cecily Q Bernales, Mary Encarnacion, Irene M Yee, Madonna de Lemos, Talitha Greenwood, Joshua D Lee, et al. Nuclear receptor NR1H3 in familial multiple sclerosis. *Neuron*, 90(5) :948–954, 2016.

[92] Rasoul Abdollahzadeh, Mahsa Sobhani Fard, Farideh Rahmani, Kaveh Moloudi, Asaad Azarnezhad, et al. Predisposing role of vitamin D receptor (VDR) polymorphisms in the development of multiple sclerosis : A case-control study. *Journal of the Neurological Sciences*, 2016.

[93] Joseph Louis Lagrange. *Mécanique analytique*. Mallet-Bachelier, 1853.

[94] FN Gumedze and TT Dunne. Parameter estimation and inference in the linear mixed model. *Linear Algebra and its Applications*, 435(8) :1920–1944, 2011.

[95] S. J. Welham and R. Thompson. Likelihood ratio tests for fixed model terms using residual maximum likelihood. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(3) :701–714, 1997.

[96] Robert B Davies. Algorithm AS 155 : The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society Series C Applied Statistics*, 29(3) :323–333, 1980.

[97] Daniel O Stram and Jae Won Lee. Variance components testing in the longitudinal mixed effects model. *Biometrics*, pages 1171–1177, 1994.

[98] Christopher H Morrell. Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics*, pages 1560–1568, 1998.

[99] D Zwillinger. Standard mathematical tables and formulae, 1996.

[100] Norman E Breslow and David G Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421) :9–25, 1993.

[101] Han Chen, Chaolong Wang, Matthew P Conomos, Adrienne M Stilp, Zilin Li, Tamar Sofer, Adam A Szpiro, Wei Chen, John M Brehm, Juan C Celedón, et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, 2016.

[102] Xihong Lin and Norman E Breslow. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91(435) :1007–1016, 1996.

[103] Xihong Lin. Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2) :309–326, 1997.

[104] Charles R Henderson. Estimation of variance and covariance components. *Biometrics*, 9(2) :226–252, 1953.

[105] Charles R Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pages 423–447, 1975.

[106] Christopher I Amos. Robust variance-components approach for assessing genetic linkage in pedigrees. *American Journal of Human Genetics*, 54(3) :535, 1994.

[107] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3) :1709–1723, 2008.

[108] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4) :348–354, 2010.

[109] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7) :821–824, 2012.

[110] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10) :833–835, 2011.

[111] Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics*, 46(2) :100–106, 2014.

[112] Yurii S Aulchenko, Dirk-Jan De Koning, and Chris Haley. Genomewide rapid association using mixed model and regression : a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, 177(1) :577–585, 2007.

[113] Gulnara R Svishcheva, Tatiana I Axenovich, Nadezhda M Belonogova, Cornelia M van Duijn, and Yurii S Aulchenko. Rapid variance components-based method for whole-genome association analysis. *Nature Genetics*, 44(10) :1166–1170, 2012.

[114] Dawei Liu, Debashis Ghosh, and Xihong Lin. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, 9(1) :292, 2008.

[115] Wan-Yu Lin, Nengjun Yi, Xiang-Yang Lou, Degui Zhi, Kui Zhang, Guimin Gao, Hemant K Tiwari, and Nianjun Liu. Haplotype kernel association test as a powerful method to identify chromosomal regions harboring uncommon causal variants. *Genetic epidemiology*, 37(6) :560–570, 2013.

[116] Iuliana Ionita-Laza, Seunggeun Lee, Vlad Makarov, Joseph D Buxbaum, and Xihong Lin. Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics*, 92(6) :841–853, 2013.

[117] Karim Oualkacha, Zari Dastani, Rui Li, Pablo E Cingolani, Timothy D Spector, Christopher J Hammond, J Brent Richards, Antonio Ciampi, and Celia MT Greenwood. Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genetic epidemiology*, 37(4) :366–376, 2013.

[118] Han Chen, James B Meigs, and Josée Dupuis. Sequence kernel association test for quantitative traits in family samples. *Genetic epidemiology*, 37(2) :196–204, 2013.

[119] Jian Yang, Teri A Manolio, Louis R Pasquale, Eric Boerwinkle, Neil Caporaso, Julie M Cunningham, Mariza de Andrade, Bjarke Feenstra, Eleanor Feingold, M Geoffrey Hayes, et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics*, 43(6) :519–525, 2011.

[120] Jian Yang, Andrew Bakshi, Zhihong Zhu, Gibran Hemani, Anna AE Vinkhuyzen, Sang Hong Lee, Matthew R Robinson, John RB Perry, Ilja M Nolte, Jana V van Vliet-Ostaptchouk, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics*, 47(10) :1114–1120, 2015.

[121] Doug Speed, Gibran Hemani, Michael R Johnson, and David J Balding. Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics*, 91(6) :1011–1021, 2012.

[122] S Hong Lee, Jian Yang, Guo-Bo Chen, Stephan Ripke, Eli A Stahl, Christina M Hultman, Pamela Sklar, Peter M Visscher, Patrick F Sullivan, Michael E Goddard, et al. Estimation of SNP heritability from dense genotype data. *American Journal of Human Genetics*, 93(6) :1151–1155, 2013.

[123] S Hong Lee, Teresa R DeCandia, Stephan Ripke, Jian Yang, Patrick F Sullivan, Michael E Goddard, Matthew C Keller, Peter M Visscher, Naomi R Wray, Schizophrenia Psychiatric Genome-Wide Association Study Consortium, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature genetics*, 44(3) :247–250, 2012.

[124] Seunggeun Lee, Larisa Miropolsky, and Michael Wu. Package 'SKAT', 2015.

[125] Stephen F Schaffner, Catherine Foo, Stacey Gabriel, David Reich, Mark J Daly, and David Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res*, 15(11) :1576–1583, 2005.

[126] Jared O'Connell, Deepti Gurdasani, Olivier Delaneau, Nicola Pirastu, Sheila Ulivi, Massimiliano Cocca, Michela Traglia, Jie Huang, Jennifer E Huffman, Igor Rudan, et al. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet*, 10(4) :e1004234, 2014.

[127] Doug Speed and David J Balding. MultiBLUP : improved SNP-based prediction for complex traits. *Genome research*, 24(9) :1550–1557, 2014.

[128] Naomi R Wray, Jian Yang, Ben J Hayes, Alkes L Price, Michael E Goddard, and Peter M Visscher. Pitfalls of predicting complex traits from snps. *Nature Reviews Genetics*, 14(7) :507–515, 2013.

[129] Hervé Perdry and Claire Dandine-Roulland. Package 'gaston' [version 1.45], 2015.

[130] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, David Ferreira, Manuel AR an Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. PLINK : a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3) :559–575, 2007.

[131] Janis E Wigginton, David J Cutler, and Gonçalo R Abecasis. A note on exact tests of Hardy-Weinberg equilibrium. *The American Journal of Human Genetics*, 76(5) :887–893, 2005.

[132] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. GCTA : a tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, 88(1) :76–82, 2011.

[133] Xiuwen Zheng, Stephanie Gogarten, Cathy Laurie, and Bruce Weir. Parallel computing toolset for relatedness and principal component analysis of SNP data, 2014.

[134] Arthur R Gilmour, Robin Thompson, and Brian R Cullis. Average information REML : an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, pages 1440–1450, 1995.

[135] 3C Study Group et al. Vascular factors and risk of dementia : design of the Three-City Study and baseline characteristics of the study population. *Neuroepidemiology*, 22(6) :316, 2003.

[136] Claire Dandine-Roulland, Céline Bellenguez, Stéphanie Debette, Philippe Amouyel, Emmanuelle Génin, and Hervé Perdry. Accuracy of heritability estimations in presence of hidden population stratification. *Scientific Reports*, 6, 2016.

[137] Alkes L Price, Michael E Weale, Nick Patterson, Simon R Myers, Anna C Need, Kevin V Shianna, Dongliang Ge, Jerome I Rotter, Esther Torres, Kent D Taylor, et al. Long-range LD can confound genome scans in admixed populations. *American Journal of Human Genetics*, 83(1) :132, 2008.

[138] Carl A Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, Andrew P Morris, and Krina T Zondervan. Data quality control in genetic case-control association studies. *Nature Protocols*, 5(9) :1564–1573, 2010.

[139] Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15 :246–263, 1886.

[140] Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I Berndt, Michael N Weedon, Fernando Rivadeneira, Cristen J Willer, Anne U Jackson, Sailaja Vedantam, Soumya Raychaudhuri, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317) :832–838, 2010.

[141] Matthew B Lanktree, Yiran Guo, Muhammed Murtaza, Joseph T Glessner, Swneke D Bailey, N Charlotte Onland-Moret, Guillaume Lettre, Halit Ongen, Ramakrishnan Rajagopalan, Toby Johnson, et al. Meta-analysis of dense genecentric association studies reveals common and uncommon variants associated with height. *American Journal of Human Genetics*, 88(1) :6–18, 2011.

[142] Ralf JP van der Valk, Eskil Kreiner-Møller, Marjolein N Kooijman, Mònica Guxens, Evangelia Stergiakouli, Annika Sääf, Jonathan P Bradfield, Frank Geller, M Geoffrey

Hayes, Diana L Cousminer, et al. A novel common variant in DCST2 is associated with length in early life and height in adulthood. *Human Molecular Genetics*, 24(4) :1155–1168, 2015.

[143] Elizabeth K Speliotes, Cristen J Willer, Sonja I Berndt, Keri L Monda, Gudmar Thorleifsson, Anne U Jackson, Hana Lango Allen, Cecilia M Lindgren, Jian'an Luan, Reedik Mägi, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, 42(11) :937–948, 2010.

[144] Gudmar Thorleifsson, G Bragi Walters, Daniel F Gudbjartsson, Valgerdur Steinthorsdottir, Patrick Sulem, Anna Helgadottir, Unnur Styrkarsdottir, Solveig Gretarsdottir, Steinunn Thorlacius, Ingileif Jonsdottir, et al. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nature Genetics*, 41(1) :18–24, 2009.

[145] Ruth JF Loos. Genetic determinants of common obesity and their value in prediction. *Best Practice & Research Clinical Endocrinology & Metabolism*, 26(2) :211–226, 2012.

[146] Iris M Heid, Anne U Jackson, Joshua C Randall, Thomas W Winkler, Lu Qi, Valgerdur Steinthorsdottir, Gudmar Thorleifsson, M Carola Zillikens, Elizabeth K Speliotes, Reedik Mägi, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nature Genetics*, 42(11) :949–960, 2010.

[147] Cecilia M Lindgren, Iris M Heid, Joshua C Randall, Claudia Lamina, Valgerdur Steinthorsdottir, Lu Qi, Elizabeth K Speliotes, Gudmar Thorleifsson, Cristen J Willer, Blanca M Herrera, et al. Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLoS Genetics*, 5(6) :e1000508, 2009.

[148] Sachiko Yoneyama, Yiran Guo, Matthew B Lanktree, Michael R Barnes, Clara C Elbers, Konrad J Karczewski, Sandosh Padmanabhan, Florianne Bauer, Jens Baumert, Amber Beitelshees, et al. Gene-centric meta-analyses for central adiposity traits in up to 57,412 individuals of european descent confirm known loci and reveal several novel associations. *Human Molecular Genetics*, page ddt626, 2013.

[149] Tinca JC Polderman, Beben Benyamin, Christiaan A de Leeuw, Patrick F Sullivan, Arjen van Bochoven, Peter M Visscher, and Danielle Posthuma. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 2015.

[150] Peter M Visscher, Stuart Macgregor, Beben Benyamin, Gu Zhu, Scott Gordon, Sarah Medland, William G Hill, Jouke-Jan Hottenga, Gonneke Willemsen, Dorret I Boomsma, et al. Genome partitioning of genetic variation for height from 11,214 sibling pairs. *American Journal of Human Genetics*, 81(5) :1104–1110, 2007.

[151] Stuart Macgregor, Belinda K Cornes, Nicholas G Martin, and Peter M Visscher. Bias, precision and heritability of self-reported and clinically measured height in Australian twins. *Human Genetics*, 120(4) :571–580, 2006.

[152] Karri Silventoinen, Sampo Sammalisto, Markus Perola, Dorret I Boomsma, Belinda K Cornes, Chayna Davis, Leo Dunkel, Marlies De Lange, Jennifer R Harris, Jacob VB Hjelmborg, et al. Heritability of adult body height : a comparative study of twin cohorts in eight countries. *Twin Research*, 6(05) :399–408, 2003.

[153] Xu Chen, Ralf Kuja-Halkola, Iffat Rahman, Johannes Arpegård, Alexander Viktorin, Robert Karlsson, Sara Hägg, Per Svensson, Nancy L Pedersen, and Patrik KE Magnusson. Dominant genetic variation and missing heritability for human complex traits : Insights from twin versus genome-wide common SNP models. *American Journal of Human Genetics*, 97(5) :708–714, 2015.

[154] Peter M Visscher, Jian Yang, and Michael E Goddard. A commentary on "common SNPs explain a large proportion of the heritability for human height" by Yang et al.(2010). *Twin Research and Human Genetics*, 13(06) :517–524, 2010.

[155] Jian Yang, Andrew Bakshi, Zhihong Zhu, Gibran Hemani, Anna AE Vinkhuyzen, Ilja M Nolte, Jana V van Vliet-Ostaptchouk, Harold Snieder, Tonu Esko, Lili Milani, et al. Genome-wide genetic homogeneity between sexes and populations for human height and body mass index. *Human Molecular Genetics*, 24(25) :7445–7449, 2015.

[156] Cathy E Elks, Marcel Den Hoed, Jing Hua Zhao, Stephen J Sharp, Nicholas J Wareham, Ruth JF Loos, and Ken K Ong. Variability in the heritability of body mass index : a systematic review and meta-regression. *Frontiers in Endocrinology*, 3, 2012.

[157] Dirk JA Smit, Michelle Luciano, Meike Bartels, Catharine EM Van Beijsterveldt, Margaret J Wright, Narelle K Hansell, Han G Brunner, G Frederiek Estourgie-van Burk, Eco JC De Geus, Nicholas G Martin, et al. Heritability of head size in Dutch and Australian twin families at ages 0–50 years. *Twin Research and Human Genetics*, 13(04) :370–380, 2010.

[158] Sergey Ermakov, Eugene Kobyliansky, and Gregory Livshits. Quantitative genetic study of head size related phenotypes in ethnically homogeneous Chuvasha pedigrees. *Annals of Human Biology*, 32(5) :585–598, 2005.

[159] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within Europe. *Nature*, 456(7218) :98–101, 2008.

[160] Hervé Perdry, Bertram Müller-Myhsok, and Françoise Clerget-Darpoux. Using affected sib-pairs to uncover rare disease variants. *Human Heredity*, 74(3-4) :129–141, 2012.

[161] Claire Dandine-Roulland and Hervé Perdry. Where is the causal variant ? on the advantage of the family design over the case–control design in genetic association studies. *European Journal of Human Genetics*, 23(10) :1357–1363, 2015.

[162] Claire Dandine-Roulland. Package 'ASPBay' [version 1.2], 2015.

[163] P Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3) :375–386, 1955.

[164] Christian P Robert and George Casella. *Monte Carlo statistical methods*, volume 319. Springer Verlag, 2004.

[165] Marie-Claude Babron, Hervé Perdry, Adam E Handel, Sreeram V Ramagopalan, Vincent Damotte, Bertrand Fontaine, Bertram Müller-Myhsok, George C Ebers, and Françoise Clerget-Darpoux. Determination of the real effect of genes identified in GWAS : the example of IL2RA in multiple sclerosis. *American Journal of Human Genetics*, 20(3) :321–325, 2011.

[166] Fuencisla Matesanz, Alfredo Caro-Maldonado, Maria Fedetz, Oscar Fernández, Roger L Milne, Miguel Guerrero, Concepción Delgado, and Antonio Alcina. IL2RA/CD2 polymorphisms contribute to multiple sclerosis susceptibility. *Journal of Neurology*, 254(5) :682–684, 2007.

[167] David A Hafler, A Compston, S Sawcer, ES Lander, MJ Daly, PL De Jager, PI De Bakker, SB Gabriel, DB Mirel, AJ Ivinson, et al. Risk alleles for multiple sclerosis identified by a genomewide study. *New England Journal of Medecine*, 357(9) :851–862, 2007.

[168] Sreeram V Ramagopalan, Carl Anderson, A Dessa Sadovnick, George C Ebers, F Matesanz, et al. Genomewide study of multiple sclerosis. *New England Journal of Medecine*, 357(21) :2199–2200, 2007.

[169] JP Rubio, J Stankovich, J Field, N Tubridy, M Marriott, C Chapman, M Bahlo, D Perera, LJ Johnson, BD Tait, et al. Replication of KIAA0350, IL2RA, RPL5 and CD58 as multiple sclerosis susceptibility genes in Australians. *Genes & Immunity*, 9(7) :624–630, 2008.

[170] F Weber, B Fontaine, I Cournu-Rebeix, A Kroner, M Knop, S Lutz, F Müller-Sarnowski, M Uhr, T Bettecken, M Kohli, et al. IL2RA and IL7RA genes confer susceptibility for multiple sclerosis in two independent European populations. *Genes & Immunity*, 9(3) :259–263, 2008.

[171] Alexander I Tröster. Refining genetic associations in multiple sclerosis. *Neurology*, 66 :1830–36, 2006.

[172] Antonio Alcina, María Fedetz, Dorothy Ndagire, Oscar Fernández, Laura Leyva, Miguel Guerrero, María M Abad-Grau, Carmen Arnal, Concepción Delgado, Miguel Lucas, et al. IL2RA/CD25 gene polymorphisms : uneven association with multiple sclerosis (MS) and type 1 diabetes (T1D). *PLoS One*, 4(1) :4137, 2009.

[173] Calliope A Dendrou, Vincent Plagnol, Erik Fung, Jennie HM Yang, Kate Downes, Jason D Cooper, Sarah Nutland, Gillian Coleman, Matthew Himsworth, Matthew Hardy, et al. Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource. *Nature Genetics*, 41(9) :1011–1015, 2009.

[174] Lisa M Maier, David E Anderson, Christopher A Severson, Clare Baecher-Allan, Brian Healy, David V Liu, K Dane Wittrup, Philip L De Jager, and David A Hafler. Soluble IL2RA levels in multiple sclerosis subjects and the effect of soluble IL-2RA on immune responses. *Journal of Immunology*, 182(3) :1541–1547, 2009.

[175] Lisa M Maier, Christopher E Lowe, Jason Cooper, Kate Downes, David E Anderson, Christopher Severson, Pamela M Clark, Brian Healy, Neil Walker, Cristin Aubin, et al. IL2RA genetic heterogeneity in multiple sclerosis and type 1 diabetes susceptibility and soluble interleukin-2 receptor production. *PLoS Genetics*, 5(1) :e1000322, 2009.

[176] Charles M Poser, Donald W Paty, Labe Scheinberg, W Ian McDonald, Floyd A Davis, George C Ebers, Kenneth P Johnson, William A Sibley, Donald H Silberberg, and Wallace W Tourtellotte. New diagnostic criteria for multiple sclerosis : guidelines for research protocols. *Annals of Neurology*, 13(3) :227–231, 1983.

[177] Gonçalo R Abecasis, Stacey S Cherny, William O Cookson, and Lon R Cardon. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30(1) :97–101, 2001.

[178] JL Binet, A Auquier, G Dighiero, Cl Chastang, H Piguet, J Goasguen, G Vaugier, G Potron, P Colona, F Oberling, et al. A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis. *Cancer*, 48(1) :198–206, 1981.

[179] Mathieu Bourgey, Hervé Perdry, and Françoise Clerget-Darpoux. Modeling the effect of PTPN22 in rheumatoid arthritis. In *BMC Proceedings*, volume 1, page S37. BioMed Central Ltd, 2007.

[180] F Clerget-Darpoux, MC Babron, B Prum, GM Lathrop, I Deschamps, and J Hors. A new method to test genetic models in HLA associated diseases : the MASC method. *Annals of Human Genetics*, 52(3) :247–258, 1988.

[181] Anne-Louise Leutenegger, Bernard Prum, Emmanuelle Génin, Christophe Verny, Arnaud Lemainque, Françoise Clerget-Darpoux, and Elizabeth A Thompson. Estimation of the inbreeding coefficient through use of genomic data. *The American Journal of Human Genetics*, 73(3) :516–523, 2003.

[182] Hollie Schmidt, Dhelia Williamson, and Allison Ashley-Koch. HLA-DR15 haplotype and multiple sclerosis : a HuGE review. *American journal of epidemiology*, 165(10) :1097–1109, 2007.

[183] RHSR Roxburgh, SR Seaman, T Masterman, AE Hensiek, SJ Sawcer, S Vukusic, I Achiti, C Confavreux, M Coustans, E Le Page, et al. Multiple sclerosis severity score using disability and disease duration to rate disease severity. *Neurology*, 64(7) :1144–1151, 2005.

[184] Farren BS Briggs, Xiaorong Shao, Benjamin A Goldstein, Jorge R Oksenberg, Lisa F Barcellos, and Philip L De Jager. Genome-wide association study of severity in multiple sclerosis. *Genes and immunity*, 12(8) :615–625, 2011.

[185] Sergio E Baranzini, Joanne Wang, Rachel A Gibson, Nicholas Galwey, Yvonne Naegelin, Frederik Barkhof, Ernst-Wilhelm Radue, Raija LP Lindberg, Bernard MG Uitdehaag, Michael R Johnson, et al. Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Human molecular genetics*, 18(4) :767–778, 2009.

[186] Iuliana Ionita-Laza, Seunggeun Lee, Vladimir Makarov, Joseph D Buxbaum, and Xihong Lin. Family-based association tests for sequence data, and comparisons with population-based association tests. *European Journal of Human Genetics*, 21(10) :1158–1162, 2013.

[187] Daniel Rabinowitz and Nan Laird. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Human heredity*, 50(4) :211–223, 2000.

[188] Steve Horvath, Xin Xu, and Nan M Laird. The family based association test method : strategies for studying general genotype–phenotype associations. *European Journal of Human Genetics*, 9(4), 2001.

[189] Christoph Lange, Edwin K Silverman, Xin Xu, Scott T Weiss, and Nan M Laird. A multivariate family-based association test using generalized estimating equations : FBAT-GEE. *Biostatistics*, 4(2) :195–206, 2003.

[190] Jack Kiefer. Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society*, 4(3) :502–506, 1953.

[191] Alastair Compston, Christian Confavreux, Hans Lassmann, Ian McDonald, David Miller, J Noseworthy, Kenneth Smith, and Hartmut Wekerle. *McAlpine's multiple sclerosis*. Churchill Livingstone Elsevier, 2005.

[192] An Goris, Ine Pauwels, and Bénédicte Dubois. Progress in multiple sclerosis genetics. *Current Genomics*, 13(8) :646, 2012.

[193] Pierre-Antoine Gourraud, Hanne F Harbo, Stephen L Hauser, and Sergio E Baranzini. The genetics of multiple sclerosis : an up-to-date review. *Immunological Reviews*, 248(1) :87–103, 2012.

[194] Noriko Isobe, Lohith Madireddy, Pouya Khankhanian, Takuya Matsushita, Stacy J Caillier, Jayaji M Moré, Pierre-Antoine Gourraud, Jacob L McCauley, Ashley H Beecham, Laura Piccio, et al. An ImmunoChip study of multiple sclerosis risk in African Americans. *Brain*, 138(6) :1518–1530, 2015.

[195] Jonathan K Pritchard and Molly Przeworski. Linkage disequilibrium in Humans : models and data. *American Journal of Human Genetics*, 69(1) :1–14, 2001.

[196] Neil Risch, Kathleen Merikangas, et al. The future of genetic studies of complex human diseases. *Science*, 273(5281) :1516–1517, 1996.

[197] Yaodong Hu, Guilherme JM Rosa, and Daniel Gianola. Incorporating parent-of-origin effects in whole-genome prediction of complex traits. *Genetics Selection Evolution*, 48(1) :1, 2016.

[198] Ji-Yuan Zhou, Jie Ding, Wing K Fung, and Shili Lin. Detection of parent-of-origin effects using general pedigree data. *Genetic epidemiology*, 34(2) :151–158, 2010.

[199] Sang Hong Lee and Julius HJ Van Der Werf. An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. *Genetics Selection Evolution*, 38(1) :1–19, 2006.

[200] Zhiwu Zhang, Elhan Ersoz, Chao-Qiang Lai, Rory J Todhunter, Hemant K Tiwari, Michael A Gore, Peter J Bradbury, Jianming Yu, Donna K Arnett, Jose M Ordovas, et al. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42(4) :355–360, 2010.

[201] Alexander Gusev, S Hong Lee, Benjamin M Neale, Gosia Trynka, Bjarni J Vilhjalmsson, Hilary Finucane, Han Xu, Chongzhi Zang, Stephan Ripke, Eli Stahl, et al. Regulatory variants explain much more heritability than coding variants across 11 common diseases. *bioRxiv*, page 004309, 2014.

[202] Anna AE Vinkhuyzen, Naomi R Wray, Jian Yang, Michael E Goddard, and Peter M Visscher. Estimation and partitioning of heritability in human populations using whole genome analysis methods. *Annual Review of Genetics*, 47 :75, 2013.

[203] Siddharth Krishna Kumar, Marcus W Feldman, David H Rehkopf, and Shripad Tuljapurkar. Limitations of GCTA as a solution to the missing heritability problem. *Proc. Natl. Acad. Sci.*, 113(1) :E61–E70, 2016.

[204] Jian Yang, Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Commentary on "Limitations of GCTA as a solution to the missing heritability problem". *bioRxiv*, page 036574, 2016.

[205] Siddharth Krishna Kumar, Marcus W Feldman, David H Rehkopf, and Shripad Tuljapurkar. Response to commentary on" limitations of gcta as a solution to the missing heritability problem". *bioRxiv*, page 039594, 2016.

[206] Emmanuelle Génin and Françoise Clerget-Darpoux. The missing heritability paradigm : A dramatic resurgence of the GIGO Syndrome in genetics. *Human Heredity*, 79(1) :10–13, 2015.

[207] Jean-Charles Lambert, Simon Heath, Gael Even, Dominique Campion, Kristel Sleegers, Mikko Hiltunen, Onofre Combarros, Diana Zelenika, Maria J Bullido, Beatrice Tavernier, Luc Letenneur, Karolien Bettens, Claudine Berr, et al. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nature Genetics*, 41(10) :1094–1099, 2009.

[208] Michael N Weedon, Hana Lango, Cecilia M Lindgren, Chris Wallace, David M Evans, Massimo Mangino, Rachel M Freathy, John RB Perry, Suzanne Stevens, Alistair S Hall, et al. Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genetics*, 40(5) :575–583, 2008.

[209] Daniel F Gudbjartsson, G Bragi Walters, Gudmar Thorleifsson, Hreinn Stefansson, Bjarni V Halldorsson, Pasha Zusmanovich, Patrick Sulem, Steinunn Thorlacius, Arnaldur Gylfason, Stacy Steinberg, et al. Many sequence variants affecting diversity of adult human height. *Nature Genetics*, 40(5) :609–615, 2008.

[210] Guillaume Lettre, Anne U Jackson, Christian Gieger, Fredrick R Schumacher, Sonja I Berndt, Serena Sanna, Susana Eyheramendy, Benjamin F Voight, Johannah L Butler, Candace Guiducci, et al. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nature Genetics*, 40(5) :584–591, 2008.

[211] Peter M Visscher. Sizing up human height variation. *Nature Genetics*, 40(5) :489–490, 2008.

[212] Timothy M Frayling, Nicholas J Timpson, Michael N Weedon, Eleftheria Zeggini, Rachel M Freathy, Cecilia M Lindgren, John RB Perry, Katherine S Elliott, Hana Lango, Nigel W Rayner, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316(5826) :889–894, 2007.

[213] Cristen J Willer, Elizabeth K Speliotes, Ruth JF Loos, Shengxu Li, Cecilia M Lindgren, Iris M Heid, Sonja I Berndt, Amanda L Elliott, Anne U Jackson, Claudia Lamina, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genetics*, 41(1) :25–34, 2009.

[214] Angelo Scuteri, Serena Sanna, Wei-Min Chen, Manuela Uda, Giuseppe Albai, James Strait, Samer Najjar, Ramaiah Nagaraja, Marco Orrú, Gianluca Usala, et al. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genetics*, 3(7) :e115, 2007.

[215] Ruth JF Loos, Cecilia M Lindgren, Shengxu Li, Eleanor Wheeler, Jing Hua Zhao, Inga Prokopenko, Michael Inouye, Rachel M Freathy, Antony P Attwood, Jacques S Beckmann, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature Genetics*, 40(6) :768–775, 2008.

Annexes

Annexe 1 : Suppléments de méthodologie pour les modèles mixtes

Les dérivées de la log-vraisemblance

Nous considérons le modèle :

$$Y = X\beta + \omega_1 + \cdots + \omega_k + e$$

avec

$$\omega_1 \sim \mathcal{N}(0, \tau_1 K_1),\ \ldots,\ \omega_k \sim \mathcal{N}(0, \tau_k K_k),\ e \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n).$$

Sous ce modèle, la log-vraisemblance est :

$$\ell(\beta, \tau_1, \ldots, \tau_k, \sigma^2) = -\frac{1}{2}\log|V| - \frac{1}{2}(Y - X\beta)'V^{-1}(Y - X\beta).$$

avec $V = \tau_1 K_1 + \cdots + \tau_k K_k + \sigma^2 \mathbb{I}_n$.

Les dérivées premières de la log-vraisemblance sont :

- ▷ Pour les effets fixes β :

$$\frac{\partial \ell}{\partial \beta} = -\frac{1}{2}(-X')(V^{-1} + V'^{-1})(Y - X\beta) = X'V^{-1}(Y - X\beta).$$

- ▷ Pour les composantes de la variance τ_i, nous avons :

$$\begin{aligned}
&\frac{\partial}{\partial \tau_i} V = K_i \\
&\frac{\partial}{\partial \tau_i} \log|V| = \mathrm{Tr}(V^{-1}K_i) \\
&\frac{\partial}{\partial \tau_i}(Y - X\beta)'V^{-1}(Y - X\beta) = -(Y - X\beta)'V^{-1}K_iV^{-1}(Y - X\beta).
\end{aligned}$$

  Donc,

$$\frac{\partial \ell}{\partial \tau_i} = -\frac{1}{2}\mathrm{Tr}(V^{-1}K_i) + \frac{1}{2}(Y - X\beta)'V^{-1}K_iV^{-1}(Y - X\beta).$$

- ▷ Pour la variance résiduelle σ^2 :

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{1}{2}\mathrm{Tr}(V^{-1}) + \frac{1}{2}(Y - X\beta)'V^{-1}V^{-1}(Y - X\beta).$$

Puis, les dérivées secondes pour les composantes de la variance sont :

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \tau_i \partial \tau_j} &= \frac{1}{2}\mathrm{Tr}(V^{-1}K_jV^{-1}K_i)\\ &\quad - \frac{1}{2}(Y-X\beta)'V^{-1}K_jV^{-1}K_iV^{-1}(Y-X\beta)\\ &\quad - \frac{1}{2}(Y-X\beta)'V^{-1}K_iV^{-1}K_jV^{-1}(Y-X\beta)\\ &= \frac{1}{2}\mathrm{Tr}(V^{-1}K_jV^{-1}K_i) - (Y-X\beta)'V^{-1}K_iV^{-1}K_jV^{-1}(Y-X\beta).\end{aligned}$$

De la même façon, nous avons :

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \tau_i \partial \sigma^2} &= \frac{1}{2}\mathrm{Tr}(V^{-2}K_i) - (Y-X\beta)'V^{-1}K_iV^{-2}(Y-X\beta),\\ \frac{\partial^2 \ell}{\partial \sigma^2 \partial \tau_j} &= \frac{1}{2}\mathrm{Tr}(V^{-1}K_jV^{-1}) - (Y-X\beta)'V^{-2}K_jV^{-1}(Y-X\beta),\\ \frac{\partial^2 \ell}{\partial \tau_i \partial \sigma^2} &= \frac{1}{2}\mathrm{Tr}(V^{-2}) - (Y-X\beta)'V^{-3}(Y-X\beta).\end{aligned}$$

Les dérivées de la log-vraisemblance restreinte

La vraisemblance restreinte pour le même modèle que précédemment s'écrit :

$$\ell^{\mathrm{re}}(\tau_1,\ldots,\tau_k,\sigma^2) = -\frac{1}{2}\log|V| - \frac{1}{2}\log\left|X'V^{-1}X\right| - \frac{1}{2}Y'PY + \text{constante}$$

avec $P = V^{-1} - V^{-1}X\left(X'V^{-1}X\right)^{-1}X'V^{-1}$.

