Evry Val d'Essonne University
Statistics and Genome Laboratory
Pharnext

PhD Thesis

presented by

Matthieu Bouaziz

for the diploma of Doctor in Applied Mathematics

Statistical methods to account for different sources of bias in Genome-Wide association studies

Defended on 22 November 2012 in front of

Reviewers:	Emmanuelle Génin (Dr)	Inserm
	David Balding (Pr)	UCL Genetics Institute
Examiners:	David-Alexandre Trégouet (Dr)	Inserm
	Jean-François Zagury (Pr)	Conservatoire National des Arts et Metiers
Supervisors:	Christophe Ambroise (Pr)	Évry Val d'Essonne University
	Mickaël Guedj (PhD)	Pharnext

Remerciements

Je tiens tout d'abord à adresser un grand merci à Christophe Ambroise et Mickaël Guedj, mes deux directeurs de thèse, pour m'avoir permis de la réaliser dans de bonnes conditions et pour avoir encadré et soutenu mes recherches durant ces trois années.

Je remercie Emmanuelle Génin, David Balding, David Tregouet et Jean-François Zagury pour avoir accepté de faire partie du jury ainsi que pour leurs remarques et leurs conseils qui ont été enrichissants pour la rédaction de ce manuscrit.

J'adresse mes remerciements à la direction de Pharnext pour m'avoir accueilli au sein de leur entreprise durant cette thèse.

J'aimerais chaleureusement remercier tous les collègues que j'ai côtoyé pour leur gentillesse et leur soutien durant ces trois années. Je remercie principalement Caroline, Marine, Fabrice, les biostats', Nicolas, Aude, Aurélie, Yannick, Esther, Patrick, Gilles, Pierre, Cyril et Michèle. Merci aussi à tous les autres collègues de Pharnext et du laboratoire, ceux qui sont partis et ceux qui sont restés.

Merci également à Flora Alarcon, François Coquet et Céline Vial pour m'avoir, à leur manière, montré la voie de la recherche.

Enfin je désire remercier ma famille, mes amis ainsi que tous ceux qui sont toujours restés à mes côtés.

Abstract

The study of high-throughput Molecular Biology offers unprecedented opportunities for understanding cell mechanisms and diseases. In particular, Genome-Wide association studies have become powerful tools to detect genetic variants associated with diseases. This PhD thesis focuses on several key aspects of the new computational and methodological problematics that have arisen with such research.

Notwithstanding the widespread use of Genome-Wide association studies, their results have been questioned, in part because of the bias induced by population stratification. Many strategies are available to account for population stratification but their performances differ according to numerous parameters. We propose a robust comparison study of these methods. Their advantages and limitations in different stratification scenarios are highlighted in order to propose practical guidelines to account for population stratification in Genome-Wide association studies.

We then focus on the inference of population structure that has many applications for genetic research. In addition to be used to account for population stratification in association studies, it can provide information for evolutionary and demographic studies. We present in this manuscript a new clustering algorithm called `SHIPS` (Spectral Hierarchical clustering for the Inference of Population Structure). This algorithm was applied to a set of simulated and real SNP datasets along with several of the mainly used algorithms in the field to propose a comparison of their performances.

Finally, the issue of multiple-testing in Genome-Wide association studies is discussed on several levels. We propose a general review of the multiple-testing corrections and discuss their validity for different study settings. We then focus on deriving gene-wise interpretation of the findings that corresponds to multiple-testing between dependent tests. We analyze the rationale of the methods designed to this end and then propose a comparison of the most used in practice in order to determine the best strategy to obtain valid gene-disease association measures.

Résumé

Les récentes avancées en Biologie Moléculaires ainsi que le développement des puces et techniques de séquençage ont mené les biologistes à considérer des études génétiques à grande échelle. En particulier, les études d'association *Genome – Wide* sont devenues un outil très performant pour détecter les variants génétiques associés aux maladies. Ce manuscrit de doctorat s'intéresse à plusieurs des aspects clés des nouvelles problématiques informatiques et statistiques qui ont émergé grâce à de telles recherches.

Les études d'association *Genome – Wide* sont très répandues cependant leurs résultats sont critiqués, en partie à cause du biais induit par la stratification des populations. De nombreuses stratégies existent pour prendre en compte la stratification des populations mais leurs performances varient. Nous proposons une étude de comparaison de ces méthodes. Leurs avantages et limites sont discutés en s'appuyant sur divers scénarios de structure des populations dans le but de proposer des conseils et indications pratiques pour prendre en compte la stratification dans les études d'association.

Nous nous intéressons ensuite à l'inférence de la structure des populations dans la recherche génétique. En plus de permettre la prise en compte de la stratification dans les études d'association, cela fournit des informations pour l'étude de l'évolution et de la démographie des populations. Nous avons développé au cours de cette thèse un nouvel algorithme appelé `SHIPS` (*Spectral Hierarchical clustering for the Inference of Population Structure*). Cet algorithme a été appliqué à un ensemble de jeux de données simulés et réels et comparé à de nombreux autres algorithmes utilisés en pratique.

Enfin, la question du test multiple dans les études d'association *Genome – Wide* est abordée à plusieurs niveaux. Nous proposons une présentation générale des méthodes de tests multiples et discutons leur validité pour différents schémas d'études. Nous nous concentrons ensuite sur l'obtention de résultats interprétables aux niveaux des gènes, ce qui correspond à une problématique de tests multiples avec des tests dépendants. Nous discutons et analysons les différentes approches dédiées à cette fin et proposons une comparaison des plus utilisées en pratique, afin de déterminer la meilleure stratégie pour obtenir des mesures d'association gène-maladie valides.

Contents

Preface

General background

Genetic disorders research uses the discovery of Genetics and Molecular Biology that decrypted the structure of the genetic information. The human genome is actually composed of several chromosomes that correspond to DNA sequences. In recent years, many advances have been made to identify the genes, that are portions of a DNA sequence, and understand their functions. The Human Genome Project highlighted that 99.9% of the information contained in the 20,000-25,000 genes that compose the genome is common for all individuals. It is then in the remaining 0.1% that is the key that differentiates individuals. As a matter of fact, certain portions of the genome are not identical from an individual to another and we call an allele a version of the genetic text of such a polymorphic portion.

In addition, research of Genetics and Genetic Epidemiology led to the differentiation of two types of genetic diseases. Monogenic diseases result from the modification of a single gene while complex (or multifactorial) diseases are the result of the combined effect of several genes and of the environment. The analysis of such complex diseases have led to the development of many new technological, computational and analytical methods. This has motivated approaches that aim to understand the underlying complex mechanisms of diseases, i.e. what are the genes, the proteins or the flawed signaling and metabolic pathways that intervene in the disease, in order to provide therapies. As a part of this process, the pharmaceutical industries, such as the French company **Pharnext**[1], aim to provide therapeutic solutions.

The classical R&D approaches to find therapeutic molecules are usually based on the "one drug, one disease" paradigm under which a single drug is used to treat a single yet often multifactorial disease. The novel strategy proposed by **Pharnext** is on the other hand based on pleotherapy. Pleotherapy aims to identify the best combination of active

[1] `http://www.pharnext.com/`

molecules in order to restore the molecular pathways perturbed in each disease and addresses the shortcomings of the standard R&D approach that has shown its limits in terms of efficacy and safety. It allows targeting several molecular 'nodes' in a disease-perturbed pathway and thus helps to increase the treatment efficacy and safety.

In this strategy, one milestone step, that is also part of the classical approaches in research of genetic diseases, is to identify genes, that are bound to have a role in the mechanism of a disease. This can be done by the use of Genome-Wide association studies that arose with the development of Genetic Epidemiology. These studies usually aim to screen large portions of the genome in order to detect genetic markers, and by extension genes, associated with diseases. More precisely, individuals affected by a disease (called cases) are compared to healthy individuals (called controls) in oder to detect genetic variations, i.e. alleles, that are significantly different between the two groups with regard to the disease. The results of such studies are used to complete the constitution of the Pharnext Genetic Association Database that is an important component to unravel diseases mechanisms.

With the recent improvement of high-throughput genotyping technologies, the usage of Genome-Wide association studies have become widespread in genetic research. These studies are however criticized as they involve complex settings and analytical methods that can lead to biased results. As a matter of fact, the high dimension of the genetic data, the simultaneous testing of many markers or the necessity to account for the complex genetic structure of human populations are, among others, tricky issues that have raised doubts about the relevance of these studies' findings.

The development of methods in Statistical Genetics is therefore very important to improve these studies, to ensure that they are correctly conducted and provide a proper interpretation of their findings. To this end, many research groups and laboratories in Genetics, Mathematics or Statistics therefore dedicate part of their work or collaborate to enhance the treatment of complex genetic data. For instance, the French laboratory **Statistics and Genome**[2] is a research unit in Statistics that focuses on networks, evolution and statistical methods for Genetics and Genomics.

This PhD was designed on the basis of a CIFRE[3] convention in collaboration between the company **Pharnext** and the **Statistics and Genome** laboratory and focuses on their practical research needs and methodological developments. We aim to provide practical indications and guidelines that answer the questions raised by the treatment and the analysis of complex genetic data, especially in the case of Genome-Wide association studies and to develop methods that allow to improve certain aspects of the genetic research.

[2] `http://stat.genopole.cnrs.fr/`

[3] CIFRE: Conventions Industrielles de Formation par la REcherche

The next section provides a brief overview of the problematics we focused on during this PhD and that we present in this manuscript.

Manuscript overview

This manuscript is organized in five chapters. Each main chapter begins with an introduction to precise the context and introduce the notions that are then developed and ends with a discussion of the results.

In a first introductory chapter we present the statistical and genetic notions that are necessary for a proper understanding of the rest of the manuscript. After an initiation to the statistical hypothesis-testing, we introduce several concepts of the Genetics, Population Genetics and Genetic Epidemiology fields. This is also the occasion to detail the notion of genetic diversity. We then present the Genome-Wide association studies by providing indications on each step of these studies and discussing the reasons that might affect their results. Finally, we propose a definition of population stratification and analyze its causes and effects on genetic association studies.

The second chapter of this manuscript focuses on accounting for population stratification in Genome-Wide association studies. Many strategies are available to account for population stratification but their performances differ according to numerous parameters. After a presentation of the classical association testing approaches and of those accounting for population stratification, we propose a robust comparison study of these methods. Their advantages and limitations in different stratification scenarios are highlighted in order to propose practical guidelines to account for population stratification in Genome-Wide association studies.

We then focus, in a third chapter, on the inference of population structure, which represents an important part of our PhD work. In addition to be used to account for population stratification in association studies, this inference can provide information for evolutionary and demographic studies. To this end, many algorithms have been proposed to cluster individuals into genetically homogeneous sub-populations. We first introduce and discuss some of these algorithms and then present a novel approach that we developed during this PhD. We therefore detail our clustering algorithm called Spectral Hierarchical clustering for the Inference of Population Structure (`SHIPS`), based on a divisive hierarchical clustering strategy allowing a progressive investigation of population structure. This method takes genetic data as input to cluster individuals into homogeneous sub-populations and estimates the optimal number of such sub-populations. `SHIPS` is then applied to a set of simulated and real SNP datasets along with several of the mainly used algorithms in the field to propose a comparison of their performances.

The fourth chapter of this manuscript is dedicated to the issue of multiple-testing in Genome-Wide association studies that is discussed on several levels. These studies aim to test many markers to detect associations and are therefore susceptible to be biased by the multiple-testing problem inherent to conducting numerous statistical tests. We propose a general review of the multiple-testing corrections and discuss their validity for different study settings.

We then look more in detail at one of the last steps of genetic association studies that is deriving gene-wise interpretation of the findings. This task demands the aggregation of the results obtained on sets of markers that are usually correlated and therefore corresponds to a specific case of multiple-testing with dependent tests. Many approaches have been developed to do so. We analyze these methods and their rationales and then propose a comparison of the most used in practice in order to determine the best strategy to obtain valid gene-disease association measures.

Finally, the last chapter corresponds to a conclusion of this manuscript. The first objective of this chapter is to review and summarize the main results presented in the manuscript. We then evoke the different perspectives that our PhD work has opened concerning the analysis of genetic data and principally the Genome-Wide association studies.

Chapter 1

Introduction

The first chapter of this manuscript introduces the statistical and biological notions necessary to the understanding of our work.

We first provide some statistical background knowledge about hypothesis-testing which is indispensable to the conduct of genetic studies.

We then define some genetic concepts such as the genome and its features. Concepts such as the mechanisms leading to diversity, the Hardy-Weinberg equilibrium or the linkage disequilibrium are presented. We then outline different fields of research in Genetics that are Population Genetics and Genetic Epidemiology. We introduce the basis of genetic studies which implies discussing the etiology of diseases, the types of genetic markers and the various possible study designs that one can encounter.

A third section is dedicated to the Genome-Wide association studies that are the main focus of this manuscript. The different steps of such studies are detailed: from the data collection and analysis to the assessment of the validity of the findings. Two important causes of bias in these association studies are highlighted: population stratification and multiple-testing issues. The analysis of these two points is part of the research conducted during this PhD and will be developed later in this manuscript.

In a last section, we present in more details the notion of population structure and population stratification. We explain its origin along with the different types of structures and determine how and why this phenomenon can induce a bias in Genome-Wide association studies.

1.1 Statistical concepts

This section introduces the notion of hypothesis-testing that is important for the understanding of the manuscript.

Definition

Hypothesis-testing is a decision procedure that uses statistical theory. It aims to determine the plausibility of a statement by analyzing how likely is a result, observed from certain data, to have occurred by chance alone or to validate the statement. First, hypotheses are made about the data and then a statistical procedure is applied to determine the probability that the result is due to chance alone.

Formally speaking, two hypotheses are considered. The hypothesis we are interested in, called the alternative hypothesis (H_1), and the hypothesis that we use to assess H_1 that is the null hypothesis (H_0). The rationale of hypothesis-testing is to assume that the data are drawn from H_0 and to evaluate how likely it is that the observed result would then occur. If it is not likely that such a result occurs under H_0 then the alternative H_1 is favored.

Usually, a test statistic ($\mathcal{S}$) is defined to assess these hypotheses. A way to consider how meaningful is $\mathcal{S}$ is to assess the probability that a particular value of this statistic ($\mathcal{S}^{obs}$) would occur by chance under the null hypothesis.

The null hypothesis is used to derive the null distribution of the test statistic. This distribution serves as a reference to describe the variability of $\mathcal{S}$ due to chance under H_0. The hypothesis-testing procedure compares the observed test statistic $\mathcal{S}^{obs}$ to the null distribution and computes a statistical confidence measure called p-value to summarize the result. The p-value is defined as the probability that a test statistic at least as large as the observed one would occur in data drawn according to H_0:

$$p\text{-value} = \mathbb{P}_{H_0}(\mathcal{S} \geq \mathcal{S}^{obs})$$

A small p-value indicates that the test statistic lies in the extremities of the null distribution which suggests that the null hypothesis does not accurately describe the observation.

Interpretation of a statistical test

In practice, determining whether an observed test statistic is statistically significant requires comparing its corresponding p-value to a confidence threshold (α) also known as level of significance. When the p-value is smaller than α, the null hypothesis is rejected with sufficient confidence. Conversely, if the p-value is above the threshold, the observation is not sufficiently inconsistent with the null hypothesis which is not rejected. The result is considered as non-significant.

Many studies use a threshold $\alpha = 5\%$, historically suggested by Fisher who argued that one should reject a null hypothesis when there is only 1 in 20 chance that it is true (Fisher 1925). It is however possible to select other thresholds as there are no particular

statistical reasons for using $\alpha = 5\%$. The choice of the significance threshold actually depends on the costs associated with false-positives and false-negatives, and these costs may differ from one experiment to another.

When a statistical test is performed, depending on whether the null hypothesis is true or false and whether the statistical test rejects or does not reject the null hypothesis, one of four outcomes will occur: **(i)** the procedure rejects a true null hypothesis (false-positive or type-I error), **(ii)** the procedure does not reject a true null hypothesis (true-negative), **(iii)** the procedure rejects a false null hypothesis (true-positive) and **(iv)** the procedure does not reject a false null hypothesis (false-negative or type-II error). The true state and the decision to accept or reject a null hypothesis are summarized in **Table 1.1**.

	H_0 is not rejected	**H_0 is rejected**
H_0 is true	true-negative $(1-\alpha)$	false-positive / Type-I error rate (α)
H_0 is false	false-negative / Type-II error rate (β)	true-positive / Power $(1-\beta)$

Table 1.1: **Outcomes of a statistical test.**

Defined that way, the p-value of a test appears simply as the probability of false-positive and the corresponding threshold α corresponds to the false-positive (or type-I error) rate.

Another value is presented in **Table 1.1** and corresponds to the type-II error rate β, that is the probability of false-negative. We usually consider $1-\beta$, the statistical power of the test instead. The statistical power corresponds to the probability of rejecting the null hypothesis when the alternative hypothesis is true. Formally speaking, it corresponds to $\mathbb{P}_{H_1}(\textit{rejecting } H_0)$ and is function of the test significance level α. The power of a statistical test is used to determine what is its potential to detect true-positives. As statistical power and significance level are dependent, a trade-off has to be found to properly manage false-positive and false-negative rates. This is a way of selecting the significance level relevant to the data studied and also to compare different approaches. Given a common significance threshold α, the best test at the level α is the one maximizing the power β. The power can be calculated exactly when the distributions are known, otherwise it requires estimation procedures. We will present in **Section 2.4.5** methods to estimate the power of a test.

1.2 Genetic concepts

1.2.1 Genome and genetic information

Genetics is a science studying heredity, that is the transmission of traits within a species from generations to generations. The first theories were proposed by Gregor Mendel in the mid-19th and determined that the transmission of traits was carried out through entities that we nowadays call **genes**. Consecutive research provided a more global understanding of genes and more generally of the genetic information. The entirety of the hereditary information of an individual is called the **genome**. It is composed of one or several **chromosomes** that are oriented sequences of 4 different molecules called nucleotides (or sometimes bases). The 4 nucleotides are the Adenine, the Thymine, the Cytosine and the Guanine and are usually represented by A, T, C or G. Such sequences of nucleotides are called **DNA** (Desoxyribo Nucleic Acid) sequences. A gene is actually a portion of a DNA sequence. **Figure 1.1** corresponds to a graphical representation of the genetic concepts presented here.

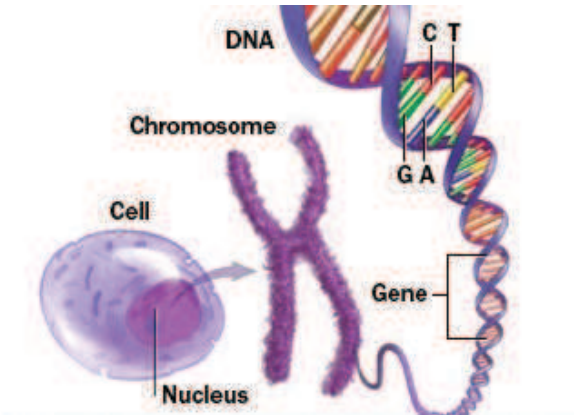

Figure 1.1: **Representation of the genetic information contained in one human cell.** This representation is the property of the Mayo Foundation for Medical Education and Research.

The alphabet {A, T, C, G} constituting the DNA is the same for all species. Also, it was discovered by J. Watson and F. Crick that the DNA is shaped as a double helix as shown in **Figure 1.1**. However the structure of the genetic information can vary between species. For example, a bacteria has only one chromosome whereas humans have 22 pairs of homologous chromosomes and 1 pair of sexual chromosomes.

Recent research initiated by the Human Genome Project[1] have led to the genotyping of the entire human genome and pointed out that there are about 20,000-25,000 genes

[1] `www.ornl.gov/hgmis/home.shtml`

present in the human DNA. These genes are more commonly called protein-coding genes. Indeed through a certain process the information contained in these genes is used to synthesize proteins that have a main role in the functioning of an organism.

Only less than 5% of the genome is constituted of genes. The rest of the genome is composed of sequences for which the function is known such as regulatory sequences for instance and a majority of sequences for which their roles are still unknown.

1.2.2 Genetic diversity and Population Genetics

The genetic diversity corresponds to the total amount of different genetic features of a species and is also called the gene pool of a species. As a matter of fact, within a species the genomes of all individuals are not identical. We call a **locus** a specified position on the genome and an **allele** a possible version of the genetic text at a given locus. We say that a locus is monomorphic when only one allele is possible (i.e. all the individuals share the same genetic text) and polymorphic when there are several possible alleles at the locus. A **haplotype** corresponds to a set of several alleles located on different loci of the same chromosome. In humans, for a given locus, each parent passes down one allele to the offspring. Each chromosome therefore carries two alleles at a given locus. We call **genotype** this combination of alleles at a locus. We also say that an individual is homozygous at the locus if the two alleles are the same and heterozygous otherwise.

The diversity of a gene pool can be assessed through several simple measures such as the proportion of polymorphic loci, the proportion of individuals carrying polymorphic loci (heterozygocity) or the number of alleles per loci.

Two main genetic mechanisms are responsible for the modification of the genetic text and therefore the existence of variety of different alleles: mutations and recombinations.

Mutation: a mutation corresponds to the sudden and spontaneous modification of a DNA sequence. It can for instance correspond to the modification of a base of a sequence, its deletion from the sequence or the insertion of a novel base in a the sequence (**Figure 1.2-A**). Mutations can occur because of external factors as well as being set off by the organism itself. A mutation can have various consequences depending on the sequence that has been altered. It can lead to no effect or to either a positive or negative effect on the organism depending on how the gene affected by the mutation is altered. The original function of the altered gene can be preserved (e.g. resulting in the same protein) or on the contrary the mutation can modify the gene function or regulations.

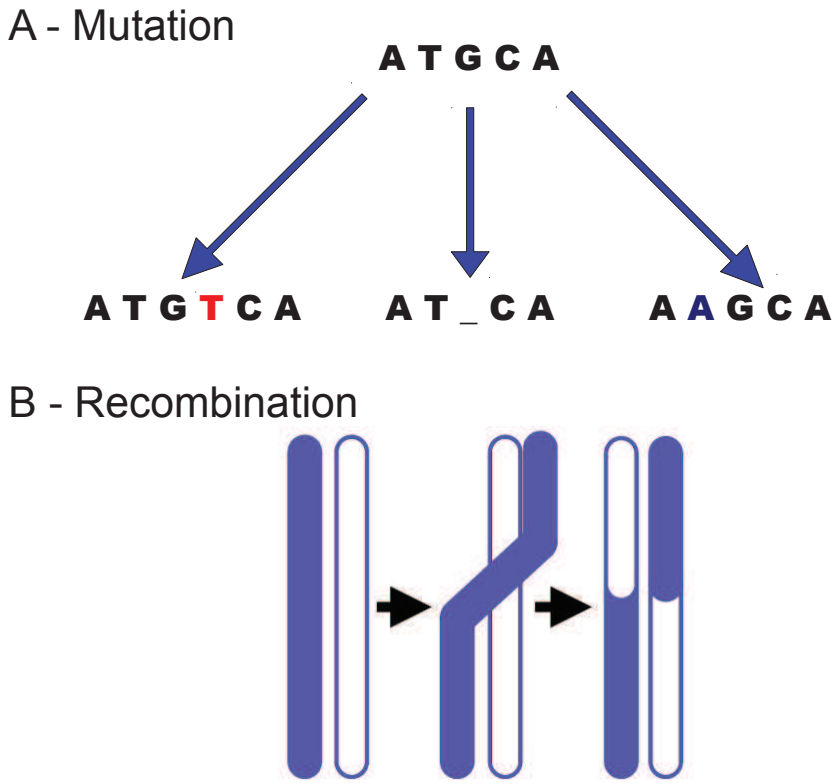

Figure 1.2: **A - Three mechanisms of mutation of a single base.** They correspond to the insertion of a novel base, the deletion of a base or the modification of a base. **B - Mechanism of recombination by crossing-over.** Two sequences are exchanged during a crossing-over.

Recombination: a recombination event refers to the exchange of nucleotide sequences between two DNA molecules. It is the case of the recombination between homologous chromosomes by crossing-over (**Figure 1.2-B**). Such a recombination generally occurs during the meiosis, that is the process leading to the formation of gametes. The probability of recombination between two loci is proportional to their distance on the sequences. As a consequence two genes distant on the DNA sequence are more likely to recombine. On the other hand, two genes located close to each other on a chromosome are likely to be inherited together. We say that these genes are linked and call this common inheritance the genetic linkage.

Population Genetics

Other mechanisms than mutations and recombinations can lead to diversify the gene pool of a species and ensure that this diversity in passed down through generations. Population Genetics studies these natural phenomena that are:

Natural selection: non-random process through which alleles (i.e. traits) are conserved or tend to be bred out from a population according to the advantages that they procure. Natural selection can for example be responsible for certain alleles to be passed down from generation to generation as they confer to their bearers advantages so they can live long enough to reproduce.

Genetic drift: mechanism explaining the frequency of alleles in populations due to chance. As a matter of fact, each parent has two possible alleles at a given locus and passes only one to its offspring. The alleles of an offspring are therefore a random sample of its parent alleles.

Population migration: this corresponds to the fact that individuals can migrate in (immigration) or out (emigration) of a population. These migrations modify the gene pool of populations. For instance after immigration, novel alleles can be introduced into a population and after emigration certain alleles can disappear from a population.

Mating process: certain species do not select their mates according to certain criterion which corresponds to a random mating. In such a situation alleles are randomly passed down to future generations. Other species have a selective mating process, causing certain traits to be favored and therefore more bound to be transmitted to next generations. Conversely, such a selective mating process can also lead to certain alleles being bred out from the gene pool.

Linkage Disequilibrium

Another important notion in Genetics is the notion of linkage disequilibrium (LD) that is not the same as the linkage introduced earlier. Linkage disequilibrium corresponds to the non-random association of certain alleles. Considering two alleles located at two different loci (not necessarily on the same chromosome), we say that these loci are in linkage disequilibrium if the probability of observing this particular combination of alleles, i.e. this haplotype, is not the same as the probability of this haplotype being randomly formed from the alleles based solely on their frequencies. In the opposite case, when the combination of two alleles is not more frequent than it would be under a random formation of the corresponding haplotype, we say that the two alleles are in linkage equilibrium.

The existence and level of linkage disequilibrium is influenced by the mechanisms previously described responsible for the genetic diversity.

Several measures exist to quantify the linkage disequilibrium. We consider two bi-allelic loci with respective alleles a/A and b/B. Let p_a, p_A, p_b and p_B be the frequencies of the different alleles and p_{ab}, p_{aB}, p_{AB} and p_{Ab} the frequencies of the four corresponding possible haplotypes. A first measure of LD is the linkage disequilibrium coefficient $\mathcal{D}$ that quantifies the deviation between the observed frequency of a haplotype from the expected if there was independence:

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

A second measure is the correlation coefficient that is a normalized version of $\mathcal{D}$ and is expressed as

$$r^2 = \frac{\mathcal{D}^2}{p_a p_A p_b p_B}.$$

This coefficient is comprised between 0 and 1 and is more interpretable than $\mathcal{D}$ which leads to a more common use in practice.

A null value of these indicators indicates that the two loci are in linkage equilibrium and any non null value indicates that there is linkage disequilibrium. The strength of LD can be quantified by the absolute value of these indicators.

Generally, linkage disequilibrium arises between loci located close to each other on the sequence and decreases as the distance between them increases. Important values of linkage disequilibrium have however been observed between distant loci ($>$ 500kb). Certain portions of the genome are in very high LD and are as a consequence passed down to future generations without any allelic alteration. We call such regions LD blocks. These blocks can be used to characterize the genome and their study is the source of many interest.

Wide international projects such as HapMap[2] and the 1000genomes[3] projects have described these LD patterns along the genome and provide the data to conduct further analyses. **Figure 1.3** presents an example of a LD pattern that is observable in the human genome.

In practice the calculation of the LD can be complicated as the allelic frequencies are usually known in a given sample but the corresponding haplotypic frequencies have to be estimated. Several algorithms, based on the Expectation-Maximization (EM) algorithm (Dempster et al. 1977, Excoffier and Slatkin 1995, Lou et al. 2003) or on Hidden Markov Models (HMM) (Stephens and Donnelly 2003, Delaneau et al. 2008), have been developed

[2] `http://www.hapmap.org`

[3] `http://www.1000genomes.org`

to conduct the estimation of these frequencies and therefore assess the measures that we have introduced. To study the linkage disequilibrium in very large genomic regions these approaches might turned out to be quite demanding in terms of time and resources. In order to have a global overview of the LD pattern a shortcut classically used is to consider the correlation between the genotypes at two loci as a measure of LD. This correlation is not the same as the correlation coefficient defined earlier but is quite similar and faster to compute. This is for example an option proposed by the software `plink`[4] to analyze genome-wide LD patterns. In addition, software such as `Haploview`[5] have been designed to calculate and provide graphical representations of the linkage disequilibrium.

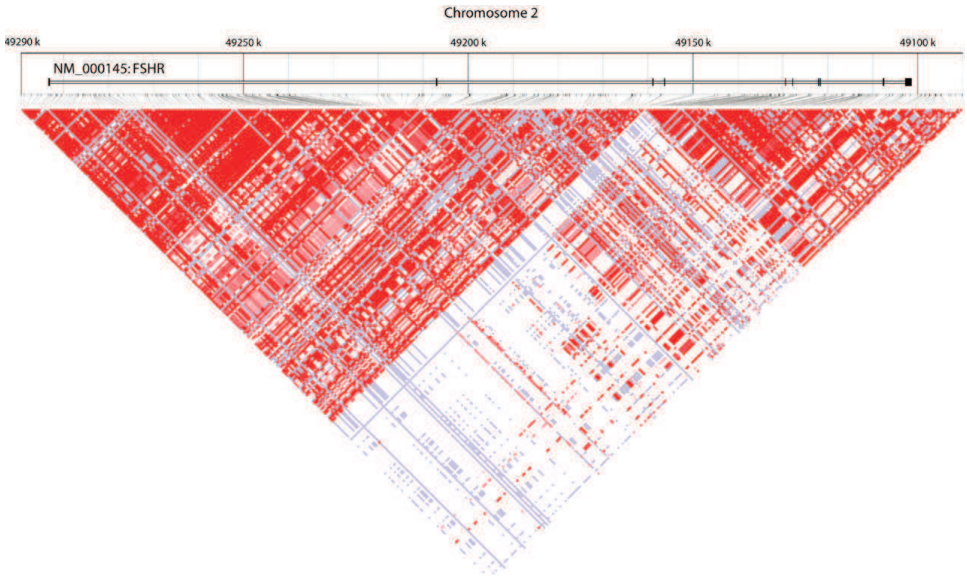

Figure 1.3: **LD pattern of a part of the chromosome 2 in humans.** The more red is a cell, the higher the linkage disequilibrium between the corresponding loci is (Simoni et al. 2008).

Hardy-Weinberg equilibrium

One of the main principle of Population Genetics is the Hardy-Weinberg equilibrium (HWE). This theory indicates that under certain assumptions, alleles and genotypes frequencies are constant. We say that these frequencies follow the Hardy-Weinberg equilibrium. The necessary assumptions to reach this equilibrium are (i) infinite population size to minimize the effect of the genetic drift, (ii) random mating process, meaning that

[4] `http://pngu.mgh.harvard.edu/~purcell/plink/`
[5] `www.broad.mit.edu/mpg/haploview/`

all possible mating are equiprobable, (iii) no mutations, no population migration and no natural selection to avoid any phenomena leading to the disappearance or conservation of certain alleles and (iv) the successive generations are discrete. Under these conditions it is possible to derive the genotypic frequencies directly from the allelic frequencies. Let us consider a bi-allelic locus with alleles a/A having frequencies p_a and $p_A = 1 - p_a$ and the possible genotypes (aa, aA, AA) with frequencies (p_0, p_1, p_2), then under the equilibrium we have:

$$\begin{cases} p_0 &= p_a^2 \\ p_1 &= 2p_a p_A \\ p_2 &= p_A^2 \end{cases} .$$

In real life, one or several of the assumptions that have been made are bound to be violated. The Hardy-Weinberg equilibrium actually defines an ideal state that is used to study the changes in the genotypic frequencies.

When an assumption is violated, one can observe a deviation from the Hardy-Weinberg proportions. According to the assumption concerned, the deviation differs. If the infinite population size or the random mating assumptions are violated then the Hardy-Weinberg proportions are no longer respected. On the other hand, if mutations, population migration or natural selection are in effect then the allelic frequencies change but the Hardy-Weinberg proportions may still be respected at each generation.

Wright proposed a model to specify the genotypic proportions when the equilibrium is no longer respected (Wright 1921). He introduced a consanguinity coefficient $\mathcal{F}$ and derived the new frequencies:

$$\begin{cases} p_0 &= p_a^2 + \mathcal{F} p_a p_A \\ p_1 &= 2p_a p_A - 2\mathcal{F} p_a p_A \\ p_2 &= p_A^2 + \mathcal{F} p_a p_A \end{cases} .$$

One can interpret this coefficient as indicating a deficit ($\mathcal{F} > 0$) or conversely an excess ($\mathcal{F} < 0$) of heterozygous individuals. When $\mathcal{F} = 0$, the population follows the Hardy-Weinberg equilibrium. **Figure 1.4** represents the genotypic frequencies at the equilibrium or with varying $\mathcal{F}$. As we will see in the rest of the manuscript this equilibrium is often assumed in genetic studies.

1.2.3 Genetic Epidemiology

Definition

Epidemiology is the study of health related events such as health patterns and their distributions, health characteristics or determinants influencing health. This science is nowadays used to understand and control diseases, identify therapeutic targets and define public health policies. Epidemiology does not aim to find causal relations between health

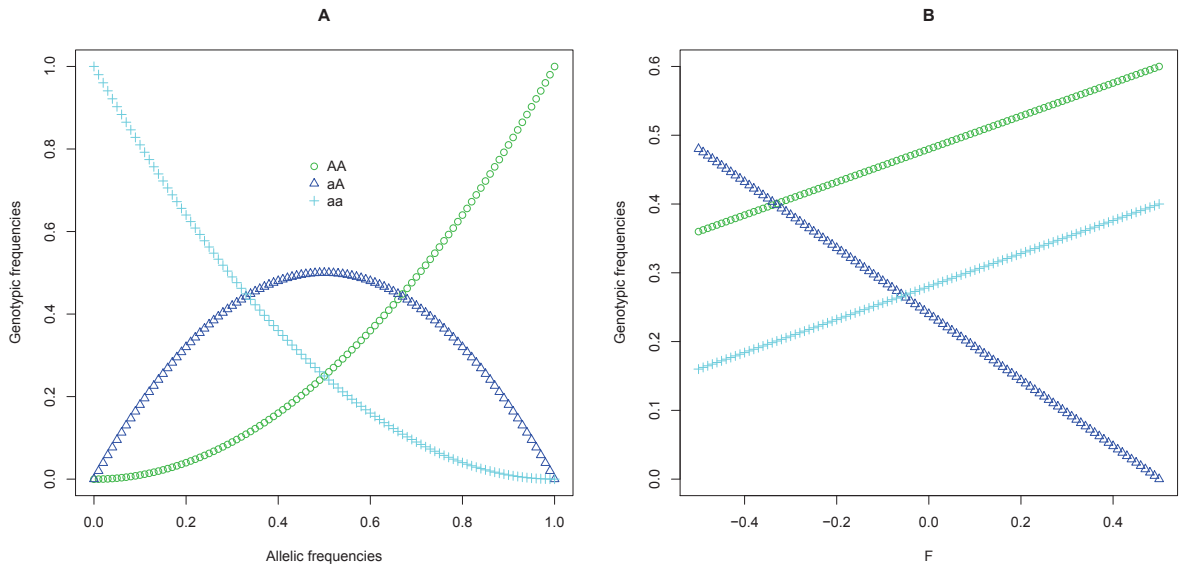

Figure 1.4: **A - Hardy-Weinberg proportions under the equilibrium. B - Deviation from the equilibrium in function of the consanguinity coefficient $\mathcal{F}$ ($p_a = 0.4$).**

events and certain determinants but assesses the likelihood of such relations in order to provide directions and guide the analyses and the treatment of these events.

Epidemiological investigations can concern the etiology (analytical studies), the outbreak (prospective studies) or surveillance (descriptive studies) of diseases.

Genetic Epidemiology combines classical Epidemiology and Genetics. It corresponds to the study of the inherited causes of diseases in families and in populations. A formal definition of Genetic Epidemiology was proposed by Morton: "a science which deals with the etiology, distribution, and control of disease in groups of relatives and with inherited causes of disease in populations".

The emergence of this science has become possible with the discoveries of Cellular and Molecular Biology. The understanding of the genome and of the different entities constituting the support of the genetic information such as the DNA, the chromosomes or the genes have created many research leads in the field of Genetic Epidemiology. Different types of genetic studies exist and can be classified in several categories according to the question they aim to answer.

Genetic risk studies or aggregation studies: is there a genetic component to the disease and what is the contribution of this component compared to that of the

environment?

Segregation studies: is the component formed of one or several genes? What is the mode of inheritance of the disease?

Linkage studies: what is the location(s) of the disease gene(s) or genomic region(s)?

Association studies: what is the allele(s) associated with the disease? We refer to these alleles as susceptibility alleles or disease susceptibility locus (DSL).

To achieve such research, the genetic epidemiological field benefits from the advances of technologies and the support of several international projects that aim to analyze the human genome. The Human Genome Project, the HapMap Project, dbSNP[6] or the 1000genomes Project have collected huge amount of genetic data, identified thousands of genes and genetic patterns that render possible the study of genetic disorders.

Disease etiology

The etiology of a disease represents all the causes it originated from. Genetic diseases can have two types of etiologies: monogenic (also called single-gene) or multifactorial (also called complex).

Monogenic diseases are the result of the modification of a single gene in the organism. These diseases are considered as rare even though they affect millions of people. To date, over 10,000 monogenic diseases have been identified such as for instance Sickle Cell Anemia, Haemophilia, Cystic Fibrosis or Huntington's disease.

Monogenic diseases follow the Mendelian laws of inheritance but it is not the case of all genetic disorders. Fisher first identified in 1918 diseases that have polygenic causes. We consider nowadays multifactorial diseases that are caused by the combined effect of several genes and environmental factors. The study of these diseases is a lot more complicated due to the nature and the multiplicity of the factors concerned. Genetic studies generally attempt to identify genes that are involved in such diseases and refer to them as susceptibility genes as their role and implication cannot be ascertained. Example of such diseases are Asthma, autoimmune diseases such as Type-1-Diabetes, cancers or heart diseases.

Genetic markers

Before the 80s, the markers considered in Genetics were the genes that encode for easily observable traits such as the blood type. Advances in Molecular Biology and the development of novel technologies to manipulate DNA have led to the identification of other

[6] http://www.ncbi.nlm.nih.gov/projects/SNP/

markers based on the variation of the DNA sequences. Many of these markers exist such as the Variable Number of Tandem Repeat (VNTRs or minisatellites), the Short Tandem Repeat (STR or microsatellites) or the Single Nucleotide Polymorphisms (SNPs).

The SNPs are among the most widely used markers. They correspond to the variation of a single base pair of a DNA sequence within a population. Most SNPs are bi-allelic, meaning that they involve two possible different alleles (**Figure 1.5**). A SNP is therefore usually identified by its locus and its two possible alleles. To date, more than 180M of SNPs have been identified and information about them are available in the dbSNP database. Also, several millions of these markers are available in large datasets such as those provided by the HapMap or the 1000genomes Projects. These markers can be located in the intergenic regions, in the non-coding regions of the genes or in the coding regions. These latter SNPs are very important as they can have a direct effect on the function of the resulting protein.

Due to the unique location of the DNA variation that represent the SNPs, they can arise in certain populations only. For this reason SNPs are very useful to differentiate and analyze different populations. These markers are also at the origin of the Genome-Wide association studies that we will present in **Section 1.3**.

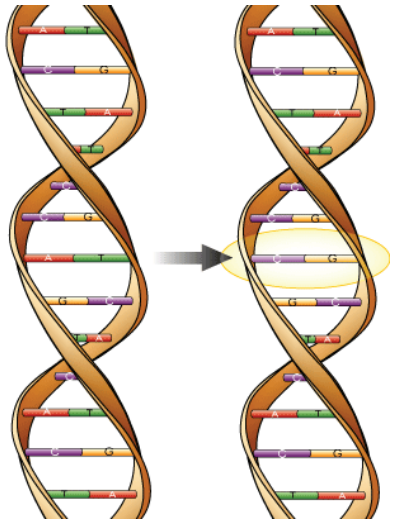

Figure 1.5: **Example of bi-allelic Single Nucleotide Polymorphism.** This representation was extracted from `http://compbio.pbworks.com/`.

Possible designs of the genetic studies

Genetics studies comprise two main types of study designs that are family-based studies and population-based studies.

The family-based study design corresponds to considering several families, trios (parents and one offspring) or brotherhood (without parents). A design including twins also exists to differentiate genetic effects from environmental effects.

As opposed to family-based studies, population-based studies refer to a design based on unrelated individuals. The most widely used is the case-control design. Considering a specific disease, a study based on such a design aims to identify genetic differences between a population of healthy individuals (controls) and a population of diseased individuals (cases).

Family-based study is the design for aggregation, segregation and linkage analyses while association studies can be based on both family or population-based designs.

In association studies, genetic markers are analyzed to determine which is the allele associated with the disease. In a family-based design, cases are considered along with their parents used as controls. Each parent passes one allele to the offspring which has the disease. A classical approach to determine the allele associated is to compare the proportions of alleles that have been inherited by the offspring to the proportions of alleles that have not been transmitted. A significant difference in these proportions can indicate a potential association. This method is formalized statistically under the name of Transmission Disequilibrium Test (TDT) (Spielman et al. 1993, Spielman and Ewens 1996).

In a population-based design, classical epidemiological approaches can be used; the genotypes of the cases are compared to those of the controls and significant differences point out whether the marker can be considered as associated with the disease. **Chapter 2** will explain more in details the strategies to assess the association between markers and diseases in case-control studies.

An important question when conducting an association analysis is which of the two designs is the best. A common bias in case-control studies can arise from an heterogeneity between the two groups. Not only these groups have to be matched according to certain features that could have an influence on the disease, such as the gender or the age, but in the situation of genetic studies, other factors have to be taken into account. We have briefly discussed the importance of the genetic diversity and the fact that certain populations can be genetically different. In case-control studies it is necessary to account for these differences, intrinsic to the population genomes, to avoid a certain bias. This phenomenon is called population stratification and will be largely discussed in **Sections 1.4 and 2.3**. Selecting parents as controls can therefore be seen as an alternative to avoid this bias. The recruitment of families is however more difficult to achieve than that of a control population, which usually leads to studies with small sample size. In addition, it is almost impossible to have access to the parents for late onset diseases. As a consequence, difficulties can arise in both designs to select the individuals included in a studies.

It has however been shown, on basis of large simulations, that the case-control ap-

proaches are more powerful than the family-based studies to identify marker-disease associations (Risch and Teng 1998, Teng and Risch 1999). Moreover important research has been conducted in recent years to improve the case-control design and take into account the different sources of errors that could weaken this approach. We will present such advances in **Chapter 2**.

From the candidate-gene to the genome-wide approach

A first approach for association studies is the candidate-gene approach. It consists in focusing on the association between a disease and a selected set of genes. These genes are usually of interest because of their potential roles in the etiology of the disease. A biological prior knowledge is therefore necessary to select these genes. For instance candidate genes can belong to a metabolic pathway known to intervene in the disease mechanism, be identified through protein-protein interactions knowledge or be located in a region that has already been pointed out by previous analyses. This information is available in public datasets (e.g. KEGG[7] for metabolic pathways and STRING[8] for protein-protein interactions) and can be consulted to identify target genes.

An alternative to the candidate-gene approach consists in screening a very large portion of the genome to search for markers associated with the disease. These Genome-Wide association studies assume no prior knowledge on which genes might be relevant. With the development of high-throughput genotyping technologies and the reduction of the genotyping costs such analyses have become more and more feasible and popular.

The choice of a strategy depends on the aim of the study and the resources at the disposal of the researchers. Genome-Wide scans are appealing especially for complex diseases that result from moderate to small effects of several genes.

1.3 Genome-Wide association studies

Genome-Wide association studies (GWASs) usually correspond to case-control studies, with unrelated individuals, that aim to scan a large portion of the genome to find associations between a disease of interest and genetic markers. We present in this section the major steps of such studies from the data collection to the analysis techniques.

[7] `http://www.genome.jp/kegg`
[8] `http://string-db.org/`

1.3.1 Data collection

Sample selection

The selection of samples participating to a genetic study is conducted among the general population of consenting individuals. The sample selection depends on the study design. As we discussed previously, selecting individuals is difficult whatever design is considered. A large number of samples allows a better statistical power to identify the genetic associations therefore it is advised to get cohorts as large as possible.

In addition to the number of individuals considered it is also very important to be sure of their states (case or control). This task can be more or less complicated according to the type of phenotype considered. Certain diagnoses are straightforward while others necessitate a more careful selection relying on precise criteria and demanding the intervention of medical specialists.

In a case-control design it is necessary that the two groups of patients are comparable. In order to ensure such a feature of the final cohort, individuals are included in a study according to certain characteristics such as the gender, the age or the ethnicity.

In definitive, the sample selection can take benefits from a medical point of view, to determine which sample is fit to be included in a cohort, but also from a statistical point of view. The selection of a proper amount of individuals, as well as the creation of homogeneous cohorts can be favored by the collaboration from the different researchers involved in a study.

Marker selection

A useful genetic marker corresponds either to an etiological locus, in which case the corresponding association is called a direct association (**Figure 1.6-A**) or is in linkage disequilibrium with one, which corresponds to a indirect association (**Figure 1.6-B**).

Genome-Wide association studies are usually conducted on SNP markers. A reason for this is the large amount of such markers available, the relatively low complexity and cost demanded to genotype them and the valuable information that they carry. These studies are based on the hypothesis known as 'common disease - common variant' which stipulates that common diseases, usually complex diseases, can be explained by the combined relatively small effects of many common variants, such as SNPs for instance. This principle highlights the necessity of disposing of a consistent and relevant set of markers.

Three main marker selection strategies can be highlighted that are (i) random selection, (ii) tagSNPs selection and (iii) gene-centered selection (Pettersson et al. 2009).

The random selection is usually adopted when no prior information is available about the disease and the potential regions of the genome that could intervene in its mechanism. With the development of high-throughput genotyping technologies, such selection can be

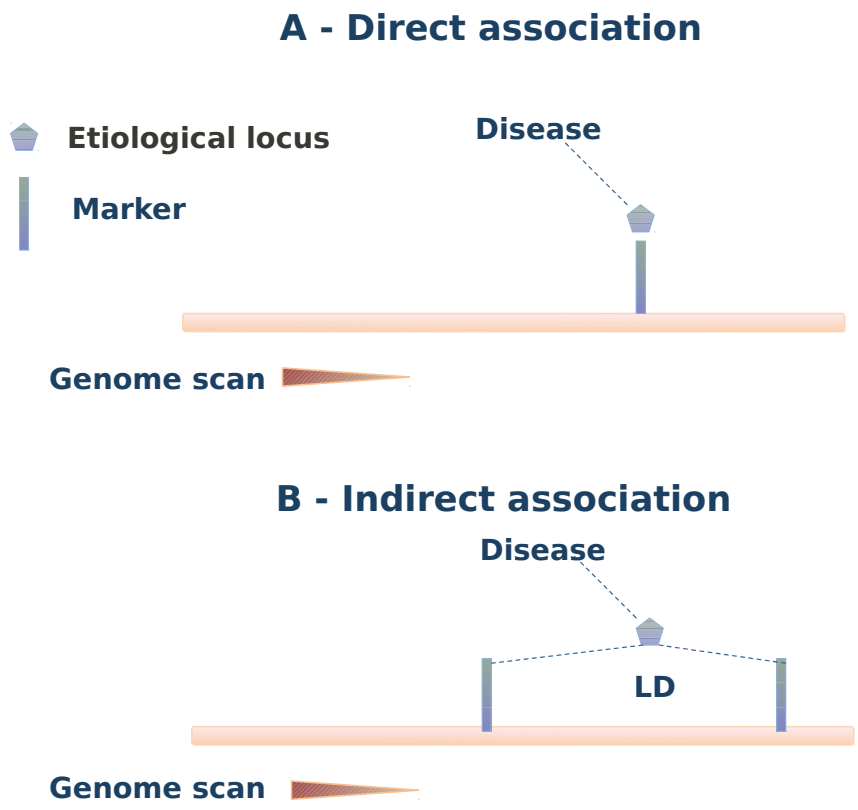

Figure 1.6: **A - Direct association:** the marker investigated corresponds to the etiological locus. **B - Indirect association:** the marker investigated is in LD with the etiological locus.

performed on a massive amount of markers. The type and cost of such technologies therefore greatly influence the selected sets of markers.

The tagSNPs selection is based on a certain type of SNPs. We have seen in a previous section that certain sets of SNPs in linkage disequilibrium form LD blocks. A tagSNP is a SNP among a LD block that carries the entirety of the genetic information of the block by itself. Thus, due to the block-like structure of the genome, the selection of such markers is sufficient to cover a large portion of the genetic information available (Balding 2006).

The gene-centered selection involves genotyping SNPs in proximity of the genes. For markers to be efficiently selected that way it is necessary to know all the genes of the genome. However, this selection process has the advantages of including biological a priori information indicating which gene might be involved in the disease etiology or was

previously highlighted in a former study.

All these marker selection strategies are questionable. While the random selection seems to cover the largest part of the genome, it is possible that certain interesting markers are missed due to the random sampling process utilized to compensate the lack of biological prior information. The tagSNPs selection has the disadvantage of focusing on SNPs belonging to LD blocks that might not be the only relevant ones. Finally, the gene-centered selection tends to select SNPs located within, or close to a gene but excludes the selection of SNPs that are located farther from the genes but may still have an effect on their roles, such as SNPs located in regulatory regions for instance. An exhaustive selection could encompass all aspects of the three strategies presented here. Note that such selection is becoming more and more possible with the development of the new generation sequencing.

SNPs selected for a study highly depend on the technology that is used and the markers that they can genotype. The *Illumina* and *Affymetrix* technologies are nowadays widely used to genotype SNP markers. These two technologies, based on different approaches, allow the genotyping of millions of SNPs that include tagSNPs and SNPs located within the genes or on other loci on the genome. In practice, it has become usual to use these chips to conduct association studies and the wide choice of markers they provide can be used as a global marker selection strategy.

One last point needs to be taken into account when selecting the markers included in a study. We have briefly evoked the problem of having cohorts of individuals genetically heterogeneous. For now, one just need to known that considering quite large sets of unlinked markers can be useful to tackle this problem. We will discuss more in detail why such sets of markers are of use to account for the genetic heterogeneity in the following.

1.3.2 Data analysis

Once the samples have been selected and genotyped for the set of markers investigated, the genetic data can be analyzed. Such analysis is usually composed of several steps. First, as usual when statistically analyzing data, a pre-processing step is necessary. Then, two strategies are available to identify the markers associated with the disease: the single-marker and the multi-marker approaches.

Pre-processing of the data

The first step of a GWAS when the data is available is a preliminary one that consists in formatting and ordering the data so that it can be properly analyzed. Given the large size of such datasets, disposing of clear and clean data files is very important to accelerate the analyses.

Also this pre-processing step focuses on determining the quality of the data (i.e. a

quality control). Certain features of the data have to be investigated in order to determine which samples or markers can reasonably be included in the analysis without leading to incoherent results.

The quality control of the markers can include:

SNP call rate: usually one assesses the amount of missing data for each marker. The SNP call rate is the proportion of genotypes per marker with non-missing data. When the call rate is too low, we might suspect a problem with the SNP genotyping step. Thus a threshold has to be set to remove these potentially poorly genotyped SNPs. Classically a threshold of 95% is used. However this threshold has to be set carefully because important markers could be spuriously removed. It is also important to determine if the missing data is missing at random, i.e. in approximately equal proportions between the cases and controls. Statistical tests are available to determine the type of missingness. If a marker contains missing data at random, one usually considers that it can be included in the study however the power to determine its association is decreased. In the case the data is not missing at random, usual approaches are to exclude the marker from the study or to impute the missing data using imputation algorithms. Example of such algorithms can be found in (Servin and Stephens 2007, Marchini et al. 2007, Li et al. 2010b).

Hardy-Weinberg equilibrium: a second feature of the markers that is verified is the Hardy-Weinberg equilibrium. In the case of a GWAS, it is reasonable to assume that the Hardy-Weinberg assumptions hold so that each marker follows the equilibrium. In the case a marker deviates from the equilibrium then it is possible that genotyping errors have been committed. A statistical test can be performed to determine which markers follow or not the equilibrium. The markers deviating from the equilibrium are usually excluded from the study.

Minor allele frequency: the minor allele frequency (MAF) of a marker represents the frequency of its less frequent allele in a given population. SNPs with low MAF have to be carefully examined. There are two reasons for this: (i) when a SNP is genotyped, the heterozygous and homozygous genotypes (Aa and AA/aa) have to be represented by the maximum of individuals to limit the genotyping errors. SNPs with low MAF can lead to low proportions of heterozygous or homozygous and therefore technical difficulties for the genotyping. (ii) SNPs with low MAF lead to statistical tests with low power to detect associations. The SNPs with very low MAF are generally removed from the study. Typically, a MAF threshold of 1-2% is applied in a lot of GWASs. Note that when the number of individuals is important, this threshold can be somehow less stringent.

In certain GWASs, and for the reasons explained above, a filtering on the number of heterozygous is conducted instead or in addition to the filtering on the MAF.

SNPs with very low MAF are also refer to as rare variant SNPs. An accurate genotyping of these markers has become possible with the new generation sequencing and particular statistical tests aim to assess their association with the diseases. These analyses are generally not part of a GWAS and constitute alternative studies.

Then, the quality control of the individuals includes:

Individual call rate: the individual call rate corresponds to the proportion of genotypes per individual with non-missing data and is controlled, like for the SNPs, at a certain threshold.

Identity by descent: analyzing the identity by descent allows to determine which samples are not independent (i.e. are related). Two or more alleles are Identical By Descent (IBD) if they are identical copies of the same ancestral allele. For each pair of individuals, the following probabilities can be estimated :

- Z_0: probability of sharing 0 allele IBD
- Z_1: probability of sharing 1 allele IBD
- Z_2: probability of sharing 2 alleles IBD

We define $\hat{\pi}$ the proportion of alleles shared IBD between a pair of individuals. This proportion indicates the relation between them:

$$\hat{\pi} = 0.5 \times Z_1 + 2 \times Z_2.$$

- $\hat{\pi} = 1$: sample duplicate or monozygotic twins
- $\hat{\pi} = 0.5$: 1^{st} degree relatives (full sibs, parent-offspring)
- $\hat{\pi} = 0.25$: 2^{nd} degree relatives (half-sibs, uncle/aunt-nephew/niece)
- $\hat{\pi} = 0.125$: 3^{rd} degree relatives (cousins, etc ...)

In a GWAS, samples have to be unrelated to ensure the validity of the association tests. Thus, individuals that are too close to each other have to be removed. A threshold is therefore applied to $\hat{\pi}$. Usually, when two individuals are related, the individual with the more missing data is removed. Furthermore, individuals that are linked with a lot of others are excluded in priority.

The IBD calculation is a very important step in a GWAS but can be quite time consuming. To that end, fast algorithms have been developed to calculate the IBD

matrices between pairs of individuals (Pong-Wong et al. 2001, Kong et al. 2008, Browning and Browning 2010).

Inbreeding coefficient: a feature of interest of the samples is the inbreeding coefficient (F). This coefficient represents the proportions of homozygous markers of an individuals. It is calculated for an individual by $F = \frac{O-E}{N-E}$, where O and E are the observed and expected numbers of homozygous genotypes for the individual and N is the number of markers considered for the calculation. Extreme values of this coefficient can indicate that the individual was not correctly genotyped.

Gender mismatch: the gender of the samples is sometimes available along with information about the sexual chromosomes. When discrepancies appear between these two information, the corresponding samples are treated with precaution as there could be other mistakes in their information.

Ancestry outliers: we have evoked that the cohorts of cases and controls have to be genetically homogeneous. Samples that appears as outliers need therefore to be identified. Methods to spot and deal with these individuals will be presented in **Chapter 2**.

Single-marker approach

An initial natural approach to determine the markers associated with the disease is to consider them one by one and to test them for association.

The main association tests based on the genotypes are the Armitage Trend test (Cochran 1954, Armitage 1955) that aims to find a linear trend between the probability of having the disease and the genotypes, the genotypic test that computes a Pearson χ^2 statistic or an exact Fisher test on the genotypic contingency table or the Hardy-Weinberg test (Nielsen et al. 1998) that considers a deviation from the equilibrium in the cases population as an indicator of association with the disease.

Other association tests are based on the alleles directly such as for instance the allelic test that is an analogue of the genotypic test on the allelic contingency table.

