NNT : 2017SACLE040

Thèse de doctorat de l'Université Paris-Saclay préparée à l'Université d'Évry

Ecole doctorale n°577
Structure et dynamique des systèmes vivants
Spécialité de doctorat : Sciences de la vie et de la santé

par

Mme Virginie Stanislas

Approches statistiques pour la détection d'épistasie dans les études d'associations pangénomiques

Thèse présentée et soutenue à Évry, le 18 Décembre 2017.

Composition du Jury :

Mme. M-L. Martin-Magniette (DR), AgroParisTech, (Présidente du jury)
M. M. Emily (MC), Agrocampus Ouest, (Rapporteur)
Mme. N. Villa-Vialaneix (CR), INRA of Toulouse, (Rapportrice)
M. J-P. Hugot (PUPH), Université Paris Diderot-Sorbonne, (Examinateur)
M. C. Ambroise (PR), Université d'Evry Val d'Essonne, (Directeur de thèse)
M. C. Dalmasso (MC), Université d'Evry Val d'Essonne, (Directeur de thèse)

À ma famille

Merci !

Même si mon nom est en première page (en gras en plus), l'aboutissement du travail de ces trois dernières années n'aurait pas été possible sans le soutien d'un grand nombre de personnes que j'aimerais ici remercier le plus possible. Peut être à l'image de l'écriture de cette thèse, il est tout aussi difficile de se lancer dans la rédaction de remerciements à la hauteur de ce à quoi ils font référence. J'espère dans tous les cas que ces quelques mots arriveront à transmettre le plus possible ma reconnaissance et à quel point j'ai pu me sentir si bien entourée au cours de ces années.

En tout premier lieu, bien sûr, je tiens à vous remercier Christophe et Cyril pour le très bon encadrement que j'ai eu la chance de recevoir tout au long de ce parcours. Je me rappelle encore du premier échange skype lorsque j'étais encore à Rennes à la recherche d'un stage et d'y avoir déjà ressenti votre écoute et l'environnement chaleureux que cela pourrait être de venir dans votre équipe et je suis vraiment heureuse d'avoir eu cette chance. Pour quelqu'un qui a autant de doute que moi, savoir que je pouvais compter sur votre soutien à tout moment a vraiment été très important et je vous remercie de m'avoir fait confiance. Merci Christophe pour tes nombreuses idées quand je ne voyais plus où aller et d'avoir aussi su mettre des limites pour ne pas que je m'égare et que je puisse rester dans les temps. Merci aussi de m'avoir fait redécouvrir l'escalade en pleine nature le temps d'un midi avec Florent et Marie. Merci Cyril pour le temps que tu as toujours accepté de me consacrer pour répondre à chacune de mes questions et inquiétudes et ceci à n'importe quel moment où elles décidaient d'arriver dans mon esprit (et parfois il peut y en avoir beaucoup). Je suis très heureuse d'avoir été ta première thésarde et j'espère que l'expérience a été aussi positive que pour moi. Merci aussi à vous deux pour les relectures tardives et les entrainements pour les présentations orales qui me permettent d'être capable de mieux m'exprimer devant un public un peu moins effrayant à mes yeux maintenant.

Je souhaite également remercier Nathalie Villa-Vialaneix et Mathieu Emily d'avoir accepté de rapporter mes travaux de thèses et pour le temps que vous avez consacré à mon travail. Je vous remercie pour vos commentaires et vos questions qui ont permis d'améliorer à la fois le manuscrit mais également d'ouvrir de nouvelles pistes de réflexions. Merci également à Marie-Laure

Martin-Magniette d'avoir présidé ce jury de thèse et à Jean-Pierre Hugot d'avoir accepté de faire partie de ce jury et d'avoir apporté un point de vue biomédical ouvrant de nouvelles voies d'applications à ce travail. Merci à vous tous pour votre présence, vos questions et l'intérêt que vous avez porté à mes travaux.

Je souhaite aussi remercier toutes les chouettes personnes que j'ai pu rencontrer tout au long de cette thèse. Merci tout d'abord à Sarah, Florent et Marie d'avoir été d'aussi bons compagnons de bureau. Merci Sarah pour toutes nos discussions et tes conseils face à mes inquiétudes pour la thèse ou sur comment gérer une classe d'étudiants. Merci Florent pour être un compagnon de bureau si serein, je te souhaite bon courage pour la fin de ta thèse. Marie, même si ce fut juste pour quelques mois, je suis très contente d'avoir eu la possibilité de les partager avec toi. Merci aussi aux autres doctorants pour les bons moments que j'ai pu passer avec vous et notamment Morgane pour m'avoir emmenée courir de si nombreuses fois et pour être venue nous rendre visite à chacun de tes passages à l'imprimante et Benjamin pour ta bonne humeur à chacun de tes passages. Merci à Frank et Margot pour toutes les pauses que j'ai pu prendre dans votre bureau et de m'avoir laissée exprimer toutes mes compétences artistiques sur vos tableaux respectifs. Merci aussi Franck pour m'avoir laissée me moquer de toi si longtemps (en vrai tu n'es pas si vieux que ça). Mais merci surtout d'avoir toujours cherché à me faire relativiser et déstresser face aux évènements incontournables de la thèse à mes yeux insurmontables (tout ce qui a pu être relatif à parler devant d'autres êtres humains, en gros). Merci aussi d'avoir pensée à me faire stresser quand j'oubliais. Maurice, Droopy ne t'oublie pas non plus, merci pour l'aide que tu m'as apportée à chaque fois que j'en ai eu besoin et ton rire communicatif qui s'entend à l'autre bout du labo. Merci également Michèle pour ta disponibilité à chacun de nos soucis, Carène pour ta bonne humeur et ton écoute. Merci aussi Édith pour tes conseils et aussi pour ces trois semaines de colloque à Seattle, ce fut une super expérience pour moi et je n'oublie pas ce 14 juillet franco-américain pas comme les autres, merci de m'avoir permis d'y assister. Merci également Valérie pour toute ton aide lors de l'organisation de la soutenance et à plein d'autres occasions. Merci aussi Guillem et Valérie d'avoir suivi un bout de mon travail en assistant à mes comités de thèses et pour tous vos conseils. Une pensée aussi pour Bernard, qui nous a quitté trop tôt mais dont la présence reste encore ancrée parmi nous. Enfin, je n'oublie pas non plus Marie (merci pour la glacière le jour J !), Vincent, Marie-Luce, Agathe, Claudine, Nathalie, Anne-Sophie et les autres membres de l'équipe, merci pour les moments que j'ai pu partager avec vous et merci de faire de Statistique et Genome un environnement si agréable à vivre.

Et après tout ce monde, un grand merci à mes amis pour simplement être là quand il le faut. Merci pour tous les moments que vous m'avez apportés au cours de ces années qui m'ont permis de retrouver mes forces quand j'en avais

besoin ou pour simplement ces moments de parenthèse pour profiter de la vie et du bonheur d'être ensemble. Je remercie donc mes colocataires Mathias et Lucie devenus mêmes mes propriétaires pour leur ténacité à vouloir me conserver sous le même toit qu'eux. Merci aussi pour avoir écouté d'une oreille bienveillante tous mes désarrois de doctorante. Mais ne vous inquiétez pas j'en trouverais de nouveaux très bientôt (ainsi que des bonnes nouvelles ne vous inquiétez pas). Merci aussi Mathias d'avoir veillé sur moi depuis maintenant plus de 5 ans (on ne peut plus compter les moisiversaire), merci pour cette amitié qui m'a permis de devenir quelqu'un de mieux (ou de moins pire à toi de voir). Merci à ma profiteuse préférée Eva, pour m'écouter me plaindre et pour pouvoir radoter ensemble les mêmes histoires indéfiniment. Ne plus vivre sous le même toit que toi me manque mais j'espère qu'on pourra trouver de nouvelles excuses pour venir habiter l'une chez l'autre. Et finalement c'était plutôt un bon choix de te laisser partir à Düsseldorf je dirais :) merci d'avoir choisi cet endroit et de m'y avoir invitée (ou de m'y avoir laissée m'incruster) si souvent. Merci aussi à tous les ensaiens restés à Paris, Camille pour tout ces bons moments avec toi, Abdel pour simplement t'entendre dire "c'est pas grave" et faire relativiser n'importe quelle situation et pour ta si grande générosité. Merci Thomas (tu m'as dit que ça suffirait d'écrire juste ça dans mes remerciements, bon j'avoue j'ai égayé un peu le contour, mais je respecte ton souhait). Merci aussi Yolaine et Thomas de rester mes parrain et marraine même 3 ans après la sortie de l'école et d'être toujours autant à mes soins. Merci aussi Cécile de m'avoir fait venir à l'aquaboulevard de si nombreuses fois. Merci Matthieu pour ces plats que tu nous as cuisinés et pour ce nouvel an à Etel. Et bien sûr, Alex pour nous faire rire aussi souvent et rendre n'importe laquelle de nos soirées aussi mémorables. Juliette, merci pour ces soirées Big Bang Theory et de t'être forcée à sortir de ton lit pour venir à ma soutenance de thèse sans même pouvoir profiter du pot. Merci aussi Lucie d'avoir fait le déplacement ce jour là et pour ton amitié depuis toutes ces années. Steph et Louise, merci pour vos encouragements et de pouvoir vous retrouver sans que rien ne change malgrès le temps qui passe. Vous allez tous me manquer mais je reviendrais vous rendre visite à Paris c'est promis. Merci aussi à mes amis éparpillés un peu partout, Marie, Lucile, Tachou et Elo pour nos retrouvailles toujours dans des lieux différents depuis le bac et qui ne sont pas prêtes de s'arrêter. Lamia, Marion et Amélie que je suis toujours heureuse de retrouver lors de mes visites dans le sud. Charlotte et Lionel pour les visites en Anjou et félicitation Lionel d'avoir réussi à venir m'écouter à temps malgré t'être retrouvé enfermé le matin même. Amandine et ta petite Alice qui n'arrête pas de grandir pour simplement faire une pause ensemble dans nos vies superactives (enfin plutôt ta vie, tu es imbattable sur ce point là) avec un verre de mojito à la main, bien sûr. Merci en tout cas, d'être devenue une personne si incontournable dans ma vie maintenant. Merci aussi à mes amis Erasmus pour le changement d'air qu'ils me font prendre à chaque occasion que j'ai de les revoir, j'y retrouve toujours une petite partie de moi qui ne doit pas appartenir à la France et qui

me remet immédiatement sur pied. Merci de tout cœur Monika, Vera, Ihsan, Maryse, Julia, Lena, Jonas et Annalisa. J'espère, mais je n'en doute pas vraiment, avoir l'occasion de vous revoir très bientôt.

Enfin, je tiens bien évidement à remercier ma famille pour son soutien sans faille. Merci d'être venus si nombreux le jour de la soutenance et pour tous vos messages d'encouragements. Un mot tout d'abord à mes grands parents pour être si présents dans ma vie et pour tous ces moments où j'ai pu me ressourcer en venant vous voir en Anjou. Je sais que vous me soutenez dans tout ce que j'entreprends et sachez que vous faites partie du voyage à chaque fois. Un grand merci aussi pour ma maman et tout ce qu'elle nous donne pour nous voir réussir dans nos vies. Je sais que tu ne t'attendais pas à me voir faire des études pendant de si longues années mais c'est en fait surement grâce au goût de la curiosité que tu as su éveiller en moi dès mon plus jeune âge. Merci pour tout ce que tu m'as apporté bien que ces mots ne seront jamais assez suffisants pour tout exprimer. Merci aussi Bernard pour m'avoir fait une place aussi importante dans ta vie, je ne l'ai pas forcément remarqué lorsque j'étais plus jeune et je n'ai pas forcément eu l'occasion de le dire mais cela a bien sûr une grande importance pour moi. Enfin merci à mes frères, Raphaël et Pierre pour toute la joie que vous m'avez toujours apportée, pour le bonheur que j'ai d'être votre grande sœur et pour simplement être là. Vous n'êtes bien sûr plus des enfants mais vous resterez toujours mes petits schtroumpfs (oui Pierre c'est un compliment).

Et enfin pour finir sur une note plus légère, je ne remercie pas Mark Zuckerberg pour le temps qu'il m'a fait perdre.

Contents

General introduction

Many human diseases have a genetic component. Monogenic disorders are the result of a single defective gene. They can be inherited or caused by a spontaneous mutation with no previous family history. Those pathologies are however rare compared with multifactorial diseases for which the effects of multiple genes are coupled with lifestyle and environmental factors. For those disorders, known as complex diseases, it is particularly challenging to establish the right etiology because it is difficult to determinate the real impact of both genetic and environmental factors. A first issue is that the influence of genetic versus environmental factors may strongly fluctuate according to the studied disease. For example, among autoimmune and chronic inflammatory disorders, psoriasis is under strong genetic influences while multiple sclerosis presents weak genetic but strong environmental pressure [Fiocchi, 2009]. Another issue is that the degree of contribution of any given gene can be extremely variable and can also depend on the presence of some specific other genes. In addition, those genes or gene combinations will not only determine the simple apparition of a particular disease but also the degree of its severity, its duration and its response to possible therapies.

The exploration of both genetic and environmental influences is the subject of numerous research and investigations. In the early 2000s, development and costs reduction of genotyping technologies have made possible the emergence of new types of genetic investigations based on the whole genome as Genome Wide Association Studies (GWAS). The main objective of GWAS is to identify relevant genes that are related to a particular phenotypic condition by the detection of genetic markers (SNPs) among the whole genome. If GWAS have shown success to identify a large number of associated markers for a plurality of diseases, they have however only been able to explain a small part of the phenotypic variations expected to result from genetic influences. This difference is known as the missing heritability. Limitations of GWAS may be explained by the various challenges that arise from the analysis of high dimensional data sets like the difficulty to take into account the global genetic architecture of the phenotypic trait considered. Indeed, GWAS are traditionally based on a univariate approach in which each genetic marker is individually tested for a possible association. If this type of analyses is computationally sustainable for a whole genome investigation, it fails to consider the possible links between

genetic markers and thus fails to correctly model the global genetic architecture. Interaction effects between genetic markers, known as epistasis, are thus suspected to play a relevant role in the regulation of phenotypic conditions and may explain a part of the missing heritability induced by GWAS [Zuk et al., 2012]. However, investigation of epistatic effects in GWAS is particularly challenging due not only to methodological and computational difficulties but also to the fact that the definition of epistasis may refer to diverse biological and statistical concepts. Inquiry of epistasis is currently a large field of investigation that requires the development of new modeling approaches able to take into account and identify those type of effects.

The main objective of this work is to propose a new statistical approach dedicated to the detection of epistasis in GWAS. More precisely we will focus on gene based epistasis which consists in considering interactions between groups of markers belonging to the same genes rather than between genetic markers themselves.

This manuscript is composed of five different sections. The first section will present some genetic notions and will introduce Genome Wide Association Studies with their limits and their remaining challenges. In a second section, we will present the statistical methodologies used throughout the manuscript as well as interaction detection approaches that have been proposed in the literature in the past years. Our new methodological approach will be present in a third section. The fourth section will present the results of various simulation studies aiming to evaluate the performance of our proposed approach. We will finish with a fifth section that will focus on the application of our approach to detect gene gene interactions in real data sets regarding several different pathologies.

Chapter 1

Genetic context of Genome Wide Association Studies

This first chapter aims to introduce the basic notions of genetics that will be used in this manuscript and the principles and limitations of Genome Wide Association Studies.

Contents

1.1 Genetic concepts

1.1.1 Genome: definition and role

What is the genome?

The genome is the complete set of genetic instructions of an organism. Each individual possesses a unique genome allowing to define his singularity among human beings. The set of genetic instructions that define the genome is stored as a code in the DNA which is present in all the cells of the organism. There are many different types of cells such as nerve cells or hair cells, with different shapes and forms but they all have the same basic parts: a membrane that defines the outer border of the cell and a cytoplasm which is a liquid material inside the cell that contains the nucleus. Genetic information is stored inside the nucleus in the form of chromosomes (cf Figure 1.1).

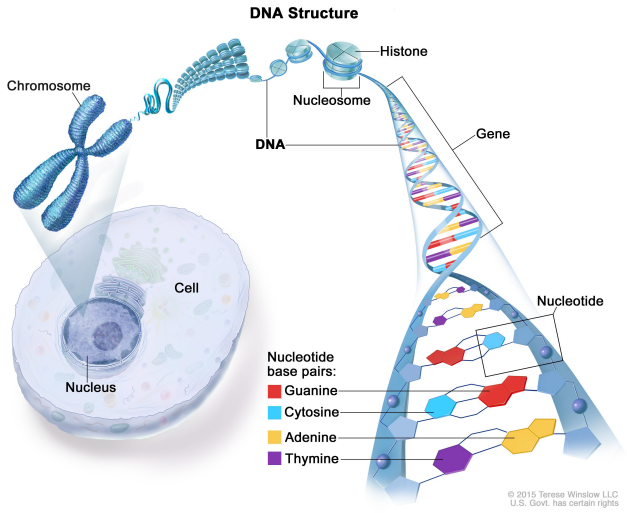


Figure 1.1: **From the cell to the DNA** from https://cancergenome.nih.gov

In humans, each cell normally contains 23 pairs of chromosomes (cf Figure 1.2). 22 of the pairs consist of non identical copies of the same chromosome, each copy presents different variants inherited from the parents and recombined during the meiosis. The two non identical chromosomes that form a pair are referred to as homologous chromosomes [Laird and Lange, 2011]. The 23rd pair contains the sex chromosomes.

The chromosomes are really long strings of DNA (deoxyribonucleic acid). DNA is made of two complementary strands. It is shaped like a ladder that has been twisted, this shape is called the double helix. The steps of the ladder are made of four distinct bases: Adenine (A), thymine (T), guanine (G) and cytosine (C). Each step of the ladder is a base pair of these four bases. As adenine

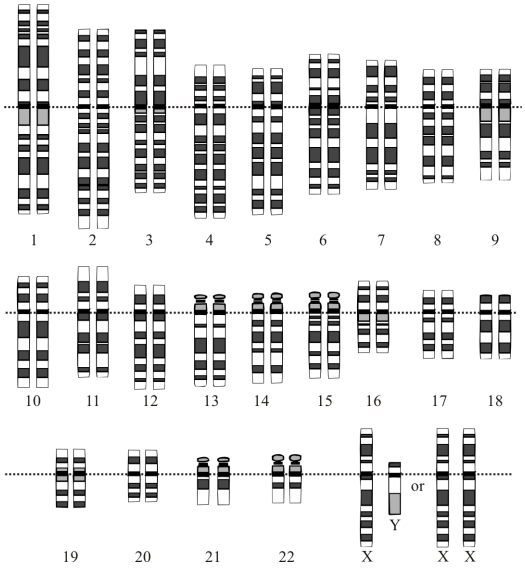


Figure 1.2: **Representation of the human genome** from National Human Genome Research Institute

(A) can only form a base pair with thymine (T) and guanine (G) can only form a base pair with cytosine (C), only two base pairs are possible. In this way, if we know the sequence of bases of one strand of DNA we will also know the sequence of bases on the other strand. The order of bases is referred to as the sequence. An example of a short sequence of a single strand of DNA is: ATTGCTCAT. It is not possible to give the exact number of base pairs contained in the whole human genome because of genetic variation between individuals, but it is known that it measures more than three billion base pairs [Richards and Hawley, 2010].

We call a locus a precise position on the genome and an allele a possible version for a given locus. More generally, an allele represents a possible version for any type of genetic variant (that could also be a gene or a haplotype for example). However, in the following, we will mainly use allele to define a version of a particular locus as in the context of this work we will mainly focus on locus variants and more particularly on SNPs (this type of variants will be presented in more details in Section 1.2.3). A locus is said to be monomorphic when all the individuals present the same allele and polymorphic when different alleles can be observed in the population at this specific locus. As every human being received two homologous copies of each chromosome from his parents, he will carry two alleles at each given locus. Each pair of alleles represents the genotype for a specific locus. For example, if we consider a locus with two possible alleles A/a, three possible genotypes can be observed: AA, Aa or aa. If both alleles are identical (genotype AA or aa) we will say that the individual is homozygous at the locus and heterozygous otherwise (genotype Aa). As genotype is the information contained within two alleles its definition will also

depend on the type of variants studied (locus, genes, haplotype...). A particular genotype at a specific genetic variant may contribute to the phenotype of an individual. The phenotype represents the observable physical characteristics of a subject as his appearance, development or behavior. In addition to the role of the genotype, the phenotype can also result from environmental factors or from the interactions between genes, environmental factors or a mixture of both. The genetic difference between two individuals can be obtained when comparing their genotypes and the importance of genetic characteristics in the manifestation of a specific phenotypic trait can be assessed when comparing the genotype of multiple individuals.

DNA to proteins

The role of DNA is the long-term storage of genetic information. Fundamentally, this genetic information is structured into genes that contain the instructions to construct proteins. Proteins are large, complex molecules that play many critical roles in the body. They do most of the work in cells and are required for the structure, function, and regulation of the body's tissues and organs. Genes are stretches of DNA that vary widely in size, some being as small as a few thousand base pairs, and some containing millions of base pairs [Laird and Lange, 2011]. The DNA sequence of the gene determine the size and the shape of the proteins it builds and thus, the diversities of their function in the organism.

When proteins are needed, the corresponding genes are transcribed into RNA, this process is known as transcription. The RNA is first processed so that non-coding parts are removed and is then transported out of the nucleus. Outside the nucleus, the proteins are built based upon the code in the RNA, this is the translation step.

1.1.2 Genetic diversity and variation

Genetic diversity could be defined as the variations within and among species. For human being, it is noted that two individuals share at least 99.5% of their genome [Levy et al., 2007]. Genetic diversity lies in the remaining 0.5 percent. This part allows to define the singularity of each individual in a population but also contains traces explaining the genetic diversity between population. Indeed, if each individual has a unique genetic code we can observe similarities between individuals that come from the same population and strong variation between population. As an example, we can consider the frequency of lactose intolerance, which is high in nearly all but those of European origin. However, before spreading in a given population a new trait is the result of a mutation or a genetic recombination. It will become a characteristic trait of the population if it presents a significant advantage face to the environmental demands.

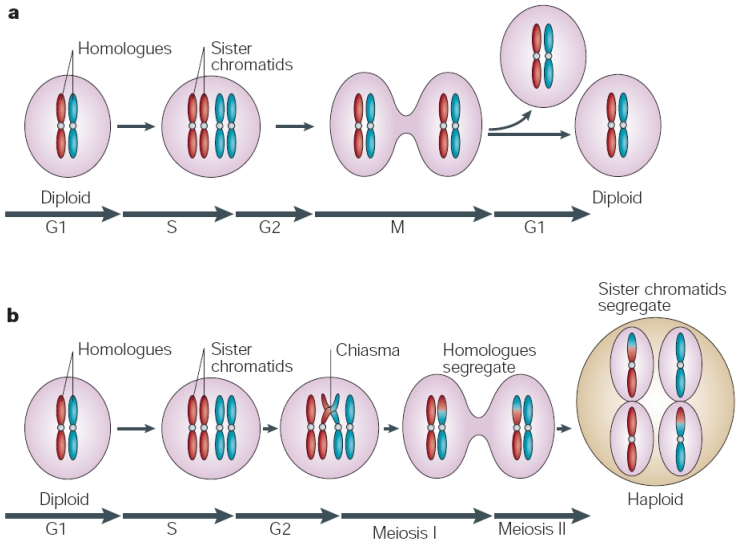


Figure 1.3: **The mitotic and meiotic cell cycles** from O'Connor [2008]. (a) Production of identical daughter cells in mitosis. (b) Formation of nonidentical gametes during meiosis.

Thus, genetic variation is the base of genetic diversity for all organisms and contributes eventually to biodiversity on a larger scale through species diversity. Genetic variation is a measure of the genetic differences that exist within a population as the differences in DNA sequences or genes between individuals. Genetic variation can occur because of mutations or genetic recombination while genetic diversity becomes important according to environmental aspects.

Mutations and genetic recombination are two events that happen during cells division. Cells divide in two ways: mitosis and meiosis as presented in Figure 1.3. Mitosis is a process where a single cell divides into two genetically identical daughter cells. In meiosis, on the other hand, a single cell divides twice in order to give rise to four cells containing half the number of chromosomes of the original cell. These new cells are our sex cells also called gametes: sperm in males, eggs in females.