Ainsi, les dérivées premières de cette vraisemblance sont :

$$\begin{aligned}\frac{\partial \ell^{\mathrm{re}}}{\partial \tau_i} &= -\frac{1}{2}\mathrm{Tr}(PK_i) + \frac{1}{2}Y'PK_iPY\\ \frac{\partial \ell^{\mathrm{re}}}{\partial \sigma^2} &= -\frac{1}{2}\mathrm{Tr}(P) + \frac{1}{2}Y'PPY.\end{aligned}$$

Puis, les dérivées secondes s'écrivent :

$$\begin{aligned}\frac{\partial^2 \ell^{\mathrm{re}}}{\partial \tau_i \partial \tau_j} &= \frac{1}{2}\mathrm{Tr}(PK_iPK_j) - Y'PK_iPK_jPY,\\ \frac{\partial^2 \ell^{\mathrm{re}}}{\partial \tau_i \partial \sigma^2} &= \frac{1}{2}\mathrm{Tr}(PK_iP) - Y'PK_iPPY,\\ \frac{\partial^2 \ell^{\mathrm{re}}}{\partial \sigma^2 \partial \tau_j} &= \frac{1}{2}\mathrm{Tr}(PPK_j) - Y'PPK_jPY,\\ \frac{\partial^2 \ell^{\mathrm{re}}}{\partial \sigma^2 \partial \tau_j} &= \frac{1}{2}\mathrm{Tr}(PP) - Y'PPPY.\end{aligned}$$

Les espérances des dérivées secondes de la vraisemblance restreinte sont :

$$\begin{aligned}
E\left(\frac{\partial^2 \ell^{\mathrm{re}}}{\partial\tau_i\partial\tau_j}\right) &= \frac{1}{2}\mathrm{Tr}(PK_jPK_i) - E\left(Y'PK_iPK_jPY\right)\\
&= \frac{1}{2}\mathrm{Tr}(PK_jPK_i) - \mathrm{Tr}(PK_iPK_jPV)\\
&= \frac{1}{2}\mathrm{Tr}(PK_jPK_i) - \mathrm{Tr}(PK_jPVPK_i)\\
&= -\frac{1}{2}\mathrm{Tr}(PK_jPK_i)
\end{aligned}$$

comme $PVP = P$ et $PX = \mathbb{O}_{n\times p}$ avec $\mathbb{O}_{n\times p}$ la matrice de 0 de taille $(n \times p)$.

De la même façon, nous avons :

$$\begin{aligned}
E\left(\frac{\partial^2 \ell^{\mathrm{re}}}{\partial\tau_i\partial\sigma^2}\right) &= -\frac{1}{2}\mathrm{Tr}(PPK_i),\\
E\left(\frac{\partial^2 \ell^{\mathrm{re}}}{\partial\sigma^2\partial\tau_j}\right) &= -\frac{1}{2}\mathrm{Tr}(PK_jP),\\
E\left(\frac{\partial^2 \ell^{\mathrm{re}}}{\partial\sigma^2\partial\sigma^2}\right) &= -\frac{1}{2}\mathrm{Tr}(PP).
\end{aligned}$$

Ainsi, la matrice de l'information observée est l'inverse de la matrice hessienne de la vraisemblance restreinte pour les paramètres $(\tau_1, \ldots, \tau_k, \sigma^2)$:

$$\begin{bmatrix}
-\frac{1}{2}\mathrm{Tr}(PK_1PK_1) + Y'PK_1PK_1PY & \cdots & -\frac{1}{2}\mathrm{Tr}(PK_1PK_k) + Y'PK_1PK_kPY & -\frac{1}{2}\mathrm{Tr}(PK_1P) + Y'PK_1PPY\\
\vdots & \vdots & \ddots & \vdots\\
-\frac{1}{2}\mathrm{Tr}(PK_kPK_1) + Y'PK_kPK_1PY & \cdots & -\frac{1}{2}\mathrm{Tr}(PK_kPK_k) + Y'PK_kPK_kPY & -\frac{1}{2}\mathrm{Tr}(PK_kP) + Y'PK_kPPY\\
-\frac{1}{2}\mathrm{Tr}(PPK_1) + Y'PPK_1PY & \cdots & -\frac{1}{2}\mathrm{Tr}(PPK_k) + Y'PPK_kPY & -\frac{1}{2}\mathrm{Tr}(PP) + Y'PPPY
\end{bmatrix}.$$

L'information d'information de Fisher est quant à elle l'inverse de l'espérance de la même matrice hessienne :

$$\begin{bmatrix}
\frac{1}{2}\mathrm{Tr}(PK_1PK_1) & \cdots & \frac{1}{2}\mathrm{Tr}(PK_1PK_k) & \frac{1}{2}\mathrm{Tr}(PK_1P)\\
\vdots & \vdots & \ddots & \vdots\\
\frac{1}{2}\mathrm{Tr}(PK_kPK_1) & \cdots & \frac{1}{2}\mathrm{Tr}(PK_kPK_k) & \frac{1}{2}\mathrm{Tr}(PK_kP)\\
\frac{1}{2}\mathrm{Tr}(PPK_1) & \cdots & \frac{1}{2}\mathrm{Tr}(PPK_k) & \frac{1}{2}\mathrm{Tr}(PP)
\end{bmatrix}.$$

Lemmes mathématiques

Les projections définies par une matrice orthogonale

Soit $A \in \mathbb{R}^{n\times(n-p)}$, $B \in \mathbb{R}^{n\times p}$, avec $\text{rank}(A) = n-p$ et $\text{rank}(B) = p$. Si $A'B = 0$, alors

$$A\left(A'A\right)^{-1}A' + B\left(B'B\right)^{-1}B' = \mathbb{I}_n$$

Preuve La projection sur $\Im A$ (l'espace vectoriel généré par les colonnes de A) est $A\left(A'A\right)^{-1}A$, pendant que la projection sur $\Im B$ est $B\left(B'B\right)^{-1}B'$. Sous nos hypothèses, $\Im A$ et $\Im B$ sont orthogonaux et leur union est l'espace total $\mathbb{R}^n$. Ce résultat nous permet d'en déduire les propriétés mathématiques suivantes. □

Deux lemma pour les matrices

Soit $X \in \mathbb{R}^{n\times p}$ de rang p, et $C \in \mathbb{R}^{(n-p)\times n}$ telle que $CC' = \mathbb{I}_p$. Si nous appliquons le résultat précédent pour $A = C'$ et $B = X$, nous obtenons

$$C'C = \mathbb{I}_n - X\left(X'X\right)^{-1}X'. \tag{6.1}$$

Si V est une matrice symétrique définie positive, avec $A = V^{\frac{1}{2}}C'$ et $B = V^{-\frac{1}{2}}X$, Nous obtenons aussi

$$V^{\frac{1}{2}}C'\left(CVC'\right)^{-1}CV^{\frac{1}{2}} = \mathbb{I}_n - V^{-\frac{1}{2}}X\left(X'V^{-1}X\right)^{-1}X'V^{-\frac{1}{2}}.$$

Par conséquent

$$C'\left(CVC'\right)^{-1}C = V^{-1} - V^{-1}X\left(X'V^{-1}X\right)^{-1}X'V^{-1}. \tag{6.2}$$

Cette matrice est notée P dans ce manuscrit.

Nous pouvons remarquer que P est singulière, $\Im P = \Im C' = (\Im X)^{\perp}$, et $PVP = P$.

Le déterminant de CVC'

Nous allons montrer que

$$\left|CVC'\right| \times \left|X'X\right| = |V| \times \left|X'V^{-1}X\right|$$

ou

$$\log\left|CVC'\right| + \log\left|X'X\right| = \log|V| + \log\left|X'V^{-1}X\right|. \tag{6.3}$$

Première preuve Pour $t \in [0,1]$, nous avons $V(t) = (1-t)\mathbb{I}_n + tV$ et donc $V(0) = \mathbb{I}_n$ et $V(1) = V$; pour tout $t \in [0,1]$, $V(t)$ est définie positive, et les fonctions suivantes sont définies :

$$\begin{aligned} f(t) &= \log\left|CV(t)C'\right| + \log\left|X'X\right| \\ g(t) &= \log|V(t)| + \log\left|X'V(t)^{-1}X\right|. \end{aligned}$$

Nous avons $f(0) = g(0)$. Si nous montrons que $f'(t) = g'(t)$, alors $f(1) = g(1)$ la première équation est prouvée. Notons que $V'(t) = V$.

Nous calculons

$$\begin{aligned} f'(t) &= \mathrm{Tr}\left(\left(CV(t)C'\right)^{-1} CVC'\right) \\ &= \mathrm{Tr}\left(C'\left(CV(t)C'\right)^{-1} C \times V\right) \\ &= \mathrm{Tr}\left(V(t)^{-1} \times V\right) - \mathrm{Tr}\left(V(t)^{-1} X \left(X'V(t)^{-1}X\right)^{-1} X'V(t)^{-1} \times V\right), \end{aligned}$$

avec l'équation 6.2. Nous avons aussi

$$\begin{aligned} g'(t) &= \mathrm{Tr}\left(V(t)^{-1} \times V\right) + \mathrm{Tr}\left(\left(X'V(t)^{-1}X\right)^{-1} \times X'\left(-V(t)^{-1} \times V \times V(t)^{-1}\right) X\right) \\ &= \mathrm{Tr}\left(V(t)^{-1} \times V\right) - \mathrm{Tr}\left(V(t)^{-1} X \left(X'V(t)^{-1}X\right)^{-1} X'V(t)^{-1} \times V\right), \end{aligned}$$

qui conclue la preuve. □

Seconde preuve Soit $D \in R^{p\times n}$ telle que les lignes de D sont une base orthogonale de $\Im X$, l'espace vectoriel engendré par les colonnes de X. Nous avons $DD' = \mathbb{I}_p$, $D'D = X(X'X)^{-1}X'$, et $CD' = 0$.

Soit $U = \begin{bmatrix} C \\ D \end{bmatrix}$; nous avons $UU' = U'U = \mathbb{I}_n$. Alors

$$UVU' = \begin{bmatrix} CVC' & CVD' \\ DVC' & DVD' \end{bmatrix}.$$

Le complément de Schur du bloc supérieur gauche est

$$S = DVD' - DVC'\left(CVC'\right)^{-1} CVD'.$$

En utilisant l'expression de P, nous obtenons facilement

$$S = DX\left(X'V^{-1}X\right)^{-1} X'D'.$$

Nous avons $|V| = |CVC'| \times |S|$. De plus,

$$\begin{aligned} |S| &= |DX| \times \left|\left(X'V^{-1}X\right)^{-1}\right| \times \left|X'D'\right| \\ &= \left|X'D'DX\right| \times \left|X'V^{-1}X\right|^{-1} \\ &= \left|X'X\right| \times \left|X'V^{-1}X\right|^{-1} \end{aligned}$$

comme $D'D = X(X'X)^{-1}X'$. □

Annexe 2 : L'astuce de la diagonalisation

Dans cette section, nous regardons uniquement le cas où $k = 1$ (nous omettrons donc les index pour $Z = Z_1$, $K = K_1$). Dans de cas, il est possible d'utiliser la transformation en composantes principales de K pour réécrire le modèle sous la forme diagonale. Cette astuce permet un gain de temps pour le calcul d'un modèle en particulier lorsque des PCs de la matrice de corrélation génétique K sont incluses dans le modèle en covariable. Afin d'appliquer cette astuce, il faut au préalable calculer la décomposition en valeurs singulières (SVD) de Z ou la décomposition en composantes principales de la matrice K.

Le modèle sous la forme diagonale

Prenons $Z \in \mathbb{R}^{n\times q}$ une matrice de génotypes centrés et réduits. Sa décomposition en valeurs singulières (encadré 1.17) s'écrit

$$Z = U\Sigma V'.$$

Avec ces notations, la décomposition en composantes principales de K s'écrit $K = U\Sigma^2 U'$. Les colonnes de la matrice orthogonale $U \in \mathbb{R}^{n\times n}$ sont les composantes principales normalisées des génotypes. Ainsi, nous décomposons la matrice U en deux blocs, $U = [U_1\, U_2]$ où U_1 est composée des p premières composantes principales incluses en covariable et U_2 les $q-p$ dernières PCs. Nous considérons alors le modèle linéaire mixte

$$Y = X\beta + Zu + e = [X_{\text{cov}}\, U_1]\,\beta + Zu + e \tag{6.4}$$

avec X_{cov} la matrice de r covariables, β les effets fixes, $u \sim \mathcal{N}(0, \tau\mathbb{I}_n)$, $e \sim \mathcal{N}(0, \sigma^2\mathbb{I}_n)$. Les variables incluses dans le modèle avec des effets fixes X sont donc composées de r « vraies covariables » (parmi lesquelles l'intercept sous la forme d'une colonne de 1) et de p composantes principales normalisées. Si nous notons $Y_D = U'Y$ et $X_D = U'X$, nous pouvons alors réécrire le modèle (6.4) sous la forme

$$Y_D = X_D\beta + U'U\Sigma V'u + e_D = X_D\beta + \Sigma w + e_D \tag{6.5}$$

où $w = V'u \sim \mathcal{N}(0, \tau\mathbb{I}_n)$ et $e_D = U'e \sim \mathcal{N}(0, \sigma^2\mathbb{I}_n)$. Nous pouvons remarquer que la matrice $X_D \in \mathbb{R}^{n\times(r+p)}$ peut se décomposer en blocs

$$X_D = \begin{bmatrix} U_1'X_0 & U_1'U_1 \\ U_2'X_0 & U_2'U_1 \end{bmatrix} = \begin{bmatrix} X_a & \mathbb{I}_p \\ X_b & 0 \end{bmatrix}. \tag{6.6}$$

Nous notons également X_c la matrice $(n \times r)$ valant $X_c = U'X_{\text{cov}} = \begin{bmatrix} X_a \\ X_b \end{bmatrix}$.

Le calcul et la maximisation de la vraisemblance restreinte

Sous le seconde modèle (6.5), la log-vraisemblance restreinte s'écrit

$$\ell(\tau, \sigma^2) = -\frac{1}{2}\left(\log|V| + \log\left|X_D'V^{-1}X_D\right| + Y_D'PY_D\right)$$

avec $V = \tau\Sigma^2 + \sigma^2\mathbb{I}_n$ et $P = V^{-1} - V^{-1}X_D\left(X_D'V^{-1}X_D\right)^{-1}X_D'V^{-1}$.

Il est alors possible de reparamétriser cette vraisemblance avec $v = \tau + \sigma^2$ et $h^2 = \frac{\tau}{\tau+\sigma^2}$. Avec ces nouveaux paramètres, la log-vraisemblance restreinte se réécrit

$$\ell(h^2, v) = -\frac{1}{2}\left(\log|V_0| + \log\left|X_D'V_0^{-1}X_D\right| + \frac{1}{v}Y_D'P_0Y_D + (n-r-p)\log(v)\right)$$

avec $V_0 = V/v = h^2\Sigma^2 + (1-h^2)\mathbb{I}_n$ et $P_0 = vP = V_0^{-1} - V_0^{-1}X_D\left(X_D'V_0^{-1}X_D\right)^{-1}X_D'V_0^{-1}$.

La dérivée de $\ell(h^2, v)$ par rapport à v est

$$\frac{\partial\ell}{\partial v}(h^2, v) = \frac{1}{2v^2}\left(Y_D'P_0Y_D - (n-r-p)v\right).$$

Cette dérivée nous donne la valeur de v qui maximise la log-vraisemblance restreinte,

$$v = \frac{1}{n-r-p}Y_D'P_0Y_D.$$

Maximiser la log-vraisemblance restreinte revient alors à minimiser l'expression suivante

$$\log|V_0| + \log\left|X_D'V_0^{-1}X_D\right| + (n-r-p)\log(Y_D'P_0Y_D) \tag{6.7}$$

qui dépend d'un seul paramètre, h^2. Elle peut être efficacement résolue à l'aide d'une méthode d'ordre 0 (par exemple la méthode « *Golden Search* » [190]).

Une méthode de calcul efficace de $\left|X_D'V_0^{-1}X_D\right|$ et P_0

Les termes les plus long à calculer sont $\left|X_D'V_0^{-1}X_D\right|$ et P_0. Nous allons donc ici regarder comment les calculer de façon efficace. Pour cela, nous découpons la matrice diagonale V_0 en deux blocs :

$$V_0 = \begin{bmatrix} V_{0a} & \mathbb{O}_{p\times(n-p)} \\ \mathbb{O}_{(n-p)\times p} & V_{0b} \end{bmatrix}$$

où $V_{0a} \in \mathbb{R}^{p\times p}$ et $V_{0b} \in \mathbb{R}^{(n-p)\times(n-p)}$ sont des matrices diagonales. Ainsi, en utilisant la décomposition de la matrice X_D (équation 6.6), nous obtenons

$$X_D'V_0^{-1}X_D = \begin{bmatrix} X_c'V_0^{-1}X_c & X_a'V_{0a}^{-1} \\ V_{0a}^{-1}X_a & V_{0a}^{-1} \end{bmatrix}. \tag{6.8}$$

Nous cherchons maintenant à calculer le déterminant de cette matrice. Pour cela, la matrice étant définie par bloc, nous pouvons utiliser le complément de Schur du bloc V_{0a}^{-1} dans la matrice précédente. Nous obtenons alors

$$\begin{aligned}\left|X_D'V_0^{-1}X_D\right| &= \left|V_{0a}^{-1}\right| \times \left|X_c'V_0^{-1}X_c - X_a'V_{0a}^{-1} \times V_{0a} \times V_{0a}^{-1}X_a\right| \\ &= \left|V_{0a}^{-1}\right| \times \left|X_b'V_{0b}^{-1}X_b\right|.\end{aligned}$$

La log-vraisemblance restreinte (équation (6.7)) peut donc se réécrire

$$\log|V_{0b}| + \log\left|X_b'V_{0b}^{-1}X_b\right| + (n-r-p)\log(Y_D'P_0Y_D). \tag{6.9}$$

Dans cette expression, seul le terme $Y_D' P_0 Y_D$ représente encore une difficulté. Afin de simplifier le calcul de ce terme, nous commençons par calculer l'inverse de la matrice (6.8) à l'aide de sa décomposition en blocs :

$$\left(X_D' V_0^{-1} X_D\right)^{-1} = \begin{bmatrix} \left(X_b' V_{0b}^{-1} X_b\right)^{-1} & -\left(X_b' V_{0b}^{-1} X_b\right)^{-1} X_a' \\ -X_a \left(X_b' V_{0b}^{-1} X_b\right)^{-1} & V_{0a} + X_a \left(X_b' V_{0b}^{-1} X_b\right)^{-1} X_a' \end{bmatrix}.$$

Nous avons maintenant tous les éléments pour calculer P_0.

$$\begin{aligned} X_D \left(X_D' V_0^{-1} X_D\right)^{-1} X_D' &= \begin{bmatrix} V_{0a} & \mathbb{O}_{p\times(n-p)} \\ \mathbb{O}_{(n-p)\times p} & X_b \left(X_b' V_{0b}^{-1} X_b\right)^{-1} X_b' \end{bmatrix} \qquad (6.10) \\ \Rightarrow V_0^{-1} X_D \left(X_D' V_0^{-1} X_D\right)^{-1} X_D' V_0^{-1} &= \begin{bmatrix} V_{0a}^{-1} & \mathbb{O}_{p\times(n-p)} \\ \mathbb{O}_{(n-p)\times p} & V_{0b}^{-1} X_b \left(X_b' V_{0b}^{-1} X_b\right)^{-1} X_b' V_{0b}^{-1} \end{bmatrix} \\ \Rightarrow P_0 &= \begin{bmatrix} \mathbb{O}_{p\times p} & \mathbb{O}_{p\times(n-p)} \\ \mathbb{O}_{(n-p)\times p} & V_{0b}^{-1} - V_{0b}^{-1} X_b \left(X_b' V_{0b}^{-1} X_b\right)^{-1} X_b' V_{0b}^{-1} \end{bmatrix} \end{aligned}$$

Grâce à cette expression, il est possible de calculer efficacement le terme $Y_D' P_0 Y_D$.

Le calcul des BLUPs

Une fois les paramètres du modèle estimés, $\left(\widehat{h^2}, \widehat{v}\right)$ ou $\left(\widehat{\tau}, \widehat{\sigma^2}\right)$, à l'aide des calculs faits dans le paragraphe 2.2.6, nous pouvons en déduire les BLUPs :

- ▷ des effets fixes β

$$\widehat{\beta} = \left(X_D' X_D\right)^{-1} X_D'(Y_D - \widehat{V}\widehat{P}Y_D) = \left(X_D' X_D\right)^{-1} X_D'(Y_D - \widehat{V_0}\widehat{P_0}Y_D)$$

  de variance $\left(X_D' V_0^{-1} X_D\right)^{-1}$,

- ▷ des effets aléatoires ω

$$\widehat{\omega} = \widehat{\tau}\Sigma\widehat{P}Y_D = \widehat{h^2}\Sigma\widehat{P_0}Y_D.$$

Ici, les BLUPs des effets fixes peuvent nécessiter un important temps de calcul. Afin d'optimiser ce temps, nous pouvons utiliser une nouvelle fois la décomposition de X_D en blocs :

$$\begin{aligned} (X_D' X_D)^{-1} &= \begin{bmatrix} (X_b' X_b)^{-1} & -(X_b' X_b)^{-1} X_a' \\ -X_a (X_b' X_b)^{-1} & \mathbb{I}_p + X_a (X_b' X_b)^{-1} X_a' \end{bmatrix} \\ \Rightarrow (X_D' X_D)^{-1} X_D' \begin{bmatrix} z_a \\ z_b \end{bmatrix} &= \begin{bmatrix} (X_b' X_b)^{-1} X_b' z_b \\ z_a - X_a (X_b' X_b)^{-1} X_b' z_b \end{bmatrix}. \end{aligned}$$

Il est nécessaire de rappeler que ces BLUPs sont ceux du modèle (6.5). Pour le modèle d'origine (6.4), les BLUPs des effets fixes sont les mêmes et ceux des effets aléatoires u se retrouvent en multipliant à gauche les BLUPs du modèle (6.4) par V, $\widehat{u} = V\widehat{\omega}$.

Le partitionnement de la variance

À partir des calculs de la section 2.3, avec $k = 1$, l'espérance de $\mathrm{ev}(Y)$ vaut

$$\mathbb{E}\left[\mathrm{ev}(Y)\right] = \mathrm{ev}(X\beta) + \tau\Psi(K) + \sigma^2.$$

Nous sommes ainsi intéressés pour calculer de l'estimation non biaisée de $\mathrm{ev}(X\beta)$:

$$\mathrm{ev}(X\widehat{\beta}) - \Psi\left(X\mathrm{Var}\left(\widehat{\beta}\right)X'\right). \tag{6.11}$$

Comme $X = UX_D$, nous avons

$$\begin{aligned} X\mathrm{var}\left(\widehat{\beta}\right)X' &= U\left(X_D\mathrm{var}\left(\widehat{\beta}\right)X_D'\right)U' \\ &= U\left(X_D\left(X_D'V_0^{-1}X_D\right)^{-1}X_D'\right)U' \end{aligned}$$

En utilisant l'écriture par blocs de $X_D\left(X_D'V_0^{-1}X_D\right)^{-1}X_D'$ données dans l'équation (6.10), nous obtenons

$$X\mathrm{var}\left(\widehat{\beta}\right)X' = U_1V_{0a}U_1' + U_2X_b\left(X_b'V_{0b}^{-1}X_b\right)^{-1}X_b'U_2',$$

et nous avons

$$\Psi\left(X\mathrm{var}\left(\widehat{\beta}\right)X'\right) = \Psi\left(U_1V_{0a}U_1'\right) + \Psi\left(U_2X_b\left(X_b'V_{0b}^{-1}X_b\right)^{-1}X_b'U_2'\right).$$

Annexe 3 : Les loadings des données 3C

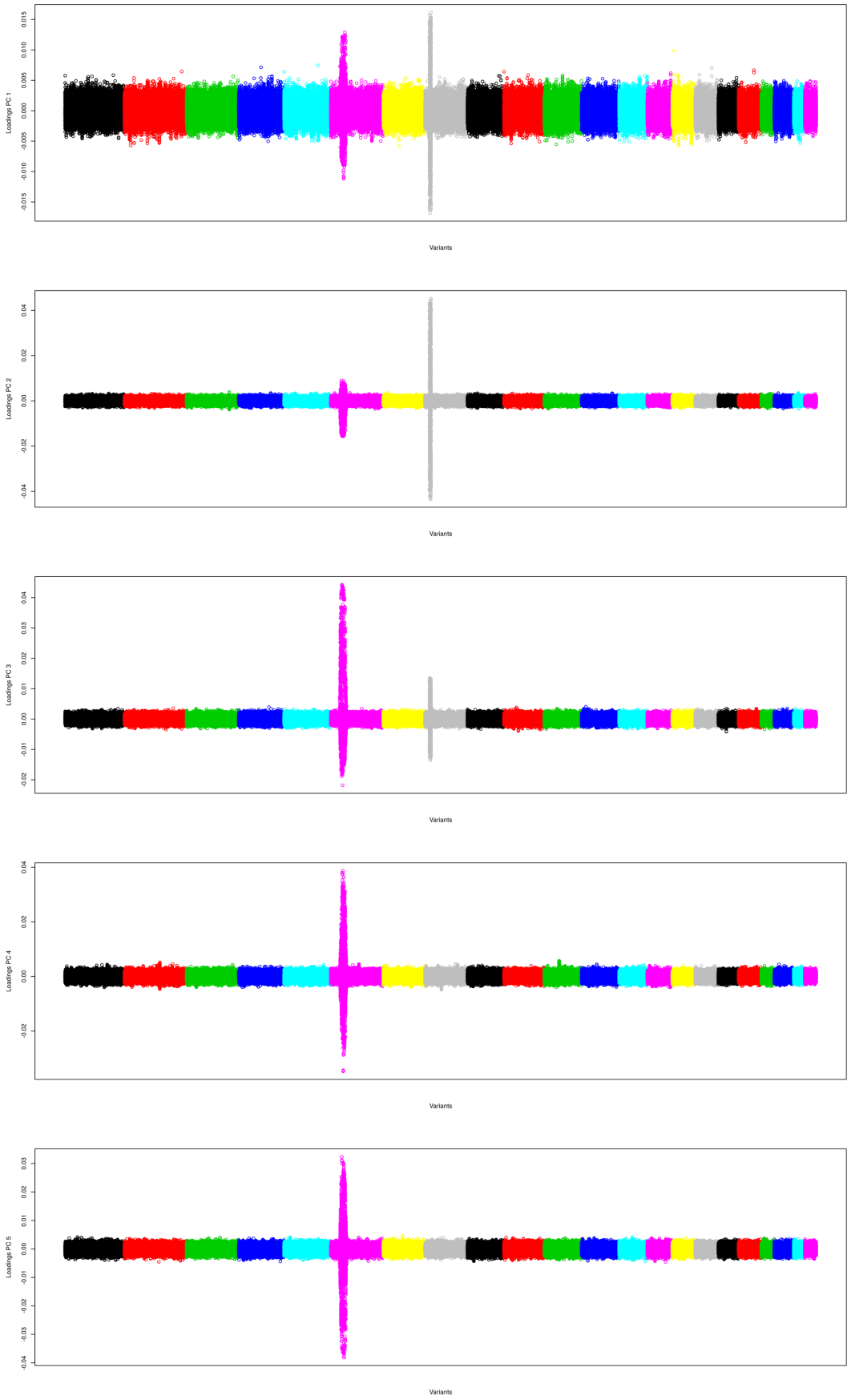


FIGURE Annexe 3.1 – Loadings des données 3C.

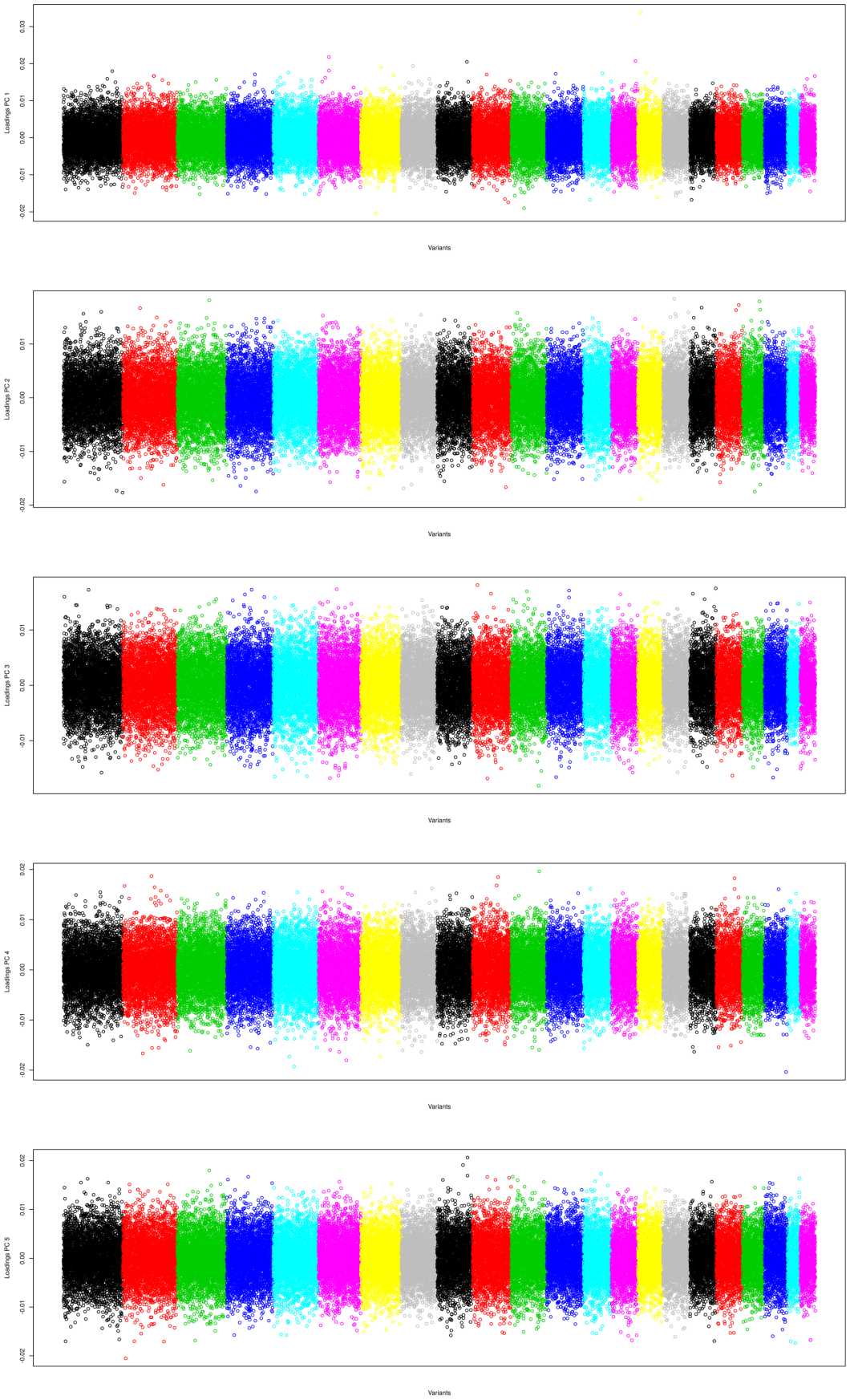


FIGURE Annexe 3.2 – Loadings des données 3C élaguées.