Some of these association tests allow to consider different modes of inheritance of the disease. These modes describe how the combination of the two alleles influence the probability of having the disease. The main modes of inheritance are additive, multiplicative, recessive or dominant. Considering the correct mode of inheritance of a marker is usually more appropriate to test for association, however this information is unknown. Generally

an additive mode is assumed.

The most used single-marker association tests will be detailed and discussed in **Chapter 2**.

Multi-marker approach

Complex diseases are generally due to the moderate or small effects of several markers. As a consequence the single-marker approach may not be enough to determine the genetic mechanisms intervening in the disease etiology. To this end, multi-marker analyses have been developed. These analyses consider sets of markers, either in close proximity on the genome or located at more distant positions, and focus on the combined effect of these markers.

We refer to as genetic interaction or epistasis when several markers have a combined effect on a disease. Considering all possible interactions of all possible markers is not feasible. Indeed the number of such possible interactions is way to important to conduct a complete investigation. Multi-marker analyses therefore attempt to 'smartly' scan the genome looking for significant sub-sets of markers that are associated with the disease. Several approaches are devoted to this task:

Small order epistasis: usually in Genome-Wide association studies, an analysis of the small order interactions (between two markers) is conducted. Certain statistical tests allow to consider more than one marker and to test for the effect of the interaction between two markers. Examples of such approaches are the Logistic Regression that will be presented in **Chapter 2** or statistical tests that compare the allelic association of two markers between cases and controls or within cases only (Hoh and Ott 2003, Zhao et al. 2006). The development of statistical procedures to analyze interactions between two loci is still a very live field in genetic research (Cordell 2009).

Haplotype analyses: a possible alternative to analyzing the genetic interactions is to consider haplotypes. Approaches have been designed to determine which haplotypes, that correspond to combinations of several SNP markers, increase the susceptibility to the disease. Some of these approaches are based on simultaneous estimations of the haplotype frequencies and association testing using EM algorithms (Zaykin et al. 2002) or stochastic EM algorithms (Tregouet et al. 2004). Others have been developed to identify haplotypes that show an excess of similarity in the controls compared to the cases (Tzeng et al. 2003, Beckmann et al. 2005).

Large order epistasis: the small order epistasis methods presented above cannot be applied to a large number of markers due to the computational cost of investigating higher

order interactions or of phasing large haplotypes. Therefore other strategies have been proposed to include a vast number of markers.

The SNP set based approaches consider sets of markers and identify the most relevant ones through the use of a statistic characterizing each set. The sets of SNPs investigated can be selected via the use of biological knowledge such as the proximity to genes or haplotype blocks (Wu et al. 2010b). Other methods also allow the SNP sets to be selected automatically such as with a moving window, according to their significance in the single marker analysis (Hoh et al. 2001) or to the global significance of the corresponding region (Guedj et al. 2006a).

Data mining approaches such as multifactor-dimensionality reduction (MDR) (Ritchie et al. 2001) are also quite popular to assess the significance of sets of markers. This method aims to pool multilocus genotypes into high-risk and low-risk groups and analyzes the datasets through cross-validation to calculate the prediction accuracy of the different sets.

The random forest method (Lunetta et al. 2004, Bureau et al. 2005) may be the method that allows the analysis of the larger amount of markers simultaneously. It performs random searches through the data by using bootstrap sampling. It generates multiple classification trees based on the markers and their ability to separate the samples in homogeneous groups. The whole set of tree is called a forest. The random forest method produces a score for each marker that measures its importance and allow the selection of several markers that have together a high capability to predict the disease.

1.3.3 Validity of the findings

Possible sources of bias and errors

Genome-Wide association studies can have several limitations that can affect the quality of the results, that are the number of false-positives or the power to detect associations. The limitations can be attenuated through a proper design of a study, a thorough quality control of the data and adequate statistical approaches for the analyses (Page et al. 2003, Pearson and Manolio 2008). Several causes of errors in GWASs can be highlighted.

Sample and marker features: the sample size is an important parameter as it directly influences the power of a study. The larger the sample size is, the greater is the power to finding association. In addition, a case-control design is based on unrelated individual, hence the necessity to control criterion such as the IBD. Genotyping errors in certain markers or samples can lead to erroneous results. The quality control of the data is therefore a crucial step in a study to detect those markers or samples that need to be put aside from the analysis.

Other features of the data can also have an effect on the power of a study. The power depends on the MAF of the markers as well as on the linkage disequilibrium between the

markers and the etiological loci in the case of indirect association (Zondervan and Cardon 2004).

Finally, the nature of the association between the marker and the disease also act on the capacity to detect association. Strong associations are more simple to uncover than weak associations. In addition, using a test based on the correct mode of inheritance usually leads to a better power.

Population stratification: when the samples are selected it is important to match certain features such as the gender or the age between cases and controls. It is also important to match the samples according to other factors leading to having genetically homogeneous groups. These factors are however unknown rendering the bias induced by the structure of population quite tricky to take into account. False-positive and false-negative associations can arise if the analyses are not carefully conducted to account for this bias. **Section 1.4** presents more in details the notion of population stratification and **Sections 2.3 and 2.4** examine solutions to take the genetic heterogeneity into account when testing for association.

Multiple-testing: whether one considers a single- or multi-marker analysis, when many statistical tests are conducted, the multiple-testing issue has to be taken into account. Many findings of GWASs are actually false-positive results due to this issue. **Section 4.1** will detail the problem and the solutions.

Replication

In a scientific process, the replication of the results is a necessary step to ensure their reliability. In Genetic Epidemiology, the replication of the results has two main objectives: confirming and providing more evidence in favor of some associations and ruling out associations that arose due to certain biases (Kraft et al. 2009).

Despite all the precautions that are taken when designing and conducting a GWAS, one cannot be sure of the pertinence of the results. These findings need to be replicated in order to provide additional evidence of their reliability. A replication is usually another study that can consider the same features than the original study or quite different ones. It is possible to consider the same markers, genotyping technologies or study designs or relatively different setups. For instance, a gene can be considered as replicated even if it was represented by different markers in different studies. In addition to being difficult to set up, the identical replication does not necessarily provide strong evidence. For instance if a bias was present in the initial study, it might then also arise in the replication due to the identical settings.

As a consequence, replication studies are preferable on similar but not identical features compared to the initial study. Therefore, the confirmation of a results can provide

evidence in favor of a finding. Nowadays, variants found in association with diseases have to be replicated in order to be published in most of the high-profile journals (Nature 1999).

Another advantage of replication is that it can extend the scope of the original findings. Replication through meta-analyses tend to reproduce the original findings on different populations. Such replications lead to generalizing results found on one population to a more general population.

Conversely, replication can have the opposite role. If a study based on a more robust design and analysis finds results in opposition with a primary study, for instance the non-replication of a marker, then the first findings can be discarded. Usually, in such situation, a careful analysis of the two studies reveals a bias that led to finding a spurious association in the first place.

The amount of genetic variants that have been replicated is still minimal (Gorroochurn et al. 2007). This enforces the reasons to develop adapted strategies to obviate the biases, induced by multiple-testing or population stratification for instance, that weight on GWASs results.

1.3.4 From the genetic markers to the genes

Genome-Wide association studies yield results at the SNP level, that are sets of SNPs associated with the disease. It is however very useful to obtain results at the gene-level, i.e. knowing which are the genes associated with the disease and the strength of the association. As a matter of fact the gene can be considered as the unit of interest in Genetics. It is for instance the genes that are considered to conduct gene-candidate study, replicate association studies, and perform pathways or gene network analyses. For this reason, a usual last step of a GWAS is the marker annotation. It corresponds to linking the markers to the genes by determining their positions on the genome compared to the genes (Cavalli-Sforza 1994).

As a result, all the genes included in a GWAS are represented by one or several markers, usually in linkage disequilibrium due to the close proximity on the genome. In order to associate to a gene a single significance measure of association it is necessary to use statistical techniques that properly combine the information of the different markers that represent it. **Chapter 4** is partly dedicated to the presentation and the analysis of these statistical methods.

1.4 Population structure and stratification

1.4.1 Origin of population structure

Population structure relates the genetic heterogeneity that exists between individuals of a population. This heterogeneity is a natural phenomenon resulting from biological and evolutionary processes. We presented in **Section 1.2.2** the different mechanisms responsible for differences in DNA sequences of individuals, i.e. mutations and recombinations, along with the processes leading to the genetic diversity, i.e. natural selection, genetic drift, population migration and mating processes. These phenomena lead in time to sub-populations genetically differing with regard to the frequency of certain alleles. For the same reasons, disease prevalences[9], allele penetrances[10] or LD patterns may vary between such groups (Cavalli-Sforza 1994).

As a result, more or less important systematic differences exist between sub-populations. The most important ones are found between ethnic and/or geographically distant groups. For instance, certain populations such as Caucasian, African and Asian ones were separated a long time ago and evolved separately so they were not touched by the exact same evolutionary processes.

1.4.2 Types of population structure

Human populations have been differentiating themselves since many generations. To relate the underlying population structures several models have been proposed, principally the **island model** (or K sub-populations model) and the **clinal model** (or isolation by distance model) (Astle and Balding 2009).

The island model is based on the rationale that a population can be decomposed in several sub-populations (also called islands). In this model, three types of structures can be highlighted: discrete structure, admixture and hierarchical structure (Li et al. 2010a).

A **discrete structure** pertains to a population formed of several discrete sub-populations. These sub-populations are discrete in the sense that they differentiated a very long time ago and that individuals tend to mate within sub-populations. For instance a population formed of Asian and African individuals.

A population is said to be an **admixture** when it is formed from the interbreeding of individuals originating from previously separated populations. As opposed to discrete structure, which corresponds to the already separated populations, we sometimes refer to admixture as continuous structure. African-American populations are examples of

[9] The total number of cases of a disease in a given population at a specific time.

[10] This is relatively to a disease, the proportions of individual carrying the alleles and having the disease.

admixed populations. Usually, the disappearance of an obstacle between populations, whether it is geographical or cultural, is a cause for the creation of admixed populations.

Finally, we call **hierarchical structure** a population that has a structure that comprises both discrete and admixed sub-populations. To a certain extent, it is possible to consider that most of the human populations can be decomposed into hierarchical structures. We therefore refer to as cryptic structure for such structure as it contains different patterns.

Figure 1.7 provides a graphical representation of these types of population structures. These representations are often used in genetic research and are based on a statistical method that is the principal component analysis thoroughly explained in **Chapter 2**.

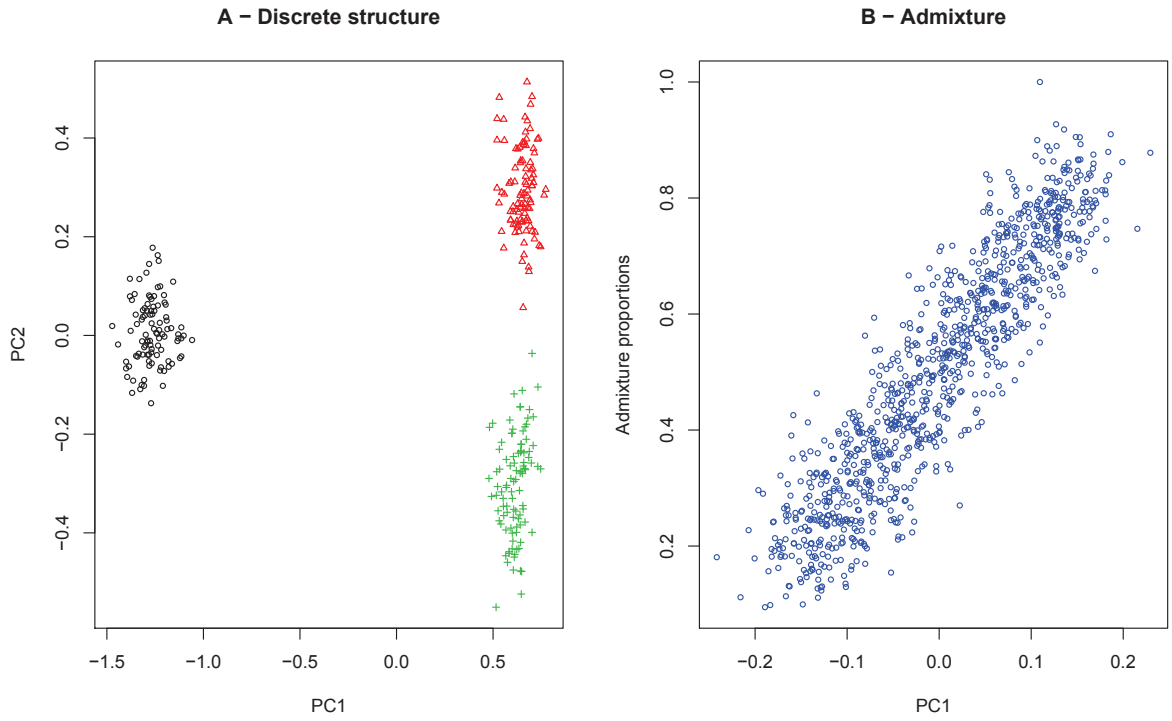

Figure 1.7: **Graphical representation of different types of population structure.** A - Discrete structure with 3 populations represented by the first 2 principal components. B - An admixture represented by the first principal component and the admixture proportions.

The clinal model does not assume the existence of sub-populations but considers that genetic variations are continuous and therefore exhibit a clinal pattern. In such a model, individuals tend to mate with those located in their vicinity. This model turns out to be a

good fit for certain populations such as European populations for instance. We can observe in **Figure 1.8** that these populations are not discrete and form a continuous gradient of genetic variations represented on the first two principal components. In addition, it is remarkable how the geographical structure of European populations is retrieved in this graph.

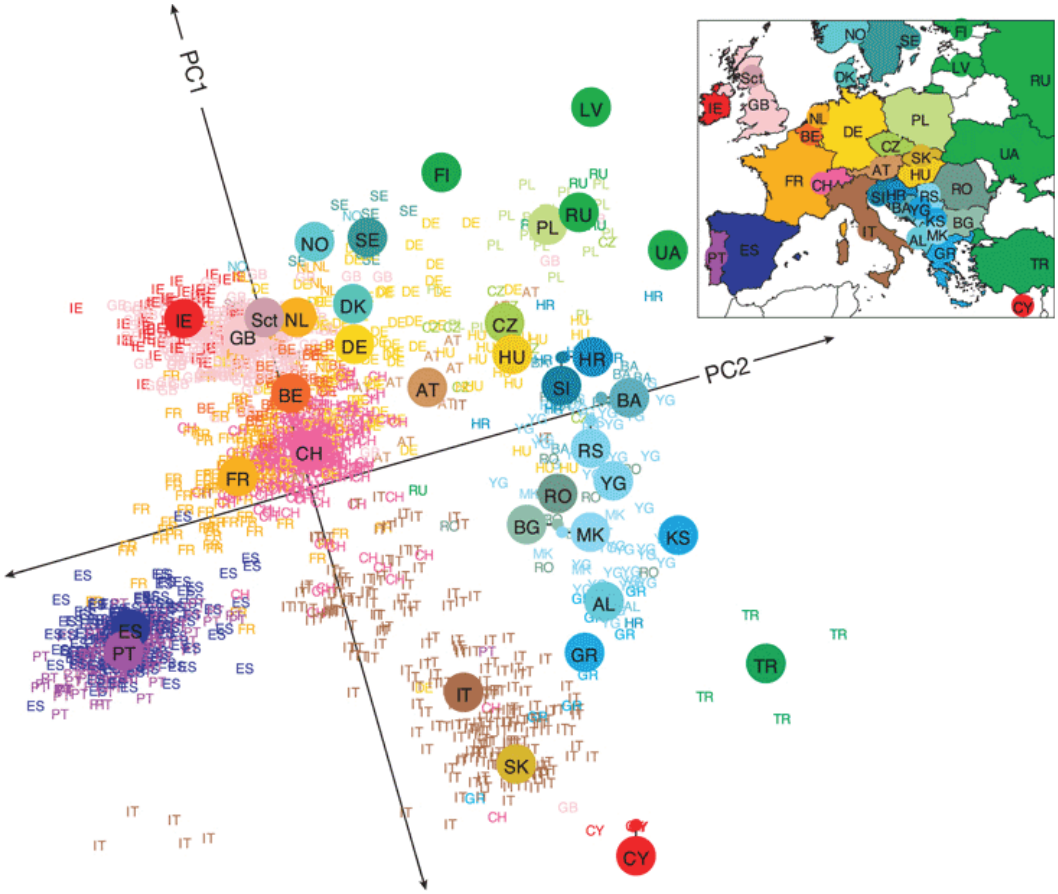

Figure 1.8: **Principal component analysis of European populations.** This plot is obtained from the first two principal components of 1,387 European individuals (Novembre et al. 2008).

We will see in the following that the type of population structure is very important as it pertains to different ancestral evolutionary processes.

1.4.3 Population stratification in GWASs

We usually refer to as population stratification when the population considered in a study is structured. We briefly indicated in previous sections that population stratification could

induce a bias in association studies. We detail in this section when and why there is actually a bias along with its effect on the findings of a study.

A confounding bias

In epidemiology we call a confounder a variable that is both correlated with the dependent variable and the outcome. The resulting bias that can induce such a variable is called a confounding bias. In association studies, the population membership[11] of the individuals can be a confounder of the relation one wish to analyze between a marker and a disease. We usually speak of population structure instead of population membership.

Population structure is therefore a confounder when it is both correlated with the disease (the phenotype) and the marker (the genotype). This means that both the links PS-genotype and PS-phenotype described in **Figure 1.9** exist.

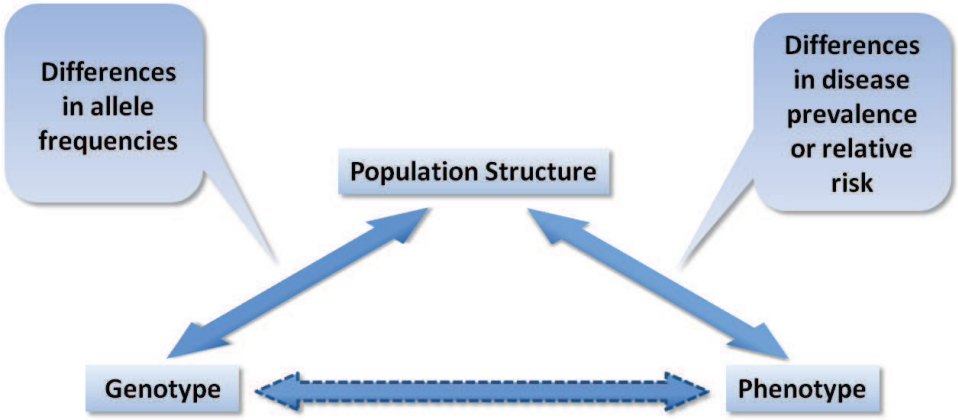

Figure 1.9: **Population structure acting as a confounder.** The association between the marker and the phenotype becomes a secondary association hidden behind the association through the confounder.

Let us consider a case-control association study with a structured population, a disease of interest and a marker investigated. There is a correlation between the population structure and the genotype because the genetic heterogeneity of the actual sub-populations leads to different allele frequencies and therefore different genotype frequencies between these sub-populations.

The link with the phenotype is due to both biological reasons and to the study design. Indeed, sub-populations may not be affected by the disease in the same way due to differences in the prevalence or in the risk of being affected. This is however not enough to

[11] i.e. the membership of each individual to its actual sub-population

create a correlation between the population structure and the phenotype. This correlation exist when, in addition, the sampling design is not appropriate, i.e. there is not the same proportion of cases and controls drawn from the sub-populations. **Figure 1.10** presents an example of a situation with two distinct sub-populations and a sampling such as there are more controls in the first one and more cases in the second one.

In such a situation, the genetic pattern of the individuals are more likely to be due to their belonging to one of the sub-populations than to their states: case or control. As a classical association test compares a marker between cases and controls, a bias can arise because the result of such a comparison may reflect the comparison of the marker between the two sub-populations and not between the two states.

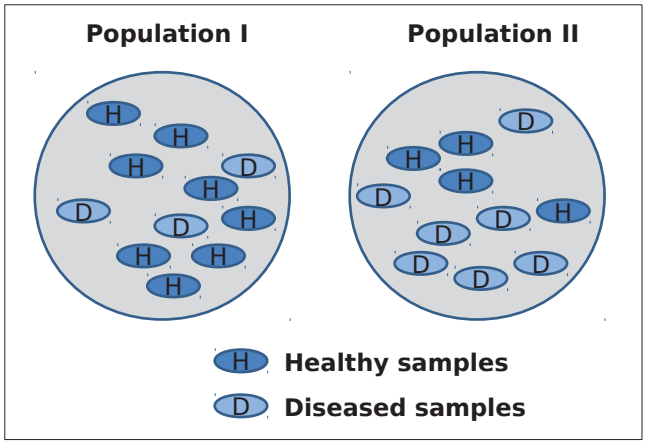

Figure 1.10: **Sampling design that would lead to a bias.** Such an imbalanced selection of cases and controls would associate the genetic pattern of the healthy individuals to population I and the genetic pattern of the diseased individuals to population II.

The case-control design is usually the more convenient to account for confounding. The usual techniques consist in matching cases and controls in appropriate proportions according to the strata defined by the confounding factors. Ideally, samples are selected within populations considered as homogeneous as possible. The issue of population stratification could then be resolved if it was possible to consider the exact same proportions

of cases and controls in each sub-population. This is however not possible in practice.

The very complex structure of human populations makes it impossible to determine truly genetically homogeneous groups of individuals and the knowledge of the actual sub-populations is not available. Even though usual study designs tend to consider samples from the same ethnicity and coming from close geographical locations, the cryptic structure of the populations may lead to the existence of sub-groups of individuals with differences in their genetic patterns. In addition, certain population structures such as admixtures do not correspond to discrete structures and render even more complicated the determination of sub-populations.

Moreover, with the recent availability of large cohorts of patients genotyped for many markers and many diseases, it becomes more and more usual to conduct association studies on structured populations.

Effect on association

Classically, to investigate the association between a marker and a disease one has to compare this marker between the cases and the controls. Because of population stratification, such a comparison may be biased and lead to finding a spurious association or to missing a genuine one (Deng 2001, Marchini et al. 2004, Freedman et al. 2004, Heiman et al. 2004).

Let us consider the situation of **Figure 1.10** where population stratification is in effect and assume that the marker investigated is not associated with the disease. A classical association test, e.g. the Trend test, compares the genotypes distribution between the cases and the controls. Because controls are mainly represented by the first sub-population, cases by the second one and due to the differences in genotypic proportions between the two sub-populations, the marker may appear to be differentially distributed between cases and controls. However, this distribution of genotypes is due to the genetic differences between the sub-populations not between the two status groups. This represents what we call a **spurious association**.

On the other hand if we assume that the marker investigated is associated with the disease, it is possible that the differences in genotype frequencies between the two sub-populations balance the differences due to the status and then hide the existing association.

A third situation can occur when the marker is associated with the disease and the differences in genotype frequencies due to stratification inflate those due to the association. In this case, the marker could appear to be strongly associated with the disease while it corresponds actually to a weak association.

The classical population-based association tests evoked in **Section 1.3.2** are vulnerable to population stratification. In presence of population structure their false-positive and false-negative rates are inflated.

1.4.4 Analysis of population structure

Identifying the underlying structure of populations is of great interest for the genetic research. It allows the study of evolutionary relationships between populations as well as learning about their demographic histories (Cavalli-Sforza 1994, Bowcock et al. 1994, Mountain and Cavalli-Sforza 1997, Pritchard et al. 2000a, Lee et al. 2009). In addition, it offers solutions to account for population stratification in association studies.

During the last decades, many statistical methods focusing on studying population structure and its consequences have been designed. Three main endgames of these methods can be outlined:

- Detecting whether a population is structured and gauging the complexity of the structure.
- Identifying the actual sub-populations and the relations of the individuals to these sub-populations.
- Accounting for population stratification in association studies.

We will present in **Chapters 2 and 3** these statistical approaches dealing with population structure. **Chapter 2** will focus on the statistical tests that allow to account for population stratification. **Chapter 3** will be dedicated to methods inferring the structure of the populations.

Chapter 2

Accounting for Population Stratification in Genome-Wide association studies

We presented in the Introduction the different steps of a Genome-Wide association study. We also emphasized the potential sources of errors that can affect the validity or the interpretability of the findings. We focus in this chapter on the bias induced by the structure of the populations. Considering a single-marker analysis strategy to look for association, we examine the approaches dealing with population stratification.

We first introduce the background of association testing, classical statistical tests to identify markers associated to diseases, along with some statistical background necessary to the understanding of more complex tests.

We then present the different strategies that exist to account for population stratification in Genome-Wide association studies. We also propose a comparison of these strategies based on a wide and realistic set of simulations covering various types of population structure. Our goal being to provide a practical answer to the question: what is the best approach to avoid the bias induced by population stratification in any genetic study? The results of this analysis have been published in PLOS ONE (Bouaziz et al. 2011).

2.1 Introduction

2.1.1 Genetic data

Genetic data available for association studies generally comprises different types of information concerning the individuals and the markers genotyped. The first information

corresponds to a table made of the genotypes of all the samples for all the markers. Information relative to the individuals are also available such as the status (case or control) and sometimes additional information such as the gender, the age or the ethnicity. The information about ethnicity is seldom available and can be more or less precise according to the database.

A last piece of information pertains to the markers and can include their possible alleles or their positions on the genome. We consider here the case of SNP markers.

These different information can be represented as tables. **Table 2.1** displays the information about the samples and the genotypes and **Table 2.2** concerns the information about the markers.

	SNP_1	SNP_2	...	SNP_p	Status	Gender	Age	Ethnicity
i_1	0	2	...	2	D	f	26	English
i_2	1	2	...	1	H	f	31	English
$\vdots$								
i_n	1	0	...	1	H	m	42	German

Table 2.1: **Samples information.** D = diseased, H = healthy, f = female, m = male.

	Name (rsID)	Chr	Position
M_1	rs123456	1	1234555
M_2	rs234567	1	1237793
$\vdots$			
M_p	rs234567	1	1237697

Table 2.2: **Markers information.** Each SNP marker has a unique rsID as identifier in international databases.

For certain of these variables, it is necessary to provide more formal statistical notations that we will conserve in the following of this manuscript. We call

$$X = (x_{ij})_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant j \leqslant p}}$$

the $n \times p$ genotype matrix for the n individuals and the p markers. Each term x_{ij} corresponds to the genotype of sample i for the marker j and is coded 0, 1 or 2 according to the number of variant alleles. For a SNP with alleles a and A where A is the variant, the possible genotypes are aa (coded 0), aA or Aa (indistinguishable and coded 1) and

AA (coded 2). This coding of SNP data is one of the most used in practice. In certain settings it corresponds to the coding for an additive model, that is usually assumed. Also it has the advantage of being interpretable in a qualitative manner, comparing the different genotypes to the reference genotype, or in a quantitative manner by measuring the amount of reference allele.

The phenotype vector is denoted by

$$Y = (y_1, \ldots, y_n),$$

where y_i is the status of sample i coded 0 (healthy) or 1 (diseased).

A specificity of genetic data is that usually the number of markers is greater than the number of individuals, $p \gg n$. This high dimension of the data requires the use certain specific statistical techniques.

2.1.2 Measures of association

When testing for association it is both important to statistically determine if the association is significant and to assess its degree. The statistical tests are used to provide evidence of the association and measures of association are used to quantify it. We present two of these measures that are the relative risk (RR) and the odds-ratio (OR). These two measures compare the probability of being affected by a disease in function of the exposure to a certain risk factor. This factor can have several levels and a reference level that corresponds to no exposure. In the case of genetic association, the risk factor can be the marker, and the different levels the genotypes aa, aA/Aa and AA with A being the variant allele. To introduce the relative risk and the odds-ratio we consider as a risk factor a marker x with the three genotypes coded 0 (the reference), 1 or 2 as different levels. We call $p_{x=i}$ the probability of being affected by the disease with the genotype i.

The relative risk is the risk of being affected by the disease relative to the exposure. It corresponds to the ratio of the probabilities of having the disease for two levels of the factor. The relative risk of the genotype i compared to the genotype 0 is

$$RR_{i/0} = \frac{p_{x=i}}{p_{x=0}}.$$

The interpretation of the relative risk is as follows. If $RR_{i/0} = 2$ then the individuals with genotype i are two times more likely than those with genotype 0 of being affected by the disease. More generally, if $RR_{i/0} = 1$ then there is no difference between the genotypes 0 and i, if $RR_{i/0} > 1$ then it is more likely to be affected by the disease with the genotype i than with the genotype 0 and if $RR_{i/0} < 1$ it is the opposite.

	Cases	Controls
$x = i$	a	b
$x = 0$	c	d

Table 2.3: **Contingency table of a risk factor compared to the disease.** The quantity a, b, c and d are the individual counts for each combination status/factor.

Using **Table 2.3**, the relative risk can be estimated by

$$RR_{i/0} = \frac{a/(a+b)}{c/(c+d)}.$$

The odds-ratio is defined as the ratio of the odds. An odd reflects the likelihood of being affected by a disease given a certain exposure. It corresponds to the ratio of the probability of being affected by the disease with a certain exposure and the probability of not being affected with this same exposure.

$$O_{x=i} = \frac{p_{x=i}}{1 - p_{x=i}}.$$

The odds-ratio of being affected by the disease is

$$OR_{i/0} = \frac{O_{x=i}}{0_{x=0}} = \frac{p_{x=i}(1 - p_{x=0})}{(1 - p_{x=i})p_{x=0}}.$$

Odds-ratios are always positive. If $OR_{i/0} = 1$ then there is not more risk of being affected with the genotype i than with the genotype 0. If $OR_{i/0} > 1$ then there is an increased risk with the genotype i compared to the genotype 0 and if $OR_{i/0} < 1$ it is the opposite.

An advantage of the odds-ratio compared to the relative risk is that it treats the two variables (status and exposure) symmetrically. This means that instead of considering the probability of being affected by the disease given the genotype, the odds-ratio can be defined using the probability of having a certain genotype given the status.

Using the **Table 2.3**, the odds-ratio can be estimated by

$$OR_{i/0} = \frac{ad}{bc}.$$

Odds-ratios are preferably used in genetic studies because they can be easily estimated. Indeed, it is also possible to estimate odds-ratios using logistic regression as we will present in the next section.

In addition, their interpretation is intimately linked to that of the association test. Determining that a marker is not associated with the disease, i.e. the genotype and the

status are independent, is equivalent to having all possible odds-ratios between the disease and the different genotypes being equal to 1.

2.1.3 Linear and logistic regressions

Regression analyses are methods that model the relationship between a dependent variable y (also called outcome) and one or several independent variables $x_1, \ldots, x_p$. Regression models estimate the conditional expectation of the dependent variable given the independent variables $\mathbb{E}(y \mid x_1, \ldots x_p)$. A regression function is used to model this expectation

$$\mathbb{E}(y \mid x_1, \ldots x_p) = f(x_1, \ldots, x_p, \boldsymbol{\beta}),$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters (or coefficients) that is estimated and quantifies the effect of each variable on the outcome. In most of the situations, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ as there is one parameter per variable.

A regression model can be re-written as

$$y = f(x_1, \ldots, x_p, \boldsymbol{\beta}) + \varepsilon,$$

where ε represent a noise also called error.

A regression analysis is based on a sample of independent individuals, for which the different variables have been observed, and that is used to estimate the unknown parameters. Classical assumptions of regression models are that (i) the errors are uncorrelated, (ii) they have means of zero conditional on the independent variables ($\mathbb{E}(\varepsilon \mid x_1, \ldots x_p) = 0$) and (iii) they have constant variances across observations ($\mathbb{V}(\varepsilon) = \sigma^2$). Also these models assume that (iv) the independent variables are indeed independent from each other and (v) they are observed without errors.

We present in this section the linear and logistic models that are used in association studies to analyze the relationship between a phenotype (y) and a marker (x).

Simple linear regression

Linear regression models assume that the function f correspond to a linear function.

$$y = \alpha + \beta x + \varepsilon$$

The parameter α corresponds to the constant independent variable also called the intercept. This model is used when the outcome is a continuous variable.

The estimation of the parameters can be conducted using the ordinary least square or the maximum likelihood methods.

It is possible to assess the significance of the parameter β, and therefore determine whether or not the corresponding variable x explains the outcome, by assuming that the errors are normally distributed and testing the hypothesis

$$H_0 : \{\beta = 0\}.$$

Usually a t-test is conducted to assess this hypothesis.

Simple logistic regression

Principle

The logistic regression is used when the outcome is a categorical variable. This regression model is therefore more suitable for association testing as the phenotype is a discrete variable. The expectation of the phenotype conditional on the genotype ($\mathbb{E}(y \mid x)$) is actually equivalent to the probability of being affected by the disease conditional on the genotype ($\mathbb{P}(Y = 1 \mid x) = p_x$). The logistic function is used to model this probability.

$$p_x = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}},$$

which is equivalent to

$$logit(p_x) = \alpha + \beta x.$$

This model leads to values of p_x always comprised between 0 and 1 which is valid for a probability.

The logistic model can also be expressed in terms of odds

$$log(O_x) = \alpha + \beta x.$$

The parameters of the logistic model can be interpreted using the odds-ratio. If we consider two genotype values, x_A and x_B, that x can take then

$$OR_{x_A/x_B} = e^{\beta(x_A - x_B)},$$

$$log(OR_{x_A/x_B}) = \beta(x_A - x_B).$$

The advantage of the logistic regression is that it therefore allows the estimation of the odds-ratio through the parameters. These parameters can be estimated using a maximum likelihood method.

Significance of the parameter
Two methods can be used to test the significance of the parameter which corresponds to the null hypothesis

$$H_0 : \{\beta = 0\}.$$

The likelihood ratio test compares the likelihood of the null model with only the intercept and an alternative model that also includes the genotype. The statistic is

$$D = -2log(\frac{\text{likelihood of the null model}}{\text{likelihood of the alternative model}}),$$

and follows a $\chi^2(df_2 - df_1)$ distribution under the null where df_1 and df_2 are the degrees of freedom of each model.

Another approach to test one coefficient is the Wald statistic

$$Z = \frac{\hat{\beta}}{\sqrt{\mathbb{V}(\hat{\beta})}},$$

that follows a standard normal distribution $\mathcal{N}(0, 1)$ under H_0.

Extension to multiple variables
It is possible to include several variables $X = (x_1, \ldots, x_p)$ into the logistic model that generally becomes

$$logit(\mathbb{P}(Y = 1 \mid x_1, \ldots, x_p)) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j,$$

where β_0 corresponds here to the intercept and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is the parameter to be estimated.

The interpretation in terms of odds-ratio can still be applied except that in such a model these measures are calculated for a certain variable x_j considering that all other variable are kept fixed at a certain level.

In order to test the parameter it is necessary to introduce certain quantities that are the score function

$$S(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \mathcal{L}(Y, \boldsymbol{\beta}),$$

and the Fisher information

$$I_f(\boldsymbol{\beta}) = -\mathbb{E}(\frac{\partial^2}{\partial \boldsymbol{\beta} \boldsymbol{\beta}'} \mathcal{L}(Y, \boldsymbol{\beta})),$$

where $\mathcal{L}(Y, \boldsymbol{\beta})$ is the log-likelihood of the model.

To test the null hypothesis of nullity of q coefficients

$$H_0 : \{ \beta_{j_1} = \ldots = \beta_{j_q} = 0 \},$$

it is possible to use the likelihood ratio test that is expressed in the same way as with one variable or the score test that is based on the statistic

$$S = S(\hat{\boldsymbol{\beta}}_{H_0})' I_f^{-1}(\hat{\boldsymbol{\beta}}_{H_0}) S(\hat{\boldsymbol{\beta}}_{H_0}),$$

where $\hat{\boldsymbol{\beta}}_{H_0}$ is the maximum likelihood estimator of β under H_0. This statistic follows a $\chi^2(q)$ distribution under the null.

2.2 Classical association tests

The main association tests can be based on the genotypic contingency table (**Table 2.4**), indicating the number of individuals for a given genotype and status, or on the allelic contingency table (**Table 2.5**), indicating the number of individuals for a given allele and status. These tables are represented using a marker with alleles a/A and genotypes aa, aA or Aa that are not differentiable and AA.

	aa	**aA/Aa**	**AA**	**Total**
Case	D_0	D_1	D_2	N_D
Control	H_0	H_1	H_2	N_H
Total	N_0	N_1	N_2	N

Table 2.4: **Genotypic table.** This table considers a SNP with alleles a and A. D_i represents the number of cases with the genotype i, H_i the number of controls with the genotype i and N_i the number of individuals with the genotype i.

	a	**A**	**Total**
Case	$D_a = 2D_0 + D_1$	$D_A = 2D_2 + D_1$	$2N_D$
Control	$H_a = 2H_0 + H_1$	$H_A = 2H_2 + H_1$	$2N_H$
Total	$N_a = 2N_0 + N_1$	$N_A = 2N_2 + N_1$	$2N$

Table 2.5: **Allelic table.** This table considers a SNP with alleles a and A. D_a and D_A represent the number of cases with the allele a or A, H_a and H_A the number of controls with the allele a or A and N_a and N_A the number of individuals with the allele a or A.

We present the three main unadjusted association tests that are the genotypic test, the Armitage Trend test and the allelic test. For all these tests we introduce both the formal definition and the interpretation in terms of logistic regression.

2.2.1 Genotypic test

The genotypic test compares the proportion of the different genotypes between the cases and the controls. It is directly based on the genotypic contingency table. This test does not assume any mode of inheritance for the marker. The null hypothesis of this test is

$$H_0 : \left\{ \begin{array}{lcl} p_{D_0} = p_0 p_D & ; & p_{H_0} = p_0 p_H \\ p_{D_1} = p_1 p_D & ; & p_{H_1} = p_1 p_H \\ p_{D_2} = p_2 p_D & ; & p_{H_2} = p_2 p_H \end{array} \right\},$$

where p_{D_i} and p_{H_i} are the proportions of diseased and healthy individuals with the genotype i, p_i is the proportion of individuals with the genotype i and p_D and p_H are the proportions of diseased and healthy individuals. The corresponding test statistic is

$$S_G = \sum_{i=0}^{2} \frac{\left(D_i - \frac{N_D N_i}{N}\right)}{\frac{N_D N_i}{N}} + \frac{\left(H_i - \frac{N_H N_i}{N}\right)}{\frac{N_H N_i}{N}}.$$

This statistic is a Pearson statistic and follows a $\chi^2(2)$ distribution under H_0.

This test corresponds to the following regression model

$$logit(p) = \alpha + \beta_1 z_1 + \beta_2 z_2,$$
$$H_0 : \{\beta_1 = \beta_2 = 0\},$$

where z_1 and z_2 are dummy variables that represent the genotypes. The genotype aa is coded $\binom{z_1=0}{z_2=0}$, aA/Aa is coded $\binom{z_1=1}{z_2=0}$ and AA is coded $\binom{z_1=1}{z_2=1}$.

2.2.2 Armitage Trend test

The Armitage Trend test (Cochran 1954, Armitage 1955) aims to find a linear trend between the probability of having the disease and the genotypes. In this test, the genotypes are ordered which leads to gaining a degree of freedom. The order of the genotypes assumes that there is a quantitative effect depending on the number of reference allele. The null hypothesis is

$$H_0 : \{\frac{p_{D_0}}{p_0} = \frac{p_{D_1}}{p_1} = \frac{p_{D_2}}{p_2}\},$$

and the test statistic can be expressed according to (Sasieni 1997) as

$$S_T = \frac{N[N(D_1 + 2D_2) - N_D(N_1 + 2N_2)]^2}{N_D N_H[N(N_1 + 4N_2) - (N_1 + 2N_2)^2]}.$$

This statistic follows a $\chi^2(1)$ under the null hypothesis.

This test corresponds to a logistic regression model using the genotype variable x coded 0, 1 or 2 such as indicated in the previous section.

$$logit(p) = \alpha + \beta x,$$

$$H_0 : \{\beta = 0\}.$$

An advantage of this test is that it allows to account for different modes of inheritance (Slager and Schaid 2001). By modifying the coding of the genotype variable x, it is possible to also consider dominant and recessive modes.

2.2.3 Allelic test

The allelic test is the analogue of the genotypic test based on the alleles instead of the genotypes. It compares the alleles counts between the cases and the controls. This test is based on the null hypothesis

$$H_0 : \left\{ \begin{array}{lcll} p_{D_a} = p_a p_D & ; & p_{H_a} = & p_a p_H \\ p_{D_A} = p_A p_D & ; & p_{H_A} = & p_A p_H \end{array} \right\},$$

and uses the Pearson test statistic

$$S_A = \frac{2N[(2D_2 + D_1)(2H_0 + H_1) - (2D_0 + D_1)(2H_2 + H_1)]^2}{2N_D N_H(2N_2 + N_1)(2N_0 + N_1)}$$

that follows a $\chi^2(1)$ distribution under H_0.

A logistic regression model using the variable z corresponding to the allele ($z = 0$ for a and $z = 1$ for A) can be associated to this test.

$$logit(p) = \alpha + \beta z,$$

$$H_0 : \{\beta = 0\}.$$

As each individual contributes to two observations (one for each allele), this test can be biased. Alternative allelic tests that obviate this bias have been designed (Guedj et al. 2006b).

2.3 Association tests accounting for population structure

In the Introduction we have indicated and explained how population stratification can bias the classical association tests, by confusing the relationship between the marker and the disease. This can lead to finding spurious associations or to missing genuine ones. Using family designs and the corresponding Transmission Disequilibrium Test can avoid this bias, however the complexity of conducting such studies have led to the necessity of developing approaches to account for stratification in case-control designs. We present in this section different strategies that allow to account for population structure while testing for association. The main strategies are the Genomic control, Structured Association, principal component analysis based approaches, regression models and meta-analyses. The next section is dedicated to a comparison of these approaches.

2.3.1 Unlinked marker selection

Many of the approaches that we will present are based on sets of unlinked markers. These markers are not in linkage disequilibrium between each other and with the disease susceptibility markers. The use of such markers allow to get information about the genetic diversity of the populations. If the markers were linked, the information would be redundant and the ancestry of the population could be incorrectly estimated. These ancestry estimations are then used to correct for stratification.

The selection of such markers is not straightforward as it is not always possible to know which markers are in LD with the disease susceptibility markers. It has however been suggested that selecting markers at random among the whole set of markers available in a study should lead to the selection of such loci. Indeed, given the massive amount of markers, the probability of picking markers in LD with the disease susceptibility markers can be considered as relatively low (Pritchard and Rosenberg 1999).

In order to ensure that all markers are in linkage equilibrium with each others, an approach called pruning has been developed. Pruning strategies use sliding windows and progressively exclude markers from the dataset to ensure that the linkage disequilibrium between the markers that are still included in the data does not exceed a certain threshold. A pruning strategy is proposed in the software `plink` and is often used in practice as a first step in many genetic studies.

Another question of interest is how many of such markers should be use to get a proper assessment of the individual ancestry. Several studies have studied that question, assessing the quality of the genetic history obtained with different numbers of unlinked markers. Each method actually needs a different amount of markers to be accurate.

A solution to reduce the number of markers and ensure the quality of the ancestry estimation is to use ancestry informative markers (AIMs). These markers are the one providing the most information about the ancestry of the individuals. Certain methods have been designed to identify such markers (Paschou et al. 2007, Zhang et al. 2009a). It is however not always possible in practice to use AIMs as these markers are often population specific and might not be suitable for all studies.

2.3.2 Genomic control

Methodology

The Genomic control (GC) is one of the first method proposed to account for stratification (Devlin and Roeder 1999). Its idea is based on the distribution of the Armitage Trend test statistic. Under the null hypothesis, this association test statistic should follow a χ^2 distribution. In the case there is population stratification this distribution is inflated so that the statistic follows a non-central χ^2. **Figure 2.1** is an example of such inflated distribution.

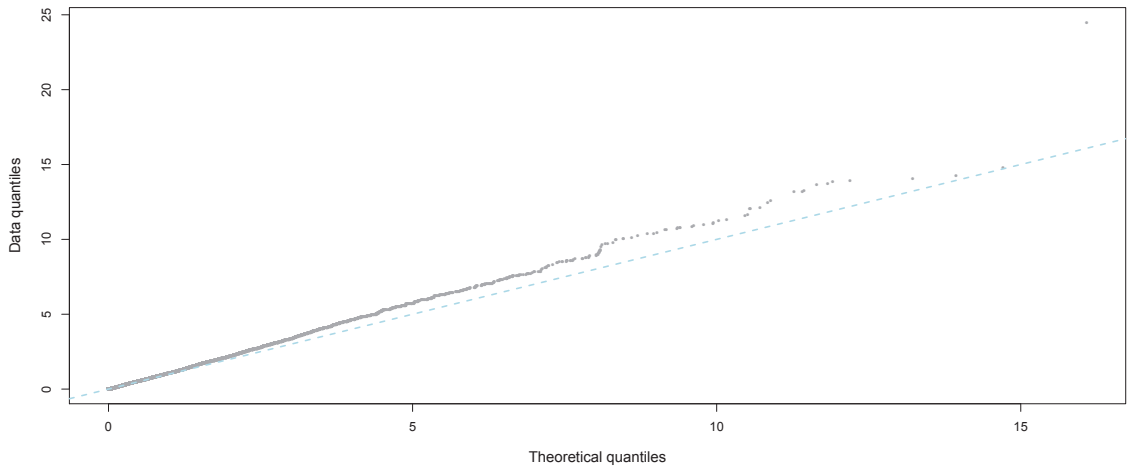

Figure 2.1: **Inflated χ^2 distribution.** This Q-Q plot shows the inflation of the test statistic distribution in presence of stratification.

The Genomic control aims to estimate a factor λ called *variance inflation factor* that represents the degree of stratification and is used to correct the observed statistic distribution to fit a χ^2 distribution.

The factor λ is calculated using the information contained in several markers included in the study and not only the one being tested.

$$\lambda = \frac{median(S_1, \ldots, S_L)}{0.456}$$

where $(S_1, \ldots, S_L)$ are the Trend test statistics for L non-associated and unlinked markers[1] and 0.456 is the median of the expected χ^2 distribution if there was no inflation. The factor λ is then used to correct the distribution of the test statistic of each marker tested so that the new distribution S_i^2/λ can be compared to a χ^2 distribution.

Comments

In addition to correct the test statistic, the factor λ can be used as an indicator of the presence and the degree of stratification. In practice, we usually consider that a λ inferior to 1.05 indicates that there is no stratification (Price et al. 2010). The higher is λ, the more important the stratification can be considered.

The Genomic control relies on one major hypothesis that the inflation factor is unique for all markers. This is the main weakness of the method as it is not likely that all the SNPs equally reflect the stratification. In addition, the number of markers necessary to properly estimate λ depends on the complexity of the structure of the population. On the other hand, this method is computationally easy to program and runs rather fast.

Alternatives to the classical Genomic control have been proposed such as the GCmean that uses an estimation of λ based directly on the mean of the test statistics instead of on the median (Reich and Goldstein 2001), the GCF that considers λ to be variable having a distribution to account for its variability (Devlin et al. 2004) or the robust Genomic control (RGC) that considers different association models to calculate the adjusted test statistics (Zheng et al. 2006).

All these Genomic control approaches offer a way to detect and quantify any kind of stratification as well as to take it into account when testing marker-disease associations.

2.3.3 Structured Association

Structured Association is a class of methods that uses unlinked markers to determine the presence of population stratification, infer details about the population structure and conduct association testing while accounting for the corresponding bias.

[1] i.e. markers that are not in linkage disequilibrium.

Detecting population stratification

Pritchard et al. proposed a statistical test to detect population stratification (Pritchard and Rosenberg 1999). This method uses L unlinked markers that are also not in linkage disequilibrium with the potential candidate markers so they are not associated with the disease. The rationale of this test for stratification is that if there is no stratification, the set of unlinked markers should not be associated with the disease. By combining the association test statistics of each of these markers $(S_1, \ldots, S_L)$, it is possible to compute a global statistic

$$S_s = \sum_{i=1}^{L} S_i,$$

that follows a χ^2 distribution under the null hypothesis

$$H_0 : \{\text{None of the } L \text{ unlinked markers is associated with the disease}\}.$$

The number of degrees of freedom of this statistic equals to the sum of the number of degrees of freedom of each individual test.

The amount of unlinked markers to conduct such a test depends on the complexity of the population structure involved.

Association testing

Performing a Structured Association test corresponds to inferring the population structure and then testing for association conditionally on the inferred structure (Pritchard et al. 2000b, Satten et al. 2001). We present here the association test proposed by Pritchard et al. and available in the software `Strat`[2]. The method to infer population structure assumes the existence of K genetically homogeneous sub-populations and assigns the individuals to these sub-populations or determines how much of each individual's genome comes from each sub-population[3]. The algorithm used by Pritchard et al. to conduct this inference is `Structure` and will be further detailed in **Chapter 3** that is dedicated to the inference of population structure. We therefore assume here that the structure of the population have been inferred and present the adjusted association test.

The main test for Structured Association, `Strat`, is based on a likelihood ratio test and the null hypothesis

$$H_0 : \{\text{The genotype frequencies are independent from the disease status}\}.$$

[2] `http://pritch.bsd.uchicago.edu/software.html`

[3] We call admixture proportions the corresponding quantities.

Let $x = (x_1, \ldots, x_n)$ be the marker tested, $y = (y_1, \ldots, y_n)$ the states of the n individuals, $q = (q_{ik})_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant k \leqslant K}}$ the admixture proportions, that are the proportions of individual i's genome coming from the $k-th$ inferred sub-population and $p = (p_{kj})_{\substack{1 \leqslant k \leqslant K \\ 0 \leqslant j \leqslant 2}}$ the frequency of genotype j in sub-population k. We also define $P_0 = \{p_{kj}; 1 \leqslant k \leqslant K, 0 \leqslant j \leqslant 2\}$ and $P_1 = \{p^y_{kj}; 1 \leqslant k \leqslant K, 0 \leqslant j \leqslant 2, y = 0 \text{ or } 1\}$ the genotype frequencies under H_0 and H_1. As a matter of fact, under the null, the genotypes are independent of the status, however under H_1 these frequencies are different between the cases and the controls.

The likelihood ratio test corresponds to the ratio of the distributions of x under H_0 and H_1

$$\Lambda(x) = \frac{\mathbb{P}_{H_1}(x; P_1, q)}{\mathbb{P}_{H_0}(x; P_0, q)}.$$

This ratio includes information about the inferred population structure. It actually corresponds to a ratio stratified between the different sub-populations.

The two distributions can be estimated by

$$\mathbb{P}_{H_0}(x_i = j \mid P_0, q, y) = \sum_{k=1}^{K} q_{ik} p_{kj},$$

$$\mathbb{P}_{H_1}(x_i = j \mid P_1, q, y) = \sum_{k=1}^{K} q_{ik} p^{y_i}_{kj}.$$

The estimation of q, that corresponds to the population structure, is conducted using the software `Structure` and the estimations of P_0 and P_1 using EM algorithms.

This test statistic has an unknown distribution, therefore its significance is empirically calculated by

$$p - value = \frac{\#(\{\Lambda(x^{(b)}) \geq \Lambda(x), b = 1 \ldots B\})}{B},$$

where $\#()$ represents the cardinal function and $(x^{(b)})_{1 \leqslant b \leqslant B}$ are B simulated random draws from under H_0.

This statistical test is computationally quite intensive. First the estimation of the parameters q, P_0 and P_1 have a high computational cost, then the computation of an empirical p-value leads to repeating these estimations many times.

Note that another Structured Association approach was proposed in (Satten et al. 2001) and simultaneously infers the population structure and performs the association test. This approach is however rarely used in practice.

2.3.4 Principal component analysis based methods

Principal component analysis (PCA) is a way of exploring data by creating axes of variation, that are linear combinations of the original variables and represent the variability of the data. Each axis of variation account for as much variability as possible that has not been accounted for in the previous axes. This property of the method allows one to analyze a reduced number of variables, called principal components (PCs), while keeping most of the relevant information. This feature is particularly important in genetic studies where the number of markers analyzed is very large.

Principal component analysis applied to genetic data

PCA method

The most used PCA-based method was introduced in 2006 (Price et al. 2006, Patterson et al. 2006). The authors suggest that main axes of variation of a PCA applied to SNP data represent the ancestral genetic variability between individuals and therefore provide a way of analyzing population structure and accounting for stratification. This method is implemented in the software `Eigensoft`[4].

Let us consider the $n \times p$ genotype matrix X as introduced in **Section 2.1.1**. We first assume that all the markers are unlinked. To conduct a PCA of this matrix it is first necessary to center and normalize the data. We therefore define the column means

$$\mu_j = \frac{\sum_{i=1}^{n} x_{ij}}{n},$$

and the observed allele frequency of each marker

$$p_j = \frac{1 + \sum_{i=1}^{n} x_{ij}}{2 + 2n},$$

where the missing entries are excluded from the computation. The new genotype matrix $\tilde{X}$ is defined so that each entry is

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\sqrt{p_j(1 - p_j)}}.$$

If the markers are not unlinked, then it is possible to regress each marker on the d preceding markers and to use the residuals of such regressions as new variables. This operation allows to account for the linkage disequilibrium between the markers. Patterson et al. suggest that considering $d = 1$ or 2 is generally sufficient.

[4] `http://genepath.med.harvard.edu/~reich/Software.htm`

One then computes a singular vector decomposition of the $n \times n$ covariance matrix $\frac{1}{p}\tilde{X}\tilde{X}'$. This yields a set of principal components $(PC_1, \ldots, PC_{n-1})$. These principal components represent the coordinates of the individuals on several sub-spaces and allow among others to account for population stratification or to cluster individuals into homogeneous sub-populations. Genetic and genealogical interpretations of these sub-spaces are also available. One can refer to (Patterson et al. 2006, McVean 2009) for further details.

Assessing the number of significant principal components

An important issue with PCA applied to genetic data is to determine the number of significant principal components, that is the number of principal components that actually describe a structure of the population. A possible technique to do so is to use the eigenvalues and the Tracy-Widom distribution (Tracy and Widom 1994). It has been demonstrated that after a suitable normalization, the largest eigenvalue follows the Tracy-Widom distribution (Johnstone 2001). Patterson et al. extended this theory to determine a statistic for the $k-th$ eigenvalue λ_k, that represents the $k-th$ axis of variation. The statistic is

$$\frac{\lambda_k - \mu(n, p, \lambda_k, \ldots, \lambda n - 1)}{\sigma(n, p, \lambda_k, \ldots, \lambda n - 1)},$$

where μ and σ are normalization functions depending on the eigenvalues and on the degree of linkage disequilibrium in the data.

The strategy to determine the number of significant principal components corresponds to testing one after another the null hypotheses

$$H_0 : \{ \text{ The } k-th \text{ principal component is not significant} \},$$

by the use of the above test statistic that follows a Tracy-Widom distribution.

Application of the PCA

A first use of these axes of variation is to provide graphical representations of the individuals. **Figures 1.7 and 1.8** are examples of such representations. This type of representation is widely used to provide a graphical overview of the population structure.

Another possible application of the principal components is to highlight outlier individuals in a genetic studies. Indeed, by plotting all the individuals of a study on several axes of variation, or by analytically looking at these axes, it is possible to determine which individuals seem to lie farther than the others. Principal component analysis is therefore often used in practice in the first steps of a genetic study (**Section 1.3.2**).

PCA and clustering of genetic data

The set of significant principal components selected to describe the structure of the population can also be used to cluster individuals into genetically homogeneous sub-populations. For instance a Gaussian Mixture model or a $k-means$ algorithm can be applied to these principal components (Lee et al. 2009). An example of such clustering method will be presented in **Chapter 3**.

Association testing

In order to account for population stratification, a statistical test using the results of the PCA applied to genetic data have been designed (Price et al. 2006). Given that the principal components represent axes of genetic variation, it is possible to project the genotypes on the main axes to obtain new variables that correspond to the proportion of the genotypes that is due to the structure of the population. One can then subtract to the original genotypes the part due to the structure and obtained adjusted genotypes x_{adj}. The same transformation is possible for the phenotypes and leads to adjusted phenotypes y_{adj}.

The computation of the adjusted data is very computationally efficient as it only consists in operations on matrices or vectors.

Using these adjusted data, a test statistic can be derived as an extension of the Armitage Trend test,

$$(n-1-m) \times corr(X_{adj}, Y_{adj})^2,$$

where n is the number of samples and m the number of retained principal components. This statistic follows a $\chi^2(1)$ distribution under the null hypothesis of no association.

PCA-like methods

Alternatives to principal component analyses are possible. The main alternatives are the multi-dimensional scaling (MDS), that uses similarity matrices between the individuals to create axes of variation (Li et al. 2008) or laplacian eigenfunctions (Zhang et al. 2009b).