Recombination

In meiosis, the pairs of chromosomes exchange DNA fragments in a process called recombination or crossing over. By the exchange of DNA sequences, genetic recombination allows producing novel sets of genetic information and thus guarantees the genetic variability between individuals. Each gamete will include both maternally and paternally genetic information. The resulting offspring will thus acquire a maximum of genetic variation by inheriting genes from all four of its grandparents [Clancy, 2008]. If we look at two specific loci,

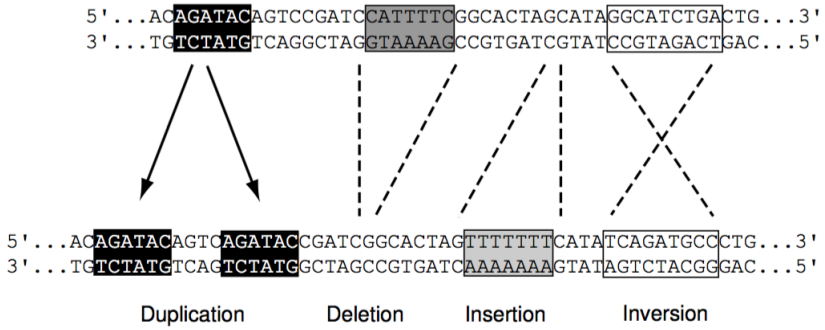


Figure 1.4: **Possible modification of a DNA sequence** from Deonier et al. [2005]

the probability of their recombination is proportional to their distance on the sequence. If they are located close to each other, they are likely to be inherited together. This phenomenon is called genetic linkage.

Mutations

Mutations are random nucleotide alterations that occur during DNA replication, which is the process of duplication of the DNA in cell division. During its replication, it is frequent that DNA undergoes chemical changes at some particular points. If most of these changes are repaired some are preserved and result in modifications in the DNA sequence. These variations are known as mutations. Mutations can occur during meiosis and thus be inherited from parents or acquired over the life of an individual during cell division, the latter being at the root of human diseases like cancer [Karki et al., 2015]. Variations in DNA sequence between individuals can take various form as displayed in Figure 1.4. For example, one or more contiguous bases in a DNA sequence may be removed (deletion), introduced (insertion), repeated in different locations (duplication) or reversed (inversion) [Deonier et al., 2005]. We can also observe point mutation, in that case, the base usually found at a particular position in the DNA is replaced by another one. Single nucleotide polymorphisms (SNP) are a type of genetic variant that emanates from point mutations (SNPs will be presented in more details in Section 1.2.3).

According to where it occurs in the genome, a genetic variation may be at the origin of genetic disorders or disease. For example, a variation in the coding sequence of a gene can lead to the alteration in the construction of the corresponding protein and then to a possible malfunction. Conversely, genetic variations may also lead to genetic improvements and have positive impacts for the individual, or they may just as well have no observable repercussion.

1.1.3 Structure in the genome

Linkage disequilibrium

As previously stated, loci that tend to be inherited together within a family are said to be genetically linked. Sets of genetically linked alleles at different loci are sometimes so close that they are rarely separated by recombination phenomenon and may persist over generations. These types of alleles combinations carried on a single chromosome are called haplotypes. Population structures can be characterized by the frequencies of various haplotypes. These frequencies can be used to reconstruct the evolutionary history of a population [Deonier et al., 2005].

Linkage disequilibrium refers to the non random association of alleles in haplotypes. Alleles of different loci are said to be in linkage disequilibrium if they occur together in a population at a higher rate than we would expect by chance and are characterized by an unusual correlation. The existence of linkage disequilibrium is influenced by the mechanisms previously described responsible for genetic variation as genetic recombination. Over time and with the succession of recombination events in a population, the stretches of founder chromosomes from the initial generation will break and will be reduced. Thus, a pair of markers will not present linkage anymore as recombination event would eventually have occurred between those points. They will then said to be in linkage equilibrium. Figure 1.5 presents these notions.

Linkage disequilibrium can be quantified by various common measures based on the coefficient of linkage disequilibrium D. If we consider two independent loci with respective alleles a/A and b/B with p_a, p_A, p_b and p_B the frequencies of the different alleles and p_{AB}, p_{Ab}, p_{aB}, p_{ab} the frequencies of the corresponding haplotypes, the coefficient of linkage disequilibrium is defined as:

$$D = p_{AB} - p_A p_B = p_{ab} - p_a p_b \tag{1.1}$$

D simply corresponds to the difference between observed and expected haplotype frequencies. It takes values between $[-1, 1]$ with $D = 0$ indicating linkage equilibrium. This measure is however highly dependent on allele frequencies at the considered loci what makes it difficult to compare the level of LD among different pairs of loci. Two other measures less sensitive to extreme allele frequencies have then been proposed as:

- $r^2 = \dfrac{D^2}{p_A p_B p_a p_b}$,
- $D' = D/D_{min}$ with $D_{min} = \begin{cases} \max\{-p_A p_B,\ -p_a p_b\} & \text{when } D < 0 \\ \min\{p_A p_b,\ p_a p_B\} & \text{when } D > 0 \end{cases}$.

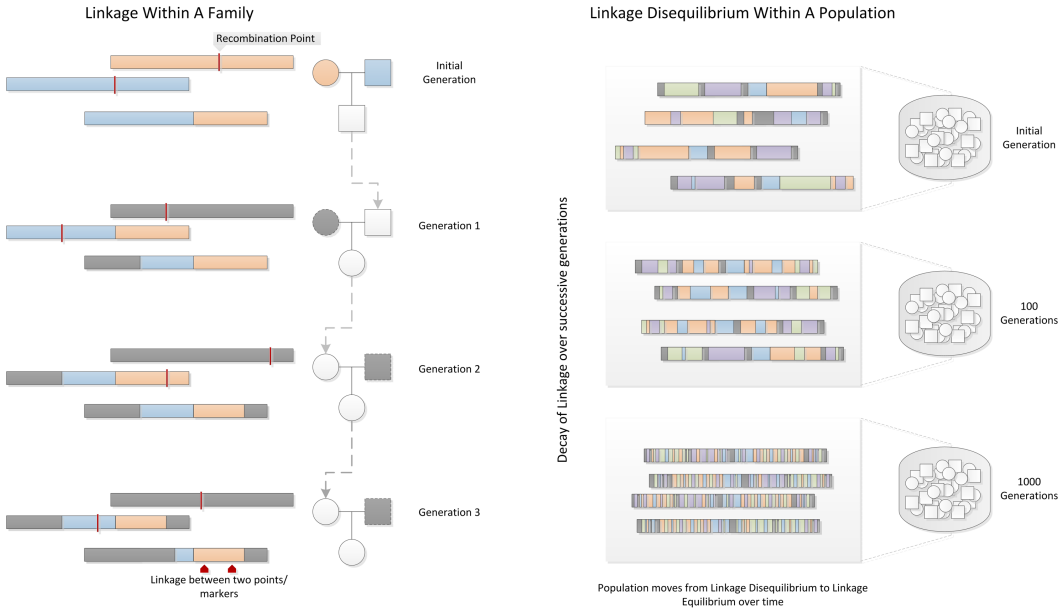


Figure 1.5: **Linkage and linkage disequilibrium** from Bush and Moore [2012] On the left is an example of linkage: two markers are said in linkage within a family when they are not broken apart by recombination events (shown as red lines). On the right is presented an example of moving from linkage disequilibrium to linkage equilibrium in a population over generation.

These two measures are the most commonly used to describe linkage disequilibrium. They range from 0 to 1. $D' = 1$ corresponds to complete LD. In this case, there are at most 3 of the 4 possible haplotypes present in the population. $r^2 = 1$ correspond to perfect LD. Perfect LD occurs when there are exactly 2 of the 4 possible haplotypes present in the population. Both situations mean that the two loci have not been separated by recombination. It also means that they have the same allele frequencies in case of perfect LD. Between these two measures, the r^2 is more commonly used to describe linkage disequilibrium in practice in association studies. It also has the particularity to be equivalent to the Pearson correlation coefficient [Mueller, 2004].

Due to recombination, linkage disequilibrium is more important between loci that are close to each other as they are more likely to be inherited together. In the genome, the presence of linkage disequilibrium between markers can lead to the formation of strongly correlated markers blocks. These blocks correspond to portions of the genome with very high levels of linkage disequilibrium and maintained by low levels of recombination. An example of block structure is presented in Figure 1.6.

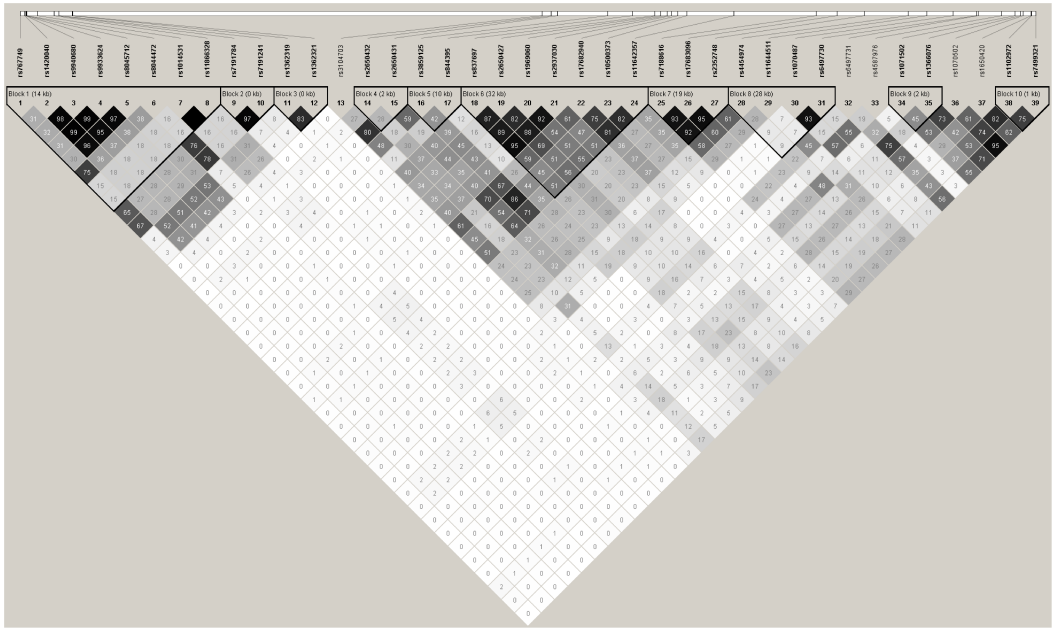


Figure 1.6: **An example of linkage disequilibrium blocks** from Zhao et al. [2013]. This plot show the LD structure of 39 SNPs, the values in squares are the pair-wise calculation of r^2.

Epistasis

Epistasis is generally defined as the interaction between different genetic markers. It corresponds to a situation in which the effect of a genetic variant depends either on the presence or the absence of another genetic variant. However the term epistasis encompasses different definitions according to disciplines and much confusion exist in the literature regarding the definition and interpretation of epistasis [Wang et al., 2011a]. In a review, Cordell [2002] point out the differences between a number of commonly used definitions of epistasis and present the limitations in moving from statistical estimates of epistatic effects to understanding genetic causation. Two main distinctions can be made in the definition of epistasis, they correspond to the concepts that are used in the disciplines of classical molecular genetics and quantitative genetics. In the first case, we will refer to biological epistasis and in the second to statistical epistasis.

Biological epistasis was first introduced by Bateson [1909] to describe the masking effect of a particular genetic variant over a second one. The expression of the effect of the second variant located at another locus is thus prevented by the presence of the first genetic variant. Table 1.1 gives an example of an epistatic effect under the definition of Bateson. Three different hair colors (white, grey, black) can be obtained according to the genotypes of a subject at two specific loci A and B. In this example, locus B is epistatic to locus A in the sense that

locus A \ locus B	bb	bB	BB
aa	White	Grey	Grey
aA	Black	Grey	Grey
AA	Black	Grey	Grey

Table 1.1: Example of Bateson [1909] definition of epistasis from Cordell [2002].

locus A \ locus B	bb	bB	BB
aa	0	0	0
aA	0	1	1
AA	0	1	1

Table 1.2: Example of a more general definition of epistasis between two loci from Cordell [2002].

the effect of locus A is observable only if the genotype at locus B is bb.

However, the definition given by Bateson [1909] is limited to a particular pattern of interactions between loci. Cordell [2002] shows the limits of the definition of biological epistasis, particularly in the case of binary traits. The author presents the example given in Table 1.2 where we can see that the locus B is epistatic to locus A but that the locus A is also epistatic to locus B. This form of epistasis is not considered in the formal definition given by Bateson where epistasis only corresponds to the masking effect of one factor on another. Numerous epistasis models that do not correspond to the definition from Bateson can be observed. Examples are given by Neuman and Rice [1992]. More generally, biological epistasis may be defined as the result of physical interactions among biomolecules making the effect of a gene dependent on the presence of one or several other genes at the individual level [Moore and Williams, 2005].

Statistical epistasis refers to the deviation from additive effects of genetic variants at different loci in a mathematical model. This definition was given by Fisher [1918] and is closer to the usual concept of statistical interaction. For example, if we consider the following model:

$$\text{logit}[P(\mathbf{y} = 1|\mathbf{x}_1, \mathbf{x}_2)] = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_1\mathbf{x}_2,$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ refer to the individual effects of each allele at locus 1 and locus 2 for a binary phenotype $\mathbf{y}$, a statistical interaction may be assessed by testing whether $\beta_3 \neq 0$. Unlike biological epistasis, statistical epistasis is defined at the population level and can be interpreted as the interindividual variability in biological epistasis among individuals [Moore and Williams, 2005]. The differences between the two definitions are illustrated in Figure 1.7. Numerous

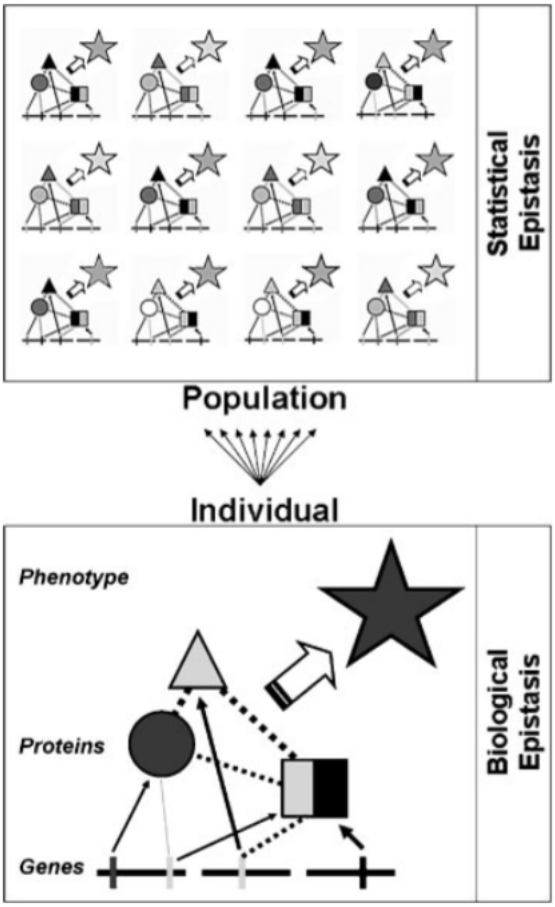


Figure 1.7: **Relationship between biological and statistical epistasis from Moore and Williams [2005]** Vertical bars refer to DNA sequence variations, circles, squares and triangles to biomolecules. The dashed lines represent the physical interactions between biomolecules leading to a particular phenotype (star). Biological epistasis happens at the individual level whereas statistical epistasis corresponds to the differences between biological epistasis in a population.

analytical approaches have been developed in the field of genetic epidemiology to detect epistasis as presented in Section 2.3. They are however based on the Fisher's definition and may not succeed in modeling the real genetic architecture. Aylor and Zeng [2008] propose an extension for estimating and interpreting epistasis that combines the strengths of classical and statistical frameworks.

In another review, Phillips [2008] distinguish tree different definition of epistasis. Compositional epistasis and statistical epistasis respectively correspond to the definition of classical molecular genetics and quantitative genetics previously describe. The third definition, functional epistasis, corresponds to the physical molecular interactions between various proteins. This definition refers to a strictly functional description of interaction without necessarily implicating a direct genetic link. Functional epistasis is sometimes understood in the former definition of biological epistasis [Moore and Williams, 2005].

Considering the variety of definitions for the term epistasis, any interpretation of genetic epistasis should be made with caution. Indeed, statistical interaction does not necessarily imply interaction on a biological level. The difficulties that exist in inferring biological meaning from statistical models have also been underlined by Trinh and Rioux [2005] or in the epidemiological literature by Thompson [1991]. They recommend extreme caution in interpreting epidemiological findings regarding the joint effects of multiple risk factors. Phillips [2008] notes that the functional basis of the potential interactions identified by the study of statistical epistatic effects in human disease has been revealed in only a few cases. According to authors such as Achkar and Fiocchi [2009], not enough resources are now at hand to perform all that is required to study epistatic interactions. They claim that true biological interactions may ultimately be better identified using molecular methods, rather than statistical methods [Cordell, 2002]. In an essay, Moore and Williams [2005] explore the relationship between biological and statistical epistasis. They explain that biological epistasis can occur without evidence for statistical epistasis notably if the individuals sampled do not present variations regarding their biomolecules interactions. However, they also present some counterexamples suggesting that inference about biological epistasis from statistical results may be possible even if particularly challenging. If all biological information from genetic, genomic, proteomic and metabolic are accessible, links between the two definitions of epistasis may be feasible.

Even if challenging the investigation of epistasis is indispensable in order to better understand the complex architecture of genetic data. Indeed, when looking for the role of genetic variants in the susceptibility of a disease it is common to see that most single variants explain individually very little of the risk for any given disease. The genetic architecture of disease is likely to be a mix of differ-

ent types of genetic effects including epistasis, gene-environment interactions, and marginal effects. Regarding epistasis, some studies have explained that this phenomenon may be ubiquitous in human biology and may even have a more important role than independent main effects [Moore, 2003; Moore and Williams, 2009]. Even if the detection of epistasis is currently a challenging task it is also a subject of a large amount of research as evidenced by the wide number of methods that have been proposed to this end in the last few years (presented in more details in Section 2.3).

1.2 Genome Wide Association Studies

With the rapid development of technologies of DNA sequencing and high-density genotyping arrays, access to genetic variants has become easier and studies that rely on a large number of genetic variants covering the whole genome are nowadays common. These studies are known as Genome Wide Association Studies (GWAS) and have been very popular since the mid-2000s. In this section, we will present association studies and the different aspects of a GWAS, from the data collection to the statistical analysis.

1.2.1 Linkage analyses and association studies

Genetic mapping refers to the variety of different methods that aim to determine the positions and the linear sequence of genes in the genome. These methods can also be used in order to map and identify genes that are involved in the susceptibility of a disease [Altshuler et al., 2008]. Disease genes investigation can be performed using genetic linkage or using association studies.

Linkage analysis

The traditional approach to map disease genes has been linkage analysis. Linkage analysis investigates "the dependence in inheritance of genes at different genetic loci, on the basis of phenotype observations on individuals" [Thompson, 2003]. More precisely, it aims at identifying chromosomal regions containing disease genes by studying cosegregation of genetic markers with the disease in families [Baron, 2001]. Cosegregation corresponds to the tendency for closely linked genes and genetic markers to be inherited together. By looking for evidence of cosegregation with other genes whose locations are already known, linkage analysis will allow determining the approximate chromosomal location of a disease gene. Indeed, markers that are transmitted through families in a manner that parallels the transmission of the disease will provide evidence for the general chromosomal location of the disease gene [Thomas, 2004]. This approach has proved to be successful for locating genes contributing to monogenic disorders. However, linkage studies are less helpful for complex traits, where multiple genes play a role in disease causation. They can also require

Figure 1.8: **Direct and indirect association in GWAS** from Hirschhorn and Daly [2005]. Here the genotyped associated genetic markers are presented in red, causal genes in green and the non genotyped marker in blue.

costly procedures for the collection of family data [Baron, 2001].

Association studies

In comparison to linkage analyses, association studies may be more economical and easier to perform. They are also known to be more powerful than linkage studies for identifying genetic polymorphisms contributing to complex diseases if the underlying causative variants are not very rare [Morris and Cardon, 2003]. The aim of association studies is to identify disease susceptibility gene variants by comparing genetic marker frequencies usually between affected and unaffected individuals. This study pattern corresponds to a case control design in which the genotype of two groups of unrelated participants is compared: people with the disease and similar people without. If the allele of a genetic marker is significantly more frequent for affected individuals, then it can be deduced that this allele is associated with the disease (directly or indirectly). Thus, linkage and association should not be confused with one another. Linkage refers to the relationship of loci, whereas association refers to the relationship of alleles at a frequency greater than predicted by chance [Pulst, 1999].

However, caution has to be taken before interpreting a significant association obtained from an association study. It can refer to a direct or an indirect association (as presented in Figure 1.8). In the first case, the genetic marker is a causal marker that is directly involved in the presence of the phenotypic trait and might implicate an associated gene. In the second case, the causal marker is not directly typed but is strongly correlated with one or several other typed genetic markers because of the presence of linkage disequilibrium. These markers in high LD with the real causal variant may present a significant statistical association that will represent an indirect association [Bush and Moore, 2012].

The first association studies focused on pre-specified genes or group of genetic markers. The objective was to test the association between a phenotypic trait and a specific region selected on the basis of previous biological knowledge or linkage analysis results [Laird and Lange, 2011]. However, for many complex

traits, the possible causal genes are not yet known, what can limit the use of candidate genes approaches.

With the completion of the Human Genome Project in 2003 and the International HapMap Project in 2005, 2007 and 2009 [Gibbs et al., 2003; The International HapMap Consortium, 2005], strong advances in the design of population-based association studies have been made possible. The Human Genome Project was an international project conducted over 15 years with the objectives to determine the sequence base pairs of the entire human genome and to obtain the complete mapping of all the genes of human beings. The HapMap project (short for "haplotype map") was an international project designed to build a catalog of human genetic variation across the genome and to characterize correlations among genetic variants [Bush and Moore, 2012]. The project allows describing haplotypes, including their locations in the genome and how common they are in different populations throughout the world. Parallel to the completion of these projects, major improvements in the genotyping technology were observed with a strong decline of genotyping expenses which make possible to genotype several hundred thousands of genomic markers across the human genome [Laird and Lange, 2011].

Genome Wide Association Studies (GWAS) emerged in this context as a new type of population-based association analysis. In GWAS, hundreds of thousands of genetic variants among the whole genome can simultaneously be investigated. The objective is to identify those variants that are associated with a phenotype of interest without being restricted to a set of pre-specified genetic markers. Since the early 2000s, GWAS have become a powerful tool for investigating the genetic architecture of complex diseases and have been successful in identifying hundreds of genetic variants [Welter et al., 2014; Li et al., 2012].

1.2.2 Sample selection in GWAS

Even if the genotyping costs strongly decline in the past years due to improvements in genotyping efficiency it is still expensive to genotype a large number of individuals. Precautions have to be taken when selecting individuals that will participate in a GWAS. The objective is to recruit a large enough number of individuals to ensure a sufficient statistical power for the identification of associations in the study while limiting too important genotyping expenses.

For quantitative phenotypes, a strategy to limit cost consists in genotyping individuals whose trait value presents important deviations from the population mean. This principle is known as "selective genotyping" and allows to increase power for detecting association [Huang and Lin, 2007]. However, as the selection of the individuals depends on phenotypic values, particular statistical methods are needed to analyze such a data set. In a case control design, the

individuals forming the two groups need to be comparable, the control group is composed of subjects that match cases according to certain characteristics such as gender, age or ethnicity [Witte, 2010]. This comparability between the two groups is necessary to limit biases and to avoid confounding factors. Regarding ethnicity, a strategy when recruiting controls may be to select people living in the same geographic area than cases or to collect information regarding the geographical origin of people such as the birth place of parents or grandparents and to check the groups' distributions [Lewis, 2002].