Annexe 4 : Étude 3C, le calcul des variances avec d'autres PCs

Dans cette annexe, nous montrons les résultats de l'analyse décrite dans la section 3 en utilisant cette fois les PCs calculées avec les données non élaguées pour corriger la stratification de population. La première figure (Annexe 4.1) donne les coefficients de corrélation entre les premières PCs et les coordonnées géographiques. Les figures Annexe 4.2 et Annexe 4.3 montrent les résultats pour :

- ▷ les coordonnées géographiques,
- ▷ les 3 traits anthropométriques pour lesquels nous avons constaté un changement des estimations avec l'inclusion des premières PCs en effets fixes (stature, la circonférence crânienne et le rapport de la circonférence de la taille sur celle des hanches).

Les résultats sont donnés plus en détails dans les tableaux Annexe 4.2, Annexe 4.1 et Annexe 4.3.

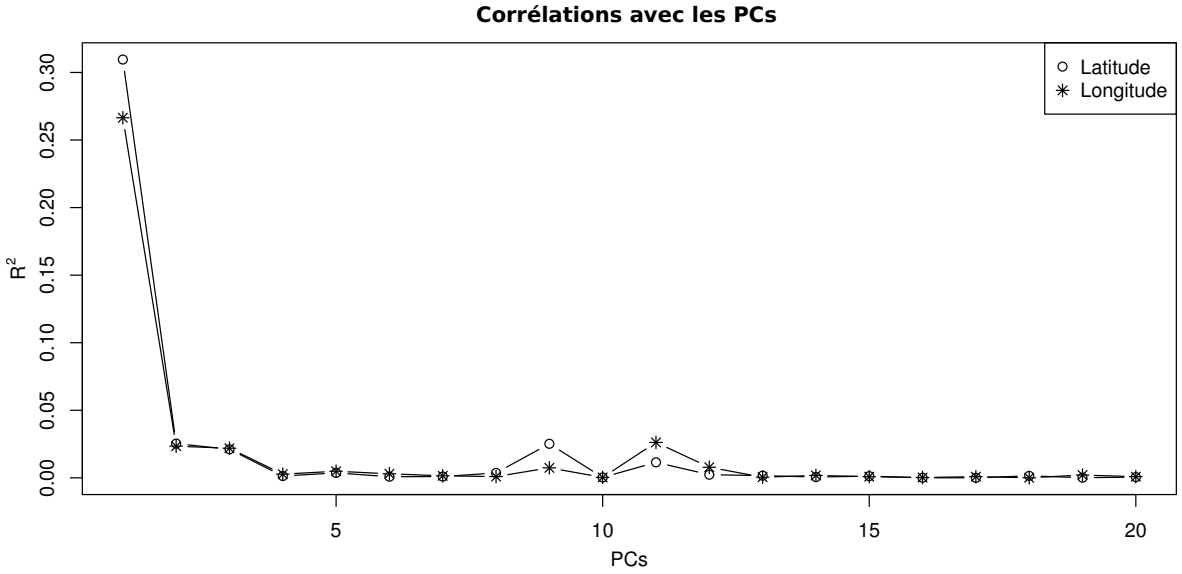


FIGURE Annexe 4.1 – Corrélation (R^2) entre les coordonnées géographiques et les 20 première PCs de la matrice de similarité calculée sur les données non élaguées.

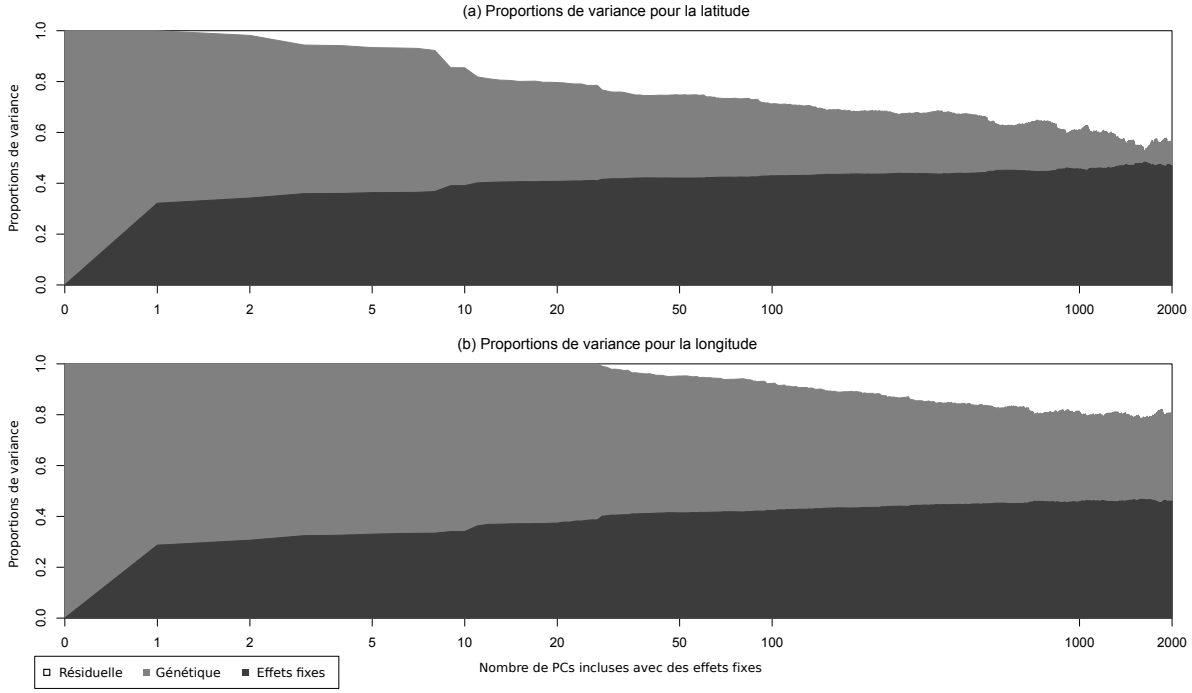


FIGURE Annexe 4.2 – Proportions de variance estimées pour les coordonnées géographiques en fonction du nombre de PCs incluses dans le modèle avec des effets fixes (PCs calculées avec la totalité des SNPs). Le blanc, le gris clair et le gris foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

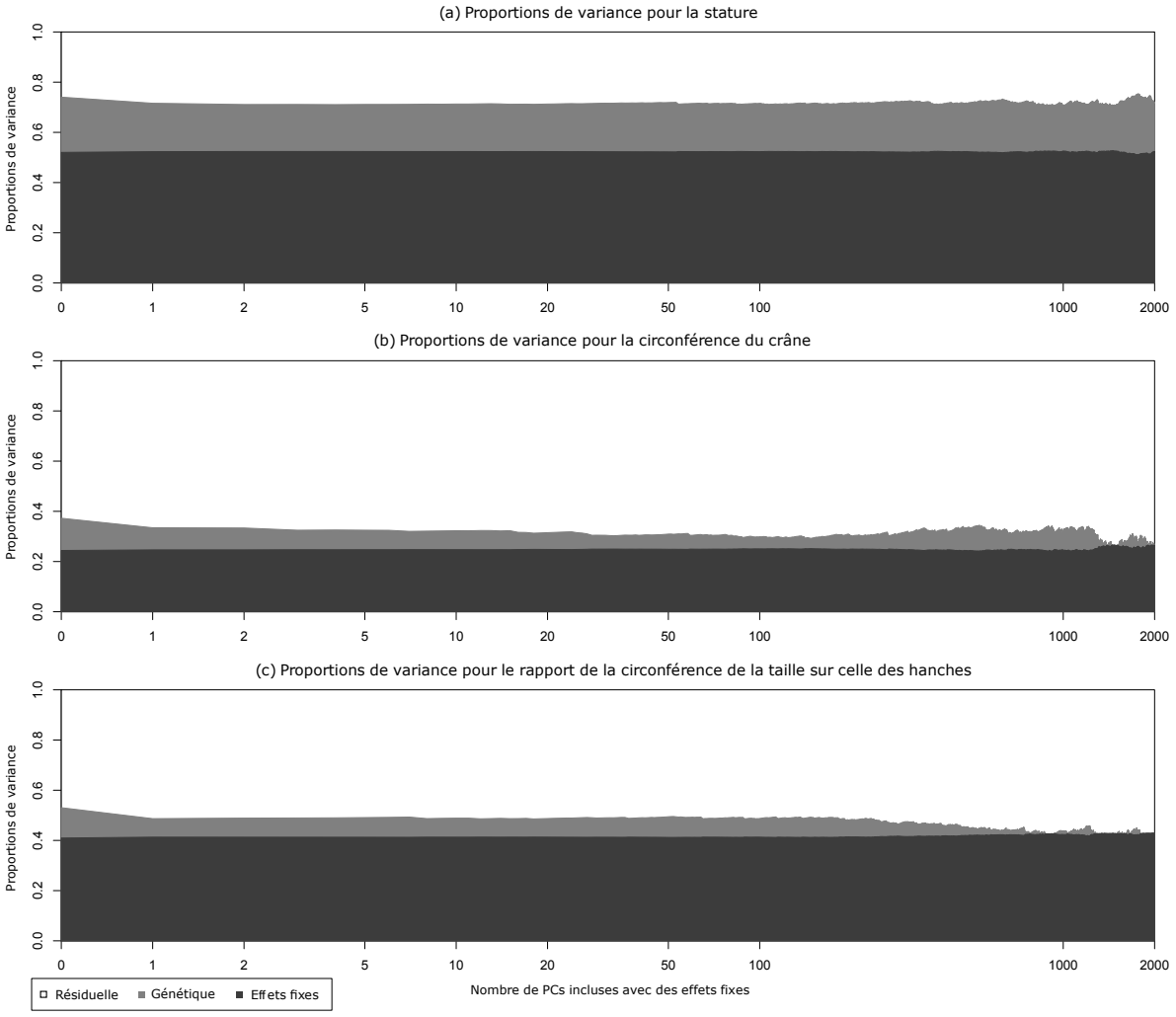


FIGURE Annexe 4.3 – Proportions de variance estimées pour (a) la stature, (b) la circonférence crânienne et (c) le ratio taille sur hanches en fonction du nombre de PCs incluses dans le modèle avec des effets fixes (PCs calculées avec la totalité des SNPs). Le blanc, le gris clair et le gris foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

	LRT	p-valeur	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma^2_P}$ (se)	$\widehat{h^2}$ (se)
0 PC	1892.63	<1e-40	4.34 (0.085)	2.9e-4 (0.089)	4.34 (0.084)	1.000
1 PC	822.44	<1e-40	3.90 (0.076)	2.6e-4 (0.072)	3.90 (0.076)	1.000
2 PCs	708.77	<1e-40	3.86 (0.075)	2.6e-4 (0.070)	3.86 (0.075)	1.000
3 PCs	600.64	<1e-40	3.82 (0.075)	2.5e-4 (0.069)	3.82 (0.074)	1.000
4 PCs	598.75	<1e-40	3.82 (0.074)	2.5e-4 (0.069)	3.82 (0.074)	1.000
5 PCs	570.04	<1e-40	3.80 (0.074)	2.5e-4 (0.068)	3.80 (0.074)	1.000
10 PCs	524.18	<1e-40	3.77 (0.074)	2.5e-4 (0.067)	3.77 (0.073)	1.000
20 PCs	377.77	<1e-40	3.65 (0.071)	2.4e-4 (0.063)	3.65 (0.071)	1.000
50 PCs	207.02	<1e-40	3.19 (0.068)	0.28 (0.060)	3.47 (0.068)	0.919 (0.061)
100 PCs	168.31	8.7e-39	2.96 (0.068)	0.46 (0.060)	3.42 (0.067)	0.867 (0.064)
500 PCs	69.56	3.7e-17	2.29 (0.069)	0.98 (0.061)	3.27 (0.067)	0.701 (0.079)
1000 PCs	38.23	3.2e-10	2.07 (0.073)	1.14 (0.063)	3.22 (0.070)	0.645 (0.095)
2 000 PCs	15.93	3.3e-5	2.02 (0.087)	1.18 (0.068)	3.20 (0.079)	0.632 (0.136)

Table Annexe 4.1 – Estimations des paramètres du modèle pour la longitude et leur erreur-type (se) en fonction du nombre de PCs incluses dans le modèle : test du rapport de vraisemblance (LRT) et la p-valeur associée, $\widehat{\tau}$, $\widehat{\sigma^2}$ et $\widehat{\sigma^2_P} = \widehat{\tau} + \widehat{\sigma^2}$ l'estimation des variances génétique, résiduelle et totale respectivement et $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ l'héritabilité estimée.

	LRT	p-valeur	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma^2_P}$ (se)	$\widehat{h^2}$ (se)
0 PC	1854.15	<1e-40	2.05 (0.040)	1.4e-4 (0.039)	2.05 (0.040)	1.000
1 PC	558.69	<1e-40	1.81 (0.035)	1.2e-4 (0.031)	1.81 (0.035)	1.000
2 PCs	424.64	<1e-40	1.74 (0.035)	0.05 (0.030)	1.79 (0.035)	0.972 (0.050)
3 PCs	312.82	<1e-40	1.61 (0.035)	0.15 (0.030)	1.76 (0.034)	0.912 (0.053)
4 PCs	307.45	<1e-40	1.60 (0.035)	0.16 (0.030)	1.76 (0.034)	0.908 (0.053)
5 PCs	292.00	<1e-40	1.57 (0.034)	0.18 (0.030)	1.75 (0.034)	0.896 (0.054)
10 PCs	177.38	<1e-40	1.28 (0.033)	0.40 (0.030)	1.69 (0.033)	0.761 (0.058)
20 PCs	114.47	5.1e-27	1.08 (0.033)	0.57 (0.030)	1.65 (0.032)	0.656 (0.062)
50 PCs	74.84	2.5e-18	0.91 (0.032)	0.70 (0.030)	1.61 (0.032)	0.563 (0.065)
100 PCs	52.25	2.4e-13	0.79 (0.032)	0.80 (0.030)	1.59 (0.031)	0.496 (0.068)
500 PCs	21.08	2.2e-6	0.60 (0.033)	0.95 (0.031)	1.55 (0.032)	0.387 (0.082)
1000 PCs	7.10	3.9e-3	0.43 (0.036)	1.08 (0.032)	1.51 (0.033)	0.286 (0.104)
2 000 PCs	1.18	0.139	0.28 (0.044)	1.20 (0.035)	1.48 (0.037)	0.189 (0.167)

Table Annexe 4.2 – Estimations des paramètres du modèle pour la latitude et leur erreur-type (se) en fonction du nombre de PCs incluses dans le modèle : test du rapport de vraisemblance (LRT) et la p-valeur associée, $\widehat{\tau}$, $\widehat{\sigma^2}$ et $\widehat{\sigma^2_P} = \widehat{\tau} + \widehat{\sigma^2}$ l'estimation des variances génétique, résiduelle et totale respectivement et $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ l'héritabilité estimée.

Trait		LRT	p-valeur	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma^2_P}$ (se)	$\widehat{h^2}$ (se)
Stature	0 PC	68.80	5.6e-17	17.48 (0.719)	20.64 (0.689)	38.12 (0.687)	0.459 (0.061)
	1 PC	41.96	4.7e-11	15.32 (0.714)	22.57 (0.691)	37.89 (0.686)	0.404 (0.064)
	5 PCs	37.06	5.7e-10	14.90 (0.714)	22.95 (0.691)	37.85 (0.686)	0.394 (0.066)
	10 PCs	37.58	4.4e-10	15.05 (0.714)	22.82 (0.691)	37.87 (0.686)	0.397 (0.066)
	20 PCs	37.02	5.8e-10	15.05 (0.716)	22.82 (0.691)	37.87 (0.687)	0.398 (0.066)
Poids	0 PC	13.92	9.6e-5	28.24 (2.32)	97.23 (2.32)	125.47 (2.11)	0.225 (0.062)
	1 PC	13.80	1.0e-4	28.26 (2.32)	97.21 (2.32)	125.47 (2.11)	0.225 (0.062)
	5 PCs	13.07	1.5e-4	27.90 (2.32)	97.53 (2.32)	125.43 (2.12)	0.222 (0.062)
	10 PCs	13.45	1.2e-4	28.38 (2.33)	97.11 (2.32)	125.49 (2.11)	0.226 (0.063)
	20 PCs	12.41	2.1e-4	27.55 (2.33)	97.82 (2.32)	125.38 (2.12)	0.220 (0.063)
BMI	0 PC	10.44	6.2e-4	3.23 (0.302)	13.09 (0.302)	16.31 (0.304)	0.198 (0.063)
	1 PC	9.34	1.1e-3	3.12 (0.302)	13.18 (0.302)	16.30 (0.304)	0.191 (0.064)
	5 PCs	9.65	9.4e-4	3.18 (0.302)	13.12 (0.303)	16.31 (0.304)	0.195 (0.064)
	10 PCs	10.21	7.0e-4	3.29 (0.303)	13.04 (0.303)	16.32 (0.304)	0.201 (0.064)
	20 PCs	8.26	2.0e-3	3.00 (0.303)	13.28 (0.303)	16.28 (0.304)	0.184 (0.065)
Circonférence crânienne	0 PC	9.85	8.5e-4	0.722 (0.078)	3.52 (0.079)	4.24 (0.080)	0.170 (0.057)
	1 PC	3.70	0.027	0.495 (0.078)	3.73 (0.079)	4.23 (0.080)	0.117 (0.062)
	5 PCs	2.67	0.051	0.437 (0.078)	3.79 (0.079)	4.22 (0.079)	0.103 (0.064)
	10 PCs	2.46	0.058	0.424 (0.078)	3.80 (0.079)	4.22 (0.080)	0.100 (0.065)
	20 PCs	1.88	0.085	0.376 (0.078)	3.84 (0.079)	4.22 (0.079)	0.089 (0.065)
Ratio taille sur hanche	0 PC	13.09	1.5e-4	9.2e-04 (8.6e-5)	3.6e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.205 (0.061)
	1 PC	3.54	0.030	5.6e-04 (8.6e-5)	3.9e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.126 (0.068)
	5 PCs	3.98	0.023	6.0e-04 (8.6e-5)	3.9e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.133 (0.068)
	10 PCs	3.65	0.028	5.6e-04 (8.6e-5)	3.9e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.125 (0.066)
	20 PCs	3.41	0.032	5.6e-04 (8.6e-5)	3.9e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.126 (0.070)

TABLE Annexe 4.3 – Estimations des paramètres du modèle pour les traits anthropologiques et leur erreur-type (se) en fonction du nombre de PCs incluses dans le modèle : test du rapport de vraisemblance (LRT) et la p-valeur associée, $\widehat{\tau}$, $\widehat{\sigma^2}$ et $\widehat{\sigma^2_P} = \widehat{\tau} + \widehat{\sigma^2}$ l'estimation des variances génétique, résiduelle et totale respectivement et $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ l'héritabilité estimée.

Annexe 5 : Les analyses complémentaires pour les traits anthropométriques

Nous donnons ici les résultats des analyses complémentaires faites sur la stature pour la circonférence crânienne et le rapport de la circonférence de la taille sur celle des hanches. Ces analyses se décomposent en quatre parties :

- ▷ l'ajout de la longitude et la latitude en covariable,
- ▷ l'inclusion d'un effet centre,
- ▷ l'étude des individus recrutés à Dijon,
- ▷ une matrice de *kinship* par chromosome [119].

Pour la dernière analyse, le BMI et le poids ont aussi été considérés.

L'ajout de la longitude et la latitude en covariable

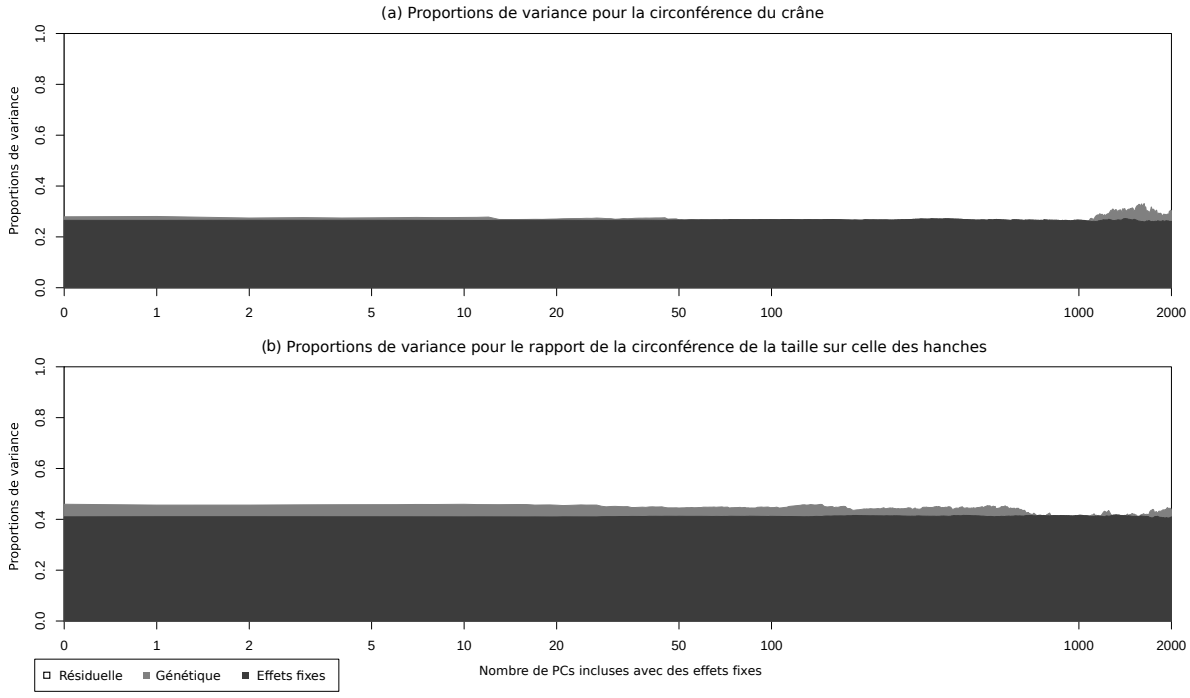


FIGURE Annexe 5.1 – Estimations des proportions de variance pour (a) la circonférence du crâne, et (b) le ratio de la circonférence de la taille sur celle des hanches avec les coordonnées géographiques en covariable en fonction du nombre de PCs incluses dans le modèle avec des effets fixes (échelle logarithmique). Le blanc et les gris clair et foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

Trait		LRT	p-valeur	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma_P^2}$ (se)	$\widehat{h^2}$ (se)
Circonférence	0 PC	0.07	0.39	0.077 (0.078)	3.95 (0.080)	4.03 (0.080)	0.019 (0.071)
crânienne	1 PC	0.09	0.38	0.084 (0.078)	3.95 (0.080)	4.03 (0.080)	0.021 (0.071)
	5 PCs	0.03	0.43	0.053 (0.078)	3.98 (0.080)	4.03 (0.080)	0.013 (0.071)
	10 PCs	0.05	0.41	0.064 (0.078)	3.97 (0.080)	4.03 (0.080)	0.016 (0.071)
Ratio	0 PC	1.35	0.12	3.8e-4 (9.1e-5)	4.2e-3 (9.3e-5)	4.5e-3 (9.3e-5)	0.084 (0.071)
taille	1 PC	1.13	0.14	3.5e-4 (9.1e-5)	4.2e-3 (9.3e-5)	4.5e-3 (9.3e-5)	0.078 (0.072)
sur	5 PCs	1.23	0.13	3.7e-4 (9.1e-5)	4.2e-3 (9.3e-5)	4.5e-3 (9.3e-5)	0.081 (0.076)
hanche	10 PCs	1.32	0.13	3.8e-4 (9.1e-5)	4.2e-3 (9.3e-5)	4.5e-3 (9.3e-5)	0.084 (0.073)

Table Annexe 5.1 – Estimations des paramètres du modèle pour deux des traits anthropométriques et leur erreur-type lorsque les coordonnées géographiques sont incluses en covariable en plus du sexe, de l'âge et d'un nombre variable de PCs. La table contient aussi le test du rapport de vraisemblance (LRT) testant la significativité de l'héritabilité et la p-valeur associée. $\widehat{\tau}$ est l'estimation de la variance génétique, $\widehat{\sigma^2}$ l'estimation de la variance résiduelle, $\widehat{\sigma_P^2} = \widehat{\tau} + \widehat{\sigma^2}$ l'estimation de la variance totale et $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ l'héritabilité estimée.

L'effet centre

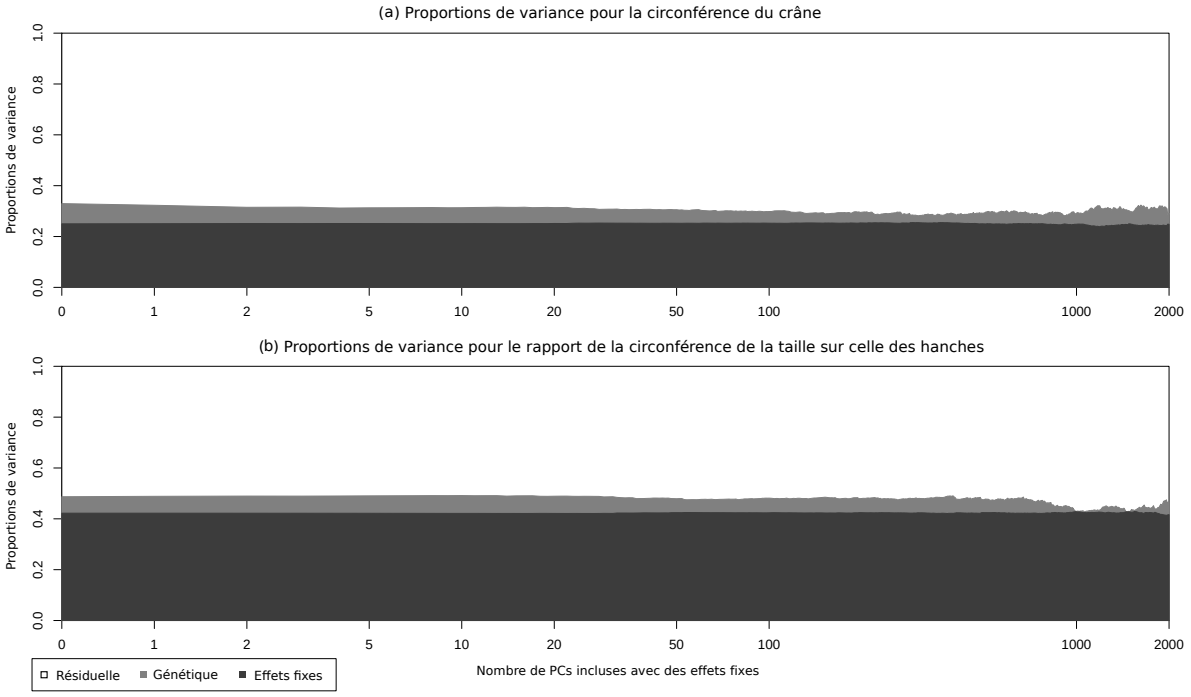


Figure Annexe 5.2 – Estimations des proportions de variance pour (a) la circonférence du crâne, et (b) le ratio de la circonférence de la taille sur celle des hanches, avec le centre inclus en covariable et en fonction du nombre de PCs incluses dans le modèle avec des effets fixes (échelle logarithmique). Le blanc et les gris clair et foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

Trait		LRT	p-valeur	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma_P^2}$ (se)	$\widehat{h^2}$ (se)
Circonférence	0 PC	2.99	**0.042**	0.446 (0.077)	3.76 (0.079)	4.21 (0.079)	0.106 (0.062)
crânienne	1 PC	2.39	0.061	0.407 (0.077)	3.80 (0.079)	4.21 (0.079)	0.097 (0.063)
	5 PCs	1.69	0.097	0.347 (0.077)	3.85 (0.079)	4.20 (0.079)	0.083 (0.064)
	10 PCs	1.73	0.094	0.351 (0.077)	3.85 (0.079)	4.20 (0.079)	0.084 (0.064)
Ratio	0 PC	2.94	**0.043**	4.9e-4 (8.4e-5)	3.9e-3 (8.5e-5)	4.4e-3 (8.6e-5)	0.112 (0.067)
taille	1 PC	3.10	**0.039**	5.1e-4 (8.4e-5)	3.9e-3 (8.5e-5)	4.4e-3 (8.6e-5)	0.115 (0.066)
sur	5 PCs	3.30	**0.035**	5.2e-4 (8.4e-5)	3.9e-3 (8.6e-5)	4.4e-3 (8.6e-5)	0.118 (0.066)
hanches	10 PCs	3.44	**0.032**	5.3e-4 (8.4e-5)	3.9e-3 (8.6e-5)	4.4e-3 (8.6e-5)	0.121 (0.067)

TABLE Annexe 5.2 – Estimations des paramètres du modèle pour deux des traits anthropométriques et leur erreur-type lorsque le centre est inclus comme covariable. Le sexe, l'âge et un nombre variable de PCs sont inclus dans le modèle. La table contient aussi le test du rapport de vraisemblance (LRT) testant la significativité de l'héritabilité et la p-valeur associée. $\widehat{\tau}$ est l'estimation de la variance génétique, $\widehat{\sigma^2}$ l'estimation de la variance résiduelle, $\widehat{\sigma_P^2} = \widehat{\tau} + \widehat{\sigma^2}$ l'estimation de la variance totale et $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ l'héritabilité estimée.

L'étude des individus recrutés à Dijon

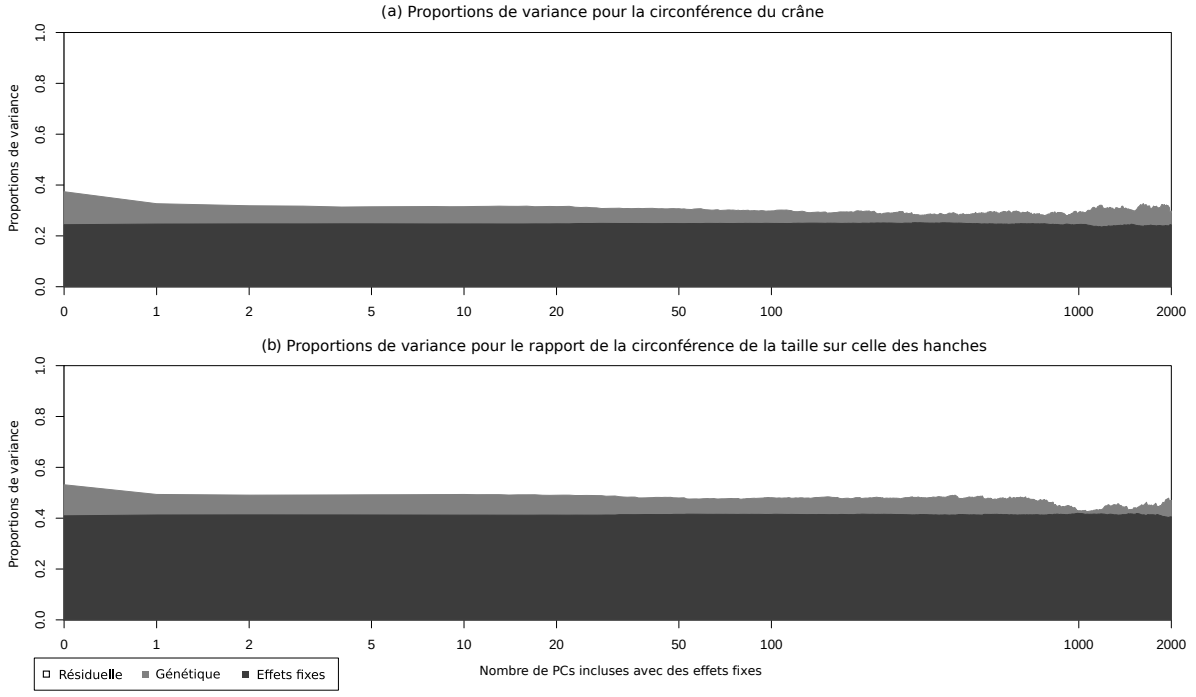


FIGURE Annexe 5.3 – Estimations des proportions de variance pour (a) la circonférence du crâne, et (b) le ratio de la circonférence de la taille sur celle des hanches lorsque seuls les individus recrutés à Dijon sont analysés en fonction du nombre de PCs incluses dans le modèle avec des effets fixes (échelle logarithmique). Le blanc et les gris clair et foncé représentent les variances résiduelles, génétiques et des effets fixes respectivement.