2.3.5 Regression models

As we have seen previously, classical association tests can be expressed in terms of logistic regression models. Several adjustments of these models are possible to account for population stratification (Setakis et al. 2006, Balding 2006). The more commonly used model corresponding to the Armitage Trend test becomes

$$\text{logit}(p_x) = \alpha + \beta x + \text{Covariate(s)}.$$

Testing the nullity of the parameter β can therefore provide a way to test for association while taking stratification into account.

Adjustment on the principal components

The adjustment on the principal components uses the results of the PCA applied to SNP data such as described in the previous section. Principal components $PC_1, \ldots, PC_m$ are added to the logistic model as covariates. The number of principal components retained can be estimated with the Tracy-Widom method or be chosen manually given the complexity of the structure. One has to note that too much principal components can weaken the analysis as the number of variables in the regression model would be too large. We will discuss this point in the comparison study that we further propose.

Adjustment on population labels

The adjustment on discrete population labels corresponds to adding a discrete variable $(a_i)_{1 \leqslant i \leqslant n}$ to the model. This variable indicates the population of origin of each sample. It can pertain to the real population of origin if this information is available or to a population estimated through the use of a clustering algorithm.

2.3.6 Meta-analyses

Another possible approach to deal with population stratification is to conduct the analyses within K sub-populations considered homogeneous and to combine the results with meta-analysis methods, such as the Fisher method or Stouffer's Z-score (Whitlock 2005).

For a given marker, the same association test is conducted within each sub-population which leads to a set of p-values $(p_i)_{1 \leq i \leq K}$. Usually the Armitage Trend test is the test preferred. To derive combined statistics, one has to assume that all the tests are independent.

Fisher's method

Fisher's statistic corresponds to

$$-2 \sum_{i=1}^{K} log(p_i),$$

and follows a $\chi^2(2K)$ distribution under the null hypothesis of no association.

Stoufler Z-score method

Stoufler's Z-score statistic is

$$\sum_{i=1}^{K} F^{-1}(1-p_i),$$

where F is the standard normal cumulative distribution and follows a standard normal distribution under the null hypothesis.

Comments

These tests imply that one needs some knowledge about the homogeneous sub-populations. They can either correspond to those provided with the data or be estimated.

A main assumption of these approaches is that the statistical tests conducted within each of these sub-populations are independent. This might not be true if, for instance, certain sub-populations are close to each other but distant with the others. One can see that the cryptic structure of certain populations can be a disadvantage for testing association with meta-analysis approaches.

2.3.7 Other possible approaches

Note that other methods accounting for stratification, less used in practice, have been proposed. They are based on phylogenetic trees (Li et al. 2010a), on the F coefficient (Zhang et al. 2009c), mixed models (Kang et al. 2010), stratification scores (Epstein et al. 2007), re-matching of cases and controls (Guan et al. 2009), randomization test (Kimmel et al. 2007), simultaneous correction of stratification and genotyping errors (Cheng and Lin 2007) or propensity scores (Zhao et al. 2009).

2.4 Comparison of different approaches

2.4.1 Introduction

Many reviews and comparison articles looking at approaches to account for population stratification have examined the potential of these methods (Pritchard and Donnelly 2001, Zhang et al. 2008, Tian et al. 2008, Wang et al. 2009, Price et al. 2010, Wu et al. 2011). They focused on certain parameters affecting the stratification such as the sampling imbalance, the minor allele frequency of the disease susceptibility locus or the sample size. Most of them did not however exhaustively considered the different types of population structures. The study that we propose in this section carefully analyzes this very parameter. We propose a comparison of the mainly used methods by considering a large panel of stratification scenarios corresponding to the different types of population structures.

Our study differ from the recent comparison proposed in (Wu et al. 2011) by the methods considered and the type of simulations conducted. In our study numerous stratified datasets are simulated based on real data so that the structures of the population is well controlled and the data are similar to the ones used in real situations. We are interested in determining which methods tend to perform well, in terms of false-positive rate and power, under various situations. More precisely we aim at providing practical indications regarding which method(s) should be used with a given structure of the population as they account properly for the stratification bias. We address these questions for unstructured populations, admixed populations, discrete and hierarchical ones. Also, we propose a solution for situations where the sampling design has led to sub-populations only composed of cases or controls that haven't been genetically matched.

First, we present the different methods that we decided to compare. Then we describe our process to simulate genetic data under various stratification scenarios. We provide precisions on the comparison strategy as well, i.e. how we estimated the statistical indicators that are the false-positive rates and powers of the methods. We then present our results and conclusions.

2.4.2 A large panel of methods compared

We decided to compare the performances of six broadly used strategies to account for stratification. First, we focused on the Genomic control (GC) (Devlin and Roeder 1999) and on the test proposed by Price et al. implemented in `Eigenstrat` (Eig) (Price et al. 2006). Then, we included a family of adjusted Logistic Regressions (Reg). A large number of types of adjustments can be considered. We decided to focus on the mainly used in practice: adjustment on the five first principal components resulting from a PCA (Reg PCs), adjustment on the real population labels when this information is precisely known (Reg Real Pop) and adjustment on estimated population labels (Reg Est Pop). These latter labels were estimated using a Gaussian mixture model clustering on the principal components[5] (Lee et al. 2009). We also studied one Meta-Analysis approach based on the Fisher method (Meta). Finally, we considered the Armitage Trend test, that does not account for stratification, as a reference to assess the level of stratification in the data. Structured Association were not included in our comparison due to the computational cost of such methods and the fact that they are not usually utilized in practice to account for stratification when testing for association.

Several additional alternatives of the Genomic control, Regressions and Meta-Analysis where investigated as well. Since their results did not turned out to be significantly different from the original approaches, we will only discuss them in **Section 2.5**.

[5] This algorithm is detailed in **Section 3.2.5**.

2.4.3 Simulation model

The simulated genetic data tha we consider are based on the island model therefore on the assumption that the population structure is organized in sub-populations. This assumption is often made to when simulating data of genetic association studies. Our simulation model follows approaches previously used (Hyam et al. 2008, Li and Li 2008, Peng and Amos 2010) and is based on the genotype frequencies of real datasets. These frequencies are used as an empirical distribution of the range of possible genotypes. Simulating this way leads to genetic patterns similar to those found in real data and therefore allows us to finely control the type of population structure. That way, we first simulate several datasets corresponding to the sub-populations of origin. Then we randomly mate each sub-population and apply a genetic model to generate diseased and healthy samples. To simulate discrete sub-populations, the populations of origin are independently mated and for admixed populations we mate these populations with each other. The final sub-populations simulated are mixed together to produce a cohort of individuals with population structure. The type of population structure depends on the original datasets selected and the parameters of the model.

The genetic model is based on Wright's model (Wright 1921) applied to a bi-allelic marker with alleles A and a. Let p_0, p_1 and p_2 be the frequencies of genotypes aa, aA/Aa and AA defined by the Hardy-Weinberg proportions

$$\left\{\begin{array}{lcl} p_0 & = & p_a^2 + \mathcal{F} p_a(1-p_a) \\ p_1 & = & 2p_a(1-p_a) - 2\mathcal{F} p_a(1-p_a) \\ p_2 & = & (1-p_a)^2 + \mathcal{F} p_a(1-p_a) \end{array}\right. ,$$

where p_a is the frequency of allele a and $\mathcal{F}$ is the consanguinity coefficient that we consider null hereafter so that the disease susceptibility locus (DSL) is under the Hardy-Weinberg equilibrium.

We then want to compute the genotype frequencies of the DSL for cases and controls p_{D_i} and p_{H_i}, $i = 0$, 1 or 2, using the disease prevalence K_p, the penetrances f_0, f_1 and f_2 of the genotypes and the mode of inheritance of the disease. The main modes of inheritance can be defined by considering the relative risk $RR_{i/0} = RR_i = \frac{f_i}{f_0}$, $i = 1$, 2 by

$$\left\{\begin{array}{ll} \text{Recessive} & RR_1 = 1 \\ \text{Additive} & RR_1 = \frac{RR_2+1}{2} \\ \text{Multiplicative} & RR_1 = \sqrt{RR_2} \\ \text{Dominant} & RR_1 = RR_2 \end{array}\right. .$$

Using $f_0 = K_p/(p_0 + RR_1.p_1 + RR_2.p_2)$, $f_1 = RR_1.f_0$ and $f_2 = RR_2.f_0$ and the Bayes formulas we can easily derive the desired frequencies.

$$\begin{array}{lcl} (p_{D_0}, p_{D_1}, p_{D_2}) & = & \left(\frac{f_0.p_0}{K_p}, \frac{f_1.p_1}{K_p}, \frac{f_2.p_2}{K_p}\right), \\ (p_{H_0}, p_{H_1}, p_{H_2}) & = & \left(\frac{(1-f_0).p_0}{K_p}, \frac{(1-f_1).p_1}{K_p}, \frac{(1-f_2).p_2}{K_p}\right). \end{array} \tag{2.1}$$

Figure 2.2 graphically represents the simulation process.

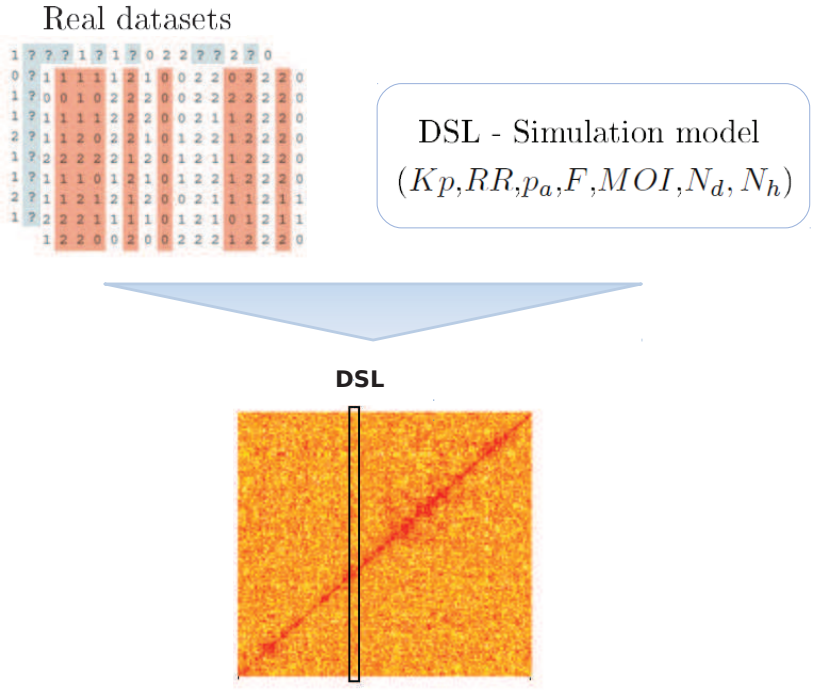

Figure 2.2: **Graphical representation of the simulation process.** Real datasets are combined with a genetic model to create a simulated dataset having one DSL and a specific genetic pattern.

2.4.4 Data sources and stratification scenarios

We simulated our data according to the model described in the previous section and using the HapMap phase III populations[6]. A number of 5,500 SNPs, with minor allele frequencies higher than 5%, were randomly chosen in equal number on each of the non-sexual chromosomes. We only considered SNPs present on an *Affymetrix GeneChip Human Mapping 500K* so that these SNPs are those commonly used in GWASs. Then, for each of our stratification scenario, some of the HapMap populations were used to simulate our final data with 5,500 SNPs and one DSL following an additive model and randomly located among the available loci.

We aimed at covering several situations as it may be harder to account for stratification with closely related populations than with very distant ones. Therefore, to get

[6] `http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/2010-08_phaseII+III`

an exhaustive assessment of the strategies we considered several scenarios corresponding to different types of population structure: no structure, admixed populations, discrete structures with populations more or less genetically close, and a hierarchical structure. The proportions of cases and controls simulated are different in the sub-populations so that the design is not a simple random sampling design[7]. Our design and the differences between the populations ascertain that we induced and controlled a bias due to population stratification.

The different scenarios that we considered are described hereafter and graphically represented in **Figure 2.3**. In addition, **Table 2.6** gives the simulation parameters of the DSLs for these scenarios.

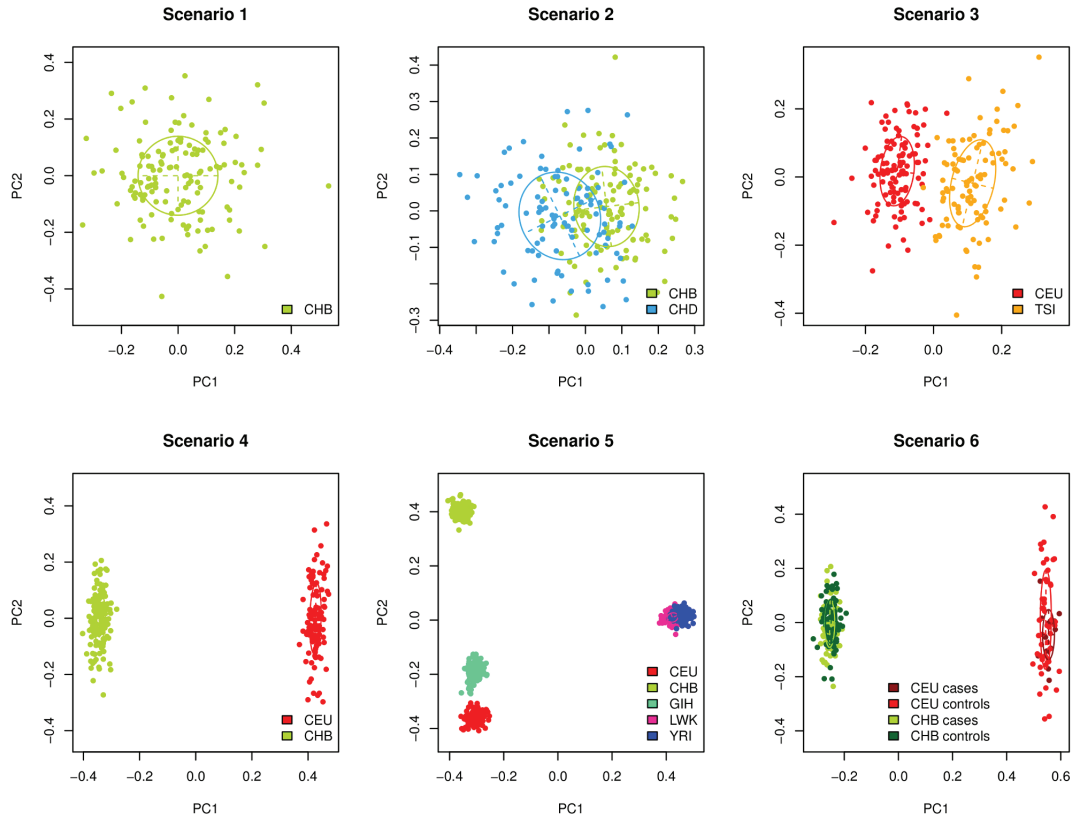

Figure 2.3: **Population structures of the different scenarios.** Samples are represented on the first two principal components (PCs) estimated on the genotype data.

[7]In such design, the probability of selecting a case or a control is the same among all sub-populations.

Scenario (Structure)	Pop	Prevalence (K_p)	MAF	# Cases	# Controls
1 (None)	*CHB*	0.05	0.3	200	200
2 (Admixture)	*CHD*	0.05	0.3	125	75
	CHB	0.01	0.2	75	125
3 (Discrete)	*CEU*	0.05	0.3	125	75
	TSI	0.01	0.2	75	125
4 (Discrete)	*CHB*	0.05	0.3	125	75
	CEU	0.01	0.2	75	125
5 (Hierarchical)	*GIH*	0.05	0.3	10	60
	LWK	0.01	0.4	30	50
	YRI	0.01	0.4	30	50
	CEU	0.05	0.2	10	60
	CHB	0.03	0.1	150	10
6 (Discrete)	*CHB*	0.05	0.3	200.r	100
	CEU	0.01	0.2	200.(1-r)	100

Table 2.6: **Simulation parameters for the stratification scenarios.** The MAF indicated corresponds to that of the disease susceptibility locus.

Scenario 1: One homogeneous population. With only one such population there is no stratification. The idea is to determine if the methods accounting for stratification are reliable when there are applied to a non-stratified population. Individuals from Han Chinese in Beijing, China (CHB) are used to simulate these data.
Scenario 2: Admixture. We considered an admixture of two originally close populations: Chinese in Metropolitan Denver, Colorado (CHD) and Han Chinese in Beijing, China (CHB) are used.
Scenario 3: Two fairly distant discrete populations. The two relatively distant discrete populations are Utah residents with Northern and Western European ancestry from the CEPH collection (CEU) and Toscans in Italy (TSI).
Scenario 4: Two very distant discrete populations. The two very distant discrete populations are Han Chinese in Beijing, China (CHB) and Utah residents with Northern and Western European ancestry from the CEPH collection (CEU).
Scenario 5: Hierarchical structure. The hierarchical structure is composed of five populations: Yoruba in Ibadan, Nigeria (YRI), Luhya in Webuye, Kenya (LWK), Han Chinese in Beijing, China (CHB), Gujarati Indians in Houston, Texas (GIH) and Utah residents with Northern and Western European ancestry from the CEPH collection (CEU).
Scenario 6: Varying proportions of cases/controls This scenario uses the same populations as scenario 4 but with a varying proportion of cases between the two subpopulations. The proportion of controls is fixed and equal in the two populations while the proportion of cases is taken with a (r, 1 - r) ratio, with r varying. When this proportion is of 0 then all the cases are in the CEU population that is the less affected by

the disease. When it is of 1 then all the cases are in the most affected population (CHB). Our goal is to observe the behavior of the methods in function of the degree of sampling imbalance and to look at whether they tend to perform well in the extreme case where all the cases come from only one of the populations. In this latter case, it is also of interest to determine if the best solution to account for population stratification is not to consider only the cohort composed of both cases and controls by excluding the samples that are not matched. The answer to this issue is particularly useful for large studies where controls with different ancestries are used to match the genotyped cases.

2.4.5 Comparison strategy

We used a statistical framework to analyze the potential of the main approaches investigated that focuses on their false-positive rates and powers.

Note that population stratification is said to lead to spurious associations but also to mask true associations. This second effect is more tricky to observe but the statistical power can be useful to do so. As it corresponds to the proportion of SNPs that have been detected associated when they were, a loss of power between a situation with no stratification and a situation with stratification means that SNPs that used to be correctly detected in the first situation are no longer in the second. This corresponds to missing associations.

Both false-positive rate and power can be expressed in function of the test statistic. However the distribution of this statistic is not always obvious so we prefer using the p-values instead. Thus the false-positive rate becomes $\mathbb{P}_{H_0}(p\text{-value} \leqslant \alpha)$ and the power $\mathbb{P}_{H_1}(p\text{-value} \leqslant \alpha)$. In our simulations, each dataset is simulated with one disease susceptibility locus, for which the degree of association is controlled, and 5,500 additional SNPs to assess the population structure. By placing ourselves under the null hypothesis, of no association, then under the alternative hypothesis, of association, we can respectively assess both false-positive rate and power of the methods. To do so, we use a Monte-Carlo method and assess the same quantity

$$\frac{\#\big(\{p\text{−value}_i \leqslant \alpha, i = 1 \dots B\}\big)}{B},$$

where $\#()$ represents the cardinal function, meaning that we count the number of p-values inferior or equal to α, and B the number of simulated datasets.

All the DSLs simulated, whether it is under the null hypothesis or the alternative, are differentiated between the sub-populations. This implies that for all the population structure scenarios considered, one DSL is simulated per sub-population. These DSL are excluded of the mating process the populations are then submitted to in order to reach

the desired type of structure. That way, the properties of the DSLs, such as the relative risk, are conserved whatever population structure is simulated (**Table 2.6**).

Note that only methods with equivalent false-positive rate can be compared in term of power. This implies that a method with high power is no better than one with low power if the first one did not maintain a correct false-positive rate.

We simulated data for several DSL relative risks ranging from 1 (no association) to 2.5 (strong association). For each relative risk a number of $B = 2{,}000$ datasets were simulated to get an accurate estimation of the statistical quality indicators. We genuinely estimated the indicators with this process as we controlled the degree of association through the simulation model. Note that there is an equivalence between the false-positive rate and the power when the relative risk is of 1. A level $\alpha = 5\%$ was chosen for all the tests. Data simulations and comparison of the strategies were performed using the software `R`[8].

2.4.6 Results

The results of the comparison are presented in this section for each scenario (**Figures 2.4 to 2.11**). **Table 2.7** summarizes the estimations of λ for the different scenarios. These estimations were conducted using the Genomic control based on the median of the Armitage Trend test statistics.

Scenario	λ
Scenario 1	1.002
Scenario 2	1.009
Scenario 3	1.065
Scenario 4	2.711
Scenario 5	9.571

Table 2.7: **Estimated λ for the different scenarios.** Classical Genomic control was used to conduct these estimations.

Scenario 1: One homogeneous population

In the first scenario, with an unstructured population, the estimation of λ was 1.002 confirming that there was no stratification. **Figure 2.4-A** presents the false-positive rate of the methods. We noted that all of the methods had a correct false-positive rate, lying within the 95% confidence bounds. `Eigenstrat` and Regressions adjusted on principal components (Reg PCs) were however the closest to the 5% level.

[8] `http://cran.r-project.org`

Figure 2.4-B provides the power curves of the different methods in function of the increasing relative risks. Powers of all the strategies were equivalent in this scenario except for Meta that was less powerful. One can note that there was no difference between an adjustment on a the real population labels and on the estimated ones. This was due to the fact that the population was so homogeneous that the clustering algorithm considered all samples to be in a unique population.

When there was no stratification, all the methods performed well and did not induce any bias. Besides, except for the Meta-Analysis, there was no loss of power when adjusting the results for stratification compared to the non-adjusted approach.

Scenario 2: Admixture

This scenario corresponded to an admixture of two close populations. The estimation of λ was 1.009 which meant that according to the Genomic control there was almost no stratification.
However, one can observe that there was still a real bias induced by population stratification as the Trend test had a false-positive rate significantly higher than 5% (**Figure 2.5-A**). This was also quite logically the case of the Genomic control as the variance inflation factor was close to 1.

`Eigenstrat` and Regressions adjusted principal components (Reg PCs) had false-positive rates reaching the upper bound of the confidence interval. Regressions adjusted on the estimated population labels (Reg Est Pop) led to a high number of false-positive findings. This might have been due to the fact that the clustering algorithm used was not accurate enough to determine the correct population labels of the individuals in the case of an admixture.

The Regression adjusted on the real population labels (Reg Real Pop) and the Meta-Analysis had a false-positive rate of almost 5%.

The analysis of the power curves (**Figure 2.5-B**) showed that the Trend test, the Genomic control and the Regression adjusted on the estimated population labels (Reg Est Pop) had the highest powers. This was however due to the inflation of the false-positive rate, also affecting the power, and therefore did not mean that these methods were more powerful. `Eigenstrat` and the Regression adjusted on the principal components were equivalent and outperformed the other methods in term of power. Regression adjusted on the real population labels (Reg Real Pop) and Meta were the less powerful methods.

In an admixture scenario, so with a very fine population structure, only `Eigenstrat`, Reg (PCs) and Reg (Real pop) were correctly correcting for stratification.

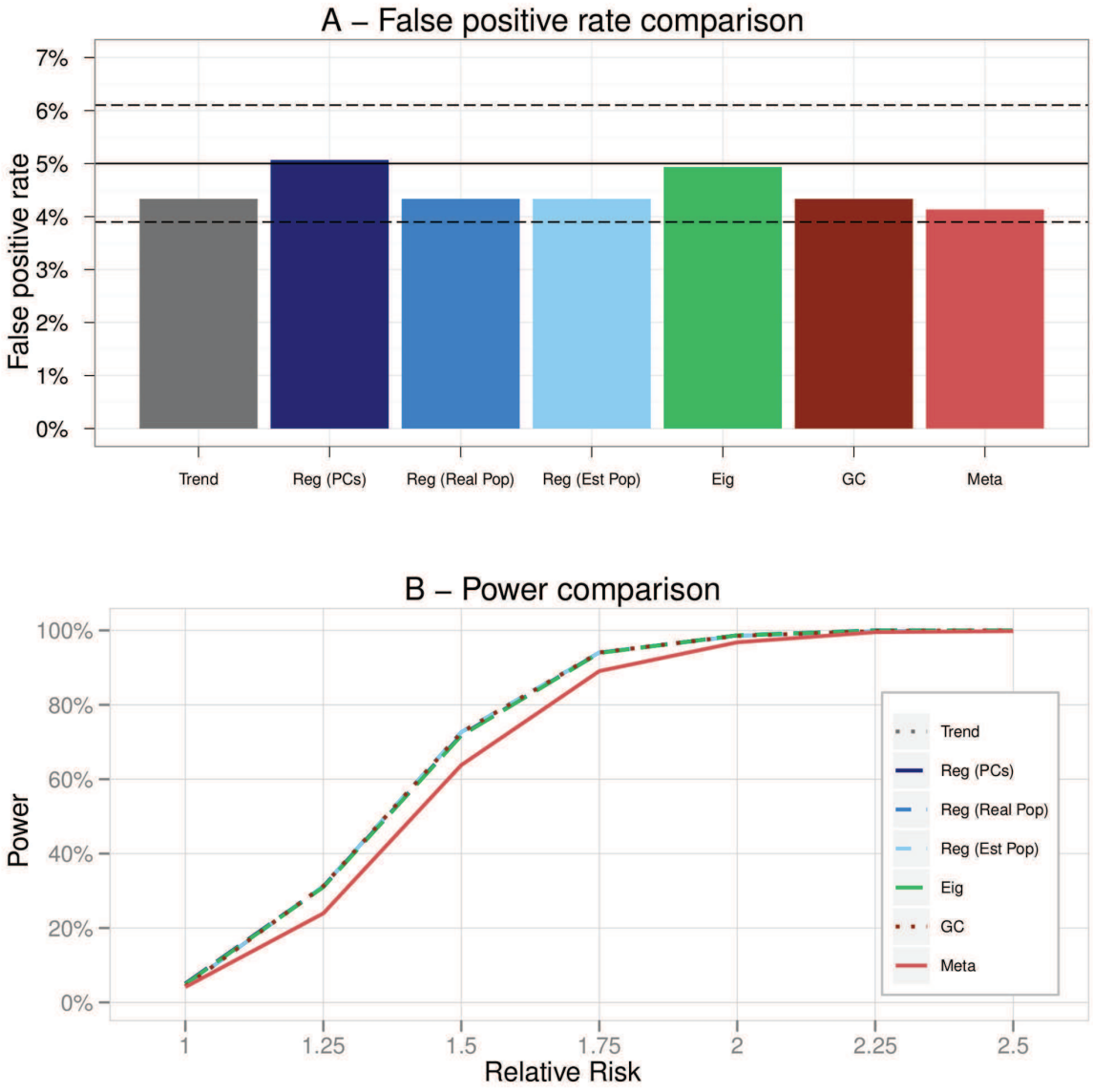

Figure 2.4: **Scenario 1 (One homogeneous population).** A - False-positive rates of the methods. The plain black line represents the 5% level at which the tests were conducted. The dashed black lines are the 95% confidence intervals for this level. B - Powers of the methods in function of the increasing relative risk.

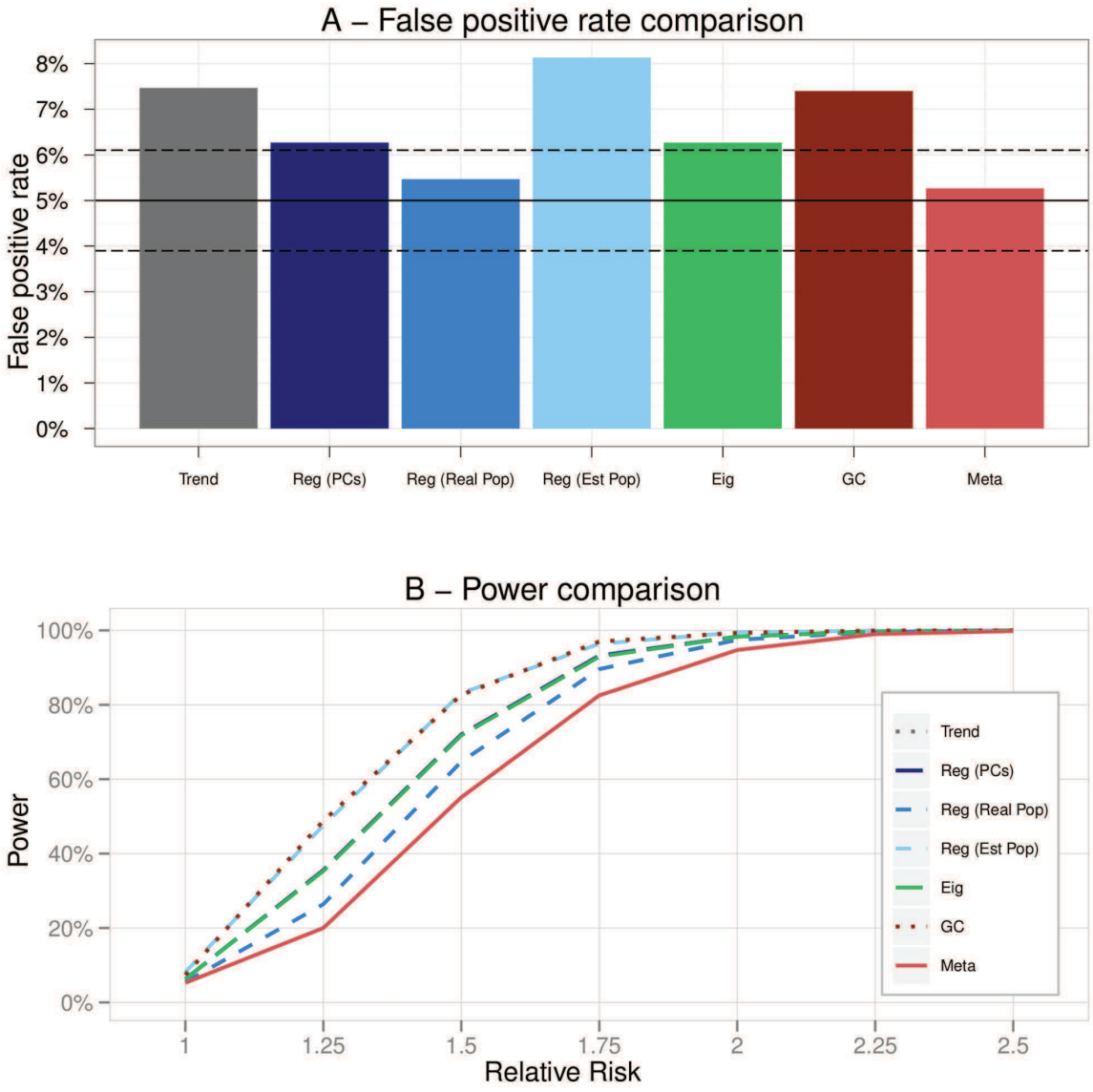

Figure 2.5: **Scenario 2 (Admixture).** A - False-positive rates of the methods. The plain black line represents the 5% level at which the tests were conducted. The dashed black lines are the 95% confidence intervals for this level. B - Powers of the methods in function of the increasing relative risk.

Scenarios 3 and 4: Discrete structures

The third scenario corresponded to two populations closely related but that were differentiable. The estimated λ was 1.065 indicating a slight stratification according to the Genomic control. Again the inflation factor was under-estimated as the false-positive rate of GC was very high such as for the Trend test. All the other methods had a correct false-positive rate (**Figure 2.6-A**).

On **Figure 2.6-B**, the power of `Eigenstrat` and the Regression methods were similar and higher than that of the Meta-Analysis.

In a situation where the populations were quite close it appeared that `Eigenstrat` and Regression based methods were the best solutions to account for stratification.

In Scenario 4, the estimation of λ was 2.711 which denoted quite an important structure of the population. In such a situation , the Trend test was very biased and had a highly inflated false-positive rate (**Figure 2.7-A**). On the other hand, the Genomic control behaved differently and became too conservative. All Regression methods were equivalent and performed as well as `Eigenstrat` both in terms of false-positive rate and power. Again the Meta-Analysis was the less powerful strategy (**Figure 2.7-B**).

Scenario 5: Hierarchical structure

Scenario 5 pertained to a more complex population structure. There were five populations and a hierarchical structure leading to an estimation of λ of 9.571. It was striking how the Trend test deviated from the 5% level by reaching almost 100% of false-positive findings under the null assumption. On the contrary, the Genomic control was very conservative due to the high value of λ. `Eigenstrat` had an inflated false-positive rate and was no longer equivalent to the adjusted Regressions. In addition, we observed that Meta was too conservative in this scenario (**Figure 2.8-A**).

The Genomic control was not powerful at all as it did not detected any association. Powers of all the Logistic Regressions were slightly smaller than that of `Eigenstrat` but this was due to the difference in false-positive rates (**Figure 2.8-B**).

In such a situation, only Logistic Regressions were capable of keeping correct false-positive rates while reaching good powers.

Scenario 6: Varying proportions of cases/controls

The sixth scenario corresponded to the same population structure as the fourth but with a varying sampling design. **Figure 2.9** presents the evolution of λ with the proportion

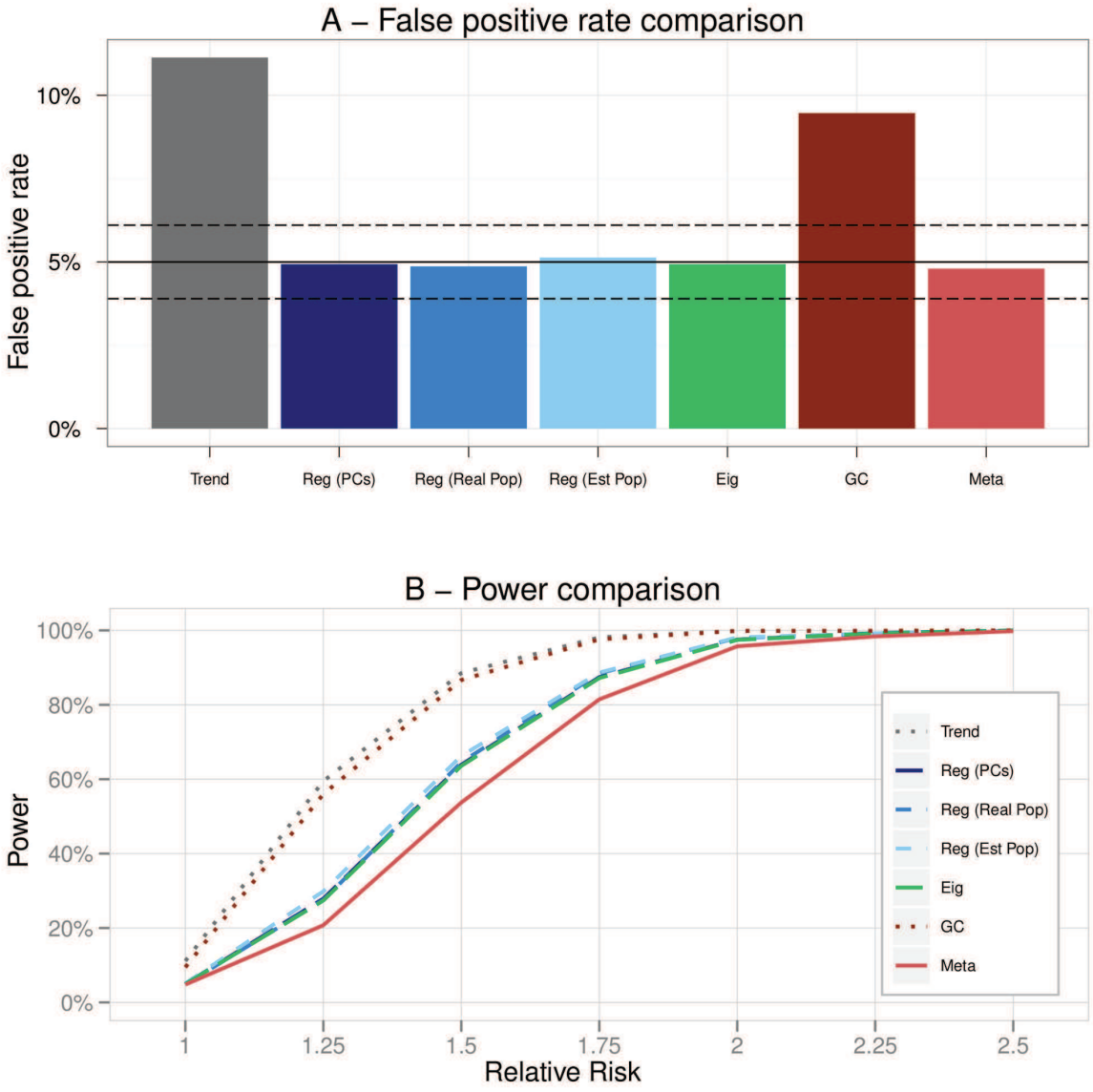

Figure 2.6: **Scenario 3 (Two fairly distant discrete populations).** A - False-positive rates of the methods. The plain black line represents the 5% level at which the tests were conducted. The dashed black lines are the 95% confidence intervals for this level. B - Powers of the methods in function of the increasing relative risk.

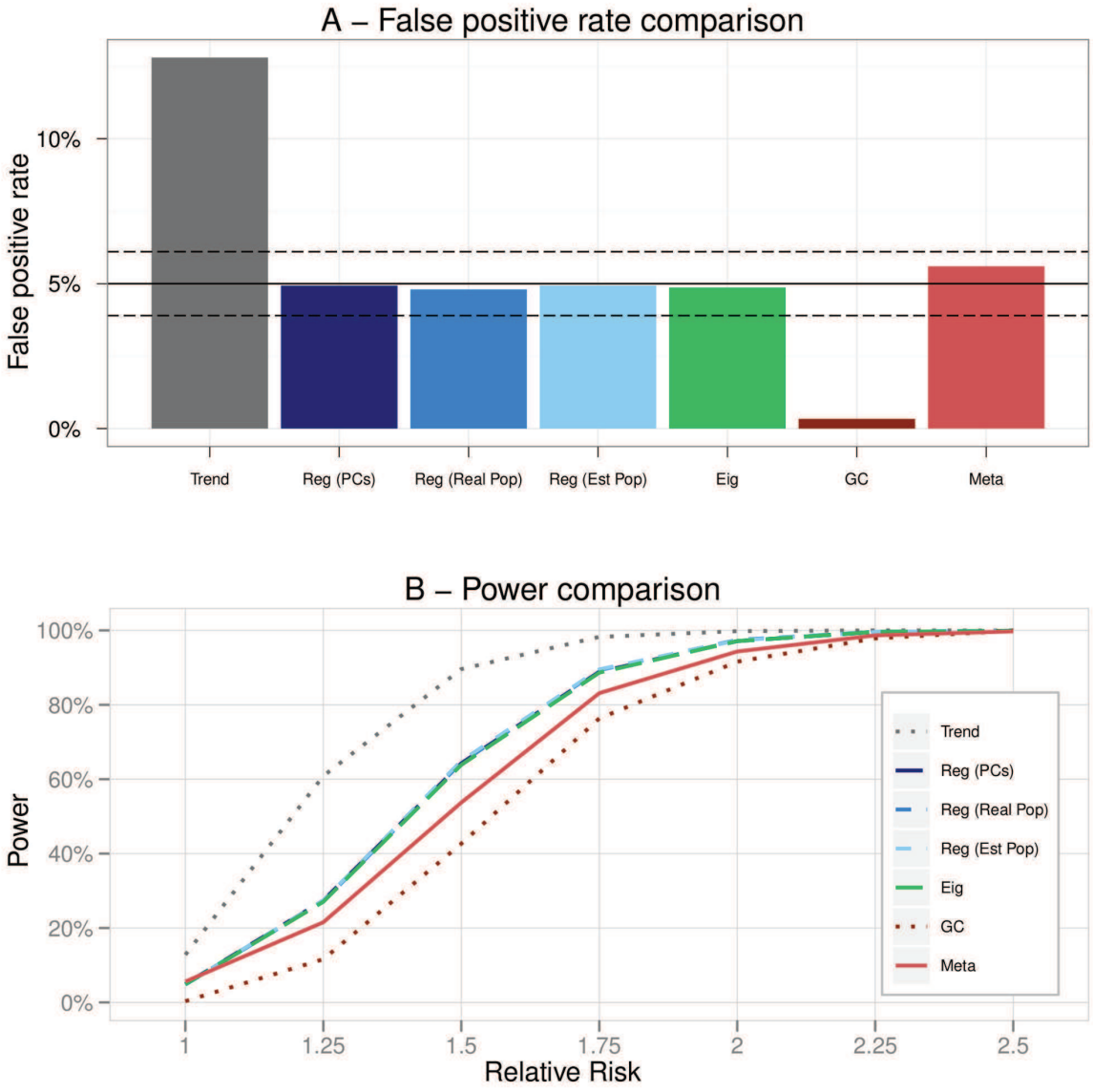

Figure 2.7: **Scenario 4 (Two very distant discrete populations).** A - False-positive rates of the methods. The plain black line represents the 5% level at which the tests were conducted. The dashed black lines are the 95% confidence intervals for this level. B - Powers of the methods in function of the increasing relative risk.

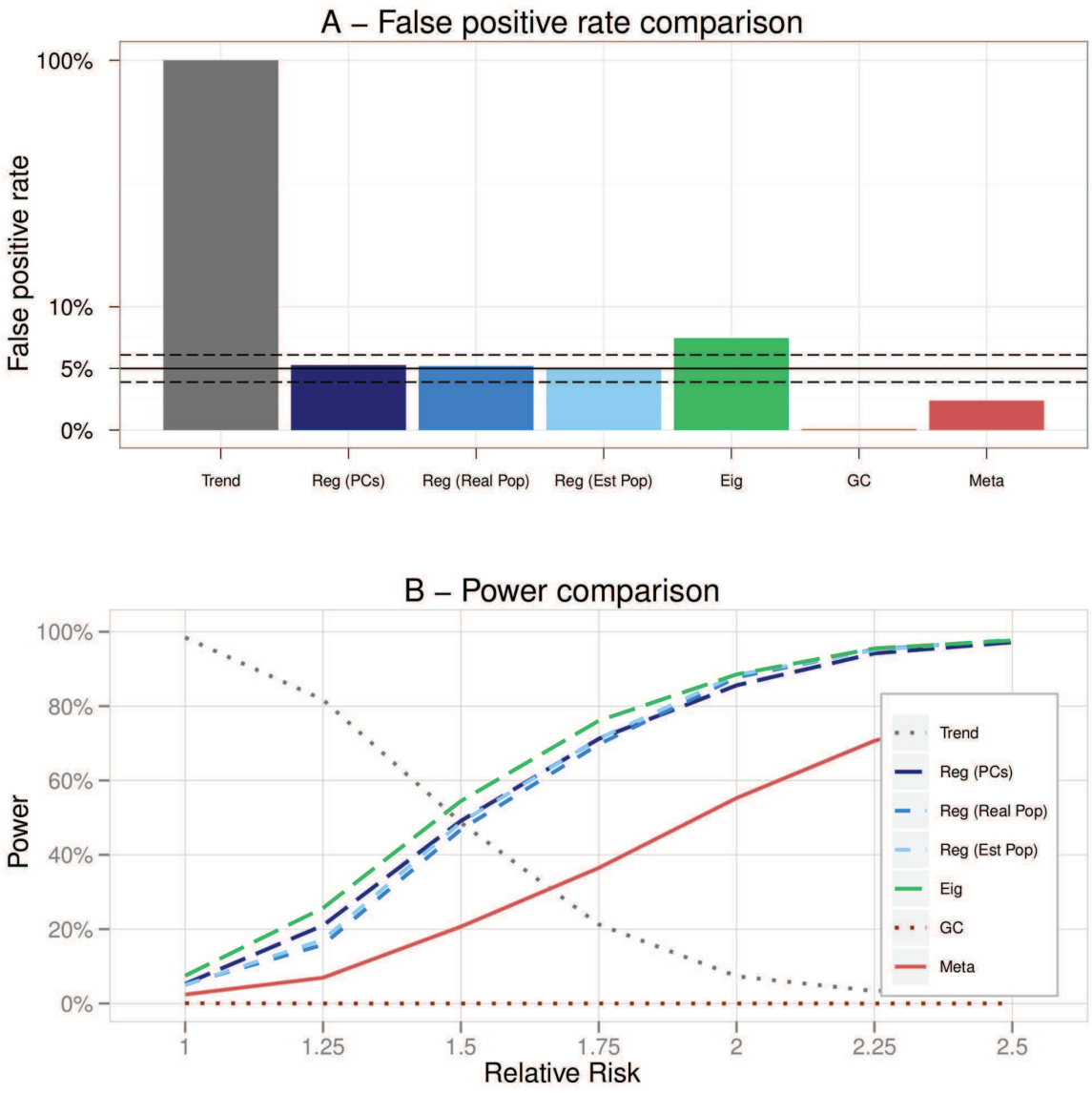

Figure 2.8: **Scenario 5 (Hierarchical structure).** A - False-positive rates of the methods. The plain black line represents the 5% level at which the tests were conducted. The dashed black lines are the 95% confidence intervals for this level. B - Powers of the methods in function of the increasing relative risk.

of cases.

Figure 2.9: **Evolution of λ for Scenario 6.** Representation of λ estimated with GC in function of the proportion of cases in the first population (pop1).

We observed that the Trend test had a correct false-positive rate only when the sampling design was balanced between the two populations otherwise it was inflated. The opposite trend was noticeable for the Genomic control (being quickly too conservative) and Meta. On the other hand, whatever the sampling design, Regressions and `Eigenstrat` globally maintained a correct false-positive rate (**Figure 2.10-A**). When the sampling was very imbalanced however, `Eigenstrat` tended to deviate from the 5% level.

The analysis of the power (**Figure 2.10-B**) showed us that powers of Regressions and `Eigenstrat` were equivalent which confirmed the result that we previously found in scenario 4.

An interesting fact was to observe the loss of power of the Trend test between the extreme situations. This confirmed that population stratification can lead to missing genuine associations. Quite logically we also retrieved the fact that if individuals are sampled in a very affected population then the power was more important than in other cases.

It is quite common in GWAS to include patients having different ancestries than the original cohort. This can be done to get larger samples or to find controls corresponding to the typed cases. A larger sample size implies a gain in power, however if the corresponding ancestries are different, population stratification could generate a bias reducing the power. If one of the group of patients with a different ancestry than the rest of the cohort is only

composed of controls (or cases), one practical question often discussed is whether it is better to exclude this cohort of the study or to keep it and account for stratification.

We answered this question by comparing the powers of the methods when all the patients were kept and when only the cohort composed of both cases and controls was kept. We focused only on Regressions and `Eigenstrat` that were the methods able to correctly correct for stratification. Whether all the cases were in the most affected or in the less affected population, we observed that the powers were the same whether the cohort composed of controls only was excluded or not. The power was not more important with more samples because of the bias due to stratification. However this bias was taken into account by the two methods so that it was not necessary to exclude a part of the patients (**Figure 2.11**).

Computational considerations

In term of execution time, the investigated methods are relatively equivalent. The Genomic control is relatively fast as it imply to test two times each SNP. Adjusted Regressions and `Eigenstrat` are quite equivalent when principal components are used to adjust the results. The necessary time to adjust on estimated population labels depends on the algorithm used to infer the population structure and can be quite fast or very time consuming.

It has been pointed out that Linear Regression can be a practical alternative to Logistic Regression as it is computationally faster, especially when there are covariates included in the models (Wu et al. 2010a). We analyzed this method as well in our study (data not shown). Linear and Logistic Regression methods seemed to be perfectly equivalent in most of the scenarios, however it appeared that the use of the dichotomous outcome that is the disease status in the Linear Regression is no longer a viable options in hierarchical populations (Scenario 5). We therefore recommend to keep using the Logistic Regression instead.

2.5 Discussion

The problem of population stratification is a serious shortcoming for GWASs raising doubts about their findings. To counteract this effect many approaches have been developed to account for stratification but it is not always clear in which situations they should be applied. Several articles have been published studying the performances of the different methods when some parameters influencing the stratification bias such as the minor allele frequency of the susceptibility locus, the degree of sampling imbalanced, the number of markers or the sample size vary (Pritchard and Donnelly 2001, Zhang et al. 2008, Tian et al. 2008, Wang et al. 2009, Price et al. 2010, Wu et al. 2011). We have decided to focus here on a parameter that has not been studied in depth and is yet quite

important that is the type of population structure itself. Indeed, one can wonder whether it is a good thing to adjust for stratification when there is no structure of the population, or whether reducing the bias is easier with distant or close populations. Also the relative performances of the most commonly used approach under these scenarios may vary differently. We compared these approaches through simulation studies by considering several scenarios of population structures. A particularity of our study is that to do so, we used a robust simulation model that is based on real genotype data so that we simulated datasets similar to the ones used in real situations.

We first determined that if there is no structure in the population, all of the studied methods correcting for stratification performed well both in term of false-positive rate and power reflecting trends previously reported (Epstein et al. 2007, Guan et al. 2009, Wu et al. 2011). Given this result and since it is quite difficult to be entirely sure that the population is sufficiently homogeneous, we recommend to always apply a correction for the stratification bias.

Concerning the type of population structure, our study also pointed out the fact that as soon as there is an admixture in the structure (Scenarios 2 and 5) then it is more delicate to correct the bias than with discrete populations.

We then highlighted methods that did not provide a good correction for stratification. First, we showed that the Genomic control failed to properly account for stratification in most of the situations. An interesting observation is that this method was not always affected in the same manner by the stratification. For genetically close populations the variance inflation factor λ was not a good indicator of the stratification level as it indicated almost no structure. This means that the Genomic control was anti-conservative. On the other hand, with relatively distant populations, this factor was overestimated, and therefore the false-positive rate was below the 5% level, rendering the Genomic control too conservative. We therefore confirm the conservativeness of the Genomic control reported in many situations (Pritchard and Donnelly 2001, Zhang et al. 2008, Dadd et al. 2009). We also studied an alternative version of the Genomic control, where the estimation of λ was based on the mean of the test statistics instead of on the median. This version provided the same results as the Genomic control presented in our study.

Second, in most of the scenarios we noted that the Meta-Analysis method was less powerful than the other alternatives. If it is however required to use a Meta-Analysis method then Fisher's method appeared as the best option. Indeed, we compared the Fisher and the Z-score methods and found that Fisher's always had a correct false-positive rate and a better power.

We therefore do not recommend the use of the Genomic control and Meta-Analyses methods to get a proper correction for stratification.

Note that it was not possible in our study to include the test implemented in the software `Strat` which is based on the results of `Structure` as the underlying algorithms are computationally very intensive (Zhang et al. 2008, Price et al. 2010). This rendered difficult to compare the test to the other methods in a robust manner. Even though it has been shown that `Strat` can provide a reasonable correction for stratification (Zhang et al. 2008), its high computational cost and complexity would lead us not to consider this test to account for stratification when conducting a GWAS.

Our results pointed out that the test implemented in the software `Eigenstrat` is a good solution to account for stratification with admixed or discrete structure which confirms the findings of (Zhang et al. 2008, Li and Yu 2008, Wu et al. 2011). On the other hand, with a hierarchical structure (Scenario 5), we found that `Eigenstrat` had a false-positive rate deviating from the 5% level which has been reported by previous studies (Li et al. 2010a, Wu et al. 2011). In the recent comparison study (Wu et al. 2011), no hierarchical structure was investigated however the inflated false-positive rate of `Eigenstrat` was reported for stratification scenarios including several populations or admixtures. Given that Regressions were able to correct the bias in a satisfactory way in this scenario it implies that `Eigenstrat` and the Logistic Regressions adjusted on the principal components are not always equivalent. This results is also outlined in (Wu et al. 2011).

Note that we included 5 principal components for the Regression adjustments and `Eigenstrat`. We decided not to use the Tracy-Widom theory to automatically determine the significant number of principal components as it usually leads to select a very high number of components (more than 20), which could lead to a poor convergence of the estimation of the Regression parameters. In practice only the first components are used. It is however of interest to look at the quality of the corrections if more or less components are considered. Additional simulations considering 1, 2, 5, 10, 20 or 50 components were conducted. They show that for a structure relatively simple to infer (Scenario 4), the number of principal components included in the models do not have an influence on the adjustments. Both the Logistic Regression and `Eigenstrat` have correct false-positive rates and comparable powers (**Figure 2.12**). When the structure of the population is more complex (Scenario 5), more components are needed to keep a reasonable false-positive rate (**Figure 2.13**). The Logistic Regression has an inflated false-positive rate if only one component is used and a better power if more than two components are used. It is interesting to note that `Eigenstrat` has a false-positive rate that is no longer outside of the confidence interval for the 5% level when many components are used (more than ten in our simulations). This however goes along with a consequent loss of power. This might be the reason why Price et al. advised a default number of ten components when using this method (Price et al. 2006). Logistic Regression is therefore more stable than `Eigenstrat` to the number of principal components used.

We also showed that the most efficient methods to account for stratification make use of Logistic Regressions. In all of the situations studied here these methods were able to maintain a proper false-positive rate and provided a good power to detect associations.

Concerning the different types of adjustments, one has to note that the Regressions adjusted on the real population labels may not be applicable in every situations since an accurate information about the sample ancestries is not always available. If the information available is not accurate enough then estimated labels may be more informative about the homogeneous sub-groups and should be used instead (Barnholtz-Sloan et al. 2008).

We also investigated alternative Regression based approaches that were not discussed in the Results section but that are closely related to the main approaches we presented. First, we investigated another method combining the use of estimated population labels and principal components to adjust the association test (Li and Yu 2008). This method was not different than using only the principal components in our data. The rationale invoked by Li et al. to use both adjustments to respectively account for discrete and admixed populations is however pertinent making this method a reasonable option when the population labels can be accurately estimated. In addition, we investigated the use of estimated population probabilities instead of the discrete labels which showed that both methods are equivalent.

Another important question is how the methods behave when the sampling proportions become more imbalanced between the sub-populations. We addressed this question in the sixth scenario that highlighted the fact that Regressions and `Eigenstrat` were the methods capable of correcting for stratification even with very imbalanced samplings. In the extreme cases where all the cases are from one population only, we observed that considering only the cohort composed of both cases and controls by excluding the cohort with controls only was as powerful as considering all the samples. This highlights that adjusted Logistic Regressions and `Eigenstrat` are performing well enough so that they can deal with extreme sampling within sub-populations.

New sequencing methods allow to focus on DSL with very low minor allele frequency ($\leqslant 1\%$). In order to determine the quality of the methods to account for stratification with such rare DSL we simulated additional datasets corresponding Scenarios 4 and 5 (**Figure 2.14 and 2.15** respectively). It appeared that the approaches considered had the same behavior than with more important minor allele frequencies but they all experienced a loss of power. This loss of power is expected when testing a non-stratified association with low minor allele frequency and our results have confirmed the findings of (Zhang et al. 2008) that is it still the case with stratification.

Finally, we expect that when the number of SNPs available in a study increases, the information about the structure of the populations and therefore the quality of the cor-

rections of all the methods also increase. This is confirmed by the comparisons conducted in (Zhang et al. 2008, Wu et al. 2011) that considered more than 10,000 SNPs. When a certain amount of SNPs is reached, usually tens of thousands, the information provided by additional SNPs becomes redundant (e.g. because of linkage disequilibrium) and the corrections are no longer better. Also, when the amount of SNPs included is not important enough, usually less than a couple of hundreds, the methods are not provided with enough information to properly account for stratification.

Method	Type	No Strat		Admixture		Discrete Strat		Hierarchical	
		FP	Power	FP	Power	FP	Power	FP	Power
Trend	None	++	++	-	.	-	.	-	.
Reg (PCs)	C	++	++	+	++	++	++	++	++
Reg (Real Pop)	D	++	++	++	+	++	++	++	++
Reg (Est pop)	D	++	++	-	.	++	++	++	++
`Eigenstrat`	C	++	++	+	++	++	++	-	.
GC	C	++	++	-	.	-	.	-	.
Meta	D	++	+	++	+	++	+	-	.

Table 2.8: **This table summarizes the results of our study in terms of false-positive rate and power.** A '++' implies a very good performance, a '+' a good performance, a '-' a bad performance and a '.' that it was not possible to assess a comparable power given that the false-positive rate was not correct. FP: false-positive rate, C: continuous, D: discrete.

To conclude, we summarize the performances of the main methods analyzed in our study for all the types of population structure (**Table 2.8**). Given the results we presented, we recommend to use, whatever the population structure, an adjusted Logistic Regression model. The adjustment on the principal components is the more advantageous as it always leads to a correction of the bias. Moreover, principal component analysis can always be applied to the genetic data without any previous knowledge on the structure. If one has some accurate information on sample labels, then a joint adjustment with the principal components should provide an even better correction. For this reason, we also focused our research on clustering algorithms to estimate population labels. **Chapter 3** will present a novel clustering strategy that we developed to this end.