1.2.3 Markers collection

The simplest genetic markers that are commonly used in GWAS are single nucleotide polymorphisms (SNPs). We will present here these markers as well as the tools that are commonly used to measure them in the genome.

Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) are the most common type of genetic variations among people. Each SNP represents a single nucleotide variation occurring at a specific position in the genome. An example of a SNP is presented in the Figure 1.9. At a specific position in the genome, the nucleotide cytosine (C) may replace the nucleotide thymine (T) for a minority of individuals. As each individual possesses two homologous copies of each chromosome (one received from each parent) each specific allele may be present in 0, 1 or 2 copies. For example, possible genotypes at a specific position will be "TT", "TC" or "CC". If T is the reference, "CC" is coded 0 (referent homozygote), "TC" is coded 1 (heterozygote) and "TT" is coded 2 (variant homozygote"). Thus, from a statistical point of view, a SNP can be viewed as a categorical variable with three categories or as a quantitative variable corresponding to the number of copies.

Figure 1.9: **Example of a single nucleotide polymorphism** from https://genomainternational.com

SNPs are abundantly widespread among the whole genome and can be found either within coding sequences of genes or within non-coding or intergenic

regions. They occur on average once in every 300 nucleotides in the human genome, which represents roughly 10 million SNPs for entire genome.

For each SNP, the minor allele frequency (MAF) refers to the frequency at which the least common allele (here T) occurs in a given population. Historically a locus was considered as polymorphic if its MAF was at least superior to 1%, this condition is not required anymore [Laird and Lange, 2011] but some distinctions still exist in the literature regarding the terminology between rare and common variants. The term SNP may generally be applied to common variants whereas rare variants may be defined as mutation [Bush and Moore, 2012].

Genotyping arrays

One way to detect SNPs across the human genome is to use SNP arrays. The first commercial SNP array used for genotyping was released in 1996 by Affymetrix and targeted about 1,500 human SNPs [Lamy et al., 2011]. Since then, many different manufacturers have developed microarrays that allow genotyping a much larger number of SNPs but they differ in terms of technology, cost and gene coverage [DiStefano and Taverna, 2011]. The most commonly used SNP arrays were produced by Illumina and Affymetrix. The chips proposed by Affymetrix print short DNA sequences as a spot on the chip that recognizes a specific SNP allele whereas Illumina's chips used a bead-based technology with slightly longer DNA sequences to detect alleles. Both number and type of SNPs considered can greatly variate from one to another chip. Regarding the number of markers, some Illumina chips can genotype approximately 2.5 million SNPs whereas other ones proposed by Affymetrix focus on smaller covering expanses with a number of included SNPs reduced to around 600,000. These latter chips are proposed at a lower price and are designed for special ethnicity (European, Asian, and African) [Ha et al., 2014].

The choice of the SNPs to consider on a SNPs array is a crucial decision. They should be selected in order to represent the maximal genetic variation in the genome. In principle, Manufacturers proposed chips where the genetic selected markers are correlated with other SNPs that are not genotyped in the same region. This strategy allows reducing the number of markers to genotype while assuring a good coverage of the whole genome but also allows avoiding redundant information. These selected markers are known as tag SNPs. More specifically, each tag SNPs is a marker in high linkage disequilibrium level with as many other SNPs as possible and is used as a representative of a LD block region. By choosing only one referring marker, the other SNPs present in the region will be tested for association indirectly (as presented in Figure 1.8). When a tag SNP presents a significant association, linked SNPs have to be investigated as possible causal SNPs. Tag SNPs have been defined

using the LD-information of the HapMap project in order to provide adequate coverage of the genome for various ethnicities. SNPs that are not considered are in strong LD with at least one genotyped tag SNP ($r^2 \geq 0.7$ or 0.8) [Laird and Lange, 2011]. Thus, even though the total number of common SNPs exceeds 10 million, considering 500,000 tag SNPs in the human genome should be sufficient to represent common variation in non-African populations [Visscher et al., 2012]. As haplotype block patterns vary between ethnicity, different sets of tag SNPs have to be considered when performing association studies. For example, African population tends to have more different haplotypes due to the occurrence of more recombination events in the population history and thus, more SNPs are needed for a study focusing on African genomes [Bush and Moore, 2012].

1.2.4 Data analysis

Data qualities control

Preprocessing is the first step that is performed once SNP array data are collected for each individual. It consists of background correction and normalization across arrays following by the summarization of feature intensities into genotype classes (AA, AB, BB). The normalization step is aimed to correct for non-biological or technical variability in microarray data. Indeed, errors and bias may arise due to many circumstances as sample preparation, hybridization, instrumental noise and other factors. Then, a classification algorithm is used to transform the measured intensities into genotypes. These preprocessing steps will not be further examined in this thesis, and we will now present quality control procedures that are required once the data are preprocessed.

Before conducting any analyses of association it is necessary to filter out samples and genetic markers that could mislead the analysis. With the huge number of genetic markers genotyped for a GWAS, even a small portion of errors can lead to important false results and quality control is thus a crucial step to undertake. Quality control procedure follows different steps:

- Sample call rates: It corresponds to the proportion of markers with non-missing data for a specific genetic sample. A low call rate may be indicative of a poor quality DNA sample and the corresponding sample should be removed from the data set. No consensual threshold is fixed, it may vary between studies according to various criteria such as the genotyping platform used, the quality of the DNA samples, or the distribution of missing genotype rates across the entire sample set [Turner et al., 2011a; Anderson et al., 2010]. Usually, individuals with a call rate $\leq 90\%$ are removed but some authors recommend using a higher threshold around $97 - 99\%$ as in the study of Wellcome Trust Case Control Consortium

[2007].

- Heterozygosity: This quality criterion corresponds to the proportion of heterozygous genotypes for a given individual. Like sample call rate, heterozygosity rate can be used as a measure of DNA sample quality. Indeed, an excessive or reduced proportion of heterozygous genotypes in a genetic sample may be indicative of DNA sample contamination or inbreeding [Anderson et al., 2010]. Individuals with excessive or reduced proportion of heterozygous genotypes are identified by considering the mean heterozygosity among the sample: Individuals with a heterozygosity rate higher or lower than 2 times the standard deviation of the mean are usually removed [Luzón-Toro et al., 2015].

- Sex inconsistencies: The conformity between the reported sex and the predicted sex is checked for each individual in the sample. The individuals presenting discordant sex information are removed.

- SNPs call rates: It corresponds to the proportion of genetic samples per marker with no missing data. Like for the sample call rates, the proportion of missing data is a good indicator of a marker quality. Classically, markers with a call rate less than 95% are removed from the data set.

- Minor allele frequencies (MAF): The minor allele frequency of a marker corresponds to the frequency at which the less frequent allele occurs in a population. Usually, GWAS are used to detect a trait association with common SNPs, namely markers with a MAF ≥ 0.01. They are however not efficient to detect rare variants. If they are analyzed in a GWAS, the corresponding statistical tests will have extremely low power [Turner et al., 2011a]. Depending on the study, SNPs with a MAF ≤ 0.01 are removed. This threshold can be increased up to 0.05 in some studies. It is worth noting that markers with a low MAF can bring technical difficulties and lead to genotyping errors.

- Identity By Descent (IBD): Identity By Descent is a metric of relatedness between two individuals. Two sequences observed in two individuals are IBD if they have been inherited from the same ancestor without recombination. In GWAS the genetic samples have to be unrelated, in order to avoid bias in the analysis. The presence of not independent samples could lead to high level of correlation and overrepresentation of alleles. Thus related samples need to be excluded from the analysis. The determination of IBD for each samples pair involves the calculation of the proportion of common alleles. If we set τ the proportion of IBD shared alleles between two individuals, we will expect that $\tau = 1$ for duplicates or monozygotic twins, $\tau = 0.5$ for first-degree relatives (full siblings, parents-offspring), $\tau = 0.25$ for second-degree relatives (half siblings, grandparent-grandchild), $\tau = 0.125$ for third-degree relatives (first

cousins). Some variations may be observed for these theoretical values due to disturbing factors such as genotyping errors or linkage disequilibrium. Usually, one individual from a pair presenting an IBD value greater to a defined threshold is removed. One value that can be chosen for this threshold is 0.1875 which is halfway between the expected proportion for third- and second-degree relatives [Anderson et al., 2010; Kevans et al., 2016].

- Hardy Weinberg Equilibrium (HWE) test: The Hardy-Weinberg principle states that allele frequencies in a population will remain constant from generation to generation in the absence of other disturbing factors. Let consider a biallelic locus with variants A and a, with the three possible genotypes: AA, Aa and aa. If p is the frequency of allele A, and q the frequency of allele a, the HWE expected that the frequency for the three genotypes will respectively be p^2 for the AA genotype, $2pq$ for the Aa genotype, and q^2 for the aa genotype and that their sum will equal 1 [Wittke-Thompson et al., 2005]. The Hardy Weinberg Equilibrium is generally illustrated with the equation: $p^2 + 2pq + q^2 = 1$. Test for the deviation of the Hardy Weinberg Equilibrium can be assessed using a Chi-Square goodness-of-fit test where the observed genotype counts are compared with the expected values under HWE. Markers deviating from the equilibrium are usually excluded from the study. Sometimes, the verification of HWE is made only among controls.

Accounting for population structure

If the use of unrelated individuals in population based GWAS brings several advantages over family studies (recruitment of individuals relatively easy, use of standard statistical analysis techniques) it can lead to a high number of false-positive findings and low power due to population stratification confounding effects [Laird and Lange, 2011]. Population stratification occurs when the study samples consist of several sub-populations that show differences in allele frequencies possibly due to different ancestry. A population stratification may have an impact on the study findings if the sub-population variation among the allele frequencies is combined with a variation of the phenotypic trait under investigation across the sub-populations [Zheng et al., 2012]. In that case, a significant association may arise for a genetic marker not because of a real association with the trait but because of allele frequency differences between the founder populations that comprise the sample. Several approaches exist for dealing with population substructure as genomic control [Devlin and Roeder, 1999], structural association [Pritchard et al., 2000] or Principal Component Analysis (PCA) [Price et al., 2006]. We will briefly present this latter method. The first principal component obtains with a PCA accounts for as much variation as possible in the data in a single component. The first principal components are supposed to be highly correlated with the geographical position of

populations throughout the world and thus may reflect the environmental and cultural variation in worldwide populations, as well as population migration [Zheng et al., 2012]. The principal components can then be incorporated into a regression model when testing for marker association.

SNPs modeling

For a bi-allelic locus there are three possible genotypes: AA, Aa and aa. A is a risk allele if this allele is more common in cases than in controls. A genetic model refers to a specific mode of inheritance and defines the relationship between genotype and phenotype. The more common models are the following [Thomas, 2004]:

- Dominant model: A is dominant over a if a single copy of the allele A increases risk. Carrying either the aA or AA genotype has an equivalent association with risk compared to the genotype aa. In this model, genotypes AA and aA are compared to the genotype aa.

- Recessive model: Two copies of the variant allele are necessary to increase the risk. Here the genotype AA is compared to genotypes aA and aa. A is recessive to a and conversely, a is dominant over A.

- Additive model: In this model, there is a linear increase in risk for each copy of the risk allele. The heterozygotes genotype aA have an intermediate effect between the effect of the two homozygotes genotypes. Here, the genotype AA is compared to genotype Aa but also to genotype aa.

- Co-dominant model: All three genotypes have different effects on disease risk but the heterozygotes genotype aA may not necessarily have an intermediate effect between the effect of AA and aa. The co-dominant model includes additive models.

We have presented these four different genetic models in the context of a binary phenotypic trait but the same models can be considered for continuous outcomes. These different models can be retrieved depending on the way that genotype data is encoded. For example if we consider the three genotypes: AA, Aa and aa for a SNP, the codage $1, 1, 0$ will correspond to a dominant model, $1, 0, 0$ to a recessive model and $2, 1, 0$ to an additive model. A co-dominant model cannot directly be coded from a single SNP, it could correspond to $1, x, 0$ with $x \in [0, 1]$ unknown. The choice of the genetic model to consider and the resulting data encoding will have an influence on the statistical power of the association test [Bush and Moore, 2012]. However, the correct mode of inheritance is usually unknown and an additive model is generally considered in most GWAS. This model will have a good power to detect additive or dominant effects but will be limited if it is a recessive model that lies under the real genotype-trait relation. A co-dominant model will be the most powerful

one to detect associations when the genetic model is not known according to Lettre et al. [2007]. In this thesis, we will consider an additive genetic model.

Single marker analysis

The most commonly used approach for detecting genetic association in GWAS is the single marker analysis. In this analysis, each marker is analyzed individually for a possible association with the phenotype. The statistical test conducted depends on various factors such as the study design (population or families), the genetic model or the type of phenotypic traits (usually binary or continuous). In the following we will focus on population based study designs that use unrelated individuals under an additive genetic model.

The standard tool for testing the association between a single marker and a continuous outcome is to use a generalized linear model (GLM) and most commonly the analysis of variance (ANOVA). ANOVA is similar to linear regression with categorical predictor variables which are genotypes and potential categorical covariates (such as sex, age, known clinical covariates, population stratification components). In the case of binary traits, the association analysis of a marker can be realized using different types of tests. One commonly used test is the Pearson's chi-squared test, that is based on allelic or genotypic contingency tables. It compares the observed alleles/genotypes counts in cases and controls with their expected values under the null hypothesis that the disease status and genotypes are independent. Another family of tests also based on contingency tables are the Cochran-Armitage Trend Tests. These tests aim to find a linear trend between the probability of having the disease and the genotypes. A fisher's exact test is often used when the MAF is small and the asymptotic normal distribution does not provide a good approximation of the null distribution of the test statistic. We can also cite the Hardy-Weinberg Disequilibrium Trend Test that is based on the difference of HWE in cases and controls to test for association [Zheng et al., 2012]. However, those tests can not take into account possible confounding effects. A more flexible approach consists of using a logistic regression model belonging to the class of GLM that permits to adjust for potential covariates. Logistic regression can also provide adjusted odds ratios as a measure of effect size. Details on the use of GLM in the detection of single marker association will be presented in Section 2.1.1.

Regardless the statistical test used, the p-values obtained are then corrected to adjust for multiple comparisons. Multiple comparison issue and procedures that allow adjusting for them will be presented in Section 2.1.2.

Results of GWAS analyses, after p-values correction, are often represented by a Manhattan plots as presented in Figure 1.10.

In these plots, genomic coordinates are displayed along the X-axis, with the

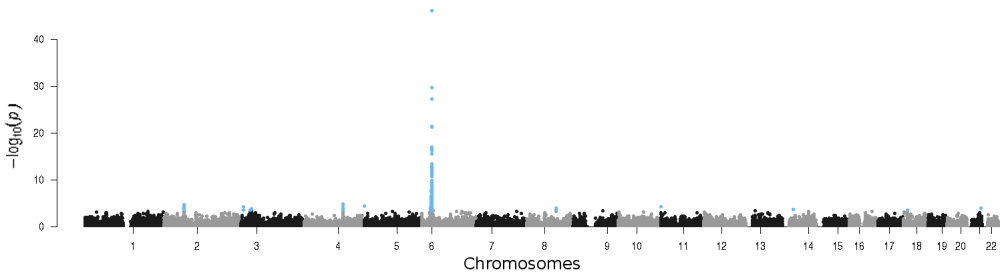


Figure 1.10: **Manhattan plot.**

negative logarithm of the association P-value for each marker (SNP) displayed on the Y-axis, meaning that each dot on the Manhattan plot represents a SNP. Because the strongest associations have the smallest p-values, their negative logarithms will be the greatest. Markers that are considered as significant after multiple testing corrections are displayed in blue. In this example, some SNPs on the chromosome 6 seem to be associated with the trait of interest.

Replications

In order to validate any significant association finds in a GWAS one or several replication studies have to be carried out. Chanock et al. [2007] suggested an ensemble of criteria in order to establish positive replication. Among their recommendations, they advise for example to use a sufficient sample size in order to provide adequate statistical power to the initial association detection. They also recommend using the same genetic model as the one reported in the initial study. Replication studies are conducted using the same design as the original study with similar phenotype and similar covariates but with an independent sample of individuals. The new subjects are selected from the same population in order to confirm the findings in the population that was targeted by the original study. If the effects are confirmed, new replication studies using datasets drawn from other populations may be realized to determine if the initial findings are specific to the first population [Bush and Moore, 2012].

1.2.5 Heritability

Complex disorders are the result of the influence of multifactorial causes where the effects of multiple genes are coupled with environmental factors. The real influence of each factor is however unknown and difficult to evaluate. One measure allowing to distinguish the role of genetic and environment is heritability. More formally, heritability corresponds to the proportion of the phenotypic variation in a particular population that is due to genetic factors.

If we consider a quantitative trait with phenotypic value P we can write its

decomposition as:

$$P = G + E + E \times G$$

where G and E correspond respectively to the genetic and environmental factors and $E \times G$ to their interaction.
Assuming independence of these different terms, we can decompose the phenotypic variance of the trait as:

$$V_P = V_G + V_E + V_{E\times G}$$

where V_G, V_E, $V_{E\times G}$ are the corresponding variance of genetic, environmental, and genotype - environment interaction terms. The genetic variance, V_G is due to all genetic effects which include additive effects V_a, dominance effects V_d and epistatic effects V_i and can be written as: $V_G = V_a + V_d + V_i$.

Heritability can be defined in two different ways:

- Narrow sense heritability: $h^2 = \frac{V_a}{V_P}$ corresponds to the fraction of the phenotypic variance of a quantitative trait that is due to additive genetic factors.
- Broad sense heritability: $H^2 = \frac{V_G}{V_P}$ corresponds to the fraction of the phenotypic variance that is due to all genetic factors including dominance, additive and epistasis effects.

In both cases, the value of heritability always lies between 0 (no genetic contribution) and 1 (all differences on a trait reflect genetic variation) but an estimation of heritability for a particular trait is never fixed. Heritability can depend on the population considered because both genetic and environmental variance can differ across populations. However, some similarities can be observed among populations. Heritability is usually high for traits such as height in human and low for fitness traits [Visscher et al., 2008]. Other divergences for heritability values can be observed between sex and overtime.

Various methods exist to estimate heritability. We will present three methods based on mixed linear models that have been reviewed by Vinkhuyzen et al. [2013] to estimate genetic variance and narrow sense heritability for a quantitative phenotype. The first type of methods that have been used are pedigree studies and more particularly twins studies. Genetic variation is estimated by the phenotypic resemblance of twin pairs and environmental variation by the phenotypic variation. This design has then been extended to families with full or halfsiblings where the narrow-sense heritability is estimated by regressing the phenotypic similarity of a pair or relatives on their genetic similarity. The authors also present a population based method where all SNPs are included in the model to estimate additive genetic variance.

Figure 1.11: **GWAS finding.** This diagram shows all SNP-trait associations with p-value $\leq 5.0 \times 10^{-8}$, published in the GWAS Catalog. Each color refer to a trait category. http://www.ebi.ac.uk/gwas/

Heritability for dichotomous traits can be estimated using a liability threshold model on the phenotype. This model assumes the existence of a quantitative trait, called the "liability score". An individual is affected if the probability that their liability score exceeds a fixed threshold. Tenesa and Haley [2013] reviewed four commonly used approaches to estimate the heritability in case of a binary phenotypic trait. These approaches include a general method for population data, a twin based approach, the Falconer's method and an approach based on generalized mixed linear models.

1.3 GWAS limitations and recommendations

During the past decade, GWAS have shown success to detect associations through LD between common SNPs and disease phenotypes [Visscher et al., 2012; Seng and Seng, 2008]. The number of publications based on GWAS increases continually making GWAS a common approach in the detection of genetic markers. New associations have then been found and reproduced among studies for varied pathologies included inflammatory bowel disease[Duerr et al., 2006], age-related macular degeneration [Klein et al., 2005], type 2 diabetes [Saxena et al., 2007; Scott et al., 2007; Zeggini et al., 2007], breast cancer [Hunter et al., 2007; Easton et al., 2007], prostate cancer [Thomas et al., 2008; Yeager et al., 2007] and even more as presented in Figure 1.11.

However, all these discoveries are tempered by several limitations [Witte, 2010; Du et al., 2012]. GWAS findings have shown to be difficult to reproduce between studies and the effects found in GWAS only explain a small part of the real phenotypic variation expected from classical family studies [Manolio et al., 2009]. For example, human height is a classic complex trait with an estimated heritability of about 80%. However the loci that have been associated with this trait in several GWAS were able to account for less than 5% of the phenotypic variance [Lettre et al., 2008; Gudbjartsson et al., 2008; Weedon et al., 2008]. These limitations can partly be explained by the characteristic of genetic data. The huge number of variants to analyses is confronted to sample size involving only hundreds of subjects, what leads to statistical challenges for estimating and identifying relevant genetic risk factors. In the following, we will present the principal limitations observed in GWAS and the solutions that have been proposed to take into account these issues.

1.3.1 Limits of GWAS

Interpretation of findings

If GWAS have allowed detecting a large number of associated genetic variants, some of these findings were sometimes unforeseen. Thus, some well-known genetic risk factors have been missing among the founding effects. Non coding SNPs seem to have a greater role in common diseases than what was expected as an important part of associated SNPs were located in non coding region in the genome without referring to specific genes [Hindorff et al., 2009] Another important limitation of GWAS findings is that a major part of identified variants present relatively moderate or even weak effects and low predictive values which can limit their clinical relevance [Witte, 2010].

Reproductibility

A second limit is that most of the GWAS findings have shown to be difficult to reproduce. Lack of reproducibility among studies may be explained first by false discoveries in GWAS. Association does not imply causation and any significant association can occur for other various reasons than a real causal genetic association. It may be possible that the significant marker is in reality associated with another factor (a confounder) which is linked to the disease but is not in the same causal pathway. For example, genetic ancestry is a major confounder in GWAS, ethnic groups often share distinct dietary habits and lifestyle characteristics that are correlated with phenotypic differences. In GWAS samples, individuals are adjusted for ancestry differences before any analysis in order to prevent the presence of many false positives. However, It has been suggested that shared environments among relatives may not be adequately taken into account. Thus, associations can emanate from diverse origins, the markers that explicitly explain the susceptibility to the disease

will be hidden among the others in an unknown proportion. Another reason explaining the lack of reproducibility among studies is the fact that studied diseases are multifactorial and a large number of markers with small effects may have an impact on the susceptibility to the disease. As small effects can be difficult to identify in such high dimensional studies as GWAS, they might be detected only on few studies.

Missing heritability

Many reasons have been suggested to explain missing heritability [Manolio et al., 2009]. A first explanation is that a large number of variants are suspected to present small or moderate effects sizes and have not been identified yet. For example, [Yang et al., 2010b] developed a new method to quantify the proportion of phenotypic variation when considering all SNPs simultaneously. They showed that a substantial proportion of genetic variation for human height was associated with common SNPs whose effects were too small to be statistically significant and have thus not been discovered yet. They showed that almost 40% of the phenotypic variance may remain hidden in undiscovered SNPs in addition to the variants that already explain about 5% of height heritability. Another important causal factor of missing heritability seems to be due to the unaccounted effect of rare variants. Indeed, GWAS are designed to detect associations between common diseases and common SNPs that present a MAF superior to 5% and analyze of variants with smaller MAF may be difficult using existing genotyping arrays. Thus, this design does not allow to consider the possible impact of rare variants that contribute to the susceptibility of common diseases [Gibson, 2012; Cantor et al., 2010]. This lack of consideration of rare variants may be one of the reasons why GWAS findings account for a little amount of the real heritability [Witte, 2010]. Rare variants identification needs the development of new methods because their detection may present a number of statistical challenges [Cantor et al., 2010]. As rare variants correspond to new mutations, their identification may be even more robust to population stratification and some authors suggest to rather use family designs [Laird and Lange, 2011]. Other causes of the unexplained heritability may be due to the fact that genetic variation is evaluated using only SNPs in GWAS whereas other variants as copy number variants may account for some trait heritability [Redon et al., 2006]. Another explanation is the lack of consideration of the role of the complex structure in genetic data by current GWAS but equally the non-consideration of gene-environment interactions [Thomas, 2010]. Finally, it has also been suggested that the non-adequacy between heritability estimated in pedigree studies and in GWAS may come from incorrect estimation measures. For example, a wrong consideration of environment in family studies [Maher, 2008] or an inappropriate definition of the global phenotypic variation to estimate heritability in GWAS.