Trait		LRT	p-valeur	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma_P^2}$ (se)	$\widehat{h^2}$ (se)
Circonférence	0 PC	10.00	**7.8e-4**	0.729 (0.078)	3.52 (0.079)	4.24 (0.080)	0.172 (0.058)
crânienne	1 PC	2.88	**0.045**	0.449 (0.078)	3.78 (0.079)	4.23 (0.080)	0.106 (0.063)
	5 PCs	1.96	0.081	0.376 (0.078)	3.85 (0.079)	4.22 (0.079)	0.089 (0.064)
	10 PCs	2.01	0.078	0.381 (0.078)	3.84 (0.079)	4.22 (0.080)	0.090 (0.064)
Ratio	0 PC	13.33	**1.3e-4**	9.3e-4 (8.6e-5)	3.6e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.207 (0.062)
taille	1 PC	4.45	**0.017**	6.2e-4 (8.6e-5)	3.9e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.138 (0.067)
sur	5 PCs	4.26	**0.020**	6.1e-4 (8.6e-5)	3.9e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.135 (0.067)
hanche	10 PCs	4.44	**0.018**	6.2e-4 (8.6e-5)	3.9e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.138 (0.067)

Table Annexe 5.3 – Estimations des paramètres du modèle pour deux des traits anthropométriques et leur erreur-type lorsque seuls les individus recrutés à Dijon sont analysés. Le sexe, l'âge et un nombre variable de PCs sont inclus dans le modèle. La table contient aussi le test du rapport de vraisemblance (LRT) testant la significativité de l'héritabilité et la p-valeur associée. $\widehat{\tau}$ est l'estimation de la variance génétique, $\widehat{\sigma^2}$ l'estimation de la variance résiduelle, $\widehat{\sigma_P^2} = \widehat{\tau} + \widehat{\sigma^2}$ l'estimation de la variance totale et $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ l'héritabilité estimée.

Une matrice de *kinship* par chromosome

Poids	0 PC	1.8e-4	0.98	4.0e-6	0.51
	1 PC	3.1e-4	0.69	3.9e-7	0.94
	2 PCs	2.3e-4	0.75	4.3e-7	0.93
BMI	0 PC	-1.1e-5	0.99	4.0e-6	0.50
	1 PC	-5.4e-4	0.52	4.1e-6	0.48
	2 PCs	-5.5e-4	0.50	4.0e-6	0.48
Circonférence	0 PC	1.2e-5	0.99	2.0e-5	**6.8e-3**
crânienne	1 PC	6.6e-4	0.14	-4.5e-6	0.15
	2 PCs	6.4e-4	0.15	-5.3e-6	0.089
Ratio	0 PC	-1.8e-4	0.90	3.7e-5	**1.6e-3**
taille sur	1 PC	-5.1e-4	0.52	1.1e-5	0.053
hanche	2 PCs	-4.1e-4	0.58	9.6e-6	0.075

Table Annexe 5.4 – Résultats de la régression de la différence des héritabilités estimées sur un seul chromosome dans un modèle avec un seul chromosome et celui avec tous les chromosomes en même temps sur la longueur des chromosomes (en mega-base). Un intercept significatif est interprété comme une indication de la présence d'apparentement cryptique. Une pente significative est interprétée comme une indication de la présence de stratification de population.

Annexe 6 : Les fréquences génétiques des paires de germains atteints

Nous rappelons les notations suivantes :

- ▷ Att_1 l'événement *"germain index atteint"*,
- ▷ Att_2 l'événement *"deuxième germain atteint"*,
- ▷ f_a et f_A les fréquences alléliques des allèles a et A du SNP A,
- ▷ IBD l'état IBD des deux germains atteints,
- ▷ G_1 et G_2 les génotypes du cas index et du second germain respectivement,
- ▷ ψ_0 le risque de base du génotype AA,
- ▷ ψ le risque allélique associé à l'allèle a.

Nous cherchons à calculer :

$$\begin{aligned}
P_{k.i} &= \mathbb{P}[G_1 = k, IBD = i|ASP] \\
&= \frac{\mathbb{P}[ASP|G_1 = k, IBD = i]\mathbb{P}[G_1 = k, IBD = i]}{\mathbb{P}[ASP]} \\
&= \frac{\mathbb{P}[ASP|G_1 = k, IBD = i]\mathbb{P}[G_1 = k]\mathbb{P}[IBD = i]}{\sum_{l,j}\mathbb{P}[ASP|G_1 = l, IBD = j]\mathbb{P}[G_1 = l]\mathbb{P}[IBD = j]}
\end{aligned}$$

Il nous reste $\mathbb{P}[ASP|G_1 = k, IBD = i]$ à calculer :

$$\begin{aligned}
\mathbb{P}[ASP|G_1 = k, IBD = i] =& \mathbb{P}[Att_1, Att_2|G_1 = k, IBD = i] \\
=& \mathbb{P}[Att_1|G_1 = k]\mathbb{P}[Att_2|Att_1, G_1 = k, IBD = i] \\
=& \psi_0\psi^k \sum_l \mathbb{P}[Att_2|Att_1, G_1 = k, G_2 = l, IBD = i]\mathbb{P}[G_2 = l|Att_1, G_1 = k, IBD = i] \\
=& \psi_0\psi^k \sum_l \psi_0\psi^l \mathbb{P}[G_2 = l|G_1 = k, IBD = i]
\end{aligned}$$

avec :

$$\begin{aligned}
&\mathbb{P}[G_2 = l|G_1 = k, IBD = 0] = \mathbb{P}[G_2 = l] = (2 - \mathbb{1}_{k=1})f_a^{2-k}f_A^{k-2} \\
&[\mathbb{P}[G_2 = l|G_1 = k, IBD = 1]]_{kl} = \begin{pmatrix} f_A & f_a & 0 \\ \frac{1}{2}f_A & \frac{1}{2} & \frac{1}{2}f_a \\ 0 & f_A & f_a \end{pmatrix} \\
&\mathbb{P}[G_2 = l|G_1 = k, IBD = 2] = \mathbb{1}_{k=l}
\end{aligned}$$

Nous en déduisons les valeurs de $P_{k.i}$ suivantes :

$$\begin{aligned}
P_{0.2} &= \tfrac{1}{\pi} \times (1-f_a)^2 \\
P_{0.1} &= \tfrac{1}{\pi} \times 2(1-f_a)^2 \times ((1-f_a)+f_a\psi) \\
P_{0.0} &= \tfrac{1}{\pi} \times (1-f_a)^2 \times ((1-f_a)^2+2(1-f_a)f_a\psi+f_a^2\psi^2) \\
P_{1.2} &= \tfrac{1}{\pi} \times 2(1-f_a)f_a \times \psi^2 \\
P_{1.1} &= \tfrac{1}{\pi} \times 2(1-f_a)f_a \times \psi \times ((1-f_a)+\psi+f_a\psi^2) \qquad (6.12) \\
P_{1.0} &= \tfrac{1}{\pi} \times 2(1-f_a)f_a \times \psi \times ((1-f_a)^2+2(1-f_a)f_a\psi+f_a^2\psi^2) \\
P_{2.2} &= \tfrac{1}{\pi} \times f_a^2 \times \psi^4 \\
P_{2.1} &= \tfrac{1}{\pi} \times 2f_a^2 \times \psi^2 \times ((1-f_a)\psi+f_a\psi^2) \\
P_{2.0} &= \tfrac{1}{\pi} \times f_a^2 \times \psi^2 \times ((1-f_a)^2+2(1-f_a)f_a\psi+f_a^2\psi^2)
\end{aligned}$$

(6.13)

avec $\pi = (2+4f_a(\psi-1)+f_a(1+f_a)(1-\psi)^2)^2$.

Annexe 7 : Étude de la Sclérose en Plaques

Dans cette annexe, nous donnons pour commencer la projection des génotypes de nos individus sur les deux premières composantes principales des Européens de 1000 Genomes. Puis, nous présentons les résultats des études d'association utilisant les modèles mixtes des données de la Sclérose en Plaques lorsque la structure familiale n'est plus corrigée avec la matrice de *kinship*, $K = 2\Phi$, mais avec une matrice de corrélation génétique (GRM). Nous regardons les résultats avec les GRM estimées sur la totalité des données génétiques et sur les données génétiques élaguées.

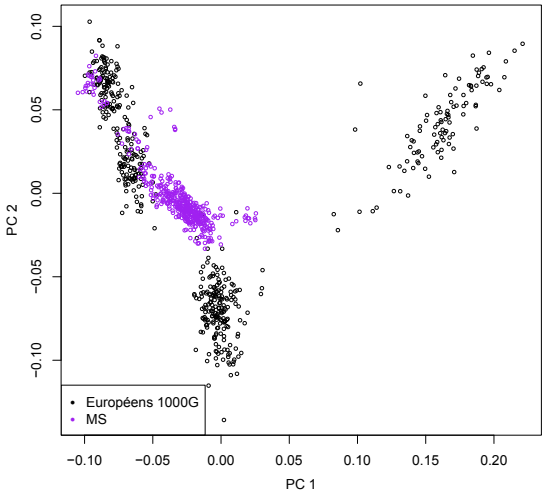


FIGURE Annexe 7.1 – Projection de nos individus sur les deux premières PCs des Européens de 1000 Genomes.

L'étude d'association du score de gravité de la Sclérose en Plaques

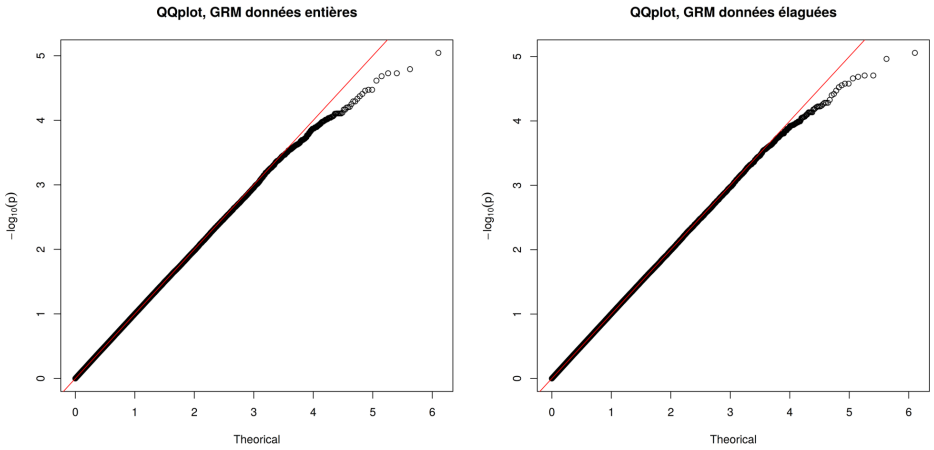


FIGURE Annexe 7.2 – QQ-plot des p-valeurs du test du score sous le modèle mixte pour l'association pour les comparer à une distribution du χ^2_1.

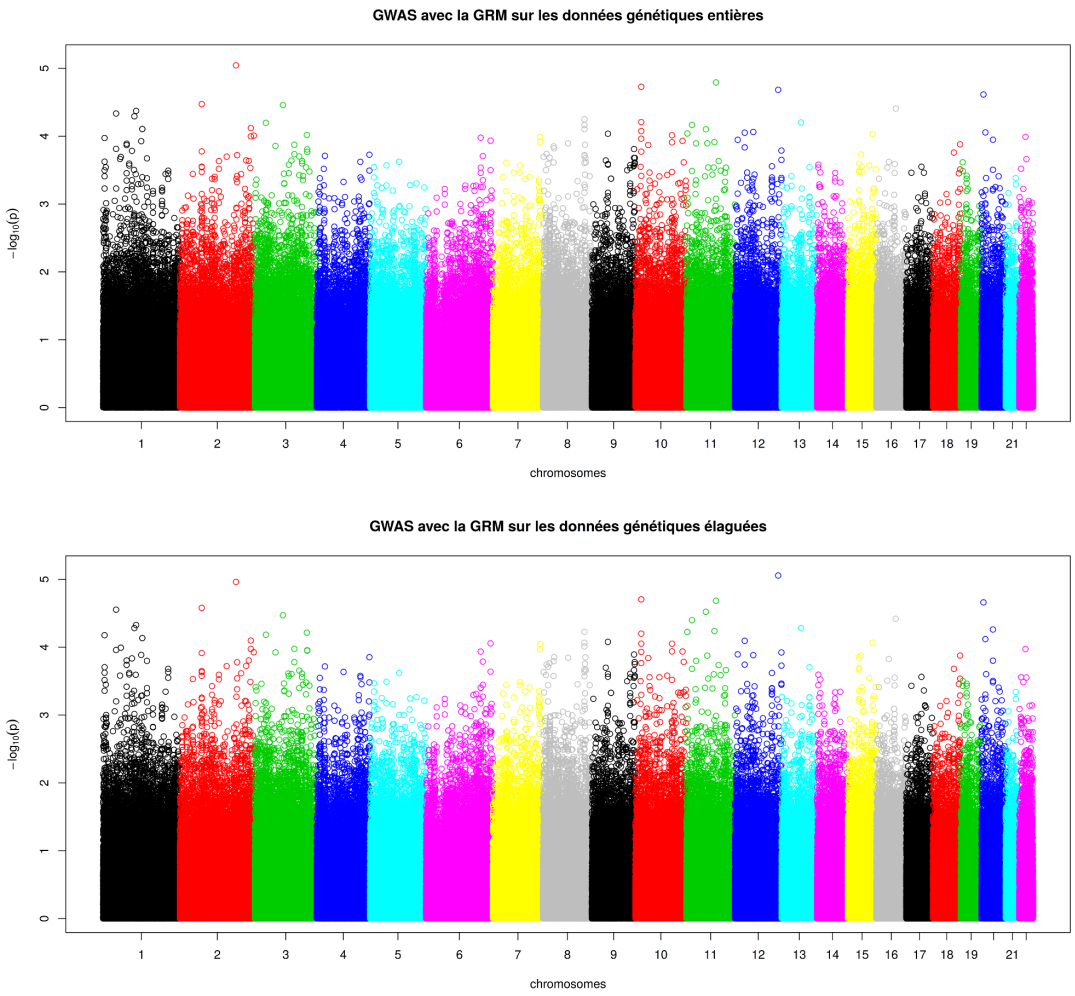


FIGURE Annexe 7.3 – P-valeurs du test du score sous le modèle mixte pour l'association. Les couleurs représentent les différents chromosomes.

L'étude d'association du statut de la Sclérose en Plaques

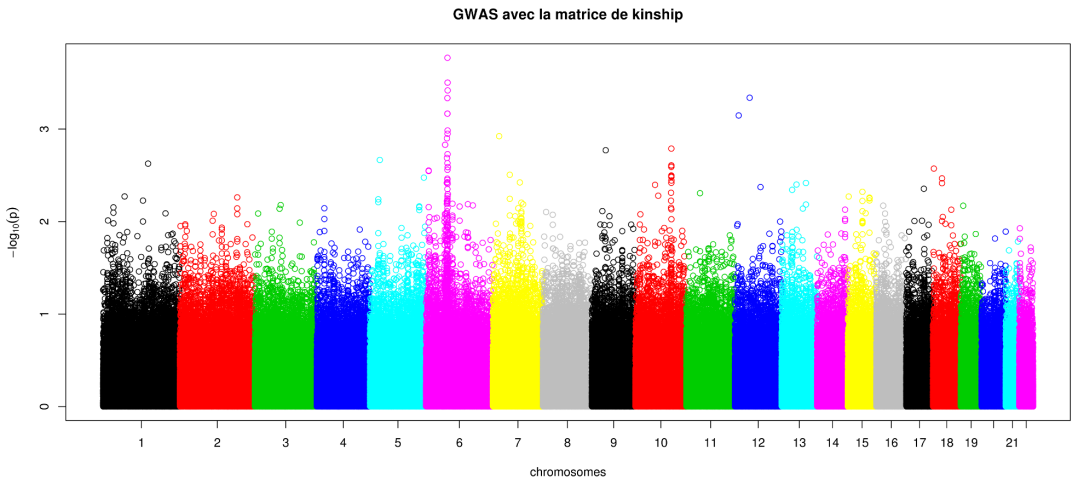


FIGURE Annexe 7.4 – P-valeurs du test du score sous le modèle mixte pour l'association. Les couleurs représentent les différents chromosomes.

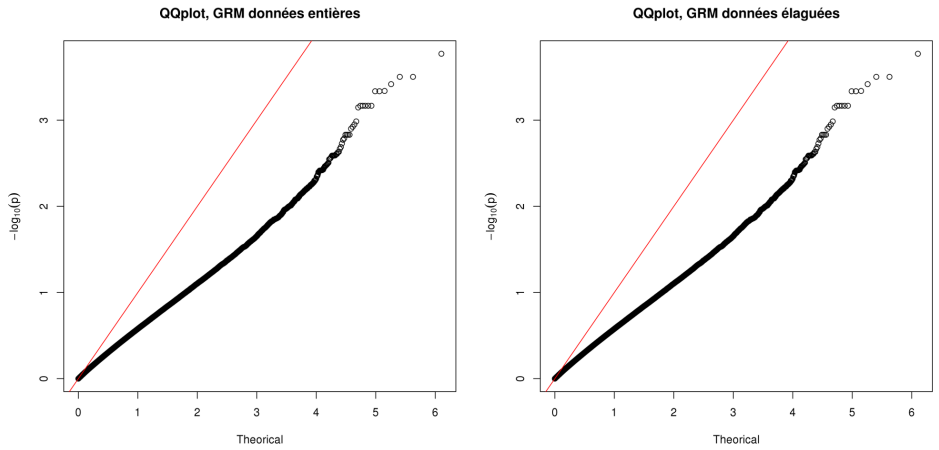


FIGURE Annexe 7.5 – QQ-plot des p-valeurs du test du score sous le modèle mixte pour l'association pour les comparer à une distribution du χ^2_1.

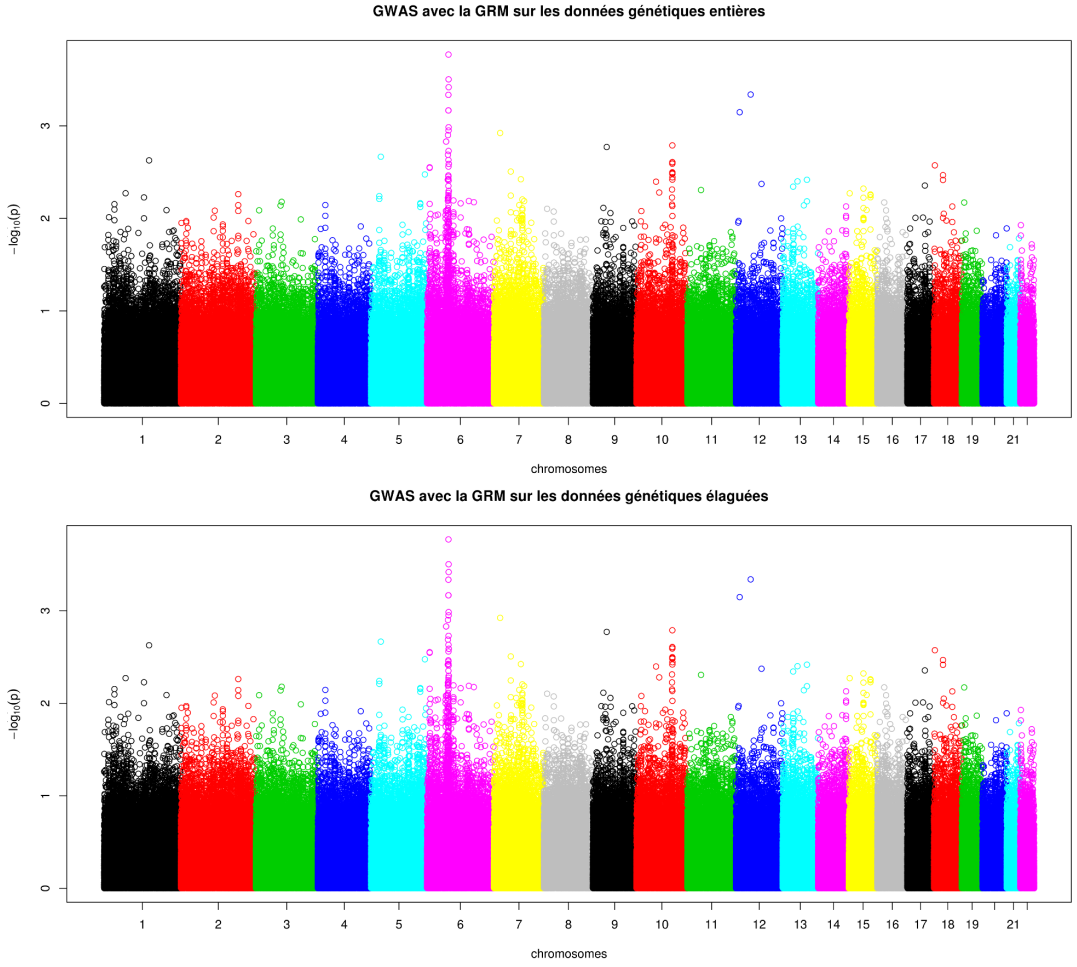


FIGURE Annexe 7.6 – P-valeurs du test du score sous le modèle mixte pour l'association. Les couleurs représentent les différents chromosomes.

Annexe 8 : Les articles

Article 1

European Journal of Human Genetics (2015) 23, 1357–1363

www.nature.com/ejhg

ARTICLE

Where is the causal variant? On the advantage of the family design over the case–control design in genetic association studies

Claire Dandine-Roulland*[,1,2] and Hervé Perdry[1,2]

Where is the causal variant?
On the advantage of the family design over the case-control design in genetic association studies

Claire Dandine-Roulland[1, 2] , Hervé Perdry[1, 2]

1 UMR-S 669, Université Paris-Sud 11, Villejuif, France

2 U669, INSERM, Villejuif, France

Running title: Where is the causal variant?

Corresponding author:

Claire Dandine-Roulland

Hôpital Paul Brousse

Bâtiment Inserm 15/16

16 avenue Paul-Vaillant-Couturier

F-94807 Villejuif

France

E-mail: claire.dandine-roulland@inserm.fr

Abstract

Many associated Single Nucleotide Polymorphisms (SNPs) have been identified by association studies for numerous diseases. However, the association between a SNP and a disease can result from a causal variant in linkage disequilibrium with the considered SNP. Assuming that the true causal variant is among the genotyped SNPs, other authors demonstrated that the power to discriminate between it and other SNPs in linkage disequilibrium is low. Here, we propose to take advantage of the information provided by family data to improve the inference on the causal variant: we exploit the linkage information provided by affected sib pairs to discriminate the causal variant from the associated SNPs. The family-based approach improves discrimination power requiring up to fives times less individuals than its case-control equivalent.

However, the main advantage of family design is the possibility to carry out the procedure one step further: the linkage information allows inference on causal variants which are not genotyped but in linkage disequilibrium with tag-SNPs displaying association, which is impossible with case-control design. By means of Bayesian methods, we estimate the linkage disequilibrium between the observed SNPs and an unobserved causal variant, as well as the allelic odds-ratio at the unobserved causal variant. The proposed procedure is illustrated on a Multiple Sclerosis family data set including genotypes of SNPs in *IL2RA*, confirming the advantage of using a family design to identify causal variants. The results of our method on this data suggest the existence of two distinct causal variants in this gene for the Multiple Sclerosis.

Keywords: Genetic Association Studies; Linkage Disequilibrium; Family Design; Bayesian Inference; Multiple Sclerosis.

Introduction

Association studies aim to identify variants associated with a disease, usually focusing on Single Nucleotide Polymorphisms (SNPs). They are able to detect the variants with modest effect which are implied in complex diseases, contrarily to linkage analysis[1]. In genome-wide association studies (GWAS), the considered variants are tag-SNPs, which capture most common SNPs of the genome through linkage disequilibrium (LD)[2]. However, the association between a SNP and a disease does not prove the causality link between the two: the association can result from a causal effect of the SNP itself or from the linkage disequilibrium with another causal variant. Consequently, a significant association signal indicates a set of correlated variants associated with the disease. Discriminating between the causal variant and variants in LD with it using case-control data was adressed by Udler et al.[3] Under the hypothesis that the causal variant is among the genotyped SNPs, the proposed method allows to select a minimal subset of potentially causal SNPs among disease associated variants. Family data convey more information than case-control data, and their use can improve the performance of this selection process; moreover, family data allow to address a limitation of the discrimination method with case-control data, which is that the causal variant is among genotyped SNP.

Here, we propose a method exploiting family data to select a minimal subset of associated SNPs, and to make inference on putative causal variants in LD with those SNPs. This method uses an association framework which takes advantage on the linkage information existing in affected sib-pairs data.[4] The first step is to select a minimal subset of potentially causal SNPs among disease associated variants; assuming that the causal SNP is among the genotyped SNPs, we compare the performance of this discrimination step with the method using case-control data.[3] The second step of the method addresses the situation where the causal variant is not directly genotyped, but is in LD with genotyped SNPs. In this situation, case-control data does not convey enough information to make the difference between a SNP in strong LD with the unobserved causal variant, and a truly causal variant. Using a sample of Affected Sib-Pairs (ASPs), the number of alleles shared Identical-By-Descent by the two affected siblings allows to make inference about the causal variant, to estimate its allelic frequencies, the allelic odds ratio (OR), and the LD between it and an observed SNP.

The advantage of the family method is illustrated on a sample of Multiple Sclerosis data. Multiple Sclerosis (MS) is a chronic autoimmune neurological disease of the central nervous system which affects about 1–2 per 1000 people in Europe and North America.[5] It is manifested by demyelination of nerve fibers in the brain, spinal cord and optic nerve. The disease is progressive and may lead to the loss of walking and eventually death. It is a multifactorial disease and has environmental and genetic factors. Several associations with genes involved in the immune response have been found. In particular, in the literature, there are associations with various genes in the *HLA* (Human Leukocyte Antigen) region, and various non-HLA genes, e.g. *CD58*, *IL2RA* and *IL7R*.[6] Our data set consists of french ASP and controls data from a previous study[7] and collected through REFGENSEP. It comprehends the genotypes of 26 SNPs in *IL2RA* on the chromosome 10. Several studies find association between MS and this gene in Caucasian populations.[8,9,10,11,12,13] The method using case-control data selects a minimal subset of 7 associated SNPs, which reduces to 3 SNPs when using family data. The second step shows that none of these SNPs is causal, and that the association signal is due to at least two different ungenotyped variants in the region.

Materials and Methods

In the first two paragraphs below we give an overview of the method proposed by Udler et al.[3] to identify causal variants in case-control association studies. The reader is referred to the original paper for details. In the third paragraph, we present the discrimination method that uses family data in two steps; selection of a subset of associated SNPs and inference about a putative causal variant not genotyped in the sample. In the last paragraph, we describe the Multiple Sclerosis dataset used to illustrate this approach.

Discrimination procedures

Consider n highly correlated variants in a genomic region. Under the hypothesis that one of these variants is causal, the aim is to select a subset of these variants that is likely to contain the causal variant. The method relies on Bayesian principles: if L_i for $i = 1,\ldots,n$ is the likelihood that the ith variant is the causal variant, the variants selected are those of index i such that

$$\frac{\max_j L_j}{L_i} > K,$$

or, equivalently,

$$2\ln(\max_j L_j) - 2\ln L_i > k,$$

where $k = \ln(K^2) = 2\ln K$. Following Udler et al.[3], we take $K = 100$ (i.e. $k = 2\ln 100 \approx 9.21$) which is interpreted as excluding variants with odds greater than 100:1.

For example, for two SNPs A and B, the SNP B is not retained if $2\ln L_A - 2\ln L_B > k$. As asymptotically

$$2\ln L_A - 2\ln L_B \approx Y_A^2 - Y_B^2$$

where Y_A^2 and Y_B^2 are association test statistics correspondings to the likelihoods L_A and L_B, such as the Armitage Trend Test statistic[14,15,16] for case-control design, and a score statistic adaptated to the family design,[4] this is equivalent to not retaining B is

$$Y_A^2 - Y_B^2 > k. \tag{1}$$

Both these association statistics Y are approximately standard normal, $Y \sim N(0,1)$, under the hypothesis of no association with the disease. Otherwise, it is approximately decentered normal: assuming that the causal variant for the disease is A with ψ its per-allele OR,

$$Y_A \sim N(\eta,1)$$

where η is a decentered parameter which will depend on the sample sizes, on the allele frequencies and on ψ. This parameter η also depends on the association statistic test used. Being in linkage disequilibrium (LD) with the variant A, the variant B is also associated with the disease. Then,

$$Y_B \sim N(r\eta,1) \quad \text{and} \quad \operatorname{cov}(Y_A, Y_B) \approx r,$$

where r is the correlation coefficient between the two variants, measuring the intensity of the LD. The distribution of the discrimination statistic (equation 1) is approximately

$$Y_A^2 - Y_B^2 \sim N(\eta^2(1-r^2), 4(1-r^2)(1+\eta^2))$$

(see details in section 1 of Supplementary Information).

Then, we can rely the power of discrimination $1-\beta$ with the decentered parameter η by

$$\eta^2(1-r^2)-z_{1-\beta}\sqrt{4(1+\eta^2)(1-r^2)}=k \tag{2}$$

where $z_{1-\beta}$ is the quantile of level $1-\beta$ of the standard normal distribution.

Discrimination with case-control data

Here we consider case-control data: the association statistic is the Armitage statistic.[14,15,16] Udler et al. show that in this case

$$\eta=\frac{\sqrt{2f_a f_A}(\psi-1)}{\sqrt{\frac{(f_a\psi+f_A)^2}{m}+\frac{\psi}{n}}} \tag{3}$$

where, m and n are the number of controls and cases, f_A and f_a the frequencies of the reference and alternative alleles, A and a, and ψ the per-allele OR of a. The demonstration is given in section 2 of Supplementary Information. Then, if we assume that the number of controls and cases are equal, the total sample size needed to achieve power $1-\beta$ is

$$n+m=\frac{\eta^2(f_A+f_a\psi)^2+\psi}{f_a f_A(\psi-1)^2} \tag{4}$$

where η^2 can be computed from β using equation 2.

Family design

Here, we propose a method in two steps. The first step is the selection of a subset of variants that is likely to contain the causal variant using the same discrimination procedure that in the case-control design, but using an association statistic designed for family data. We compute the power of discrimination of this procedure, assuming that the causal SNP is among the genotyped SNPs.

The second step uses the selected variants to make inference on causal variants in LD with them, relying on Bayesian principles. This step allows to retrieve information on a causal variant even if it is not genotyped.

First step: discrimination with family data

First, we use the same discrimination statistic (equation 1) based on a statistic Y which has been proposed for ASPs and controls.[4] The data considered include genotypes of controls,

genotypes of the index cases and the number of Identical-By-Descent (IBD) alleles in each sib-pair. Hereafter, we denote the three possible genotypes by the number of alternative alleles: 0, 1 and 2.

We denote n_{ki} the number of ASPs in which the index genotype is k and the number of IBD alleles is i, m_k the number of controls with genotype k, and n and m the total number of affected sib-pairs and controls. The association statistic is $Y = U / \sqrt{\sigma^2}$ where U is the score

$$U = \left(\sum_{k,i \in \{0,1,2\}} (2+i) n_{ki} \right) \hat{f} + \frac{1}{2} \sum_{k,i \in \{0,1,2\}} (2+i) k n_{ki}$$

with

$$\hat{f} = \frac{1}{2(m+n)} \left(\Sigma_{i \in \{0,1,2\}} n_{1i} + 2 \, \Sigma_{i \in \{0,1,2\}} n_{2i} + m_1 + 2m_2 \right)$$

the estimator of the alternative allele frequency, and

$$\sigma^2 = \frac{\frac{1}{4} \times (1 - \hat{f}) \hat{f} (19m + n - 1) n}{n + m}$$

the estimator of the variance of U under the hypothesis of no association. In absence of association, the distribution of Y is standard normal $Y \sim N(0,1)$.

We consider the causal variant A with allele frequencies f_A and f_a and OR ψ. The association statistic is decentered: $Y_A \sim N(\eta,1)$, where η is approximately

$$\eta = \frac{E(U_A)}{\sqrt{E(\hat{\sigma}_A^2)}}. \tag{5}$$

Formulas for $E(U_A)$ and $E(\hat{\sigma}_A^2)$ (depending on f_A, f_a, ψ and sample sizes n and m) are given in section 3 of Supplementary Information. Then, we can calculate power of discrimination for a given set of parameters, or total sample size needed to achieve a given power, with equation 2.