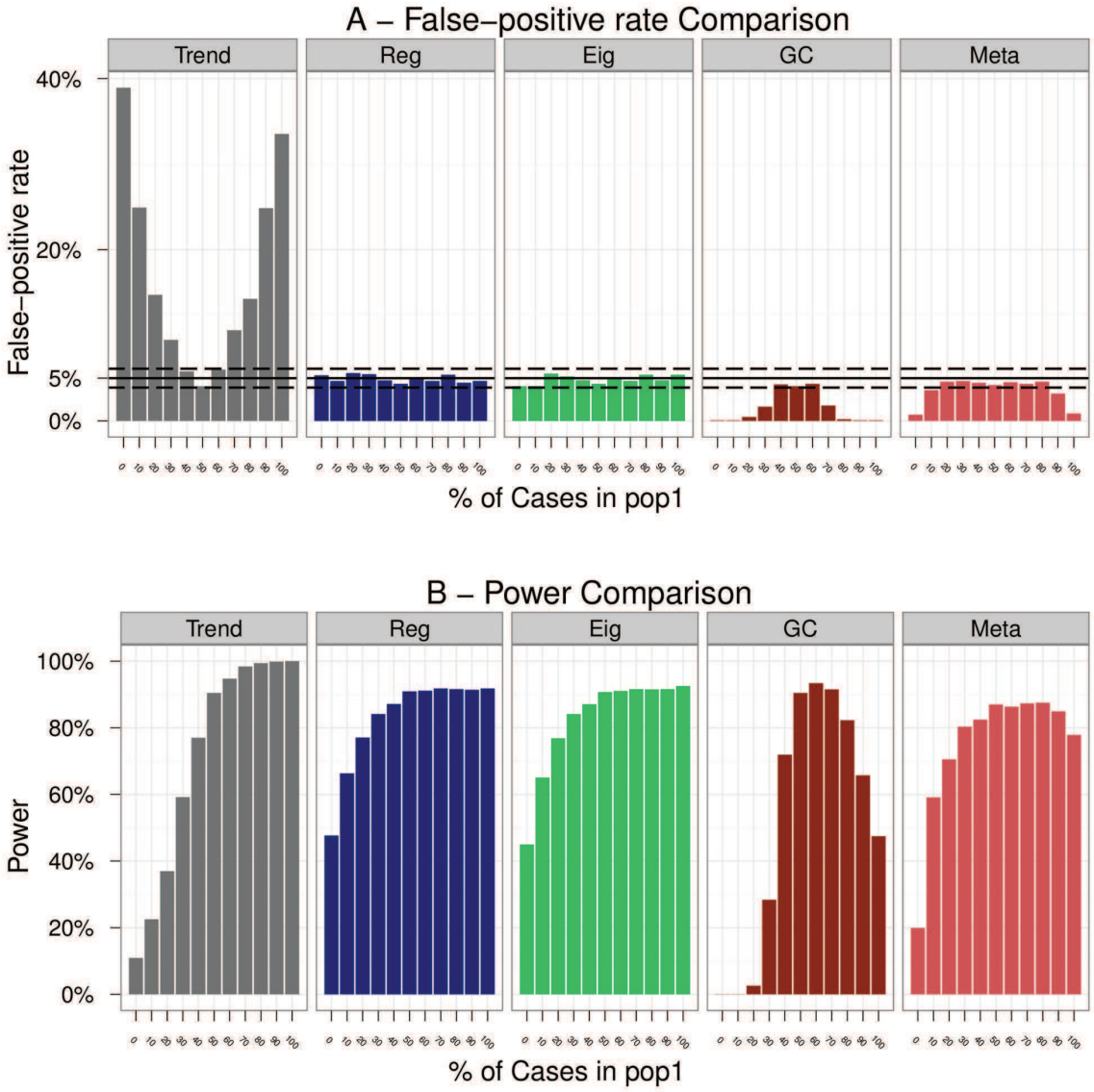

Figure 2.10: **Scenario 6 (Varying proportions of cases/controls).** A - False-positive rates of the methods. The plain black line represents the 5% level at which the tests were conducted. The dashed black lines are the 95% confidence intervals for this level. B - Evolution of the power (with RR = 2) of the methods in function of the proportion of cases in the first population (pop1). Note that all the Regression methods being equivalent for this scenario, we summarize the results for these methods under the name 'Reg' only.

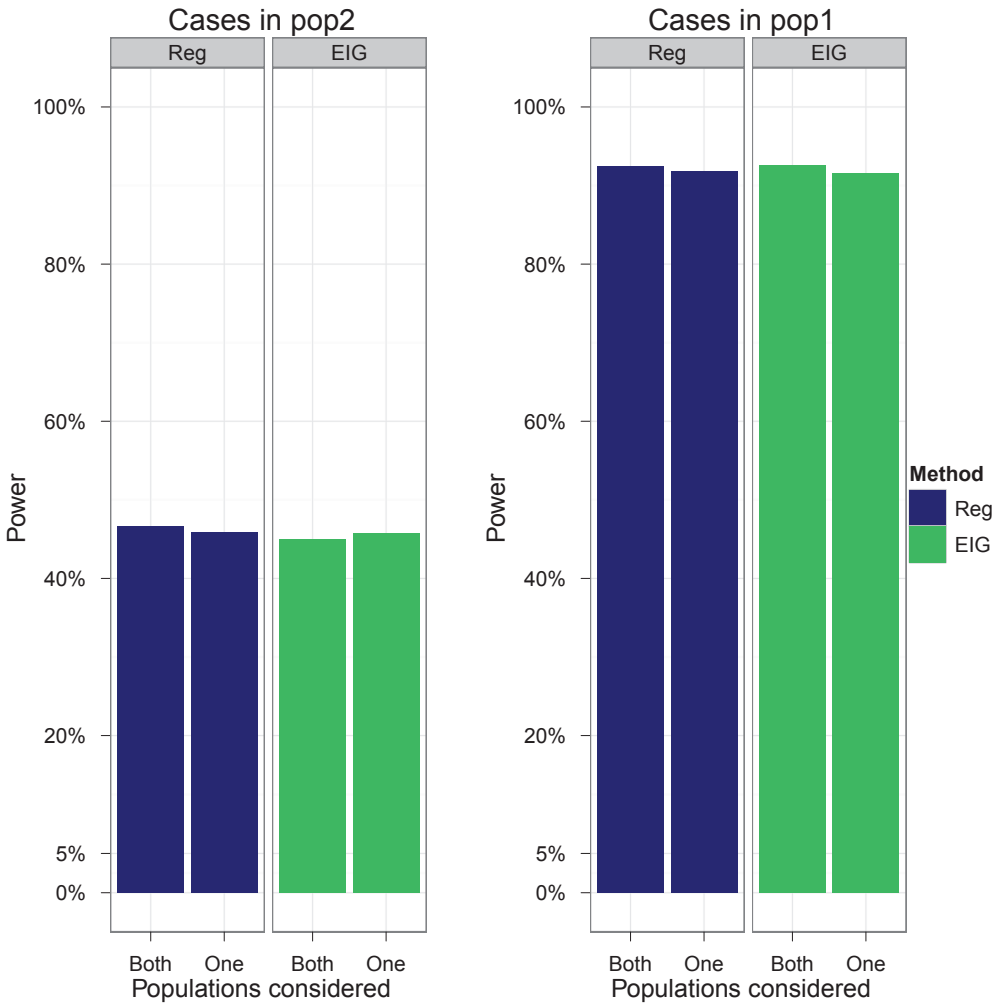

Figure 2.11: **Power comparison for scenario 6 with one or two cohorts.** The powers of Reg and EIG are represented when keeping the two populations (Both) or when excluding the population with only controls (One). On the left hand all the cases are in pop2 (less affected by the disease) and on the right hand all the cases are in pop1 (more affected by the disease).

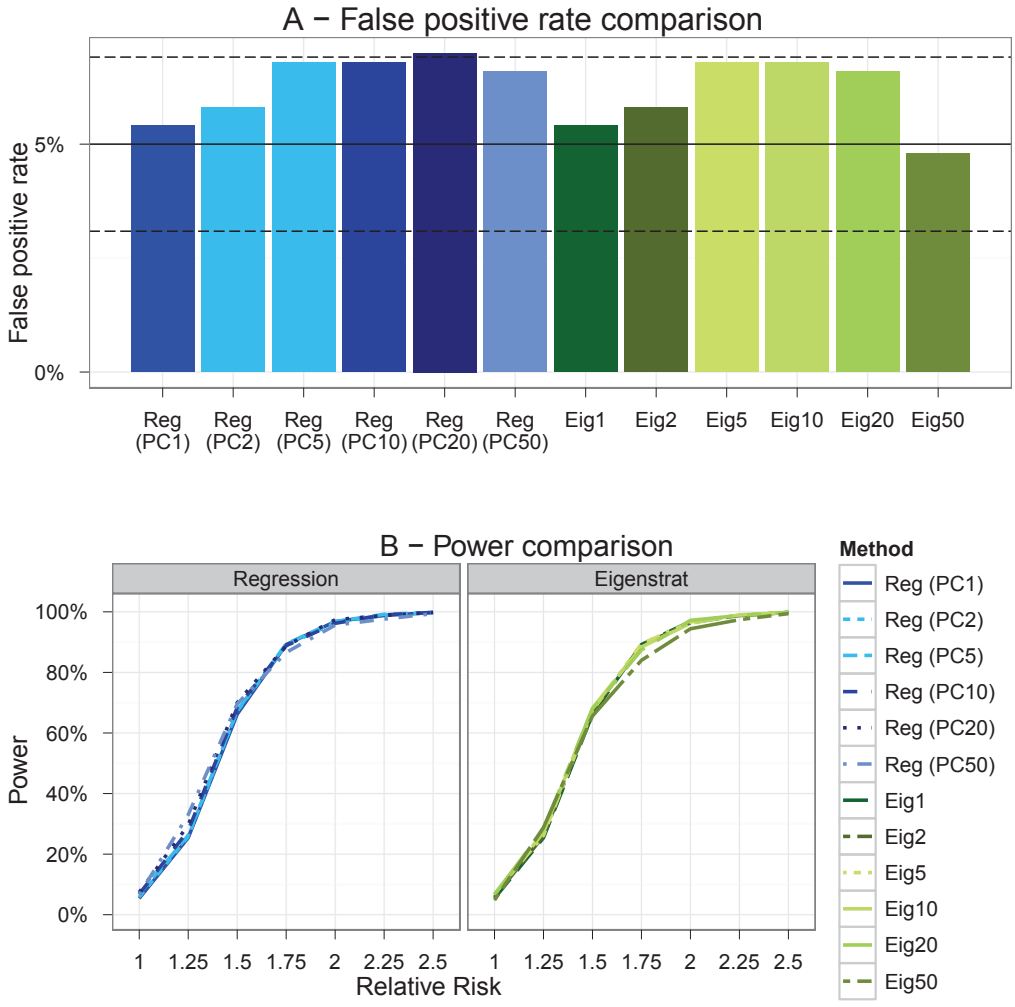

Figure 2.12: **Comparison of principal component based methods with varying numbers of components included in the models (Scenario 4).** A - False-positive rates of the methods. The plain black line represents the 5% level at which the tests were conducted. The dashed black lines are the 95% confidence intervals for this level. B - Powers of the methods in function of the increasing relative risk.

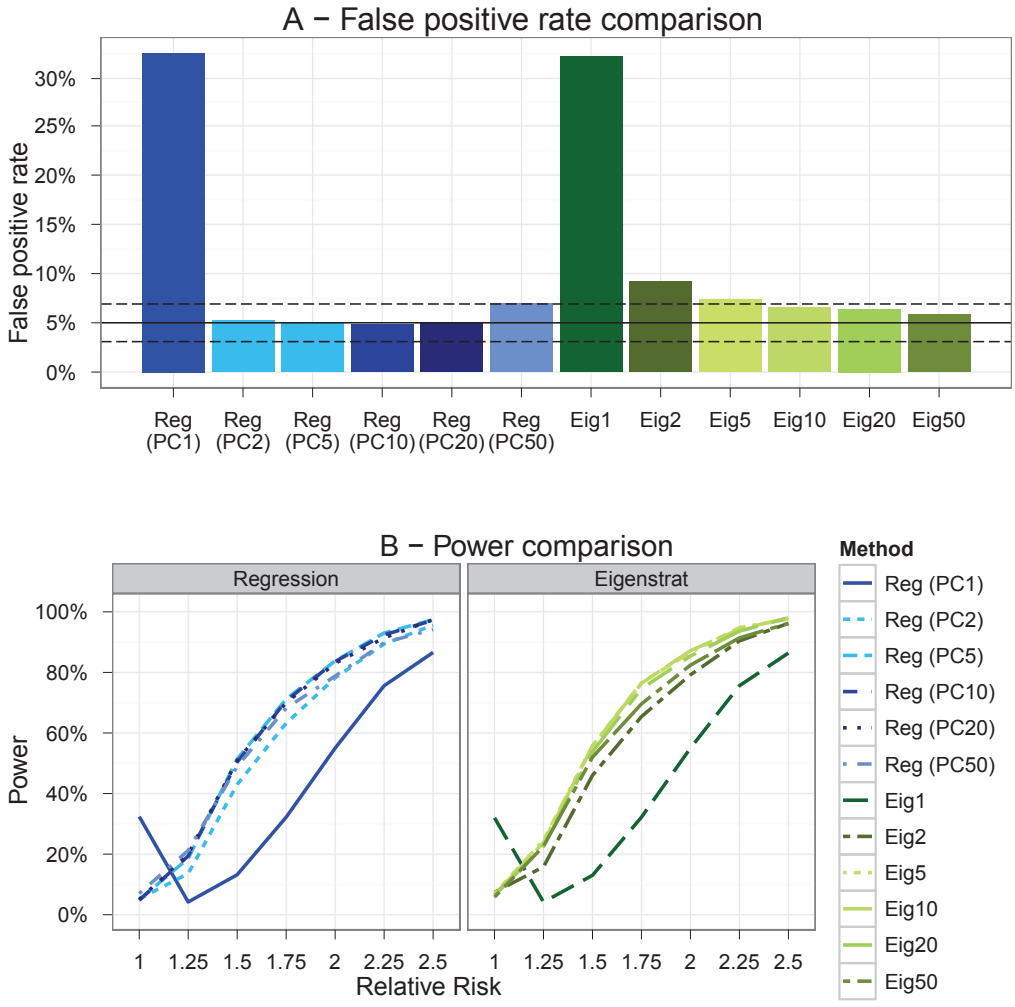

Figure 2.13: **Comparison of principal component based methods with varying numbers of components included in the models (Scenario 5).** A - False-positive rates of the methods. The plain black line represents the 5% level at which the tests were conducted. The dashed black lines are the 95% confidence intervals for this level. B - Powers of the methods in function of the increasing relative risk.

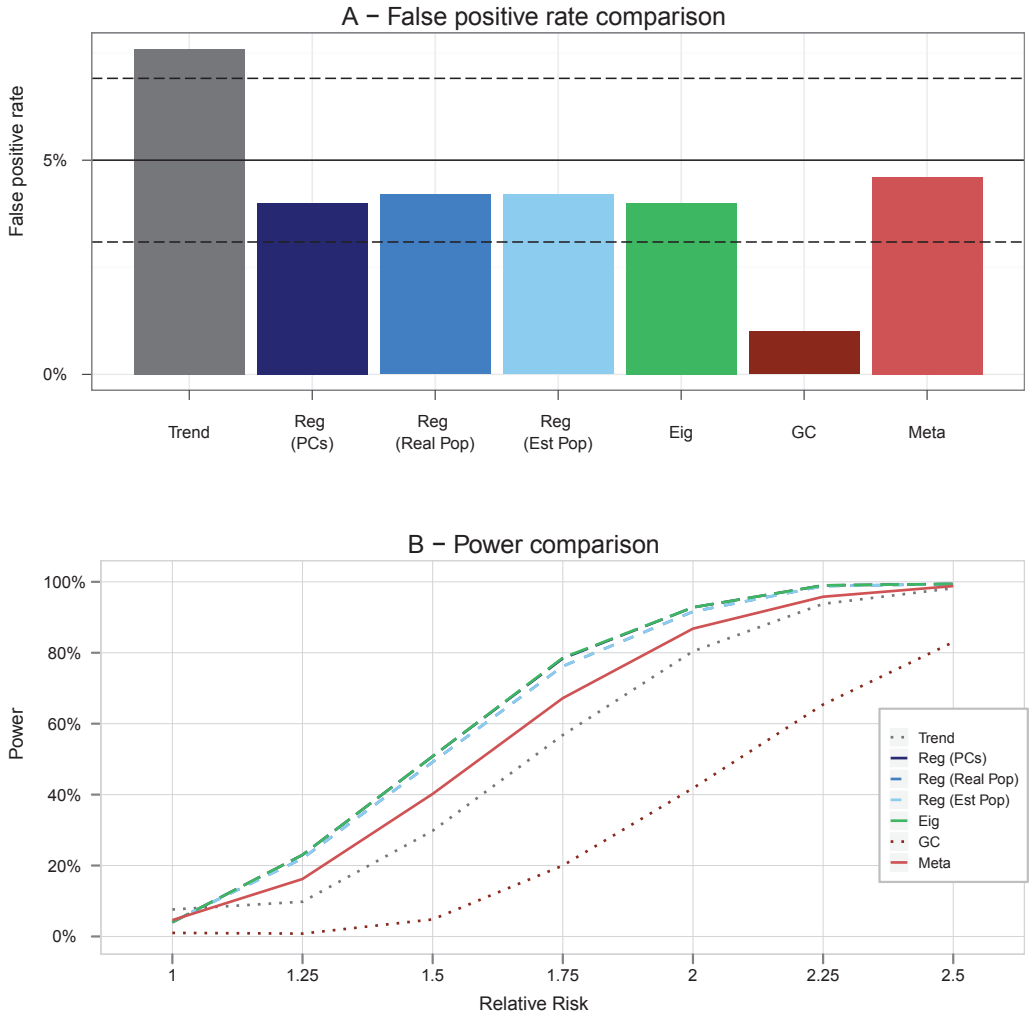

Figure 2.14: **Comparison of the methods with low minor allele frequency (Scenario 4).** A - False-positive rates of the methods. The plain black line represents the 5% level at which the tests were conducted. The dashed black lines are the 95% confidence intervals for this level. B - Powers of the methods in function of the increasing relative risk.

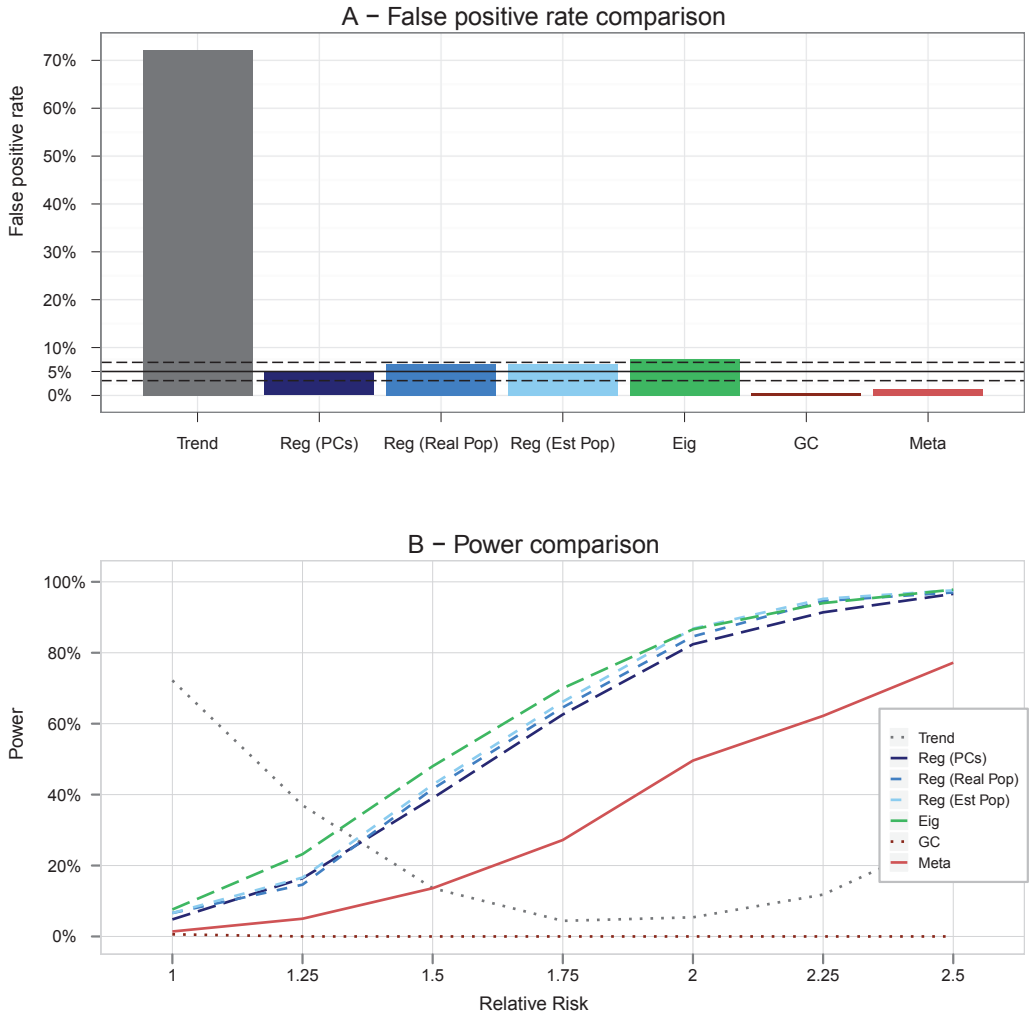

Figure 2.15: **Comparison of the methods with low minor allele frequency (Scenario 5).** A - False-positive rates of the methods. The plain black line represents the 5% level at which the tests were conducted. The dashed black lines are the 95% confidence intervals for this level. B - Powers of the methods in function of the increasing relative risk.

Chapter 3

Inference of Population Structure

In this third chapter we analyze the issues arising to infer the structure of populations along with the different strategies that have been designed to achieve this task. We also present an important part of this PhD work that is the development of an innovative clustering algorithm to determine and analyze population structure.

In a first section we introduce what it means to infer the population structure along with the difficulties that one can encounter to do so. Indeed the type of population structure and the type of data can have an influence on the inference techniques.

We then present the main strategies aiming to uncover the structure of populations and the statistical concept behind them.

The third section will pertain to the method that we developed: Spectral Hierarchical clustering for the Inference of Population Structure (`SHIPS`).

We then compare this novel strategy to the main existing approaches and discuss the different methods. Our novel clustering approach has been published in PLOS ONE (Bouaziz et al. 2012b).

3.1 Introduction

Why the inference of population structure

Identifying the underlying structure of populations is often of use for genetic research. It allows the study of evolutionary relationships between populations as well as learning about their demographic histories (Cavalli-Sforza 1994, Bowcock et al. 1994, Mountain and Cavalli-Sforza 1997, Pritchard et al. 2000a, Lee et al. 2009).

Such analyses are also of a great interest for population-based genetic studies such as Genome-Wide association studies. Notwithstanding the widespread usage of GWASs,

we have seen that their findings have been criticized partly because they are vulnerable to population stratification. Inferring the structure of the populations can therefore be helpful to identify whether there is indeed a structure or to define homogeneous clusters of individuals that can later be use to correct the association test and account for stratification. We presented in the **Chapter 2** several adjusted methods that use homogeneous clusters of individuals to correct for stratification. For instance these clusters can be used for adjusted Regressions as well as for Structured Association or Meta-Analyses.

Two main general approaches

Two major strategies have been developed to infer the structure of the populations that are parametric model-based clustering and non-parametric clustering.

Model-based clustering approaches make numerous assumptions on the genetic data and use statistical inference methods to assign individuals to sub-populations. Many of these parametric approaches exist such as for instance `Structure` (Pritchard et al. 2000a), `Admixture` (Alexander et al. 2009, Alexander and Lange 2011), `BAPS` (Corander et al. 2003) or `FRAPPE` (Tang et al. 2005). These parametric methods are the most commonly used in practice to infer population structure.

Non-parametric approaches have the advantage over parametric ones of making fewer assumptions on the data. The main non-parametric methods are `AWclust` (Gao and Starmer 2007) using a distance-based hierarchical clustering or `ipPCA` (Intarapanich et al. 2009) using iterative principal component analysis (PCA).

It is also possible to apply clustering algorithms, such as for instance a Gaussian mixture model-based clustering, to the principal components resulting from a PCA applied to genetic data (Lee et al. 2009). We refer to this particular method as `PCAclust` in the following.

Issues when inferring the structure

A major difficulty to infer the structure of populations is the dimension of the data. Indeed, the number p of marker is very important and usually far greater than the number of samples n. It is as a consequence difficult to genetically cluster individuals due to this constraint. Certain methods use a pruning technique to reduce the number of markers previously to conduct a clustering. Others reduce the dimension of the data by considering principal component analysis or similarity matrices.

The type of population structure also influences clustering method strategies. For discrete structure is it natural to provide a discrete classification of the individuals in homogeneous sub-populations. For admixtures, such a classification is no longer appropriate. Indeed, there are no discrete sub-populations to be identified in admixtures. Instead, several methods estimate what we call admixture proportions. Given a certain number of hypothetical ancestral populations, the admixture proportions represent the proportions of each individual's genome coming from each of these ancestral populations. We briefly introduced this notion when we presented the Structured Association method. Usually parametric algorithms are more adapted to the estimation of admixture proportions than non-parametric algorithms.

3.2 Approaches to infer population structure

We present in this section the main approaches to infer the structure of populations.

3.2.1 Structure

`Structure` is a parametric algorithm that uses Bayesian statistical inference to cluster individuals from genotype data or to determine admixture proportions (Pritchard et al. 2000a). It is part of the Structured Association method that we introduced in **Section 2.3.3**. Different statistical models are associated with each endgame of the method. Both models assume that the Hardy-Weinberg equilibrium is in effect in the estimated sub-populations. The other assumptions on the data concern the distributions of the different parameters indicated hereafter.

The model without admixture

Let K be the number of sub-populations from which were sampled the n individuals, $Z = (Z_1, ..., Z_n)$ the unknown vector of population labels and $P = (p_{kjl})_{\substack{1 \leqslant k \leqslant K \\ 1 \leqslant j \leqslant p \\ 1 \leqslant l \leqslant 2}}$ the frequency of allele l at locus j in population k. X represents here the genotype matrix of bi-allelic unlinked markers.
Bayesian inference is used to obtain the distribution of $Pr(Z, P \mid X)$. This is done using the posterior distribution

$$Pr(Z, P \mid X) \propto Pr(Z)Pr(P)Pr(X \mid Z, P),$$

where an uniform prior is chosen for Z and a Dirichlet prior for P

$$\forall\ i = 1 \dots n,\ \forall\ k = 1 \dots K,\ P(Z_i = k) = 1/K,$$

$$\forall\ k = 1 \dots K,\ \forall\ j = 1 \dots p,\ \forall\ l = 1 \dots 2,\ p_{kjl} \sim \mathcal{D}(\lambda_1, \lambda_2),$$

and $Pr(X \mid Z, P)$ is given by

$$\forall\, i = 1 \ldots n,\ j = 1 \ldots p,\ \forall\, l = 1 \ldots 2,\ Pr(x_{ij}^{(a)} = l \mid Z, P) \text{ is defined by } p_{z_i j l},$$

where $x_{ij}^{(a)}$ is the copy of allele a for sample i at locus j.

The posterior distribution $Pr(Z, P \mid X)$ cannot be computed exactly but observations $(Z^{(1)}, P^{(1)})$, ..., $(Z^{(M)}, P^{(M)})$ can be approximated using Markov Chain Monte Carlo (MCMC) estimations.

The model with admixture

To account for admixture, a new parameter is introduced in the model. The parameter $Q = (q_{ik})_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant k \leqslant K}}$ represents the proportion of individual i's genome that originated from population k. For each individual, this parameter follows a Dirichlet prior distribution as well. The quantity that is estimated with MCMC simulations are then $(Z^{(1)}, P^{(1)}, Q^{(1)})$, ..., $(Z^{(M)}, P^{(M)}, Q^{(M)})$.

Estimation of the number of clusters

The two models previously introduced allow one to estimate the populations of origin of the individuals in the case of a known number of populations K. In practice, this number is unknown and has to be estimated. Pritchard et al. propose a way of estimating K using a Bayesian approach, therefore the posterior distribution

$$Pr(K \mid X) \propto Pr(X \mid K) Pr(K).$$

Problems arise to compute $Pr(X \mid K)$ reason why they also propose an *had oc* solution that is selecting the number of clusters K maximizing an estimation of $log(Pr(X|K))$. This information is given in the output of the program along with the admixture proportions or the population labels.

Extension of the initial software

Several extensions of the original version of `Structure` have been developed. They include a novel model, called linkage model, allowing markers to be in linkage disequilibrium. Also they propose the F-model that assumes that different prior distributions for the allele frequencies are possible to account for the dependence of the frequencies between population (Falush et al. 2003).

Another extension of the algorithm was also designed to consider a special type of markers, the dominant markers (Falush et al. 2007).

Finally, the last addition to the original method includes information about the sampling location of the individuals when it is available and informative (Hubisz et al. 2009).

Advantages and limitations

The advantages of the method proposed by Pritchard et al are that it provides rather reliable estimations of the population of origin. It also allows for admixture in a way that can be useful to account for cryptic population structure. We presented the algorithm for bi-allelic markers, however it can be use with markers having more than two alleles.

The majors drawbacks of this method are the computational time and the reliability of the number of estimated populations (K). The MCMC runs of the algorithm are rather long to compute which makes it difficult to apply `Structure` on large datasets.
The problem of the number of population is a more delicate one. One has to keep in mind that the clusters estimated by `Structure` may not represent the real sub-populations but hypothetical ancestral populations. Recent algorithms such as Structurama (Huelsenbeck and Andolfatto 2007) allow a better estimation of K and can be combined with the clusterings estimated by the program to provide more accurate estimations of the population ancestries.

Also, like every parametric model, `Structure` is quite sensitive to the priors used. It is important to have some minimal knowledge about the data to consider prior distributions and parameters that fit.

3.2.2 Admixture

Statistical model

`Admixture` is another very popular parametric algorithm that estimates admixture proportions (Alexander et al. 2009, Alexander and Lange 2011). `Admixture` uses the same statistical model as `Structure` however instead of sampling priors using MCMC estimations, the program directly maximizes the likelihood. `Admixture` assumes the Hardy-Weinberg equilibrium between the unlinked markers.

To estimate the parameters Q and P, the algorithm maximizes the log-likelihood

$$\mathcal{L}(Q,P) = \sum_{i,j}\{x_{ij} log\big(\sum_{k=1}^{K} q_{ik}p_{kj1}\big) + (2-x_{ij}) log\big(\sum_{k=1}^{K} q_{ik}(1-p_{kj1})\big)\}.$$

In order to accelerate the estimation several statistical techniques are employed. A block relaxation algorithm is used to conduct the optimization (de Leeuw 1994). The convergence of this algorithm is accelerated using a quasi-Newton method and the standard-errors of the parameters are calculated using a moving block bootstrap method.

Estimation of the number of clusters

The estimation of the number of clusters K is conducted by a cross-validation technique. This method partitions the initial data into several subsets that are used to estimate K and to validate the estimation.

Extension of the algorithm

`Admixture` is a recent program however an extension has already been developed to include information about the sample ancestry into a supervised model.

Comments

`Admixture` has rapidly become a popular algorithm to infer the population structure. The computational time is greatly improved compared to `Structure` as well as the number of parameters involved that is diminished as no Bayesian estimation is conducted.

As `Admixture` uses unlinked markers, the authors advise to use a pruning method, to reduce the linkage disequilibrium between the markers, prior to performing the ancestry inference.

3.2.3 Other parametric approaches

Several other parametric methods exist and can be consulted in (Satten et al. 2001, Wang 2003, Corander et al. 2004, Purcell and Sham 2004, Tang et al. 2005, Chen et al. 2006, Reeves and Richards 2009, Shringarpure and Xing 2009). These methods are based on Bayesian models, mixed models or latent class models. It is also interesting to cite an algorithm recently published that is based on a parametric similarity matrix (Lawson et al. 2012). Many of these programs are quite time consuming or are not able to assess certain cryptic structure which is the reason why the development of novel clustering algorithm is an ongoing research matter.

3.2.4 AWclust

The `AWclust` (Allele sharing distance and Ward's minimum of variance hierarchical clustering) algorithm is a non-parametric method based on a hierarchical clustering (Gao and Starmer 2007 2008). This method is composed of three steps.

1. A distance matrix between all pairs of individuals is computed. This is a dissimilarity matrix based on the allele sharing distance (ASD). The dissimilarity at SNP l between samples i et j is

$$d_{i,j}(l) = \begin{cases} 0 & \text{if same genotype} \\ 1 & \text{if one common allele} \\ 2 & \text{if no common allele} \end{cases} .$$

2. A hierarchical clustering is applied to the distance matrix. Initially each individual forms a single cluster. The clusters are progressively merged following Ward's minimum of variance criterion until all samples are in the same cluster.

3. The estimation of the number of clusters K is conducted using a gap statistic method (Tibshirani et al. 2000). The novel algorithm that we developed during this PhD is also based on a gap statistic. We will present this estimation method in the next section dedicated to the `SHIPS` algorithm.

   Note that the `AWclust` algorithm limits the maximum number of clusters to 16 in order to reduce the computational cost of the method.

3.2.5 Principal component analysis and clustering

We presented in **Chapter 2** the principal component analysis and evoked the possibility of using the axes of variation to cluster individuals into homogeneous sub-populations. Several clustering algorithms can be used on the principal components such as a classical K-means or a Gaussian mixture model (GMM) clustering (Lee et al. 2009). We present here the method called `PCAclust` that uses this latter clustering strategy.

PCAclust method

The `PCAclust` algorithm is composed of the 3 following steps.

1. A principal component analysis is applied to the genotype data to compute the principal components (PCs). This step is conducted using the software `Eigensoft` using the LD option that replaces each SNP by the residual of its regression on the two preceding SNPs.

2. A set of significant PCs, (PC_1,...,PC_m), is selected using the Tracy-Widom statistic at a 5% level.

3. A Gaussian mixture model clustering algorithm is applied to the selected PCs to cluster the samples. As a matter of fact the principal components are normally distributed which leads to a good fit with the GMM clustering. **Figure 3.1** is an

example of the normality of the principal components that also shows the mixture of the normal distributions between the three populations considered.

The estimated number of clusters is computed so that the likelihood of the model is maximized.

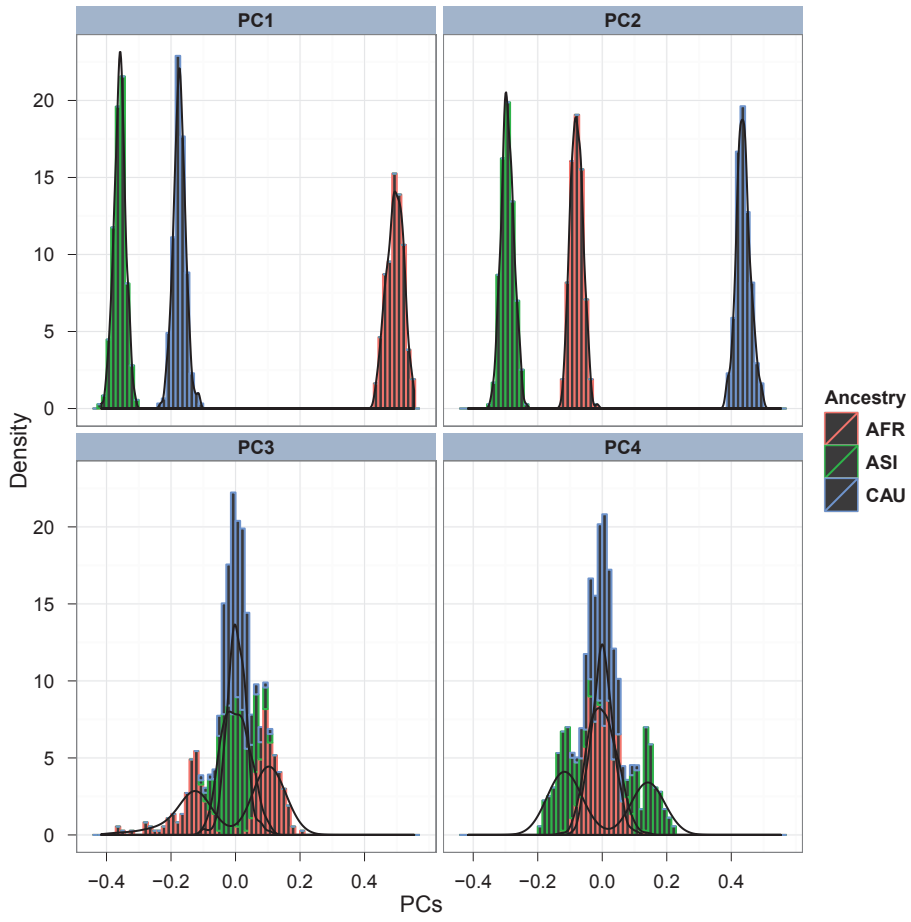

Figure 3.1: **Normal distribution of the principal components.** A genetic data composed of three HapMap populations is used. The three main ethnicities are represented by the populations CEU (Caucasian), MKK (African) and CHB (Asian). The principal components are plotted as histograms and display a mixture pattern of 3 distributions representing each ethnicity. Each of these distribution is colored according to the ethnicity.

3.3 SHIPS: Spectral Hierarchical clustering for the Inference of Population Structure

We present in this section our novel non-parametric distance-based clustering approach based on a divisive hierarchical clustering method. Our method is based on the idea that it might not be possible to uncover all of the structure in the data when applying a clustering algorithm just once. Fine population structures may not be detected as the corresponding sub-populations are hidden within the major sub-populations detected by the first run of the algorithm.

We therefore implemented a robust statistical framework to iteratively apply a clustering algorithm to the data and so analyze in depth the genetic patterns of the studied populations. This corresponds to a divisive hierarchical clustering strategy. Based on a pairwise distance matrix, the algorithm progressively divides the original population in two sub-populations by the use of a spectral clustering algorithm. The process is then iterated in each of the two sub-populations and so on. This leads to the construction of a binary tree, where each node represents a group of individuals. To determine the final clusters, a tree pruning procedure and an estimation of the optimal number of clusters are applied. In such an approach, both the final clustering of the individuals and the number of clusters are estimated by the method. We call our method 'Spectral Hierarchical clustering for the Inference of Population Structure' (`SHIPS`).

3.3.1 The SHIPS algorithm

`SHIPS` can be described in several steps that are graphically represented in **Figure 3.2**.

1. Computation of a distance matrix that is a similarity matrix S between each pair of individuals. This matrix is used for the next steps of the algorithm.

2. Creation of a binary tree. Each population is divided in two sub-populations and so on (**Figure 3.2-A**).

3. Pruning of the tree to keep only the relevant branches corresponding to the relevant divisions (**Figure 3.2-B**).

4. Estimation of the optimal number of clusters K to determine which clusters of the tree are the final ones (**Figure 3.2-C**).

3.3.2 Similarity matrix

`SHIPS` is based on a spectral clustering algorithm. A similarity matrix is therefore necessary to apply this clustering method. We decided to consider a similarity matrix based

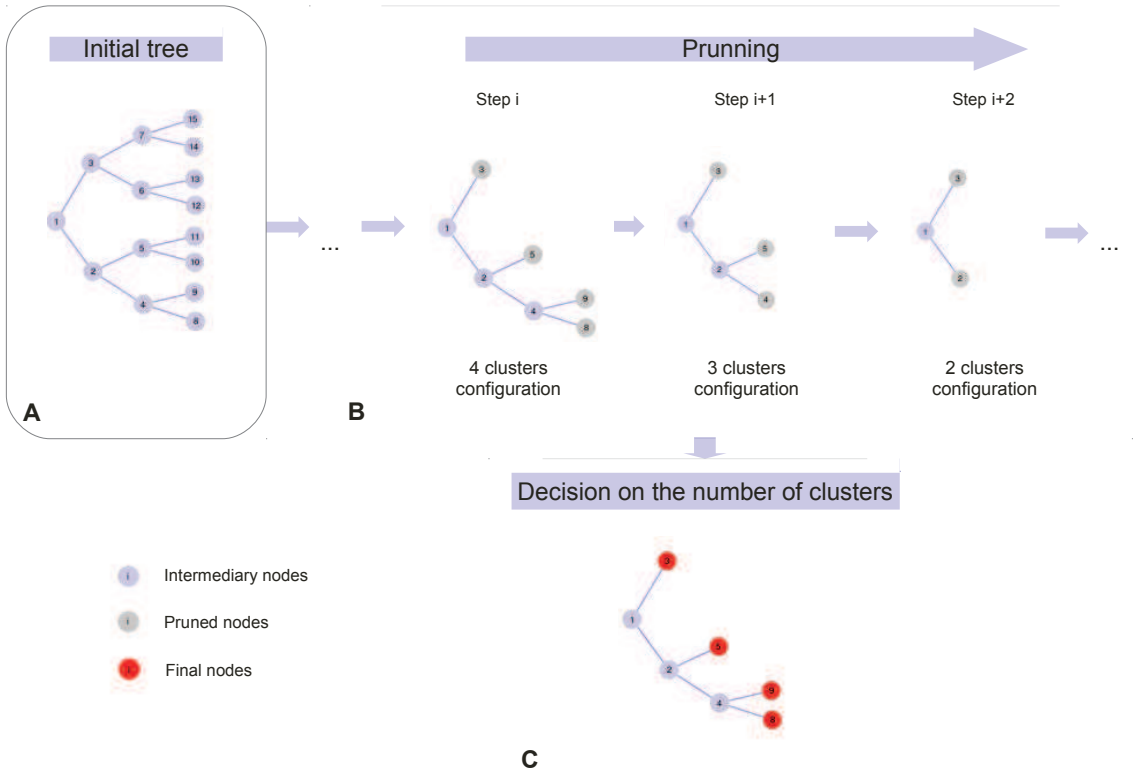

Figure 3.2: **Graphical example of the SHIPS algorithm.** After that the initial binary tree is built (A), the pruning procedure leads at the end of each step to a possible clustering of the individuals. In this example the data is clustered in four, then three then two clusters (gray nodes) at step i, $i+1$ and $i+2$ respectively (B). The final clusters decided by the gap statistic correspond to the ones of the four classes clustering (red nodes) (C).

on the allele sharing distance that has been previously used to identify genetic patterns among populations (Mountain and Cavalli-Sforza 1997, Gao and Starmer 2007). This matrix represents how close the genomes of each pair of individuals are. The similarity at SNP l between samples i and j is calculated as follows

$$s_{i,j}(l) = \begin{cases} 2 & \text{if same genotype} \\ 1 & \text{if one common allele} \\ 0 & \text{if no common allele} \end{cases} .$$

The total similarity between samples i and j is

$$s_{i,j} = \sum_{l=1}^{p} s_{i,j}(l) = \sum_{l=1}^{p} (2 - |x_{il} - x_{jl}|),$$

where x_{il}, x_{jl} are the sample genotypes coded 0, 1 or 2 according to the number of reference alleles present at the locus l. The final matrix $S{=}(s_{i,j})_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant j \leqslant n}}$ is a squared matrix of dimension $n \times n$, n being the number of individuals.

One has to note that any pairwise similarity matrix could be used in the algorithm instead of the one presented here. Examples of such matrices, based for instance on haplotypes instead of genotypes, are presented in (Li and Yu 2008, Browning and Browning 2010, Lawson et al. 2012, Lawson and Falush 2012). We decided the choice of this similarity measure as it is fast to compute and it led to high empirical performances of the algorithm.

3.3.3 Creation of a binary tree with successive spectral clustering algorithms

The binary tree produced by `SHIPS` is obtained by successively dividing each population in two sub-populations using a spectral clustering algorithm. Spectral clustering methods cluster points using eigenvectors of matrices derived from the initial data. We decided to use the version of this method proposed by Ng et al. (Ng et al. 2001, Luxburg 2007) that is the normalized spectral clustering described in the three following steps.

First, the similarity matrix S computed in the previous section is transformed into its normalized laplacian L with

$$L = I - D^{-1/2} W D^{-1/2},$$

where $W = S - diag(S)$, I is the identity matrix

and D is a diagonal degree matrix such as each diagonal term $d_i = \sum_{j=1}^{n} w_{ij}$.

In a second step, a singular vector decomposition of the laplacian $L = U \Lambda U'$ is computed and the m first eigenvectors (U_1,..., U_m) are normalized to get new vectors (T_1,..., T_m), with norms of 1, defined by

$$T_i = \frac{U_i}{\sqrt{\sum_{i=1}^{m} U_i^2}}.$$

These vectors are used to cluster the points, i.e. divide a population in two sub-populations. Note that m represents here the number of desired clusters so $m = 2$ in the case of the

SHIPS algorithm.

In a third step, a clustering algorithm is applied to the new vectors (T_1, T_2) to create the two sub-populations. We decided to use a Gaussian mixture model (GMM) clustering after determining empirically that the usual $k-means$ clustering algorithm is less robust than the GMM clustering when applied to our genetic data. The GMM clustering is used in the way the $k-means$ would be, that is by strictly fixing the number of estimated clusters to $m = 2$.

If the population that we wish to split in two sub-populations is deemed homogeneous by the algorithm, the GMM clustering creates two clusters, one with all the samples and an empty one. This is a termination criterion that defines the end of a branch of the tree, called a terminal node. In extreme cases, the terminal nodes are all composed of a unique sample of the original population which ensures the convergence of the tree building step of the algorithm.

3.3.4 Pruning of the tree

The divisive strategy of SHIPS consists in dividing the original population in two sub-populations with the spectral clustering algorithm previously described and to iterate this procedure within each sub-population. This process leads to the computation of a binary tree (**Figure 3.2-A**). It is however noticeable that certain divisions are not relevant enough in terms of separating really distinct genetic populations. As a result, a pruning procedure is applied to the tree to progressively suppress the nodes, and the corresponding branches, that are the less relevant. This procedure creates several nested trees, each corresponding to a possible clustering of the individuals with a decreasing number of clusters (**Figure 3.2-B**). At the last step of the pruning, all the samples are in the same cluster.

The strategy of tree pruning that we use is the reduced error pruning. A quality indicator is defined and calculated for each node of the tree. This indicator is based on the sum of the squared similarities of a node and of its leaves. We define the function calculating the sum of squared similarities within a node A by

$$SW(A) = \sum_{i,j \in A} s_{i,j}^2,$$

where $s_{i,j}$ is the similarity previously introduced between samples i and j.

Considering a tree T, the quality of a node G which has the leaves $L(G) = (L_1, \ldots, L_d)$ is defined by

$$qual(G \mid \mathtt{T}) = SW(G) - \sum_{k=1}^{d} SW(L_k).$$

In terms of inter-cluster sums the quality can be expressed by

$$qual(G \mid \mathtt{T}) = \sum_{1 \leqslant k < k' \leqslant d} \sum_{i \in L_k, j \in L_{k'}} s_{i,j}^2,$$

which corresponds to the sum of squared similarities between the leaves of G.

At each step, the node with the lowest quality value, $G_{pruned} = \arg\min_{G \in \mathtt{T}} qual(G \mid \mathtt{T})$, is pruned along with the subtree which it is the root. The indicators are recalculated after each step to account for the new topology of the tree.

3.3.5 Estimation of the optimal number of clusters

Principle

The optimal number of clusters K is regarded as a variable that is estimated using Tibshirani et al.'s gap statistic (Tibshirani et al. 2000). This method compares a quality indicator calculated on the result of a clustering in k classes of a dataset of interest and the value that this indicator would take under its null distribution, that is when the same clustering algorithm is applied to cluster a null reference dataset in k classes also.

A range of possible numbers of clusters, $k = 1 \dots k_{max}$, is thus investigated and for each an indicator W_k is calculated. The gap statistic is defined for a clustering with k clusters by

$$Gap(k) = E[W_k] - W_k,$$

and estimated by

$$\widehat{Gap(k)} = E^*[W_k] - W_k = \frac{1}{B} \sum_{b=1}^{B} W_{kb}^* - W_k,$$

where $E^*[W_k]$ represents the expectation from the null distribution and therefore the W_{kb}^* are the quality indicators calculated on B simulated null reference datasets. The simulation process for these datasets is described hereafter.

Several possible estimations of the optimal number of clusters K exist (Dudoit and Fridlyand 2002). The one we use is $\hat{K}$, the smallest k such as

$$Gap(k) \geq Gap(\tilde{k}) - s_{\tilde{k}},$$

where $\tilde{k} = \arg\max_k Gap(k)$ and $s_k = sd((W_{kb}^*)_{1 \leqslant b \leqslant B}).\sqrt{(}1 + 1/B)$. Note that the factor $\sqrt{(}1 + 1/B)$ accounts for the simulation error of the W_{kb}^*.

Quality indicator

Let $(C_k)_{k=1,\dots,k_{max}}$ be possible clusterings of the samples in the data with k clusters in a clustering C_k. These clusterings are in our algorithm the ones determined at each step of the pruning (**Figure 3.2-B**). We call W_k the quality indicator calculated on the clustering C_k. If we denote $C_k = (D_1, \dots, D_k)$, where D_r is the $r-th$ cluster of C_k, then the indicator that we consider is

$$W_k = \sum_{r=1}^{k} \frac{1}{2.|D_r|}.\Sigma(D_r),$$

where $\Sigma(D_r)$ is the sum of the squared dissimilarities between the samples of the $r-th$ cluster of C_k and $|D_r|$ its cardinal (i.e. the number of samples in D_r). The dissimilarities are calculated like for the `AWclust` algorithm that is by inverting the values (0 if the samples have the same genotypes and 2 if they have no common alleles) compared to the similarities.
In the classical version of the gap statistic, the logarithm of W_k is used however several alternatives have recently been investigated (Mohajer et al. 2011). We decided to use the aforementioned criterion as we observed that it led to a better estimation of the number of clusters for both the simulated and real genetic data that we used to assess the method.

Simulation under the null distribution

To compute null reference datasets we simulate datasets with a number of variables and individuals identical to the ones of the original datasets. Each variable was taken uniformly within $\{0, 1, 2\}$ to match the SNPs values of the original datasets. Simulated that way, the null datasets correspond to data where there is no structure of the population. This simulation choice is also the one made in the algorithm `AWclust` that uses a gap statistic method. Note that theoretically it is not necessary to match all of the features of the data, such as for example the minor allele frequency of each SNP, when simulating under the null. This choice of simulation model was motivated by the empirical performances of the corresponding gap statistic to estimate accurate numbers of clusters in our applications.

Adequacy of SHIPS and the gap statistic

`SHIPS` has the advantage of producing in one run of the algorithm nested clusterings of the samples for $k = 1 \dots k_{max}$ which renders faster the computation of the gap statistic. Note also that the quality indicator used in the gap statistic is based on a dissimilarity matrix while `SHIPS` uses a similarity matrix. This actually does not imply the computation of a new matrix, as the dissimilarity and the similarity matrix are linearly related. The gap statistic is therefore well suited to determine the optimal number of clusters with this new method.

3.3.6 Implementation

The SHIPS algorithm was implemented in R and the Mclust package was used within the spectral clustering steps to apply Gaussian mixture model clustering. A R package is freely available at http://stat.genopole.cnrs.fr/logiciels/SHIPS.

This algorithm takes as input parameters a SNP matrix of dimension $n \times p$ where n is the number of individuals and p the number of SNPs. Each entry of the matrix is coded 0, 1 or 2 given the number of reference alleles present at each locus for each sample. It is also necessary to indicate the maximum number of clusters to be investigated (denoted here k_{max}) and the number of null datasets simulated (B here) to apply the gap statistic. A default value of $B = 20$ is set in the program.

3.4 Comparison of SHIPS to other approaches

We now propose several applications of the SHIPS algorithm to SNP datasets. We considered five scenarios of simulated population structures. The software Genome (Liang et al. 2007) was used to simulate these data of increasing complexity. We also applied the method to a simulated admixed dataset that was produced using the simulation model presented in **Section 2.4.3**. In addition, we evaluated the performances of the algorithm on two real datasets, namely data from the HapMap project (Consortium 2005) and the Pan-Asian dataset (Ngamphiw et al. 2011). A comparison of our method SHIPS and some of the main approaches that are Structure, Admixture, AWclust, and PCAclust was also conducted on these datasets.

3.4.1 Evaluation of the methods

Methods included in the comparison

We compared SHIPS to some of the most commonly used clustering algorithms in the genetic field. We first considered the parametric approaches Structure and Admixture. Also we included a non-parametric approach, namely AWclust, and finally we added the alternative clustering strategy PCAclust to the comparison.

SHIPS was used with the default parameters, i.e. 20 null datasets simulated for the gap statistic. A reasonable maximum number of clusters was considered for all the methods, for instance, when analyzing a dataset with 10 (known) sub-populations we investigated up to 20 possible sub-populations.

The version 2.3.2.1 of Structure was downloaded from http://pritch.bsd.uchicago.edu/structure.html and used with 5,000 burn-ins, 5,000 runs, the admixture model and no LD model. Structure provides a way of estimating the optimal number of clusters

K through the model likelihood however it has been demonstrated that this method had shortcomings compared to more recent algorithms such as for instance Structurama (Huelsenbeck and Andolfatto 2007) that allows a better estimation of K. To consider the best use of `Structure`, we therefore decided to opt for a way of estimating the number of clusters that advantages this method. In our comparison strategy a criterion is used to compare the different programs and we considered an estimated K for `Structure` that optimizes this criterion. Also, as `Structure` provides admixture proportions under the admixture model, we decided as it is usually done that an individual was assigned to the estimated population it has the highest probability to belong. Note that with this assignment method, certain clusters computed by the admixture model might not have any individuals assigned to them. In such a situation we considered the estimated number of clusters to be the effective number of sub-populations after the assignment procedure.

The program `Admixture` was downloaded from `http://www.Genetics.ucla.edu/software/admixture`. The estimation of the number of clusters was conducted using the minimum of cross-validation error with the default parameter of 5 fold cross-validation. Like with `Structure`, we obtained discrete clusterings with this program by assigning an individual to the population it has the highest probability to belong.

The version 2.0 of `AWclust` was downloaded from `http://AWclust.sourceforge.net` and used with the default parameters and 20 simulations for the computation of the gap statistic. The estimated number of clusters was determined using the maximum of the gap statistic.

The method `PCAclust` was recoded as it is not available as a software. The PCA was conducted using the software `Eigensoft`. The `R` package `Mclust` was used to apply GMM clustering to the set of relevant principal components selected with the use of the Tracy-Widom statistic. The optimal number of clusters was estimated using the likelihood computed by `Mclust`.

Population structure scenarios

We assessed `SHIPS` and the other methods on several datasets. We considered simulated datasets where the structures of the populations were controlled, a simulated admixed dataset and real datasets to determine the performances of the different approaches in real situations. For all of these scenarios small datasets of thousands of markers and large datasets of hundreds of thousands of markers were considered. We used several replicates for the small data in order to account for the simulation process or the markers sampling. Only one was used for the large scenarios due to the computational cost of certain algorithms.

Simulated datasets

We simulated datasets using the software `Genome` based on the coalescent approach and assuming an island model of population structure. We considered a first model M1 with no structure of the population in order to determine which methods are capable of uncovering that the data is not structured. We then considered 4 structured models, M3, M5, M10 and M20 with respectively 3, 5, 10 and 20 sub-populations and increasing complexities of population histories. **Figure 3.3** presents the population histories of these models and **Table 3.1** the detail of the sampling. Each small dataset is composed of 5,000 SNPs and each large dataset of 200K SNPs simulated in equal number on each of the non-sexual chromosomes. Ten datasets were simulated and analyzed by the algorithms for each small scenario. The results are then averaged over these datasets. Note also that for computational purposes, `Structure` was only applied to five small datasets and was not applied to the large ones.

Model	Samples per sub-population
Model M1 (1 sub-population)	100
Model M3 (3 sub-populations)	100
Model M5 (5 sub-populations)	50
Model M10 (10 sub-populations)	50
Model M20 (20 sub-populations)	30

Table 3.1: **Details of the simulated models.**

Simulated admixed datasets

In order to assess the performances of the various algorithms on more realistic situations we simulated a discrete admixed dataset corresponding to the model named Madx. Two real populations from the HapMap phase III data[1], namely the Han Chinese from China (CHB) and the Utah residents with Northern and Western European ancestry from the CEPH collection (CEU), were used in an evolutionary model to produce an admixed population. The evolutionary model consists in randomly mating samples from each of the two original populations and to iterate this process over time. The final dataset is composed of the two original populations (CEU and CHB) and the admixed simulated one (named XY). The detail of the sampling is provided in **Table 3.2**. Like for the other simulated datasets we considered small data of 5,000 SNPs with ten replicates and one large data of 200K SNPs.

HapMap dataset

We also focused on the potential of the methods when applied to real datasets. We

[1]The full HapMap dataset is available at `http://hapmap.ncbi.nlm.nih.gov/downloads`.