Single SNP analysis

Another limit of the use of GWAS univariate analysis in the research of significant genetic markers is the fact that these analyses lack to consider the complete genetic architecture of the disease. For one specific regression equation corresponding to one marker, coefficients will be forming ignoring the role of any remaining markers. But, if the studied marker is linked with any other one in the whole data set, it can lead to an inaccurate estimation of the individual marker effect on the phenotype. As it is well-known, genetic markers rarely work in an individual way, they present strong correlation among themselves, notably due to linkage disequilibrium, and even shown epistatic effects. These effects cannot be taken into account in a univariate analysis framework and may be explanatory causes to missing heritability. Beside that the univariate framework will fail to consider these type of effects, it may also induce misleading estimates of direct association. As said by Cordell [2002] "if the effect of one locus is altered or masked by effects at another locus, power to detect the first locus is likely to be reduced and elucidation of the joint effects at the two loci will be hindered by their interaction". Not considering the complex genetic structures as multiple interactions between markers may thus be an explanation for the missing heritability [Haig, 2011].

1.3.2 Solutions and recommendations

Moore and Williams [2009] present several recommendations in order to improve the usefulness of genetic association results. Among their suggestions is the need to develop new powerful analytical methods that are able to cope with the complexity of genetic architecture. This need of new methods is suggested by other authors [Du et al., 2012; Witte, 2010; Moore and Ritchie, 2004], they also explain that more complex statistical analysis tools will be required to detect epistasis and rare variants now that data on less common variants become available. Thus, in the past years, a large number of methods have been developed in order to respond to the numerous challenges that have arisen in GWAS analysis and are still the subject of important investigations.

Facing the limit of GWAS in investigating the individual role of each genetic marker, looking for combinations of SNPs would be more efficient regarding the complexity of the genetic architecture of common disease. However, in the high dimensional context of GWAS, a multi-locus analysis presents numerous statistical and computational challenges and is more complex than single variant analysis [Moore and Ritchie, 2004]. Group level analyzes have been considered to this end. These methods seek to simultaneously analysis sets of genetic markers. The first methods were based on haplotypes but have rapidly been supplanted by gene-based approaches [Neale and Sham, 2004]. Group level approaches will be presented more thoroughly in Section 2.1.3.

Additionally, a plurality of statistical methods has been developed to cope with GWAS issues. These methods are based on diverse theoretical foundations whether it be regression based methods, machine learning [Dasgupta et al., 2011; Szymczak et al., 2009] or even bayesian methods [Stephens and Balding, 2009]. An example can be presented with mixed models that are used to different ends in GWAS. Indeed, approaches based on mixed models have been proposed in the context of association testing, heritability estimates, and correction for population stratification [Dandine-Roulland and Perdry, 2015]. In the case of association testing, random effects are assigned to a set of genetic variants. SKAT [Wu et al., 2011] is one of the methods based on this definition. Designed to detect sets of common or rare variants in a region, it showed powerful results against traditional rare-variant association tests. Regarding correction for population stratification, the use of mixed models can present more efficient results than resorting to Principal Component Analysis [Kang et al., 2010].

Moore and Williams [2009] also recommend to not forget the principles of classical genetics, as the use of pedigree based studies, that may have been overshadowing by the emergence of new genomics methodologies as GWAS. This idea to not only consider population GWAS design is expressed by Laird and Lange [2011]. They suggest to rather use family designs to identify rare variants. Indeed, as they correspond to new mutations their identification may be even more robust to population stratification and a population design may not be appropriate.

Finally, regarding the particular case of epistasis, it is needed to improve the knowledge of biological and statistical epistasis and its role in human health and disease. Epistasis has not been sufficiently investigated in current studies mostly because of the methodological difficulties to properly explore interactions in genetic data sets [Carlborg and Haley, 2004]. But, even if detection of epistatic effects is challenging, it is greatly assumed that this type of effects should play a relevant role in the regulation of a number of phenotypic traits and may explain a part of the missing heritability. Indeed, Zuk et al. [2012] suggest that heritability may be underestimated in GWAS because current estimates of heritability are defined on the assumption that traits do not involve genetic interactions. They thus proposed a model that takes into account epistatic interactions in relation to Crohn's disease. They found that 80% of the missing heritability could be due to genetic interactions. A better comprehension of epistatic effects and their extent are necessary in order to understand the complete architecture of genetic data. To this end, the development of new detecting approaches is required as well as the development of better experimental methods in order to confirm statistical epistasis findings [Moore and Williams, 2009]. These authors also recommend the integration of systems biology into human genetics as it may be the more efficient way to

narrow the division between biological and statistical epistasis. The development of methods to detect epistasis is particularly challenging. Indeed, if the marginal analysis of hundreds of thousands of SNP predictors is possible, with already several methodological constraints, examining all pair-wise or higher-order combinations of SNPs in a GWAS that can include one million of genetic markers is a lot more complex and demanding process even for highly efficient algorithms. The investigation of epistasis requires new statistical methods less computationally intensive [Cantor et al., 2010]. Another challenge concerns the selection of genetic markers to consider in the analysis [Moore and Ritchie, 2004]. One strategy consists in considering only SNPs that have been detected as individually significant but this strategy will include bias in the analysis. Another possibility is to select SNPs on the basis of knowledge of biological function.

In Section 2.3 we will present methods that have been developed to detect epistasis in GWAS. We will first present methods that focus on markers interaction before introducing group level interaction methods. Chapter 3 will be devoted to the presentation of a new approach to detect epistasis at the gene level that we recently published [Stanislas et al., 2017].

Chapter 2

Methodological background for epistasis investigation

In recent years a number of methods for studying epistasis have been proposed and reported in various reviews [Niel et al., 2015; Wei et al., 2014; Steen, 2012]. They vary in terms of their data analysis (either based on the whole genome or on a pre-selected set of genetic markers [Sun et al., 2014]) and their statistical methodology (bayesian, frequentist, machine learning or data mining). Most of them focus on single-locus interactions whereas some others are rather interested in considering interactions at a group level. This chapter will focus on the presentation of some of these various existing approaches with a particular interest on methods considering group level.

As the various strategies are based on different statistical methodologies, we will start this chapter with a presentation of the methodological framework required in the context of epistasis investigation. We will first present the traditional methods that are routinely used in GWAS to test association between genetic variants and a phenotype while in a second section, we will introduce statistical methods adapted to the context of high dimensional data set. The third section of this chapter will be dedicated to the presentation of peculiar approaches used to detect epistasis.

Contents

2.1 Statistical methods used to detect association in GWAS

2.1.1 Generalized Linear Model

Presentation

We consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon \tag{2.1}$$

where $\mathbf{y}$ is a n vector of continuous outcomes, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is a $n \times p$ design matrix of p explanatory variables, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$ the coefficient vector and ϵ an error term generated from a normal distribution with zero mean $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.
The parameters $\boldsymbol{\beta}$ are traditionally estimated with the ordinary least squares (OLS) that minimize the least square error,

$$\hat{\boldsymbol{\beta}}^{OLS} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mathbf{X}_i\boldsymbol{\beta})^2 \tag{2.2}$$

with the corresponding solution for the estimation of the coefficients:

$$\hat{\boldsymbol{\beta}}^{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Linear regression model is a simple, useful and widely used tool for predicting quantitative responses. But despite its simplicity this model is based on restrictive assumptions which are not appropriate in diverse settings. Among those assumptions is the normality of the error terms. However in many biological data analyses the normality assumption is not tenable, for example in the case of count or binary data, discrete distributions are more adequate. Another assumption is that the error terms are constant $\mathrm{Var}(\epsilon_i) = \sigma^2$ for all observations i but, it is not rare to observe non-constant variances (or heteroscedasticity), with for instance variances depending of the mean of the data. Linear regression model makes also the assumption that the relationship between the predictors and the outcomes is strictly linear. Predictors that present non linear association have to be transformed in order to be incorporated in a linear model.

Generalized linear model (GLM) extends the traditional linear model framework to address these issues. GLM allows the dependent variable to have a non-normal distribution by considering any probability distribution from the exponential family. Linear relationship between the explanatory variables and the mean of the distribution function, $\boldsymbol{\mu} = \mathrm{E}[\mathbf{y}|\mathbf{X}]$, is allowed by the use of a specified link function $g(.)$. Heteroscedasticity issues can also be taken into account by using a variance function that describes how the variance depends

on the mean $\text{Var}(y) = \phi f(\mu)$ with ϕ a dispersion parameter and $f(\mu)$ a function of the mean.

For a set of p variables a GLM can be presented as:

$$g(\text{E}[\mathbf{y}|\mathbf{X}]) = \mathbf{X}\boldsymbol{\beta} \tag{2.3}$$

with $g(.)$ the link function that describes how the mean, $\boldsymbol{\mu}$, depends on the linear predictor.

The parameters $\boldsymbol{\beta}$ of a GLM are usually estimated by maximum-likelihood. Maximum-likelihood estimation entails finding the set of parameters for which the probability of the observed data is greatest. By denoting $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)$, the vector of unknown GLM parameters, we can write $\hat{\boldsymbol{\theta}}$ the maximum-likelihood estimator as the solution to the maximization problem:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, L(\boldsymbol{\theta}; \mathbf{x}_1, \ldots, \mathbf{x}_p), \tag{2.4}$$

with $L(\boldsymbol{\theta}; \mathbf{x}_1, \ldots, \mathbf{x}_p)$ the likelihood function.

The classical linear regression model is a special case of GLM supposing that the distribution function is normal $\mathbf{y}|\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I)$ with constant variance ($f(\mu) = 1$ and $\phi = \sigma^2$) and using the identity as link function, namely, $g(\mu) = \mu$. The normal likelihood at its maximum can be found as an explicit function of the observed data $x_1, \ldots, x_p$ and can be shown to be the same estimator than the one obtained with the least squares approach. Maximizing the likelihood of a normal response is equivalent to minimizing the least squares criterion presented in Equation (2.2).

When the phenotype y is binary, as for case-control studies, only two values $\{0, 1\}$ are possible for the outcome. Instead of modeling the response y directly, we rather search to model $g(p)$ with p the probability that y belongs to one of the two categories and note $p(\mathbf{x}) = P(y = 1|\mathbf{x})$ the probability that $y = 1$. The regression of y on $\mathbf{x}$ is then a conditional probability with $\text{E}[y|\mathbf{x}] = p(\mathbf{x})$. To model $p(\mathbf{x})$ we need a function that produce output between $[0, 1]$ for all $\mathbf{x}$ as the logistic function:

$$p(\mathbf{x}) = \frac{\exp(\mathbf{x}\boldsymbol{\beta})}{1 + \exp(\mathbf{x}\boldsymbol{\beta})}. \tag{2.5}$$

This expression can be rewritten after a slight modification as:

$$\frac{p(\mathbf{x})}{1 - p(\mathbf{x})} = \exp(\mathbf{x}\boldsymbol{\beta}) \tag{2.6}$$

where the quantity $p(\mathbf{x})/(1 - p(\mathbf{x})) = \text{odds}(\mathbf{x})$ corresponding to the odds of presenting the phenotype $y = 1$ depending on $\mathbf{x}$. By taking the logarithm of

both side we obtain the logistic regression model:

$$\log\left(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right) = \mathbf{x}\boldsymbol{\beta} \tag{2.7}$$

where the left part of the equation defines the logit function, linear in $\mathbf{x}$.

The logistic regression model can be recovered using a GLM with the logit as link function $g(\mu) = \text{logit}(\mu) = \log[\mu/(1-\mu)]$. The distribution function is generally chosen to be the Bernoulli distribution $y_i \sim \mathcal{B}(\mu)$. Here the variance is not constant and depend of the mean with $f(\mu) = \mu(1-\mu)$ and $\phi = 1$. Unfortunately and unlike linear regression, it is not possible to find a closed-form expression for the coefficients $\boldsymbol{\beta}$ that maximize the likelihood function in a logistic regression model. Instead, iterative algorithms are needed such as the Newton-Ralphson algorithm or the iterated re-weighted least squares algorithm.

Interpretation of the coefficients in logistic regression

We consider the simple model:

$$g(\mathrm{E}[y|x]) = \beta_0 + \beta_1 x. \tag{2.8}$$

In linear regression, β_1 refers to the average change in y for a unit change in x. In logistic regression, the coefficient β_1 is much less interpretable, it refers to the change in the log of the odds that the outcome $y = 1$ occurs.

To interpret the link between the variables and the outcome in a logistic model it is usually more common to resort to odds ratio. It can be interpreted as the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure. If x is a binary variable, the odds ratio is defined as the ratio of the odds to present $y = 1$ for $x = 1$ to the odds to present $y = 1$ for $x = 0$:

$$OR = \frac{\text{odds}(x=1)}{\text{odds}(x=0)} = \frac{p(x=1)/(1-p(x=1))}{p(x=0)/(1-p(x=0))} = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1). \tag{2.9}$$

If the variable x is continuous the odds ratio can be defined as:

$$OR = \frac{\text{odds}(x+1)}{\text{odds}(x)} = \exp(\beta_1). \tag{2.10}$$

Here, $\exp(\beta_1)$ is the odds ratio associated with a one-unit increase in the exposure. In the case of a multivariate model the interpretation of each $\exp(\beta_j)$ is an estimate of the odds ratio between y and the corresponding variable x when the values of the others variables in the model are held fixed.

An $OR = 1$ means that the exposure does not affect odds of outcomes. $OR \geq 1$ means that the exposure is associated with higher odds of outcome while $OR \leq 1$ means that the exposure reduces the odds of outcome. Odds ratio has to be accompanied by the confidence interval to be interpreted. If the confidence interval for the OR includes 1, the OR is not statistically significant.

Odds ratio should not be confused with risk ratio that are defined as:

$$RR = \frac{p(x=1)}{p(x=0)}. \tag{2.11}$$

However when the occurrence of $y = 1$ is rare (as it is the case for phenotypic traits corresponding to rare disease) the odds ratio is very similar to the risk ratio. Indeed, in that case $(1 - p(x)) \approx 1$ and thus $OR \approx RR$.

Accuracy of the model, R^2

Once the model's parameters estimated, the interest is to determine the quality of the fit by measuring how well the observed outcomes are replicated by the fitted model. In linear regression, the goodness of a fit is typically assessed using the coefficient of determination R^2. This statistic represents the proportion of empirical variability in the response that is explained by the regressor variables $\mathbf{X}$ and is defined as:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.12}$$

with $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ the predicted values. By definition, R^2 always takes on value between 0 and 1. A R^2 value close to 1 would indicates that a large proportion of the variability in the response has been explained by the regression. The criterion for what is considered as a good R^2 value will depend on the field of the application and on how important is the role of other unmeasured factors on the response. According to the circumstance, a small R^2 value might be a synonym of a good fit.

However there are some dangers of relying on R^2 too heavily when interpreting multiple regression output. For example, one limit of the R^2 is that it always increases as more variables are added in the model. As an attempt to take into account this phenomenon, an adjusted measure of the R^2 has been proposed. The adjusted R^2 is a modification that adjusts for the number of variables in a model relative to the number of data points and can be calculated as:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{2.13}$$

with p the number of variables and n the sample size. When a new variable is added in the model, R^2_{adj} increases only if the new variable improves the model more than would be expected by chance and might decrease if the effect does

not improve the model. The R^2_{adj} can be used as an indicator of the fit quality when we compare two models composed of a different number of predictors.

In logistic regression analysis, there is no agreement upon analogous measure, but different propositions have been made. We will focus here on the McFadden's R^2 which is defined as:

$$R^2_{McF} = 1 - \frac{\log(L_c)}{\log(L_{\text{null}})} \tag{2.14}$$

where L_c denotes the maximized likelihood value from the current fitted model, and L_{null} the corresponding value for the null model with only an intercept and no covariates. The ratio of the likelihoods suggests the level of improvement over the null model offered by the full model.

An adjusted version of McFadden's R^2 that mirror the adjusted R^2 in linear regression has also been proposed. The adjusted McFadden's R^2 subtract p, the number of parameters in the model:

$$R^2_{McF,adj} = 1 - \frac{\log(L_c) - p}{\log(L_{\text{null}})}. \tag{2.15}$$

As with adjusted R^2, the adjusted McFadden's R^2 can decrease with the addition of a predictor that does not add sufficiently to the model, even if the R^2_{McF} increases slightly.

Utilization in GWAS

In GWAS, GLM is a standard tool for detecting associations between genetic markers and a phenotype. The approach consists to use a GLM for each single marker x as in the following:

$$g(\mathrm{E}[y|x, \mathbf{e}]) = \beta_0 + \beta_x x + \boldsymbol{\beta}_{\mathbf{e}}\mathbf{e} \tag{2.16}$$

where β_x is the parameter of interest quantifying the association between the genetic marker and the mean of the phenotype. $\mathbf{e}$ is typically a vector of covariates, such as age and gender. The link function $g(.)$ is chosen depending on the phenotype distribution. As binary traits (case-control) are more common in GWAS, it is often a logistic regression which is performed. A significance test on the parameter β_x will determine the possible association with the phenotype y.

In the context of GWAS, this strategy leads to testing hundred of thousand hypotheses, which will cause concerns such as multiple testing. In order to control the probability of false positive findings that strongly increase with the number of tests, it is required to use multiple testing procedures.

2.1.2 Multiple hypothesis testing correction

In the case of a single hypothesis, a null hypothesis H_0 (in GWAS context, no association between marker x_j and y) is tested versus an alternative hypothesis H_1 (the marker x_j is associated with the phenotype y). H_0 is rejected in favor of H_1 whenever the test statistic lies in a specified rejection region.
The rejection region of a statistical test is chosen with a determined level for the Type I error α (which corresponds to the probability to reject H_0 when is true) traditionally set to 0.05. If possible, the test is also chosen in order to maximize the probability $1 - \beta$ to reject H_0 when this hypothesis is really false, which means in the context of GWAS, to detect an association when it really exists. This probability defines the power of the test.

When realizing m multiple simultaneous independent tests with the same rejection rule, the probability of having at least one Type I error when all the hypotheses are false is $1 - (1 - \alpha)^m$. As the number m will get larger, this probability will soon get close to 1. As an example, if one thousand of markers are simultaneously investigated with α fixed to 0.05, 50 markers will be expected to be detected as significantly associated with the trait even if no real association is actually present.

Multiple testing issue is the subject of abundant investigations and various multiple testing procedures have now been proposed. Their objective is to control a Type I error risk that takes into account the number of tested hypotheses. These procedures vary in their construction but often share the cost of a reduction in the power of the individual tests as they will seek to make these tests more conservative in order to minimize the number of Type I errors [Austin et al., 2014].

Different rates to control the Type I error risks have been defined, the two most important used by multiple testing procedures are the Family-Wise Error Rate (FWER) and the False Discovery Rate (FDR). By setting V the number of false discoveries, which corresponds to the Type I error, and R the number of rejected hypotheses, also called discoveries, the two criteria are defined as:

- $FWER = P(V \geq 1)$
  the probability to observe at least one Type I error,
- $FDR = E(Q)$, where $Q = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases}$
  the expected proportion of false positives.

Bonferroni correction [Bland and Altman, 1995] was one of the first procedure to be proposed. This correction controls the FWER and is based on a simple idea which consists in rejecting any hypothesis whose p-value $\leq \frac{\alpha}{m}$, with m the

number of tested hypotheses. This ensures that the overall Type I error rate of α is maintained when performing the m hypothesis tests simultaneously. Although simple to use, this method is however often too conservative, especially when there is a broad number of tests to conduct as in GWAS. Indeed, when the number of tests becomes too large, very few p-values are able to pass the threshold which could lead on a high rate of false negatives and ultimately, a lack of power of the method.

Controlling the FDR is, on the contrary, a less stringent condition that leads to more powerful procedures than the ones based on FWER control. This is explained by the fact that FDR is smaller than or equal to the FWER from the moment where there are some true H_1 among the tested hypotheses, with equality of the rates when all the hypotheses are null. Thus, methods that control the FWER will automatically control the FDR with less stringent and more powerful results for FDR controlling procedures.

Benjamini and Hochberg [1995] proposed a procedure that controls the FDR. The method is based on ordered p-values, that is, for a set of m tested hypotheses, with $p_1, p_2, \ldots, p_m$ the corresponding p-values, we considered $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ the ordered p-values and $H_{(i)}$ the null hypothesis related to $p_{(i)}$. Let k be the largest i for which $p_{(i)} \leq \frac{i}{m}\alpha^*$, the procedure will reject all the null hypotheses $H_{(i)}$ for which $i = 1, 2, \ldots, k$. Here α^* correspond to the level of which the procedure controls the FDR. For example, if $\alpha^* = 0.05$, 5% of the tests considered as significant have in reality a real null hypothesis and are then false discoveries. The difference with FWER is that α^* control the number of false discoveries whereas α control the probability of making only one false discovery.

2.1.3 Group analysis

By considering each marker individually, GWAS univariate strategies fail to consider the possible link between the genetic variants. Instead of using a separate model for each marker, an alternative approach will be to directly consider a whole set of markers in a simultaneous way, in order to better examine the complete structure of the data. However due to the high dimensional context, traditional multivariate model are not reliable if we do not restrain the analysis to a reduced number of genetic markers. To do so, an idea consists in grouping SNPs together by defining a SNP set and analyzing those SNPs using a multivariate approach. However, according to the chosen grouping strategy, SNP sets can sometimes still be excessively large. Thus, one can for example choose to analyze only a set of defined representative markers among the set as, for example, tag SNPs [Stram, 2004; Bakker et al., 2005]. Another strategy consists in using a function for each set that represents the combined effects

of the belonging SNPs. Both strategies can also be considered together by first selecting representative markers in the set and represent the global effect of these selected SNP with a specific function. Then, once the representative markers or the functions are determined, SNP sets can be analyzed for possible disease association. With the use of adapted grouping multivariate models, several SNP sets can be simultaneously considered. In the following we present SNP sets analysis with more details.

Using SNP sets can provide various advantages over the standard analysis of individual SNPs. It can be a way to increase statistical power and reproducibility, genetic effects are more easily detected when SNP effects are aggregated together. Multiple SNPs modeling also allows to take into account the possible correlation among SNPs and thus consider linkage disequilibrium (LD) in the analysis. Not to mention that this strategy can simplify the multiple testing problem by reducing the number of variables [Fridley and Biernacka, 2011]. Depending on the grouping unit chosen, SNP sets analysis can also improve interpretability. For example, as genes are the functional unit of the genome, results may be more biologically interpretable if genes are used as grouping unit.

Various grouping strategies can be used to define a SNP set. One example is to define SNP sets using haplotype blocks. Grouping via haplotype blocks is attractive because they make explicit the use of the LD information [Schaid, 2004b] and various haplotype disease association methods have been proposed [Schaid, 2004a; Epstein and Satten, 2003; Su et al., 2008]. Unfortunately, haplotypes are not directly observable, and require the use of costly molecular haplotyping methods to be generated. These methods are not widely used as they add significant computational expense. Algorithms that are based on statistical theories to infer haplotypes are preferred for haplotype reconstructions. Various haplotype inference methods exist as for example the ones reviewed by Niu [2004]. Haplotype blocks can be generated using softwares like Haploview [Barrett et al., 2005] that proposes to use several commonly used block definitions. However, this plurality of definitions lead to different haplotype block patterns that can be difficult to interpret. Other group definition using LD information have also been proposed, for example, Dehman et al. [2015] proposed a grouping approach allowing to detect LD block structure in GWAS. This approach uses a Ward's hierarchical clustering with a specific dissimilarity based on a LD measure. The algorithm is built with the constraint that only adjacent groups and/or variables can be grouped together.

Alternatively, one can consider gene set or gene pathway set as grouping strategies by jointly taking all SNPs that are located in or near a gene or a gene pathway. Several research report have argued for the merit of gene level methods over haplotype and SNP based approaches [Neale and Sham, 2004; Jorgenson

and Witte, 2006; Lehne et al., 2011]. First, gene level analysis would provide more biologically interpretable results as they represent the functional units in the genome and are defined with more explicit boundaries. Then, consistent replication at the haplotype level can be more difficult than at the gene level as suggested by Neale and Sham [2004]. Furthermore, Jorgenson and Witte [2006] explain that variants in genes may have high probability of being functionally more important than those that occur outside of a gene. However, we still have to note that this strategy will omit to consider intergenic regions as well as many disease-associated SNPs that do not lie in defined genes. Not to mention that gene level analysis depends on the current knowledge of gene definition and remaining gaps in knowledge may prevent definition of appropriate gene sets [Fridley and Biernacka, 2011].