Second step: Bayesian inference on the causal variant

Secondly, we propose to use the SNPs selected by the discrimination step to retrieve information on the causal variants of the region. Let's assume that a variant B in linkage disequilibrium with the causal variant A is observed. In this case, we want to make inferences on A, in particular to estimate the LD between A and B, and the OR of A. This task is undoable with case-control data, as an OR for variant B can always be computed which explains fully the

observations under the hypothesis that B is the causal variant. However, the advantage of family data lies in the linkage information provided by the IBD state of the sib-pairs, which allows to discriminate between observations made directly at a causal variant A, and observations made at a variant B in LD with A.

In section 4 of Supplementary Information, we write a likelihood for the family data $L(\psi(f_a, f_b, d)$ (where ψ is the OR in A, f_a and f_b the alternative allele frequencies in A and B and d is the LD between A and B). In section 5 of Supplementary Information, we show that all parameters are identifiable, provided that $\psi > 1$ and $d \neq 0$.

This likelihood can be used to define the posterior distribution of parameters, from which we sample using Metropolis-Hastings algorithm[17] (cf section 6 of Supplementary Information for details). In particular, we can estimate the posterior distribution of the disequilibrium $r^2 = d^2 / (f_a(1-f_a)f_b(1-f_b))$. We also find simultaneous credibility regions for f_a and f_b, or for ψ and r^2, using the posterior joint density of these parameters as estimated from the values sampled by the Metropolis-Hastings algorithm.

Multiple Sclerosis data

These two methods of discrimination are illustrated on Multiple Sclerosis data described in full details in Babron et al.[7] This data include 26 tag-SNPs on the *IL2RA* gene for french families with at least one affected child collected through REFGENSEP. All affected people were reviewed by a board-certified neurologist and diagnosed according to Poser criteria.[18] All individuals signed informed consent in accordance with the European Union and Country Laws and the Helsinki Convention. The sample comprises 522 trio families (one affected with two living parents) and 101 multiplex families (at least two affected sibs).

The trio families are used to create pseudo-control genotypes consisting of the alleles untransmitted by the parents to their affected child. Pseudo-control genotypes are known to represent general population genotypes.[19] Affected sib-pairs are obtained from multiplex families, randomly selecting two affected sibs in each family. The IBD states are calculated using the software Merlin[20], which calculates the probability of each IBD state. Only affected sib-pairs for which one IBD state has probability higher than 0.8 are kept, assigning the IBD state with

probability exceeding 0.8 to them.

Overall, the dataset comprises 522 pseudo-controls and 82 affected sib-pairs with case index genotypes and IBD states. In addition to applying the family-based descrimination method on the dataset, we will also use the case-control method on the 82 index sibs as cases and the 522 pseudo-controls.

Results

Power of the family and case-control discrimination methods

The power of the two discrimination methods depends on the expression of the decentered parameter η (equation 2). In figure 1, we display η values for an OR ψ varying from 1 to 5. For all OR, the η parameter in a family design is higher than that of a case-control design.

The total sample size needed required to achieve 90 power to exclude variants at 100:1 odds assuming an equal number of controls and unrelated cases or ASPs for different values of alternative allele frequency, odds ratio ψ and linkage disequilibrium r^2 is reported in figure 2. For identical parameters, the family discrimination method needs a smaller sample size than the case-control method. For example, when $\psi = 3$, $r^2 = 0.9$ and $f_a = 0.1$, the case-control method needs the genotypes of 1500 controls and 1500 cases, whereas the family discrimination method needs only the genotypes of 300 controls and 300 sib-pairs (genotype of the index case and IBD state, which can be obtained with a low density genotyping of the second sib).

Inference on a causal variant with family data

Using Metropolis-Hastings algorithm, we simulate data composed of 1000 ASPs and 1000 controls. The theoretical distribution used for these simulations is described in section 4 of Supplementary Information. The posterior distribution of f_a, f_b, ψ, and r^2, obtained from 10^7 distribution samples, are displayed in figures 3 and 4. Each shade of gray represents the credibility region for one level. The lightest gray corresponds to all sampled values.

In figure 3, the data are simulated under a model with total linkage disequilibrium ($r^2 = 1$), alternative allele frequencies $f_a = f_b = 0.2$, and an OR $\psi = 2$ for the causal variant A. The 95 credibility regions of f_b, f_a, ψ, and r^2 are approximately $[0.17, 0.215]$, $[0.08, 0.29]$, $[1.8, 3]$, and $[0.3, 1]$ respectively. They contain the true values of parameters, and the mode of

the distribution is near to the true values. Note that the allele frequency is best estimated at the variant which is directly observed, which corresponds to a certain amount of uncertainty on r^2.

In figure 4, the data are simulated with $r^2 = 0.8$, $f_a = 0.435$, $f_b = 0.448$, and $\psi = 3$. The 95 credibility regions of f_b, f_a, ψ and r^2 are $[0.415, 0.47]$, $[0.3, 0.55]$, $[2.5, 5]$, and $[0.5, 1]$ respectively. Again, they contain the true values of parameters. Interestingly, although the causal variant is not directly observed, some inference of its characteristics is possible.

Application to Multiple Sclerosis data

Discrimination methods

The values of association statistics and p-values for the two discrimination methods at all SNPs are displayed in the table 1. For the SNPs with the smallest p-values (SNPs 1 – 4, 21 and 24), the family based p-values are lower than the case-control ones. However, the use of family data does not decrease p-value for all SNPs: for example, the SNP reported in the literature[21,22,23,24,25] as associated with Multiple Sclerosis, rs2104286 (SNP 23 in our numbering), is not associated using case-control data, and adding the IBD information does not decrease its p-value. After Bonferroni correction, the association is significant association only with rs3118470 (SNP 24) for both case-control and family designs, and, with rs12359875 (SNP 1) for family design only.

To apply discrimination methods on these data, we compute the difference of association statistics between the most associated SNP, i.e. SNP 24, and others (table 1). Comparing these values with the threshold $k = 9.210$, we select the set of SNPs 1, 2, 3, 4, 20, 21 and 24 as likely to contain the causal variant, using the case-control discrimination method, while the selected set contains only SNPs 1, 4 and 24, for the family discrimination method.

Metropolis-Hastings on SNPs 24 and 1

Applying the Metropolis-Hastings algorithm on the most associated SNP, i.e. SNP 24, the posterior distributions are displayed in figure 5. The frequency f_b corresponds to the SNP 24 and f_a to the hypothetical causal variant. The 95 credibility region of f_b and OR ψ are $[0.24, 0.31]$, and $[1.4, 3.2]$ respectively. The linkage disequilibrium r^2 is not well estimated, since its 95 credibility region containing almost all possible values. Finally, for the parameter f_a, the disease allele frequency is bimodal, with two modes near 0.3 and 0.8.

Also applying the Metropolis-Hastings algorithm on the second associated SNP, i.e. SNP 1, the posterior distributions are displayed in figure 6. Graphically, the 95 credibility regions of f_b, f_a, ψ, and r^2 are approximately $[0.68, 0.76]$, $[0.6, 0.95]$, $[1.5, 5]$, and $[0.1, 1]$ respectively. The mode of f_a is around 0.8.

We have also applied the Metropolis-Hastings algorithm on rs9663421 (SNP 4) which is in the subset of SNPs selected by the family method. The results are similar with those obtained for SNP 1. This is consistent with the observed LD between the SNPs 1 and 4 in our data ($r^2 = 0.85$).

Discussion

Nowadays, research on complex diseases focuses on massive case-control designs neglecting family designs. However, the joint use of linkage and association information in families allows efficient designs for complex diseases. Using linkage information in association studies results not only in a gain of power in association testing, but also in an increased ability to estimate the risk conferred by the allelic variants, as illustrated in previous papers on Rheumatoid Arthritis[26,27] and Multiple Sclerosis.[7] The MASC method[28] was developed to exploit all information in family data. The association test from Perdry et al.[4] is built on the same idea.

In this paper, we have shown that sib-pairs provide a gain of power to discriminate between several SNPs associated with a disease. For example, with an OR of 3, $r^2 = 0.9$ and 0.1 alternative allele frequency, the family method needs five times less individuals than the case-control method to achieve similar power. This illustrates well the gain of information provided by family data, as the sib-pairs test uses simultaneously association information comparing control and index cases and linkage information through the IBD. Note that in many cases, the IBD information is already available from previous linkage studies that have been performed using the same sib-pair sample. If it is not available, it can be obtained through low density genotyping which has a negligible cost as compared to the high density genotyping of the index cases.

Moreover, Udler et al.'s method for case-control data assumes that the causal variant is genotyped, which is unlikely to be true when using tag-SNPs. Imputation methods, that have been widely used in GWAS, can help to reach a fine enough mapping scale. Nevertheless, we have shown that with family data, thank to the IBD information, we can capture information on the unobserved causal variants through the linked observed variants. This was first done formally, by proving the identifiability of the parameters (section 5 of Supplementary Information). This

allows in theory to assess whether the observed variant is the causal variant, or if it is only in linkage disequilibrium with the causal variant. Sampling from the posterior distribution of the different parameters allows to obtain credibility regions for allele frequencies, per-allele risk and linkage disequilibrium between observed and causal variants. We showed on simulated data that these credibility regions are reasonably accurate.

These methods were illustrated on a real data set, consisting of *IL2RA* genotypes on Multiple Sclerosis cases and controls, which was first considered in Babron et al.[7] Both discrimination methods provide subsets of SNPs which likely contain the causal variants. However, the subset obtained using case-control data contains seven SNPs, whereas the subset obtained using the family method contains only three SNPs. This reduced size illustrates the increase of discrimination power. Additionally, the family data was additionally used to estimate the risk allele frequency and the OR of a putative causal variant in *IL2RA*: when using rs12359875 (SNP 1), the method infers a causal variant with a risk allele frequency around 0.8, likely different from SNP 1; the value of the LD between SNP 1 and the causal variant and the allelic OR are not well identified. When using rs3118470 (SNP 24), the posterior distribution of the risk allele frequency of the causal variant is bimodal, with one mode around 0.8 and another around 0.3. This pleads in favour of the existence of a second causal variant with a risk allele frequency near 0.3, in LD with SNP 24 but not with SNP 1, while the first causal variant with a risk allele frequency near 0.8 is in LD with both SNPs. In the previous study,[7] the association signal in *IL2RA* was best captured by an haplotype of rs2256774 (SNP 22) and rs3118470 (SNP 24). Besides that we agree on the fact that the association signal is not due to a single SNP in the region, it is difficult to compare our results with their results, as our approach does not allow to consider several markers at the same time. Also, we considered only a multiplicative risk model, which was not the case in the previous study.[7]

Our approaches could be extended in these directions: considering haplotypes instead of isolated SNPs, droping the multiplicative risk hypothesis, and allowing for multiple disease alleles. If this can improve the disease model inference, a compromise has however to be found between the complexity of the model considered, and the amount of available information. Considering larger nuclear families or even multiplex families can be a solution to improve the precision of the inference. An other interesting strategy would be to use the difference of LD pattern between cases and controls in the vicinity of the causal variant[29] at the same time as the

IBD information.

Identifying variants helps both to improve disease risk prediction, and to uncover biological mechanisms involved in human diseases. A better statistical modeling of the effect of the variants in an associated genomic region is a crucial step on this way. In this regard, family design should not be neglected.

Software

The proposed method is accessible in the R package ASPBay available on the Comprehensive R Archive Network (CRAN).

Acknowledgments

We wish to thank gratefully Françoise Clerget-Darpoux for many fruitful discussions. We thank all the patients who generously participated in and the physicians constituting the REFGENSEP (Réseau français de la génétique de la sclérose en plaques). We also thank all the authors of Babron et al.[7] for their help on example data set, in particular Marie-Claude Babron. We also thank Rémi Kazma for a careful reading of the manuscript. Hervé Perdry was partly funded by a grant from ARSEP fondation (Fondation pour l'Aide à la Recherche sur la Sclérose en Plaques).

Conflict of interest

The authors declare no conflict of interest.

Supplementary information

Supplementary information accompanies this paper on European Journal of Human Genetics website.

References

1. Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**(5281): 1516–1517.
2. Manolio T: Genomewide association studies and assessment of the risk of disease. *N Engl J Med* 2010; **363**(2).

3. Udler MS, Tyrer J, Easton DF: Evaluating the power to discriminate between highly correlated snps in genetic association studies. *Genet. Epidemiol.* 2010; **34**(5): 463–468.
4. Perdry H, Müller-Myhsok B, Clerget-Darpoux F: Using affected sib-pairs to uncover rare disease variants. *Hum Hered* 2012; **74**(3-4): 129–141.
5. Compston A, Confavreux C, Lassmann H, *et al.*: McAlpine's multiple sclerosis. Churchill Livingstone Elsevier, 2005.
6. Goris A, Pauwels I, Dubois B: Progress in multiple sclerosis genetics. *Curr Genomics* 2012; **13**(8): 646.
7. Babron MC, Perdry H, Handel AE, *et al.*: Determination of the real effect of genes identified in GWAS: the example of IL2RA in multiple sclerosis. *Am J Hum Genet* 2011; **20**(3): 321–325.
8. Matesanz F, Caro-Maldonado A, Fedetz M, *et al.*: IL2RA/CD2 polymorphisms contribute to multiple sclerosis susceptibility. *J. Neurol.* 2007; **254**(5): 682–684.
9. Ramagopalan SV, Anderson C, Sadovnick AD, Ebers GC, Matesanz F, *et al.*: Genomewide study of multiple sclerosis. *N Engl J Med* 2007; **357**(21): 2199–2200.
10. Hafler DA, Compston A, Sawcer S, *et al.*: Risk alleles for multiple sclerosis identified by a genomewide study. *N Engl J Med* 2007; **357**(9): 851–862.
11. Ramagopalan SV, Anderson C, Sadovnick AD, Ebers GC: Genomewide study of multiple sclerosis. *N Engl J Med* 2007; **357**(21): 2199–2200.
12. Rubio J, Stankovich J, Field J, *et al.*: Replication of KIAA0350, IL2RA, RPL5 and CD58 as multiple sclerosis susceptibility genes in australians. *Genes Immun* 2008; **9**(7): 624–630.
13. Weber F, Fontaine B, Cournu-Rebeix I, *et al.*: IL2RA and IL7RA genes confer susceptibility for multiple sclerosis in two independent european populations. *Genes Immun* 2008; **9**(3): 259–263.
14. Armitage P: Tests for linear trends in proportions and frequencies. *Biometrics* 1955; **11**(3): 375–386.
15. Sasieni PD: From genotypes to genes: doubling the sample size. *Biometrics* 1997; **53**(4): 1253–1261.
16. Slager S, Schaid D: Case-control studies of genetic markers: Power and sample size approximations for armitage's test for trend. *Hum Hered* 2001; **52**(3): 149–153.
17. Robert CP, Casella G: Monte Carlo statistical methods, volume 319. Springer Verlag, 2004.
18. Poser CM, Paty DW, Scheinberg L, *et al.*: New diagnostic criteria for multiple sclerosis: guidelines for research protocols. *Ann Neurol.* 1983; **13**(3): 227–231.

19. Thomson G: Mapping disease genes: family-based association studies. *Am J Hum Genets* 1995; **57**(2): 487.
20. Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* 2001; **30**(1): 97–101.
21. Tröster AI: Refining genetic associations in multiple sclerosis. *Neurology* 2006; **66**: 1830–36.
22. Alcina A, Fedetz M, Ndagire D, *et al.*: IL2RA/CD25 gene polymorphisms: uneven association with multiple sclerosis (ms) and type 1 diabetes (t1d). *PLoS One* 2009; **4**(1): e4137.
23. Dendrou CA, Plagnol V, Fung E, *et al.*: Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource. *Nature Genet.* 2009; **41**(9): 1011–1015.
24. Maier LM, Anderson DE, Severson CA, *et al.*: Soluble IL2RA levels in multiple sclerosis subjects and the effect of soluble IL-2RA on immune responses. *J. Immunol.* 2009; **182**(3): 1541–1547.
25. Maier LM, Lowe CE, Cooper J, *et al.*: IL2RA genetic heterogeneity in multiple sclerosis and type 1 diabetes susceptibility and soluble interleukin-2 receptor production. *PLoS Genet.* 2009; **5**(1): e1000322.
26. Binet J, Auquier A, Dighiero G, *et al.*: A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis. *Cancer* 1981; **48**(1): 198–206.
27. Bourgey M, Perdry H, Clerget-Darpoux F: Modeling the effect of PTPN22 in rheumatoid arthritis. In BMC proceedings, volume 1. BioMed Central Ltd, 2007; p. S37.
28. Clerget-Darpoux F, Babron M, Prum B, Lathrop G, Deschamps I, Hors J: A new method to test genetic models in HLA associated diseases: the MASC method. *Ann Hum Genet.* 1988; **52**(3): 247–258.
29. Bochdanovits Z, Simón-Sánchez J, Jonker M, Hoogendijk WJ, van der Vaart A, Heutink P: Accurate prediction of a minimal region around a genetic association signal that contains the causal variant. *Eur. J. Hum. Genet.* 2013; **22**(2): 238–242.

Figures

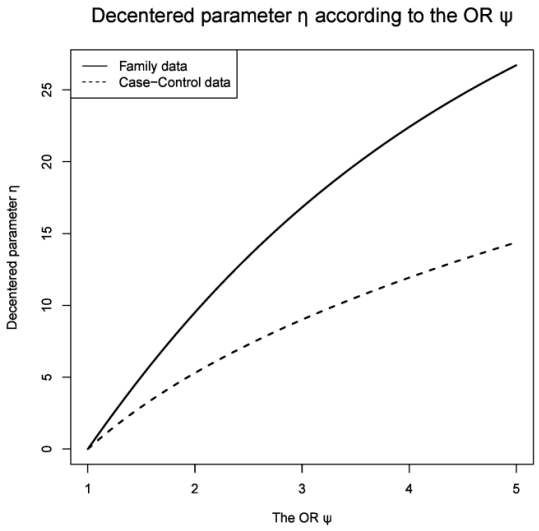


Figure 1: The parameter η according to the OR ψ for 0.1 alternative allele frequency, $r^2 = 0.8$ and 500 cases and 500 controls.

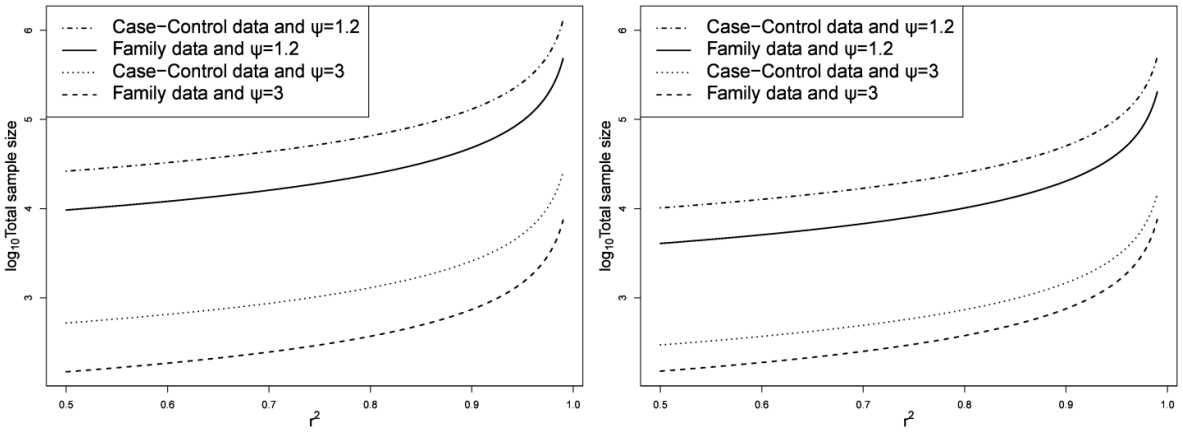


Figure 2: The sample size needed to achieve 90 power to exclude variants at 100:1 odds is plotted as function of r^2, for various values of ψ and for (a) $f_a = 0.1$ and (b) $f_a = 0.5$. The number of cases and controls are assumed to be equivalent.

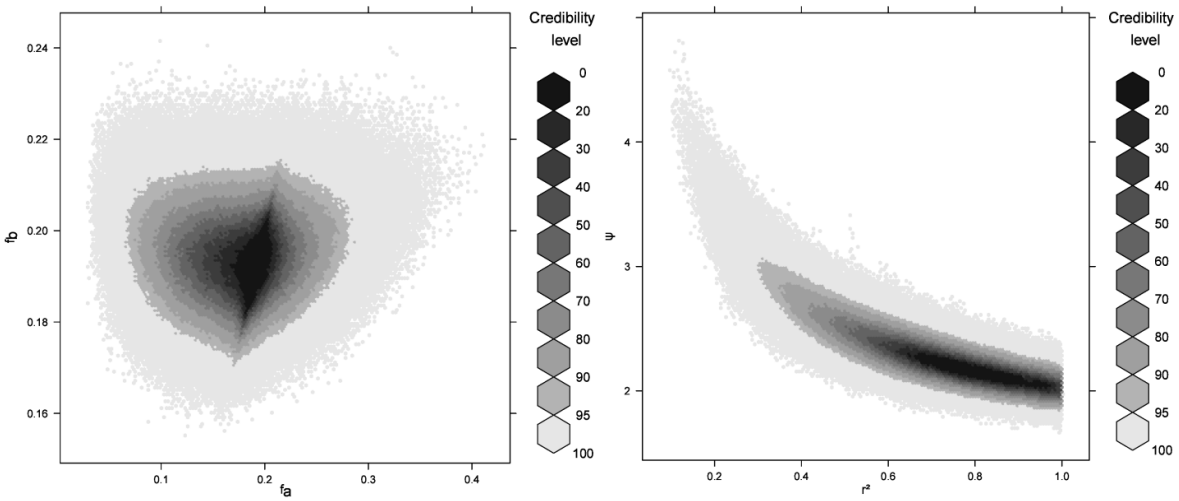


Figure 3: Posterior distributions with Metropolis-Hastings for a simulated sample of 1000 sib-pairs and 1000 controls. The parameters used for simulation are $r^2 = 1$, $f_a = f_b = 0.2$ and $\psi = 2$.

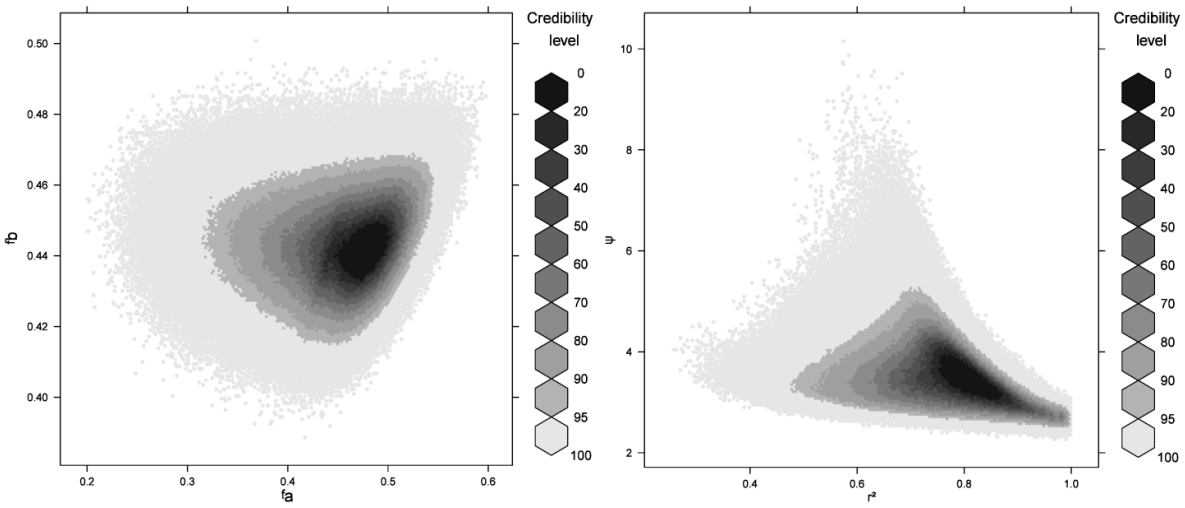


Figure 4: Posterior distributions with Metropolis-Hastings for a simulated sample of 1000 sib-pairs and 1000 controls. The parameters used for simulation are $r^2 = 0.8$, $f_a = 0.435$, $f_b = 0.448$ and $\psi = 3$.

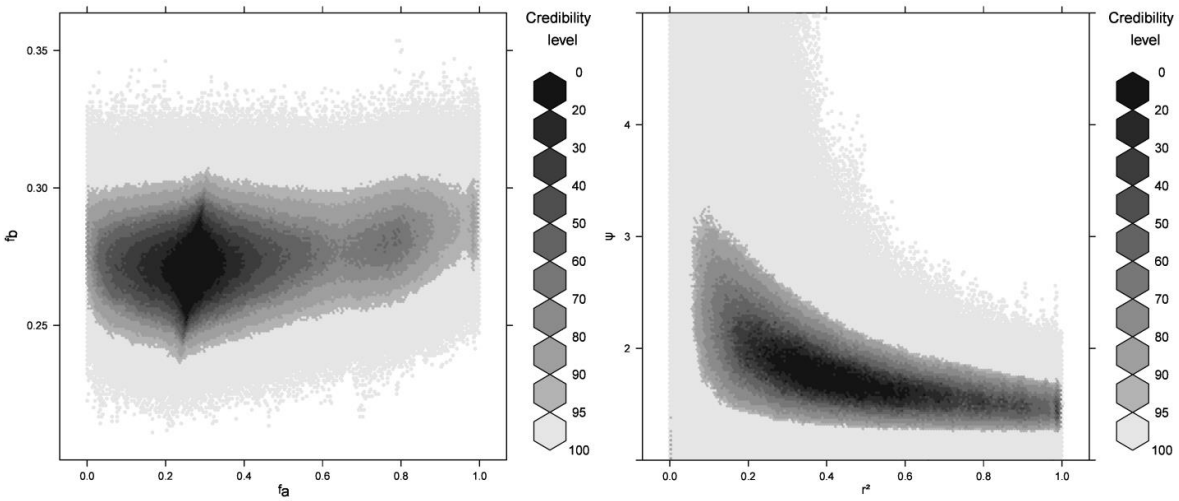


Figure 5: Posterior distributions with Metropolis-Hastings using SNP 24 (rs3118470).

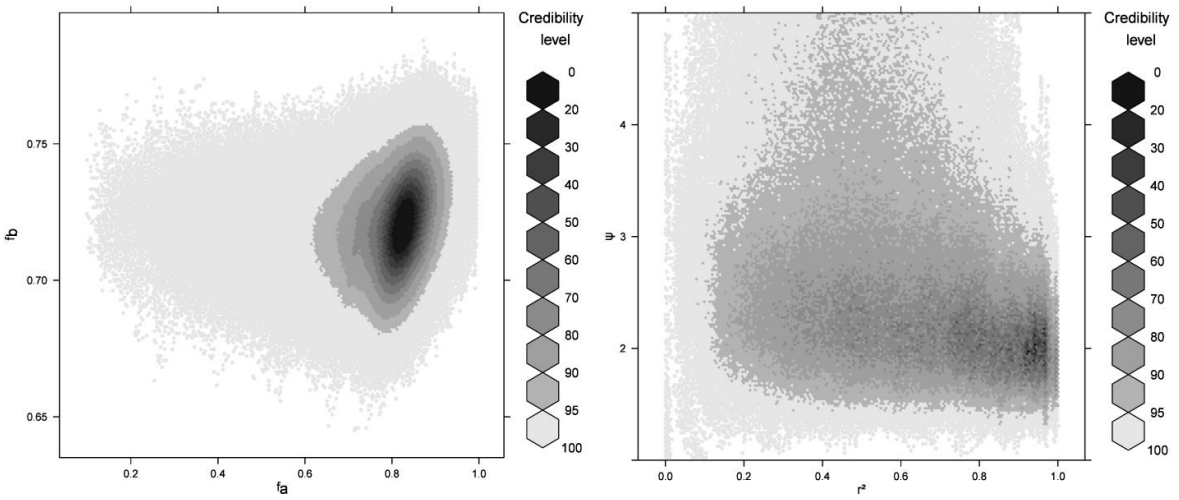


Figure 6: Posterior distributions with Metropolis-Hastings using SNP 1 (rs12359875).

Tables

Table 1: Association test statistics and ρ-values; discrimination statistics of all SNPs with SNP 24

SNP name		Association test statistics and *p* -values				Discrimination statistic with SNP 24	
		Case-control	*p* -value	Family	*p* -value	Case-control	Family
SNP 1	rs12359875	8.75	0.0031	**9.63**	**0.0019**	**2.617**	**5.383**
SNP 2	rs12722605	4.04	0.044	4.88	0.027	**7.323**	10.133
SNP 3	rs12244380	2.25	0.13	3.11	0.078	**9.117**	11.910
SNP 4	rs9663421	5.28	0.022	7.62	0.0058	**6.086**	**7.399**
SNP 5	rs12722596	1.00	0.32	1.70	0.19	10.364	13.315
SNP 6	rs2386841	0.04	0.84	0.01	0.93	11.324	15.009
SNP 7	rs12722588	1.33	0.25	0.64	0.42	10.031	14.374
SNP 8	rs2076846	0.10	0.75	0.21	0.65	11.264	14.805
SNP 9	rs12722561	0.13	0.72	0.27	0.60	11.235	14.743
SNP 10	rs6602392	0.20	0.65	0.26	0.61	11.163	14.757
SNP 11	rs7072398	0.29	0.59	0.53	0.47	11.074	14.484
SNP 12	rs11256456	0.05	0.82	0.002	0.96	11.316	15.014
SNP 13	rs11256457	0.01	0.92	0.17	0.68	11.356	14.843
SNP 14	rs4749924	1.97	0.16	2.11	0.15	9.398	12.904
SNP 15	rs11598648	0.02	0.88	0.31	0.58	11.342	14.702
SNP 16	rs11256497	0.04	0.85	0.13	0.71	11.325	14.883
SNP 17	rs791587	2.03	0.15	3.05	0.081	9.339	11.967
SNP 18	rs791589	0.38	0.54	0.16	0.69	10.989	14.860
SNP 19	rs791590	0.37	0.54	0.61	0.44	10.995	14.409
SNP 20	rs10905669	2.31	0.13	2.39	0.12	**9.051**	12.629
SNP 21	rs2476491	3.72	0.054	4.13	0.042	**7.642**	10.889
SNP 22	rs2256774	0.85	0.36	1.48	0.22	10.513	13.534
SNP 23	rs2104286	1.26	0.26	0.35	0.55	10.110	14.663
SNP 24	rs3118470	**11.37**	**0.00075**	**15.02**	**0.00011**	**0.000**	**0.000**
SNP 25	rs12722489	0.33	0.57	0.44	0.51	11.035	14.581
SNP 26	rs12722486	0.10	0.74	0.004	0.95	11.259	15.013

Article 2

Human Heredity

Original Paper

Hum Hered 2015;80:196–206
DOI: 10.1159/000447634

Published online: September 1, 2016

The Use of the Linear Mixed Model in Human Genetics

Claire Dandine-Roulland Hervé Perdry

CESP, Inserm, Université Paris-Sud, Université Paris-Saclay, Villejuif, France

The Use of the Linear Mixed Model in Human Genetics

Claire Dandine-Roulland and Hervé Perdry*

CESP, Inserm, Univ. Paris-Sud, Université Paris-Saclay, Villejuif, France

*Corresponding author. CESP, Hôpital Paul-Brousse, Avenue Paul-Vaillant-Couturier, F-94807 Villejuif. herve.perdry@u-psud.fr

Abstract

In the beginning of this paper, we give a short but detailed review of the methods used to deal with Linear Mixed Models (Restricted Likelihood, AIREML algorithm, Best Linear Unbiased Predictors, etc), with a few original points. The second part describes three common applications of the Linear Mixed Model in contemporary human genetics: association testing (pathways analysis or rare variants association tests), genomic heritability estimates, and correction for population stratification in Genome Wide Association Studies. We also consider the performance of Best Linear Unbiased Predictors for prediction in this context, through a simulation study for rare variants in a short genomic region, and through a short theoretical development for genome-wide data. For each of these applications, we discuss the relevance and the impact of modeling genetic effects as random effects.