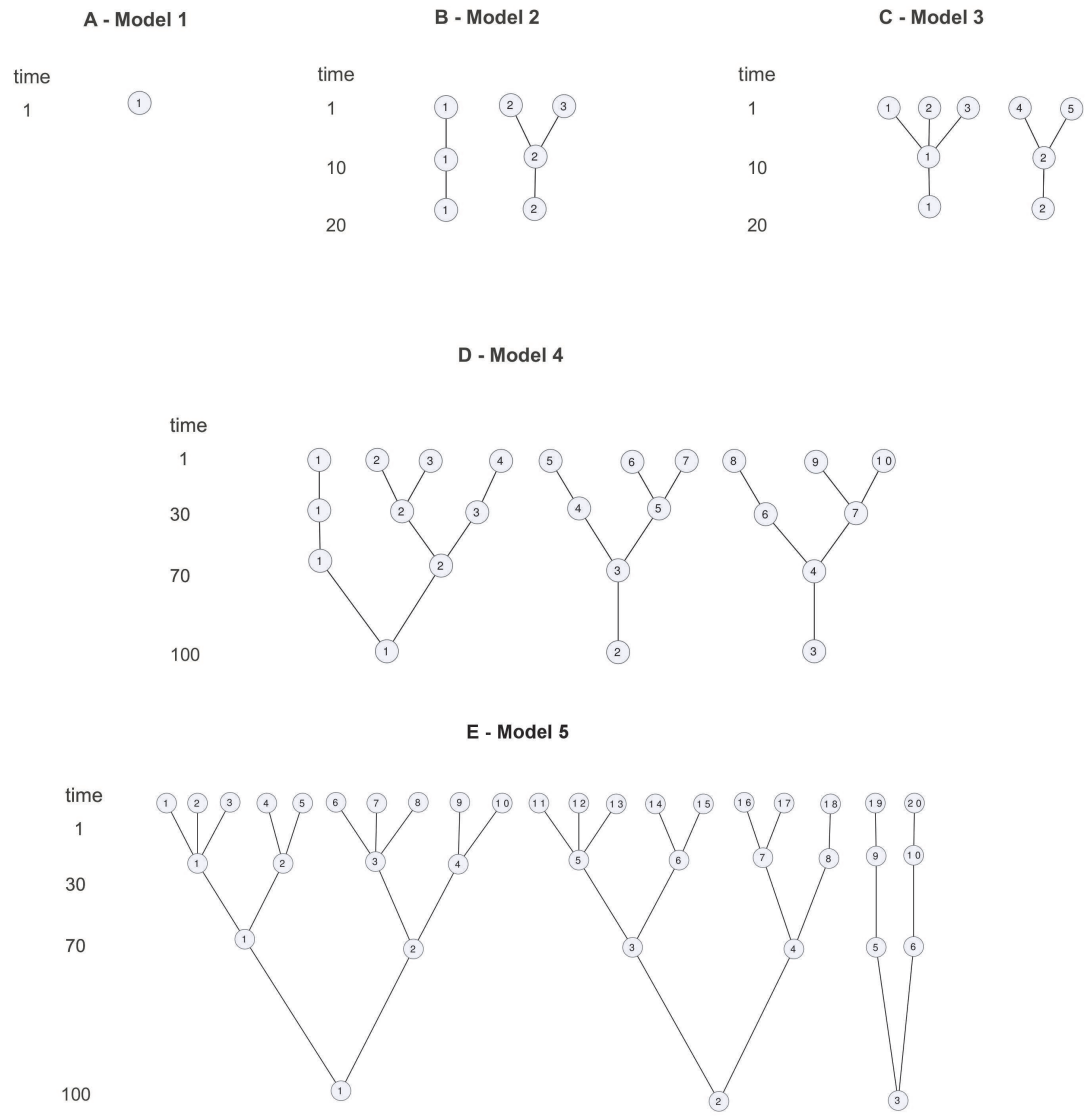

Figure 3.3: **Population history trees used to generate the simulated datasets.** A - One population. B - Three sub-populations. C - Five sub-populations. D - Ten sub-populations. E - Twenty sub-populations.

Population	# Samples
CEU	112
CHB	137
Admixed (named XY)	100

Table 3.2: **Details of the admixed dataset.**

first considered the HapMap phase III dataset with 9 populations and 1,087 individuals (**Table 3.3**). **Figure 3.4** is a graphical representation of these populations on the principal components space. We considered small data with 20,000 SNPs and large data with 220K SNPs randomly chosen among the whole set of SNPs available and in equal number on each of the non-sexual chromosomes. To account for the SNPs sampling, twenty replicates of the small HapMap data were considered to assess the methods, except for `Structure` that was only applied to five datasets.

Population	Ethnicity	# Samples
CEU	Utah residents with Northern and Western European ancestry	112
CHB	Han Chinese in Beijing, China	137
CHD	Chinese in Metropolitan Denver, Colorado	109
GIH	Gujarati Indians in Houston, Texas	101
JPT	Japanese in Tokyo, Japan	113
LWK	Luhya in Webuye, Kenya	110
MKK	Maasai in Kinyawa, Kenya	156
TSI	Toscani in Italia	102
YRI	Yoruba in Ibadan, Nigeria	147

Table 3.3: **Details of the HapMap dataset.**

Pan-Asian dataset

The PASNPi consortium provides the genotype data of 75 Pan-Asian and HapMap populations with 1928 individuals and 54,794 SNPs[2]. Among all these populations, certain main groups, defined by the countries of origin, can be highlighted. We focused on 10 subpopulations formed by 443 individuals, from each of these groups (**Table 3.4**, **Figure 3.5**) and refer to these data as the Pan-Asian datasets. Like for the HapMap data, we selected 20,000 SNPs randomly chosen in equal number on each of the non-sexual chromosomes among the initial dataset for the small data (with twenty replicates) and the whole set

[2]The complete PANSNPi dataset is available at `http://www4a.biotec.or.th/PASNP`.

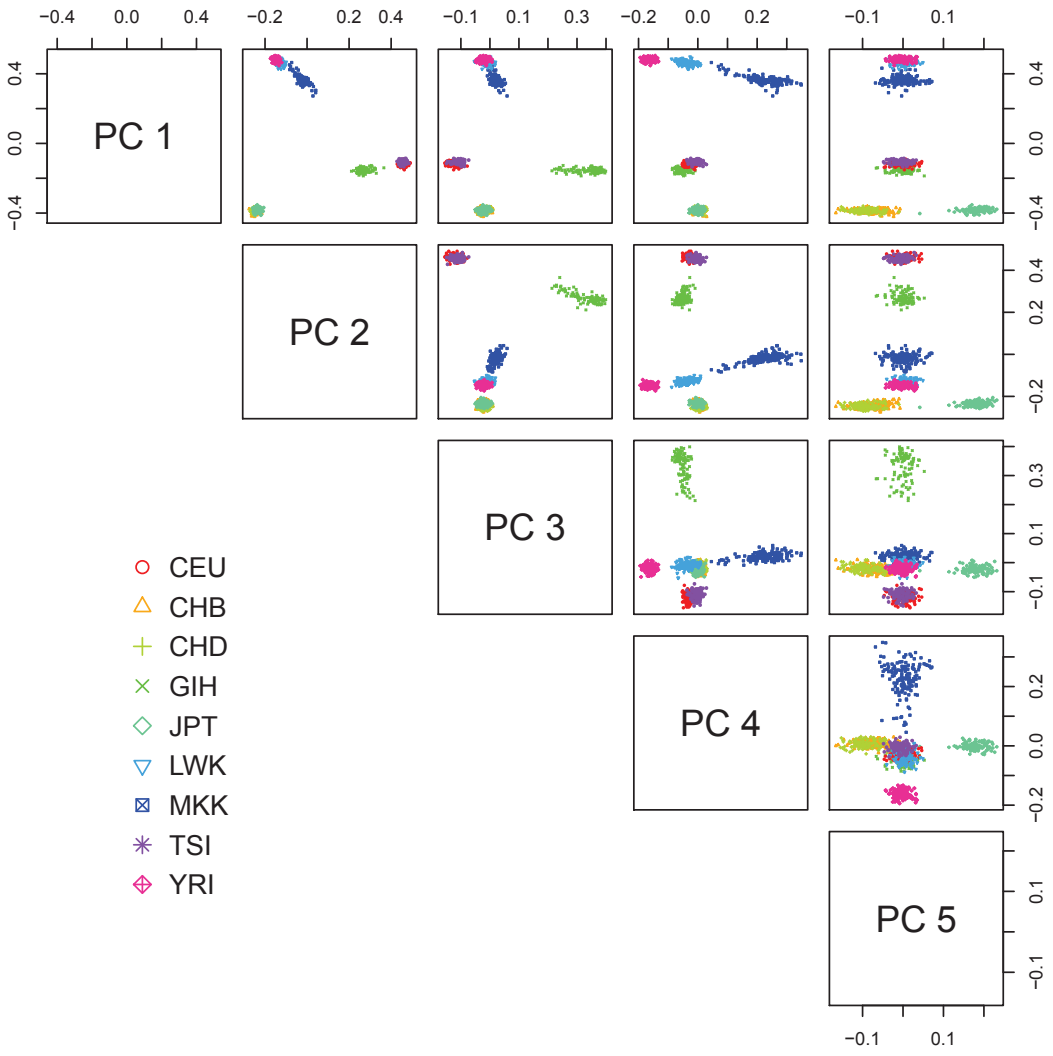

Figure 3.4: **Representation of the 9 populations of the HapMap dataset.** This scatter-plot uses the first five principal components of a dataset with 20K SNPs. This graph is only intended to present the general genetic pattern of the dataset and does not exhaustively represent the capability of the PCA to separate the populations.

of SNPs for the large data. For the reasons indicated previously, `Structure` was only applied to five small replicates.

Population	Ethnicity	# Samples
CN.WA	Wa, China	56
ID.JA	Javanese, Indonesia	34
IN.TB	Mongoloid features, India	23
JP.ML	Japanese ,Japan	71
KR.KR	Koreans, Korea	90
MY.JH	Negrito, Malaysia	50
PI.AT	Ati, Philippines	23
SG.ID	Indian, Singapore	30
TH.MA	Mlabri, Thailand	18
TW.HA	Chinese, Taiwan	48

Table 3.4: **Details of the Pan-Asian dataset.**

Assessing the clustering quality

To assess the potential of a clustering method it is important to focus on both the sample assignments and the estimated number of clusters. The quality indicator usually considered is the accuracy, that is the proportion of individuals that were assigned to the correct populations. This indicator focuses only on the one-to-one relationship between estimated clusters and true populations. We decided not to retain this criterion as it does not exhaustively describe the quality of a clustering method's assignments and does not account correctly for the estimated number of clusters.

The indicator we selected to account for both the assignments and the estimation of the number of clusters is the adjusted Rand index (Rand 1971). This index is calculated using the contingency table of two clusterings U and V (**Table 3.5**) with the formula

$$\text{adjusted Rand index} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}},$$

where a_i, b_i and n_{ij} are defined in **Table 3.5**.

This index focuses on all pairs of samples and considers whether they have correctly been assigned to the same population or correctly been assigned to different populations. That way, in addition to the accuracy criterion, the adjusted Rand index takes into account the fact that certain samples should not be clustered together. The adjusted Rand index is comprised between 0 and 1, a value of 1 meaning a perfect clustering. Note that if there is only one cluster in the data and that a clustering method properly uncovers

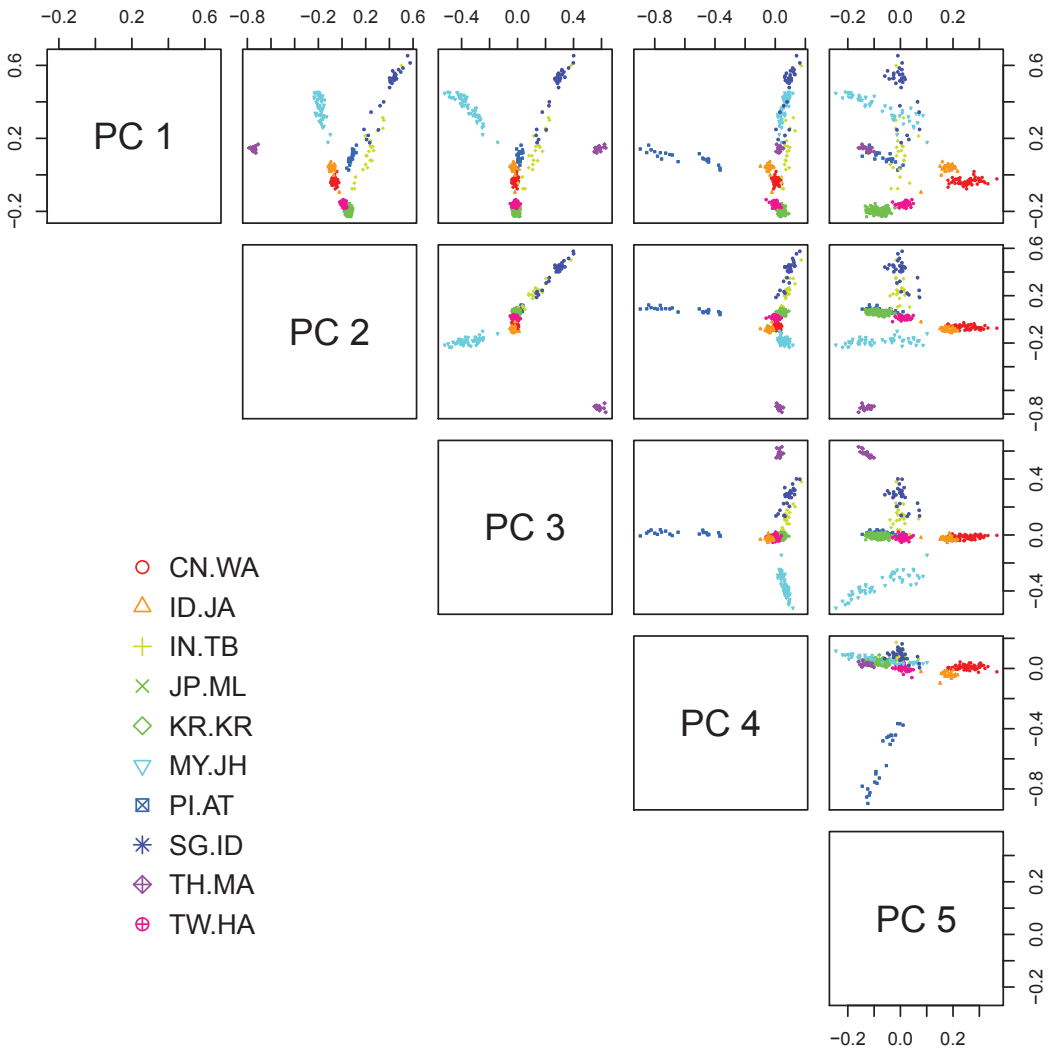

Figure 3.5: **Representation of the 10 populations of the Pan-Asian dataset.** This scatter-plot uses the first five principal components of a dataset with 20K SNPs. This graph is only intended to present the general genetic pattern of the dataset and does not exhaustively represent the capability of the PCA to separate the populations.

	V_1	V_2	...	V_C	$Sums$
U_1	n_{11}	n_{12}	...	n_{1C}	a_1
U_2	n_{21}	n_{22}	...	n_{2C}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
U_R	n_{R1}	n_{R2}	...	n_{RC}	a_R
$Sums$	b_1	b_2	...	b_C	N

Table 3.5: **Contingency table between two clustering U and V.** a_i and b_i are the numbers of samples in the $i-th$ clusters U_i of U and V_i of V and n_{ij} the number of samples in the $i-th$ cluster U_i of U and the $j-th$ cluster V_j of V.

such a structure the Rand index is theoretically not defined. Given that the structure is perfectly estimated in such a case we consider a value of 1 for the Rand index.

For simulated datasets we compared, via the adjusted Rand index, the clusterings proposed by the different methods to the true population labels that are available through the simulation process. For the admixed and the real datasets, no true population labels exist. As a consequence we provide two quality measures that are the quality index using as comparison partitions the population labels provided with the datasets (e.g. CHB or CHD in HapMap) and the partitions produced by `Admixture`. We selected `Admixture` as it is one of the most widely used methods for the estimation of population structure. Also we represent the admixture proportions of all the methods with barplots. For discrete clusterings these proportions are either 0 or 1.

3.4.2 Results

Several small datasets and one large dataset were investigated for each simulated or real scenario. The average Rand indexes and the average estimated numbers of clusters are the indicators we are interested in. **Figure 3.6** presents these values for all the methods applied to small datasets and **Figure 3.7** for the large datasets. In addition, **Figure 3.8** provides examples of the graphical representations of the criterion used by `SHIPS` to estimate the number of clusters K on the small datasets and **Table 3.6** the average numbers of principal components retained by the algorithm `PCAclust` in each scenario.

Data / Model	M1	M3	M5	M10	M20	Madx	HapMap	Pan-Asian
Small data	0	17	12	51	72	9	49	64
Large data	0	6	5	28	49	25	70	99

Table 3.6: **Number of PCs selected with PCAclust.**

Simulated datasets

Model M1 (1 sub-population)

For the model M1, with only one population, `SHIPS` was always able to correctly determine the correct number of one cluster for both all the small and large datasets. This was also the case of `Structure` and `PCAclust`. As a consequence these three methods perfectly assigned all the individuals to the correct population and had a Rand index of 1. On the other hand, `Admixture` was only able to determine that there was no structure in the small datasets, estimating $K = 1$, but not in a large dataset producing $K = 2$. This is bound to be due to the number of SNPs that led the algorithm to determine a more complicated structure. `AWclust` properly determined that there was one cluster in 7 small replicates out of 10, but the average number of estimated clusters is $K = 2$. On the large dataset, this latter method correctly estimated the number of clusters as the amount of SNPs allowed the `AWclust`'s gap statistic to be more accurate.

Model M3 (3 sub-populations) and M5 (5 sub-populations)

The performances of `SHIPS`, `Structure` and `AWclust` were comparable for these models. An average number of 3 and 5 clusters was respectively estimated for all small and large replicates of the models M3 and M5 (except for `Structure` that was not applied to large datasets). These three methods mis-classified in average less than 3 individuals leading to Rand indexes higher than 0.99. `PCAclust` was able to estimate the correct number of 3 sub-populations in 8 small replicates out of 10 small datasets of the model M3 and in 5 replicates for the model M5. When the number of SNPs increased to 200K, `PCAclust` was able to correctly estimate K and led to perfect sample assignments. The clustering proposed by `Admixture` on these models were not consistent with the true populations. Indeed, this method identified the maximum number of clusters to be the optimal one, that is 10 in our case. Larger sample sizes did not improve these results.

Model M10 (10 sub-populations)

The model M10, with 10 populations, pertains to a more complex structure of the data. In this scenario `SHIPS`, `Structure` and `AWclust` succeeded in perfectly estimating K and assigning all individuals to the correct populations for both small and large datasets. `PCAclust` estimated a mean number of 6 clusters for the small data, 4 for the large data as it was not able to separate certain populations. `Admixture` again over-estimated the number of clusters ($K = 18$ for small data and $K = 17$ for large data). We investigated up to 20 clusters but the algorithm did not converge for values of K greater than those estimated.

Model M20 (20 sub-populations)

In this last simulated model, with the more complex structure and 20 populations, both `SHIPS` and `Structure` evaluated the correct number of clusters for all replicates and

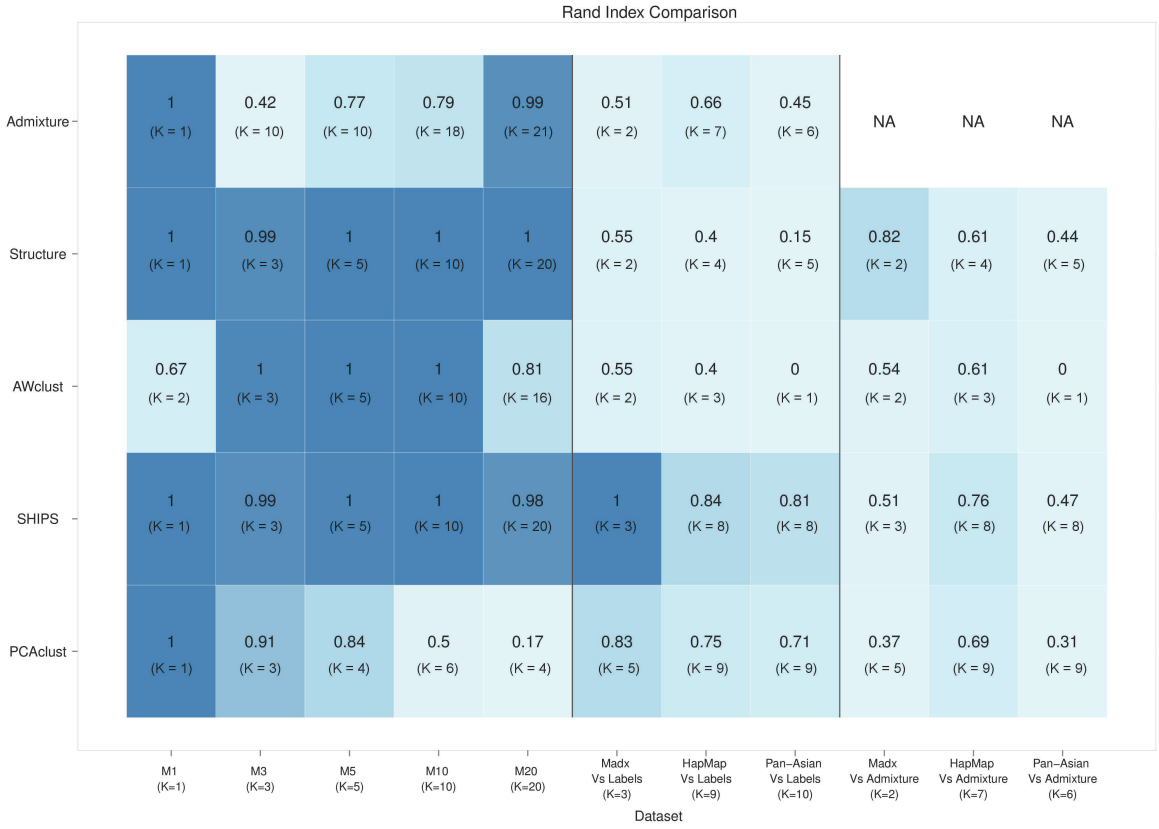

Figure 3.6: **Comparison of the clustering methods on the small datasets.** Average Rand indexes over all small replicates are indicated for each method and each model along with the estimated numbers of clusters in parenthesis. The darker a cell color is, the better the corresponding clustering is.

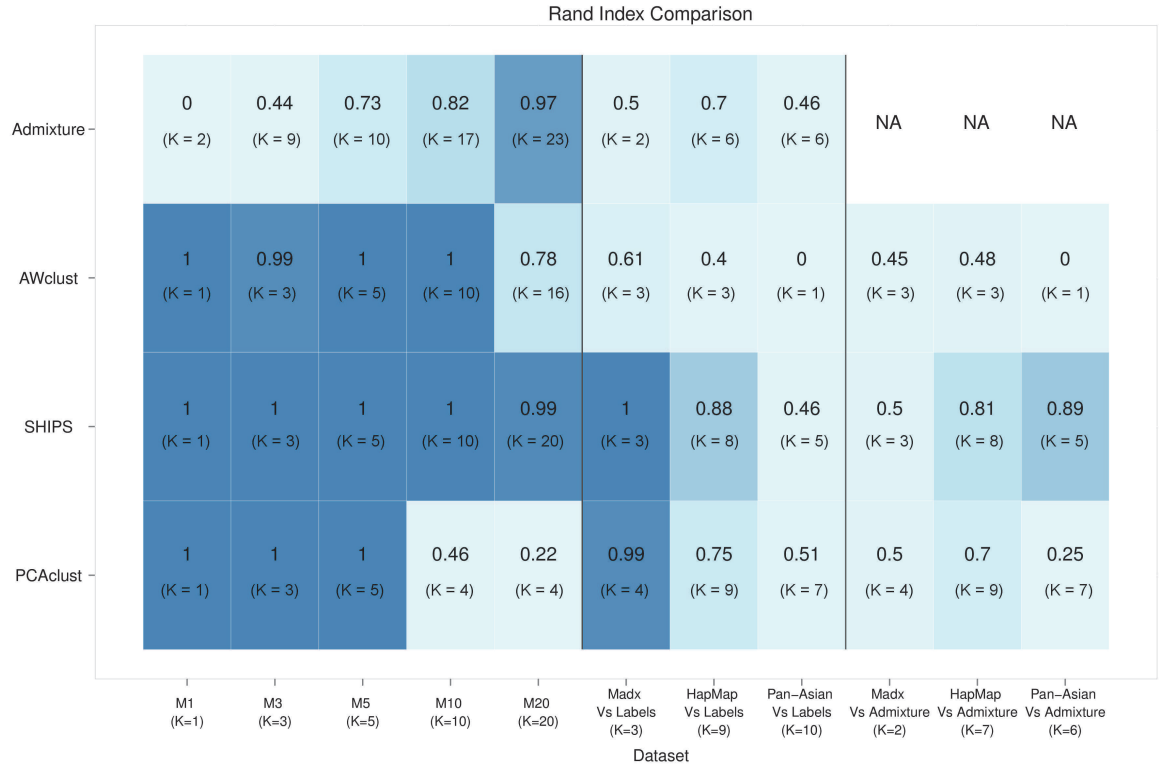

Figure 3.7: **Comparison of the clustering methods on the large datasets.** Rand indexes are indicated for each method and each model along with the estimated numbers of clusters in parenthesis. The darker a cell color is, the better the corresponding clustering is. The software Structure was not applied to large datasets due to a too large computational cost.

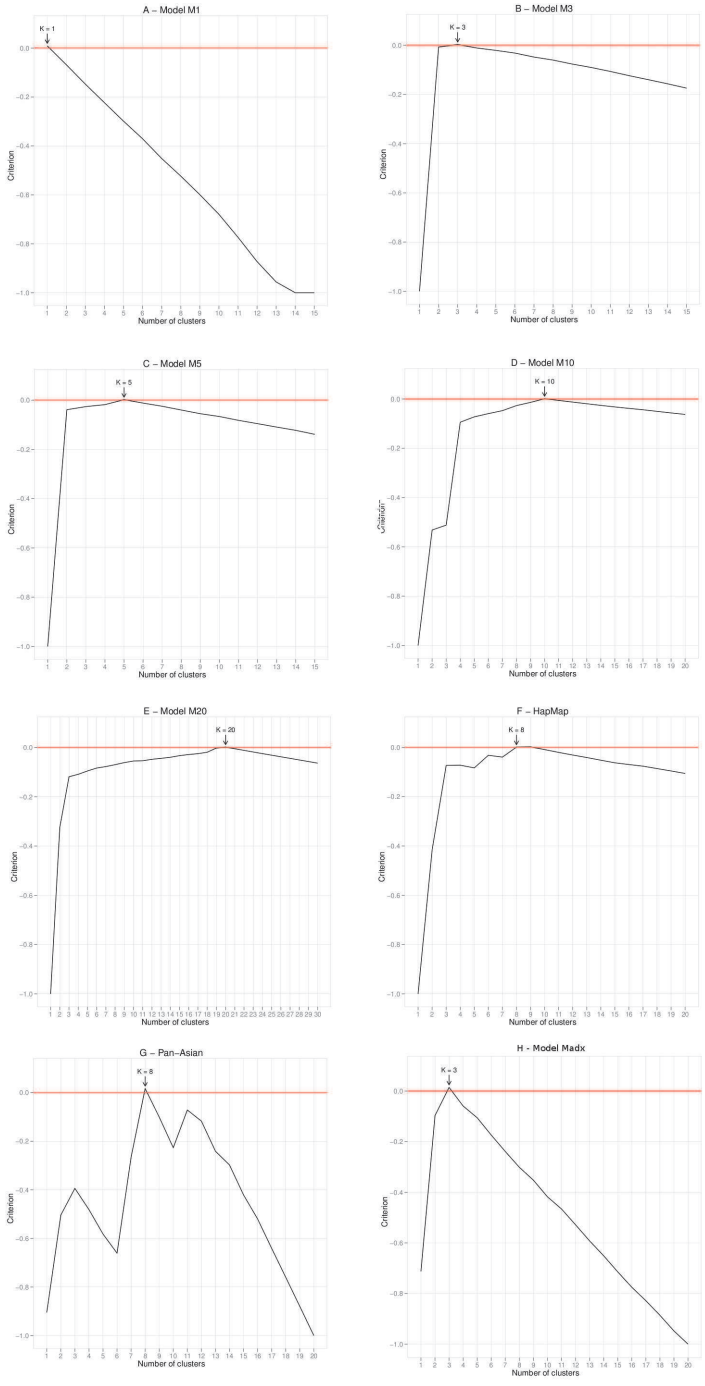

Figure 3.8: **Estimation of the number of clusters with SHIPS.** Representation of the criterion used to determine the estimated number of clusters $crit(k) = Gap(k) - Gap(\tilde{k}) + s_{\tilde{k}}$. The estimated K is the smallest k such as $crit(k) \geq 0$. The black curves represent the criterion and the red lines the 0 threshold. The first replicate of each small dataset was used to produce these plots.

completed an individual assignment very consistent with the true populations. `AWclust` and `PCAclust` underestimated the number of clusters. `AWclust` only allows to estimate a maximum of 16 clusters that was reached for this complex dataset. One could wonder if the clustering assignments would have been better if the maximum number of clusters was more flexible. On the other hand, `PCAclust` was not able to detect the structure of this dataset. Only 4 clusters in average were identified in the small and large datasets as many populations were not separated thus leading to a low Rand index close to 0.2. For both small and large datasets `Admixture` estimated 21 clusters and almost perfectly assigned all the individuals to the correct populations. Even though these clusterings are quite accurate, it is noticeable that 21 was the maximum number of clusters for which the algorithm converged. In other words, it is possible that if the convergence could have been reached for greater values of K, the number of clusters could have been over-estimated again.

`SHIPS` and `Structure` were the most accurate methods when applied to simulated datasets both in terms of estimating the correct number of clusters K and assigning individuals consistently with the true population labels. The performances of the other methods were a little less, especially for `Admixture` that always over-estimated K and `PCAclust` that usually under-estimated it. It is also noticeable that for all of the methods the results are generally comparable between the large and the small datasets.

Admixed and real datasets

In order to assess the quality of the clustering methods we were also interested in looking at admixed and real datasets, more representative of the ones encountered in genetic studies. We present the average results over the different small and large replicates, along with details on the assignments performed. In order to account for the fact that there is no 'true' structure in real datasets, we considered both the population labels and the labels produced by the program `Admixture` as structures (also called partitions) of reference. Barplots of the admixture proportions of the different methods are presented for the admixed datasets (**Figures 3.9 and 3.10**), the HapMap datasets (**Figures 3.11 and 3.12**) and the Pan-Asian datasets (**Figures 3.14 and 3.15**). In addition, **Figures 3.13 and 3.16** display the graphical results of `SHIPS` for the small HapMap and Pan-Asian data.

An admixed population

`SHIPS` identified 3 distinct populations for the admixed datasets that are the two populations of origin (CEU and CHB) and the one simulated as an admixture. `Structure`, `Admixture` and `AWclust` detected two populations. The admixture proportions displayed in **Figure 3.9** show that `Admixture` and `Structure` estimated almost the same ancestries for the individuals, with the admixed population (XY) having a genome coming approx-

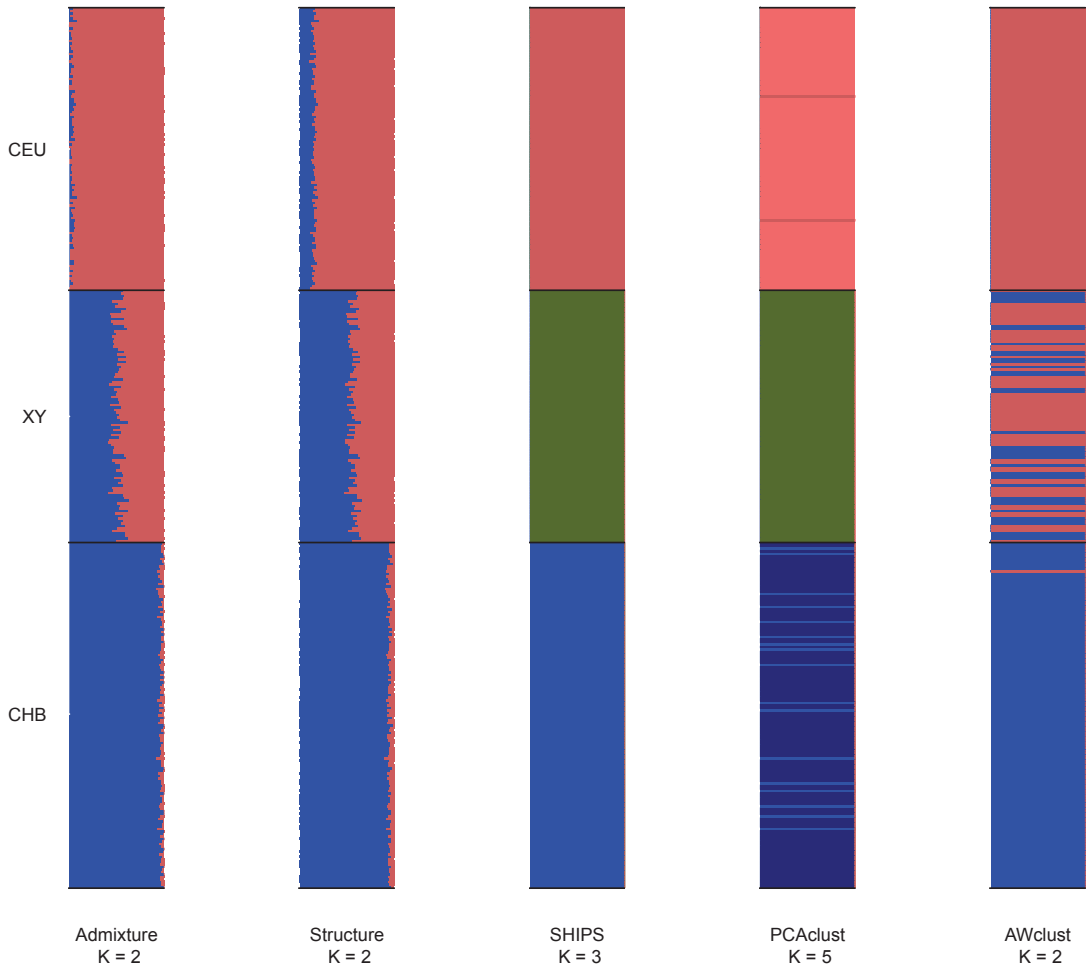

Figure 3.9: **Barplots of the admixture proportions for the small admixed data.** The first small dataset was used to produce this plot. Populations are separated by black lines and assigned with a unique color that is approximately reported on the barplot of each method. For the discrete methods the admixture proportions are either 0 or 1.

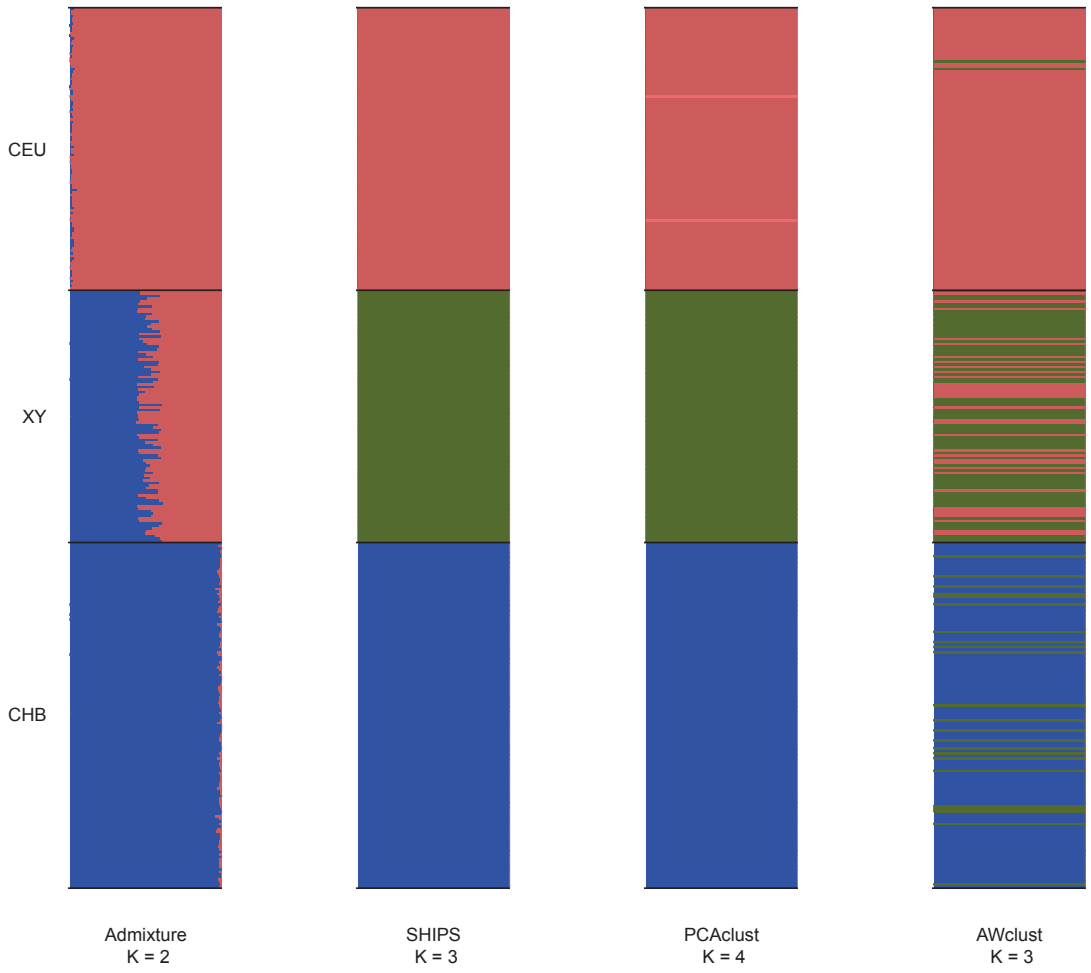

Figure 3.10: **Barplots of the admixture proportions for the large admixed data.** Populations are separated by black lines and assigned with a unique color that is approximately reported on the barplot of each method. For the discrete methods the admixture proportions are either 0 or 1.

imately in equal part from the CHB and CEU populations. These proportions correctly match those used in our simulation model. `AWclust` resulted in a split of the admixed population in function of these admixture proportions. On the other hand, `PCAclust` estimated 5 clusters that correspond to the 3 distinct populations identified by `SHIPS` and two small clusters being sub-populations of the CHB and CEU populations.
In terms of quality indexes, when comparing to the population labels, `SHIPS` and `PCAclust` performed the best as they identified the 3 main discrete populations. When comparing the results to `Admixture`, `Structure` is the closest in such a setting and `SHIPS` and `AWclust` are in agreement at about 50% as they assigned the samples from the admixed population to another population being a cluster of admixed, CEU or CHB individuals.
The results are quite similar on the large admixed dataset except for `PCAclust` that did not find small sub-clusters within the CHB populations (**Figure 3.10**).
It is interesting to notice that there are two kinds of behaviors to cluster the admixed individuals. Certain methods assigned them to the populations of origin they are the closest genetically speaking and others created a specific admixed cluster. These two behaviors of the methods are understandable given the nature of the admixture that we considered in this simulation. Indeed, we simulated a discrete admixture, meaning that the admixed samples, even though originating from the CHB and CEU populations, form a discrete cluster. The nature of this structure is therefore more challenging for discrete clustering algorithms such as `SHIPS` and `AWclust` but also quite favorable to discrete assignments compared to 'real life' admixtures that are usually continuous. The results produced by `Structure` and `Admixture` have to be interpreted in the sense that with a continuous admixture only the admixture proportions can properly relate the structure as there would be no discrete cluster to be identified. Further analyses of these algorithms on continuous admixture would reveal more precisely the behaviors of the algorithms with such a population structure and complete the partial results presented here.

HapMap 9 populations

Considering all 20 small replicates, `SHIPS` was able to identify 8 clusters in average (**Figure 3.8**). Certain populations such as the two Chinese populations (CHD and CHB) were not entirely differentiated in some datasets. Also, two of the African populations YRI and LWK were sometimes assigned to the same cluster. Results were similar on the large dataset. In both cases, an average Rand index of about 0.8 was reached when using the population labels as reference (**Figures 3.6 and 3.7**). `PCAclust` estimated 9 clusters by assigning CHB and CHD to the same cluster and splitting certain populations such as GIH or the African ones into several clusters. `Structure` and `AWclust` produced clusterings less consistent with the population labels. `Structure` identified the three main ethnicities, that are African, Caucasian and Asian plus the GIH population. Note that this population derives from the Asian and Caucasian one. `AWclust` was only able to detect the three main ethnicities. These two latter methods have therefore a relatively low Rand index (0.4) compared to the population labels.

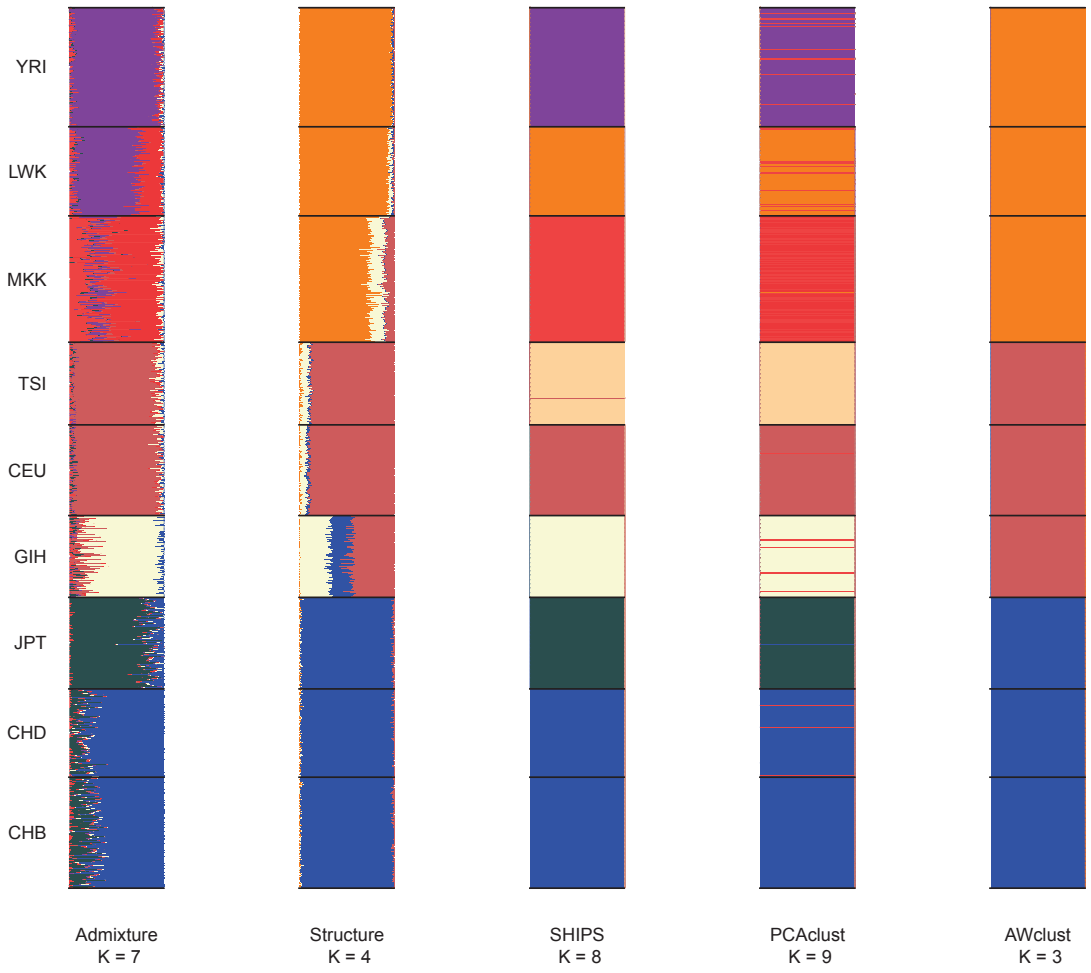

Figure 3.11: **Barplots of the admixture proportions for the small HapMap data.** The first small dataset was used to produce this plot. Populations are separated by black lines and assigned with a unique color that is approximately reported on the barplot of each method. For the discrete methods the admixture proportions are either 0 or 1.

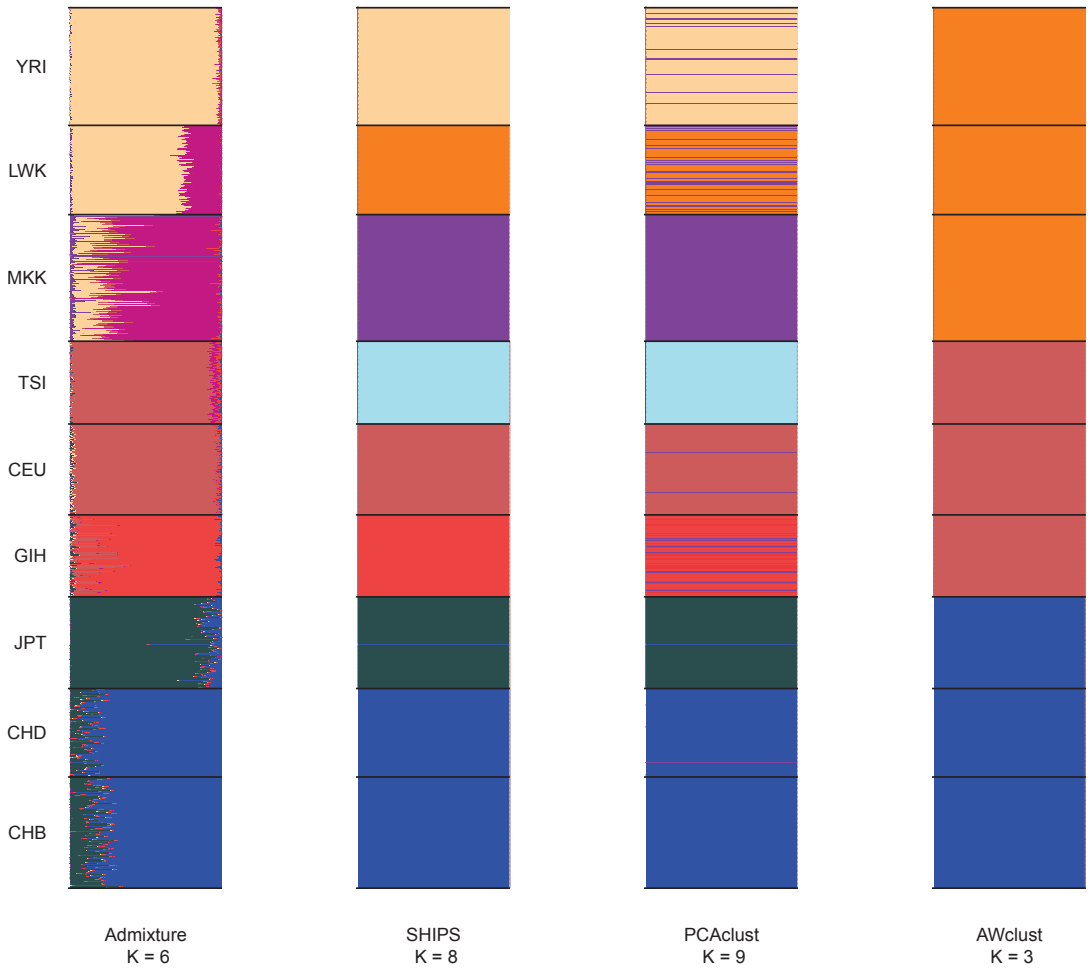

Figure 3.12: **Barplots of the admixture proportions for the large HapMap data.** Populations are separated by black lines and assigned with a unique color that is approximately reported on the barplot of each method. For the discrete methods the admixture proportions are either 0 or 1.

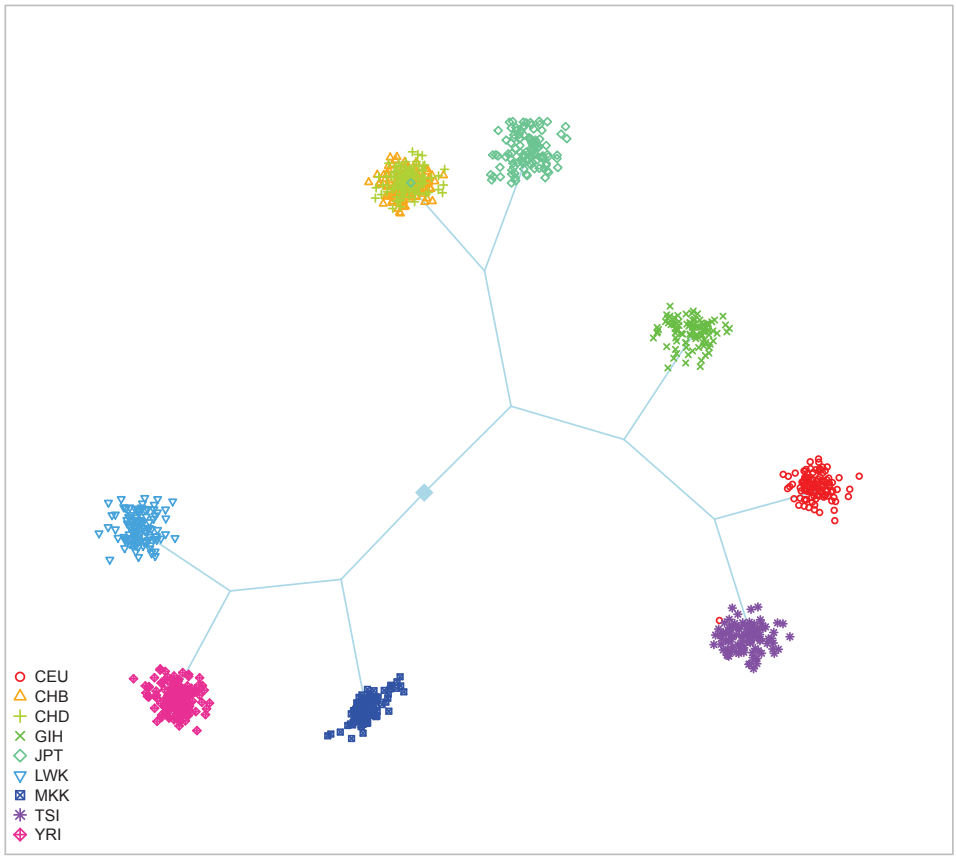

Figure 3.13: **SHIPS tree of the 9 HapMap populations.** This representation is an output produced by SHIPS. The tree structure corresponds to the successive divisions conducted by the algorithm. Each final cluster is represented by a scatter-plot of its members. We colored here the individuals according to the population labels.

Admixture estimated 7 ancestral populations in the small datasets. As we can observe on **Figure 3.11**, according to Admixture, the CHB and CHD populations share a very close ancestry, which can explain why SHIPS and the other methods did not split these populations. The JPT population has a common ancestry with the Chinese populations but with different admixture proportions. SHIPS and PCAclust were able to differentiate this population from CHB and CHD but not Structure and AWclust. Among the 7 ancestral populations detected by Admixture, one is specific to the GIH population. In addition, Structure uncovered the same admixture pattern which validates the clusterings of SHIPS and PCAclust that differentiated the GIH population. It is noticeable that even though the admixture proportions of the Caucasian population CEU and TSI are very close, SHIPS and PCAclust were able to separate them in two distinct clusters. The behavior of the methods is however different on the African populations. The 3 corresponding populations share the same 3 ancestries in different proportions. SHIPS differentiated these 3 populations correctly whereas PCAclust created a fourth cluster composed of samples from each of these populations. When observing the admixture proportions of the samples clustered into this additional group, there seems to be no common pattern and therefore this split appears to be inconsistent with the structure of the population. As a result SHIPS is the method that agrees the most with Admixture (Rand index = 0.76) followed by PCAclust (Rand index = 0.69), Structure (Rand index = 0.61) and AWclust (Rand index = 0.61).
On the large dataset, results are quite similar except that Admixture estimated 6 ancestral populations. The corresponding assignments were however more consistent with the population labels. The same observation can be made for SHIPS and as a consequence the quality indicator of our new method improved whether we compared it to the population labels or to Admixture.

Pan-Asian 10 populations

We first describe the results for the small datasets. In average, over all the small Pan-Asian datasets, SHIPS estimated 8 clusters. In the majority of the replicates the population from India (IN.TB) was clustered with the Philippines (PI.AT) or Singapore (SG.ID) and the populations from China (CN.WA) and Indonesia (ID.JA) or Japan (JP.ML) were assigned to the same cluster. These clusterings of the data are quite consistent with the labels of the populations and as a consequence SHIPS has the highest Rand index of 0.81 with this reference partition. PCAclust estimated 9 clusters. The CN.WA population was split in several clusters and often assigned to the same clusters as samples from SG.ID and IN.TB or PI.AT and MY.JH. Several other populations were separated according to the population labels and therefore the quality index with this reference is of 0.71. Structure identified 5 ancestral populations. The corresponding discrete clustering is however quite distant from the population labels. Indeed, only the MY.JH, TH.MA and part of the SG.ID populations are separated. As a consequence the Rand index compared to the population labels is quite low. Likewise, AWclust has a null Rand index as this

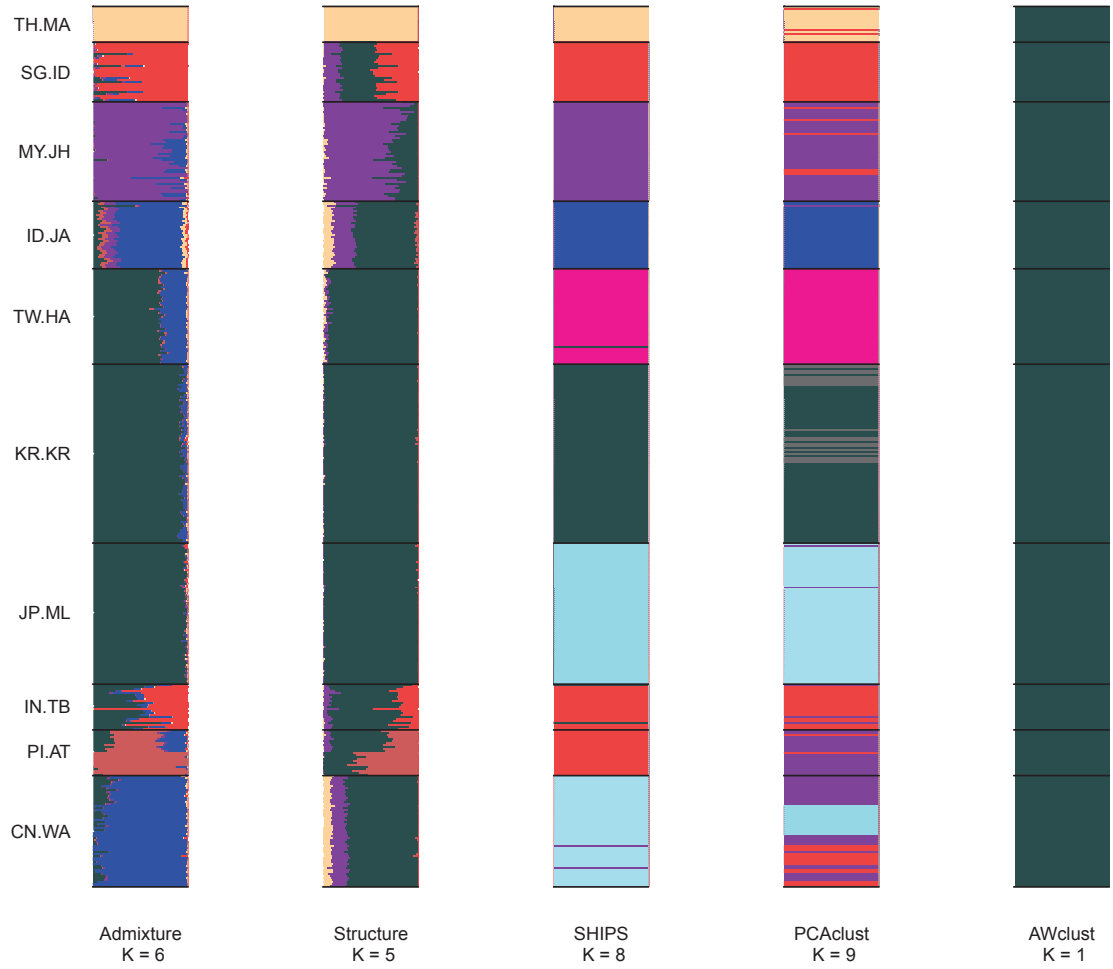

Figure 3.14: **Barplots of the admixture proportions for the small Pan-Asian data.** The first small dataset was used to produce this plot. Populations are separated by black lines and assigned with a unique color that is approximately reported on the barplot of each method. For the discrete methods the admixture proportions are either 0 or 1.