Grouping based on strategies beyond the ones that we have considered are also possible, as using a moving windows across the genome, but overall, the gene set strategy is one of the most commonly used approach. In any case, use of additional grouping strategies can provide alternative and better coverage of the human genome.

Once defined, a first idea to analyze SNP sets could be to perform a logistic regression (or linear regression, according to the type of phenotype variable) with the SNPs markers included as covariates. However, if this method can be more powerful in some situations than testing each marker individually it might suffer from low power due to high degrees of freedom, that increases with the number of modeled SNPs. One can thus seek to limit the number of markers when creating SNP sets and this for additional reasons. Indeed, when all genetic markers are combined in a SNP set, there is a risk that SNPs with no genetic contribution dilute the set signal by adding potential noise in the analysis and thus leading to low power. The pre-selection of the markers to use in the analysis can, for example, be done by considering only tag SNPs that capture the LD information [Qin et al., 2006; Carlson et al., 2004]. Additionally, Wu and Cui [2014] propose a diSNP selection method that select genetic variants by considering their relative information contribution to a disease status. However, even if reducing the set to a pre-selected number of representative markers can improve power by considering only important SNPs, SNP sets can still remain large and other methods than multivariate regression to analyze the set will be necessary for a better efficiency.

Various strategies have been proposed for SNP sets analysis as the ones listed by Lily Wang [2011]. Among the propositions, one consists in computing gene scores based on individual SNPs tests [Schaid et al., 2012; Lehne et al., 2011]. Other authors propose to use reduction based analytical methods, such as principal components and kernels [Fridley and Biernacka, 2011]. For example, one proposition consists in modeling the effect of a SNP set by using a semipara-

metric regression model with a non parametric kernel function $h(.)$ modeling the overall SNP effects of the set [Liu et al., 2007, 2008; Kwee et al., 2008; Wu et al., 2010b]:

$$g(\mathrm{E}[\mathbf{y}|\boldsymbol{x}, \mathbf{E}]) = \beta_0 + h(\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_p) + \boldsymbol{\beta}_{\mathbf{E}}\mathbf{E}.$$

$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ correspond to the SNPs belonging to a defined SNP set and $\mathbf{E}$ to a matrix of potential confounding covariates. Note that the proposed model can incorporate multiple SNP sets by using several kernel function $h_1(.), \ldots, h_s(.)$. The use of a kernel function is appealing because it allows to incorporate prior information into the analysis in the form of weights. Traditional procedures for GLM models are not applicable in a semiparametric framework. A way to estimate the parameters of the model is to use the least square kernel machine (LSKM) procedure [Kwee et al., 2008]. LSKM may be a tough procedure to use but instead, an alternative consists to resort to restricted maximum likelihood (REML) procedures, that are typically applied to mixed models, as it was noted that LSKM can be represented by a specific form of a linear mixed model. Liu et al. [2007] remarked that the estimation of $\hat{h}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{E}}$ with the LSKM procedure correspond to the respective estimation of random and fixed effects of a peculiar linear mixed model. Wu et al. [2010b] showed that the logistic kernel-machine regression outperforms individual SNP analysis when the causal SNP is correlated with multiple typed SNPs. In 2011 the authors introduced another kernel approach, the sequence kernel association test (SKAT) [Wu et al., 2011]. In SKAT, the relative importance of each marker can be controlled by the use of a specific weight. Those weights can be defined in various ways, for example, as the authors propose to focus their analysis on rare variants they used a function depending of the minor allele frequency to define the weights.

A number of approaches that use Principal Component Analysis (PCA) have also been proposed [Zhang and Wagener, 2008; Wang and Abbott, 2008; Gauderman et al., 2007]. The idea is to realize a PCA on the SNPs belonging to the same set and to retain the first component as the representative variable of the SNP set. The component captures the overall set structure and thus takes into account the LD among SNP. Components of each set are then incorporated in a regression model in the following way:

$$g(\mathrm{E}[\mathbf{y}|\boldsymbol{x}, \mathbf{E}]) = \beta_0 + \boldsymbol{\beta}_{\mathbf{T}}\mathbf{T} + \boldsymbol{\beta}_{\mathbf{E}}\mathbf{E}.$$

with $\mathbf{T}$ the matrix of components representing the studied SNP sets and $\mathbf{E}$ a matrix of potential confounding covariates. Using PCA to represent genes, Chen et al. [2010] extend the previous model by using regularized regression. They propose the GRASS algorithm that combine both Ridge and LASSO regression.

2.2 Statistical methods adapted to high dimensional context

The use of traditional multivariate methods such as GLM would be relevant as long as the set of genetic markers remains small enough (typically with $p < n$), which hardly ever happens in the high dimensional context of genomic data. In the case of linear regression, when $p > n$, the matrix $\mathbf{X}^T\mathbf{X}$ is not invertible and the coefficients $\hat{\boldsymbol{\beta}}$ are not defined. Additionally, when $p \simeq n$, the coefficients can be determined but the risk is to create overfitted GLM models that will be too close and dependent on the data. Additionally, Those models will lead to inaccurate regression coefficients and p-values, producing conclusions that exaggerate minor fluctuations in the data and will fail to generalise to other datasets.

In this section we will focus on the presentation of alternative approaches adapted to the context of high dimensional data sets. We will first present penalized regression methods which extend standard regression techniques so that a large number of possibly correlated variables may be analyzed. Both reliable and easily interpretable new estimators have been proposed in this category. In this work we will particularly focus on LASSO estimators that can simultaneously perform coefficient estimation and variable selection.

Other ideas to bypass high dimensionality issues when using regression methods is to use dimensionality reduction methods. In genetics it is common that a phenotypic trait can be due to the agglomeration of a large number of little contributions brought by a wide set of genetic markers. As each variable can have its own importance to explain the outcome, the simple selection of a set of variables to fit a chosen model will not capture the global variability of the data. Not to mention that it will be a difficult task to compose the reduced set of variables as we do not really known which variables are relevant. Dimensionality reduction methods propose to create new components from a set of variables by transforming the data in the high-dimensional space to a space of fewer dimensions.

Finally, in order to take into account complex data structure, machine learning approaches can be useful in high dimensional context. They can be used to consider complex relationships between variables and we will finish this section with a presentation of few machine learning approaches.

2.2.1 Shrinkage methods

In genetics, it is usually suspected that the phenotypic trait is only related to a subset of p genetic markers. In this case, we suppose that the coefficient vector $\boldsymbol{\beta}^*$ is sparse, with only a small number of coefficients β_j^*, $j \in \{1 \dots p\}$

not equal to zero. Searching for associated markers can thus be viewed as a variable selection problem. In the context of high dimensional data sets, the use of traditional selection algorithms is limited when p becomes too large. New estimators dealing with high dimensionality issues, as penalized regression, have since been proposed and outperformed traditional algorithms in term of predictions. In a comparative study, Ayers and Cordell [2010] examine the performance of a variety of penalized regressions in selecting SNPs as predictors in GWAS with standard single locus analysis and simple forward stepwise regression. They concluded that variable selection techniques based on penalized regressions outperform single-SNP analysis and stepwise selection.

In this section we will present the LASSO [Tibshirani, 1996] along with the Group LASSO [Yuan and Lin, 2006] both allowing to create sparse models by simultaneously estimating coefficients and selecting predictors.

LASSO

The Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani [1996] is a constrained version of ordinary last squares. The approach is based on the addition of a penalty on the sum of the absolute values of the coefficients known as L_1 norm, $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|$. The L_1 penalty allows the LASSO to yield sparse solution vectors by shrinking to zero some of the coefficients and so to perform estimation of the coefficients and variable selection at the same time. The L_1 norm is the smallest norm for convex problems that yields a sparse solution [Hastie et al., 2015]. Other penalized regression methods have been developed as the Ridge regression [Hoerl and Kennard, 1970] which resorts to a L_2 penalty. The Ridge regression allows to compute an unique estimator $\hat{\boldsymbol{\beta}}$ when $p \geq n$ or when the explanatory variables are highly correlated making the matrix $\mathbf{X}^T\mathbf{X}$ not invertible. However, this estimator will not allow to realize variable selection as it will usually keep all model parameters.

We note $\mathbf{X}$ a matrix with p variables. The LASSO estimator $\hat{\boldsymbol{\beta}}^{l1}$, considering a penalty parameter $\lambda \geq 0$, can be defined as the solution of the following optimization problem:

$$\hat{\boldsymbol{\beta}}^{l1} = \underset{\beta}{\operatorname{argmin}} \left(\underbrace{\sum_{i=1}^{n} (y_i - \mathbf{X}_i\boldsymbol{\beta})^2}_{\text{phenotype explanation}} + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{\text{parsimony explanation}} \right). \tag{2.17}$$

The regularization parameter λ controls the model parsimony and determines the penalization. When $\lambda > 0$, the LASSO estimates regression coefficients and, by potentially setting some coefficients β_j to 0, simultaneously selects

variables. For a large enough λ all coefficients go to 0 and no marker is selected. On the contrary, when using a fairly small λ value, all markers are kept.

The LASSO can be generalized to GLM models with similar properties by penalyzing the negative likelihood with the L_1 norm, in that case the optimization problem is written as:

$$\hat{\boldsymbol{\beta}}^{l1} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} -\log L(y_i; \mathbf{X}_i \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right). \tag{2.18}$$

Unlike OLS or Ridge estimators, there is no explicit formula for LASSO estimator as the LASSO penalty is not differentiable. However, efficient algorithms, as least-angle regression (LARS) algorithm [Efron et al., 2004], are available for computing the entire path of solutions according to the various possible λ values.

Cross-Validation

To select a value of λ, cross-validation is mostly used [Shao, 1993]. The approach allows to choose a λ value that minimize the prediction error of the estimator $\hat{\boldsymbol{\beta}}^{l1}$. To do so, cross-validation randomly partitions the sample data into K equally sized subsets that will successively be used as validation set. For each k, the model is fit using the data remaining in the $K-1$ subsets and the mean squared prediction error for the subset k is computed as followed:

$$MSE_k = \frac{1}{|\Gamma_k|} \sum_{i \in \Gamma_k} (\hat{y}_i - y_i)^2,$$

where $\hat{y}_i$ corresponds to the estimation obtained for the subjects belonging to the removed subset k, noted as the ensemble Γ_k. The cross-validation prediction error, is estimated by the average of the K MSE:

$$CV = \frac{1}{K} \sum_{k=1}^{K} MSE_k$$

When $K = n$, the cross-validation corresponds to the leave-one-out where each value y_i is simultaneously used as validation set.

The optimal λ, noted $\hat{\lambda}$, is chosen as the one that minimize the cross-validation prediction error for different models that used a λ among a set of possible values $(\lambda_1, ..., \lambda_m)$:

$$\hat{\lambda} = \underset{\lambda \in (\lambda_1, ..., \lambda_m)}{\operatorname{argmin}} \; CV(\lambda)$$

If the LASSO can be applied in a context where $p > n$ it would be limited in the sense that no more than n variables could be selected by the procedure. The LASSO will then be adequate if there is a parsimonious assumption among the data, that means if we suppose that only a small set of predictors have a real influence on the outcome. Another limit is that the LASSO is based on the hypothesis of low dependency between explanatory variables and does not perform well with highly structured variables. If there is a group of correlated variables, the LASSO will tend to select only one variable or a few of them from the group. In the context of genetic data, where the markers are supposed to be structurally grouped due to biological phenomenon such as linkage disequilibrium or epistasis, this issue may restrict the use of the LASSO and make the estimators not convenient in its initial form. In this perspective, LASSO methods have been extended with the use of structuring penalties, as Group LASSO, that allow considering preconceived knowledge regarding markers connection.

Group LASSO

Yuan and Lin [2006] introduced the Group LASSO penalty which allows to take into account the group structure of data. This penalty is an in-between the L_1 and L_2 penalties. In group LASSO, an L_2 norm of the coefficients associated with each group of variables is used. Thus, this procedure acts like the LASSO but at the group level: depending on the penalty coefficient λ, an entire group of predictors may drop out of the model. All the variables that belong to the same group are all equal or all different from zero. This allows to realize group variable selection. The importance of the penalization depends on the parameter λ. As for the LASSO, the value of the parameter lambda to retain is the one that minimizes the prediction error and is determined by cross-validation.

We note $\mathbf{X}$ a matrix with p variables belonging in G blocks of respective lengths $p_1, p_2, \ldots, p_g$ with $\sum_{g=1}^{G} p_g = p$. Group LASSO estimator $\hat{\boldsymbol{\beta}}^{group}$ [Yuan and Lin, 2006] is defined, for a $\lambda \geq 0$, as the solution of the following optimization problem:

$$\hat{\boldsymbol{\beta}}^{group} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (y_i - \mathbf{X}_i \boldsymbol{\beta})^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g} ||\boldsymbol{\beta}^g||_2 \right) \tag{2.19}$$

where $\boldsymbol{\beta}^g$ is a p_g vector of regression coefficients belonging to the group g with $\boldsymbol{\beta} = (\boldsymbol{\beta}^1, \ldots \boldsymbol{\beta}^G)$.

As for the LASSO, the Group LASSO estimator is fitted with an algorithm and does not have an explicit formula. In order to use Group LASSO with binary data, Meier et al. [2008] have proposed the Group LASSO for logistic regression. They propose an efficient algorithm suitable in the case of high

dimensional data set that can be applied to generalized linear models allowing to find $\hat{\boldsymbol{\beta}}^{group}$ that respond to the optimization problem:

$$\hat{\boldsymbol{\beta}}^{group} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} -\log L(y_i; \mathbf{X}_i \boldsymbol{\beta}) + \lambda \sum_{g=1}^{G} \sqrt{p_g} ||\boldsymbol{\beta}^g||_2 \right).$$

Another limit of the LASSO method lies in the interpretation of the respective roles of the remaining variables. One property of LASSO is that asymptotically the probability that the support resulting of the regression $\hat{S}$ contains the real support S^* is close to one. However the number of variables contained in the regression support can be very large in comparison with the real one [Huang et al., 2012].

$$P(S^* \subset \hat{S}) \underset{n \to +\infty}{\longrightarrow} 1 \text{ but } |\hat{S}| \gg |S^*|$$

As it is not possible to directly compute p-values or confidence intervals for LASSO estimators, it is difficult to distinguish the variables that have a real explanatory role.

Penalized regression has been used in the context of genetic data providing new results with sometimes adaptation of the presented method. One example is the Supervised Group LASSO proposed by Ma et al. [2007] that takes into account the cluster structure in gene expression data. The proposed approach consists of two steps. Important genes within each cluster are first selected using the LASSO and, in a second step, important clusters are selected using Group LASSO. By realizing selection at both the gene and the cluster scale, this approach thus allows to handle the presence of large cluster.

2.2.2 Methods using transformed variable input

Dimensionality reduction methods propose to reduce the dimensionality by projecting the data to a lower dimensional subspace. The projection is realized on a set of new components obtained by transforming the original data in a space of fewer dimensions. The data transformation can be based on various criteria giving the root of diverse methods that we will now present.

Principal Component Analysis and regression

Considering a group of p variables $\mathbf{X}_{n,p}$, *Principal Component Analysis* (PCA) is a statistical procedure that allows to transform a set of correlated variables as $\mathbf{X}$ into a set of uncorrelated artificial new variables $\mathbf{T}_{n,p}$, called principal components, by using an orthogonal transformation of the data. Those principal components are ordered depending of the amount of the total variance they represent with the first few ones retaining generally most of the variation present in the original data set. To transcribe the best explanation of $\mathbf{X}$, we

search components that will maximize the variance of the data with the following orthogonal transformation, $\mathbf{T} = \mathbf{XW}$ where $\mathbf{W}$ is a $p \times p$ matrix of loading vectors. We usually consider to work on the centered and standardized matrix $\mathbf{X}$ in order to center the cloud of data on the origin and to assign equal weights to all the variables. As $\mathbf{WW}^T = \mathbf{I}$, the original matrix $\mathbf{X}$ can be decomposed as $\mathbf{X} = \mathbf{TW}^T$. Each component s can be noted as

$$\mathbf{t}_s = \mathbf{X}\mathbf{w}_s = w_s^1\mathbf{X}^1 + w_s^2\mathbf{X}^2 + \cdots + w_s^p\mathbf{X}^p$$

where $\mathbf{w}_s$ is a p-vector that maximizes the variance of a linear combination of the original predictor variables under the norm constraint $\|\mathbf{w}_s\| = 1$

$$\mathbf{w}_s = \underset{\|\mathbf{w}_s\|=1}{\operatorname{argmax}} \operatorname{Var}(\mathbf{X}\mathbf{w}_s) = \underset{\|\mathbf{w}_s\|=1}{\operatorname{argmax}} \mathbf{w}_s^T\mathbf{X}^T\mathbf{X}\mathbf{w}_s. \tag{2.20}$$

For each $s > 1$, orthogonality constraints are added ensuring to construct uncorrelated components. For each $\mathbf{w}_s$, the result of this optimization problem gives as solution

$$\mathbf{X}^T\mathbf{X}\mathbf{w}_s = \lambda_s\mathbf{w}_s \ , \ s = 1, \ldots, p.$$

where λ_s are the eigenvalues of the $p \times p$ covariance matrix $\mathbf{X}^T\mathbf{X}$ with $\mathbf{w}_s$ the associated eigenvectors. As each $\mathbf{w}_s$ is chosen to have an unit length with the norm constraint $\|\mathbf{w}_s\| = 1$, λ_s will directly correspond to the variance of $\mathbf{t}_s$. The first component $\mathbf{t}_1$ will be constructed using the eigenvector associated with the largest eigenvalue λ_1, $\mathbf{t}_2$ using the eigenvector associated with the second largest eigenvalue λ_2 and so on until the last component $\mathbf{t}_p$. Subsequently, each succeeding component will have the highest variance possible under an orthogonality constraint with the preceding ones.

PCA is frequently used as a way to reduce the dimensionality of a given data set $\mathbf{X}_{n,p}$ composed of p variables by only reckoned on the k ($k < p$) first components $\mathbf{T}_{n,k}$ that represent the original set of variables in a lower dimensional space. The value k may be chosen using various criteria as ensuring that the k components illustrate a certain amount of the total variance of $\mathbf{X}$. The resulting set of k components constructs with PCA can then be use as new predictors in a regression model as a solution to high dimensionality issues.

$$\mathbf{y} = \boldsymbol{\beta}_0 + \sum_{j=1}^{k} \boldsymbol{\beta}_j\mathbf{t}_j + \boldsymbol{\epsilon} \tag{2.21}$$

The orthogonality of the components allows to eliminate possible multicolinearity problem among the original set of variables $\mathbf{X}$. This approach that combined both PCA and regression is called *Principal Components Regression* (PCR) and is particularly compelling when there are strong correlations between different explanatory variables or a group structure among the data.

Partial Least Square

If PCA finds components that describe as much as possible of the variation in $\mathbf{X}$, the response $\mathbf{y}$ is not considered in the construction process. Therefore, if there is a lot of variation in $\mathbf{X}$ that is not correlated with the response it is possible that the component might not be adequate to describe $\mathbf{y}$. Moreover, useful predictive information for $\mathbf{y}$ may be discarded as noise when constructing those components.

Partial Least Square (PLS) also constructs a set of new linear combinations of the original variables to use in a regression model but, unlike PCA, PLS identifies these new features in a supervised way by using the outcomes $\mathbf{Y}_{n,q}$ (which can be multidimensional with $q \geq 1$) in addition to $\mathbf{X}_{n,p}$. Specifically, PLS regression searches for a set of orthogonal components (called latent variables) by performing a simultaneous decomposition of $\mathbf{X}$ and $\mathbf{Y}$ with the constraint that these latent variables explain as much as possible of the covariance between $\mathbf{X}$ and $\mathbf{Y}$. The decomposition of $\mathbf{X}$ and $\mathbf{Y}$ can be presented as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T$$

where $\mathbf{T}$, $\mathbf{U}$ are respectively the latent variables and $\mathbf{P}$, $\mathbf{Q}$ respectively the loading matrix of $\mathbf{X}$ and $\mathbf{Y}$. The latent variables $\mathbf{T}$ are linear combination of the original predictor variables such as $\mathbf{T} = \mathbf{X}\mathbf{W}$ with $\mathbf{W}$ a matrix of loading vectors. Those loading vectors are calculated in order to maximize the covariance between the latent variables $\mathbf{T}$ and $\mathbf{U}$ under the norm constraints $\|\mathbf{w}_1\| = 1, \|\mathbf{q}_1\| = 1$ and the orthogonality constraint $\mathbf{t}_i \perp \mathbf{t}_j \ \forall i \neq j$.
We can notice the difference between $\mathbf{W}$ in PLS and in PCA. For both PLS and PCA $\mathbf{W}$ defines the transformation of $\mathbf{X}$ to components $\mathbf{T}$. However in PCA $\mathbf{W}$ reflects the variance structure of $\mathbf{X}$ whereas in PLS $\mathbf{W}$ reflects the covariance structure of $\mathbf{X}$ and $\mathbf{Y}$.