Keywords: Linear Mixed Models, Rare Variant Association Testing, Heritability, Population Stratification, Prediction

Introduction

One important incentive for the Linear Mixed Model development was the genetic evaluation of dairy cattle, in which the breeding value of bulls is estimated from their daughter cows milk yield [1, 2]. Widely used in animal genetics, these models have proved useful in human genetics in the linkage era [3]; they also proved useful in epidemiology, where modeling, for example, center effects as random effects is a powerful way to deal with heterogeneity among centers of ascertainment. It is however only in the last few years that these models have been widely used in human genetics, for association studies and for heritability estimates.

In this paper, we first give a short but detailed review of the methods used to deal with Linear Mixed Models (Restricted Likelihood, AIREML algorithm, Best Linear Unbiased Predictors, etc). For this exposition, we mainly used Searle, Casella and McCulloch classical textbook [4], but some points are original at least in their presentation (the interpretation of the EM algorithm as a gradient ascent and the derivation of the Best Linear Unbiased Predictors formulas). Some points are only sketched in the main text, but full details are given in Supplementary Material. This material is classic, but necessary for a good understanding the applications in human genetics, and we think that a comprehensive and short presentation of these methods can be useful to readers yet unfamiliar with some aspects of the theory.

In the sequel, we describe three common applications of the Linear Mixed Model in contemporary human genetics: association testing (pathways analysis or rare variants association tests), genomic heritability estimates, and correction for population stratification in Genome Wide Association Studies. For each of these applications, we discuss the relevance of modeling the genetic effects as random effects. We also consider the performances of the Linear Mixed Model for prediction purposes in two cases: predicting the contribution of the rare variants harbored in a genomic region to a quantitative phenotype, and predicting a quantitative phenotype from genome wide data.

In the discussion, we discuss briefly the epistemological status of the Linear Mixed Model in human genetics.

The Linear Mixed Model

Notations and model

Consider n individuals; for each of these individuals, a trait or phenotype Y_i is measured, together with some covariates $X_{i1}, \ldots, X_{ip}$ that may influence the value of Y_i. The model includes effects of some of these covariates as parameters $\beta_1, \ldots, \beta_p$, called *fixed effects*. In our applications, the covariates with fixed effect will be clinical or environmental covariates (age, sex, exposure...), or in some cases the genotypes of a SNP potentially associated with the phenotype. The effects of a second set of covariates denoted $Z_{i1}, \ldots, Z_{iq}$ (in our applications, the genotypes of the individuals in q loci) are assumed to be drawn independently from a normal distribution: these are the *random effects* $u_1, \ldots, u_q \sim \mathcal{N}(0, \tau)$. The variance τ of this distribution is a parameter of the model; all the random effects $u_1, \ldots, u_q$ are modeled by this single parameter. It is important to note that the random effects are independent of the index i of the individual; one can consider that they are drawn at the beginning of the random experiment. The relationship between Y and the covariates is linear:

$$Y_i = X_{i1}\beta_1 + \cdots + X_{ip}\beta_p + Z_{i1}u_1 + \cdots + Z_{iq}u_q + e_i$$

for all $i = 1, \ldots, n$, where the residual error terms $e_1, \ldots, e_n$ are drawn independently from a normal distribution $\mathcal{N}(0, \sigma^2)$.

This model is best written under matrix form:

$$Y = X\beta + Zu + e, \tag{1}$$

with $Y = (Y_1, \ldots, Y_n)'$, X the $n \times p$ matrix of covariates with fixed effects $\beta = (\beta_1, \ldots, \beta_p)'$, Z the $n \times q$ matrix of covariates with random effects $u = (u_1, \ldots, u_q)' \sim \mathcal{N}(0, \tau I_q)$, and the residual error vector $e = (e_1, \ldots, e_n)' \sim \mathcal{N}(0, \sigma^2 I_n)$. The model can easily be extended to handle several matrices $Z_1, Z_2, \ldots$ with random effects of variances $\tau_1, \tau_2, \ldots$. For the sake of simplicity, we won't consider this case in the main text, but the interested reader is referred to the supplementary material.

An equivalent way to write the model is

$$Y = X\beta + \omega + e, \tag{2}$$

where $\omega = Zu$; thus, $\omega \sim \mathcal{N}(0, \tau K)$ with $K = ZZ'$. The vector Y is drawn from a multivariate normal distribution with expected value $X\beta$ and variance $V = V(\tau, \sigma^2) = \tau K + \sigma^2 I_n$. In the sequel, we will simply denote this matrix V, omitting the parameters τ and σ^2.

Maximum (Restricted) Likelihood Estimates

The model parameters to be estimated are β, τ and σ^2. We will see later that the random effects can also be retrieved.

The Likelihood

The Likelihood, or more precisely the Log-Likelihood, follows from the density of the multivariate normal distribution $\mathcal{N}(X\beta, V)$, evaluated in Y. It is

$$\ell(\beta, \tau, \sigma^2) = -\frac{1}{2}\log|V| - \frac{1}{2}(Y - X\beta)'V^{-1}(Y - X\beta). \tag{3}$$

There is no closed form for the Maximum Likelihood Estimates (MLEs). However, it comes readily from the gradient in β that for fixed values of σ and τ, the Likelihood is maximal for

$$\widehat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y. \tag{4}$$

We can use this to derive a Profile Likelihood for τ and σ^2, simply by by plugging $\widehat{\beta}$ in the Likelihood. One obtains

$$\ell^{pr}(\tau, \sigma^2) = \ell(\widehat{\beta}, \tau, \sigma^2) = -\frac{1}{2}\log|V| - \frac{1}{2}Y'PY \tag{5}$$

where

$$P = V^{-1} - V^{-1}X\left(X'V^{-1}X\right)^{-1}X'V^{-1}, \tag{6}$$

so that $Y - X\widehat{\beta} = VPY$. From the definition of P, one sees that $PVP = P$, which is used for the derivation of (5). The matrix P will play a important role in the sequel.

MLEs of τ and σ^2 can be obtained by maximizing either the Likelihood or the Profile Likelihood. However, this estimates are biased, because the expected value $X\beta$ of Y is estimated together with the variance components τ and σ^2; this is the same phenomenon that produces the bias in the MLE of the variance of a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with unknown parameters μ and σ^2 (cf Supplementary Information 3 for details).

The preferred approach is to use the Restricted Likelihood, which is obtained by projecting the vector Y on a subspace of $\mathbb{R}^n$ chosen such that the expected value of the projected vector is null. The expected value being known, the variance components are estimated without bias.

The Restricted Likelihood

The reader is referred to Supplementary Information 3 for full details on the material presented in this section. To define the Restricted Likelihood, we first pick a matrix of contrasts $C \in \mathbb{R}^{(n-p)\times n}$ such that $CX = 0$ and $CC' = I_{n-p}$ (the lines of C are an orthonormal basis of the vector space orthogonal to the space generated by the columns of X).

The vector CY follows a Gaussian distribution with mean $CX\beta = 0$ and variance $CVC' = \tau CKC' + \sigma^2 I_{n-p}$. The Restricted Likelihood is obtained by plugging these in equation (3), which becomes

$$\ell^{re}(\tau, \sigma^2) = -\frac{1}{2}\log|CVC'| - \frac{1}{2}Y'C'(CVC')^{-1}CY.$$

The crucial result is that this likelihood doesn't depend of the choice made for C: for all such contrast matrix, we have $C'(CVC')^{-1}C = P$ with P as in (6), and

$$\ell^{\mathrm{re}}(\tau, \sigma^2) = -\frac{1}{2}\log|V| - \frac{1}{2}\log\left|X'V^{-1}X\right| - \frac{1}{2}Y'PY. \tag{7}$$

Note that this expression bears some intriguing similarity with the Profile Likelihood (5). The estimates of τ and σ^2 obtained by maximization of the Restricted Likelihood are called Restricted Maximum Likelihood Estimates (REML).

Iterative algorithms for Likelihood maximization

The maximization of the Restricted Likelihood can be achieved with several methods. The EM algorithm is particularly appealing, as the constraints $\tau, \sigma^2 > 0$ are verified at each step.

We show in Supplementary Information 5 that it can be interpreted as a gradient ascent, as it each iterated step can be written

$$\tau_{r+1} = \tau_r + \left(\frac{2\tau_r^2}{r_K}\right)\frac{\partial \ell^{\text{re}}}{\partial \tau}\left(\tau_r, \sigma_r^2\right)$$
$$\sigma_{r+1}^2 = \sigma_r^2 + \frac{2\left(\sigma_r^2\right)^2}{n}\frac{\partial \ell^{\text{res}}}{\partial \sigma^2}\left(\tau_r, \sigma_r^2\right).$$

The convergence of this algorithm is however very slow, but it can be useful to perform a few steps, to get closer to the maximum before using a second-order method.

The two classical second-order methods for likelihood optimization, the Newton-Raphson Algorithm and the Fisher Scoring Algorithm, both converge in a small number of steps, but each step is computationally intensive. The Average Information Restricted Maximum Likelihood Algorithm (AI-REML) [5, 6] is a compromise between these two algorithms, which has the advantage of being significantly less computationally intensive (Supplementary Information 5).

Variance Component Testing

The null hypothesis $H_0 : \tau = 0$ can be tested by a Score Test, which involves the following statistic:

$$Q = (Y - X\widehat{\beta})' K (Y - X\widehat{\beta}) \tag{8}$$

where $\widehat{\beta}$ is the estimation of β under the null hypothesis (that is, a linear model). The distribution of Q is asymptotically normal, however for small samples it is more accurate to use a linear combination of chi-squares with one degree of freedom:

$$Q \sim \lambda_1 \chi^2(1) + \cdots + \lambda_n \chi^2(1),$$

where the coefficients $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of a $n \times n$ symmetric matrix depending on K and X (cf Supplementary Information 4 for details).

Estimating fixed effects, predicting random effects

The vector of fixed effects β can be estimated by plugging in formula (4) the REML estimates of σ and τ:

$$\widehat{\beta} = \left(X'\widehat{V}^{-1}X\right)^{-1} X'\widehat{V}^{-1}Y.$$

The values taken by the random terms u (or $\omega = Zu$) and ε in the model can then be *predicted*; their prediction $\widehat{u}$ and $\widehat{\varepsilon}$ must verify

$$Z\widehat{u} + \widehat{\varepsilon} = Y - X\widehat{\beta} = \widehat{V}\widehat{P}Y.$$

The values of $\widehat{u}$ and $\widehat{\varepsilon}$ are those which maximize the joint density of u and ε under the above constraint, that is (cf Supplementary Information 6)

$$\begin{aligned} \widehat{u} &= \widehat{\tau} Z'\widehat{P}Y \\ \widehat{\varepsilon} &= \widehat{\sigma}^2 \widehat{P}Y. \end{aligned} \tag{9}$$

Note that the predicted genomic value is $\widehat{\omega} = Z\widehat{u} = \widehat{\tau}K\widehat{P}Y$. These quantities are known under the acronym BLUPs, or more precisely eBLUPS, for (empirical) Best Linear Unbiased Predictors.

Note that $\widehat{u}$ is in the vector space spanned by the lines of Z. If $n \ll q$, as it is the case in many genetics applications, this is a small subspace of $\mathbb{R}^q$, and $\widehat{u}$ is merely a prediction of the projection of u on this subspace. This makes its usefulness dubious, in particular for prediction purposes.

Applications in Human Genetics

Pathways and rare variants association

The mixed model allows to test the association between a phenotype and a set of several SNPs; this is among the most popular applications of mixed models in human genetics. It has been first proposed for genetic pathways [7] or as an alternative to classical GWAS [8], but it is in the context of rare variants [9] that it really met the success under the name of Sequence

Kernel Association Test (SKAT), as it can be more powerful than the Burden tests that have been proposed previously [10–12]. We focus on this last case.

Consider n individuals, genotyped in q rare variants located in a genomic region to be tested for association with the phenotype Y; the genotypes are encoded by $g_{ij} = 0, 1$ or (seldom) 2, according to the rare allele counts. They are stored in a $n \times q$ matrix G. The following linear model, similar to (1), is used:

$$Y = X\alpha + GWu + \varepsilon \tag{10}$$

where X is a $n \times p$ matrix of covariates (including a column of ones for the intercept), W is a diagonal matrix of weights $w_1, \ldots, w_q$, $u \sim \mathcal{N}(0, \tau I_q)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. The role of W is to put a weight on each variant, depending on its rare allele frequency; the authors suggest to use $w_j = (1 - p_j)^{24}$, where p_j is the frequency of the rare allele at loci j. In this way, the potential effect of a the rarest variants is higher than the one of the most common variants.

The parameter τ models the association between Y and G; the null hypothesis of no association is H_0: $\tau = 0$, to test versus H_1: $\tau > 0$. The test statistics is as in (8),

$$Q_{\text{SKAT}} = (Y - X\widehat{\beta})' K (Y - X\widehat{\beta})$$

with $K = GW^2G'$.

The use of this model might seem surprising, as the hypothesis that the effects of the genotypes are random is rather unnatural. However, the test procedure which is derived from it is unquestionably correct: when $\tau = 0$ the model reduces to a simple linear model, in which the genotypes in G have no effect on Y. This is under this hypothesis that the asymptotic distribution of Q_{SKAT} is derived, and it is valid without doubt (if the trait is normally distributed, and with the classical caveat that population stratification might be a confounder). The mixed model can be here seen as mathematical trick to build a test for all variants at once; if the postulated relation between the rare allele frequency and its potential effect is close to the truth (which is of course debatable), the test will be powerful.

Many extensions of this test have been proposed, among which in particular SKAT-O which can be more powerful when the hypotheses of SKAT on the variants effects are not satisfied [13], an association test for both rare and common variants together [14], and an

association test for family samples [15].

Heritability, then and now

We first give a simple presentation of Fisher narrow-sense heritability [16, 17]. Consider a quantitative phenotype Y, partly determined by a large number q of unlinked di-allelic (autosomal) loci according to the model

$$Y = \alpha_0 + G^{(1)}\alpha_1 + \cdots + G^{(q)}\alpha_q + \varepsilon. \tag{11}$$

The genotypes $G^{(1)}, \ldots, G^{(q)}$ at the q causal loci are encoded by $0, 1$ or 2 according to the alternate allele counts; $\alpha_1, \ldots, \alpha_q$ are (small) fixed effects, and ε is a random term, including the contribution of the environment, which is assumed to be Gaussian with variance σ^2. This model assumes that there is neither gene-environment correlation, nor gene-environment interaction.

The genotypes $G^{(1)}, \ldots, G^{(q)}$ of a random individual are random, hence $G = G^{(1)}\alpha_1 + \cdots + G^{(q)}\alpha_q$ is a random term; if q is large enough, it is approximately Gaussian. It is assumed to be independent of the environment term ε. Denote its variance by $\mathrm{var}(G) = \tau$. The variance of Y is then $\mathrm{var}(Y) = \tau + \sigma^2$, and the proportion of variance due to genetic effects is the narrow-sense heritability

$$h^2 = \frac{\tau}{\tau + \sigma^2}. \tag{12}$$

More elaborate models can include dominance effects, epistasis terms, etc, [17], but this model is sufficient for the present discussion. Now consider two related individuals A and B, with coefficient of relationship Φ_{AB}. Their phenotypes are

$$Y_A = G_A + \varepsilon_A,$$

$$Y_B = G_B + \varepsilon_B.$$

The q loci considered being unlinked, it is easy to check that $\mathrm{cov}(G_A, G_B) = 2\Phi_{AB}\tau$. If one

assumes $\mathrm{cov}(\varepsilon_A, \varepsilon_B) = 0$ (no shared environment), then the variance of $Y = (Y_A, Y_B)'$ is

$$\mathrm{var}(Y) = \begin{bmatrix} \sigma^2 + \tau & 2\Phi_{AB}\tau \\ 2\Phi_{AB}\tau & \sigma^2 + \tau \end{bmatrix} = \tau \begin{bmatrix} 1 & 2\Phi_{AB} \\ 2\Phi_{AB} & 1 \end{bmatrix} + \sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \tau K + \sigma^2 I_2.$$

More generally, if $Y = (Y_1, \ldots, Y_n)'$ is a vector of phenotypes measured in individuals with coefficients of relationship Φ_{ij}, and the *kinship matrix* K is the matrix of general term $2\Phi_{ij}$, then $\mathrm{var}(Y) = \tau K + \sigma^2 I_n$. Equivalently, Y follows a mixed model as in (2), $Y = \mathbb{1}_n \beta + \omega + \varepsilon$, with $\mathrm{var}(\omega) = \tau K$ and $\mathrm{var}(\varepsilon) = \sigma^2 I_n$; τ and σ^2 can be estimated by REML, allowing to compute an estimate of h^2.

This presupposes the use of related individuals to estimate the heritability. The assumption of independence of the environment terms is a serious issue: the presence of shared environment will lead in upper biased estimates of h^2. Methods have been proposed to take into account the presence of shared environment, the most popular being the use of monozygotic and dizygotic twins in "twin studies" [17, 18].

We now turn to the "genomic heritability" theory, which allows to derive heritability estimates from population samples – which arguably wipes out the shared environment issue: while the individuals are not close relatives, which discards the existence of family environment, some gene-environment correlation might still be present. Many slightly different methods have been proposed [6, 19–22]. Here, we present a short account of the nowadays common practice, relying on softwares like GCTA. Consider n (unrelated) individuals genotyped in q autosomal SNPs. The "raw" genotypes g_{ij} are encoded by 0, 1, or 2 as above. The matrix Z of standardized genotypes is the matrix with general term

$$z_{ij} = \frac{g_{ij} - 2p_j}{\sqrt{2p_j(1-p_j)}}, \tag{13}$$

where p_j is the (empirical) frequency of the alternate allele. The term $2p_j$ is the mean of g_{ij} across individuals $i = 1, \ldots, n$, and $2p_j(1-p_j)$ is its expected variance under Hardy-Weinberg Equilibrium. Thus, the columns of Z are centered, with variance (approximately) equal to one. Consider then a mixed model as in (1):

$$Y = \mathbb{1}_n \beta + Zu + \varepsilon, \tag{14}$$

with $u \sim \mathcal{N}\left(0, \frac{\tau}{q} I_q\right)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$; then $\mathrm{var}(Y) = \tau K + \sigma^2 I_n$, with $K = \frac{1}{q} Z Z'$; the model parameters are estimated by REML, and the heritability h^2 is estimated by $\frac{\hat{\tau}}{\hat{\tau}+\hat{\sigma}^2}$.

When some covariates with an effect on the phenotype are known, they can be included in the model with a fixed effect, by a term $X\beta$. In this case, it is customary to report the heritability as $\frac{\hat{\tau}}{\hat{\tau}+\hat{\sigma}^2}$, ignoring the variance explained by the covariates [23]. We showed in [24] that it is possible to report a partition of the total variance of Y in three components (variance due to the presence of the covariates, to the random genomic effects, and residual variance).

For related individuals, the matrix K is an estimate of the kinship matrix. For unrelated individuals, it is usually called the Genetic Relationship Matrix (GRM), or Genomic Relationship Matrix. This may serve as a justification of the method, K being seen as an estimate of a genealogical relatedness between individuals, however small it may be. In this optic, the model (14) would be a mere mathematical artifact, the interest of which would only be to produce the "right" variance structure and to allow to estimate the different variance components. As the true proportion of genome sharing between relatives can differ significantly from its expected value [25,26], this may even be more accurate than using the kinship matrix computed from the genealogical information.

However, the temptation is strong to slip into an acceptance of the mixed model interpretation, despite the magnitude of the reversal from a model in which random genotypes have fixed effects, to a model in which the genotypes are considered fixed, but with a random effect. This reversal of the origin of randomness has strange consequences: the hypothesis of absence of linkage between the causal loci is no longer needed in the mixed model setting; it is replaced by the hypothesis of independence of the random effects. While Fisher's Model did not postulate anything about the effect sizes of the causal loci (except that they are small enough to produce an approximately normal distribution), the mixed model assumes that their expected values depend on the loci allele frequencies. The absence of gene-environment interaction and correlation is implicit in the hypothesis made on the distribution of the residual term, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

For the authors who accept the mixed model interpretation, the efficiency of the model (14) is due to its ability to incorporate a polygenic component through the random effects vector u. The standardization in equation (13) implies that the allelic effect at SNP j has variance $\mathrm{var}(u_j) \propto \frac{1}{p_j(1-p_j)}$, which is, in this interpretation, arbitrary. A different relationship

between the allele frequencies, e.g. of the form $\mathrm{var}(u_j) \propto \left(p_j(1-p_j)\right)^{\alpha}$ for different values of α, can then be postulated [22]; alternatively, the SNPs can be binned according to their MAF, and the model extended in order to have a variance parameter for each MAF bin [19]. The connexion between heritability estimates computed by such methods and the original definition of Fisher is, in our view, not totally clear.

Among the limitations of the method, its sensitivity to the presence of population stratification has been remarked [27]. A common solution is to incorporate in the model the Principal Components (PCs) of the whole genome data with fixed effect; this may not always be sufficient. We showed in [24], using nationwide French data from the Three-City study [28], that the estimated heritability of geographical coordinates of the birth place is equal to 1, and remains large even after inclusion of up to a thousand PCs.

The BLUPs $\widehat{u}$ have poor potential for predictive purposes [26, 29]. We will come back on this topic later in this paper. We noted earlier that the BLUPs are at best an estimate of the projection of the true effect vectors u (of dimension q) on a subspace of dimension n; this is certainly one of the reasons why their predictive performances are bad. Another problem is the heterogeneity of the effect sizes across the genome; to overcome this issue, it has been proposed to dynamically define genomic regions with different variance parameters [30]. Another possibility would be to include the SNPs with Genome Wide significant effects with fixed effects in the model, under the form of a polygenic risk score.

It must be noted that, even in the context of dairy cattle breeding where the environment is more homogeneous and the genetic diversity is smaller, it has been shown that the success of marker assisted genomic selection is due to the usage of closely related individuals, and that the accuracy of the prediction decreases drastically with the relationship coefficient between the individuals [31, 32]. This seems to imply that trait prediction using unrelated samples of humans is doomed to remain poorly efficient.

Taking into account Population Stratification in Association Testing

Because of the small effects of the frequent variants targeted in Genome Wide Association Studies (GWAS) (which have most of the time only indirect effects due to linkage disequilibrium with an ungenotyped causal variant) and because of the large number of such variants under consideration, it is necessary to integrate in these GWAS a large number of samples to

obtain a satisfying statistical power. This large sample size, in turn, makes it necessary to take population stratification into account to avoid spurious association.

A popular solution is the Principal Component Regression (PCR) which controls for population stratification by incorporating a few Principal Components (PCs) of the whole genome data in the model [33]:

$$Y = X\beta + PC_1\gamma_1 + \cdots + PC_k\gamma_k + \varepsilon \tag{15}$$

where the matrix X includes clinical covariates and the genotype of a SNP to be tested for association (by a Wald test or a Likelihood Ratio Test). Here, the effects $\gamma_1, \ldots, \gamma_k$ of the PCs are fixed effects. The number k of PCs is small, typically $k = 10$ or 20. As the presence of population stratification is reflected by the first PCs of the genomic data [34, 35], this reduces greatly the impact of population stratification.

Other authors [36–40] have proposed to use the mixed model (1) $Y = X\beta + Zu + \varepsilon$ for the same purpose, where Z is, as in (13), the standardized matrix of all genotypes, possibly excluding the genomic region in which lies the SNP under consideration. The term Zu is often interpreted as a "polygenic component", modeling the effect of the whole genome.

However, this model is related to the PCR through the Singular Value Decomposition (SVD) of Z [41]. The SVD of Z is $Z = U\Sigma L'$, with U a $n \times n$ orthogonal matrix, Σ a $n \times n$ diagonal matrix, and L a $q \times n$ matrix (with $L'L = I_n$). The PCs are then the columns of $U\Sigma$, and the corresponding loadings are the columns of L. Let $w = (w_1, \ldots, w_n)' = L'u$; the distribution of w is a multivariate Gaussian distribution with variance $L'(\tau I_q)L = \tau I_n$. Then, (1) can be rewritten

$$\begin{aligned} Y &= X\beta + Zu + \varepsilon \\ &= X\beta + (U\Sigma)(L'u) + \varepsilon \\ &= X\beta + PC_1 w_1 + \cdots + PC_n w_n + \varepsilon, \end{aligned} \tag{16}$$

where $w = (w_1, \ldots, w_n) \sim \mathcal{N}(0, \tau I_n)$.

Thus, a simple interpretation of the mixed model is that it extends the PCR by taking into account *all* the PCs of Z, with random instead of fixed effects. Its efficiency can be interpreted as its ability to model the population stratification through a large number of PCs, while avoiding the overfitting phenomenon that would make model (15) inefficient with k too

large.

This mixed model method has been found more efficient than the PCR in controlling the presence of population stratification [40, 42], although the PCR can be better in presence of environmental confounders [42]. It is possible to add a few PCs with fixed effects to the Linear Mixed Model (16) [42]:

$$Y = X\beta + PC_1\gamma_1 + \cdots + PC_k\gamma_k + PC_{k+1}w_{k+1} + \cdots + PC_n w_n + \varepsilon,$$

with fixed effects $\gamma_1, \ldots, \gamma_k$ and random effects $w_{k+1}, \ldots, w_n \sim \mathcal{N}(0, \tau)$.

The two possible interpretations of the model (polygenic component, or extension of the PCR) lead to different choices for Z: the polygenic component interpretation leads to include in Z the largest number of variants possible, including rare variants [33], whereas in PCR it is recommended to remove long-range LD regions, and even to keep only a set of variants in low mutual LD, e.g. $r^2 < 0.2$ [43, 44]. The impact of this choice doesn't seem to have been much commented in the literature.

Prediction with mixed models

We will shortly assess the prediction performances of the Linear Mixed Models in two situations: predicting the contribution of either a small set of rare variants or of the whole genome to a quantitative trait.

Rare variants

The Linear Mixed Model (10) is used by SKAT to test association between rare variants and a quantitative trait. It would then be natural to use BLUPs to estimate the variants effects, and for trait prediction.

To assess the performances of this method, we used a simulation procedure close to the one used in [9]. We relied on the haplotype dataset available in the R package SKAT [45], described in [9], which is composed of 10 000 haplotypes of a 200 kb region simulated using a coalescent model [46]. In a random subregion of the desired length (5 kb or 20 kb), a proportion π (π = 10% or 5%) of causal variants are drawn in the set of rare variants (minor allele

frequency < 5%), and $n = 2000$ or $n = 5000$ phenotypes are generated under the linear model

$$Y = G\beta + \varepsilon$$

where G is the matrix of genotypes at the rare variants, $\beta_j = 0.4\left|\log_{10}(\mathrm{maf}_j)\right|$ for causal variants as in [9], $\beta_j = 0$ for all other variants, and ε is normally distributed with standard deviation chosen such that the causal variants in the region explain a proportion α of the total variance varying from $\alpha = 1\%$ to $\alpha = 5\%$.

After fitting the Linear Mixed Model (10), the BLUPs $\widehat{u}$ for the SNP effects u can be computed and used for prediction. To measure the prediction accuracy, we define adjustment coefficients of $\widehat{Y}$ on $G\beta$ and on Y:

$$R^2_{\mathrm{gen}} = \frac{E\left(\left(\widehat{Y} - G\beta\right)^2\right)}{\mathrm{var}(G\beta)} \quad \text{and} \quad R^2_{\mathrm{tot}} = \frac{E\left(\left(\widehat{Y} - Y\right)^2\right)}{\mathrm{var}(Y)}$$

As $R^2_{\mathrm{tot}} = \alpha R^2_{\mathrm{gen}}$, we report only R^2_{gen}, the adjustment coefficient for $G\beta$, the genetic contribution of the region, which is easier to interpret. In our simulations, these values were estimated using the whole set of haplotypes, assumed to be representative of a whole population. We report the results in figure 1, together with the power of the test when type I error is fixed to 5%.

Some interesting trends are apparent in these results: both the power and R^2_{gen} decrease when the proportion α of explained variance decreases, and when the size of the considered region increases. This latter effect is much stronger for R^2_{gen} than for the power, which decreases moderately. As a result, the test can have an acceptable power (60% for $\alpha = 2\%$ and $n = 2000$ in a 20kb region) while the predictive capabilities of the model are poor ($R^2_{\mathrm{gen}} = 0.2$ in the same situation).

It must be noted that these estimates are optimistic, as we used the simulation model from [9] which is well suited for analysis with the Linear Mixed Model. A departure from linearity, with e.g. epistasis with an unobserved region, or G×E interaction with unobserved environment, could dramatically hinder the performances of the method.

Whole Genome data

We consider here the use of the model (14) for prediction of a centered phenotype Y ($\beta = 0$). Let us divide randomly a sample of n individuals in two parts: a learning sample of n_1 individuals, with a $n_1 \times q$ standardized genotypes matrix Z_1, and a test sample of size n_2, with a $n_2 \times q$ standardized genotypes matrix Z_2. The GRM matrix of the whole sample is

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}$$

with $K_{ij} = \frac{1}{q} Z_i Z_j'$. For the sake of simplicity, we assume here that the values of τ and σ^2 are known (in practice, their estimates $\widehat{\tau}$ and $\widehat{\sigma}^2$ would be used).

Using the model $Y_1 = Z_1 u + \varepsilon$ to estimate the genetic effects u, we can predict the phenotypes of the second sample by $\widehat{Y_2} = Z_2 \widehat{u}$. Using equation (9) it comes that

$$\begin{aligned} \widehat{Y_2} &= \tau K_{21} \left(\tau K_{11} + \sigma^2 I_{n_1}\right)^{-1} Y_1 \\ &= K_{21} \left(K_{11} + \frac{1-h^2}{h^2} I_{n_1}\right)^{-1} Y_1. \end{aligned}$$

Note that this is simply the expected value of Y_2 conditionally on the observed value of Y_1, assuming $Y = (Y_1', Y_2')' \sim \mathcal{N}(0, \tau K + \sigma^2 I_n)$.

We define again two adjustement coefficients

$$R^2_{\text{gen}} = \frac{E\left(\left(\widehat{Y}_{2i} - Z_{2i} u\right)^2\right)}{\text{var}(Z_{2i} u)} \quad \text{and} \quad R^2_{\text{tot}} = \frac{E\left(\left(\widehat{Y}_{2i} - Y_{2i}\right)^2\right)}{\text{var}(Y_{2i})} = h^2 R^2_{\text{gen}}.$$

As $E\left(\widehat{Y}_{2i}\right) = E(Y_{2i}) = 0$, we have

$$R^2_{\text{gen}} = \frac{2\,\text{cov}\left(\widehat{Y}_{2i}, Z_{2i} u\right) - \text{var}\left(\widehat{Y}_{2i}\right)}{\text{var}(Z_{2i}) u}.$$

If the value Y_2 of the phenotype of the individuals of the test sample is known, $\text{cov}\left(\widehat{Y}_{2i}, Z_{2i} u\right) = \text{cov}\left(\widehat{Y}_{2i}, Y_{2i}\right)$ can be estimated by

$$\frac{Y_2' \widehat{Y_2}}{n_2}.$$

However, even in the absence of measured phenotypes, the expected value of the above ex-

pression can be computed; it is equal to

$$\frac{1}{n_2}\tau \operatorname{tr}\left(K_{21}\left(K_{11}+\frac{1-h^2}{h^2}I_{n_1}\right)^{-1}K_{12}\right). \tag{17}$$

On the other hand, we have

$$\operatorname{var}\left(\widehat{Y}_2\right)=\tau K_{21}\left(K_{11}+\frac{1-h^2}{h^2}I_{n_1}\right)^{-1}K_{12},$$

hence the expression (17) is also an estimation of the variance of the components of Y_2. Finally, we can estimate the expected value of R^2_{gen} by

$$E\left(R^2_{\text{gen}}\right)=\frac{1}{n_2}\operatorname{tr}\left(K_{21}\left(K_{11}+\frac{1-h^2}{h^2}I_{n_1}\right)^{-1}K_{12}\right). \tag{18}$$

This formula allows to compute the expected value of R^2_{gen} and $R^2_{\text{tot}}=h^2R^2_{\text{gen}}$ using solely a matrix GRM matrix K, and the heritability h^2. As soon as the test sample size n_2 exceeds a few hundreds individuals, this value is very stable, the choice of the individuals composing the two sub-samples making very little difference.