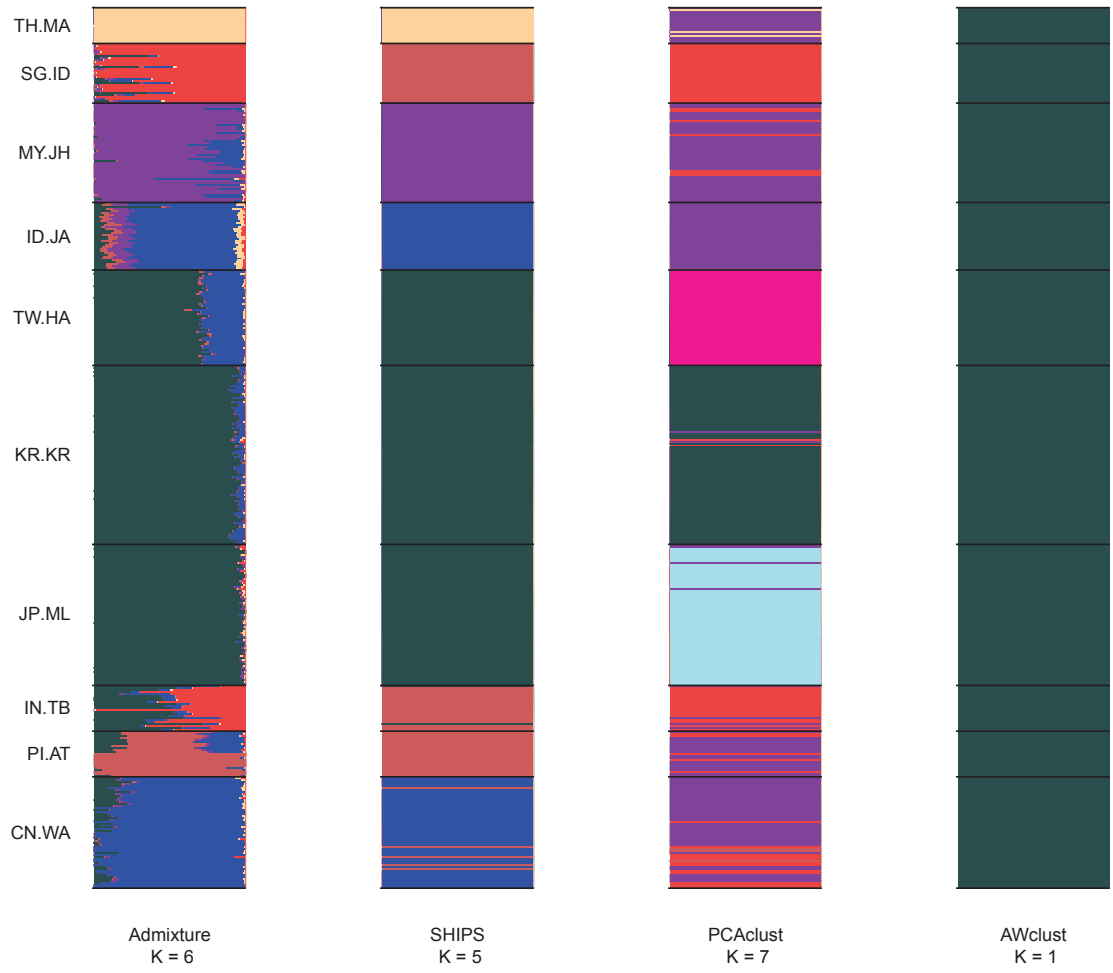

Figure 3.15: **Barplots of the admixture proportions for the large Pan-Asian data.** Populations are separated by black lines and assigned with a unique color that is approximately reported on the barplot of each method. For the discrete methods the admixture proportions are either 0 or 1.

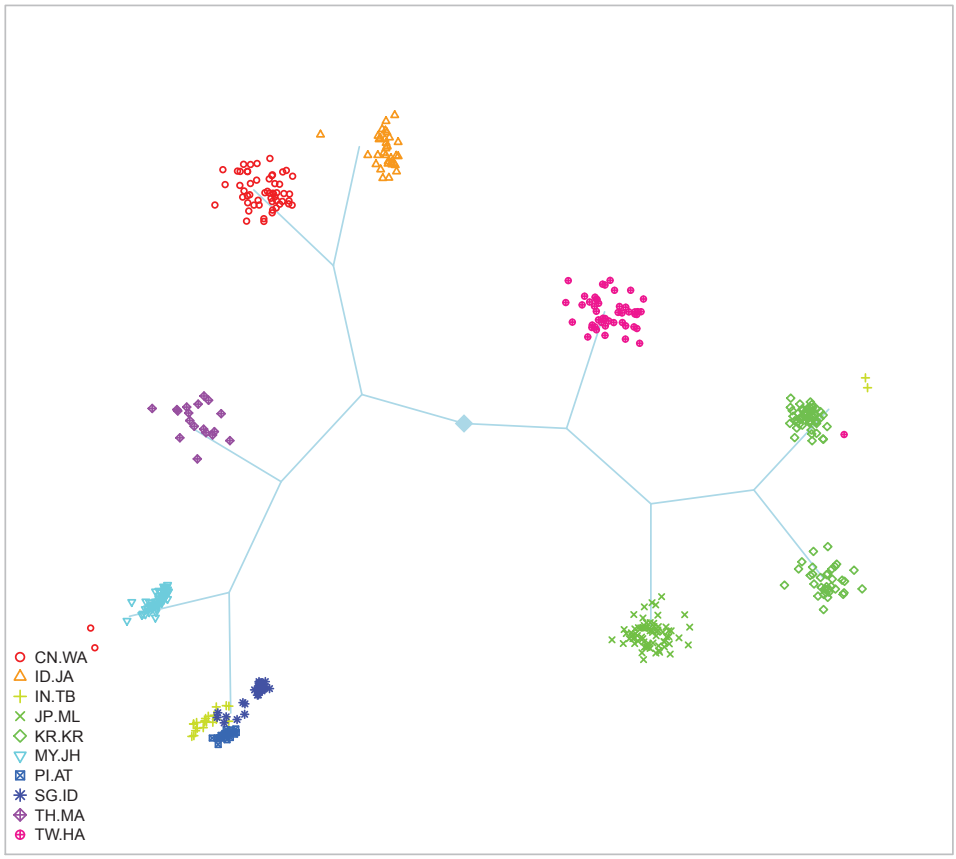

Figure 3.16: **SHIPS tree of the 10 Pan-Asian populations.** This representation is an output produced by SHIPS. The tree structure corresponds to the successive divisions conducted by the algorithm. Each final cluster is represented by a scatter-plot of its members. We colored here the individuals according to the population labels.

method did not determine any structure in the data. `Admixture` found 6 ancestral populations. The populations IN.TB, JP.ML, KR.KR and TW.HA were assigned to the same cluster like CN.WA and ID.JA. This results in a Rand index of 0.45. When analyzing the admixture proportions (**Figure 3.14**) we observe that `SHIPS` assigned the populations IN.TB and CI.AT to the same cluster whereas these populations share quite different ancestries. On the other hand, this novel algorithm differentiated the TW.HA, KR.KR and JP.ML populations that have closely related ancestries. `PCAclust` also assigned these populations to different clusters but had a lower Rand index than `SHIPS` compared to the `Admixture` partitions as the additional cluster detected by this method does not match the admixture proportions.

On the large datasets, `SHIPS` and `PCAclust` estimated fewer clusters than on the small datasets. `SHIPS` estimated 5 clusters and `PCAclust` 7 clusters. These differences resulted in `SHIPS` identifying a structure very close to that estimated by `Admixture` (Rand index of 0.89) while `PCAclust`'s clustering was less in agreement with `Admixture` (Rand index of 0.25). On the other hand, `PCAclust` was closer to the population labels partition than `SHIPS`. One has to note that when setting the number of clusters manually, `SHIPS` and `PCAclust` estimated the same structure than on the small datasets. These different behaviors of the methods are therefore due to the size of the dataset that influenced the estimations of the number of clusters.

The analysis of the real datasets pointed out that compared to the population labels as reference partitions, `SHIPS` was the most efficient method to uncover the population structures followed by `PCAclust`. Even though `SHIPS` produces discrete clusterings, this novel algorithm reached the most important agreement with the clusterings estimated by widely used methods such as `Admixture`.

3.5 Discussion

We have proposed a novel clustering approach to infer the genetic structure of populations from SNPs data. `SHIPS` is based on a divisive hierarchical clustering procedure and a pruning strategy followed by the use of the gap statistic to estimate the final number of clusters K.

Comments on the different strategies

`SHIPS` has proven to be an accurate and precise method to estimate both relevant optimal numbers of clusters as well as for producing assignments consistent with the reference partitions of the data considered. In the simulated datasets, K was always correctly estimated and only few individuals were mis-assigned. The structures identified for the admixed dataset (K=3), the HapMap (K=9) and the Pan-Asian (K=10) datasets were remarkably close to the population labels or the partitions estimated by the program `Admixture`.

The other algorithms considered had less regular performances, either missing the structure of the complex simulated data or of the real datasets. A possible explanation of these results depends on the algorithms' methods to estimate the number of clusters or on the parameters utilized for each algorithm. It is interesting to observe that even though `Structure` and `Admixture` are based on the same model their performances are notably different. In the simulated datasets, `Structure` was able to estimate the correct K for each dataset. On the other hand, `Admixture` always over-estimated the number of clusters by selecting the higher K investigated. This led to poor performances of `Admixture` in the first simulated scenarios (M1 and M3) and relatively satisfying ones in the final scenarios (M5, M10 and M20) as the correct number of clusters corresponded to the maximum K for which the method converged and therefore the estimated K. Given that when manually setting K to the true values, `Admixture` identified the true structures of the data, the estimation of the number of clusters through cross-validation can be identified as the cause of the poor clustering quality of the algorithm in the simulated datasets. We considered different cross-validation methods that are 5, 10 and 15 fold cross-validation, and obtained the same estimations of K (data not shown). It therefore appears that the cross-validation method is not fit in such settings to estimate the number of clusters. These results confirm certain limitations of the cross-validation criterion that had already been pointed out (Alexander and Lange 2011, Lawson et al. 2012). We used in our comparison an improved version of `Structure` by considering an estimated K maximizing the quality criterion thus leading to more correct estimation of K. However, one has to note that the estimation method originally used in `Structure`, that is the maximum likelihood, led to correctly identifying the structure of the simulated data (data not shown). The opposite conclusions can be drawn for real datasets (HapMap and Pan-Asian). `Admixture`

estimated values of K close to the ones defined by the population labels while `Structure` under-estimated the values of K compared to both the population labels and `Admixture`. The cross-validation method used in `Admixture` is more appropriate for real complex datasets however there are no efficient way to estimate a correct K for `Structure`. This is due to the fact that even when setting manually K, `Structure` produced clusterings with empty clusters and therefore could not identified more populations than we presented in the Results section. For example, only the three main ethnicities plus the GIH population were identified in the HapMap data while other methods such as `SHIPS` or `Admixture` were able to differentiate the Asian, Caucasian or African populations. A possible explanation for `Structure`'s results is that, even though the algorithm converged properly, a too short burn-in period and too few runs of the algorithm were used for such complex data. These choices were however made due to the very high computational time of the program.

`AWclust` generally uncovered the structure of the small and large simulated datasets but failed to properly analyze the real datasets. Whether we considered the population labels or the partitions produced by `Admixture` as reference for the real datasets, `AWclust`'s clusterings were not in agreement with these references. Only the three main ethnicities were detected in the HapMap data and no structure in the Pan-Asian data due to the fact that the optimal estimated number of clusters were under-estimated. It is however interesting to notice that when manually setting the number of clusters, the sample assignments were more consistent with both the population labels or the results of `Admixture`. This can be explained by the gap statistic used by the algorithm that was not able to select the correct values of K while the hierarchical clustering could separate certain populations. 20 simulations for the gap statistics may not have been enough though the same number was used with `SHIPS` that more correctly estimated K. These results highlight the quality of the version of the gap statistic that we used in the `SHIPS` algorithm.

In addition to the individuals clustering, both `SHIPS` and `AWclust` provide tree structures that allow the analysis of the relationship between populations. The corresponding graphical representations, presented in **Figures 3.13 and 3.16** for `SHIPS`, are quite similar to dendrograms produced by `AWclust`. The differences are that in `SHIPS` the lengths of the branches have no meaning and the individuals of the final clusters are plotted to represent their dispersion. The analysis of these two kinds of graphical representations were quite similar in our comparisons. For example, we observed in the simulated datasets, that for basic population structures (model M3 and M5), the trees provided by `SHIPS` and `AWclust` properly related the genetic histories of the populations. For more complex datasets, mainly the major population differentiations and some of the finer separations led to tree branches consistent with the population histories represented in **Figure 3.3**. Also, these representations can provide indications on the genetic distance of the real populations. For instance, we observed on **Figure 3.13** that the Caucasian and Asian populations are first separated from the African ones and then separated from each other.

The method `PCAclust` selected the number of principal components to be used for

the clustering using the Tracy-Widom statistic (**Table 3.6**). Many components (more than 25) were determined significant for the complex simulated datasets M10 and M20. This led to clusterings rather inaccurate as the estimated numbers of clusters were greatly under-estimated for both the small and large datasets. If fewer PCs were kept, e.g. only five, the estimated K would have been more exact (data not shown). This indicates that too many PCs add a non-negligible noise to the data provided to the GMM clustering and therefore that the PCs selection method of `PCAclust` could be improved.
The performances of this method are however better when applied to real datasets, especially when compared to the population labels. When comparing the clusterings produced by `PCAclust` to `Admixture`, the results are more mitigated. PCAclust estimated more clusters than `Admixture` and split populations that this latter algorithm considered coming from the same ancestral populations. A reason might be that even though the two algorithms are somehow linked (Lawson and Falush 2012), the methods to estimate the numbers of clusters are quite different.

The methods discussed here are composed of two parts to analyze the structure of the populations. The first corresponds to the quality to assign individuals to relevant clusters and the other is the ability to estimate a proper optimal number of clusters K. If a potential value of K is unknown, it is important that the clustering method estimates a proper K otherwise even with accurate sample assignment capabilities the resulting clustering may not be relevant. Among all the algorithms that we investigated, `SHIPS` was the only one that had satisfying performances for both these features of clustering methods in all the scenarios investigated. `SHIPS` did not fail to uncover the structure in simulated datasets like `Admixture` and `PCAclust` and did not miss the fine complex separation of the populations in real datasets like `Structure` or `AWclust`.

In terms of ease of use of the algorithms, the non-parametric ones generally have the advantage of demanding fewer input parameters than parametric approaches. In addition to the data, `SHIPS` needs the maximal number of clusters investigated and the number of null simulations for the gap statistics. Usually parametric algorithms need a lot of input parameters, often pertaining to the underlying statistical models and therefore more complicated to set. This is the case of `Structure`, however `Admixture` needs only the maximal number of clusters and the parameter to conduct the cross-validation.

Considering the computation time of the algorithms, `PCAclust` is the faster, e.g. taking less than an hour when applied to the Pan-Asian data. `SHIPS` and `Admixture` take a couple of hours while `AWclust` takes close to a day and `Structure` several days. Even though `PCAclust` is the fastest algorithm that we considered in our comparison, one has to note that the program does not come as a package and has to be recoded. The other methods that we considered have the advantage of being freely available in the form of packages.

Particularities of SHIPS

Several particularities of the SHIPS algorithm can be highlighted. The divisive strategy is based on the rationale that a clustering method has to be applied iteratively to the sub-populations in order to detect the cryptic structures that are hidden behind the main structure of the data. SHIPS finely investigates each estimated cluster to determine if it can be divided into several relevant sub-clusters. This division procedure, that is equivalent to the construction of a binary tree, is conducted by the use of a spectral clustering that takes as input a similarity matrix. This similarity matrix has to be computed only once for all the data and sub-matrices corresponding to the sub-clusters investigated can be extracted at each step. This renders the construction of the tree a fast and efficient part of the algorithm. One has to note that the individuals assignment part of the SHIPS algorithm is intimately linked to the choice of a proper similarity matrix. We decided to consider a matrix based on the allele sharing distance as it is computationally fast to compute and led to accurate clustering results. It is however possible to use different matrices that could lead to even better clustering performances (Lawson and Falush 2012). It has been demonstrated that matrices based solely on the allele sharing distance can have low power for the identification of population structure compared to more elaborate distances taking into account other features of the data such as the dependencies between the markers or the relatedness between the samples. Example of such distances can be found in (Browning and Browning 2010, Lawson et al. 2012) and could easily be used with SHIPS. Indeed, a flexibility of the SHIPS algorithm is that a large variety of similarity matrices can be used to conduct the samples assignment.

The pruning procedure leads to several possible clusterings of the samples. These configurations are all nested within each other. This allows in one run of the algorithm to get for all possible K the corresponding clusterings. This information is useful if the user does not desire to use the estimation procedure of K and wants to manually look at the clustering possibilities. The hierarchical clustering of AWclust proposes the same option, while software such as Admixture, Structure or PCAclust have to be applied each time for each possible number of clusters. In addition, this allows a fast application of the gap statistic that needs all clustering options for varying numbers of clusters.

SHIPS does not use the same version of the gap statistic than the one used in AWclust. As explained in **Section 3.3.5**, we decided not to consider the logarithm of the within-cluster sum of squares but directly the sum of squares. This indicator showed better empirical performances to estimate the optimal K. Given that AWclust was sometimes able to infer the structure of certain data when manually setting a value for K but that the version of the gap statistic used in the program failed to do so, we are confident in our choice of statistic. This gap statistic is rather precise but, like all gap statistics, a time consuming method to estimate the number of clusters. Certain methods, such as AWclust, therefore limit the maximum number of clusters investigated in order to accelerate the whole clustering process. We decided not to make this limitation in the SHIPS package

in order to let the user of the program the choice of a reasonable maximum number of clusters.

Also, we determined through several experiments that repetitive applications of the `SHIPS` algorithm to the same dataset leads to the same clustering results. This robustness of the algorithm confirms that `SHIPS` is a powerful tool to detect population structure.

The novel clustering approach presented in this chapter was applied to SNP data. It produces accurate clustering results and is therefore a promising method to uncover the genetic structure of many populations.

Chapter 4

Multiple-testing Issues in Genome-Wide association studies

We refer to multiple-testing issues when not one but several statistical tests are performed simultaneously. When only one statistical test is conducted, the probability of having a false-positive is controlled at the level of significance α. When more than one test are performed, α can no longer be interpreted as the probability of false-positive of the overall tests but rather as the expected proportion of tests providing false-positive results.

Let us consider, for instance, m tests performed in a single experiment. With just one test ($m = 1$) performed at the usual 5% significance level, there is a 5% chance of incorrectly rejecting H_0. However, with $m = 20$ tests in which all the null hypotheses are true, the expected number of such false rejections is $20 \times 0.05 = 1$. Now, with $m = 100,000$ tests, the expected number of false-positives is 5,000 which is much more substantial.

As a consequence, the control of the proportion of false-positives in the context of multiple-testing is a crucial issue and the false-positive rate does not appear adapted anymore to define a confidence threshold to apply to the p-values. We therefore need multiple-testing correction procedures aiming at controlling certain more relevant statistical confidence measures.

Genome-Wide association studies involve testing many markers and are therefore sensible to this issue. The first section of this chapter is dedicated to a complete analysis of the classical multiple-testing correction approaches in genetic studies. We present and discuss the Family-Wise Error Rate (FWER), the False-Discovery Rate (FDR) or the local-FDR. This review of the classical correction procedures has been published in Methods in Molecular Biology (Bouaziz et al. 2012a).

We then focus on a particular issue in GWASs that is the determination of the genes associated with the disease. We introduced in **Section 1.3.4** this problem that con-

sists in considering several markers in linkage disequilibrium. Improved multiple-testing corrections and alternative methods accounting for the variable dependency between the markers as well are therefore needed in such situations. We present the main strategies to derive gene-wise measurements and compare them in a simulation study.

4.1 Multiple-testing and Genetic association studies

As we discussed previously, multiple-testing issues are a potential bias of association studies. Many statistical approaches, generally referred to as multiple-testing procedures, have been developed to deal with multiple-testing and the inherent problem of false-positives. They consist in reassessing probabilities obtained from statistical tests by considering more interpretable and suited statistical confidence measures than the usual p-values.

Several recent reviews dealt with the multiple-testing problem in the context of large-scale molecular studies (Balding 2006, Rice et al. 2008, Moskvina and Schmidt 2008, van den Oord 2008, Noble 2009, Chen et al. 2010). Our goal in this section is to provide an intuitive understanding of these confidence measures, an idea on how they are computed and some guidelines about how to select an appropriate measure for a given experiment. First we give some statistical basics about multiple-testing issues. Then we describe and discuss the main confidence measures (FWER, FDR and local-FDR). We also provide information about the p-values distribution which is an important point seldom considered in practice.

4.1.1 Introduction

When one considers several tests conducted in a single experiment, α can no longer be interpreted as the probability of false-positive of the whole set of tests but as the expected proportion of false-positive results. The generic situation is the following: when n statistical tests are performed, depending on whether each hypothesis tested is true or false and whether the statistical tests reject or does not reject the null hypotheses, each of the m results will fall in one of four outcomes defined in **Table 1.1**. This leads to the equivalent **Table 4.1** corresponding to multiple-tests, indicating the actual number of false-positives and false-negatives (fp and fn) instead of their respective rates of occurrence (α and β).

In the case of Genome-Wide association studies, the number of false-positives is higher than the expected number of true discoveries and unfortunately, it is not possible to know which null hypothesis is correctly or incorrectly rejected. It is therefore necessary to consider alternative confidence measures instead of the false-positive rate. More pertinent measures such as the FWER, the FDR and the local-FDR have been developed to this end and are presented in the following sections.

	$\mathbf{H_0}$ **is not rejected**	$\mathbf{H_0}$ **is rejected**	**Total**
H_0 is true	true-negatives (tn)	false-positives (fp)	$m_0 = tn + fp$
H_0 is false	false-negatives (fn)	true-positives (tp)	$m_1 = fn + tp$
Total	$m_U = tn + fn$	$m_R = fp + tp$	m

Table 4.1: **Outcomes of m statistical tests performed at the level α.**

4.1.2 Family-Wise error rate

Definition

The first alternative confidence measure proposed to handle the multiple-testing problem is referred to as the Family-Wise Error-Rate (FWER) criterion. It is defined as the probability of falsely rejecting at least one null hypothesis over the collection of hypotheses (or family) that is being considered for joint testing:

$$\text{FWER} = \mathbb{P}_{H_0}(fp \geq 1 \text{ at the level } \alpha).$$

Controlling the FWER at a given level is to control the probability of having at least one false-positive, which is very different from the false-positive rate. In practice, as the number (m) of tests increases, the false-positive rate remains fixed at the level α whereas the FWER generally tends toward 1.

Several procedure exist to control the FWER; here we introduce the Bonferroni and the Sidak procedures that can be considered as the reference methods.

The Bonferroni procedure

Certainly the simplest and most widely used method to deal with multiple-testing is to control the FWER by applying the Bonferroni adjustment (Bonferroni 1935 1936) which accounts for the number of tests. It is based on the following simple relation between the p-value of a test i (p_i), the number of tests performed (m) and the FWER at the level p_i (FWER_i):

$$\begin{aligned}
\text{FWER}_i &= \mathbb{P}_{H_0}(fp \geq 1 \text{ at the level } p_i) \\
&= \mathbb{P}_{H_0}(\{\text{test 1 is } fp \text{ at the level } p_i\} \text{ or} \ldots \text{or } \{\text{test } m \text{ is } fp \text{ at the level } p_i\}) \\
&\leq \sum_{k=1}^{m} \mathbb{P}_{H_0}(\{\text{test } k \text{ is } fp \text{ at the level } p_i\}) \\
&\leq mp_i,
\end{aligned}$$

because for each test i, $\mathbb{P}_{H_0}(\text{test } i \text{ is } fp \text{ at the level } p_i) = p_i$. The new confidence values $p_i^{\text{Bonf}} = mp_i$ correspond to the p-values adjusted by the Bonferroni procedure, and

represent upper bounds of the FWER_i. Controlling the FWER at a 5% level requires to apply a 5% threshold to the adjusted p-values corresponding to the product of each p-value with the number of tests. One can also prefer to apply a threshold of $5\%/m$ to the unadjusted p-values. For instance, to ensure that the FWER is not greater than 5% when performing 100 tests, each result can be considered as significant only if the p-value is less than $0.05/100 = 0.0005$. It makes no difference in term of results.

The major advantage of the Bonferroni procedure is that it is simple and straightforward to calculate and can easily be used in any multiple-testing application (Rice et al. 2008). However some authors argue that one major disadvantage of the Bonferonni procedure is that it over adjusts the p-values, resulting in a control of the FWER slightly more stringent than expected.

The Sidak procedure

An alternative was developed in the case of independent tests by Sidak (Sidak 1967) and is based on:

$$\begin{aligned} \text{FWER}_i &= \mathbb{P}_{H_0}(fp \geq 1 \text{ at the level } p_i) \\ &= 1 - \mathbb{P}_{H_0}(fp = 0 \text{ at the level } p_i) \\ &= 1 - (1 - p_i)^m \end{aligned}$$

This adjustment results in a more precise control of the FWER. However this approach assumes that all the tests performed are independent and may therefore not be suitable for every situation while the Bonferonni procedure does not make any assumption about the relation between the tests. Moreover in the case of a large number of markers tested, such as in Genome-Wide association studies, and small p-values in which we are interested in, the $1-(1-p_i)^m$ proposed by Sidak can be reasonably approximated by the mp_i proposed by Bonferroni.

Comments on the FWER

As a matter of fact, tests are often dependent when testing genetic markers that are statistically associated by linkage disequilibrium over the genome . In such situations, a practical and common alternative is to approximate the exact FWER using a permutation procedure (Balding 2006, Rice et al. 2008). Here, the genotype data are retained but the phenotype labels are randomized over the individuals to generate a dataset that has the observed LD structure but satisfies the null hypothesis of no association with the phenotype. By analyzing many of such datasets, the exact FWER can be approximated. The method is conceptually simple but can be computationally intensive, particularly as it is specific to a particular dataset and the whole procedure has to be repeated if the data is somehow altered (Balding 2006). Moreover it requires complex programming skills and statistical knowledge. However, the availability of relatively inexpensive and

fast computers and the use of built-in permutation utilities available for several genetic programs (such as `plink`) obviate these problems to some extent.

Actually the main disadvantage with the use of the FWER as a confidence measure is its unreliability with certain data sizes. This procedure works well in settings involving a few tests (e.g. 10-20, usual for candidate gene studies) and even when the number of tests is somewhat larger (e.g. a few hundreds as in genome-wide micro-satellite scans) (Rice et al. 2008). Yet the control of the FWER is not ideal when the number of tests is very large, as the level of significance becomes too stringent and as a result true associations may be overlooked and a consequent loss of test power occurs (Moskvina and Schmidt 2008). For instance in the context of Genome-Wide association studies, if 1 million genetic markers are tested, the p-value threshold for each marker must be set to 5×10^{-8} to control the FWER at 5% which is very low.

Even though such levels certainly would provide a safeguard against false-positives, they would also lead to an unacceptable number of false-negatives, particularly for complex traits where the loci effects are expected to be moderate at best. Consequently, less stringent confidence measure-based methods are designed such as the Holm procedure (Holm 1979) or the weighted Holm procedure (Dalmasso et al. 2008) to find a proper balance between false-positives and false-negatives in large-scale association studies. This constitutes one of the burning methodological issues in contemporary Genetic Epidemiology and statistical Genetics (Rice et al. 2008).

4.1.3 Analyzing the distribution of p-values to understand the basis of more advanced multiple-testing corrections

Mixture distribution of p-values

Analyzing the distribution of p-values provides a way to determine how many tests are declared under the null hypothesis (H_0) or under the alternative (H_1) and therefore to assess the multiple-testing problem. A very interesting result is that under H_0, the null distribution of p-values ($\mathcal{D}_0$) corresponds to a standard Uniform distribution on the interval $[0, 1]$ (**Figure 4.1-A**).

On the other hand, the alternative distribution of p-values ($\mathcal{D}_1$) under H_1 corresponds to a distribution that tends to accumulate toward 0 (**Figure 4.1-B**). In practice, one deals with p-values drawn under H_0 and H_1 which corresponds to a mixture distribution $\mathcal{D}$ of $\mathcal{D}_0$ and $\mathcal{D}_1$ (**Figure 4.1-C**) (McLachlan and Peel 2000):

$$f = \pi_0 f_0 + \pi_1 f_1,$$

where f, f_0 and f_1 are respectively the probability density functions defining the distributions $\mathcal{D}$, $\mathcal{D}_0$ and $\mathcal{D}_1$; π_0 and π_1 are the proportion of p-values generated under H_0 and H_1 respectively, with $\pi_0 + \pi_1 = 1$.

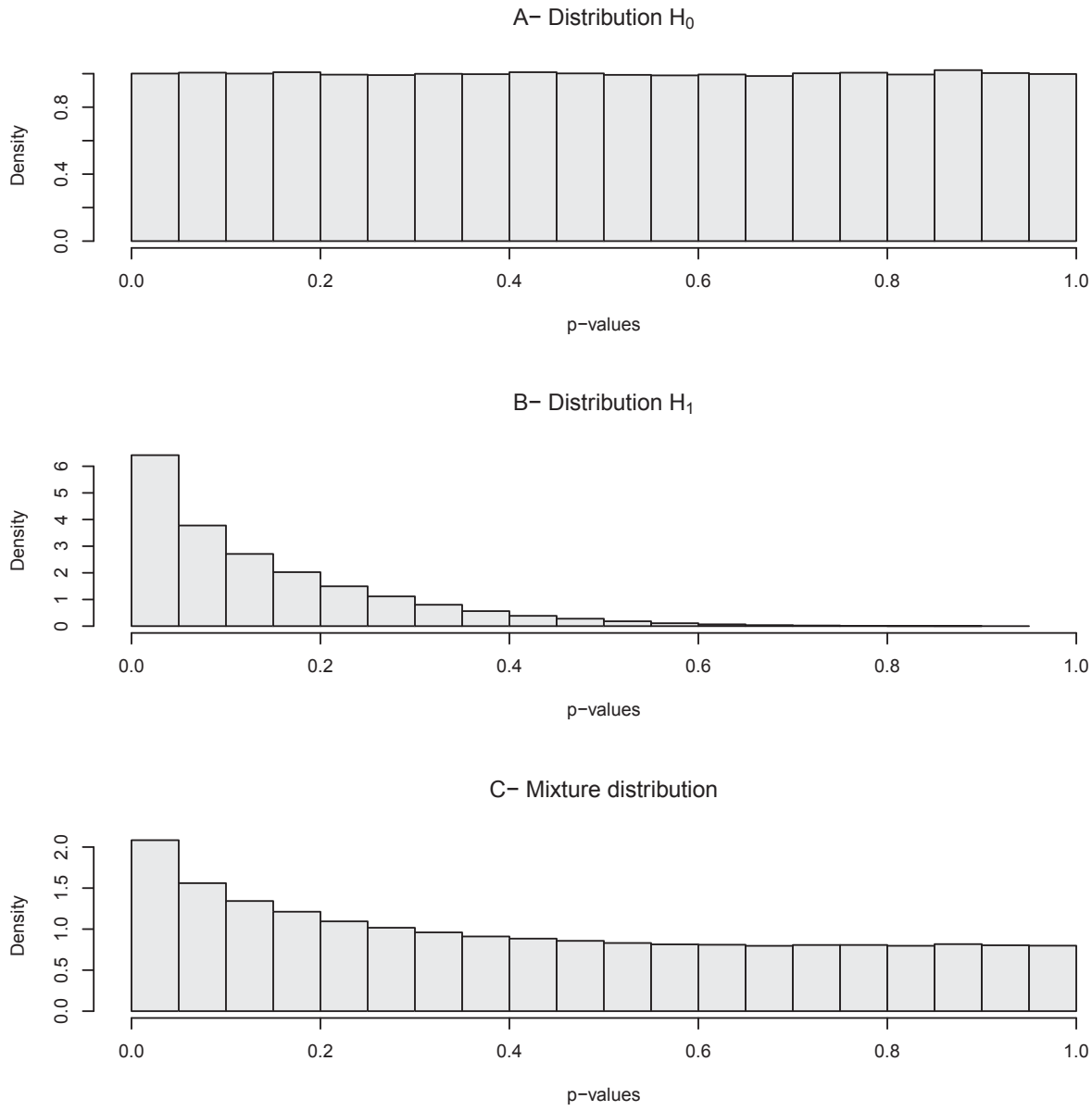

Figure 4.1: **A - Distribution of p-values under H_0. B - Distribution of p-values under H_1. C - Mixture distribution of p-values. Mixtures of Uniform and Beta distributions were used to simulate these p-value distributions.**

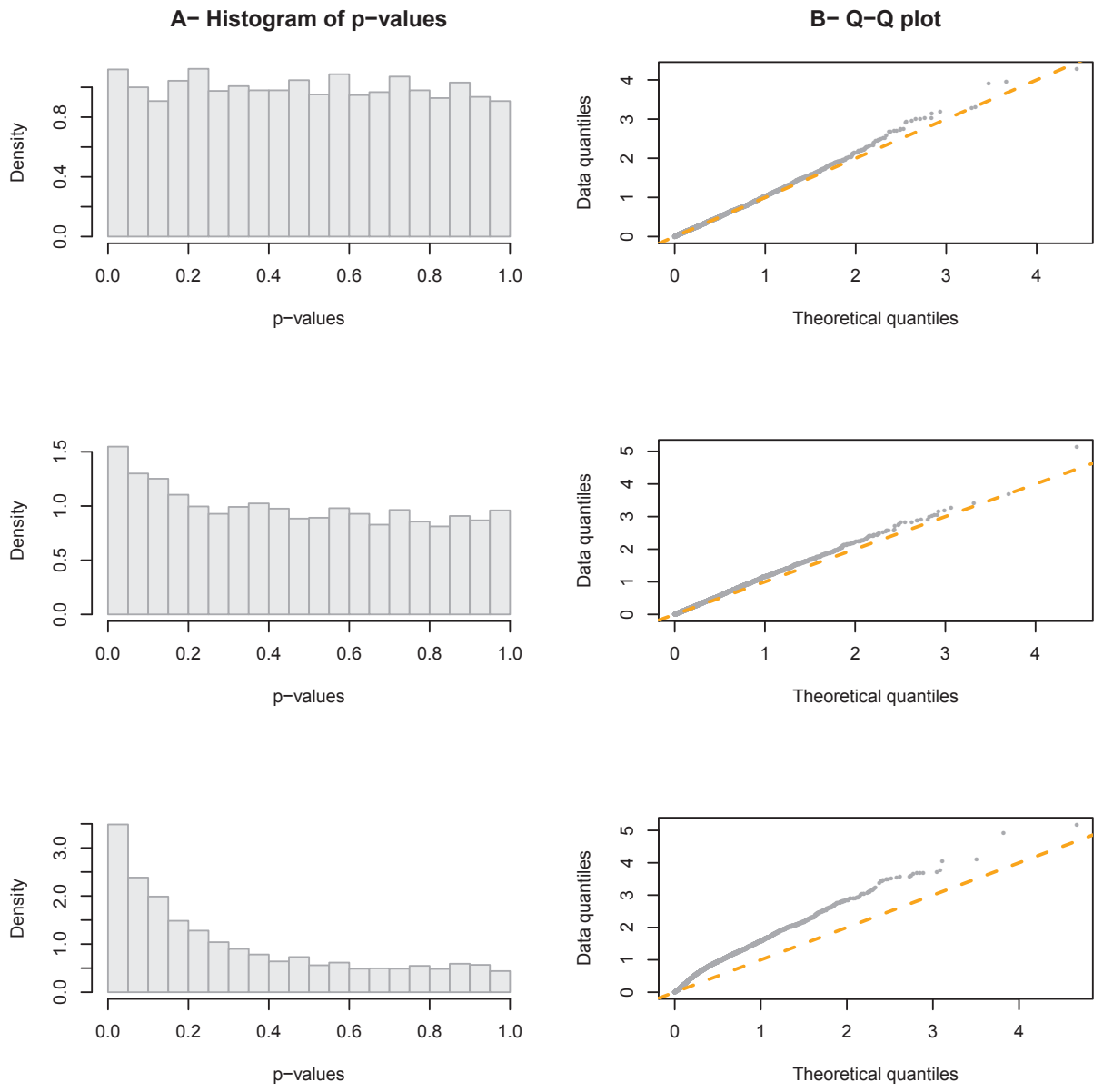

Figure 4.2: **Histogram of p-values (A) and Q-Q plots (B) for three increasing proportion of markers under H_1 ($\pi_1 = 2\%$, 10%, and 50%). Mixtures of Uniform and Beta distributions were used to simulate these p-value distributions.**

Finding evidence of the existence of true-positives

An usual question addressed at the outset of the analysis of large-scale genetic data is whether there is evidence that any of the markers tested and declared as significant at a given confidence level α is a true-positive. Investigating the distribution of p-values can help finding an answer.

Plotting the observed distribution of p-values

Plotting the distribution of p-values is a first intuitive approach to assess approximately the evidence of true-positives. Indeed, a simple histogram of the p-values can indicate whether the distribution is made of a mixture of p-values drawn under H_0 and H_1 (**Figure 4.1-C**) or is composed of p-values drawn under H_0 only (**Figure 4.1-A**). In the case the distribution is a mixture then some of the tests detected as significant may be true findings. **Figure 4.2-A** represents the histograms of p-values obtained for three increasing proportions of H_1.

An alternative and widely used approach is to compute a Quantile-Quantile plot (Q-Q plot) of the p-values comparing their distribution to the standard uniform distribution expected under H_0. If the two distributions are similar, the points in the Q-Q plot will approximately lie on the line $y = x$. If the observed quantiles are markedly more dispersed than the Uniform quantiles, this suggests that some of the significant results may be true-positives. **Figure 4.2-B** represents Q-Q plots of p-values obtained for the three increasing proportions of H_1.

Estimating the proportion of markers under H_0 and H_1

As we discussed previously, plotting the distribution of p-values can provide a qualitative and approximative criterion for assessing the proportion of genetic markers associated with the disease. Also, the actual proportions of p-values drawn under H_0 and H_1 (π_0 and $\pi_1 = 1 - \pi_0$, respectively), which are too rarely considered, can provide such information. They are also important to assess the False-Discovery-Rate, a confidence measure that will be presented in the next section. Furthermore a reliable estimate of π_0, the number of truly null tests, is of great relevance for calculating the sample size when designing the study (Wang and Chen 2004, Jung 2005). A variety of methods have been proposed to estimate π_0 based generally on statistical techniques such as mixture model estimation (Pounds and Morris 2003, McLachlan et al. 2006, Markitsis and Lai 2010), non-parametric methods (Mosig et al. 2001, Scheid and Spang 2004, Langaas and Ferkingstad 2005, Lai 2007) and Bayesian approaches (Liao et al. 2004).

The most simple and adopted method is inspired from the approach proposed by Storey and Tibshirani (Storey and Tibshirani 2003). The idea is to estimate π_0 from a part of the distribution between a given value λ and 1 where the distribution $\mathcal{D}_1$ drawn

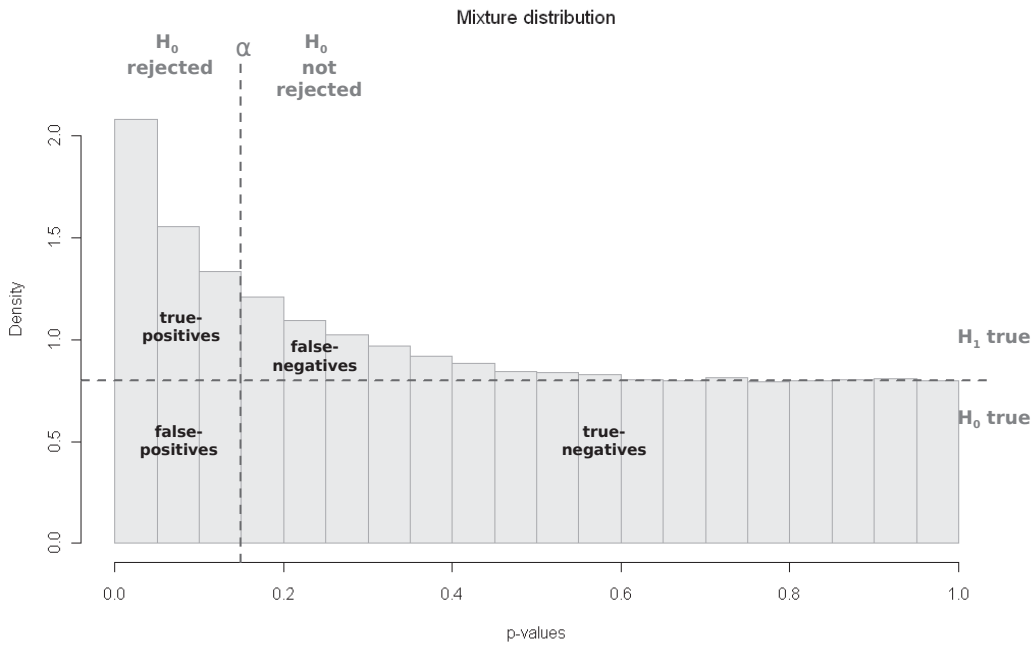

Figure 4.3: **Mixture distribution of p-values with the corresponding proportions of true-positives, true-negatives, false-positives, and false-negatives at the level α. A mixture of Uniform and Beta distributions was used to simulate this p-value distribution.**

under H_1 is negligible, and then normalized between 0 and 1:

$$\hat{\pi}_0(\lambda) = \frac{\text{number of } p\text{-values} > \lambda}{m \times (1 - \lambda)},$$

where m is the number of genetic markers tested. Some authors showed that a value of λ of 0.5 is a good compromise between bias and variability to assess π_0. An application of this procedure to the three distributions represented in **Figure 4.2** estimates π_0 of 98%, 90% and 50%, and hence π_1 of 2%, 10% and 50% respectively.

The Poisson test: To assess whether there are some true alternative hypotheses, it is common to perform a Poisson test. Indeed, when it is assumed that the tests are independent, a Poisson distribution can be used as a model for the number of significant

results at a given level α. If one assumes that all the m markers are under the null hypothesis then it is possible to determine, via the level α, the expected number of markers that should come out as significantly associated m_R^{exp}. Given the observed number of markers associated m_R^{obs}, the Poisson distribution allows us to compute a p-value indicating if m_R^{obs} is significantly greater than m_R^{exp}, in which case it is likely that some of the markers associated are true-positives. For example, if 500 independent tests are performed, each at the level $\alpha = 0.05$, we expect $m_R^{exp} = 500 \times 0.05 = 25$ significant tests to occur when all null hypotheses are true. If we observe $m_R^{obs} = 38$ significant markers, the p-value of such an observation is given by the Poisson distribution with mean 25, and is equal to 5.7×10^{-3}. So if we observe 38 significant markers out of 500 markers tested at a 5% level, it is very likely that some of them truly are associated with the disease. An application of this procedure to the three distributions represented in **Figure 4.2** based on 2,000 tests results in p-values of 0.12, 3×10^{-5} and 2×10^{-16}. A drawback of this approach is that it assumes that the tests are independent. In the case they are not, this method over-states the evidence that some of the alternative hypotheses are true when the test statistics are positively correlated, which commonly occurs in practice in Genome-Wide association studies due to LD. In such a situation, an equivalent test taking dependencies into account can be implemented based on permutations as we explained for the FWER, but remains computationally expensive.

Notes on the mixture distribution of p-values

Understanding and investigating the mixture distribution of p-values drawn under H_0 and H_1 is seldom considered in the analysis of large-scale genetic data. However it can provide a substantial amount of information. Plotting the distribution of p-values, assessing the proportion of markers under H_0 and H_1 and conducting a test based on the Poisson distribution are useful confidence indicators to confirm whether there is actually something to find in the data. **Figure 4.3** provides a graphical representation of a mixture distribution of p-values along with the potential outcomes of the corresponding statistical tests. Moreover the mixture model assumption and the estimation of the proportion of markers under H_0 and H_1 are very important to understand in order to apply more advanced multiple-testing correction approaches.

All these approaches make the theoretical, and often observed in practice, assumption that the distribution under H_0 is a standard uniform distribution. Any strong deviation from this expected distribution can alert about the appropriateness of the statistical test chosen for the analysis.

4.1.4 False-Discovery rate

Definition

Controlling the FWER is widely used, however, preventing against any single false-positive in large-scale genetic studies leads often to too stringent p-value corrections and therefore many missed findings. For these reasons, there is a certain reluctance to use FWER-based multiple-testing corrections such as the one proposed by Bonferonni (Moskvina and Schmidt 2008). In practice, most researchers would reasonably accept a higher risk of having a false-positive in return for a greater statistical power (Balding 2006). To overcome the limitations of the FWER, Benjamini and Hochberg introduced a more intuitive statistical concept as an alternative: the False-Discovery-Rate (FDR) (Benjamini and Hochberg 1995). FDR-based multiple-testing corrections use the following idea: instead of considering that one wants to be sure at 95% that none of the tests declared as significant is a false-positive (i.e. considering the FWER), the FDR method focuses on the expected proportion of truly null hypotheses that are falsely rejected.

Considering the m tests presented in **Table 4.1**, the total number of rejections (m_R) and the total number of non-rejections ($m_U = m - m_R$) at the level α are observable. However the values of tp, fp, tn and fn are unknown. With these notations, Benjamini and Hochberg defined the FDR, that is the expected proportion of tests falsely declared significant among all tests declared significant, as:

$$\text{if } m_R = 0,\ FDR = 0$$
$$\text{otherwise, } FDR = \mathbb{E}\left(\frac{fp}{fp+tp}\right) = \mathbb{E}\left(\frac{fp}{m_R}\right).$$

Intuitively, if 100 tests are predicted to be significant, i.e. $n_R = 100$, and if the FDR is controlled at 5% then $fp = 5$ of these tests should be false-positives.

The Benjamini-Hochberg procedure

The definition of the FDR proposed by Benjamini and Hochberg (Benjamini and Hochberg 1995) is the most often used as it comes with a simple procedure to apply it. For each test i, the FDR corresponding to the level p_i (FDR_i) is controlled by the expected number of test statistics declared as significant under H_0 ($m \times p_i \times \pi_0$) over the observed number of test statistics declared as significant (m_{Ri}) at the level p_i[1]:

$$\text{FDR}_i \leq \frac{m \times p_i \times \pi_0}{m_{Ri}}.$$

To obtain a correct estimation of the FDR it is therefore necessary to have a precise estimation of the proportion π_0. This implies having some knowledge about the alternative

[1] m_{Ri} corresponds to the number of tests with p-values smaller than p_i

hypothesis which is not always the case. The idea proposed by Benjamini and Hochberg to avoid the explicit calculation of π_0 is to use the fact that $\pi_0 \leq 1$ and therefore:

$$\mathrm{FDR}_i \leq \frac{m \times p_i}{m_{Ri}}.$$

The Benjamini-Hochberg procedure, like the FWER correction, corresponds to applying the threshold α to a set of adjusted p-values $p_i^{\mathrm{BH}} = \frac{m \times p_i}{m_{Ri}}$. In other words, if we order the m tests by their increasing p-values $(p_{(1)}, \ldots, p_{(m)})$, this procedure is equivalent to finding the largest k such as $p_{(k)} \leqslant \frac{k}{m}\alpha$ and rejecting the k first ordered null hypotheses.

Comments on the FDR

The main advantage of FDR estimation is that it allows to identify a set of candidate positives, of which a weak and controlled proportion are likely to be false-positives. The false-positives within the candidate set can then be identified in a follow-up study. Note that the FDR and the false-positive rate are often mistakenly equated, but their difference is actually very important (Storey and Tibshirani 2003). Given a level of confidence for declaring the test statistics significant, the false-positive rate is the rate that test statistics obtained under H_0 are called significant by chance whereas the FDR is the rate that significant test statistics are actually null. For instance, a false-positive rate of 5% means that on average 5% of the test statistics drawn under H_0 in the study will be declared as significant and a FDR of 5% means that among all test statistics declared as significant, 5% of these are actually drawn from H_0 on average.

One possible problem with the procedure of Benjamini and Hochberg is that when considering a sorted list of p-values in ascending order, it is possible for the FDR associated with the p-value at rank m to be higher than the FDR associated to the one at rank $m+1$ (Noble 2009). This non-monotonicity[2] can make the resulting FDR estimates difficult to interpret.

Moreover, a main assumption in the Benjamini-Hochberg approach is that the tests are independent. As already discussed before, this may not be the case when analyzing large-scale genetic data. To account for such situations, Benjamini and Yekutieli (Benjamini and Yekutieli 2001) developed a quite similar procedure where the FDR is controlled by:

$$\mathrm{FDR}_i \leq \frac{m \times p_i}{m_{Ri} \times c(m)},$$

where $c(m)$ is a function of the number of tests depending on the correlation between the tests. If the tests are positively correlated, $c(m) = 1$, there is no difference with the approach presented before. Otherwise if the tests are negatively correlated, $c(m) = \sum_{k=1}^{m} \frac{1}{k}$.

[2] i.e. the FDR does not consistently increase

Although the Benjamini-Hochberg procedure is simple and sufficient for many studies, one can argue that an upper bound of 1 for π_0 leads to a loss of precision in the estimation of the FDR. Such estimations are actually probably conservative with respect to the proportion of test statistics drawn under H_0 and H_1. This means that if the classical method estimates that the FDR associated with a collection of p-values is 5%, then on average the true FDR is lower than 5%.

Consequently, a variety of more sophisticated methods introducing the estimation of π_0 have been developed to achieve more accurate FDR estimations. Depending on the data, applying such methods may make a big difference or almost no difference at all (Noble 2009). To this end, Storey introduced the q-value defined to be the FDR analogue of the p-value (Storey and Tibshirani 2003). Storey's procedure can be considered as analogue to the procedure of Benjamini and Hochberg except that it incorporates an estimate of the null proportion π_0.

Finally, as mentioned in the previous section, FDR-based corrections strongly depend on the assumption that the p-values are uniformly distributed if there is no association, which may not always be the case in practice. Methods providing more exact estimations of the FDR in such situations exist and rely on more advance statistical and algorithmic notions (Wojcik and Forner 2008).

4.1.5 Local false-discovery rate

Definition

The FDR criterion introduced in the previous section has received a great focus during the last decades due to its lower conservativeness compared to the FWER. The FDR is defined as the mean proportion of false-positives among the list of rejected hypotheses. It is therefore a global criterion that cannot be used to assess the reliability of a specific genetic marker. More recently, a strong interest has been devoted to the local version of the FDR, called 'local-FDR' (Efron and Tibshirani 2002) and denoted hereafter fdr. We briefly present this new confidence measure in the following.

The idea is to quantify the probability for a given null hypothesis to be true according to the specific p-value of each genetic marker tested. For a given p-value p_i associated to a test i:

$$\begin{aligned}
\text{fdr}_i &= \mathbb{P}(\text{test statistic } i \text{ is under } H_0 \text{ knowing } p_i) \\
&= \frac{\mathbb{P}(H_0)\mathbb{P}_{H_0}(p_i)}{\mathbb{P}(p_i)} \\
&= \frac{\pi_0 f_0(p_i)}{f(p_i)} \\
&= \frac{\pi_0 f_0(p_i)}{\pi_0 f_0(p_i) + \pi_1 f_1(p_i)},
\end{aligned}$$

where π_0 and π_1 are the proportions of p-values generated under H_0 and H_1 respectively, and f_0, f_1 and f the density functions corresponding to the distributions $\mathcal{D}_0$, $\mathcal{D}_1$ and $\mathcal{D}$ as described in **Section 4.1.3**. In general, the local-FDR is more difficult to estimate than the FWER and the FDR due to the difficulty in estimating density functions. Many approaches have been proposed that are fully parametric (Allison et al. 2002, Pounds and Morris 2003, Liao et al. 2004, McLachlan et al. 2006), semi-parametric (Robin et al. 2007) that is implemented in a `R` package `Kerfdr` (Guedj et al. 2009), Bayesian (Broet et al. 2004, Newton et al. 2004) or empirical Bayes (Efron and Tibshirani 2002), and most of them rely on the mixture model assumption described in **Section 4.1.3**.

Comments on the local-FDR

The main advantage of the local-FDR over the more classical FDR measure, is that it assesses for each genetic marker its own measure of significance. In this sense it appears more intuitive and precise than the FDR. However its estimation requires to determine with precision π_0 and the distribution under H_1 ($\mathcal{D}_1$) which requires more advanced statistical skills and the use of fully developed algorithms. In addition, like for the other alternative confidence measures that we presented, most of the algorithms to assess the local-FDR assume the independence of the tests.

4.1.6 Conclusions

Genome-Wide association studies are attractive but susceptible to certain sources of bias such as the one induced by multiple-testing. To this end, multiple-testing corrections have been developed, based on alternative measures of significance adapted to genome-wide studies. Nowadays, FWER, FDR and local-FDR constitute key statistical concepts that are widely applied in the study of high-dimensional data.

The FWER controls the probability of having one or more false-positives and turns out to be too stringent in most of the situations. As a more intuitive alternative, the FDR considers the proportion of significant results that are expected to be false-positives. More recently through the local-FDR, authors have proposed to consider the actual probability for a result of being under H_0 or H_1. Considering the fact that several confidence measures can be applied to account for multiple-testing, a question that naturally arises is which method one should use and also if a FWER correction, due to its stringency, is even appropriate. Noble provided a practical solution to the problem (Noble 2009). Based on the same rationale motivating the choice of a significance threshold, choosing which multiple-testing correction method to use depends on the cost associated with false-positives and false-negatives. For example, if one's follow-up analyses will focus on only few experiments, then a FWER-based correction is appropriate. Alternatively, if one plans on performing a collection of follow-up experiments and tolerate having some of them that fail, then the FDR correction may be more appropriate. Finally, if one is inter-

ested in following up on a single gene, local-FDR may be precisely the more suited method.

Correlations between genetic markers, due to LD for instance, is still a difficult issue when considering multiple-testing. Such local dependencies between the tests lead to a smaller number of independent tests than markers examined.

As we have seen, some approaches allow taking dependent tests into account such as permutation-based methods (Benjamini and Yekutieli 2001, Moskvina and Schmidt 2008, Wojcik and Forner 2008), but require far more advanced programming skills and statistical knowledge than the simple Bonferroni or Benjamini-Hochberg procedures. Alternative procedures aim to identify the effective number of independent tests conducted in order to derive confidence measures that account for these dependencies (Li and Ji 2005). Such approaches are not yet applied to whole genome scans but are considered when looking at smaller sets of markers, usually in high LD. We will focus on such approaches to derive gene-level p-values instead of SNP-level ones in **Section 4.2**.

On the other hand, theoretical and simulation studies suggest that multiple-testing corrections assuming the independence of the tests perform quite well in cases of weak positive correlations, which is common in many large genetic studies (Benjamini and Yekutieli 2001, Moskvina and Schmidt 2008).

This issue of dependency in the analysis of large-scale molecular data is likely to be a live field of research in the near future. As statistical techniques are still developing to account for the complexities involved in controlling false-positives in exploratory researches, independent replications and validations still remain necessary steps in the discovery process (van den Oord 2008).

4.2 Gene-level association

4.2.1 Introduction

Gene-wise interpretation of GWASs

Genome-Wide association studies yield results at the SNP level, that are sets of SNPs associated with the disease. A secondary step of these studies is usually to derive a gene-level interpretation of the findings. As a matter of fact, in Genetics, the gene is often considered as the unit of interest as the analyses of the functional mechanisms of a disease are generally based on genes and their products such as RNA or resulting proteins (Jorgenson and Witte 2006). Determining the genes associated with the disease opens the door to a lot of additional research such as targeting genes of interests for candidate-gene studies or replicate association studies. Also, it allows the consideration of biological information, such as pathways or protein interactions, in the analysis of GWASs (Neale and Sham 2004).

For instance, enrichment analysis such as performed by the method Gene Set Enrichment Analysis (GSEA) (Subramanian et al. 2005) aims to determine sets of genes involved in a common biological process (e.g. defined in the Gene Ontology[3]) or biological pathways (e.g. defined in the KEGG database). Such an analysis is possible through the use of functional information that is only available at the gene or protein level. It is therefore necessary to obtain association information at the gene-level. To derive a gene-level measure of significance, such as a test statistic or a p-value, one needs to combine the results of all the SNPs corresponding to the gene.