The first pair of vectors $\mathbf{w}_1$, $\mathbf{q}_1$ is obtained following the following optimization criterion with both centered and standardized $\mathbf{X}$ and $\mathbf{Y}$:

$$\begin{aligned}(\mathbf{w}_1, \mathbf{q}_1) = \underset{\|\mathbf{w}_1\|=\|\mathbf{q}_1\|=1}{\operatorname{argmax}} \operatorname{Cov}(\mathbf{t}_1, \mathbf{u}^1) &= \underset{\|\mathbf{w}_1\|=\|\mathbf{q}_1\|=1}{\operatorname{argmax}} \operatorname{Cov}(\mathbf{X}\mathbf{w}_1, \mathbf{Y}\mathbf{q}_1) \\ &= \underset{\|\mathbf{w}_1\|=\|\mathbf{q}_1\|=1}{\operatorname{argmax}} \mathbf{w}_1^T\mathbf{X}^T\mathbf{Y}\mathbf{q}_1\end{aligned}$$

When the latent vectors $\mathbf{t}_1$ is found, we compute $\mathbf{p}_1$ the vector of regression coefficients between $\mathbf{t}_1$ and $\mathbf{X}$ and $\mathbf{q}_1$ the vector of regression coefficients between $\mathbf{t}_1$ and $\mathbf{Y}$ as:

$$\mathbf{p}_1 = \frac{\mathbf{X}^T\mathbf{t}_1}{\mathbf{t}_1^T\mathbf{t}_1}$$

$$\mathbf{q}_1 = \frac{\mathbf{Y}^T\mathbf{t}_1}{\mathbf{t}_1^T\mathbf{t}_1}$$

The data matrix $\mathbf{X}$ and $\mathbf{Y}$ are then deflated, which means that the information related to the latent variable $\mathbf{t}_1$ is subtracted from the data matrices $\mathbf{X}$ and $\mathbf{Y}$ as:

$$\mathbf{X}_d = \mathbf{X} - \mathbf{t}_1 \mathbf{p}_1^T$$
$$\mathbf{Y}_d = \mathbf{Y} - \mathbf{t}_1 \mathbf{q}_1^T$$

With $\mathbf{X}_d$ and $\mathbf{Y}_d$ the deflated matrices. The estimation of the next component is calculated using the same procedure starting from $\mathbf{X}_d$ and $\mathbf{Y}_d$ and so on until the last component $\mathbf{t_k}$. After every iteration, vectors $\mathbf{w}$, $\mathbf{t}$, $\mathbf{p}$ and $\mathbf{q}$ are saved as columns in matrices $\mathbf{W}, \mathbf{T}, \mathbf{P}$ and $\mathbf{Q}$, respectively. Different algorithms can be used for calculating a PLS model, as NIPALS, SIMPLS or the kernel algorithm.
Finally, we can regress $\mathbf{Y}$ on $\mathbf{T}$ using the following model:

$$\mathbf{Y} = \boldsymbol{\beta}_0 + \mathbf{TQ} + \epsilon = \boldsymbol{\beta}_0 + \sum_{j=1}^{k} \mathbf{q}_j \mathbf{t}_j + \boldsymbol{\epsilon} \tag{2.22}$$

In PLS, the latent variables are simultaneously calculated as the partial regression are executed which give a better predictive power to the method in comparison with PCR.
We can remark that the criterion to optimize with PLS is a compromise between the ACP and CCA criteria:

$$\mathrm{Cov}(\mathbf{X}\mathbf{w}_s, \mathbf{Y}\mathbf{q}_s) = \underset{\text{CCA}}{\mathrm{Cor}(\mathbf{X}\mathbf{w}_s, \mathbf{Y}\mathbf{q}_s)} \sqrt{\underset{\text{PCA of } \mathbf{X}}{\mathrm{Var}(\mathbf{X}\mathbf{w}_s)}} \sqrt{\underset{\text{PCA of } \mathbf{Y}}{\mathrm{Var}(\mathbf{Y}\mathbf{q}_s)}}$$

Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) aims to investigate the presence of relations between two groups of quantitative variables respectively noted $\mathbf{X}_{n,p}$ and $\mathbf{Y}_{n,q}$ with $q \leq p$. The method consists in finding q couples of variables presented in the matrices $\mathbf{U}_{n,q}$ and $\mathbf{V}_{n,q}$ of maximal correlation where $\mathbf{U}$ and $\mathbf{V}$ are respectively a linear combination of $\mathbf{X}$ and $\mathbf{Y}$.
These new sets of variables can be decomposed as:

$$\mathbf{U} = \mathbf{XW}$$

$$\mathbf{V} = \mathbf{YQ}$$

with $\mathbf{W}_{p,q}$, $\mathbf{Q}_{q,q}$ the matrices of loading vectors. As $q \leq p$ only q couples of variables can be constructed with CCA.
Each step s will consist in finding the couple of vectors $(\mathbf{w}_s, \mathbf{q}_s)$ (and then the couple of variables $(\mathbf{u}_s, \mathbf{v}_s)$) of maximal correlation and uncorrelated with the previous couples:

$$(\mathbf{w}_s, \mathbf{q}_s) = \mathrm{argmax}\, \mathrm{Cor}(\mathbf{u}_s, \mathbf{v}_s) = \mathrm{argmax}\, \frac{\mathrm{Cov}(\mathbf{u}_s, \mathbf{v}_s)}{\sqrt{\mathrm{Var}(\mathbf{u}_s)}\sqrt{\mathrm{Var}(\mathbf{v}_s)}}$$

By setting the constraint $\text{Var}(\mathbf{u}) = \text{Var}(\mathbf{v}) = 1$ we obtain the following optimization problem :

$$(\mathbf{w}_s, \mathbf{q}_s) = \underset{\text{Var}(\mathbf{u})=\text{Var}(\mathbf{v})=1}{\text{argmax}} \text{Cov}(\mathbf{u}_s, \mathbf{v}_s) = \underset{\text{Var}(\mathbf{u})=\text{Var}(\mathbf{v})=1}{\text{argmax}} \text{Cov}(\mathbf{X}\mathbf{w}_s, \mathbf{Y}\mathbf{q}_s)$$
$$= \underset{\text{Var}(\mathbf{u})=\text{Var}(\mathbf{v})=1}{\text{argmax}} \mathbf{w}_s^T \mathbf{X}^T \mathbf{Y} \mathbf{q}_s$$

The result of this optimization problem gives as solution:

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X}\mathbf{w}_s = \lambda^2 \mathbf{w}_s \tag{2.23}$$

with $\mathbf{w}_s$ the eigenvectors corresponding to the maximal eigenvalue λ^2. λ^2 corresponds to the canonical correlation coefficient of $\mathbf{u}_s$ and $\mathbf{v}_s$. Knowing $\mathbf{w}_s$ we are able to determine $\mathbf{q}_s$ as $\mathbf{q}_s = \frac{(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X}}{\lambda}\mathbf{w}_s$ and then the couple ($\mathbf{u}_s$, $\mathbf{v}_s$).
The s-th pair of canonical variables is the pair of linear combinations $\mathbf{u}_s$ and $\mathbf{v}_s$ having unit variances, which maximizes the correlation among all choices that are uncorrelated with the previous $s-1$ canonical variable pairs.

2.2.3 Machine Learning

Machine learning is an ensemble of methods designed to model complex relationships between variables and to detect patterns in data without always specifying a particular genetic model. In genetics, these methods can show advantages for taking into account relationship and structure. A learner is a ML algorithm that has been trained to make the best possible prediction on new data. In this section we will present some machine learning approaches frequently used in genetic as ensemble learning approaches, support vector machines and neural networks.

Ensemble Learning

Ensemble learning methods are a set of approaches that combine a set of models in order to improve their global performance. Starting from a weak learner which has a performance that is only slightly better than random, usually a classifier or a regression function, they use repeatedly the weak learner on various versions of a data set and combine an overall final prediction by using a weighted vote of the predictions obtained on the multiple estimated models.

Tree based methods Tree-based methods as regression tree (CART) [Breiman et al., 1984] are often used as base learners. Tree-based methods seek to construct partitions from a set of individuals by splitting their predictor space

using a recursive algorithm. The segmenting criterion at each node consists in minimizing the variance or the heterogeneity intra-class of the two new constructed regions until a stopping criterion is reached. The prediction of a single observation will be obtained by considering the mean (or the mode for binary outcome) of the observations belonging in the same region. Tree-based methods are simple and useful for interpretation but, when they are considered as a single model, are not competitive with the best supervised learning approaches in terms of prediction accuracy and stability. If the data change a little, the tree picture can be very different. Working on aggregating trees will allow to avoid these problems. By producing multiple trees which are then combined in order to obtain a single prediction, tree aggregation will lead to wide improvements in prediction accuracy, at the expense of some loss in interpretation.

Bagging and Random Forests In the late 1990's, Breiman [1996] proposed the bagging approach (short for bootstrap aggregating) that aggregates tree results obtained on B bootstrap samples of the original data set. The bootstrap, introduced by Efron [1979], consists to draw sample sets of the same size than the original data set by performing an uniform sampling with replacement. Thereby, some observations may be repeated in each sample sets. This approach allows to obtain a set of simulated samples having the same distribution as the real data making the bootstrap particularly useful in absence of any information about the distribution. In bagging, the idea is then to consider B bootstrap samples from the data and to fit a tree on each of these sample. Predictions are obtained by averaging the B trees results for regression or by assigning class per majority votes for classification. In 2001, Breiman proposed Random Forests as a special version of bagging that uses a small tweak allowing to create decorrelated trees [Breiman, 2001]. As in bagging, B trees are built on bootstrapped samples. But when building these trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors. Usually, we choose $m = \sqrt{p}$ for categorical outcome and $m = \frac{p}{3}$ for continuous outcome. Predictions are then obtained using the same idea as in bagging.

Boosting Another method to generate an ensemble learning is boosting [Freund, 1995], which works as an ensemble method in a similar way as bagging by combining the results of a set of training models. Here, each base learner is constructed using a weighted version of the data, with the weights depending on the results of the previous base learner. Thus, while bagging simultaneously trains separate models on bootstrap samples before averaging their prediction, Boosting, in another way, works iteratively. At each step, it uses information brought by the previous iteration to train the model and consider as a final result a weighted combination of all trained model. At each iteration the weight coefficients depend on the performance of the previous models, and are read-

justed in order to give more weights to misclassified examples allowing the next models to focus on difficult examples. Boosting is an approach that can be applied on a large set of statistical methods in order to improve their predictions results. It was originally designed for solving classification problems, but can also be extended to regression. For regression problems, boosting is a form of "functional gradient descent" [Elith et al., 2008]. At each step we search to minimize a loss function, as the squared error $(y_i - f(\mathbf{x}_i))^2$ in the case of continuous outcomes, and the weights of each model are updated to minimize that error.

Support Vector Machines

Support Vector Machines (SVM) are a class of supervised learning algorithms first proposed by Vapnik and colleagues in the late 1970 [Vapnik, 1979]. Firstly designed for classification problems SVM have been then generalized to regression analysis. Originally SVM are a generalization of maximal margin classifier, a method that consists in finding a hyperplane that separates a dataset between two perfectly separated classes. Considering the margin M, which corresponds to the minimum distance between the hyperplane and the training observations, the objective is to find a hyperplane with an optimal marge M that correctly separates the data while being the farthest from the training observations. However this method is restricted to linear separable cases and does not apply in many situations. In order to accommodate non-linear class boundaries, SVM propose an extension of maximal margin classifier by using kernels. A kernel can be defined as a function K that measures the effect of a couple $(\mathbf{x}_i, \mathbf{x}_j)$ trough its correlation or its distance and then quantifies its similarity. In the linear case, the kernel K is $K(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} x_{ik}x_{jk}$. Common kernels that allow to consider non-linear transformation include:

$$\text{dth-Degree polynomial: } K(\mathbf{x}_i, \mathbf{x}_j) = \gamma(c + \sum_{k=1}^{p} x_{ik}x_{jk})^d,$$

$$\text{Radial basis: } K(\mathbf{x}_i, \mathbf{x}_j) = exp(-\gamma \sum_{k=1}^{p}(x_{ik} - x_{jk})^2),$$

with $c \in \mathbb{N}$, $d \in \mathbb{N}$ and $\gamma > 0$.

Neural Network

Neural Network are computational models inspired by the way biological neural networks process information in the human brain. The basic unit in a neural network is the neuron. In the simplest possible neural network with only one neuron considered, an output y is computed from a set of inputs $x_1, \ldots, x_k$ using a defined function f. Weights $w_1, \ldots, w_k$ are assigned to the k inputs and reflect of their relative importance. The output of the neuron can be written as $y = f(w_1x_1 + \cdots + w_kx_k)$ with f a non-linear activation function usually defined

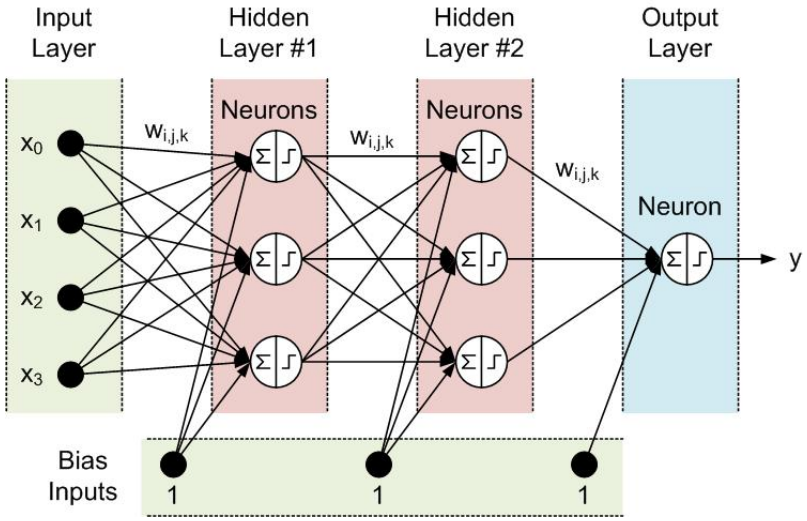


Figure 2.1: **A multi layer perceptron having two hidden layers** from http://griswoldandco.com/

using a piecewise linear function or a sigmoid function, such as the hyperbolic tangent or the logistic function. Many neurons can be considered in a neural network, each of them serving as inputs or outputs for other neurons, and are organized in layers. The input layer represents the initial variables, whereas the output layer, which often consists of a single neuron, generates an output signal corresponding to the response variable y. The input layer communicates with one or more hidden layers that will in turn link to the output layer using weighted connections. An example of the architecture of a neural network with two hidden layers is presented in the figure 2.1. The bias inputs correspond to intercept terms.

2.3 Methods to detect epistasis

Given the challenges of its investigation, epistatis is a topic of great interest in genetic and has been the subject of a large amount of research in the past few years. Numerous methods from various theoretical backgrounds have been proposed and have been reviewed in several papers [Niel et al., 2015; Wei et al., 2014; Steen, 2012; Hu et al., 2014; Shang et al., 2011; Cordell, 2009; An et al., 2009]. Given the huge number and the diversity of the methods it is difficult to categorize the methodologies used. Various criteria have been used in the previously mentioned reviews with several overlaps between the different classification schemes. In this section we will only present some of the approaches developed to detect epistasis following the classification presented in the Figures 2.2 and 2.3. We will first introduce the type of epistasis investigation that can be conducted (exhaustive or filtering) before focusing on methods designed

to detect single-locus interactions methods and ending with methods that have been proposed at the group level.

2.3.1 Exhaustive or filtering based investigation

The first issue that arises when investigating for epistasis effects in GWAS is dealing with the high dimensional context. For example, for a GWAS including one million of SNPs, 5×10^{11} pairwise interactions have to be investigated. Two types of strategies can be conducted in this context, an hypothesis-free approach where all pairs of markers are exhaustively investigated for the search of epistasis or an hypothesis-driven approach where the search is restricted to specific subset of genetic markers selected according to some prior filtering hypothesis [Wei et al., 2014].

Exhaustive search

A first idea to detect interactions between a couple of genetic markers may be to resort to a parametric model in a similar approach than the one used to detect single association in GWAS. Thus, in the context of the detection of epistasis effects the model should be:

$$g(\mathrm{E}[\mathbf{y}|\mathbf{x}]) = \beta_0 + \beta_1\mathbf{x}_1 + \boldsymbol{\beta}_2\mathbf{x}_2 + \boldsymbol{\beta}_3\mathbf{x}_1\mathbf{x}_2. \tag{2.24}$$

The test for interaction is then based on the coefficient $\boldsymbol{\beta}_3$ with $g()$ the link function defined depending of the phenotype distribution. Software as PLINK [Purcell et al., 2007] propose this type of model to detect interactions in GWAS. A specific set of SNPs can be selected by the user but otherwise all pairwise combinations of SNPs can be tested. In all cases, p-values obtained with this strategy will have to be corrected using a multiple correction procedure in order to prevent the presence of a too large number of false positives as many significant results will be generated. Using such a parametric model to detect interactions will however pose several limits. First, the huge number of tests to perform if all pairwise interactions have to be considered will generate a considerable computational burden and will be time consuming. Second, only interactions with very strong effect will be able to pass the multiple test threshold and thus be detected under this framework. Finally, this type of model does not allow to consider the global structure of genetic data as only two markers are considered in each model. Using an unique model that contains all main and interaction effects will however not be reliable given the high dimensional context. Besides the initial parametric framework proposed in PLINK and despite the methodological difficulties, other exhaustive approaches have been developed to investigate epistasis in the whole genome and will be presented in the section dedicated to SNP-SNP approaches.

Filtering based search

In order to limit the computational burden issue of exhaustive investigation, a strategy is to limit the analysis to a reduced number of genetic markers before testing for interaction. Various filtering approaches may be used in this way [Sun et al., 2014]. Common filtering steps are based on statistical results, biological knowledge or knowledge extracted from the dataset using data mining and data integration techniques.

- **Statistical filtering:** Filtering based on statistical results includes the ideas to keep only SNPs with significant marginal effects or regarding SNP genotype frequencies [Ackermann and Beyer, 2012]. Choosing SNPs under the assumption that they present significant main effects follows the concept of strong heredity [Chipman, 1996]. This principle stipulates that both predictors of a significant interaction should be marginally significant. Several two step methods based on this type of filtering have been proposed [Kooperberg and LeBlanc, 2008; Wu et al., 2009; Evans et al., 2006; Marchini et al., 2005] and show good power. However filtering SNPs on the basis of the presence of their main effect will bias the analysis by only considered one type of interaction. SNPs that are associated to the phenotype of interest solely by their interaction effects will not be retained in the analysis. In order to capture the effects of SNPs that do not necessarily manifest main effect but may interact with other SNPs in an epistasis network, other filtering approaches built on the relaxed assumption of weak heredity, where only one effect has to be marginally significant, have been proposed [Bien et al., 2013; Gao et al., 2013; Li et al., 2014].

- **Biological knowledge filtering:** A second filtering approach consists to use biological knowledge as for example, resorting to SNPs belonging to a same pathway or encoding for particular proteins that are involved in relevant interaction. In the latter case, information regarding protein-protein interactions can be found using online database like IntAct [Kerrien et al., 2012], BioGRID [Chatr-Aryamontri et al., 2015] or STRING [Franceschini et al., 2013]. Public databases allowing to retrieve information about genes and thus SNPs involving in specific pathway are also available like KEGG [Kanehisa et al., 2012], BioCarta [Nishimura, 2001] or Reactome [Fabregat et al., 2016]. We can also cite the software program Biofilter [Bush et al., 2009] that catalogs biological information by integrating data from multiple databases. Filtering using biological knowledge may be a good strategy to improve the inference of biological epistasis from statistical interaction results and will also have the advantage to be easily interpretable. Several studies successfully identified interactions using this type of filter [Ma et al., 2012; Turner et al., 2011b] however as the previous filtering approach, filtering SNPs under biological assumption will bias the analysis. Only interactions between SNPs

for which biological knowledge is established will be analyzed, thus preventing the detection of novel interactions [Ritchie, 2011]. Furthermore, online databases are not complete and the understanding of biological pathways limited. Certain pathways are more intensely investigated than others which can bias the choice of the SNPs to analyze.

- **Data mining filtering:** A third strategy consists in selected SNPs using data mining techniques such as Relief algorithms. The first version of these algorithms was proposed by Kira and Rendell [1992] and has now been extended to several variants. Relief algorithms are nearest neighbor based approaches that compute a proximity measure between individuals on the basis of their genetic similarity. The power of each SNP to select is evaluated according to their capacities to distinguish between close individuals. Among the algorithms that have been proposed we can cite, ReliefF [Kononenko, 1994], Tuned ReliefF (TuRF) [Moore and White, 2007] or Spatially Uniform ReliefF (SURF) [Greene et al., 2009].

2.3.2 SNP - SNP interactions detecting methods

Single-locus is the most common level at which interactions are considered and a large number of approaches have been proposed using various methodologies. In this section we will present a sample of these methods. We chose to classify these methods in the following way. First we will introduce the regression based approaches, which may be exhaustive or filter based. Then, the methods based on an interaction test that evaluates the difference between case and control regarding criteria such as linkage disequilibrium or odds ratio will be presented. We then consider methods based on machine learning or data mining and we will eventually mention other families of methods such as entropy or bayesian modeling. This classification of single locus interaction detecting methods is illustrated in the Figure 2.2 that presents for each category a non exhaustive list of corresponding methods.

Regression based methods

Generalized linear model is a common framework to investigate epistasis. Logistic models are commonly used as many studies have case control designs. Typically, a model is defined for each couple of genetic markers and the interaction effect is assessed using a likelihood ratio test. However this framework can only consider one interaction at the time and will be confronted to computational and statistical burdens if a too large number of variants are considered. Therefore filtering on significant main effect [Kooperberg and LeBlanc, 2008] or on biological information [Emily et al., 2009] is usually first conducted. Other authors proposed to resort to penalized regression in order to consider multiple genetic markers and their interactions in a single model. Commonly a L_1

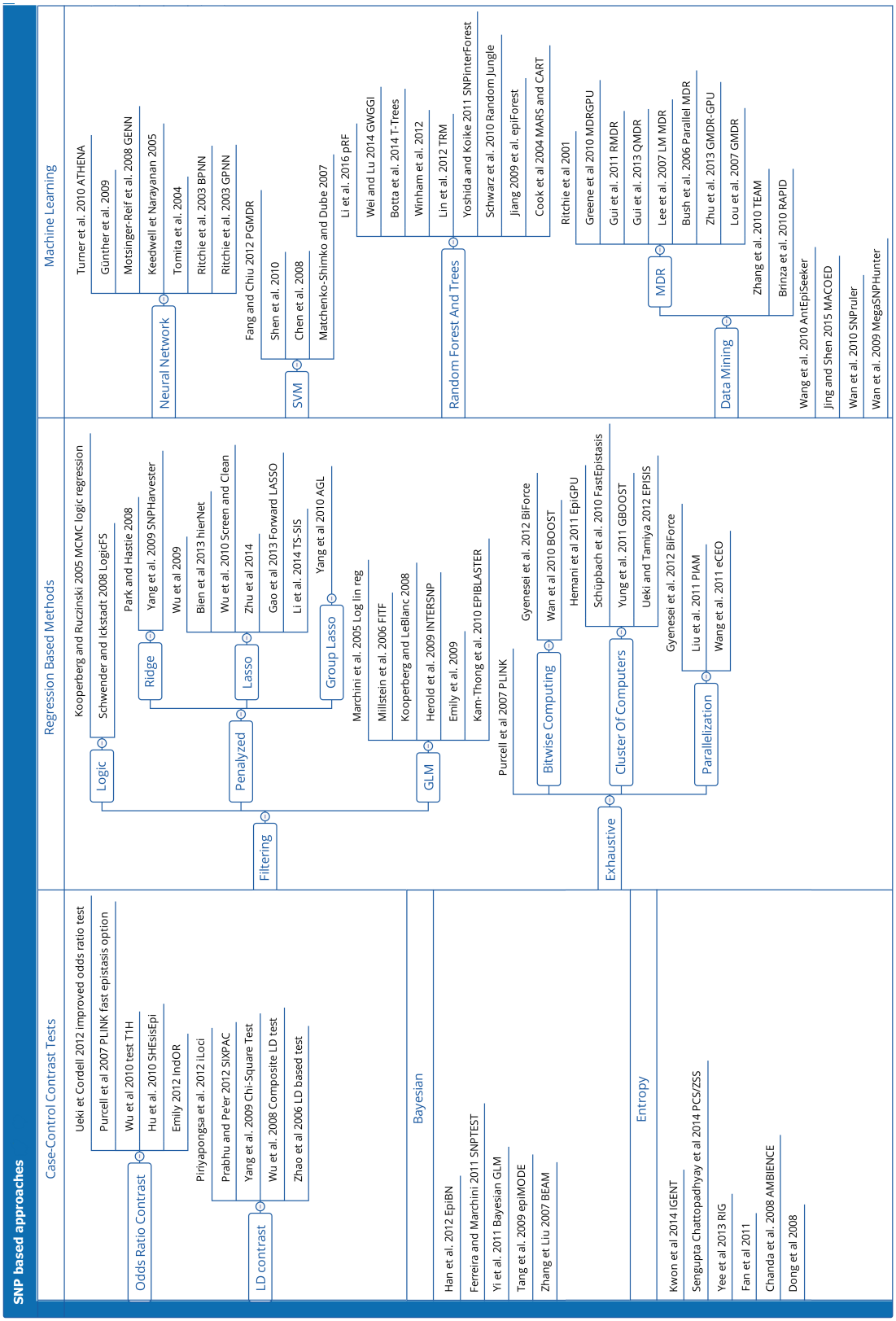


Figure 2.2: Classification of the presented methods for detecting epistasis between single markers

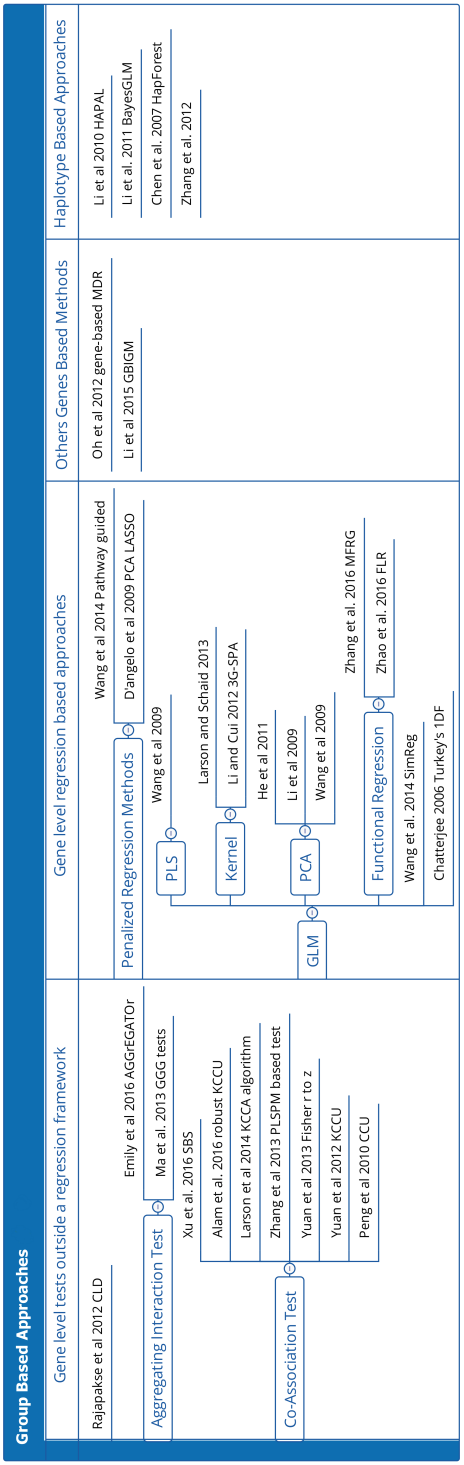


Figure 2.3: **Classification of the presented methods for detecting epistasis between groups**

penalty (LASSO) is used in the penalized model as it allows to jointly perform estimation and variable selection but several authors also propose to resort to a L_2 penalty (Ridge) combined to a variable selection procedure that allows to take into account colinearity among variables. Penalized methods also need to use a preliminary filtering step as a penalized regression model loses relevance when confronted to a too large number of variables. Besides traditional filtering strategies several authors [Gao et al., 2013; Li et al., 2014] proposed to select SNPs following a weak hierarchical assumption (in which only one predictor of a pairwize interaction needs to be marginally significant) which allows to capture epistasis effects for the SNPs that only contribute to the phenotypic trait through their combination with other genetic markers. Thus, a first step aims at identifying SNPs with significant marginal effects and later test them for possible interaction with all other genetic markers either they show marginal effect or not. However, other authors are more prudent with the weak hierarchical assumption and prefer to consider only SNPs showing main effects [Zhu et al., 2014]. Other methods chose to use their own filtering strategy as SNPHarvester [Yang et al., 2009a] that develops a PathSeeker algorithm that aims to identify disease-associated SNP groups in which SNP interactions are latter investigated using Ridge regression and variable selection. Logic regression is also used in the context of epistasis investigation [Schwender and Ickstadt, 2008; Kooperberg and Ruczinski, 2005]. In this framerwork, sets of variables are converted into Boolean expressions before being introduced in a regression model. These Boolean expressions can be represented as trees and are defined using classification rules.