A very rough approximation can be obtained by noting that the matrix $h^2K_{11}+(1-h^2)I_{n_1}$ is close to I_{n_1} (the off-diagonal terms in K_{11} being typically very small), thus

$$E\left(R^2_{\text{gen}}\right)\simeq h^2\kappa n_1 \tag{19}$$

where κ is the variance of the terms in K_{12}, ie the variance of the off-diagonal terms in K. This simplistic approximation can't be valid for large values of n_1, as it is unbounded. Using GRM matrices to compute the value of (18) for several learning sample sizes n_1 seems inevitable.

It is however difficult to obtain GRM matrices computed with large real data sets from unstratified or moderately stratified populations. The natural distribution for random covariance matrices is the Wishart distribution; comparison with real data prove however that the spectrum of GRM matrices is not similar to the spectrum of Wishart matrices, making this model inappropriate. Instead, we propose the following model for random GRM matrices: let

$$K=U\Lambda U'$$

where U is a random $n \times n$ orthogonal matrix and $\Lambda = \mathrm{diag}(\lambda_1, \ldots \lambda_n)$ is a diagonal matrix. Tedious but elementary computations show that, for Λ fixed, the expected value of the diagonal terms of K is $E(\lambda) = \frac{1}{n}\sum_i \lambda_i$, while the off-diagonal term are centered with variance $\simeq \frac{1}{n}\mathrm{var}(\lambda) = \frac{1}{n}\sum_i (\lambda_i - E(\lambda))^2$. The values of the λ_i were thus taken in distributions similar to the distribution observed on real data, with the constraint $E(\lambda) = 1$ and $\mathrm{var}(\lambda) = n\kappa$ where $\kappa = 1.63 \cdot 10^{-5}$ is the variance observed on the GRM computed on the nationwide French data in [24]. Numerical experiments show that for a fixed matrix Λ, the value of R^2_{gen} (equation 18) is very concentrated around its mean across different random orthogonal matrices U.

This procedure was used to generate GRM matrices of size $n = n_1 + n_2$ with n_1 up to $n_1 = 25\,000$, and $n_2 = 1000$. The resulting values of $E\left(R^2_{\mathrm{gen}}\right)$ are plotted on figure 2. For a learning sample size n_1 up to 5000, they are very similar to the values obtained with real data from [24]. For $n < 10\,000$, $E\left(R^2_{\mathrm{gen}}\right)$ grows almost linearly with n_1, in surprisingly good accordance with the approximation (19).

The estimated prediction performances are, to say the least, very poor — and here again, this an optimistic scenario, as all computations are performed under the Linear Mixed Model. As observed for the rare variants, the lower the heritability, the less efficient is the learning process.

Discussion

We gave a synthetic presentation of the Linear Mixed Model (LMM) and of the tools used for analyzing models in this theory. We then presented three widespread applications in Human Genetics, which have in common that allelic effects are modeled as random effects. In the most commons applications of LMMs, random effects are used when repeated measurements are taken on random individuals or in random groups: classical examples include daughter cows of a random sire, repeated measures on an individual, batch effect, center effect, etc. Their modeling as Gaussian random variables is justified by the fact that these effects are likely to be the resultant of multiple unknown small effects. In contrast, modeling the allelic effect of a SNP as random is unnatural; one could argue that the SNP is the produce of a random mutation, but anyway in our experiments the SNPs are not drawn at random. Besides, two of these methods assume that these random effects standard deviation is proportional

to the inverse of the standard deviation of the genotype encoded by the allele counts; this supposition is at best mathematically convenient.

When used for the rare variants association tests (SKAT and extensions), the LMM can be seen as a simple way to construct a test, which could even be constructed without the LMM [47, 48]. A simulation study showed that when dealing with small enough genomic regions, the BLUPs could even be used for prediction purposes. One must be careful in interpreting these results, as these simulations were performed under a favourable scenario; in the presence of departures from linearity (in particular, interaction with unobserved genetic or environmental factors), both the testing and prediction performances would be damaged.

Another application the LMM is to correct for population stratification in Genome Wide Association Studies, as an alternative to Principal Component Regression. Both strategies aim at incorporating the whole genome in the model, either by summarizing it by its first few Principal Components, or, for the LMM, by modeling the effect of each SNP as a random effect. Alternatively, the LMM can be seen as a generalization of the Principal Component Regression in which all Principal Components are included in the model, with random effects. In any case, the efficiency of both methods is due to their ability to take into account the genetic proximity of the individuals which pertain to the same population stratum.

A statistical model can be seen simply as a mathematical tool to construct a test for a particular hypothesis, such as the independence between a phenotype and a set of genotypes. In this case, the model needs only to give a satisfying description the observations distribution when restricted to the null hypothesis; this is enough to produce an unbiased test. But to use this model for prediction purposes (in our case, prediction of the phenotype from the genotypes), it is necessary that the model gives a good approximation of the observations distribution in the general case, not only under the null hypothesis. Finally, to find some descriptive value to the estimations of the parameters of the model, such as the heritability h^2, one must consider that the model is an accurate description of reality.

The ability of the LMM to produce useful tests in human genetics does not imply that it describes accurately the relation between the genome and the phenotype. One must be very cautious when using it for anything else than testing association. The computation of heritability estimates with a LMM, which was the third application we presented, is the more troublesome application in our eyes. It would be much less speculative to use the same model

to test whether the genomic proximity of the individuals is related to their phenotypic similarity – with, as always, the caveat that such a relation might be due to population stratification, and that it might be difficult to account for stratification [24]. But taking the model for granted and attaching too much value to the heritability estimates it produces is more than audacious.

We also considered the question of prediction using genome-wide data: simulating realistic Genetic Relationship Matrices (calibrated using the magnitude of relationships observed in real data), we showed that, even under the model assumptions, the expected prediction performances are poor when weakly related individuals are used.

Our discussion focused on the modeling of genetic effects as random in the Linear Mixed Model, without much discussion of the hypotheses of additivity, absence of interaction terms, etc., which of course are in themselves very objectionable [49, 50]. We also restricted to quantitative traits; most of our remarks remain valid when Generalized Linear Mixed Models are used to model a binary outcome.

References

[1] Henderson CR. Estimation of variance and covariance components. Biometrics 1953; 9:226–252.

[2] Henderson CR. Best linear unbiased estimation and prediction under a selection model. Biometrics 1975;423–447.

[3] Amos CI. Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 1994;54:535.

[4] Searle SR, Casella G, McCulloch CE. Variance components, volume 391. John Wiley & Sons, 2009.

[5] Gilmour AR, Thompson R, Cullis BR. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 1995;1440–1450.

[6] Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 2011;88:76–82.

[7] Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinform 2008;9:292.

[8] Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet 2010; 86:929–942.

[9] Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 2011;89:82–93.

[10] Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res 2007; 615:28–56.

[11] Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 2009;5:e1000384.

[12] Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 2008;83:311–321.

[13] Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 2012;13:762–775.

[14] Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet 2013;92:841–853.

[15] Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol 2013;37:196–204.

[16] Fisher RA. The correlation between relatives on the supposition of mendelian inheritance. Transactions of the Royal Society of Edinburgh 1918;52:399–433.

[17] Falconer DS. Introduction to quantitative genetics. DS Falconer, 1960.

[18] Scarr S. Environmental bias in twin studies. Eugenics Quarterly 1968;15:34–40.

[19] Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Lee SH, Robinson MR, Perry JRB, Nolte IM, van Vliet-Ostaptchouk JV, Snieder H, Study TLC, Esko T, Milani L, Magi R, Metspalu A, Hamsten A, Magnusson PKE, Pedersen NL, Ingelsson E, Soranzo N, Keller MC, Wray NR, Goddard ME, Visscher PM. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet 2015;.

[20] Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM. Common SNPs explain a large proportion of the heritability for human height. Nat Genet 2010;42:565–569.

[21] Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 2011;88:294–305.

[22] Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. Am J Hum Genet 2012;91:1011–1021.

[23] Visscher PM, Hill WG, Wray NR. Heritability in the genomics era—concepts and misconceptions. Nat Rev Genet 2008;9:255–266.

[24] Dandine-Roulland C, Bellenguez C, Debette S, Amouyel P, Génin E, Perdry H. Accuracy of heritability estimations in presence of hidden population stratification. submitted (2016);.

[25] Leutenegger AL, Prum B, Génin E, Verny C, Lemainque A, Clerget-Darpoux F, Thompson EA. Estimation of the inbreeding coefficient through use of genomic data. Am J Hum Genet 2003;73:516–523.

[26] Speed D, Balding DJ. Relatedness in the post-genomic era: is it still useful? Nat Rev Genet 2015;16:33–44.

[27] Browning SR, Browning BL. Population structure can inflate SNP-based heritability estimates. Am J Hum Genet 2011;89:191.

[28] Group CS, et al. Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population. Neuroepidemiology 2003;22:316.

[29] Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet 2013;14:507–515.

[30] Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res 2014;24:1550–1557.

[31] Habier D, Fernando R, Dekkers J. The impact of genetic relationship information on genome-assisted breeding values. Genetics 2007;177:2389–2397.

[32] Habier D, Tetens J, Seefried FR, Lichtner P, Thaller G. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Select Evol 2010;42:5.

[33] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–909.

[34] Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD. Genes mirror geography within Europe. Nature 2008;456:98–101.

[35] Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, Fabianova E, Foretova L, Georges M, Janout V, Kabesch M, Krokan HE, Elvestad MB, Lissowska J, Mates D, Rudnai P, Skorpen F, Schreiber S, Soria JM, Syvanen AC, Meneton P, Hercberg S, Galan P, Szeszenia-Dabrowska N, Zaridze D, Génin E, Cardon LR, Lathrop M. Investigation of the fine structure of European populations with applications to disease association studies. Eur J Hum Genet 2008;16:1413–1429.

[36] Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet 2012;44:821–824.

[37] Aulchenko YS, De Koning DJ, Haley C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics 2007;177:577–585.

[38] Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Methods 2011;8:833–835.

[39] Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. Efficient control of population structure in model organism association mapping. Genetics 2008; 178:1709–1723.

[40] Kang HM, Sul JH, Service SK, Zaitlen NA, Kong Sy, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. Nat Genet 2010;42:348–354.

[41] Hoffman GE. Correcting for population structure and kinship using the linear mixed model: theory and extensions. PloS one 2013;8:e75707.

[42] Zhang Y, Pan W. Principal component regression and linear mixed model in association analysis of structured samples: competitors or complements? Genet Epidemiol 2015; 39:149–155.

[43] Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV, Ge D, Rotter JI, Torres E, Taylor K, Goldstein DB, Reich D. Long-range LD can confound genome scans in admixed populations. Am J Hum Genet 2008;83:132–135.

[44] Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc 2010;5:1564–1573.

[45] Lee S, Miropolsky L, Wu M. Package SKAT, 2015. URL `https://cran.r-project.org/web/packages/SKAT/index.html`.

[46] Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. Genome Res 2005;15:1576–1583.

[47] Schaid DJ. Genomic similarity and kernel methods i: advancements by building on mathematical and statistical foundations. Human heredity 2010;70:109–131.

[48] Schaid DJ. Genomic similarity and kernel methods ii: methods for genomic information. Human heredity 2010;70:132–140.

[49] Nelson RM, Pettersson ME, Carlborg Ö. A century after Fisher: time for a new paradigm in quantitative genetics. Trends in Genetics 2013;29:669–676.

[50] Génin E, Clerget-Darpoux F. The missing heritability paradigm: a dramatic resurgence of the GIGO Syndrome in genetics. Hum Hered 2015;79:10–13.

[51] Lin X. Variance component testing in generalised linear models with random effects. Biometrika 1997;84:309–326.

[52] Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. Biometrics 2007;63:1079–1088.

[53] Davies RB. Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. J R Stat Soc Ser C Appl Stat 1980;29:323–333.

Figures

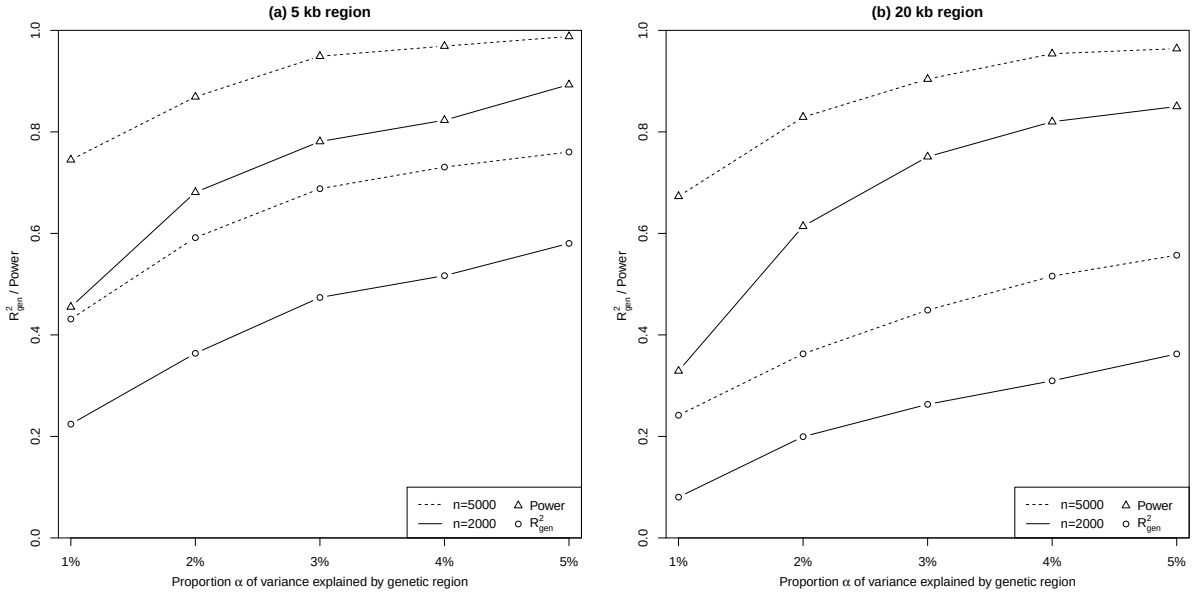


Figure 1: Power of SKAT and prediction performances using BLUPs, for two genomic regions of 5 kb and 20 kb.

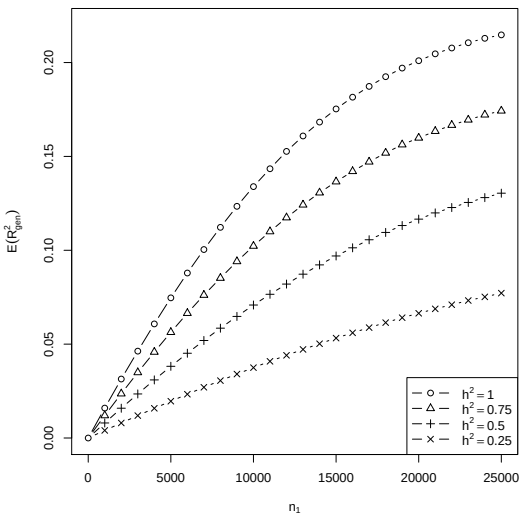


Figure 2: Using BLUPs on genome-wide data: expected adjustement coefficient as a function of the learning sample size n_1, for $h^2 = 0.25$, 0.5, 0.75 and 1.

Article 3

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Accuracy of heritability estimations in presence of hidden population stratification

Received: 09 February 2016
Accepted: 29 April 2016
Published: 25 May 2016

Claire Dandine-Roulland[1], Céline Bellenguez[2], Stéphanie Debette[3,4,5], Philippe Amouyel[2], Emmanuelle Génin[6,7,8] & Hervé Perdry[1]

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Accuracy of heritability estimations in presence of hidden population stratification

Received: 09 February 2016
Accepted: 29 April 2016
Published: 25 May 2016

Claire Dandine-Roulland[1], Céline Bellenguez[2], Stéphanie Debette[3,4,5], Philippe Amouyel[2], Emmanuelle Génin[6,7,8] & Hervé Perdry[1]

The heritability of a trait is the proportion of its variance explained by genetic factors; it has historically been estimated using familial data. However, new methods have appeared for estimating heritabilities using genomewide data from unrelated individuals. A drawback of this strategy is that population stratification can bias the estimates. Indeed, an environmental factor associated with the phenotype may differ among population subgroups. This factor being associated both with the phenotype and the genetic variation in the population would be a confounder. A common solution consists in adjusting on the first Principal Components (PCs) of the genomic data. We study this procedure on simulated data and on 6000 individuals from the Three-City Study. We analyse the geographical coordinates of the birth cities, which are not genetically determined, but the heritability of which should be overestimated due to population stratification. We also analyse various anthropometric traits. The procedure fails to correct the bias in geographical coordinates heritability estimates. The heritability estimates of the anthropometric traits are affected by the inclusion of the first PC, but not by the following PCs, contrarily to geographical coordinates. We recommend to be cautious with heritability estimates obtained from a large population.

The heritability of a quantitative phenotype is the proportion of its variance explained by genetic factors. The concept of heritability can be traced back to the pioneering works of Galton in the nineteenth century[1]; its modern definition is due to Fisher in[2]. Most recent works focus on the narrow-sense heritability which is the proportion of variance explained by additive genetic effects.

Heritability estimates have long been based on family data[3]. Twin studies[4] were very popular, as they provide an easy way to take into account the shared environment in families, which can bias the estimates if unaccounted for in the analyses. However the possibility of a bias due to difference in shared environment in monozygotic and dizygotic twins remained[4,5].

The research of the genetic polymorphisms causing the variability of the phenotypes with a strong genetic component was carried notably through numerous Genome-Wide Association Studies (GWAS). Hundreds of Single Nucleotide Polymorphisms (SNPs) have been found associated with complex traits[6,7]. However the variance explained by these SNPs is lower than the genetic variance predicted by the family studies; the unexplained variance was called the "missing heritability"[8–13].

The missing heritability problem triggered the development of methods using genome-wide data to obtain heritability estimates from unrelated individuals[13,14]. These methods have now mostly superseded family based methods; they are often advocated and perceived as providing unbiased estimates, in particular because, as the individuals are unrelated, there is no shared environment[15]. It was however demonstrated early that the presence of population stratification inflates heritability estimates[16]. A common practice to correct this bias is to include a few Principal Components (PCs) of the estimated kinship matrix in the model with fixed effects[17–21].

The common usage is to include 10 or 20 PCs in the model with fixed effects. There are however no theoretical grounds on which this number of PCs is chosen; more PCs could be included without damaging the estimates. In

[1]CESP, Inserm, Univ. Paris-Sud, Université Paris-Saclay, Villejuif, France. [2]UMR1167 - Labex Distalz, Univ. Lille, Inserm, CHU Lille, Institut Pasteur de Lille, F-59000 Lille, France. [3]Université de Bordeaux, Bordeaux, France. [4]Inserm U1219, Bordeaux, France. [5]CHU de Bordeaux, Bordeaux, France. [6]Inserm, UMR 1078, Brest, France. [7]Université Bretagne Occidentale, Brest, France. [8]Centre Hospitalier Régional Universitaire, Brest, France. Correspondence and requests for materials should be addressed to C.D.-R. (email: claire.dandine-roulland@inserm.fr)

	Family data	Twin data	Genomic data
Height	0.92[49]	0.68 to 0.94[50–53]	0.44 to 0.62[14,35,53–57]
Weight	–	0.37[53]	0.19[35], 0.26[53]
BMI	0.24 to 0.81 (mean 0.46)[58]	0.47 to 0.90 (mean 0.75)[58]	0.16 to 0.27[35,53,55,57]
		0.28[53]	
		0.45 to 0.84[59]	
Waist circumference	–	0.15[53]	0.16[53], 0.17 (men or women)[57,†]
Hip circumference	–	–	0.23 (men)[†], 0.19 (women)[57,†]
Waist-to-hip ratio	–	–	0.16 (men)[†], 0.18 (women)[57,†]
Head circumference	0.66[60]	0.75[59]	–
Skeletal traits (including head circumference)	–	0.59[52]	–

Table 1. Heritability estimations in the litterature. †Adjusted on the BMI and stratified on the sex. Ref. 58 is a meta-analysis of 88 twin studies and 27 family studies.

this paper, we explore the variability of genomic heritability estimates with the number of PCs included as fixed effects in the model, and the ability of this method to effectively correct for population stratification.

We use simulated data to assess the properties of the method, which we also apply on the Three-City (3C) data[22]. The 3C study is a longitudinal study including 9294 French individuals aged 65 years or older, recruited between 1999 and 2001 in the cities of Bordeaux, Dijon and Montpellier. The longitude and latitude of the birth cities provide examples of "purely stratified" traits, which are not determined by the genome but should be very correlated to the first PCs[23]. Thus, the "naive" genomic heritability estimates of these traits should be artificially high; including PCs in the model should reduce these estimates. Our aim is to find how many PCs have to be included to get a genomic heritability estimate close to zero.

We also analyse anthropometric phenotypes: height which is the historical example of a highly heritable trait[2,24], weight, Body Mass Index (BMI), head circumference and waist-to-hip ratio which are also expected to display significant heritability. GWAS studies discovered many variants associated with human height[25–27], and with obesity related traits such as weight, BMI, waist or hip circumference and waist-to-hip ratio[28–33]. The heritability of all these traits have been estimated multiple times through twin studies, familial studies, or genomic data; Table 1 summarizes some estimates from the literature. We compute their genomic heritability on the 3C data set, and consider how it evolves when including PCs with fixed effects.

Results

Using linear mixed models, we estimate variance components and heritabilities h^2 of several traits. The decompositions of phenotypic variance are presented in graphics with number of Principal Components included with fixed effects in the model as abscissae and the proportion of explained variance as ordinates. The estimated proportion of variance explained by the fixed effects is displayed in dark gray. The light gray and white colors are respectively the estimated proportions of genetic and residual variances.

For the 3C phenotypes, we also give Tables with numerical value of the parameter estimates, their standard errors, and the likelihood ratio test for H_0: $h^2 = 0$.

Simulated data. Phenotypes were simulated under the linear mixed model, with a non-zero effect on the first 10 PCs (see details in Methods section). They were analysed with a number p of PCs included in the model with fixed effects varying from 0 to 2000.

Figure 1 displays the mean of estimated variance proportions, and their standard deviations, computed on 100 simulation replicates. From 0 to 2000 PCs are included with fixed effects. We note that as soon as 10 PCs or more are included in the model, the true proportions of variance (20% for the PCs, 40% for each of the random terms) are well estimated (with a standard error close to 0.05). In contrast, when the number of included PCs is lower than 10, the proportion of genetic variance is overestimated. When up to 2000 PCs are included, the means of estimates are stable, while the standard deviation increase slowly.

Three-City data. Variance estimations have also been made on several quantitative traits from the 3C study. First of all, we considered as quantitative phenotypes the longitude and the latitude of the birth cities. Geographical coordinates of all birth cities are displayed on the map of France in the first panel of Supplementary Fig. S1. We also estimated the heritability of several anthropometric phenotypes: height, weight, BMI, waist-to-hip ratio and head circumference.

The mean and the standard deviation of each analyzed variable are given in Table 2, stratified on the sex. The sex has been included as a covariate when estimating the heritability of the anthropometric phenotypes.

Longitude and latitude. The variance decompositions of longitude and latitude are displayed in the two panels of Fig. 2, where we included up to $p = 2000$ PCs in the model with fixed effects. When no PCs are included in the model, it is estimated that 100% of the variance is genetic both for longitude and latitude. When adding PCs to the model, this estimated proportion decreases slowly for the longitude (first panel of Fig. 2). It is estimated close to 54% for $p = 50$ PCs, and to 26% for $p = 2000$. In contrast, the estimated proportion of genetic variance of latitude

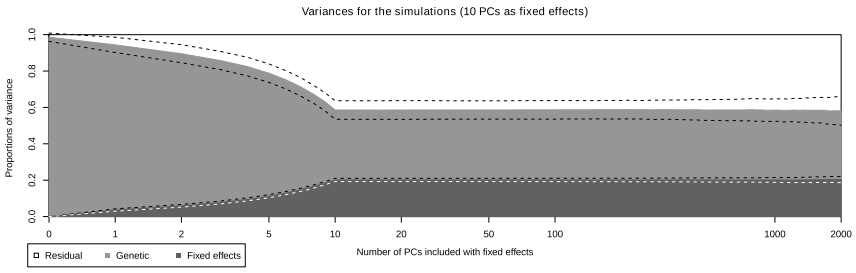


Figure 1. Estimated proportions of variance for the simulated data, depending on the number of PCs included in the model (log-scale). The white, light gray and dark gray are respectively the residual, genetic and fixed effects estimated means on 100 replicates. The dashed lines represent the mean ±1 standard deviation.

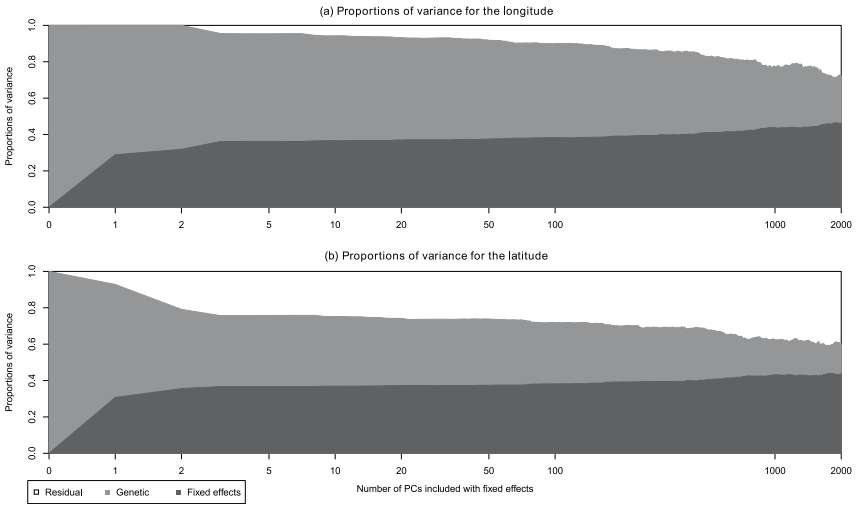


Figure 2. Estimated proportion of variance for the geographical coordinates, depending on the number of PCs included in the model (log-scale). The white, light gray and dark gray are respectively the residual, genetic and fixed effects variances.

	Men ($N = 2298$)			Women ($N = 3495$)			All ($N = 5793$)		
Phenotype	**Mean**	**Sd**	***n***	**Mean**	**Sd**	***n***	**Mean**	**Sd**	***n***
Age	74.15	5.56	2298	74.39	5.49	3495	74.30	5.52	5793
Latitude	46.78	1.73	2090	46.76	1.63	3171	46.77	1.67	5261
Longitude	3.32	2.40	2090	3.35	2.45	3171	3.34	2.43	5261
Height	169.58	6.35	2290	156.60	6.17	3461	161.77	8.91	5751
Weight	75.58	11.27	2292	62.58	11.32	3485	67.74	12.97	5777
BMI	26.27	3.53	2288	25.52	4.36	3457	25.82	4.06	5745
Head circumference	57.75	2.05	2243	55.37	2.07	3414	56.32	2.37	5657
Waist-to-hip ratio	0.95	0.07	2117	0.84	0.07	3180	0.88	0.09	5297

Table 2. Descriptive statistics of the 3C quantitative traits.

decreases sharply (to 39% of the variance) with the first few PCs included in the model; it then decreases slowly, as observed for latitude: it is still close to 36% for $p = 50$ PCs, decreasing to 16% for $p = 2000$. For both geographical coordinates, the first PC explains a non-negligible proportion of variance (29% and 31% for longitude and latitude respectively). Once the first PC is included, the proportion of variance explained by the PCs increases slowly with p. This observation is consistent with the correlation values of the geographical coordinates with the first few PCs (Supplementary Fig. S2).

Precise figures are given in Tables 3 and 4 for longitude and latitude heritability respectively. These Tables give, for various values of p, the estimates of σ^2, τ, and h^2, together with their standard error, and the likelihood ratio

	LRT	*p*-value	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma_P^2}$ (se)	$\widehat{h^2}$ (se)
0 PC	1892.63	<1e-40	4.34 (0.085)	0.00029 (0.089)	4.34 (0.084)	1.000
1 PC	563.65	<1e-40	3.83 (0.075)	0.00025 (0.069)	3.83 (0.075)	1.000
2 PCs	405.08	<1e-40	3.71 (0.072)	0.00025 (0.065)	3.71 (0.072)	1.000
3 PCs	221.81	<1e-40	3.28 (0.069)	0.24 (0.061)	3.53 (0.069)	0.931 (0.060)
4 PCs	219.54	<1e-40	3.27 (0.069)	0.26 (0.061)	3.52 (0.069)	0.928 (0.060)
5 PCs	219.56	<1e-40	3.27 (0.069)	0.25 (0.061)	3.52 (0.069)	0.928 (0.060)
10 PCs	204.70	<1e-40	3.19 (0.069)	0.32 (0.061)	3.50 (0.068)	0.910 (0.061)
20 PCs	193.57	<1e-40	3.12 (0.069)	0.37 (0.061)	3.49 (0.068)	0.894 (0.062)
50 PCs	177.83	<1e-40	3.00 (0.068)	0.46 (0.061)	3.46 (0.068)	0.868 (0.063)
100 PCs	157.09	2.4e-36	2.87 (0.068)	0.56 (0.061)	3.43 (0.067)	0.838 (0.065)
500 PCs	87.59	4.0e-21	2.35 (0.070)	0.97 (0.063)	3.32 (0.068)	0.707 (0.073)
1000 PCs	43.24	2.4e-11	1.90 (0.072)	1.30 (0.065)	3.20 (0.069)	0.594 (0.087)
2000 PCs	14.00	9.1e-5	1.51 (0.083)	1.61 (0.072)	3.12 (0.077)	0.485 (0.124)

Table 3. Model parameter estimates for the longitude and their standard error, depending on the number of PCs included in the model, likelihood ratio test statistics (LRT) to test significance of heritability and their *p*-values[34]. $\widehat{\tau}$ is the estimated genetic variance, $\widehat{\sigma^2}$ the estimated residual variance, $\widehat{\sigma_P^2} = \widehat{\tau} + \widehat{\sigma^2}$ the estimated total variance, and $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ estimated heritability.

	LRT	*p*-value	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma_P^2}$ (se)	$\widehat{h^2}$ (se)
0 PC	1854.15	<1e-40	2.05 (0.040)	0.00014 (0.039)	2.05 (0.040)	1.000
1 PC	295.37	<1e-40	1.60 (0.035)	0.18 (0.031)	1.78 (0.035)	0.897 (0.054)
2 PCs	126.24	1.4e-29	1.14 (0.033)	0.55 (0.031)	1.69 (0.033)	0.676 (0.061)
3 PCs	98.87	1.3e-23	1.03 (0.033)	0.64 (0.031)	1.66 (0.032)	0.616 (0.063)
4 PCs	99.13	1.2e-23	1.03 (0.033)	0.64 (0.031)	1.67 (0.032)	0.617 (0.063)
5 PCs	99.37	1.0e-23	1.03 (0.033)	0.64 (0.031)	1.67 (0.032)	0.618 (0.063)
10 PCs	95.50	7.4e-23	1.00 (0.033)	0.65 (0.031)	1.66 (0.032)	0.608 (0.063)
20 PCs	88.35	2.7e-21	0.98 (0.033)	0.68 (0.031)	1.66 (0.032)	0.589 (0.063)
50 PCs	83.73	2.8e-20	0.96 (0.033)	0.69 (0.031)	1.65 (0.032)	0.582 (0.064)
100 PCs	68.81	5.4e-17	0.89 (0.033)	0.75 (0.031)	1.64 (0.032)	0.543 (0.066)
500 PCs	38.96	2.2e-10	0.73 (0.034)	0.86 (0.031)	1.59 (0.033)	0.459 (0.074)
1000 PCs	15.26	4.7e-5	0.52 (0.035)	1.02 (0.032)	1.54 (0.033)	0.338 (0.086)
2000 PCs	5.22	0.0112	0.44 (0.042)	1.11 (0.037)	1.55 (0.038)	0.284 (0.122)

Table 4. Model parameter estimates for the latitude and their standard error, depending on the number of PCs included in the model, likelihood ratio test statistics (LRT) to test significance of heritability and their *p*-values. $\widehat{\tau}$ is the estimated genetic variance, $\widehat{\sigma^2}$ the estimated residual variance, $\widehat{\sigma_P^2} = \widehat{\tau} + \widehat{\sigma^2}$ the estimated total variance, and $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ estimated heritability.

test for the null hypothesis H_0: $h^2 = 0$. We first note that for all values of p up to $p = 2000$, the Likelihood Ratio Test (LRT) is significant for the heritability of both longitude and latitude (note that the LRT asymptotically follows a $\frac{1}{2}\chi^2(0) : \frac{1}{2}\chi^2(1)$ mixture[34]; the 5% significance threshold is 2.70).