Statistical issues and types of approaches

Computing a single p-value per gene raises several statistical issues. First, several SNPs are usually genotyped within a gene and combining the results of each individual SNP test outcome corresponds to a multiple-testing situation. This fact is often observed in practice as large genes tend to come out as the most significant in most of the studies. In addition, markers within a gene are usually closely located on the genome and therefore likely to be in linkage disequilibrium. This LD pattern of a gene leads to a situation of multiple-testing with dependent tests. As a consequence, statistical tests that aim to derive gene-level p-values have to account for both these features of the structure of a gene.

The most widely used approach to compute a single p-value per gene is to consider the most significant p-value of its SNPs (Torkamani et al. 2008). This method is however biased as it does not account for the two statistical issues described above (Hong et al. 2009). The use of permutation procedures is however possible to computed an unbiased p-value per gene using this approach. These permutation procedures are however quite time consuming to reach a sufficient level of precision which has led to the development of more advanced multiple-testing corrections that allow to re-assess SNP p-values adjusted on the total number of SNPs of the gene and the LD pattern between them (Li and Ji 2005, Li et al. 2011). The most significant adjusted p-value can then be selected to represent the gene.

Such an approach tends to focus on genes that are affected by at least one significant mutation. It is also of interest to obtain a gene-level measure that considers the situation where the gene is affected by an accumulation of significant mutations (Lehne et al. 2011). To that end, other strategies have been designed to include the information contained in all the SNPs simultaneously. A natural approach is to consider another statistic than the minimum to apply to the SNP p-values or test statistics such as for instance the mean or a quartile (Moskvina et al. 2011, Lehne et al. 2011). These methods may however also imply the computation of permutations to assess the gene p-value. A recent range of statistical methods aiming to assess the global significance of sets of markers has been

[3] `http://www.geneontology.org`

proposed and can be used to derive gene-level p-values in this simultaneous-testing perspective (Goeman et al. 2005, Wang and Elston 2007, Chapman and Whittaker 2008, Pan 2009).

We present in **Section 4.2.2** these different approaches and then propose in **Section 4.2.3** a comparison of the most used in practice. We based our analysis of the different methods on a set of realistic simulations by considering various gene sizes, LD patterns and strengths of association. We aim at providing indications regarding which tests correctly account for the number of markers and the LD pattern by assessing the false-positive rate and the power. Also we are interested in determining the gain in power that approaches considering several mutations per gene can have over those focusing on a single mutation per gene.

4.2.2 Strategies to derive gene-level p-values

We present here the main approaches to derive gene-level p-values from a set of SNPs in the case of a case-control genetic study.

Let Y be the phenotype of the n individuals and G a gene with T_G and p_G its test statistic and p-value that we aim to assess. Let $X = (X_1, \ldots, X_m)$ be the m markers that compose the gene G with $(S_1, \ldots, S_m)$ and $(p_1, \ldots, p_m)$ their association test statistics and p-values.

The test that we consider aim to assess the null hypothesis

$$H_0 : \{\text{The gene } G \text{ is not associated with the disease.}\}.$$

Permutation procedure

Principle

A permutation procedure is a method that consists in assessing the unknown distribution of a test statistic under a null hypothesis H_0 by permuting the labels of the samples. We consider that we estimated a statistic $T_G = f(S_1, \ldots, S_m)$, where f is a function that can correspond to the mean for instance, and that the distribution of T_G under H_0 is unknown. A permutation procedure computes B novel sets of SNP statistics $(S_1^1, \ldots, S_m^1), \ldots, (S_1^B, \ldots, S_m^B)$ each based on the association test where the sample labels have been permuted. Then, B statistics $T_G^1, \ldots, T_G^B$ are calculated on each of the new SNP statistics set. The permutation process breaks the association between the markers and the disease, therefore between the gene and the disease, without modifying the features of the data. The new statistics $T_G^1, \ldots, T_G^B$ therefore represent the distribution of T_G under H_0.

It is then possible to empirically compute a gene p-value

$$p_G = \frac{\#(\{T_G^i \geqslant T_G, i = 1 \dots B\})}{B},$$

where #() represents the cardinal function.

Comments
This procedure produces accurate estimations given that a reasonable number of simulated data is considered. This however renders the estimation process time consuming. Indeed, for each simulated data, it is necessary to compute all the SNP statistics and the gene statistic.

In practice permutation methods are not always applicable to test a large number of genes which leads to the necessity for faster statistical tests.

Multiple-testing corrections accounting for dependent tests

Several multiple-testing corrections that account for dependency between the tests have been proposed including some based on permutations. We introduce in this section a faster method that assesses the number of effectively independent tests (M_{eff}) (Li and Ji 2005). The different multiple-testing procedures introduced in **Sections 4.1.2 and 4.1.4** are based on the number of statistical tests conducted m. The idea of the approach of Li and Ji is to substitute to this number of tests, the effective number of tests M_{eff}.

Effective number of tests
The effective number of tests can be seen as the number of independent sets of correlated tests, which corresponds to the number of independent LD blocks of SNPs. This estimation uses the correlation matrix C between the markers that serves as an estimation of the linkage disequilibrium pattern and the eigenvalues of this matrix $(\lambda_1, \dots, \lambda_m)$ that represents degrees of correlation between the markers.

The effective number of tests is calculated as follow

$$M_{eff} = \sum_{i=1}^{m} f(|\lambda_i|),$$

where

$$f(x) = \mathbb{I}(x \geqslant 1) + (x - \lfloor x \rfloor),$$

$\mathbb{I}(x \geqslant 1)$ is the indicator function that gives 1 if $x \geqslant 1$ and 0 otherwise and $\lfloor x \rfloor$ is the floor function that gives the largest integer not greater than x.

Application to the FWER

We introduced in **Section 4.1.2** the Family-Wise Error-Rate correction via the Sidak procedure. If one considers the markers as independent then each marker p-value p_i can be adjusted for multiple-testing by

$$p_{i,adj}^{Sidak} = 1 - (1 - p_i)^m,$$

with m being the number of markers.

In order to also account for the dependency between the markers the idea is to replace the number of markers by the effective number M_{eff} which leads to

$$p_{i,adj}^{Sidak} = 1 - (1 - p_i)^{M_{eff}}.$$

Application to the FDR

Another multiple-testing method that can be adjusted for the dependency between the tests is the False-Discovery Rate. The Benjamini-Hochberg procedure to estimate the FDR presented in **Section 4.1.4** corresponds to calculating adjusted p-values

$$p_{i,adj}^{BH} = \frac{m.p_i}{m_{R_i}},$$

where m_{R_i} is the number of markers with p-values lower than p_i.

In order to account for the dependency it is possible to consider a novel correction

$$p_{i,adj}^{BH} = \frac{M_{eff}.p_i}{M_{eff_{R_i}}},$$

where $M_{eff_{R_i}}$ is the estimated number of independent markers among the set of markers with p-values lower than p_i.

Comments

These two methods are based on the calculation of correlation matrices between the markers. It is faster to compute these matrices than using permutation procedures, especially with the use of efficient software such as `plink` that includes special algorithms to compute correlations between sets of SNPs.

Adjusted Meta-analyses

Meta-analysis methods allow to aggregate several p-values from independent tests to produce a global measure of significance over all the test (see **Section 2.3.6**). When the m tests are independent, the overall Fisher statistic

$$T^{Fisher} = -2 \sum_{i=1}^{m} log(p_i),$$

follows a $\chi^2(2m)$ distribution under H_0.

When the tests are dependent the distribution of this statistic under the null is no longer a $\chi^2(2m)$ but has a mean equal to $2m$ and a variance σ^2 (Brown 1975) with

$$\sigma^2 = 4m + 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} cov(-2log(p_i), -2log(p_j)).$$

It has been proposed to estimate this variance in order to derive a test statistic that follows a χ^2 distribution (Moskvina et al. 2011). The covariance term of the variance σ^2 can be estimated by

$$cov(-2log(p_i), -2log(p_j)) = \rho_{ij}.(3.25 + 0.75\rho_{ij}),$$

where $\rho_{ij} = cor(X_i, X_j)$. This estimation uses a Gaussian quadrature method.

Finally the corrected test statistic

$$T_{adj}^{Fisher} = \frac{4m.T^{Fisher}}{\sigma^2},$$

follows a $\chi^2_{8m^2/\sigma^2}$ distribution under H_0.

Regression Models

Regression models that have been detailed in **Section 2.3.5** provide the possibility to include several markers in a single model to test their combined effect. Several of these models have been adapted to the analysis of sets of markers that corresponds to the analysis of a gene.

Uncorrected multi-Regression

A first approach is to consider a regression model that includes all the markers

$$\text{logit}(\mathbb{P}(Y = 1 \mid X_1, \ldots, X_m)) = \beta_0 + \sum_{j=1}^{m} \beta_j X_j.$$

It is possible to test the overall effect of all the markers, i.e. of the gene with

$$H_0 : \{\beta_1 = \ldots \beta_m = 0\},$$

by the use of a classical score test statistic

$$T^{Reg} = S(\hat{\beta}_{H_0})' I_f^{-1}(\hat{\beta}_{H_0}) S(\hat{\beta}_{H_0}),$$

where $\hat{\beta}_{H_0}$ is the maximum likelihood estimator of $\beta = (\beta_1, \ldots, \beta_m)$ under H_0, $S(\hat{\beta}_{H_0})$ the derivative in $\beta = \hat{\beta}_{H_0}$ of the log-likelihood and I_f the Fisher information. This statistic follows a $\chi^2(m)$ distribution under the null hypothesis.

This test accounts for the number of markers but does not consider the dependency between them. As a matter of fact, the dependency between the markers is bound to affect the validity of the model as correlated covariates violate the assumption of regression models.

Goeman test

The Goeman test is related to the classical association score test presented for the uncorrected multi-logistic Regression but assumes an empirical Bayesian model (Goeman et al. 2005). This model uses independent priors for the marker parameters which corresponds to considering that β is random with mean $\mathbb{E}(\beta) = 0$ and covariance $cov(\beta) = \sigma^2 I$ where I is the identity matrix. In this framework, testing the hypothesis

$$H_0 : \beta = 0$$

is equivalent to testing

$$H_0 : \sigma^2 = 0.$$

The Goeman test statistic corresponds to the empirical Bayesian score test statistic of the logistic model which can be expressed as (Goeman et al. 2005)

$$T^{Goeman} = \frac{1}{2}(U'U - trace(I_f)),$$

where $U = X'(Y - \bar{Y} \times \mathbf{1})$ with $\bar{Y} = \sum_{i=1}^{n} Y_i$ and $\mathbf{1}$ is a column vector of size n with all elements equal to 1.

By expressing the classical score test statistic such as suggested in (Chapman et al. 2003)

$$T^{Reg} = U'V^{-1}U,$$

where $V = cov(U)$, one can see the close relation with the Goeman statistic.

Also, a development of the Fisher information matrix expression leads to

$$I_f = \bar{Y}(1 - \bar{Y}) \times (X - \bar{X})'(X - \bar{X}),$$

so that finally, the Goeman test statistic has the form (Chapman and Whittaker 2008)

$$T^{Goeman} = \frac{1}{2}\Big((Y - \bar{Y} \times \mathbf{1})'XX'(Y - \bar{Y} \times \mathbf{1}) - Y(1 - \bar{Y}) \times trace\big((X - \bar{X})'(X - \bar{X})\big)\Big).$$

The test statistic has an unknown distribution under the null hypothesis. A permutation procedure could then be used to estimate the null distribution however as we

indicated earlier this can be quite time consuming. An alternative to derive the null distribution of this statistic has been proposed and is based on a decomposition of the test statistic (Chapman and Whittaker 2008). If one uses permutations then only the part $U'U$ of the statistic is random, the other terms being fixed. By noticing that the distribution of U is known under H_0 Chapman et al. proposed a way to obtain the null distribution using simulations. Under H_0, U follows a normal distribution $\mathcal{N}(0, I_f)$ as a consequence it is possible to simulate many U and $U'U$ and to reconstitute the statistic T_{Goeman}. A p-value can then be assessed like for permutation procedures, by estimating the proportions of simulated test statistics larger than the observed statistic.

Technically this test does not account for dependency between the markers. It ignores this correlation as opposite with test who assume the independence. The idea is to consider that if several markers are in LD and that if among them one is associated with the disease, then the others are bound to be associated as well so that some sort of average of the associations can summarize the signal in the corresponding block.

Marginal score test

Pan et al. have developed another statistical test, also closely related to the classical score test, but that is based on the marginal effect of each marker (Pan 2009). This test, like the Goeman test, ignores the correlation between the markers and computes an averaged signal. The statistic of this test is

$$T^{MargST} = U' Diag(I_f)^{-1} U.$$

Contrary to the usual score test, a diagonal matrix is used which favors the marginal effect of each marker. Pan et al. also demonstrated that this test statistic follows a quadratic distribution that can be approximated by $a\chi^2(d) + b$ where

$$a = \frac{\sum_{j=1}^{m} c_j^3}{\sum_{j=1}^{m} cj^2},\ b = \sum_{j=1}^{m} c_j - \frac{\left(\sum_{j=1}^{m} c_j^2\right)^2}{\sum_{j=1}^{m} c_j^3} \text{ and } d = \frac{\left(\sum_{j=1}^{m} c_j^2\right)^3}{\left(\sum_{j=1}^{m} c_j^3\right)^2},$$

where $(c_1, \ldots, c_m)$ are the eigenvalues of the matrix $W Diag(W)^{-1}$ with

$$W = Diag(I_f)^{-1} I_f Diag(I_f)^{-1}.$$

Close alternatives to this marginal score test have been designed and are based on different normalizations of the classical score test. Indeed, the classical score test is normalized by V^{-1} while the marginal score test uses $Diag(I_f)^{-1}$.

Other approaches

Several other approaches have been proposed to tackle the question of deriving gene-level p-values. They can be based on alternative function to estimate the effective number of

independent tests (Li et al. 2011), on random effect Regressions (Tzeng and Zhang 2007), on the elasticNet method (Friedman et al. 2010), on selecting tagSNP within the genes (Huang et al. 2011), or on Fourier transform of the genotypes to reduce the degree of freedom of the tests (Wang and Elston 2007).

4.2.3 Comparison of approaches to obtain gene-level information

We now propose a comparison of the main approaches dedicated to derive gene-level p-values. Several comparisons have been proposed to assess these strategies (Chapman and Whittaker 2008, Pan 2009, Ballard et al. 2010, Lehne et al. 2011). Many of these reviews focus on regression models: Chapman et al. proposed a comparison of several regression models, Pan et al. designed and compared different regression score-like tests, Ballard et al. focused on the differences between regression with fixed and mixed effects. On the other hand, Lehne et al. compared the different types of approaches that are considering the minimum of p-value of the markers or another function such as the mean or a quartile.

We decided to perform a global analysis of all these methods including at the same time all the different types of approaches presented in **Section 4.2.2**. Our goal is to determine the different behaviors of the strategies when applied to various types of genes. We aim to highlight the approaches that are properly capable to account for the size of the gene and the pattern of LD. Also we are interested in evaluating the difference in practice between methods focusing on the stronger SNP association signal and those accounting for all the signals simultaneously.

Our comparison is based on a set of realistic simulations where the number of markers, the LD patterns and the strength of the SNP associations are controlled.

Methods included in the comparison

We considered 11 usual methods to include in our comparison. We first selected several multiple-testing methods that are the False-Discovery rate and Sidak procedure without correction for dependent tests (FDR and Sidak) along with the same methods with the correction for dependent tests (FDR.corr, Sidak.corr). We also considered the meta-analyses methods with and without the adjustment for dependency (Fisher and Fisher.corr). We then focused on several Regression models that are the uncorrected Regression (Reg), Goeman test (Goeman) and the marginal score test (MargST). This choice of methods among all the possible Regression models was motivated by the results of several previous comparison studies (Chapman and Whittaker 2008, Pan 2009). It has been demonstrated that the Goeman test and the marginal score test have improved performances over the other Regression based tests. The uncorrected Regression test can be useful as a reference for the Regression methods as it corresponds to the simpler and more classical approach.

In addition, we considered one permutation based method (perm) that is based on the mean of the SNP test statistics. We aim at answering the question: does the gain in power of such a method compensates its high computational cost? Finally we included in the comparison the minimum of the SNP p-values method (minP) that is the most used in practice. This method is not adjusted for multiple-testing and dependency and is as a consequence used as a reference method to assess the quality of the other approaches. This latter method is also of use to demonstrate the crucial necessity of using adjusted methods instead of the raw minimum of the p-values.

Simulation model

We applied the different strategies to several simulated genes. Our simulation process is based on a two-stage strategy. In a first step, the structure of LD of a gene is determined and then the SNPs corresponding to this structure are simulated. We call G a gene and assume that it is composed of m SNPs.

Determination of the LD structure

First, one has to set the numbers p of LD blocks and q of independent SNPs that compose the gene. Then our algorithm randomly simulates the numbers $m_j, j = 1, \ldots, p$ of SNPs per block so that $\sum_{j=1}^{p} m_j + q = m$. Finally, positions on the gene G are randomly selected to fit the p LD blocks and the q independent markers.

Simulation of the markers

Markers are simulated using the simulation model presented in **Section 2.4.3** that allows to specify the degree of association.

The simulation of independent markers is straightforward as the model can directly be applied to the provided simulation parameters.

The simulation of LD blocks uses iterative permutations of the genotypes between several markers. As a matter of fact, it is possible to create a correlation between two SNPs by permuting the values of one of them. For example, let us consider the Pearson correlation coefficient between two SNPs x and y

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_x\sigma_y}.$$

One can calculate the contribution to this coefficient of a pair of values (x_i, y_i) as presented in **Table 4.2**.

It is then possible to estimate the variation of correlation between x and y that is created when two values of the SNP y are permuted (**Table 4.3**). Using these tables, one can create a correlation or decorrelate two SNPs.

In order to simulate a LD block, we first simulate several SNP markers with the provided parameters and select a marker of reference. The global pattern of LD is created

$\mathbf{x_i}$	$\mathbf{y_i}$	Contribution c_i to ρ_{xy}
0	0	0
0	1	$-\bar{x}$
0	2	$-2\bar{x}$
1	0	$-\bar{y}$
1	1	$1-\bar{x}-\bar{y}$
1	2	$2-2\bar{x}-\bar{y}$
2	0	$-2\bar{y}$
2	1	$2-\bar{x}-2\bar{y}$
2	2	$4-2\bar{x}-2\bar{y}$

Table 4.2: **Contribution of a pair of values (x_i, y_i) to the correlation ρ_{xy}.** Each pair (x_i, y_i) contributes to the correlation by $\frac{c_i+\bar{x}\bar{y}}{(n-1)\sigma_x\sigma_y}$.

	Case 1 **i j**	Case 2 **i j**	$\lvert\Delta\rho\rvert$
x	1 0	1 0	1
y	0 1	1 0	
x	1 2	1 2	1
y	2 1	1 2	
x	0 2	0 2	4
y	2 0	0 2	
x	2 1	2 1	2
y	0 2	2 0	
x	2 1	2 1	1
y	0 1	1 0	
x	0 1	0 1	2
y	2 0	0 2	

Table 4.3: **Absolute difference of correlation between two SNPs when the values of one of them (here y) are permuted between a pair of individuals.** This table compares the correlation between the SNPs x and y between a situation represented by Case 1 and a situation represented by Case 2 where the values of the SNP y have been permuted between the samples i and j. Each difference has to be divided by $(n-1)\sigma_x\sigma_y$. The sign of the difference of correlation can be changed by inverting Case 1 and Case 2.

by simultaneously correlating all the markers to the reference marker and by considering a certain amount of common permutations. Technically one has to provide a range of possible correlations for the LD block and the common permutation procedure ensures

that all the pairwise correlations of the block are within this range.

If the initial marker is associated with the disease, then we expect that all the markers simulated in LD with it will also be associated with the disease. In order to ensure the degree of association within the desired relative-risk bounds, it is possible to permute the genotypes within the case or control samples only.

Such a simulation process may be time consuming as it implies a careful choice of the permutations to ensure a global LD pattern and control the associations. It has however the advantage of leading to more realistic LD patterns that those considered in the simulations of (Wang et al. 2007) as the disease associated marker is not necessarily in the center of the LD block and the correlations do not decrease with the distance to this locus.

This simulation process eventually allows to simulate several independent markers and several LD blocks for a single gene by controlling the degree of LD and the strength of association.

Gene structure scenarios

We considered several scenarios pertaining to different parameters susceptible to influence the strategies analyzed. We simulated several gene sizes (3, 10 and 30 SNPs), different patterns of LD (no LD, moderate LD and high LD) and different patterns of association that corresponded to the number of markers, when there was no LD, or of blocks of markers, when there was LD, that were associated with the disease (several markers/blocks, one marker/block, no marker associated). We refer in the following to multiple markers/blocks association as strong association and to single marker/block association as weak association. Genes simulated with no marker associated were used to assess the false-positive rate while genes with associated markers were used to assess the power.

Each gene was simulated on 500 cases and 500 controls. The relative risk for the non-associated markers was fixed at 1 and was randomly chosen between 1.1 and 1.3 for the associated markers. The pattern of moderate LD implied a correlation between the markers of a block comprised between 0.4 and 0.7 and the high LD pattern referred to a correlation between 0.7 and 1. Each simulated gene was replicated $10,000$ times in order to provide accurate estimates of the false-positive rates and powers of the methods.

Figures 4.4 to 4.8 present the different patterns of LD and association when there is one in function of the size of the genes. We can observe the strong patterns of association on **Figures 4.5 and 4.7** where at least more than two markers/blocks are associated with the disease and the weak patterns in **Figures 4.6 and 4.8** where only one marker/block is associated with the disease.

Comparison strategy

For each scenario and each pattern of LD we assessed the false-positive rate and the power on $B = 10,000$ datasets using the method described in **Section 2.4.5**. This corresponds to estimate the quantity

$$\frac{\#\big(\{p\text{−value}_i \leqslant \alpha, i = 1 \dots B\}\big)}{B},$$

where the threshold α is set at 5%.

In order to present the results of the comparison of the methods we decomposed these simulations in several scenarios in function of the pattern of LD of the genes. Scenario 1 corresponded to no LD, Scenario 2 to a moderate LD and the Scenario 3 to a high LD. Each scenario therefore included several gene sizes and strengths of association. The difference between weak and strong association was only considered for 10 and 30 SNPs as it was not possible to significantly vary the number of associated markers with 3 SNPs. **Table 4.4** details the different scenarios indicating for each the gene size and the pattern of association.

Scenario	LD	dº Assoc	# SNPs	# Blocks (Assoc)	# SNPs (Assoc)
Scenario 1	None	Strong	3	0 (0)	3 (1)
		Strong	10	0 (0)	10 (2)
		Weak	10	0 (0)	10 (1)
		Strong	30	0 (0)	30 (5)
		Weak	30	0 (0)	30 (1)
Scenario 2	Moderate	Strong	3	1 (1)	0 (0)
		Strong	10	2 (2)	2 (0)
		Weak	10	2 (1)	2 (0)
		Strong	30	5 (3)	3 (0)
		Weak	30	5 (1)	3 (0)
Scenario 3	High	Strong	3	1 (1)	0 (0)
		Strong	10	2 (2)	2 (0)
		Weak	10	2 (1)	2 (0)
		Strong	30	5 (3)	3 (0)
		Weak	30	5 (1)	3 (0)

Table 4.4: **Details of the simulation scenarios when an association is simulated.** We represent here for each scenario and for each gene size the strength of association that is the number of SNPs or blocks of SNPs associated with the disease. Note that to each scenario and gene size also corresponds a situation where no SNPs are associated in order to estimate the false-positive rate.

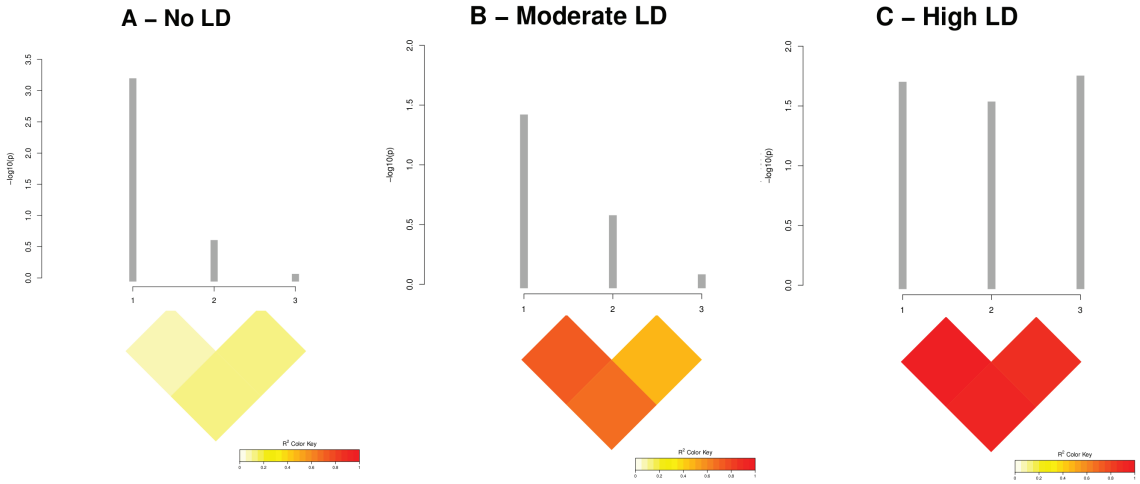

Figure 4.4: **LD and association pattern for 3 SNPs** The bottom part of the graph corresponds to the LD pattern between the markers. The top part is a plot of the association signal ($-log_{10}(p\text{-value})$) calculated using the Armitage Trend test for each marker.

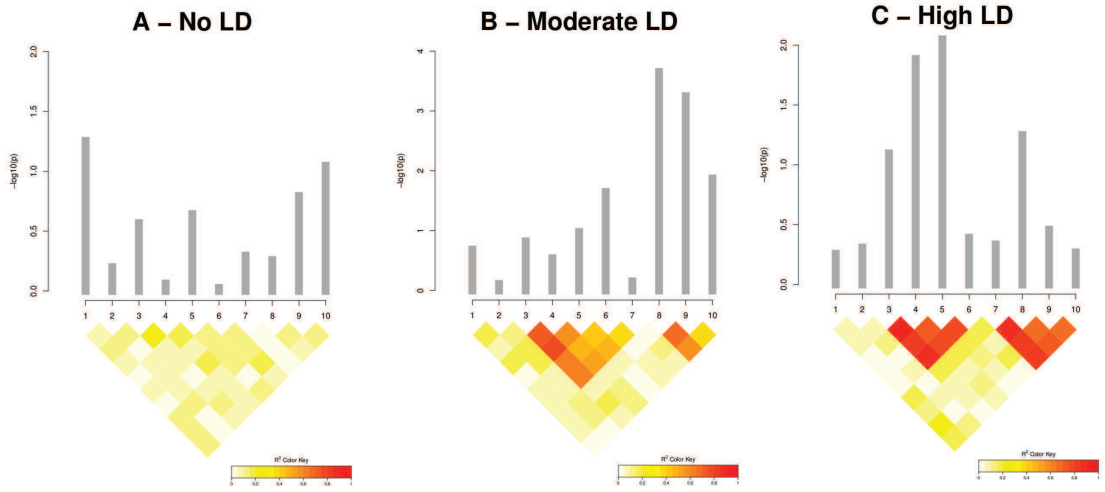

Figure 4.5: **LD and strong association pattern for 10 SNPs.** The bottom part of the graph corresponds to the LD pattern between the markers. The top part is a plot of the association signal ($-log_{10}(p\text{-value})$) calculated using the Armitage Trend test for each marker.

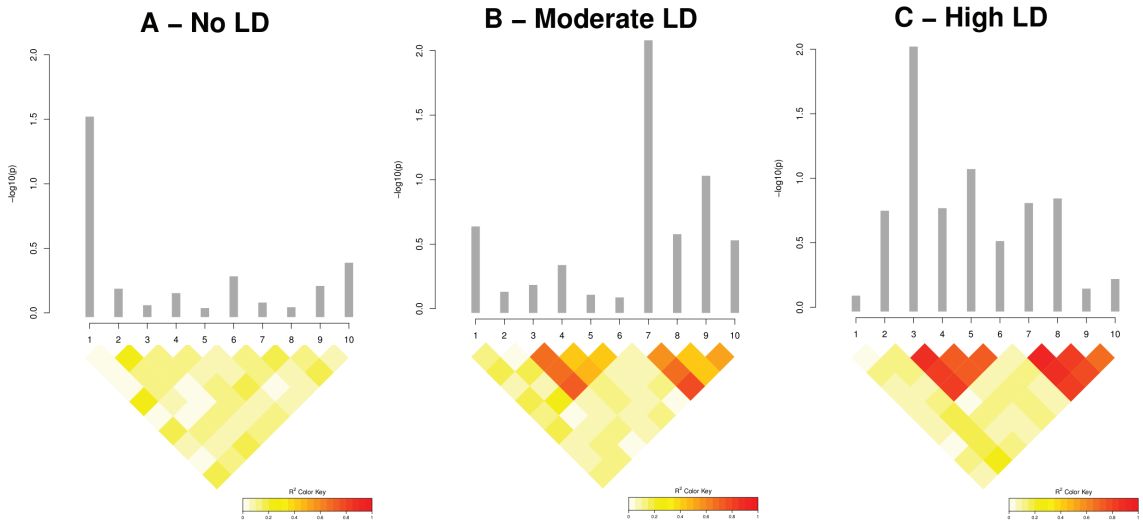

Figure 4.6: **LD and weak association pattern for10 SNPs.** The bottom part of the graph corresponds to the LD pattern between the markers. The top part is a plot of the association signal ($-log_{10}(p\text{-value})$) calculated using the Armitage Trend test for each marker.

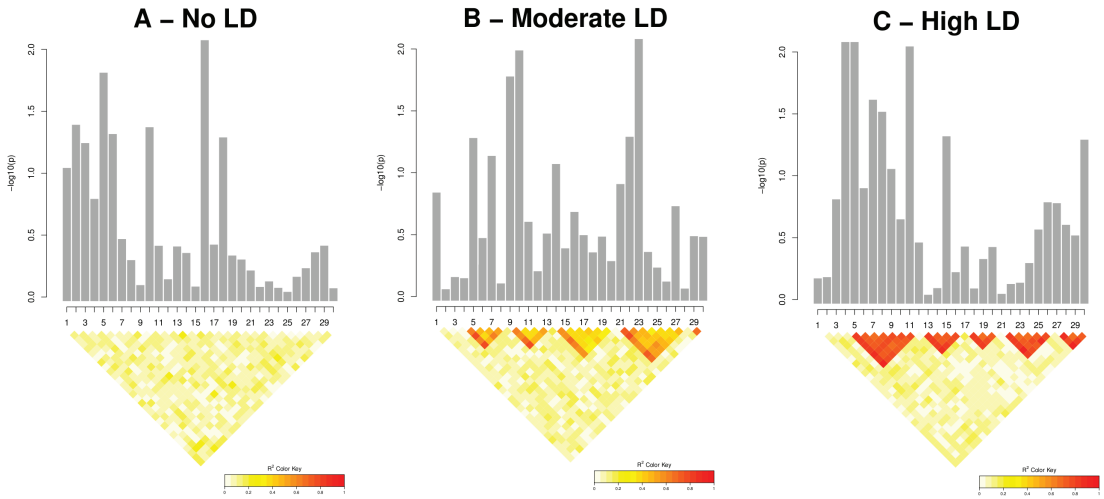

Figure 4.7: **LD and strong association pattern for 30 SNPs.** The bottom part of the graph corresponds to the LD pattern between the markers. The top part is a plot of the association signal ($-log_{10}(p\text{-value})$) calculated using the Armitage Trend test for each marker.

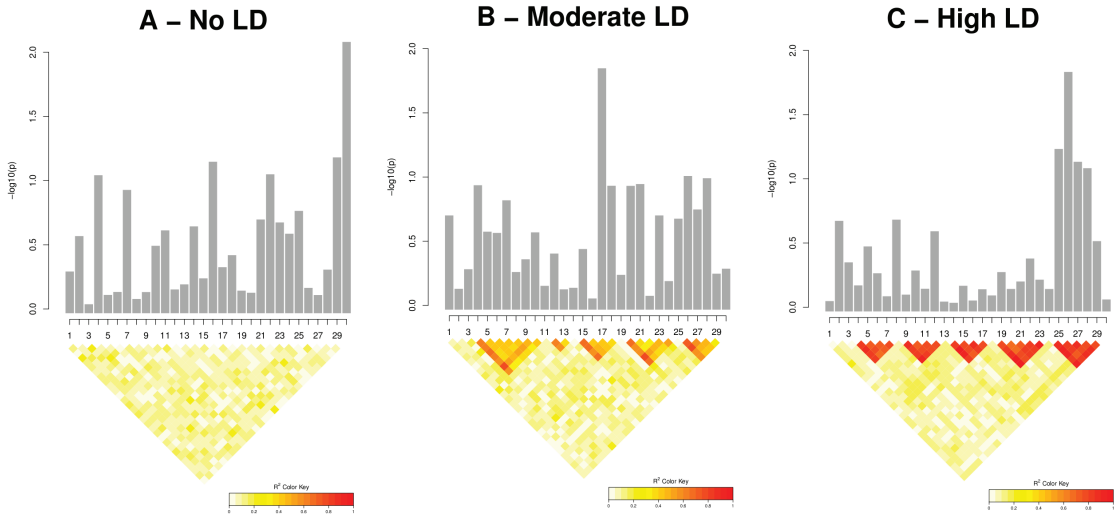

Figure 4.8: **LD and weak association pattern for 30 SNPs.** The bottom part of the graph corresponds to the LD pattern between the markers. The top part is a plot of the association signal ($-log_{10}(p$-value)) calculated using the Armitage Trend test for each marker.

Results

We present here the results of these comparisons for the different scenarios using **Figures 4.9 to 4.13** that are graphical representations of the false-positive rates and powers of the methods.

Scenario 1

In this first scenario, with no LD, one can note that the method using the minimal p-value of the SNPs (minP) had an inflated false-positive rate increasing with the number of markers (**Figure 4.9-A, C and E**). This directly highlighted the fact that it is necessary to account for the number of markers composing a gene. Otherwise, the false-positive rates of the other methods were all within the 95% confidence interval, except for the uncorrected Logistic Regression (Reg) that had a slightly inflated false-positive rate with 30 SNPs. This might have been due to the fact that Logistic models are sensitive to the number of covariates included. These results indicate that almost all the methods that account for the dependencies between the variables can be applied to sets of independent markers without having an inflated false-positive rate.

The powers of the methods were quite comparable except for the permutation strategy (Perm) that had lower power compared to the majority of the methods. Also, the Goeman test (Goeman) and the marginal score test (MargST) seemed a little more powerful than the other approaches (**Figure 4.9--B, D and F**). These results were relatively constant with the number of markers of the genes.

Scenario 2

The second scenario pertained to a moderate LD pattern between the markers. When only 3 SNPs were considered, the uncorrected meta-analysis method (Fisher) had an inflated false-positive rate and on the contrary, the corrected meta-analysis method (Fisher.corr) was conservative (**Figure 4.10-A**). These tendencies increased with the number of markers. The other methods had correct false-positive rates whatever the number of SNPs considered, except for Reg that was slightly anti-conservative. With 3 SNPs, the analysis of the powers show that the multiple-testing corrections adjusted for the dependencies between the markers (FDR.corr and Sidak.corr) along with Goeman and MargST were more powerful than the other approaches. This result was not replicated with 10 and 30 SNPs for the multiple-testing corrections. This might have been due to their false-positive rates that were reaching the upper bound of the 95% confidence interval. One can also note that Fisher.corr had a relatively high power even though this method was conservative when 30 SNPs were considered (**Figure 4.10-E and F**).

Figure 4.11 compares the powers of the methods for 10 and 30 SNPs between strong association and weak association. We can logically observe that the powers of all the methods were more important for strong associations. Considering only the methods with

correct false-positive rates, the most powerful were Goeman and MargST. Also when there were 30 SNPs, the relative increase of power of these methods was the most important. Even though the permutation method is supposed to estimate the exact distribution of the test statistic under H_0, the method was not the more powerful.

Scenario 3

With a high LD pattern, the behaviors of the methods were quite constant for all numbers of SNPs. FDR, Sidak and Fisher.corr were conservative while Fisher, FDR.cor and Sidak.corr had an inflated false-positive rate (**Figure 4.12**). Also, like in the previous scenarios, Reg had an inflated false-positive rate for 30 SNPs genes. Again, Goeman and MargST were the more powerful methods.

When comparing the powers between weak and strong association we observed that minP and Fisher had the highest power due to their very inflated false-positive rates (**Figure 4.13**). If we focus on the methods with correct false-positive rates then Goeman and MargST were once again the most powerful methods to detect gene-disease associations.

4.2.4 Discussion

We have conducted a comparison of several classically used approaches to detect gene-disease associations. Based on simulated genes with various gene sizes, linkage disequilibrium and association patterns, we analyzed the false-positive rates and powers of the methods. Several behaviors of the methods can be highlighted and seem to be emphasized when the complexity of the gene structure increases.

The first result that we highlighted is that the usual method consisting in considering the minimal p-value of the SNPs composing a gene has a highly inflated false-positive rate that increases with the number of SNPs and the LD pattern. This confirms results that have already been discovered in previous study relating the weakness of the most classical approach and highlights the importance of accounting for both the number of SNPs and the LD pattern (Hong et al. 2009).

The original Fisher method is conservative when there is a LD structure in the genes. On the other hand, after correction for dependencies between the markers this method becomes anti-conservative. This may imply that the correction factor used to correct the χ^2 distribution is too high and that the estimation of this factor could somehow be improved. On can also note that the power of Fisher.corr is quite important while this method has a low false-positive rate. It is possible that an adjustment of the false-positive rate of this method could lead to an improved approach with correct false-positive rate and a high power.

A comparable results is observable for the multiple-testing corrections. The methods that do not adjust for the dependencies between the markers are slightly conservative when

10 or 30 SNPs are considered while the corrected versions have inflated false-positive rates. This once again highlights the difficulty to estimate the proper number of independent SNP blocks to adjust the methods.

The permutation method is naturally able to maintain a correct false-positive rate. This method has however a relatively low power compared to its computational cost. One possible explanation is that 1000 permutations were not enough to assess correct null distributions. This should however be enough as the false-positive rates were correctly controlled.

In all the scenarios the uncorrected Regression method has an inflated false-positive rate when 30 SNPs are considered. This may be due to the high number of variables included in the model that affects the parameters estimation.

Finally the performances of methods such as the Goeman test and the marginal score test (MargST) are the highest. These two methods have a correct false-positive rate and reach the highest power to detect associations. The rationale that consists in ignoring the correlation between the markers and averaging the association signals within the LD blocks appears to be well suited with the gene testing issue. One downside of these approaches might however be that they are known to have a high power when the markers are positively correlated and associated in the same direction otherwise the power may be lesser. Such structures of the genes are however bound to appear in real situations (Moskvina and Schmidt 2008, Chapman and Whittaker 2008). In addition, these methods allow a certain gain in power when a gene is associated to a disease through more than one marker or block of markers. We would therefore recommend the use of one of these two statistics to assess gene-level p-values.

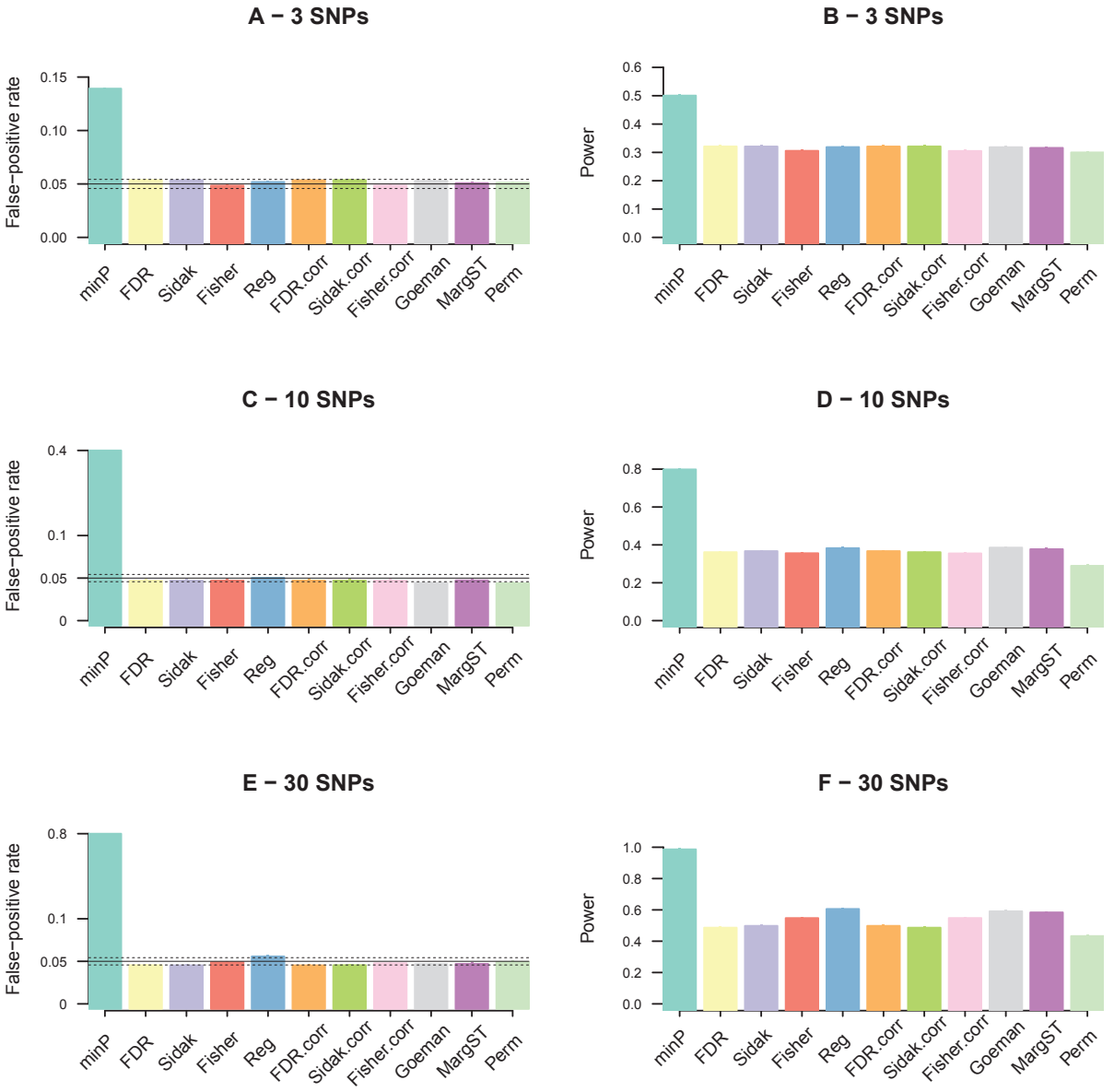

Figure 4.9: **False-positive rates and powers of the methods (Scenario 1).** This graph shows on the left panel the evolution of the false-positive rates for Scenario 1 (no LD) with 3 SNPs (A), 10 SNPs (B) and 30 SNPs (C). The plain black line represents the 5% level at which the tests were conducted. The dashed black lines are the 95% confidence intervals for this level. On the right panel is the evolution of the powers with 3 SNPs (D), 10 SNPs (E) and 30 SNPs (F). For the scenarios with 10 and 30 SNPs only the strong association are represented.

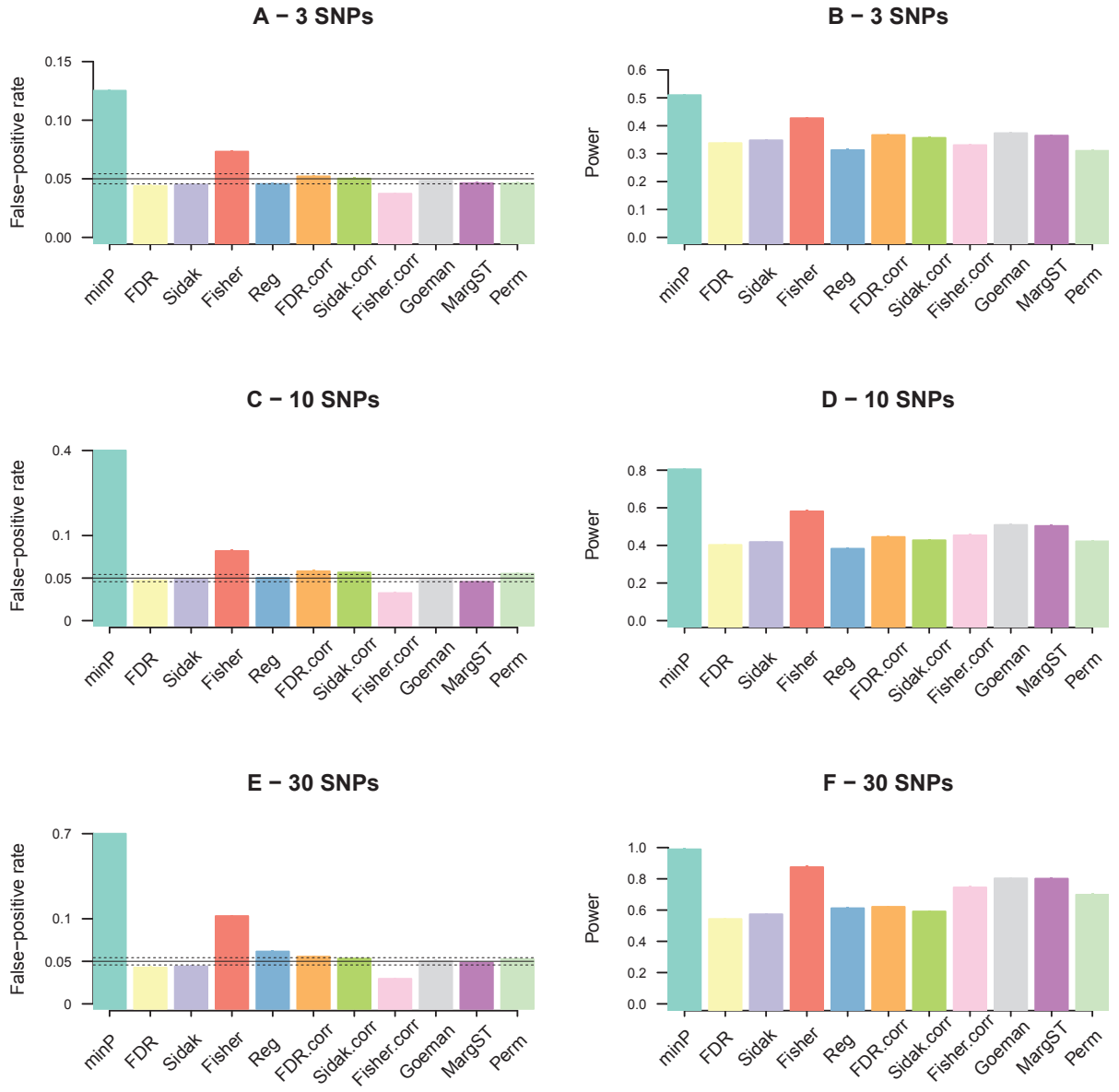

Figure 4.10: **False-positive rates and powers of the methods (Scenario 2).** This graph shows on the left panel the evolution of the false-positive rates for Scenario 2 (moderate LD) with 3 SNPs (A), 10 SNPs (B) and 30 SNPs (C). The plain black line represents the 5% level at which the tests were conducted. The dashed black lines are the 95% confidence intervals for this level. On the right panel is the evolution of the powers with 3 SNPs (D), 10 SNPs (E) and 30 SNPs (F). For the scenarios with 10 and 30 SNPs only the strong associations are represented.

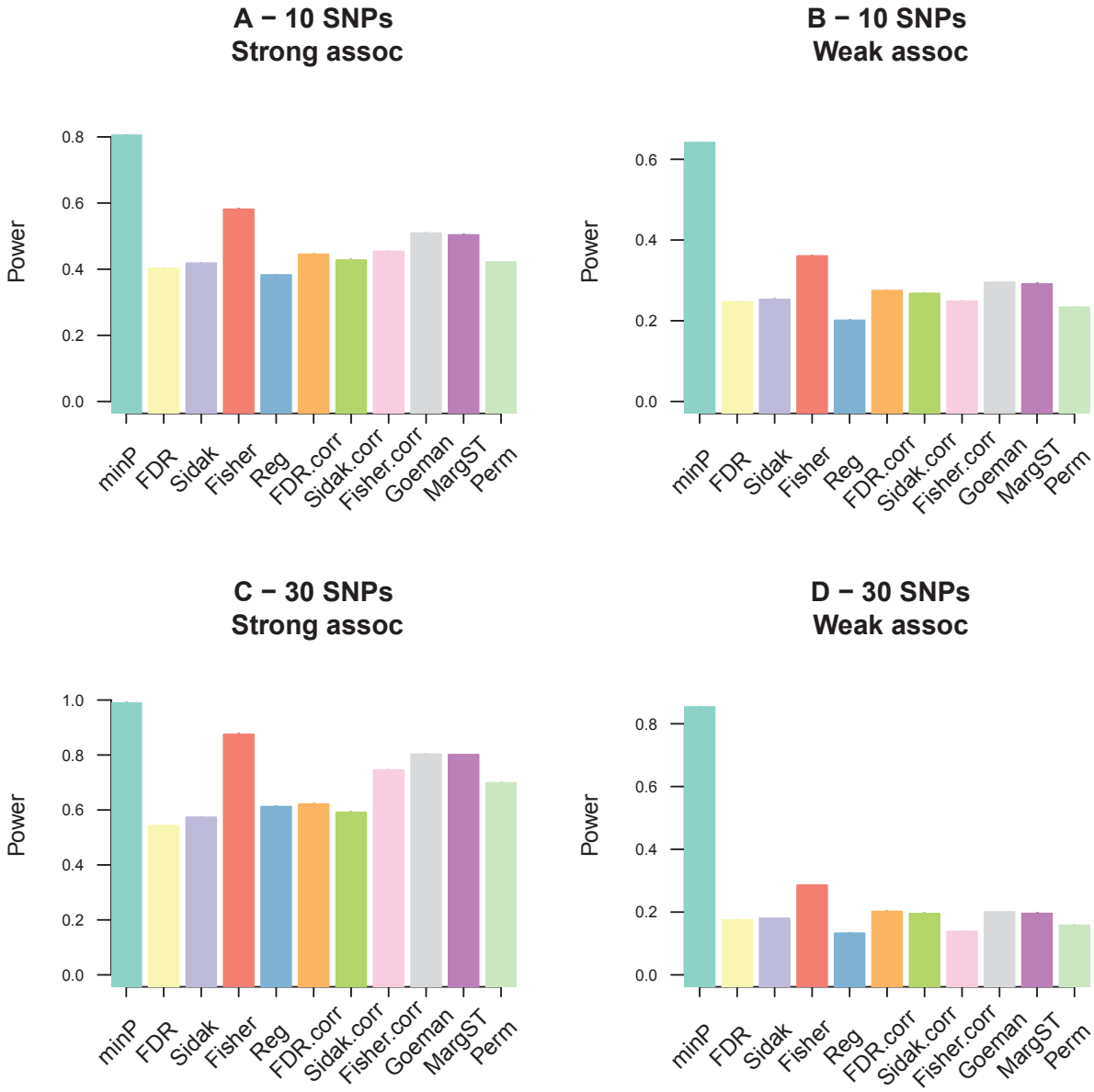

Figure 4.11: **Comparison of the powers between strong and weak association patterns (Scenario 2).** This graph shows on the left panel the evolution of the powers for the strong associations with 10 SNPs (A) and 30 SNPs (B). On the right panel are the powers for the weak associations with 10 SNPs (C) and 30 SNPs (D).

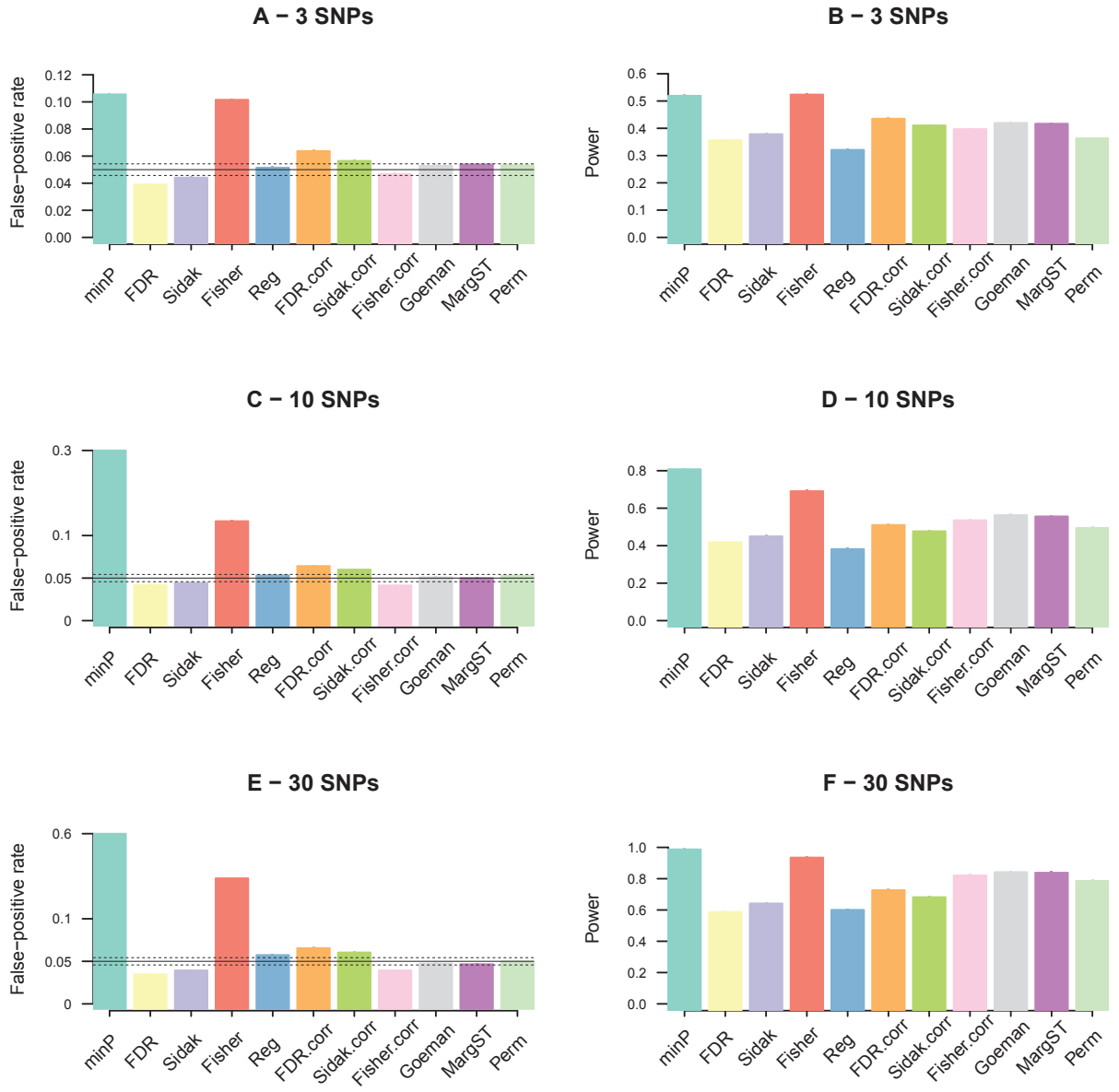

Figure 4.12: **False-positive rates and powers of the methods (Scenario 3).** This graph shows on the left panel the evolution of the false-positive rates for Scenario 3 (high LD) with 3 SNPs (A), 10 SNPs (B) and 30 SNPs (C). The plain black line represents the 5% level at which the tests were conducted. The dashed black lines are the 95% confidence intervals for this level. On the right panel is the evolution of the powers with 3 SNPs (D), 10 SNPs (E) and 30 SNPs (F). For the scenarios with 10 and 30 SNPs only the strong associations are represented.