Regarding exhaustive approaches, besides the initial parametric framework proposed in PLINK [Purcell et al., 2007] that we described previously, other methods have been developed to investigate epistasis in the whole genome. These approaches resort to different strategies in order to avoid as much as possible computational issues. For example, a bitwise computing in which SNP data are first transformed in a bitwise structure will allow a better memory efficiency and a better computing speed. This strategy has been used in approaches such as BOOST [Wan et al., 2010a] and BiForce [Gyenesei et al., 2012]. Other strategies consist in resorting to multiple processing cores unit as GPGPU (General-Purpose computation on Graphics Processing Units) or parallel computing that will both allow to accelerated computational performance and thus reduce the computational burden of a genome wide epistasis investigation.

Case-control contrast tests

Interaction between two genetic markers can be investigated using case control contrast tests. The methods that we present here are either based on a difference test between cases and controls in linkage disequilibrium (LD) or in

odds ratio.

Two unlinked markers in a population will present linkage equilibrium if they are independent. However, the existence of an interaction effect between two unlinked markers will lead to a non independence of the two markers in the disease population and will in turn create linkage disequilibrium. Following this idea Zhao et al. [2006] developed a test for interaction between unlinked markers based on the difference in LD between cases and controls. Others methods based on the same concept have later been proposed [Wu et al., 2008; Yang et al., 2009b], with some, as SIXPAC [Prabhu and Pe'er, 2012] or iLOCi [Piriyapongsa et al., 2012], designed to realize an exhaustive investigation of interactions.

Among the methods based on the difference in odds ratio we can cite the exhaustive fast epistasis option that has been implemented in PLINK [Purcell et al., 2007] in order to screen for epistasis in a faster way. Other methods based on this framework have been developed, for example, Wu et al. [2010c] proposed a pathway based genome wide interaction analysis with a new measure of interaction based on pseudohaplotype odds ratio. Latter, Ueki and Cordell [2012] investigated this method just as the fast epistasis option and proposed adjusted version of the statistics as well as two new statistical tests. Emily [2012] proposed an independence-based odds ratio (IndOR) statistic that relies on a biological definition of epistasis. They define epistasis as a variation in the level of dependency between two genetic markers and resort to odds ratio to evaluate this variation between cases and controls.

Machine Learning and data mining

An important number of methods based on machine learning or data mining have also been investigated as reviewed by Koo et al. [2013], Upstill-Goddard et al. [2012] or McKinney et al. [2006] Theoretical concept of some commonly used machine learning approaches are detailed in the section 2.2.3. Using this family of methods, a larger type of data structures can be modeled. Interaction terms for modeling data are not explicitly defined and the epistatic effect is most of the time tested implicitly by examining the combination effect simultaneously.

A popular data mining approach is the Multifactor Dimensionality Reduction (MDR) [Ritchie et al., 2001]. For a given set of genetic markers, n SNPs are selected. The possible multifactor combinations of these n markers are represented in cells in n-dimensional space and each cell is assigned a case-control ratio. Each multifactor cell is labelled as "high risk" if the case-control ratio exceeds some threshold and "low risk" otherwise. This classification allows to reduce the dimensionality of the predictors to one dimension. The new one-

dimensional variable is then evaluated for its ability to classify and predict disease status. The procedure is repeated for each possible n markers combination and the optimal n combination is chosen using cross-validation and permutation testing. A best model is chosen for each value of n. From the set of best models, the model which minimizes the prediction error and/or maximizes the cross-validation consistency is selected. MDR is now a reference in the epistasis detection field and multiple extensions have been proposed as reviewed by Gola et al. [2016]. For example, to make the method applicable to continuous outcome [Gui et al., 2013] or to boost the computing time performance with parallel implementation [Bush et al., 2006] or graphics processing unit [Zhu et al., 2013; Greene et al., 2010]. Others extensions are presented in Figure 2.2.

When using machine learning predictors or classifiers, the detection of epistasis can be considered as finding the best combination of SNPs from a given dataset which can produce the highest prediction accuracy [Chen et al., 2008]. Pattern recognition methods such as Support Vector Machines (SVM) or Neural Networks (NN) have been used in this context. Thus Shen et al. [2010] proposed a two stage method to detect gene-gene interactions that first uses a model selection method SVM with L_1 penalty to identify promising SNPs and interaction effects and then calculate the p-values associated using a logistic regression method. Chen et al. [2008] apply SVM with various kinds of combinatorial optimization methods to detect interactions. Regarding NN, one way to identify interactions is to use two-SNP NN models in which two locus are used as input. The best neural network model is chosen using an optimization criteria [Turner et al., 2010]. One particular challenge when using NN is the designation of an appropriate NN architecture and the use of classical NN based on backpropagation algorithms have shown variable success in genetic epidemiology [Motsinger-Reif et al., 2008]. In this context, machine learning algorithms such as genetic programming or grammatical evolution have been proposed to optimize NN architecture and have been used by several authors in order to detect non linear interactions among genetic markers. Thus Ritchie et al. [2003] develop the Genetic Programming Neural Network (GPNN) that resorts to genetic programming in order to optimises both the inputs and the architecture of NN for a given data set. The authors show that GPNN presents a better power in comparison with a traditional backpropagation neural network (BPNN). Motsinger-Reif et al. [2008] introduced Grammatical Evolution Neural Network (GENN) and show that their approach outperformed GPNN. More recently Turner et al. [2010] proposed ATHENA, a new tool that proposes several fundamental modifications to usual NN. The different strategies evaluated by the authors was to resort to an alternative tree-based GE crossover strategy, to use an hybrid backpropagation-GENN training algorithm and finally to determine if the incorporation of biological knowledge from external sources such as Biofilter [Bush et al., 2009] would increase performances. They

show that the three strategies increases performance of NN in detecting and modeling epistasis.

Random Forest (RF) can also be used to capture interactions between SNPs. RF allows for interactions between SNPs as it builds paths departing from the root of a tree through various nodes. A path can then be perceived as a sequence of predictor variables that includes potential interactions between them [Schwarz et al., 2010]. Thus, RF can detect associated SNPs from a set of genetic markers while implicitly taking into account the effects of interactions among them. RFs give variable importance measures (VIM) for each predictor that can be used to rank SNPs. These measures can be used to filter SNPs to investigate for interactions. However, Winham et al. [2012] explored the ability of VIMs to capture interaction effects and shown that these measure can fail to detect interaction effects in high-dimensional data sets and that the use of RF may be limited in the context of epistasis detection. Among the methods that have been proposed to detect interaction we can cite the software Random Jungle [Schwarz et al., 2010] which is a fast implementation of RF. More recently, Li et al. [2016] proposed to detect pure interactions between SNPs using permuted random forest (pRF). The method identifies top interacting SNP pairs by estimating how much power the pairwise interactions influence a random forest classification model. A more complete list of methods based on either Trees or RF can be found in Figure 2.2 as well as other machine learning approaches.

Other existing family of approaches

The different types of approaches that we have presented through this section are not exhaustive of all the family of methods that have been proposed to investigate epistasis in GWAS. Methods based on distinct foundations such as entropy or Bayesian modeling have also been developed in the past few years. However these families of methods exceed the interests of this manuscript and will not be further investigated. Several references of methods based on these two families can be found in Figure 2.2 and more information regarding entropy based methods can be found in the systematic review recently published by Ferrario and König [2016].

2.3.3 Groups interaction detecting methods

In Section 2.1.3 we presented the advantages to resort to a group based analysis in order to detect associations in GWAS along with several corresponding SNP set based approaches. As a reminder, this type of analysis presents the advantages to simplify the multiple testing problem, to present a better statistical power while allowing to take into account the correlation structure among

SNPs. A group approach analysis can also be used in the context of epistasis investigation. As in group association analysis, SNP sets to investigate for epistasis can be defined in various ways, for example using genes or haplotype blocks definition, but most of the methods like to resort to gene as grouping unit as they may be easier to define and should allow to obtain more biologically interpretable results.

Even if interactions between individual markers is the most investigated strategy, several gene-gene methods have been proposed as presented in Figure 2.3. Some existing approaches proposed at the SNP level have been extended in order to be able to consider sets of markers. For example Li et al. [2015] proposed a Gene Based Information Gain Model (GBIGM) as an extension of entropy based method, while Oh et al. [2012] extended the popular data mining approach, Multifactor Dimensionality Reduction (MDR) to detect gene interaction. Overall, most of the gene-gene based approaches adhere to one of the following strategies: some authors proposed to use a test at the gene level after aggregating SNP-SNP interaction tests, other proposed to use multidimensional approaches on the sets of genetic markers and either directly to define a test at the gene level or to resort to a regression model that incorporates summarizing variables.

Gene level interaction tests defined outside a regression framework

One strategy to investigate gene-gene interaction consists in aggregating into a gene level test the p-values that have been obtained at the SNP-SNP level. Among the methods that allow to combine the p-values of the SNPs within a gene we can cite: the truncated product method (tProd) [Zaykin et al., 2002] that removed p-values larger than a defined cutoff from consideration. The truncated tail strength test statistic (tTS) [Jiang et al., 2011] that follow the same idea. Li et al. [2011b] proposed GATES as an extended Simes test and Liu et al. [2010] proposed a minimum p-value test (minP) in the program VEGAS (versatile gene-based association study). Ma et al. [2013] used these four different methods to aggregate all pairwise p-values obtained for a couple of gene in a single interaction test and proposed four new gene-based gene-gene interaction (GGG) tests. In the same way, [Emily, 2016] proposed A Gene-based GEne-Gene interActTiOn test (AGGrEGATOr) based on a minP procedure.

Using another framework Rajapakse et al. [2012] developed the Composite Linkage Disequilibrium method (CLD) which is a gene-based test of interactions for case-control studies that takes into account the LD information by using a test based on the comparison of the covariance structure between cases and controls.

Other gene-based tests that focus on the difference between cases and controls have been developed. In this fashion, Peng et al. [2010] used a canonical

correlation-based U-statistic model (CCU) to detect co-association in case-control studies. The concept of co-association refers to the joint effect of two genes in the contribution to a trait. More precisely it refers to the extent to which this joint effect differs from the main effects of each gene on a trait. In co-association, this difference is not only due to the presence of interactions but also depends of the correlation between the genes and co-association is measured using correlation based approaches. For example, in CCU, the idea is to test for two given genes the difference between canonical correlation coefficient computed by Canonical Correlation Analysis (CCA) among cases and among controls. Their work was subsequently extended by Yuan et al. [2012] and Larson et al. [2014] who proposed a new test termed KCCU that is based on kernel CCA (KCCA) and allows to detect nonlinear correlation between genes. A robust version of this latter proposition, robust KCCU, was introduced by Alam et al. [2016]. In 2013, two other methods were proposed in the context of co-association detection that were not based on CCA. Zhang et al. [2013] proposed to use a Partial Least Squares Path Modeling approach (PLSPM) to detect co-association between genes. The test is defined on the standardized difference between cases and controls of path coefficient obtained using PLSPM for the gene pair. Yuan et al. [2013] proposed a co-association measure based on the difference of correlation coefficients between cases and controls for two genes and used a Fisher r-to-z transformation for testing the difference. Fisher r-to-z, CCU, KCCU and PLSPM-based statistics were proposed by the same group of authors that recently developed a novel score-based statistic (SBS) [Xu et al., 2016] that captures the effect of covariance between two genes and that presents good performance in comparison with the previous approaches.

Gene level regression based approaches

Regression based approaches resort to a summarizing step to obtain information at the gene level. In more recent methods, filters or penalized models are used to make the method applicable to a large number of genes, while older methods are only applicable to two or a very limited number of genes. For the summarizing step, most methods resort to a principal component (PC) approach, but each method has its specific characteristics.

Chatterjee et al. [2006] harnessed Tukey's one-degree-of-freedom method to investigate interaction between two genes. Their method is based on the assumption that the SNPs included in each gene region act as surrogates for an underlying biological phenotype. The genotypic information for the gene region is extracted as a single component by a weighted sum of all SNPs. The weights are determined according to the SNP's correlation with the trait. The product of the two sums is then introduced as the gene-gene interaction term into a logistic model, where marginal effects are represented by the respective sums. Building on this idea, Wang et al. compared two different interaction tests [Wang et al., 2009]. On the one hand, they used Principal Component

Analysis (PCA) to summarize SNP information within a gene, and on the other hand they used Partial Least Squares (PLS) to extract components that summarize, first, the information among SNPs in a gene and, second, the correlation between SNPs and the outcome of interest. They then proposed an interaction test based on either the first PC or the first PLS component for each gene, and were able to show that the PCA and PLS methods often outperformed Tukey's one-degree-of-freedom method. But it is worth noting that the main objective of these three methods was to improve the detection of associations in the presence of gene-gene interactions, rather than to identify the interactions themselves. Other approaches based on principal component analysis have since been proposed for epistasis detection [He et al., 2011; Li et al., 2009]. Li et al. [2009] proposed selecting, as the gene representation, PCs that are able to explain at least 80% of the variation. The PCA based approach proposed by He et al. [2011] allows to take LD information into account. LD information is used to weight genotype scores which are then aggregated using principal components. Other methods than PCA or PLS have been used to summarize genetic information in a gene. For example SimReg [Wang et al., 2014a] resorts to genetic similarity which, unlike PCA or PLS, utilizes all information within a gene. For each gene, the multi-marker information is summarized using a specific genetic similarity definition between subjects. The interaction is modeled by taking the product of the genetic similarities of the two genes and is then integrated in a regression model. More recently, Zhao et al. [2016] proposed a Functional Logistic Regression (FLR) model to detect interaction between two genomic regions. In this approach, functional data analysis techniques are used to reduce the dimensionality of the data and interaction between all possible pairs of variants in two regions are tested using an overall test in a logistic regression model. This model was also investigated by Zhang et al. [2016] for multivariate quantitative traits in a new methods termed Multiple Functional Regression (MFRG) model.

These methods have been defined on the consideration of one couple of genes but they can be used to explore the presence of interactions among a large number of genes using correction for multiple testing as done by Zhao et al. [2016] and Zhang et al. [2016]. The idea of considering a multiple number of genes to examine for interaction was also investigated by different authors who proposed different filtering based approaches that reduce the gene-gene search space by eliminating unimportant genes. Li and Cui [2012] proposed a model-based kernel machine method, termed Gene-centric Gene-Gene interaction with Smoothing-sPline ANOVA (3G-SPA), as a two-step procedures that first performs a search for gene pairs contributing to the overall phenotypic variations and then test significant pairs for interaction effects. In this approach, a smoothing spline-ANOVA decomposition method is used to decompose the main and interaction effects of two genes. The interaction is modeled by a cross-product of two kernel functions that captures the joint variation

of SNPs within a gene. The authors explain that the kernel machine method is mathematically equivalent to a linear mixed effects model, main and interaction effects are evaluated by testing the significance of different variance components. [Larson and Schaid, 2013] proposed an adaptation of 3G-SPA applicable to case-control studies. Another attractive alternative is offered by penalized regression methods that select a subset of important predictors out of a large number of potential predictors. These methods operate by shrinking the size of the coefficients. The coefficients of predictors with little or no apparent effect are forced to be set to zero, reducing the effective degrees of freedom and in many cases making model selection possible. A few approaches using penalized models have been proposed. D'Angelo et al. [2009] combined principal component analysis and LASSO penalized regression. Wang et al. [2014b] used a principal component analysis combined with an L1 penalty, with adaptive weights based on gene size, pathway support and effect size.

Haplotype-haplotype interaction investigation

Even if investigating epistasis at the haplotype level is a less common approach due to the issues brought by haplotype reconstruction, several haplotype based methods have been proposed. Let us cite for example Zhang et al. [2012] who proposed a method based on stratified contingency tables using a case-control genetic association design. HAPAL [Li et al., 2010], that resort to a penalized logistic regression framework with an adaptive LASSO penalty or even HapForest [Chen et al., 2007], a forest based approach. As last example, we can also cite BayesGLM a method based on a Bayesian hierarchical generalized linear model [Li et al., 2011a].

Chapter 3

A novel Group LASSO based approach to detect epistasis in GWAS

In this chapter we will present a new penalized regression framework aiming to detect interactions between groups of variables in the context of GWAS. We will first introduce the general regression model before describing the proposed approach in more details. Overall the proposed approach is composed of different steps. The first step consists in specifying a grouping of markers. Once groups are determined, new interaction variables can be defined for each possible couple and can be introduced in the regression model. The novelty of this approach lies in the definition of group interactions. We will present several manners to define those new variables using various statistical approaches. The last step of the method consist in estimating the coefficients of the regression model using a Group LASSO penalty.

Contents

3.1 Introduction

We thereafter present new group modeling approaches based on regression analysis to detect epistasis interaction in GWAS. As presented in the previous chapter, generalized linear model (GLM) is often used for GWAS: the phenotype is considered as a random variable $\mathbf{y}$ whose conditional expectation can be written as a function of covariates $\boldsymbol{X}$, usually the matrix of genetic markers.

For who is interested in investigating the presence of interactions between genetic markers, it is necessary to introduce a new set of variables $\boldsymbol{Z}$ characterizing those interactions and adding them in the GLM model as covariates as follows:

$$g(E[\mathbf{y}|(\boldsymbol{X},\boldsymbol{Z})]) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma}. \tag{3.1}$$

In this model, $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ denotes the vector of trait values for n individuals with $\mathbf{X} \in \mathbb{R}^{n\times p}$, the genotype matrix whose columns correspond to the p SNPs. $\mathbf{X}_i$ is then a p-dimensional vector of covariates for observation $i \in \{1,\dots,n\}$ and for $j \in \{1,\dots,p\}$. We usually consider an additive coding scheme in which the genotype value of each SNP j from individual i is denoted $X_{ij} \in \{1,2,3\}$. $\boldsymbol{Z}_i$ is the ith line of the matrix of interactions and $\boldsymbol{\gamma}$ a parameter vector of appropriate dimension.

In the context of a group analysis where SNPs are arranged together following a chosen aggregation strategy, we are no longer studying markers individually but as groups. The interaction variables will be then constructed in order to represent those groups interactions. Going back to the model, we consider the genetic variants in $\boldsymbol{X}$ as structured in G groups. Each group is described by a given number of SNPs p_g where $\sum_g p_g = p$. $\boldsymbol{X}^g$ denotes the submatrix of $\boldsymbol{X}$ whose columns are the p_g SNPs of group g. $\boldsymbol{\beta}$ is the coefficients vector of individual markers defined as:

$$\boldsymbol{\beta} = \left(\underbrace{\beta_{1,1}, \beta_{1,2}, ..., \beta_{1,p_1}}_{group_1}, ..., \underbrace{\beta_{G,1}, ..., \beta_{G,p_G}}_{group_G} \right)^T.$$

The main effect of each group is modeled through the sum of the effects of all corresponding SNPs. Concerning interaction effects, we compute new variables representing interactions between two specific groups. The matrix of interactions is thus structured into $G(G-1)/2$ submatrices:

$$\boldsymbol{Z} = [\boldsymbol{Z}^{11} \cdots \boldsymbol{Z}^{rs} \cdots \boldsymbol{Z}^{G(G-1)/2}]$$

where $\boldsymbol{Z}^{rs}$ describes the interactions between the two groups r and s. The parameter vector $\boldsymbol{\gamma}$ is accordingly structured into sub-vectors $\boldsymbol{\gamma}^{rs}$.

3.2 Groups modeling

As presented in Section 2.1.3, a common strategy to group SNP markers is to use previous biological knowledge. For a considerable part of genetic markers, it is possible to obtain information on pathways, haplotypes or genes in which it can belong and thus, to group together markers that share the same information. This type of biological information can be obtained using public databases. In GWAS grouping SNPs by genes is one of the most common strategy. In this thesis, we resort to the material provided by the National Center for Biotechnology Information (NCBI) in order to obtain gene information for each genetic marker and we will only retain the SNPs that belong to one unique gene. Another grouping strategy that would allow to keep a larger number of SNPs in the analysis would be to use LD block structure as proposed by Dehman et al. [2015]. This clustering approach uses a LD based distance to define blocks of strongly correlated genetic markers among the genome. However, one limit of this type of approaches is that the group structure is less interpretable than genes. Thus, we will mainly resort to a gene definition to determine the group structure of our data sets.

3.3 Modeling interactions

Once the groups are defined, the second step will consist to construct interaction variables. We will present some of the various modeling strategies using genes as groups, but any other grouping definition can be used.

3.3.1 General model

Let us consider two genes r and s described respectively by p_r and p_s SNPs. A first idea to represent the interaction between the two genes could be to consider all possible SNP products of both genes. With this idea, a possible interaction term describing the epistasis between the two genes could be:

$$\boldsymbol{Z}_i^{rsT}\boldsymbol{\gamma}^{rs} = \sum_{j=1}^{p_r}\sum_{k=1}^{p_s}\gamma_{jk}^{rs}X_{ij}^{r}X_{ik}^{s}. \tag{3.2}$$

We hereafter set $\boldsymbol{F}^{rs} = \{X_{ij}^{r}X_{ik}^{s}\}_{i=1\cdots n}^{j=1,\cdots,p_r;k=1,\cdots,p_s}$, the matrix of all pairwise SNPs products. In this case the submatrix of interactions is $\boldsymbol{Z}^{rs} = \boldsymbol{F}^{rs}$ and $\boldsymbol{\gamma}^{rs} = \{\gamma_{jk}^{rs}\}$ is a vector of size p_rp_s. However, this strategy would lead to the creation of a too extensive number of interaction variables to add in the model. For example, if we consider that the genes r and s are both composed of 6 markers, we will expect 36 variables in $\boldsymbol{Z}^{rs}$. Knowing that the data set surely contains more than two genes, the number of parameters in the model would obviously be too large to be reliably estimated. For this reason it is necessary to find other representative definitions that allow to generate a reduced

number of interaction variables for each couple and thus to reduce the number of parameters to estimate in the model.