We also note that the standard error of the various estimates increases with p, but within reasonable bounds: for example the estimated heritability of the longitude is $\widehat{h^2} = 1$ with standard error 0.04 for $p = 0$, $\widehat{h^2} = 0.59$ with standard error 0.09 for $p = 1000$, and $\widehat{h^2} = 0.48$ with standard error 0.12 for $p = 2000$.

Anthropometric traits. We analysed height, weight, BMI, head circumference and waist-to-hip ratio. For these traits, sex and age were included in the model as covariates.

In Fig. 3, the estimated proportions of variance for height, head circumference and waist-to-hip ratio are displayed. Sex and age alone are responsible for 52%, 24% and 41% of the variance of height, head circumference and waist-to-hip ratio respectively. Without population stratification correction ($p = 0$), the genetic variance is 22%, 13% and 12% of the total variance; the inclusion of a single PC in the model ($p = 1$) decreases these values to 19%, 8% and 8% respectively. Including up to $p = 50$ PCs does not modify much these proportions. For p up to 2000, the variance proportion estimates fluctuate around the previously cited values.

The estimated proportions of variance for weight and BMI are displayed in Supplementary Fig. S3: the inclusion of PCs with fixed effect does not impact much the estimated proportions of genetic variance, which are 16% and 19% respectively.

Table 5 gives for the five anthropometric phenotypes and various values of p, the estimates of σ^2, τ, and h^2, together with their standard error, and the likelihood ratio test for the null hypothesis H_0: $h^2 = 0$.

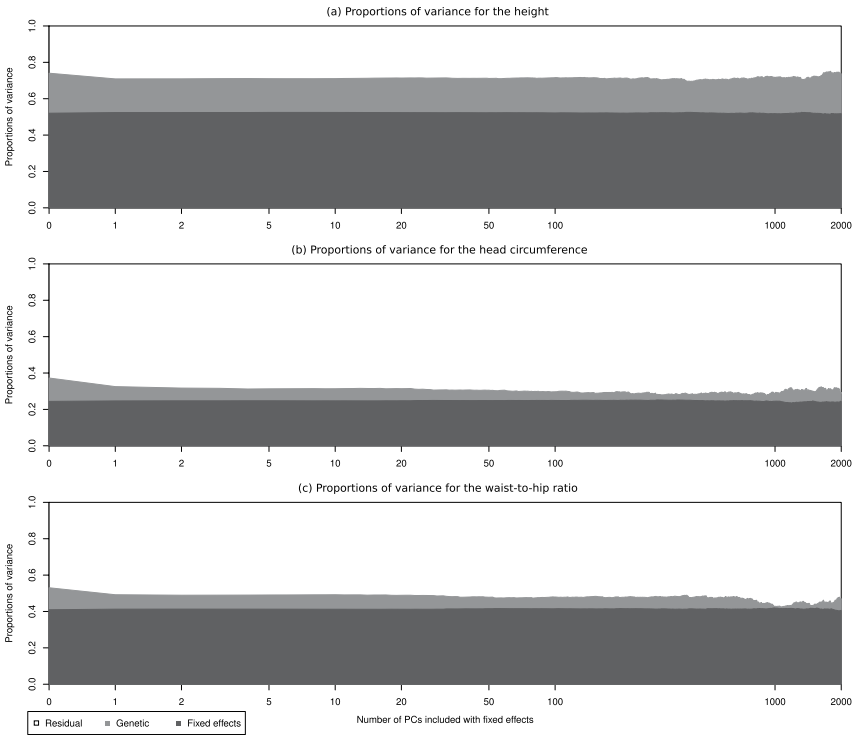


Figure 3. Estimated proportion of variance for (**a**) height, (**b**) head circumference, and (**c**) waist-to-hip-ratio, depending on the number of PCs included in the model (log-scale). The white, light gray and dark gray are respectively the residual, genetic and fixed effects variances.

Without stratification correction, height heritability is estimated to 46% with standard error of 6.1%; as soon as one PC is included in the model, this value drops to 39% (standard error 6.6%). Inclusion of more PCs (up to $p = 20$) almost does not change the estimated values nor their standard errors. The likelihood ratio tests for $h^2 = 0$ are significant.

Heritability of weight and BMI seem to be almost unaffected by population stratification correction. In the two cases, heritability is significantly positive. It is estimated close to 22% (standard error 6%) and 19% (standard error 6%).

Head circumference heritability is estimated to 17% (standard error 5.7%) without population stratification correction. With inclusion of $p = 1$ PC in the model, it is only 10%, which drops to 9% for $p = 2$ (with a standard error of 6.3% or 6.4%). The likelihood ratio test is no longer significant.

Waist-to-hip ratio heritability is estimated to 20% (standard error 6.1% without population stratification correction. This value drops to 13% (standard error 6.7%) with the inclusion of one PC in the model. The likelihood ratio tests stay significant.

Discussion

Our analyses of simulated data with up to 2000 PCs included as fixed effect show that the heritability estimates precision are not much impacted when an important number of PCs are included in the model with fixed effects. Of course, the size of the sample matters, however as soon as a few thousands individuals are included in the analysis, taking $p = 100$ or 500 is not harmful for the precision of the estimates. There is no practical reason to limit to small values of p.

We were surprised to obtain a genomic heritability estimate of 100% for latitude and longitude: as they are correlated to the first few PCs, we were expecting a positive heritability estimate, but not that large. We were also expecting the estimated heritability to vanish (or at least to become very small) after inclusion of a few PCs in the model. Instead of that, if the latitude heritability drops to 68% after inclusion of the first two PCs, adding more PCs only results in a slow decrease, and the heritability remains highly significant. The longitude behaves in a worse manner, as there is no initial drop as observed for the longitude. In both cases, the slow decrease of heritability is accompanied by a slow increase of the proportion of total variance due to the PCs included with fixed effects. The Likelihood Ratio Test (LRT) shows that the heritability is significantly positive for all values of p.

Browning *et al.*[16] obtained similar results with simulated case/control data based on the WTCCC case/control data, with an extremely disequilibrated ascertainment scheme in which 90% of individuals from Scotland and Wales were assigned to be cases, and only 10% of individuals from England, the controls being the remaining

Phenotype		LRT	*p*-value	$\widehat{\tau}$ (se)	$\widehat{\sigma^2}$ (se)	$\widehat{\sigma_P^2}$ (se)	$\widehat{h^2}$ (se)
Height	0 PC	68.80	5.6e-17	17.48 (0.719)	20.64 (0.689)	38.12 (0.687)	0.459 (0.061)
	1 PC	36.32	8.4e-10	14.70 (0.712)	23.12 (0.690)	37.83 (0.685)	0.389 (0.066)
	5 PCs	36.57	7.4e-10	14.73 (0.712)	23.03 (0.690)	37.76 (0.685)	0.390 (0.066)
	10 PCs	36.60	7.2e-10	14.77 (0.712)	23.00 (0.690)	37.78 (0.686)	0.391 (0.066)
	20 PCs	37.87	3.8e-10	15.10 (0.715)	22.75 (0.691)	37.85 (0.686)	0.399 (0.066)
Weight	0 PC	13.92	9.6e-5	28.24 (2.32)	97.23 (2.32)	125.47 (2.11)	0.225 (0.062)
	1 PC	13.34	1.3e-4	27.91 (2.32)	97.53 (2.32)	125.44 (2.12)	0.222 (0.062)
	5 PCs	12.50	2.0e-4	27.19 (2.32	98.21 (2.32)	125.40 (2.13)	0.217 (0.062)
	10 PCs	12.07	2.6e-4	26.77 (2.32)	98.59 (2.32)	125.36 (2.13)	0.214 (0.062)
	20 PCs	12.10	2.5e-4	26.91 (2.32)	98.56 (2.32)	125.47 (2.13)	0.214 (0.063)
BMI	0 PC	10.44	6.2e-4	3.23 (0.302)	13.09 (0.302)	16.31 (0.304)	0.198 (0.063)
	1 PC	9.26	1.2e-3	3.11 (0.302)	13.19 (0.302)	16.30 (0.304)	0.191 (0.064)
	5 PCs	8.56	1.7e-3	3.00 (0.301)	13.29 (0.303)	16.30 (0.304)	0.184 (0.064)
	10 PCs	8.47	1.8e-3	2.99 (0.302)	13.30 (0.303)	16.30 (0.304)	0.184 (0.064)
	20 PCs	8.52	1.8e-3	3.01 (0.302)	13.29 (0.303)	16.31 (0.305)	0.185 (0.064)
Head Circumference	0 PC	9.85	8.5e-4	0.722 (0.078)	3.52 (0.079)	4.24 (0.080)	0.170 (0.057)
	1 PC	2.79	0.047	0.442 (0.078)	3.78 (0.079)	4.23 (0.079)	0.104 (0.063)
	2 PCs	2.16	0.071	0.393 (0.078)	3.83 (0.079)	4.22 (0.079)	0.093 (0.064)
	5 PCs	1.88	0.085	0.368 (0.078)	3.85 (0.079)	4.22 (0.079)	0.087 (0.064)
	10 PCs	1.93	0.082	0.373 (0.078)	3.85 (0.079)	4.22 (0.079)	0.088 (0.064)
	20 PCs	1.92	0.083	0.372 (0.078)	3.85 (0.079)	4.22 (0.080)	0.088 (0.064)
Waist to Hip Ratio	0 PC	13.09	1.5e-4	9.2e-4 (8.6e-5)	3.6e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.205 (0.061)
	1 PC	4.29	0.019	6.0e-4 (8.6e-5)	3.9e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.135 (0.067)
	5 PCs	4.11	0.021	5.9e-4 (8.6e-5)	3.9e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.133 (0.067)
	10 PCs	4.28	0.019	6.1e-4 (8.6e-5)	3.9e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.136 (0.067)
	20 PCs	3.94	0.024	5.8e-4 (8.6e-5)	3.9e-3 (8.7e-5)	4.5e-3 (8.7e-5)	0.131 (0.066)

Table 5. Model parameter estimates for the anthropometric phenotypes and their standard error, depending on the number of PCs included in the model, likelihood ratio test statistics (LRT) to test significance of heritability and their *p*-values. $\widehat{\tau}$ is the estimated genetic variance, $\widehat{\sigma^2}$ the estimated residual variance, $\widehat{\sigma_P^2} = \widehat{\tau} + \widehat{\sigma^2}$ the estimated total variance, and $\widehat{h^2} = \widehat{\tau}/(\widehat{\tau} + \widehat{\sigma^2})$ estimated heritability.

individuals. In their response, Goddard *et al.*[15] pointed out that this was an extreme scenario; it is however realistic to imagine that a quantitative trait is under the influence of an environmental factor which varies with latitude or longitude. The heritability estimate of such a trait would be severely overestimated.

Heritability estimates obtained for anthropometric traits are globally compatible with the results from the literature. It is interesting to note that height and waist-to-hip ratio seem sensitive to population stratification, even if marginally. Head circumference is more affected, as its estimated heritability drops from 17% to 9% with the inclusion of the first two PCs, and the LRT is no longer significant. All these traits are weakly but significantly correlated with the geographical coordinates of the birth cities (cf Supplementary Table S5), but so is BMI which does not display this behaviour. We note that when including more PCs in the analysis of these traits, there is no linear trend comparable to the one observed in the case of the geographical coordinates. This can bring back some confidence in the method, which was dented by the previous analyses.

We performed additional analyses were we included the geographical coordinates as covariates when analysing height, head circumference, and waist-to-hip ratio (Supplementary Fig. S4 and Supplementary Table S1): the drop in heritability when the first PCs are added is no longer observed, which seems to indicate that the effect of the first PCs was due to a geographical stratification. Heritability of height drops to 37% (instead of 39%), which is a marginal change; as previously, heritability of head circumference is not significantly positive. Heritability of waist-to-hip ratio is estimated close to 8%, and is no longer significantly positive, while it was previously estimated to 13%, significant with $p = 2$: this falls in line with our previous findings on the fact that the inclusion of the first PCs may not be sufficient to fully correct for population stratification. One should consider seriously to include the geographical coordinates of the place of birth with fixed effect in the model, when they are available.

Epidemiologists use to take into account possible center effects by including indicator variables for the centers. We performed an additional analysis of height, head circumference and waist-to-hip ratio with fixed effects for the centers (Supplementary Fig. S5 and Supplementary Table S2). As observed for the previous analysis incorporating geographical coordinates, the drop in heritability after including the first PCs almost disappears (although not completely). The final value after inclusion of 10 PCs is however somewhat higher than the one obtained with the geographical coordinates. We also performed an analysis on the individuals from the largest center, Dijon, alone (Supplementary Fig. S6 and Supplementary Table S3). The drop in heritability after including the first PC is then quite noticeable. This discards the hypothesis that the impact of the PCs on the heritability estimates is due

to a center effect. Note however that it is impossible to tell whether the first PCs effect is due to some geographic environment, or to some genes with a north-south or west-east allelic gradient.

We made the choice of using LD-pruned data to compute the PCs included for population stratification. We could have used the whole data to this aim; this does not alter significantly our conclusions. The results of such analyses are displayed in the Supplementary Information 2.

A procedure has been proposed to detect the presence of cryptic relatedness and population stratification[35]. It consists in computing, in one hand, the 22 heritabilities due to each chromosome, one at time, and on the other hand, the same quantities by analysing all the chromosomes together. In absence of cryptic relatedness and of population stratification, these two estimates should be (roughly) the same. The procedure then regresses the difference between those two estimates on the length of the chromosomes. A significantly positive intercept is interpreted as due to the presence of cryptic relatedness, while a significantly positive slope is interpreted as due to population stratification. We applied this procedure on all real data (Supplementary Table S4). This procedure diagnoses well the presence of a population stratification for longitude, latitude, height, head circumference, and waist-to-hip ratio. However, for anthropometric traits, when one PC is added in the model, the procedure would conclude that the population stratification is correctly taken into account. Also, when analysing the latitude, after the inclusion of 10 PCs, the slope is only borderline significant. This suggests that heritability estimates of a trait depending on the latitude, for example through the sun exposure, could have an important bias which would be difficult to detect by this method.

A recent work argued that considering the genotype matrix as fixed in the linear mixed model, while in reality the genotypes are sampled from the population, is a cause of unreliability of genomic heritability estimates, and that this is aggravated by the presence of population stratification[36]; the conclusions of this work are debatable[37]. However, and more generally, the assumptions of the linear mixed model are unrealistic[38]. They certainly are not fulfilled when it comes to the geographical coordinates of the birth cities; we do not know to what extent they are more realistic for other traits. This alone should compel us to take any genomic heritability estimate with a grain of salt.

Methods

Estimation of the heritability. We consider the linear mixed model

$$Y = X\alpha + Zu + \varepsilon$$

with $Y \in \mathbb{R}^n$ a vector of phenotypes, $X \in \mathbb{R}^{n\times p}$ the matrix of covariates included with fixed effects, $\alpha \in \mathbb{R}^p$ the fixed effect vector, $Z \in \mathbb{R}^{n\times q}$ the standardized genotype matrix, $u \sim \mathcal{N}\left(0, \frac{\tau}{q}\mathbb{I}_q\right)$ the genetic effect vector ($u \in \mathbb{R}^q$) and $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbb{I}_n)$ the error vector.

The variance components τ and σ^2 are estimated by Restricted Maximum Likelihood (REML)[39–42]. We used the R package Gaston[43] and GCTA[44]. Example of commands are given in Supplementary Methods (Paragraph 1).

The heritability is usually estimated by $h^2 = \frac{\tau}{\tau+\sigma^2}$ – which ignores the variance of the phenotype which is explained by the covariates included in X. This is tantamount to defining the heritability as the proportion of genetic variance in the phenotype variance yet unexplained by known covariates, as in ref. 9. However, it is also possible to split the variance of Y in three parts: the variance explained by the covariates in X, the genetic variance τ, and the remaining variance σ^2.

Estimating the variance due to the covariates by the empirical variance of the components of $X\widehat{\beta}$ would result in upward biased estimates. We show in Supplementary Methods (Paragraph 2) how to estimate it without bias. We will report both h^2 and this decomposition of the variance in three components.

Correcting for population stratification. To correct for population stratification, we include the first p principal components of the kinship matrix estimated only on a submatrix Z_1 of Z, obtained by pruning SNPs from Z, to retain approximately 50000 SNPs in low linkage disequilibrium[45].

$$Y = X\alpha + \sum_{i=1}^{p} PC_i\beta_i + Zu + \varepsilon \tag{1}$$

with PC_i the ith principal component vector ($n \times 1$) and β_i the fixed effect of the ith PC.

We will use values of p varying from 0 to 2000 (for $n = 5793$).

The Three-City genomic data. The 3C study[22] is an ongoing French population-based longitudinal study which started in 1999. Participants were randomly selected from electoral rolls of the cities of Bordeaux, Dijon and Montpellier. To be eligible, they had to be aged 65 years or older, living in one of the recruitment cities, and not institutionalized. A total of 9294 individuals are included. Participants were genotyped with Illumina Human610-Quad BeadChip in the Centre National de Génotypage as described in ref. 46. The 3C data can be shared for an ancillary study, subject to approval by the 3C-Study Steering Committee (http://www.three-city-study.com/genetic-studies.php).

Quality Control was performed using PLINK[47] as described hereafter. Duplicate individuals and individuals with a discordance between genetic and clinical sex were discarded. In the sequel, only autosomal SNPs were considered. Individuals with more than 5% of missing genotypes, or with a proportion of heterozygous genotypes more than 3 standard deviations away from the observed mean were removed. Individuals with non-European ancestry were excluded using Principal Component Analysis on Hapmap individuals. Individuals known to be born outside mainland France were removed. To eliminate cryptic relatedness, we removed an individual from

all pairs with a genetic relatedness superior to 0.025. Finally, SNPs with a call-rate below 99% or Hardy-Weinberg threshold of 10^{-8} were removed.

After Quality Control, there are 5793 individuals (1499, 3676 and 618 from Bordeaux, Dijon and Montpellier respectively) and 509931 autosomal SNPs left.

The Principal Components for population stratification correction are computed on LD-pruned data, where we kept only SNPs with a minor allele frequency higher than 5%, and in mutual LD inferior to 0.1. We also removed SNPs in the long-range LD regions defined in ref. 48. The final LD-pruned dataset includes 49277 SNPs. The distribution of the first two PC depending of recruitment cities are given in Supplementary Fig. S1; the first PC correlates with the collection center. This observation is not surprising, as the geographical coordinates differ from one center to another.

Two types of traits are available; the longitude and the latitude of the birth cities which were retrieved from their zipcode (see first panel of Supplementary Fig. S1) and several anthropometric phenotypes: height, weight, BMI, waist-to-hip ratio and head circumference.

Simulated phenotypes. We simulated phenotypes according to the model (1), using the 3C genomic data: Z is the 5793×509931 matrix of standardized genotypes, and the PCs are computed on the LD-pruned data. The $p = 10$ first principal components are included. The coefficients $\beta_1, \ldots, \beta_p$ have been chosen such that each PC explains 2% of the phenotype variance, (thus, the 10 PCs together explain 20% of the variance). The remaining variance is equally distributed between genetic and environmental effects (thus, $\sigma^2 = \tau$, $h^2 = 0.5$, and the proportion of total variance is 40% for each).

References

1. Galton, F. *Hereditary genius* (Macmillan and Company, 1869).
2. Fisher, R. The correlation between relatives on the supposition of mendelian inheritance. *Trans. R. Soc. Edinb. Earth Sci.* **52,** 399–433 (1918).
3. Falconer, D. S. *Introduction to quantitative genetics* (Oliver & Boyd, 1960).
4. Kempthorne, O. & Osborne, R. H. The interpretation of twin data. *Am. J. Hum. Genet.* **13,** 320 (1961).
5. Scarr, S. Environmental bias in twin studies. *Eugenics Q.* **15,** 34–40 (1968).
6. Feero, W. G., Guttmacher, A. E. & Manolio, T. A. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* **363,** 166–176 (2010).
7. Donnelly, P. Progress and challenges in genome-wide association studies in humans. *Nature* **456,** 728–731 (2008).
8. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456,** 18–21 (2008).
9. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* **9,** 255–266 (2008).
10. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci.* **109,** 1193–1198 (2012).
11. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
12. Gusev, A. *et al.* Quantifying missing heritability at known gwas loci. *PLos Genet.* **9,** e1003993 (2013).
13. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88,** 294–305 (2011).
14. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42,** 565–569 (2010).
15. Goddard, M., Lee, H., Yang, J., Wray, N. & Visscher, P. Response to browning and browning. *Am. J. Hum. Genet.* **89,** 193–195 (2011).
16. Browning, S. R. & Browning, B. L. Population structure can inflate SNP-based heritability estimates. *Am. J. Hum. Genet.* **89,** 191–193 (2011).
17. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).
18. Zhang, Y. & Pan, W. Principal component regression and linear mixed model in association analysis of structured samples: competitors or complements? *Genet. Epidemiol.* **39,** 149–155 (2015).
19. Janss, L., de Los Campos, G., Sheehan, N. & Sorensen, D. Inferences from genomic models in stratified populations. *Genetics* **192,** 693–704 (2012).
20. Segura, V. *et al.* An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* **44,** 825–830 (2012).
21. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11,** 459–463 (2010).
22. Group, C. S. *et al.* Vascular factors and risk of dementia: design of the three-city study and baseline characteristics of the study population. *Neuroepidemiology* **22,** 316 (2003).
23. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456,** 98–101 (2008).
24. Galton, F. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland* **15,** 246–263 (1886).
25. Allen, H. L. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467,** 832–838 (2010).
26. Lanktree, M. B. *et al.* Meta-analysis of dense genecentric association studies reveals common and uncommon variants associated with height. *Am. J. Hum. Genet.* **88,** 6–18 (2011).
27. van der Valk, R. J. *et al.* A novel common variant in DCST2 is associated with length in early life and height in adulthood. *Hum. Mol. Genet.* **24,** 1155–1168 (2015).
28. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42,** 937–948 (2010).
29. Thorleifsson, G. *et al.* Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat. Genet.* **41,** 18–24 (2009).
30. Loos, R. J. Genetic determinants of common obesity and their value in prediction. *Best Pract. Res. Clin. Endocrinol. Metab.* **26,** 211–226 (2012).
31. Heid, I. M. *et al.* Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat. Genet.* **42,** 949–960 (2010).
32. Lindgren, C. M. *et al.* Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLos Genet.* **5,** e1000508 (2009).

33. Yoneyama, S. *et al.* Gene-centric meta-analyses for central adiposity traits in up to 57,412 individuals of european descent confirm known loci and reveal several novel associations. *Hum. Mol. Genet.* **23,** 2498–2510 (2014).
34. Stram, D. O. & Lee, J. W. Variance components testing in the longitudinal mixed effects model. *Biometrics* 1171–1177 (1994).
35. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43,** 519–525 (2011).
36. Kumar, S. K., Feldman, M. W., Rehkopf, D. H. & Tuljapurkar, S. Limitations of GCTA as a solution to the missing heritability problem. *Proc. Natl. Acad. Sci.* **113,** E61–E70 (2016).
37. Yang, J., Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Commentary on "Limitations of GCTA as a solution to the missing heritability problem". *bioRxiv* (Date of access: 04/04/2016) URL http://biorxiv.org/content/early/2016/01/20/036574 (2016).
38. Génin, E. & Clerget-Darpoux, F. The missing heritability paradigm: A dramatic resurgence of the GIGO Syndrome in genetics. *Hum. Hered.* **79,** 10–13 (2015).
39. Lee, S. H. & Van Der Werf, J. H. An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. *Genet. Sel. Evol.* **38,** 1–19 (2006).
40. Liu, D., Ghosh, D. & Lin, X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics* **9,** 292 (2008).
41. Gilmour, A. R., Thompson, R. & Cullis, B. R. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* 1440–1450 (1995).
42. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8,** 833–835 (2011).
43. Perdry, H. & Dandine-Roulland, C. Package R 'gaston', [version 1.4]. URL https://cran.r-project.org/web/packages/gaston/index.html (2015).
44. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).
45. Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nat. Protoc.* **5,** 1564–1573 (2010).
46. Lambert, J.-C. *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat. Genet.* **41,** 1094–1099 (2009).
47. Purcell, S. *et al.* Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
48. Price, A. L. *et al.* Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* **83,** 132 (2008).
49. Visscher, P. M. *et al.* Genome partitioning of genetic variation for height from 11,214 sibling pairs. *Am. J. Hum. Genet.* **81,** 1104–1110 (2007).
50. Macgregor, S., Cornes, B. K., Martin, N. G. & Visscher, P. M. Bias, precision and heritability of self-reported and clinically measured height in Australian twins. *Hum. Genet.* **120,** 571–580 (2006).
51. Silventoinen, K. *et al.* Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res.* **6,** 399–408 (2003).
52. Polderman, T. J. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* **47,** 702–709 (2015).
53. Chen, X. *et al.* Dominant genetic variation and missing heritability for human complex traits: Insights from twin versus genome-wide common snp models. *Am. J. Hum. Genet.* **97,** 708–714 (2015).
54. Visscher, P. M., Yang, J. & Goddard, M. E. A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang *et al. Twin Res. Hum. Genet.* **13,** 517–524 (2010).
55. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47,** 1114–1120 (2015).
56. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91,** 1011–1021 (2012).
57. Yang, J. *et al.* Genome-wide genetic homogeneity between sexes and populations for human height and body mass index. *Hum. Mol. Genet.* ddv443 (2015).
58. Elks, C. E. *et al.* Variability in the heritability of body mass index: a systematic review and meta-regression. *Front. Endocrinol.* **3,** 29 (2012).
59. Smit, D. J. *et al.* Heritability of head size in Dutch and Australian twin families at ages 0–50 years. *Twin Res. Hum. Genet.* **13,** 370–380 (2010).
60. Ermakov, S., Kobyliansky, E. & Livshits, G. Quantitative genetic study of head size related phenotypes in ethnically homogeneous chuvasha pedigrees. *Ann. Hum. Biol.* **32,** 585–598 (2005).

Acknowledgements

We thank the staff and participants of the 3C Study for their important contributions. The 3C Study is conducted under a partnership agreement between INSERM, Victor Segalen–Bordeaux II University and Sanofi-Aventis. The Fondation pour la Recherche Médicale funded the preparation and initiation of the study. The 3C Study is also supported by the Caisse Nationale Maladie des Travailleurs Salariés, Direction Générale de la Santé, Mutuelle Générale de l'Education Nationale (MGEN), Institut de la Longévité, Conseils Régionaux de Aquitaine et Bourgogne, Fondation de France and the French Ministry of Research–INSERM Programme Cohortes et Collections de Données Biologiques. This work was supported by the National Foundation for Alzheimer's Disease and Related Disorders, the Institut Pasteur de Lille, the Centre National de Génotypage and the LABEX (Laboratory of Excellence program investment for the future) DISTALZ - Development of Innovative Strategies for a Transdisciplinary approach to ALZheimer's disease.

Author Contributions

C.D.-R., C.B., E.G. and H.P. designed the study. S.D. and P.A. collected and managed 3C data. C.B. and C.D.-R. performed the quality control. C.D.-R. performed the analyses and prepared the tables and figures. C.D.-R. and H.P. wrote the manuscript. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Dandine-Roulland, C. *et al.* Accuracy of heritability estimations in presence of hidden population stratification. *Sci. Rep.* **6**, 26471; doi: 10.1038/srep26471 (2016).

Modélisation de la composante génétique des maladies humaines : Données Familiales et Modèles Mixtes

Mots Clés : Génétique Statistique, Modèles mixtes, Données Familiales, Données en population, Héritabilité

Le modèle linéaire mixte a été formalisé il y a plus de 60 ans. Celui-ci permet d'estimer un modèle avec des effets fixes équivalents à ceux du modèle linéaire classique et des effets aléatoires. Ce type de modélisation, d'abord utilisé en génétique animale, est depuis quelques années largement utilisé en génétique humaine. Les utilisations de ce modèle sont nombreuses. En effet, il peut être utilisé en étude de liaison, d'association, pour l'estimation de l'héritabilité ou encore dans la recherche d'empreinte parentale et peut s'adapter à des données familiales ou en population. Le but de mon doctorat est d'exploiter différentes méthodes basées sur les modèles mixtes d'abord sur des données génétiques en population puis sur des données génétiques familiales.
Dans un premier temps, nous explorons dans ce manuscrit la théorie des modèles linéaires mixtes et leur utilisation en génétique. Nous adaptons aussi certaines méthodes pour les appliquer à notre recherche. Ce travail a donné lieu au développement informatique d'un package R permettant d'utiliser ces modèles dans le cadre des études génétiques.
Dans un deuxième temps, nous utilisons les modèles linéaires mixtes pour l'estimation de l'héritabilité dans une étude en population française, l'étude Trois-Cités. Nous disposons dans cette étude des génotypes des tag-SNPs habituellement utilisés dans les études d'association ainsi que des lieux de naissance et de plusieurs traits anthropométriques quantitatifs tels que la taille. L'objectif est alors d'étudier la présence et la prise en compte dans l'analyse de stratification de population dans cette étude. Dans ce manuscrit, nous analysons les coordonnées géographiques des lieux de naissance. Nos résultats mettent en évidence la difficulté pour corriger correctement la stratification de population avec les méthodes classiques dans certains cas. Nous analysons ensuite les traits anthropométriques en particulier la taille dont nous estimons l'héritabilité à 39% dans la population de l'étude Trois-Cités.
Dans la dernière partie de ce manuscrit, nous nous concentrons sur les données familiales. Nous montrons le gain d'information que peut apporter ce type de données dans la recherche des variants causaux. Puis, nous explorons l'utilisation des modèles mixtes sur des données familiales en appliquant certaines des méthodes associées dans la recherche de signaux d'association pour la Sclérose en Plaques, une maladie auto-immune, en utilisant un échantillon d'une centaine de familles nucléaires avec au moins deux germains atteints. Nous avons alors mis en évidence l'inadéquation des méthodes classiques basées sur les modèles mixtes à ce type de données. Afin de mieux comprendre ce biais de sélection et de le corriger, plus d'investigations sont nécessaires.

Modelisation of genetic risk in human diseases : Family Data and Mixed Models

Keywords : Genetic Statistics, Mixed models, Familial Data, Population Data, Heritability

Linear mixed models have been formalized 60 years ago. These models allow to estimate fixed effects, as in the linear models, and random effects. First used in animal genetics, this type of modelling have been widely used in human genetics since a few years. Mixed models can be used in many genetic analysis ; linkage and association studies, heritability estimations and Parent-of Origin effects studies for population or familial data. My thesis' aim is to investigate mixed models based methods, for genetic data in population and, for familial genetic data.
In the first part of my thesis, we investigated the mixed model statistical theory and their multiple uses in human genetics. We also adapted methods for our own work. An R package have been created which permits to analyze genetic data in R environment with mixed models.
In a second part, we applied mixed models on Three-Cities data, a French longitudinal study, to estimate heritability of several traits. For this analysis, we have access to tag-SNPs typically used in genome-wide association studies, birthplaces and several anthropometric traits. The aim of our study is to analyze presence of population stratification and evaluate methods to correct it. In the one hand, we analyzed birthplace geographic coordinates and showed that the correction for population stratification by classical method is not sufficient in this case. In the other hand, we analyzed anthropometric traits, in particular the height for which we estimated heritability to 39% in Three-Cities study population.
In the last part, we focused on family data. In a first work, we exploited familial information in causal variant research. In a second work, we explored mixed models uses for familial data, in particular association study, on Multiple Sclerosis data. We showed that mixed model methods can not be used without taking account the ascertainment scheme : in our data, all families have at least two affected sibs. To understand and correct this phenomenon, more investigations are needed.