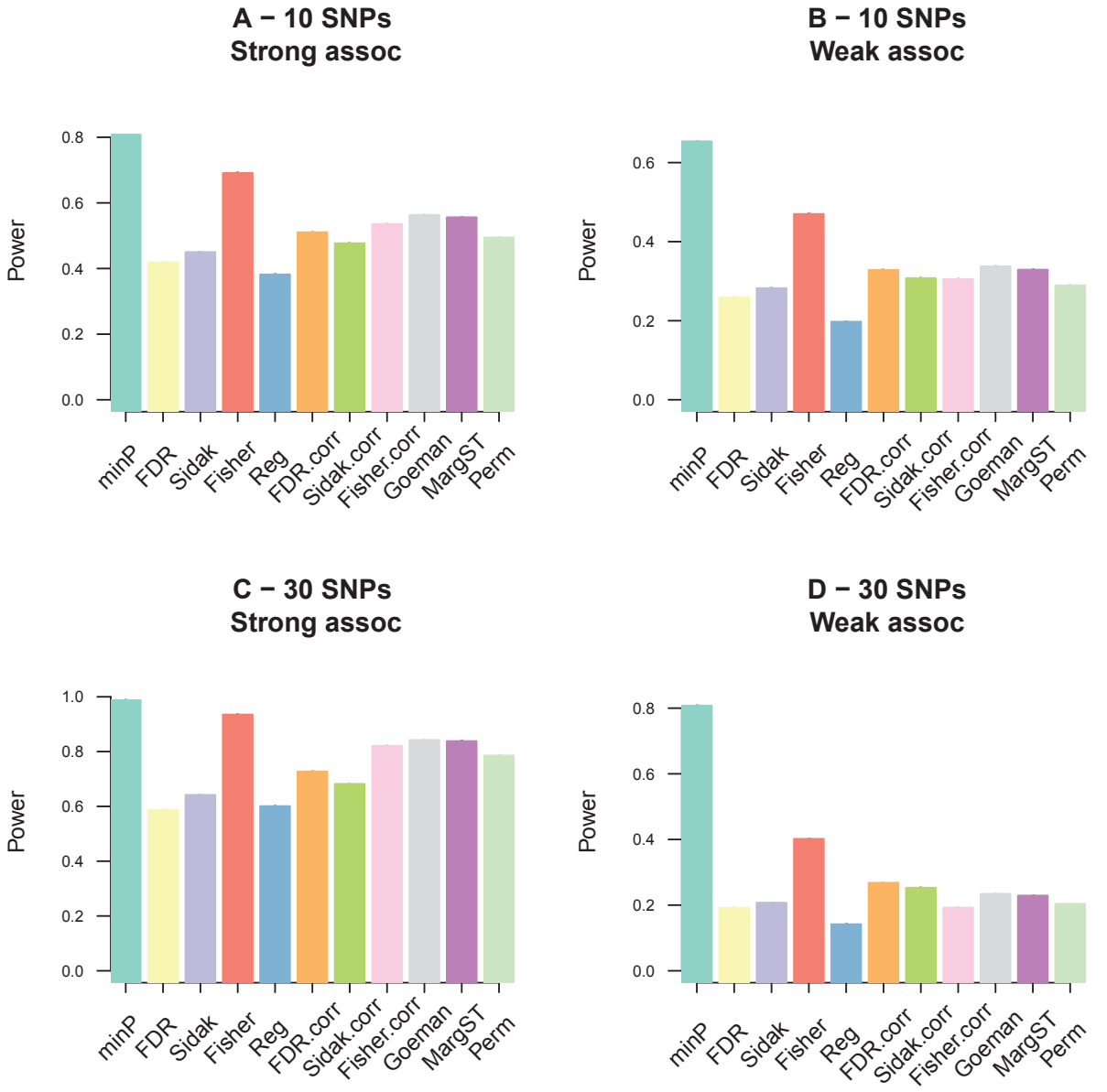

Figure 4.13: **Comparison of the powers between strong and weak association patterns (Scenario 3).** This graph shows on the left panel the evolution of the powers for the strong associations with 10 SNPs (A) and 30 SNPs (B). On the right panel are the powers for the weak associations with 10 SNPs (C) and 30 SNPs (D).

Chapter 5

Conclusions

Recent advances in Molecular Biology and improvements in micro-array and sequencing technologies have led biologists toward high-throughput genetic studies. In particular, Genome-Wide association studies have become widely used to identity markers and genes associated with diseases. As a matter of fact, the conduct of these studies represents a milestone step in the process leading to understanding the mechanisms of diseases and the development of therapeutic solutions. This PhD work focused on these studies and the related problematics in order to answer to practical interrogations of a pharmaceutical company such as **Pharnext** and of the **Statistics and Genome** laboratory. We aimed to provide indications and guidelines that answer to the questions raised by the treatment and the analysis of complex genetic data and to develop methods that allow to improve certain aspects of the genetic research.

This final chapter is dedicated to presenting the main points that we evoked in this manuscript along with the conclusions of our work regarding the different aspects of the GWASs that we analyzed and the methodological developments that we conducted. This chapter is also an opportunity to indicate further perspectives of our work.

5.1 General conclusions

Population stratification has been shown to bias the results of Genome-Wide association studies leading to false-positive or false-negative findings. Many strategies have been designed to account for stratification and it is not always obvious in practice which one should be used. We therefore conducted analyses to elucidate this point by focusing on the main approaches that are the Genomic control, methods based on Principal Components Analysis, Regression models and Meta-analyses.

We compared these different strategies on a set of simulated datasets corresponding to various stratification scenarios. These analyses allowed us to raise several conclusions.

First we determined that when there is no stratification, none of the methods that we considered induced a bias and that their respective powers were preserved. This highlights the fact that it is therefore advised to always correct for stratification. We also determined that admixture structures are more tricky to take into account than discrete structures. Finally, we showed that the Regression method adjusted on the axes of variation or on accurate population labels is the best solution to derive association measure while taking the structure of the populations into account.

This latter result highlighted the importance of powerful clustering algorithms to infer population structure. We therefore studied the issues encountered in the inference of population structure and designed a novel clustering algorithm. One of the major difficulties to cluster genetic data is the dimension of the data. Indeed, one often faces a number of variables, the markers, that is far superior to the number of individuals.

The method that we developed during this PhD is called Spectral Hierarchical clustering for the Inference of Population Structure (`SHIPS`). It is a non-parametric approach that uses a unique pairwise similarity matrix in order to reduce the dimension of the data. Our algorithm is based on the idea that it might not be possible to uncover all of the structure in the data when applying a clustering algorithm just once. Fine population structures may not be detected as the corresponding sub-populations are hidden within the major sub-populations detected by the first run of the algorithm. We therefore implemented a robust statistical framework to iteratively apply a spectral clustering algorithm to the data via a divisive hierarchical clustering strategy and so analyze in depth the genetic patterns of the studied populations. We compared our algorithm to some of the most used in practice and determined that it reaches very satisfying performances both in terms of individuals assignments and estimation of the optimal number of clusters. These results are promising for the future use of `SHIPS`.

A final point we were interested in is the issue of multiple-testing. In GWASs, many statistical tests are conducted and as a consequence a certain amount of false-positive findings arise due to the multiple-testing issues. The use of an appropriate correction is therefore necessary to ensure the validity of the results. We reviewed the different classical approaches that are the Family-Wise Error Rate, the False-Discovery Rate and the local-FDR. A question that naturally arose was which method one should use? We explained that the choice of a correction approach actually depends on the cost associated with false-positive and false-negative findings as well as the type of follow-up experiments that might be conducted on the association results.

Another problematic that was raised in the analysis of Genome-Wide association studies data is how to obtain gene-wise interpretation of the findings, i.e how to obtain a measure of significance for a gene. Deriving such a measure implies combining the information of all the markers composing a gene which leads to a multiple-testing situation with dependent tests as these markers are bound to be in LD. We presented the classical

approach that consists in using the most significant p-value of the markers. Deriving gene-level p-value with this method tends to favor large genes and the permutation procedure that can be used to correct for this bias is rather time consuming. This point highlighted the practical dilemma of finding a test that properly accounts for the structure of the gene and do not necessitate permutations. This has led to the development of many other approaches that we presented and compared through several simulations. Our study pointed out the performances of Regression based methods such as the Goeman test or the marginal score test.

Eventually, the different results that we obtained allowed to draw some interesting conclusions regarding certain aspects of the conduct of Genome-Wide association studies that can be added to the many research and discussions on this topic. In addition, we designed a novel clustering algorithm that showed promising results and is susceptible to be further used to cluster genetic data.

5.2 Perspectives

The research work that we presented in this PhD has led to several conclusions and also pointed out some interesting research perspectives.

Genetic interactions

We mainly focused on the association between a single marker and a disease. However, as we presented in the Introduction, it is also possible to look at the combined effect of several markers, that is called epistasis. In particular, the interaction between two markers is a topic of interest in GWASs. We noted that the analysis of these interactions raises several interrogations. First the definition of an interaction is relatively vague and as a consequence quite different strategies are available to assess them. For instance the main approaches correspond to Logistic or Linear Regression models, methods based on the differential correlation between the markers or uses the large interaction $\times$phenotype contingency table. A complexity in testing interactions is that the amount of test that has to be performed is considerable. If a dataset is composed of p markers then there are $\frac{p(p-1)}{2}$ interactions that need to be investigated. The development of fast statistical procedures is therefore indispensable to allow a complete screening of the genome. A second issue in the practice of the analysis of epistasis is that population stratification is rarely accounted for. With the Regression models it could seem natural to add covariates like for the single marker association test however for the differential correlation tests and the other models that exist this option is no longer possible. Certain approaches have been proposed but have yet to be thoroughly analyzed. We have designed several methods to test the interactions based for instance on the adjusted phenotypes proposed by Price et al. (see **Section 2.3.4**) in order to account for stratification. Due to the very long

computational time demanded by the simulation procedures to assess the methods, some work is still needed to assess these strategies.

Another point concerning the interactions that we think might be worth investigating concerns the gene-gene interactions. We presented how to derive a gene-level p-value in this manuscript and, following the same rationale that highlighted the interest of this work, it could also be useful to be able to derive gene$\times$gene-level p-values. A first work has been proposed in (Li and Ji 2005) and uses multiple-testing corrections for dependent variables such as those introduced in **Chapter 4**. In the case of interactions, the estimation of an effective number of independent tests is however more complicated. They proposed using the mutual information to replace the correlation of the initial procedure which is a lot more time consuming and therefore not so likely to be used in practice. We feel that there is a need to develop innovative methodologies to analyze the gene$\times$gene interactions and this work is well suited as a continuity of the one we conducted during this PhD.

Rare variants analysis

With the development of new generation sequencing, biologists are looking beyond common genetic variations and are focusing on rare variants that can contribute substantially to common diseases. As a matter of fact, associations highlighted by Genome-Wide associations studies can account for only a small fraction of the heritability of diseases. The analysis of rare variants is thought to be a solution to uncover more about diseases mechanisms.

A first problematic that arises with rare variants is finding out their associations with diseases. Indeed the low frequency of these variants renders the classical association tests not powerful to detect associations. We showed in **Chapter 2** that the power of the usual statistical techniques are very diminished when applied to rare variants. As a consequence novel association tests have to be designed and the issue of population stratification may have to be taken into account in a possibly different way.

Also, an interesting topic of research concerns the contribution of rare variants to the analysis of population structure. It has been shown that the information about the structure of the populations contained in rare variants is not the same as that contained in common genetic variation (Baye et al. 2011). Further analyses can be considered to assess more precisely the information hidden within rare and common variants and how these types of markers should be used to provide the best inference possible.

Going further with the SHIPS algorithm

The algorithm that we developed has shown promising performances. We think that this algorithm could be enhanced by continuing working on certain aspects.

First, the similarity matrix that we considered is quite simple and does not consider all the information that is available in a genetic dataset. As it is somehow at the basis of the clustering results, developing a more complete matrix could increase the accuracy of the method.

The `SHIPS` algorithm produces in one run several possible clusterings of the data for various numbers of clusters. This property allows the use of many methods to estimate the number of clusters. We considered the gap statistic as it provided the best empirical performances. One downside of this method is its computational time that is the longer part of the `SHIPS` algorithm. Investigating other methods to estimate the number of clusters could lead to finding a solution that would render faster the program.

One last perspective with `SHIPS` is its application to other types of data. Indeed, this method has been designed for genetic SNP data but technically corresponds to a divisive hierarchical clustering and could be suitable to cluster any kind of data. To do so a proper similarity matrix is necessary and the quality criterion used in the algorithm may have to be modified to fit data that are not composed of discrete values in $\{0, 1, 2\}$. We aim to apply the `SHIPS` algorithm to several data, such as micro-array or RNA-seq gene expression data, in order to design a more general version of the algorithm.

Contributions

Articles

- ◇ Bouaziz, M., Paccard, C. Guedj, M., and Ambroise, C. (2012). SHIPS: Spectral Hierarchical clustering for the Inference of Population Structure in Genetic Studies. PLOS ONE, 7(10):e45685.

- ◇ Bouaziz, M., Ambroise, C., and Guedj, M. (2011). Accounting for population stratification in practice: a comparison of the main strategies dedicated to Genome-Wide association studies. PLOS ONE, 6(12):e28845.

Book chapter

- ◇ Bouaziz M., Jeanmougin M., Guedj M. (2012). multiple-testing in large-scale genetic studies, In "Data Production and Analysis in Population Genomics" (Bonin A, Pompanon F eds), Methods in Molecular Biology Series, Humana Press.

Collaborations

This PhD also offered the opportunity to collaborate with researchers in the genetic field. Principally we contributed to the analyses of genetic data concerning systemic sclerosis for some studies conducted by Pr. Y. Allanore and Pr. P. Dieudé and their teams. These works led to several publications (Dieudé et al. 2011, Coustet et al. 2011 2012, Koumakis et al. 2012)

- ◇ Coustet, B., Bouaziz, M., Dieudé, P., Guedj, M., Bossini-Castillo, L., Agarwal, S., Radstake, T., Martin, J., Gourh, P., Elhai, M., Koumakis, E., Avouac, J., Ruiz, B., Mayes, M., Arnett, F., Hachulla, E., Diot, E., Cracowski, J.-L., Tiev, K., Sibilia, J., Mouthon, L., Frances, C., Amoura, Z., Carpentier, P., Cosnes, A., Meyer, O., Kahan, A., Boileau, C., Chiocchia, G., and Allanore, Y. (2012). Independent replication and metaanalysis of association studies establish tnfsf4 as a susceptibility

gene preferentially associated with the subset of anticentromere-positive patients with systemic sclerosis. J Rheumatol, 39(5):997-1003.

◇ Koumakis, E., Wipff, J., Dieudé, P., Ruiz, B., Bouaziz, M., Revillod, L., Guedj, M., Distler, J. H. W., Matucci-Cerinic, M., Humbert, M., Riemekasten, G., Airo, P., Melchers, I., Hachulla, E., Cusi, D., Wichmann, H.-E., Hunzelmann, N., Tiev, K., Caramaschi, P., Diot, E., Kowal-Bielecka, O., Cuomo, G., Walker, U., Czirjak, L., Damjanov, N., Lupoli, S., Conti, C., Muller-Nurasyid, M., Muller-Ladner, U., Riccieri, V., Cracowski, J.-L., Cozzi, F., Bournia, V. K., Vlachoyiannopoulos, P., Chiocchia, G., Boileau, C., and Allanore, Y. (2012). $Tgf\beta$ receptor gene variants in systemic sclerosis- related pulmonary arterial hypertension: european caucasian patients. results from a multicentre EUSTAR study of European Caucasian patients. Ann Rheum Dis. 71(11):1900-3.

◇ Dieudé, P., Bouaziz, M., Guedj, M., Riemekasten, G., Airo, P., Muller, M., Cusi, D., Matucci-Cerinic, M., Melchers, I., Koenig, W., Salvi, E., Wichmann, H. E., Cuomo, G., Hachulla, E., Diot, E., Hunzelmann, N., Caramaschi, P., Mouthon, L., Riccieri, V., Distler, J., Tarner, I., Avouac, J., Meyer, O., Kahan, A., Chiocchia, G., Boileau, C., and Allanore, Y. (2011). Evidence of the contribution of the x chromosome to systemic sclerosis susceptibility: association with the functional irak1 196phe/532ser haplotype. Arthritis Rheum, 63(12):3979-3987.

◇ Coustet, B., Dieudé, P., Guedj, M., Bouaziz, M., Avouac, J., Ruiz, B., Hachulla, E., Diot, E., Cracowski, J.L., Tiev, K., Sibilia, J., Mouthon, L., Frances, C., Amoura, Z., Carpentier, P., Cosnes, A., Meyer, O., Kahan, A., Boileau, C., Chiocchia, G., and Allanore, Y. (2011). C8orf13-blk is a genetic risk locus for systemic sclerosis and has additive effects with bank1: results from a large french cohort and meta-analysis. Arthritis Rheum, 63(7):2091-2096.

Seminar

◇ *European Mathematical Genetics Meeting.* Talk (2012). Gottingen, Germany.

◇ *Ecole Génomique et Modélisation* Talk (2012). Maffliers, France.

◇ *Statistical Methods for Post-Genomic Data.* Talk (2011). Paris, France.

◇ *Ecole Génomique et Modélisation* Talk (2011). Maffliers, France.

OPEN ACCESS Freely available online

PLOS ONE

SHIPS: Spectral Hierarchical Clustering for the Inference of Population Structure in Genetic Studies

Matthieu Bouaziz[1,2]*, Caroline Paccard[1], Mickael Guedj[1], Christophe Ambroise[2]

1 Department of Biostatistics, Pharnext, Paris, France, **2** Statistics and Genome Laboratory, University of Evry Val d'Essonne, UMR CNRS 8071 - USC INRA, Evry, France

Abstract

Inferring the structure of populations has many applications for genetic research. In addition to providing information for evolutionary studies, it can be used to account for the bias induced by population stratification in association studies. To this end, many algorithms have been proposed to cluster individuals into genetically homogeneous sub-populations. The parametric algorithms, such as Structure, are very popular but their underlying complexity and their high computational cost led to the development of faster parametric alternatives such as Admixture. Alternatives to these methods are the non-parametric approaches. Among this category, AWclust has proven efficient but fails to properly identify population structure for complex datasets. We present in this article a new clustering algorithm called Spectral Hierarchical clustering for the Inference of Population Structure (SHIPS), based on a divisive hierarchical clustering strategy, allowing a progressive investigation of population structure. This method takes genetic data as input to cluster individuals into homogeneous sub-populations and with the use of the gap statistic estimates the optimal number of such sub-populations. SHIPS was applied to a set of simulated discrete and admixed datasets and to real SNP datasets, that are data from the HapMap and Pan-Asian SNP consortium. The programs Structure, Admixture, AWclust and PCAclust were also investigated in a comparison study. SHIPS and the parametric approach Structure were the most accurate when applied to simulated datasets both in terms of individual assignments and estimation of the correct number of clusters. The analysis of the results on the real datasets highlighted that the clusterings of SHIPS were the more consistent with the population labels or those produced by the Admixture program. The performances of SHIPS when applied to SNP data, along with its relatively low computational cost and its ease of use make this method a promising solution to infer fine-scale genetic patterns.

Citation: Bouaziz M, Paccard C, Guedj M, Ambroise C (2012) SHIPS: Spectral Hierarchical Clustering for the Inference of Population Structure in Genetic Studies. PLoS ONE 7(10): e45685. doi:10.1371/journal.pone.0045685

Editor: Thomas Mailund, Aarhus University, Denmark

Received November 10, 2011; **Accepted** August 24, 2012; **Published** October 12, 2012

Funding: The authors have no funding or support to report.

Competing Interests: MB, CP and MG are employees of Pharnext, Paris. There are no patents, products in development or marketed products to declare. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: matthieu.x.bouaziz@gmail.com

Introduction

Population structure relates the genetic heterogeneity that exists between individuals of a population. This heterogeneity is a natural phenomenon resulting from biological and evolutionary processes such as for instance natural selection, genetic drift, populations migrations or mating processes [1]. These phenomena lead in time to sub-populations genetically differing with regard to the frequency of certain alleles. For the same reasons, disease prevalences or allele penetrances may vary between such groups. These systematic differences between sub-populations can be more or less important. The most identifiable are found between ethnic and/or geographically distant groups.

Identifying the underlying structure of populations is often of use for genetic research. It allows the study of evolutionary relationships between populations as well as learning about their demographic histories [2–6].

Such analyses are also of a great interest for population-based genetic studies such as Genome-Wide Association Studies (GWASs). Notwithstanding the widespread usage of GWASs, their findings have been criticized partly because they are vulnerable to population stratification. This corresponds to the bias induced in situations where the studied populations are genetically heterogeneous and the sampling of cases and controls is imbalanced between the various ancestries. Population stratification is known to lead to finding spurious associations or to missing genuine ones [7–11]. Inferring the structure of the populations can therefore be helpful to identify whether there is indeed a structure or to define homogeneous clusters of individuals that can later be used to correct the association test and account for stratification.

Two major strategies have been developed to infer the structure of the populations that are parametric model-based clustering and non-parametric clustering. Model-based clustering approaches make numerous assumptions on the genetic data and use statistical inference methods to assign individuals to sub-populations. Many of these parametric approaches exist such as for instance Structure [5], Admixture [12,13], BAPS [14] or FRAPPE [15]. These parametric methods are more commonly used to infer population structure. It has however been pointed out that they have some drawbacks such as the complexity of the underlying statistical models and of the assumptions that have to been made on the data. Also, the program Structure is known to have a very high computational cost. Non-parametric approaches have the advantage over parametric ones of making fewer assumptions on the data. For example most of these methods do not assume the Hardy-Weinberg equilibrium between genetic markers. In addition, such approaches involve few parameters to be estimated [16]. The main non-parametric methods are Awclust [17] using a

OPEN ACCESS Freely available online

PLoS one

Accounting for Population Stratification in Practice: A Comparison of the Main Strategies Dedicated to Genome-Wide Association Studies

Matthieu Bouaziz[1,2]*, Christophe Ambroise[2], Mickael Guedj[1]

1 Department of Biostatistics, Pharnext, Paris, France, **2** Statistics and Genome Laboratory, University of Evry Val d'Essonne, UMR CNRS 8071 - USC INRA, Evry, France

Abstract

Genome-Wide Association Studies are powerful tools to detect genetic variants associated with diseases. Their results have, however, been questioned, in part because of the bias induced by population stratification. This is a consequence of systematic differences in allele frequencies due to the difference in sample ancestries that can lead to both false positive or false negative findings. Many strategies are available to account for stratification but their performances differ, for instance according to the type of population structure, the disease susceptibility locus minor allele frequency, the degree of sampling imbalanced, or the sample size. We focus on the type of population structure and propose a comparison of the most commonly used methods to deal with stratification that are the Genomic Control, Principal Component based methods such as implemented in Eigenstrat, adjusted Regressions and Meta-Analyses strategies. Our assessment of the methods is based on a large simulation study, involving several scenarios corresponding to many types of population structures. We focused on both false positive rate and power to determine which methods perform the best. Our analysis showed that if there is no population structure, none of the tests led to a bias nor decreased the power except for the Meta-Analyses. When the population is stratified, adjusted Logistic Regressions and Eigenstrat are the best solutions to account for stratification even though only the Logistic Regressions are able to constantly maintain correct false positive rates. This study provides more details about these methods. Their advantages and limitations in different stratification scenarios are highlighted in order to propose practical guidelines to account for population stratification in Genome-Wide Association Studies.

Citation: Bouaziz M, Ambroise C, Guedj M (2011) Accounting for Population Stratification in Practice: A Comparison of the Main Strategies Dedicated to Genome-Wide Association Studies. PLoS ONE 6(12): e28845. doi:10.1371/journal.pone.0028845

Editor: Thomas Mailund, Aarhus University, Denmark

Received July 21, 2011; **Accepted** November 16, 2011; **Published** December 21, 2011

Funding: This study was funded by Pharnext SA, Paris, France and the Genome and Statistics Laboratory, Evry, France. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors declare that they received funding from Pharnext SA, Paris. This does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials.

* E-mail: matthieu.x.bouaziz@gmail.com

Introduction

Genome-wide association studies (GWAS) have become a widely used approach for gene mapping of complex diseases. With the development of high throughput genotyping technologies many markers are available to conduct these studies. The most common study design is the case-control design using unrelated individuals. The relevance of the results of such large scale genetic studies is however questioned. Indeed certain biases arise when conducting a GWAS, leading to false discoveries. As a consequence, only few associations are consistently and convincingly replicated [1]. There can be many causes to such spurious findings and non-replications [2–4]. It is broadly considered that failure to account for the bias induced by population stratification is one of them. This phenomenon occurs when the sampling has been made within non genetically homogeneous populations, i.e. there are systematic differences in allele frequencies due to ancestry and the baseline disease risk are different between the actual subpopulations. This can lead to finding spurious associations or to missing genuine ones [5–8]. Accounting for population stratification has nowadays become a necessary step in the conduct of a GWAS, especially with the development of very large studies such as the ones undertaken by international consortia. These studies indeed gather many cohorts of cases and controls, not always matched, with different ancestries.

The most used association test to detect an association is Armitage's Trend test. This test statistic follows a χ^2 distribution under the null hypothesis of no association. In case of population stratification, this distribution is inflated and the test statistic follows a non-central χ^2 distribution. Several main approaches exist to account for population stratification in GWAS: Genomic Control [9,10], Principal Component Analysis (PCA) based methods [11,12], Regression models [4,13], and Meta-Analyses. Genomic Control aims at correcting the Trend test statistic inflated null distribution by estimating an inflation factor, usually called λ, using many markers. In practice we usually consider that a λ inferior to 1.05 indicates that there is no stratification [14]. The main assumption of this method is that the inflation factor is the same for all markers. PCA-based methods use markers to define continuous axes of variation, called principal components, that reduce the data to few variables containing most of the information about the genetic variability. These axes often relate the spatial distribution of the ancestries of the samples. Using such methods, Price et al. propose an association test to account for

Chapter 13

Multiple Testing in Large-Scale Genetic Studies

Matthieu Bouaziz, Marine Jeanmougin, and Mickaël Guedj

Abstract

Recent advances in Molecular Biology and improvements in microarray and sequencing technologies have led biologists toward high-throughput genomic studies. These studies aim at finding associations between genetic markers and a phenotype and involve conducting many statistical tests on these markers. Such a wide investigation of the genome not only renders genomic studies quite attractive but also lead to a major shortcoming. That is, among the markers detected as associated with the phenotype, a nonnegligible proportion is not in reality (false-positives) and also true associations can be missed (false-negatives). A main cause of these spurious associations is due to the multiple-testing problem, inherent to conducting numerous statistical tests. Several approaches exist to work around this issue. These multiple-testing adjustments aim at defining new statistical confidence measures that are controlled to guarantee that the outcomes of the tests are pertinent.The most natural correction was introduced by Bonferroni and aims at controlling the family-wise error-rate (FWER) that is the probability of having at least one false-positive. Another approach is based on the false-discovery-rate (FDR) and considers the proportion of significant results that are expected to be false-positives. Finally, the local-FDR focuses on the actual probability for a marker of being associated or not with the phenotype. These strategies are widely used but one has to be careful about when and how to apply them. We propose in this chapter a discussion on the multiple-testing issue and on the main approaches to take it into account. We aim at providing a theoretical and intuitive definition of these concepts along with practical advises to guide researchers in choosing the more appropriate multiple-testing procedure corresponding to the purposes of their studies.

Key words: Multiple testing, Genetic, Association, Biostatistics, GWAS, Bonferroni, FWER, FDR

1. Introduction

During the last decade, advances in Molecular Biology and substantial improvements in microarray and sequencing technologies have led biologists toward high-throughput genomic studies. In particular, the simultaneous genotyping of hundreds of thousands of genetic markers such as single nucleotide polymorphisms (SNPs) on chips has become a mainstay of biological and

François Pompanon and Aurélie Bonin (eds.), *Data Production and Analysis in Population Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 888, DOI 10.1007/978-1-61779-870-2_13,

Bibliography

Alexander, D. H. and Lange, K. (2011). Enhancements to the admixture algorithm for individual ancestry estimation. *BMC Bioinformatics*, 12:246. (Not cited.)

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, 19(9):1655–1664. (Not cited.)

Allison, D. B., Gadbury, G., Heo, M., Fernandez, J., Lee, C.-K., Prolla, T. A., and Weindruch, R. A. (2002). Mixture model approach for the analysis of microarray gene expression data. *Comput. Statist. and Data Analysis*, 39:1–20. (Not cited.)

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11:375–386. (Not cited.)

Astle, W. and Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471. (Not cited.)

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7:781–791. (Not cited.)

Ballard, D. H., Cho, J., and Zhao, H. (2010). Comparisons of multi-marker association methods to detect association between a candidate region and disease. *Genet Epidemiol*, 34(3):201–212. (Not cited.)

Barnholtz-Sloan, J. S., McEvoy, B., Shriver, M. D., and Rebbeck, T. R. (2008). Ancestry estimation and correction for population stratification in molecular epidemiologic association studies. *Cancer Epidemiol Biomarkers Prev*, 17(3):471–477. (Not cited.)

Baye, T. M., He, H., Ding, L., Kurowski, B. G., Zhang, X., and Martin, L. J. (2011). Population structure analysis using rare and common functional variants. *BMC Proc*, 5 Suppl 9:S8. (Not cited.)

Beckmann, L., Thomas, D. C., Fischer, C., and Chang-Claude, J. (2005). Haplotype sharing analysis using mantel statistics. *Hum Hered*, 59(2):67–78. (Not cited.)

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerfull approach to multiple testing. *JRSSB*, 57(1):289–300. (Not cited.)

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188. (Not cited.)

Bonferroni, C. (1935). *Studi in Onore del Professore Salvatore Ortu Carboni*, chapter Il calcolo delle assicurazioni su gruppi di teste, pages 13–60. (Not cited.)

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Publicazioni del R Instituto Superiore de Scienze Economiche e Commerciali de Firenze*, 8:3–62. (Not cited.)

Bouaziz, M., Ambroise, C., and Guedj, M. (2011). Accounting for population stratification in practice: a comparison of the main strategies dedicated to genome-wide association studies. *PLoS One*, 6(12):e28845. (Not cited.)

Bouaziz, M., Jeanmougin, M., and Guedj, M. (2012a). Multiple testing in large-scale genetic studies. *Methods Mol Biol*, 888:213–233. (Not cited.)

Bouaziz, M., Paccard, C., Guedj, M., and Ambroise, C. (2012b). Ships: Spectral hierarchical clustering for the inference of population structure in genetic studies. *PLoS One*, 7(10):e45685. (Not cited.)

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R., and Cavalli-Sforza, L. L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368(6470):455–457. (Not cited.)

Broet, P., Lewin, A., Richardson, S., Dalmasso, C., and Magdelenat, H. (2004). A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics*, 20:2562–2571. (Not cited.)

Brown, M. B. (1975). A method for combining non-independent one-sided tests of significance. *Biometrics*, 31:978–992. (Not cited.)

Browning, S. R. and Browning, B. L. (2010). High-resolution detection of identity by descent in unrelated individuals. *Am J Hum Genet*, 86(4):526–539. (Not cited.)

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P., and Eerdewegh, P. V. (2005). Identifying snps predictive of phenotype using random forests. *Genet Epidemiol*, 28(2):171–182. (Not cited.)

Cavalli-Sforza, Menozzi, P. (1994). The history and geography of human genes. *Princeton University Press, Princeton, NJ*. (Not cited.)

Chapman, J. and Whittaker, J. (2008). Analysis of multiple snps in a candidate gene or region. *Genet Epidemiol*, 32(6):560–566. (Not cited.)

Chapman, J. M., Cooper, J. D., Todd, J. A., and Clayton, D. G. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered*, 56(1-3):18–31. (Not cited.)

Chen, C., Forbes, F., and Francois, O. (2006). Fastruct: model-based clustering made faster. *Molecular Ecology Notes*, 6:980?983. (Not cited.)

Chen, J. J., Roberson, P. K., and Schell, M. J. (2010). The false discovery rate: a key concept in large-scale genetic studies. *Cancer Control*, 17(1):58–62. (Not cited.)

Cheng, K. F. and Lin, W. J. (2007). Simultaneously correcting for population stratification and for genotyping error in case-control association studies. *Am J Hum Genet*, 81(4):726–743. (Not cited.)

Cochran, W. G. (1954). Some methods of strengthening the common chi-square tests. *Biometrics*, 10:417–451. (Not cited.)

Consortium, I. H. (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320. (Not cited.)

Corander, J., Waldmann, P., Marttinen, P., and Sillanpaa, M. J. (2004). Baps 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, 20(15):2363–2369. (Not cited.)

Corander, J., Waldmann, P., and Sillanpaa, M. J. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics*, 163(1):367–374. (Not cited.)

Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*, 10(6):392–404. (Not cited.)

Coustet, B., Bouaziz, M., Dieudé, P., Guedj, M., Bossini-Castillo, L., Agarwal, S., Radstake, T., Martin, J., Gourh, P., Elhai, M., Koumakis, E., Avouac, J., Ruiz, B., Mayes, M., Arnett, F., Hachulla, E., Diot, E., Cracowski, J.-L., Tiev, K., Sibilia, J., Mouthon, L., Frances, C., Amoura, Z., Carpentier, P., Cosnes, A., Meyer, O., Kahan, A., Boileau, C., Chiocchia, G., and Allanore, Y. (2012). Independent replication and metaanalysis of association studies establish tnfsf4 as a susceptibility gene preferentially associated with the subset of anticentromere-positive patients with systemic sclerosis. *J Rheumatol*, 39(5):997–1003. (Not cited.)

Coustet, B., Dieudé, P., Guedj, M., Bouaziz, M., Avouac, J., Ruiz, B., Hachulla, E., Diot, E., Cracowski, J.-L., Tiev, K., Sibilia, J., Mouthon, L., Frances, C., Amoura, Z., Carpentier, P., Cosnes, A., Meyer, O., Kahan, A., Boileau, C., Chiocchia, G., and Allanore, Y. (2011). C8orf13-blk is a genetic risk locus for systemic sclerosis and has additive effects with bank1: results from a large french cohort and meta-analysis. *Arthritis Rheum*, 63(7):2091–2096. (Not cited.)

Dadd, T., Weale, M. E., and Lewis, C. M. (2009). A critical evaluation of genomic control methods for genetic association studies. *Genet Epidemiol*, 33(4):290–298. (Not cited.)

Dalmasso, C., Génin, E., and Trégouet, D.-A. (2008). A weighted-holm procedure accounting for allele frequencies in genomewide association studies. *Genetics*, 180(1):697–702. (Not cited.)

de Leeuw, J. (1994). Block relaxation algorithms in statistics. *In Information systems and data analysis (eds. H Bock et al.)*, pages 308–325. (Not cited.)

Delaneau, O., Coulonges, C., and Zagury, J.-F. (2008). Shape-it: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics*, 9:540. (Not cited.)

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *J. R. Statist. Soc.*, 39:1–38. (Not cited.)

Deng, H. W. (2001). Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. *Genetics*, 159(3):1319–1323. (Not cited.)

Devlin, B., Bacanu, S.-A., and Roeder, K. (2004). Genomic control to the extreme. *Nat Genet*, 36(11):1129–30; author reply 1131. (Not cited.)

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, 55(4):997–1004. (Not cited.)

Dieudé, P., Bouaziz, M., Guedj, M., Riemekasten, G., Airo, P., Muller, M., Cusi, D., Matucci-Cerinic, M., Melchers, I., Koenig, W., Salvi, E., Wichmann, H. E., Cuomo, G., Hachulla, E., Diot, E., Hunzelmann, N., Caramaschi, P., Mouthon, L., Riccieri, V., Distler, J., Tarner, I., Avouac, J., Meyer, O., Kahan, A., Chiocchia, G., Boileau, C., and Allanore, Y. (2011). Evidence of the contribution of the x chromosome to systemic sclerosis susceptibility: association with the functional irak1 196phe/532ser haplotype. *Arthritis Rheum*, 63(12):3979–3987. (Not cited.)

Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome biology*, 3(7). (Not cited.)

Efron, B. and Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays. *Genet Epidemiol*, 23(1):70–86. (Not cited.)

Epstein, M. P., Allen, A. S., and Satten, G. A. (2007). A simple and improved correction for population stratification in case-control studies. *Am J Hum Genet*, 80(5):921–930. (Not cited.)

Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, 12(5):921–927. (Not cited.)

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587. (Not cited.)

Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes*, 7(4):574–578. (Not cited.)

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd, Edinburgh, 11th ed.(rev.) edition. (Not cited.)

Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. A., Patterson, N., Gabriel, S. B., Topol, E. J., Smoller, J. W., Pato, C. N., Pato, M. T., Petryshen, T. L., Kolonel, L. N., Lander, E. S., Sklar, P., Henderson, B., Hirschhorn, J. N., and Altshuler, D. (2004). Assessing the impact of population stratification on genetic association studies. *Nat Genet*, 36(4):388–393. (Not cited.)

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, 33(1):1–22. (Not cited.)

Gao, X. and Starmer, J. (2007). Human population structure detection via multilocus genotype clustering. *BMC Genet*, 8:34. (Not cited.)

Gao, X. and Starmer, J. D. (2008). Awclust: point-and-click software for non-parametric population structure analysis. *BMC Bioinformatics*, 9:77. (Not cited.)

Goeman, J. J., van der Geer, S. A., and van Houwelingen, H. (2005). Testing against a high dimensional alternative. *Journal Of Royal Statistical Society*, 68:477–493. (Not cited.)

Gorroochurn, P., Hodge, S. E., Heiman, G. A., Durner, M., and Greenberg, D. A. (2007). Non-replication of association studies: "pseudo-failures" to replicate? *Genet Med*, 9(6):325–331. (Not cited.)

Guan, W., Liang, L., Boehnke, M., and Abecasis, G. R. (2009). Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. *Genet Epidemiol*, 33(6):508–517. (Not cited.)

Guedj, M., Robelin, D., Hoebeke, M., Lamarine, M., Wojcik, J., and Nuel, G. (2006a). Detecting local high-scoring segments: a first-stage approach for genome-wide association studies. *Stat Appl Genet Mol Biol*, 5:Article22. (Not cited.)

Guedj, M., Robin, S., Celisse, A., and Nuel, G. (2009). Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation. *BMC Bioinformatics*, 10:84. (Not cited.)

Guedj, M., Wojcik, J., Della-Chiesa, E., Nuel, G., and Forner, K. (2006b). A fast, unbiased and exact allelic test for case-control association studies. *Hum Hered*, 61(4):210–221. (Not cited.)

Heiman, G. A., Hodge, S. E., Gorroochurn, P., Zhang, J., and Greenberg, D. A. (2004). Effect of population stratification on case-control association studies. i. elevation in false positive rates and comparison to confounding risk ratios (a simulation study). *Hum Hered*, 58(1):30–39. (Not cited.)

Hoh, J. and Ott, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet*, 4(9):701–709. (Not cited.)

Hoh, J., Wille, A., and Ott, J. (2001). Trimming, weighting, and grouping snps in human case-control association studies. *Genome Res*, 11(12):2115–2119. (Not cited.)

Holm, S. (1979). A simple sequentially rejective multiple test procedure,. *Scandinavian Journal of Statistics*, 6:65–70. (Not cited.)

Hong, M.-G., Pawitan, Y., Magnusson, P. K. E., and Prince, J. A. (2009). Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Hum Genet*, 126(2):289–301. (Not cited.)

Huang, H., Chanda, P., Alonso, A., Bader, J. S., and Arking, D. E. (2011). Gene-based tests of association. *PLoS Genet*, 7(7):e1002177. (Not cited.)

Hubisz, M., Falush, D., Stephens, M., and Pritchard, J. (2009). Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, 9:1322–1332. (Not cited.)

Huelsenbeck, J. P. and Andolfatto, P. (2007). Inference of population structure under a dirichlet process model. *Genetics*, 175(4):1787–1802. (Not cited.)

Hyam, M. C., Hoggart, C. J., O'Reilly, P. F., Whittaker, J. C., Iorio, M. D., and Balding, D. J. (2008). Fregene: simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics*, 9:364. (Not cited.)

Intarapanich, A., Shaw, P. J., Assawamakin, A., Wangkumhang, P., Ngamphiw, C., Chaichoompu, K., Piriyapongsa, J., and Tongsima, S. (2009). Iterative pruning pca improves resolution of highly structured populations. *BMC Bioinformatics*, 10:382. (Not cited.)

Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann Stat*, 29:295–327. (Not cited.)

Jorgenson, E. and Witte, J. S. (2006). A gene-centric approach to genome-wide association studies. *Nat Rev Genet*, 7(11):885–891. (Not cited.)

Jung, S.-H. (2005). Sample size for fdr-control in microarray data analysis. *Bioinformatics*, 21(14):3097–3104. (Not cited.)

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*. (Not cited.)

Kimmel, G., Jordan, M. I., Halperin, E., Shamir, R., and Karp, R. M. (2007). A randomization test for controlling population stratification in whole-genome association studies. *Am J Hum Genet*, 81(5):895–905. (Not cited.)

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P. I., Ingason, A., Steinberg, S., Rafnar, T., Sulem, P., Mouy, M., Jonsson, F., Thorsteinsdottir, U., Gudbjartsson, D. F., Stefansson, H., and Stefansson, K. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet*, 40(9):1068–1075. (Not cited.)

Koumakis, E., Wipff, J., Dieudé, P., Ruiz, B., Bouaziz, M., Revillod, L., Guedj, M., Distler, J. H. W., Matucci-Cerinic, M., Humbert, M., Riemekasten, G., Airo, P., Melchers, I., Hachulla, E., Cusi, D., Wichmann, H.-E., Hunzelmann, N., Tiev, K., Caramaschi, P., Diot, E., Kowal-Bielecka, O., Cuomo, G., Walker, U., Czirjak, L., Damjanov, N., Lupoli, S., Conti, C., Muller-Nurasyid, M., Muller-Ladner, U., Riccieri, V., Cracowski, J.-L., Cozzi, F., Bournia, V. K., Vlachoyiannopoulos, P., Chiocchia, G., Boileau, C., and Allanore, Y. (2012). $tgf\beta$ receptor gene variants in systemic sclerosis-related pulmonary arterial hypertension: results from a multicentre eustar study of european caucasian patients. *Ann Rheum Dis*. (Not cited.)

Kraft, P., Zeggini, E., and Ioannidis, J. P. A. (2009). Replication in genome-wide association studies. *Stat Sci*, 24(4):561–573. (Not cited.)

Lai, Y. (2007). A moment-based method for estimating the proportion of true null hypotheses and its application to microarray gene expression data. *Biostatistics*, 8(4):744–755. (Not cited.)

Langaas, Metteand Lindqvist, B. H. and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to dna microarray data. *Journal of the Royal Statistical Society: Series B*, 67:555–572. (Not cited.)

Lawson, D. J. and Falush, D. (2012). Population identification using genetic data. *Annu Rev Genomics Hum Genet*. (Not cited.)

Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genet*, 8(1):e1002453. (Not cited.)

Lee, C., Abdool, A., and Huang, C.-H. (2009). Pca-based population structure inference with generic clustering algorithms. *BMC Bioinformatics*, 10 Suppl 1:S73. (Not cited.)

Lehne, B., Lewis, C. M., and Schlitt, T. (2011). From snps to genes: disease association at the gene level. *PLoS One*, 6(6):e20133. (Not cited.)

Li, C. and Li, M. (2008). Gwasimulator: a rapid whole-genome simulation program. *Bioinformatics*, 24(1):140–142. (Not cited.)

Li, J. and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)*, 95(3):221–227. (Not cited.)

Li, M., Reilly, C., and Hanson, T. (2008). A semiparametric test to detect associations between quantitative traits and candidate genes in structured populations. *Bioinformatics*, 24(20):2356–2362. (Not cited.)

Li, M., Reilly, M. P., Rader, D. J., and Wang, L.-S. (2010a). Correcting population stratification in genetic association studies using a phylogenetic approach. *Bioinformatics*, 26(6):798–806. (Not cited.)

Li, M.-X., Gui, H.-S., Kwan, J. S. H., and Sham, P. C. (2011). Gates: a rapid and powerful gene-based association test using extended simes procedure. *Am J Hum Genet*, 88(3):283–293. (Not cited.)

Li, Q. and Yu, K. (2008). Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genet Epidemiol*, 32(3):215–226. (Not cited.)

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010b). Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*, 34(8):816–834. (Not cited.)

Liang, L., Zollner, S., and Abecasis, G. R. (2007). Genome: a rapid coalescent-based whole genome simulator. *Bioinformatics*, 23(12):1565–1567. (Not cited.)

Liao, J. G., Lin, Y., Selvanayagam, Z. E., and Weichung, J. S. (2004). A mixture model for estimating the local false discovery rate in dna microarray analysis. *Bioinformatics*, 20(16):2694–2701. (Not cited.)

Lou, X.-Y., Casella, G., Littell, R. C., Yang, M. C. K., Johnson, J. A., and Wu, R. (2003). A haplotype-based algorithm for multilocus linkage disequilibrium mapping of quantitative trait loci with epistasis. *Genetics*, 163(4):1533–1548. (Not cited.)

Lunetta, K. L., Hayward, L. B., Segal, J., and Eerdewegh, P. V. (2004). Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet*, 5:32. (Not cited.)

Luxburg, U. V. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 14. (Not cited.)

Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nat Genet*, 36(5):512–517. (Not cited.)

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–913. (Not cited.)

Markitsis, A. and Lai, Y. (2010). A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes. *Bioinformatics*, 26(5):640–646. (Not cited.)

McLachlan, G., Bean, R., and Ben-Tovim Jones, L. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, 22:1608–1615. (Not cited.)

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley. (Not cited.)

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genet*, 5(10):e1000686. (Not cited.)

Mohajer, M., Englmeier, K.-H., and Schmid, V. J. (2011). A comparison of gap statistic definitions with and without logarithm function. *CoRR*, abs/1103.4767. (Not cited.)

Mosig, M. O., Lipkin, E., Khutoreskaya, G., Tchourzyna, E., Soller, M., and Friedmann, A. (2001). A whole genome scan for quantitative trait loci affecting milk protein percentage in israeli-holstein cattle, by means of selective milk dna pooling in a daughter design, using an adjusted false discovery rate criterion. *Genetics*, 157(4):1683–1698. (Not cited.)

Moskvina, V., O'Dushlaine, C., Purcell, S., Craddock, N., Holmans, P., and O'Donovan, M. C. (2011). Evaluation of an approximation method for assessment of overall significance of multiple-dependent tests in a genomewide association study. *Genet Epidemiol*, 35(8):861–866. (Not cited.)

Moskvina, V. and Schmidt, K. M. (2008). On multiple-testing correction in genome-wide association studies. *Genet Epidemiol*, 32(6):567–573. (Not cited.)

Mountain, J. L. and Cavalli-Sforza, L. L. (1997). Multilocus genotypes, a tree of individuals, and human evolutionary history. *Am J Hum Genet*, 61(3):705–718. (Not cited.)

Nature (1999). Freely associating. *Nat Genet*, 22(1):1–2. (Not cited.)

Neale, B. M. and Sham, P. C. (2004). The future of association studies: gene-based analysis and replication. *Am J Hum Genet*, 75(3):353–362. (Not cited.)

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5:155–176. (Not cited.)

Ng, A., Jordan, M., and Weiss, Y. (2001). On spectral clustering: analysis and an algorithm. *Advances in Neural Information processing Systems*, 14:849 – 856. (Not cited.)

Ngamphiw, C., Assawamakin, A., Xu, S., Shaw, P. J., Yang, J. O., Ghang, H., Bhak, J., Liu, E., Tongsima, S., and Consortium, H. U. G. O. P.-A. S. (2011). Pansnpdb: the pan-asian snp genotyping database. *PLoS One*, 6(6):e21451. (Not cited.)

Nielsen, D. M., Ehm, M. G., and Weir, B. S. (1998). Detecting marker-disease association by testing for hardy-weinberg disequilibrium at a marker locus. *Am J Hum Genet*, 63(5):1531–1540. (Not cited.)

Noble, W. S. (2009). How does multiple testing correction work? *Nat Biotechnol*, 27(12):1135–1137. (Not cited.)

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., and Bustamante, C. D. (2008). Genes mirror geography within europe. *Nature*, 456(7218):98–101. (Not cited.)

Page, G. P., George, V., Go, R. C., Page, P. Z., and Allison, D. B. (2003). "are we there yet?": Deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. *Am J Hum Genet*, 73(4):711–719. (Not cited.)

Pan, W. (2009). Asymptotic tests of association with multiple snps in linkage disequilibrium. *Genet Epidemiol*, 33(6):497–507. (Not cited.)

Paschou, P., Ziv, E., Burchard, E. G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M. W., and Drineas, P. (2007). Pca-correlated snps for structure identification in worldwide human populations. *PLoS Genet*, 3(9):1672–1686. (Not cited.)

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet*, 2(12):e190. (Not cited.)

Pearson, T. A. and Manolio, T. A. (2008). How to interpret a genome-wide association study. *JAMA*, 299(11):1335–1344. (Not cited.)

Peng, B. and Amos, C. I. (2010). Forward-time simulation of realistic samples for genome-wide association studies. *BMC Bioinformatics*, 11:442. (Not cited.)

Pettersson, F. H., Anderson, C. A., Clarke, G. M., Barrett, J. C., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2009). Marker selection for genetic case-control association studies. *Nat Protoc*, 4(5):743–752. (Not cited.)

Pong-Wong, R., George, A. W., Woolliams, J. A., and Haley, C. S. (2001). A simple and rapid method for calculating identity-by-descent matrices using multiple markers. *Genet Sel Evol*, 33(5):453–471. (Not cited.)

Pounds, S. and Morris, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19:1236–42. (Not cited.)

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–909. (Not cited.)

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nat Rev Genet*, 11(7):459–463. (Not cited.)

Pritchard, J. K. and Donnelly, P. (2001). Case-control studies of association in structured or admixed populations. *Theor Popul Biol*, 60(3):227–237. (Not cited.)

Pritchard, J. K. and Rosenberg, N. A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet*, 65(1):220–228. (Not cited.)

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000a). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959. (Not cited.)

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000b). Association mapping in structured populations. *Am J Hum Genet*, 67(1):170–181. (Not cited.)

Purcell, S. and Sham, P. (2004). Properties of structured association approaches to detecting population stratification. *Hum Hered*, 58(2):93–107. (Not cited.)

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850. (Not cited.)

Reeves, P. A. and Richards, C. M. (2009). Accurate inference of subtle population structure (and other genetic discontinuities) using principal coordinates. *PLoS One*, 4(1):e4269. (Not cited.)

Reich, D. E. and Goldstein, D. B. (2001). Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol*, 20(1):4–16. (Not cited.)

Rice, T. K., Schork, N. J., and Rao, D. C. (2008). Methods for handling multiple testing. *Adv Genet*, 60:293–308. (Not cited.)

Risch, N. and Teng, J. (1998). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases i. dna pooling. *Genome Res*, 8(12):1273–1288. (Not cited.)

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., and Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*, 69(1):138–147. (Not cited.)

Robin, S., Bar-Hen, A., Daudin, J.-J., and Pierre, L. (2007). A semi-parametric approach for mixture models: Application to local false discovery rate estimation. 51:5483–5493. (Not cited.)

Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics*, 53(4):1253–1261. (Not cited.)

Satten, G. A., Flanders, W. D., and Yang, Q. (2001). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet*, 68(2):466–477. (Not cited.)

Scheid, S. and Spang, R. (2004). A stochastic downhill search algorithm for estimating the local false discovery rate. *IEEE/ACM Trans Comput Biol Bioinform*, 1(3):98–108. (Not cited.)

Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet*, 3(7):e114. (Not cited.)

Setakis, E., Stirnadel, H., and Balding, D. J. (2006). Logistic regression protects against population structure in genetic association studies. *Genome Res*, 16(2):290–296. (Not cited.)

Shringarpure, S. and Xing, E. P. (2009). mstruct: inference of population structure in light of both genetic admixing and allele mutations. *Genetics*, 182(2):575–593. (Not cited.)

Sidak, Z. (1967). Rectangular confidence region for themeans of multivariate normal distributions. *Journal of the American Statistical Association*, 62:626–633. (Not cited.)

Simoni, M., Tempfer, C. B., Destenaves, B., and Fauser, B. C. J. M. (2008). Functional genetic polymorphisms and female reproductive disorders: Part 1: Polycystic ovary syndrome and ovarian response. *Hum Reprod Update*, 14(5):459–484. (Not cited.)

Slager, S. L. and Schaid, D. J. (2001). Case-control studies of genetic markers: power and sample size approximations for armitage's test for trend. *Hum Hered*, 52(3):149–153. (Not cited.)

Spielman, R. S. and Ewens, W. J. (1996). The tdt and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet*, 59(5):983–989. (Not cited.)

Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *Am J Hum Genet*, 52(3):506–516. (Not cited.)

Stephens, M. and Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*, 73(5):1162–1169. (Not cited.)

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–9445. (Not cited.)

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550. (Not cited.)

Tang, H., Peng, J., Wang, P., and Risch, N. J. (2005). Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol*, 28(4):289–301. (Not cited.)

Teng, J. and Risch, N. (1999). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. ii. individual genotyping. *Genome Res*, 9(3):234–241. (Not cited.)

Tian, C., Gregersen, P. K., and Seldin, M. F. (2008). Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet*, 17(R2):R143–R150. (Not cited.)

Tibshirani, R., Walther, G., and Hastie, T. (2000). Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society, Series B*, 63:411–423. (Not cited.)

Torkamani, A., Topol, E. J., and Schork, N. J. (2008). Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, 92(5):265–272. (Not cited.)

Tracy, C. and Widom, H. (1994). Level-spacing distributions and the airy kernel. *Commun Math Phys*, 159:151–174. (Not cited.)

Tregouet, D. A., Escolano, S., Tiret, L., Mallet, A., and Golmard, J. L. (2004). A new algorithm for haplotype-based association analysis: the stochastic-em algorithm. *Ann Hum Genet*, 68(Pt 2):165–177. (Not cited.)

Tzeng, J.-Y., Devlin, B., Wasserman, L., and Roeder, K. (2003). On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am J Hum Genet*, 72(4):891–902. (Not cited.)

Tzeng, J.-Y. and Zhang, D. (2007). Haplotype-based association analysis via variance-components score test. *Am J Hum Genet*, 81(5):927–938. (Not cited.)

van den Oord, E. J. C. G. (2008). Controlling false discoveries in genetic studies. *Am J Med Genet B Neuropsychiatr Genet*, 147B(5):637–644. (Not cited.)

Wang, D., Sun, Y., Stang, P., Berlin, J. A., Wilcox, M. A., and Li, Q. (2009). Comparison of methods for correcting population stratification in a genome-wide association study of rheumatoid arthritis: principal-component analysis versus multidimensional scaling. *BMC Proc*, 3 Suppl 7:S109. (Not cited.)

Wang, J. (2003). Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics*, 164(2):747–765. (Not cited.)

Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet*, 81(6):1278–1283. (Not cited.)

Wang, S.-J. and Chen, J. J. (2004). Sample size for identifying differentially expressed genes in microarray experiments. *J Comput Biol*, 11(4):714–726. (Not cited.)

Wang, T. and Elston, R. C. (2007). Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am J Hum Genet*, 80(2):353–360. (Not cited.)

Whitlock, M. C. (2005). Combining probability from independent tests: the weighted z-method is superior to fisher's approach. *J. EVOL. BIOL*, 18:1368?1373. (Not cited.)

Wojcik, J. and Forner, K. (2008). Exactfdr: exact computation of false discovery rate estimate in case-control association studies. *Bioinformatics*, 24(20):2407–2408. (Not cited.)

Wright, S. (1921). Systems of mating. *Genetics*, 6(2):111–178. (Not cited.)

Wu, C., DeWan, A., Hoh, J., and Wang, Z. (2011). A comparison of association methods correcting for population stratification in case-control studies. *Ann Hum Genet*, 75(3):418–427. (Not cited.)

Wu, J., Devlin, B., Ringquist, S., Trucco, M., and Roeder, K. (2010a). Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genet Epidemiol*, 34(3):275–285. (Not cited.)

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010b). Powerful snp-set analysis for case-control genome-wide association studies. *Am J Hum Genet*, 86(6):929–942. (Not cited.)

Zaykin, D. V., Westfall, P. H., Young, S. S., Karnoub, M. A., Wagner, M. J., and Ehm, M. G. (2002). Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered*, 53(2):79–91. (Not cited.)

Zhang, F., Wang, Y., and Deng, H.-W. (2008). Comparison of population-based association study methods correcting for population stratification. *PLoS One*, 3(10):e3392. (Not cited.)

Zhang, F., Zhang, L., and Deng, H.-W. (2009a). A pca-based method for ancestral informative markers selection in structured populations. *Sci China C Life Sci*, 52(10):972–976. (Not cited.)

Zhang, J., Niyogi, P., and McPeek, M. S. (2009b). Laplacian eigenfunctions learn population structure. *PLoS One*, 4(12):e7928. (Not cited.)

Zhang, Y., Xiao, X., and Wang, K. (2009c). Accommodating population stratification in case-control association analysis: a new test and its application to genome-wide study on rheumatoid arthritis. *BMC Proc*, 3 Suppl 7:S111. (Not cited.)

Zhao, H., Rebbeck, T. R., and Mitra, N. (2009). A propensity score approach to correction for bias due to population stratification using genetic and non-genetic factors. *Genet Epidemiol*, 33(8):679–690. (Not cited.)

Zhao, J., Jin, L., and Xiong, M. (2006). Test for interaction between two unlinked loci. *Am J Hum Genet*, 79(5):831–845. (Not cited.)

Zheng, G., Freidlin, B., and Gastwirth, J. L. (2006). Robust genomic control for association studies. *Am J Hum Genet*, 78(2):350–356. (Not cited.)

Zondervan, K. T. and Cardon, L. R. (2004). The complex interplay among factors that influence allelic association. *Nat Rev Genet*, 5(2):89–100. (Not cited.)