Thereafter we will present different methods to compute interaction variables while controlling the dimension of $\boldsymbol{Z}^{rs}$ for each couple (r, s).

3.3.2 Interaction variables construction

In order to compute interaction variable matrices of low dimension, we choose to use and compare different alternatives. The first following propositions are based on various statistical methodologies as dimensionality reduction methods or machine learning which have been more widely presented in Section 2.1. The last one is an original approach that we termed Gene-Gene Eigen-Epistasis.

Principal Component Analysis

Applying PCA on each gene can allow to reduce the number of SNPs to a fixed number of representative components. Considering gene r described by p_r SNPs, we can compute the matrix $\mathbf{T}^r_{n,q}$ of the first q principal components

$$\mathbf{T}^r = \mathbf{X}^r \mathbf{W}^r,$$

where $\mathbf{W}^r$ is a $p_r \times q$ matrix of loading vectors. Using $\mathbf{T}^r$ and $\mathbf{T}^s$ instead of $\mathbf{X}^r$ and $\mathbf{X}^s$ in the computation of the interaction allows the number of parameters relative to each interaction to be controlled. This control is achieved by choosing the number of principal components q. The PCA model that we describe draws upon ideas in [Zhang and Wagener, 2008]. The interaction term takes the form

$$\mathbf{Z}_i^{rsT} \boldsymbol{\gamma}^{rs} = \sum_{j=1}^{q} \sum_{k=1}^{q} \gamma_{jk}^{rs} T_{ij}^r T_{ik}^s.$$

Relating this expression to the general form of the interaction term $\mathbf{F}_i^{rs}$ described above, we can see that performing PCA prior to computing the interactions is a means of constraining the linear interaction term of Equation (3.2).
The submatrix of interactions is $\mathbf{Z}^{rs} = \{T_{ij}^r T_{ik}^s\}_{i=1\cdots n}^{j=1,\cdots,q;k=1,\cdots,q}$, and $\boldsymbol{\gamma}^{rs} = \{\gamma_{jk}^{rs}\}$ is a vector of size q^2 describing the interaction between genes r and s. In particular, if a single principal component is chosen ($q = 1$), there will be only one parameter to estimate per interaction.

Partial Least Squares

Wang et al. [2009] proposed an alternative method for integrating interactions between genes using a PLS approach. Let $(\mathbf{X}^r, \mathbf{X}^s)$ be the genotypic matrix for the given pair of genes (r, s). Their approach computes the components that maximize $\text{cov}^2(\mathbf{X}^r\mathbf{w}, \mathbf{Lq})$, with $\mathbf{L} = (\mathbf{y}, \mathbf{X}^s)$ and $(\mathbf{w}, \mathbf{q})$ the weight vectors. The

interaction of a couple of genes (r, s) is then represented by the first q latent variables $\mathbf{T}^r_{n,q}$:

$$\mathbf{Z}_i^{rsT}\boldsymbol{\gamma}^{rs} = \sum_{j=1}^{q} \gamma_j^{rs} T_{ij}^{rs}.$$

In this approach phenotypic information is retained when the interaction variables are constructed.

Canonical Correlation Analysis

Some methods based on CCA have been developed to detect interactions such as the canonical correlation-based U statistic (CCU) [Peng et al., 2010] and the kernel versions of CCU [Larson et al., 2014; Yuan et al., 2012]. These methods also consider the phenotype in their construction, but can be applied only to case-control designs.
In our setting we propose to use CCA in the following way: for two genes r and s, we define new variables $\mathbf{U}^r$ and $\mathbf{V}^s$ which are linear combinations of the original variables $\mathbf{X}^r$ and $\mathbf{X}^s$:

$$\begin{cases} \mathbf{U}^r = \mathbf{X}^r\mathbf{W}^r, \\ \mathbf{V}^s = \mathbf{X}^s\mathbf{Q}^s \end{cases}$$

where $\mathbf{W}^r$, $\mathbf{Q}^s$ are the matrices whose columns define the weight vectors, which are solution of the CCA. We propose to code the interaction of a couple of genes (r, s) by the first q component couples of a CCA:

$$\mathbf{Z}_i^{rsT}\boldsymbol{\gamma}^{rs} = \sum_{j=1}^{q} \gamma_j^{rs} U_{ij}^r V_{ij}^s.$$

Machine Learning

We also investigate the use of machine learning approaches to compute interaction variables. For each couple of genes (r, s), we define one model using the corresponding SNPs and the phenotype and use the predicted phenotype as an interaction variable. Predictions are obtained using either random forest, boosting, support vector machine or neural network as presented bellow.

- *Random Forest*
  A random forest is constructed on a train sample of the data ($\mathbf{X}^r$, $\mathbf{X}^s$, $\mathbf{y}$) for each couple of genes. 500 bootstrap samples are used to generate the trees of a random forest. At each occurring split in a tree, m random predictors are used to generate the split. If the phenotype $\mathbf{y}$ is continuous $m = \frac{p_r + p_s}{3}$ and if $\mathbf{y}$ is categorical $m = \sqrt{p_r + p_s}$ with p_r the number of

SNPs in gene r and p_s the number of SNPs in genes s. Alternatively, we can also construct a random forest for each couple using the matrix $\boldsymbol{F}^{rs}$, representing all pairwise SNPs products between the two genes r and s, rather than the simple SNPs matrix $[\mathbf{X}^r, \mathbf{X}^s]$.

- *Boosting*
  When using boosted regression trees, for each couple of genes r and s, 2000 trees are generated with a maximum of 4 splits for each tree. For continuous outcomes, the loss function chosen is the squared error $(y_i - f(\mathbf{X}_i))^2$ whereas for binary outcomes we chose the AdaBoost exponential loss function $\exp[-(2y_i - 1)f(\mathbf{X}_i)]$ or the logistic loss function $-2(y_i f(\mathbf{X}_i) - \log[1 + \exp[f(\mathbf{X}_i)]])$.

- *Support Vector Machine*
  Using a support vector machine algorithm, we first consider the data $(\mathbf{X}^r, \mathbf{X}^s, \mathbf{y})$. The model is then fit on a train sample of this data using a specific kernel. Among the available kernels we chose to consider linear kernel, polynomial kernel of degree $d = 3$ or $d = 5$ and radial basis kernel.

- *Neural Networks*
  Finally, we investigate the use of neural networks to define interaction variables. We consider as input layer all the SNPs contained in the genes r and s. We define one hidden layer with k neurons and we fix $k = 2$ to correspond to the number of genes in the neural network.

Gene-Gene Eigen-Epistasis

We propose an original approach for modeling interactions. The general idea is to consider the interaction variable between the two genes r and s as a function $f_{\mathbf{u}}(\mathbf{X}^r, \mathbf{X}^s)$ parameterized by $\mathbf{u}$.
Ideally, we would like to choose a function that meets some of the following conditions:

- The interaction variable between genes r and s brings information which is different from the information brought by the genes themselves. From an algebraic point of view, this could be stated as $f_{\mathbf{u}}(\mathbf{X}^r, \mathbf{X}^s)$ belongs to the orthogonal subspace spanned by the columns of $\mathbf{X}^r$ and $\mathbf{X}^s$.
- The interaction variable brings maximal information about the phenotype, which we could formulate as $f_{\mathbf{u}}(\mathbf{X}^r, \mathbf{X}^s)$ and the phenotype have a maximum covariance.

With these considerations, we can propose two ways to estimate $\mathbf{u}$. The first way is to minimize the covariance between the interaction function and the genotype and the second to maximize the covariance between the interaction function and the phenotype.

If we consider the function f to be linear, both problems become easily tractable and each one has only one solution. Setting

$$\mathbf{Z}^{rs} = f_{\mathbf{u}}(\mathbf{X}^r, \mathbf{X}^s) = \mathbf{F}^{rs}\mathbf{u},$$

where $\mathbf{F}^{rs} = \{X^r_{ij} X^s_{ik}\}_{i=1\cdots n}^{j=1\cdots,p_r;k=1,\cdots,p_s}$ and $\mathbf{u} \in \mathbb{R}^{p_r p_s}$ we obtain, depending on the chosen criterion, the two following problems:

- *Minimization of the correlation with the genotype (G-GEE$_{\text{c1}}$)*

  $$\hat{\mathbf{u}} = arg \min_{\mathbf{u},\|\mathbf{u}\|=1} \hat{\text{cov}}^2(\mathbf{X}, f_{\mathbf{u}}(\mathbf{X}^r, \mathbf{X}^s)),$$

  with the genotype matrix $\mathbf{X} = (\mathbf{X}^r, \mathbf{X}^s)$ that corresponds to the SNPs of both genes r and s,

  $$\min_{\mathbf{u},\|\mathbf{u}\|=1} ||\hat{\text{cov}}[\mathbf{F}^{rs}\mathbf{u}, \mathbf{X}]||^2 = \min_{\mathbf{u},\|\mathbf{u}\|=1} ||\mathbf{u}^T\mathbf{F}^{rsT}\mathbf{X}||^2 = \min_{\mathbf{u},\|\mathbf{u}\|=1} \mathbf{u}^T\mathbf{F}^{rsT}\mathbf{X}\mathbf{X}^T\mathbf{F}^{rs}\mathbf{u} \quad .$$

  In this first case, the solution $\mathbf{u}$ is the eigenvector corresponding to the smallest eigenvalue of the matrix $\mathbf{F}^{rsT}\mathbf{X}\mathbf{X}^T\mathbf{F}^{rs}$. We then use the projection of the matrix $\mathbf{F}^{rs}$ on $\mathbf{u}$ as the interaction variable $\mathbf{Z}^{rs} = \mathbf{F}^{rs}\mathbf{u}$.

- *Maximization of the correlation with the phenotype (G-GEE$_{\text{c2}}$)*

  $$\hat{\mathbf{u}} = arg \max_{\mathbf{u},\|\mathbf{u}\|=1} \hat{\text{cov}}^2(\mathbf{y}, f_{\mathbf{u}}(\mathbf{X}^r, \mathbf{X}^s)),$$

  with:

  $$\max_{\mathbf{u},\|\mathbf{u}\|=1} ||\hat{\text{cov}}[\mathbf{F}^{rs}\mathbf{u}, \mathbf{y}]||^2 = \max_{\mathbf{u},\|\mathbf{u}\|=1} ||\mathbf{u}^T\mathbf{F}^{rsT}\mathbf{y}||^2 = \max_{\mathbf{u},\|\mathbf{u}\|=1} \mathbf{u}^T\mathbf{F}^{rsT}\mathbf{y}\mathbf{y}^T\mathbf{F}^{rs}\mathbf{u} \quad .$$

  The solution $\mathbf{u}$ is the eigenvector corresponding to the largest eigenvalue of the matrix $\mathbf{B} = \mathbf{F}^{rsT}\mathbf{y}\mathbf{y}^T\mathbf{F}^{rs}$, this eigenvector being the vector $\mathbf{F}^{rsT}\mathbf{y}$. Indeed, if we note $\mathbf{u} = \mathbf{F}^{rsT}\mathbf{y}$, then $\mathbf{B} = \mathbf{u}\mathbf{u}^T$ and we can easily see that only one eigenvalue λ exist for $\mathbf{B}$:

  $$\mathbf{B}\mathbf{u} = \lambda\mathbf{u}$$
  $$\mathbf{u}\mathbf{u}^T\mathbf{u} = \lambda\mathbf{u}$$

  with $\mathbf{u}^T\mathbf{u}$ a scalar that define λ. The complexity of computing $\mathbf{u}$ is therefore in $O(np_rp_s)$. We then use the projection of the matrix $\mathbf{F}^{rs}$ on $\mathbf{u}$ as the interaction variable $\mathbf{Z}^{rs} = \mathbf{F}^{rs}\mathbf{u}$.

With this second approach, the resulting Eigen-Epistasis vector $\mathbf{Z}$ is the linear combination of all the SNP-SNP interactions being the most correlated with the phenotype. The vector $\mathbf{u}$ can be interpreted as a weight vector defining the importance of each pairwise interaction. In its construction, G-GEE$_{c2}$ has similarities with PLS. The main difference lies in the original design matrix. PLS searches for components that maximize $\hat{\text{cov}}^2(\mathbf{X}^r\mathbf{u}, \mathbf{y}\mathbf{X}^s v)$, whereas G-GEE$_{c2}$ retains the component that maximizes $\hat{\text{cov}}^2(\mathbf{y}, \mathbf{F}^{rs}\mathbf{u})$, with $\mathbf{F}^{rs}$ the matrix of all pairwise interactions between the two genes r and s. Like PLS, G-GEE$_{c2}$ takes phenotypic information into account in the construction of the interaction variables.

3.4 Estimation of coefficients

Once the groups are defined and the interaction variables created, we are now interested in estimating the parameters of the model (3.1). In the context of large data sets with group structure, we choose to use a Group LASSO approach [Yuan and Lin, 2006]. A group includes either the SNPs of a given gene, or interaction terms relative to a given gene-pair interaction. In the particular case of linear regression, the model parameters are estimated by:

$$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\text{argmin}} \left(\sum_i (y_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\boldsymbol{\gamma})^2 + \lambda \left[\sum_g \sqrt{p_g} ||\boldsymbol{\beta}^g||_2 + \sum_{rs} \sqrt{p_r p_s} ||\boldsymbol{\gamma}^{rs}||_2 \right] \right) \quad , \tag{3.3}$$

whereas in the case of logistic regression:

$$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\text{argmin}} \left(\sum_i -\log L(y_i; \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}) + \lambda \left[\sum_g \sqrt{p_g} ||\boldsymbol{\beta}^g||_2 + \sum_{rs} \sqrt{p_r p_s} ||\boldsymbol{\gamma}^{rs}||_2 \right] \right) \quad , \tag{3.4}$$

In both cases the parameter λ is selected by cross-validation.

The use of Group LASSO will allow to obtain a list of significant groups whose coefficients are different from zero. However, as explained in Section 2.2.1, it is difficult to determine the relative importance of each of the selected groups. In order to improve estimation accuracy and to obtain p-values for each selected group, we will resort to the adaptive ridge cleaning approach proposed by Bécu et al. [2017] which is briefly described below.

Setting $\boldsymbol{H\theta} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma}$, and $J(\boldsymbol{\theta})$ as either $J(\boldsymbol{\theta}) = \sum_i (y_i - \mathbf{H}_i\boldsymbol{\theta})^2$ or $J(\boldsymbol{\theta}) = \sum_i -\log L(y_i; \mathbf{H}_i\boldsymbol{\theta})$ depending on $\mathbf{y}$ considered, the Group LASSO presented in the equations 3.3 and 3.4 can be written as:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \left(J(\boldsymbol{\theta}) + \lambda \left[\sum_g \sqrt{p_g} ||\boldsymbol{\theta}^g||_2 \right] \right)$$

with g representing either a gene or interaction group.

The screen and clean procedure proposed by Bécu et al. [2017] is a two-stage method. The Group LASSO model is first fitted on half of the data during the screening stage. The set of selected groups is the support $\hat{S}$. Then, the coefficients of these candidate groups are introduced into a ridge regression model fitted on the second half of the data with a specific penalty that allows to take into account the group structure:

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \ ; \ \theta_j=0 \text{ if } j \notin \hat{S}}{\operatorname{argmin}} \left(J(\boldsymbol{\theta}) + \mu \left[\sum_g \sum_{j \in g} \frac{\lambda \sqrt{p_g}}{||\hat{\boldsymbol{\theta}}^g||_2} \theta_j^2 \right] \right) \quad .$$

with g representing either a gene or interaction group. In this equation $\hat{\boldsymbol{\theta}}^g$ represents the regression coefficients obtained by the Group LASSO with penalty parameter λ.

We can see that the cleaning stage retains the magnitude of the coefficients $\hat{\boldsymbol{\theta}}^g$ estimated in the screening stage thus encouraging a larger penalization of the small coefficients over the larger ones. This second stage will allow to select groups of variables among the ones that have passed the first screening stage. For each of these groups, the significance of the regression coefficients is estimated using a Fisher statistic as described below.

Let

$$F_g = \frac{\sum_i (y_i - \hat{y}_i^{\omega})^2 - \sum_i (y_i - \hat{y}_i^{\Omega})^2}{\sum_i (y_i - \hat{y}_i^{\Omega})^2},$$

the F-statistic for the group g with $\hat{y}^{\Omega}$ the predicted values using all groups $g \in \hat{S}$ and $\hat{y}^{\omega}$ the predicted values obtained without the group g. As the usual F-test is not exact for Ridge regression, Bécu et al. [2017] proposed to estimate the distribution of the F-statistic under the null hypothesis using the matrix $\boldsymbol{H}$ with permuted elements of the columns corresponding to group g :

$$F_g^* = \frac{\sum_i (y_i - \hat{y}_i^{\omega})^2 - \sum_i (y_i - \hat{y}_i^{\Omega*})^2}{\sum_i (y_i - \hat{y}_i^{\Omega*})^2},$$

with $\hat{y}^{\Omega*}$ the predicted values calculated using all groups $g \in \hat{S}$ on the matrix $\boldsymbol{H}^*$ of permuted values for columns corresponding to group g. Empirical p-values are then obtained for each group as $p_g = \frac{1}{B} \#\{F_g \leq F_g^*\}$ with B the number of permutations.

The p-values obtained for each group are then corrected for multiple testing by the use of Benjamini and Hochberg procedure.

Chapter 4

Evaluation and application of interaction modeling

In this chapter we will present simulation studies that aim to compare the different modeling strategies described in Chapter 3. In the first part of this chapter we will introduce the various designs used to generate both genotypes and phenotypes. The second and third sections will focus on the presentation of the obtained results. We will finish this chapter by an application of the most promising methods on two real datasets related to Ankylosing spondylitis and Crohn's disease in order to investigate the presence of gene interactions.

Contents

4.1 Genotype and phenotypes designs for simulation studies

Different types of data sets will be considered. First, regarding the genotype, we will differentiate cases where data are completely generated (and where all parameters can be controlled) from cases where realistic data are used. The latter allow to consider more complex and real genetic structures. Simulated genotypes will be used in Section 4.3.1 to define the most meaningful parameters to use for properly comparing gene epistasis detection methods. The set of parameters that will be investigated are the following: size of the data set, values of the coefficients allocated to marginal and interacting effects respectively, correlation among SNPs in a gene, minor allele frequencies (MAF) of causal SNPs, values of the coefficient of determination R^2, number of components to construct interaction variables when using variable reduction based methods but equally the type of genetic effects considered and the presence or absence of the corresponding marginal effect for each simulated interaction. Methods will then be evaluated using either simulated genotypes or real data. In addition to the type of genotype considered, the simulation of phenotypes will vary according to the simulation model considered but also according to the type of outcome (continuous or binary).

4.1.1 Genotypes

Simulated Genotypes

Genotype simulation design was adapted from the model used in [Wu et al., 2009] with an extension to control the MAF of each SNP. The n lines of the genotype matrix are an i.i.d. sample from a multivariate random vector $\boldsymbol{X}_i \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. The correlation matrix $\boldsymbol{\Sigma}$ is block diagonal, each block corresponding to a gene. The correlation between two variables belonging to the same gene is ρ while all other correlations are null. To each SNP (column of the genotype matrix) is randomly assigned a MAF p from a uniform distribution between 0.05 and 0.5. For each causal SNP, the MAF is changed to a fixed value (0.1, 0.2 or 0.4). The genotype frequencies derived from the Hardy-Weinberg equation are then used to discretize X_{ik} values to 0, 1 or 2. In practice, X_{ik} is set to 0 if $X_{ik} > q_{(2p-p^2);N(0,1)}$, X_{ik} is set to 2 if $X_{ik} < q_{(p^2);N(0,1)}$ and X_{ik} is set to 1 otherwise.

Genotypes from real data set

We will consider genotypes coming from a real GWAS data set of the International Genetics of Ankylosing Spondylitis (IGAS) study [Cortes et al., 2013] containing 116,513 SNPs for 763 subjects (more details are given in Section 4.4.2). For each SNP we obtained gene affiliation using the NCBI2R package

[Melville, 2012] which annotates lists of SNPs with current information from NCBI. We considered only SNPs located within a single gene in order to form groups of genes without overlap. We identified 7,216 genes in the data set for a total number of 63,340 SNPs. At each iteration we will focus on a set of six genes randomly sampled from the 7,216 ones. The size of each data set will thus vary depending on the sampled genes.

4.1.2 Phenotype simulation

Different schemes can be used to generate phenotype vectors. In a first phase we will focus on the generation of continuous phenotype vectors with a particular consideration for the error term generation. We will then detail how the phenotype simulation models can be adapted to binary output.

"Wang Pathway" model

A first scheme corresponds to the model proposed by Wang et al. [Wang et al., 2014b] (which we will refer to hereafter as the "Wang Pathway" model):

$$y_i = \beta_0 + \sum_g \beta_g \left(\sum_{k \in \mathcal{C}} X_{ik}^g \right) + \sum_{rs} \gamma_{rs} \left(\sum_{(j,k) \in \mathcal{C}^2} X_{ij}^r X_{ik}^s \right) + \epsilon_i, \tag{4.1}$$

where $\mathcal{C}$ and $\mathcal{C}^2$ are respectively the set of causal SNPs and causal interactions, and ϵ_i a random Gaussian variable. For each causal gene g, we set a specific number of causal SNPs and a coefficient β_g is assigned to the standardized sum of these causal SNPs. In the same way, for the interactions, all the causal SNPs from a causal pair (r, s) are pairwise multiplied and a coefficient γ_{rs} is assigned to the standardized sum of the product.

PCA model

A second scheme for simulating phenotypes is based on the following model:

$$y_i = \beta_0 + \sum_g \beta_g \left(\sum_{k \in \mathcal{C}} X_{ik}^g \right) + \sum_{rs} \gamma_{rs} \left(\sum_{(j,k) \in \mathcal{C}^2} C_{ij}^r C_{ik}^s \right) + \epsilon_i. \tag{4.2}$$

The difference with the first model concerns the simulation of the interaction effects. In the second model the interaction effect for a causal pair (r, s) is defined as the product of the first PCA component $\boldsymbol{C}_{.1}^r$ of gene r and the first PCA component $\boldsymbol{C}_{.1}^s$ of gene s.

Error term

In both models, ϵ_i is generated independently from a $\mathcal{N}(0, \sigma^2)$, with σ^2 determined from the coefficient of determination R^2 that calibrates the strength of

the association. Both simulation models can be written as

$$y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \boldsymbol{\gamma} + \epsilon_i$$

where $\boldsymbol{X}$ is the marginal effect genotype matrix and $\boldsymbol{Z}$ the interaction effect matrix.
Let us denote $\boldsymbol{H\theta} = [\boldsymbol{X}, \boldsymbol{Z}] \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix}$ and

$$\begin{aligned} R^2 &= \frac{\sum(\boldsymbol{H_i\theta} - \bar{y})^2}{\sum(\boldsymbol{H_i\theta} + \epsilon_i - \bar{y})^2} \\ &= \frac{\sum(\boldsymbol{H_i\theta} - \bar{y})^2}{\sum(\boldsymbol{H_i\theta} - \bar{y})^2 + \sum \epsilon_i^2 + \sum 2(\epsilon_i(\boldsymbol{H_i\theta} - \bar{y}))} \\ &= \frac{\sum(\boldsymbol{H_i\theta} - \bar{y})^2}{\sum(\boldsymbol{H_i\theta} - \bar{y})^2 + n\ \hat{\text{var}}(\epsilon_i) + 2n\ \hat{\text{cov}}(\epsilon_i, \boldsymbol{H_i\theta} - \bar{y})}. \end{aligned}$$

We remark that:

$$\begin{aligned} 2n\ \text{cov}(\epsilon_i, \boldsymbol{H_i\theta} - \bar{y}) &= 2n\ \text{cov}(\epsilon_i, \boldsymbol{H_i\theta} - \frac{\sum_j y_j}{n}) \\ &= 2n\ \text{cov}(\epsilon_i, \boldsymbol{H_i\theta}) - \sum_j \frac{2n}{n} \text{cov}(\epsilon_i, y_j) \\ &= 0 - 2\ \text{cov}(\epsilon_i, \epsilon_i) = -2\sigma^2 \end{aligned}$$

Thus, replacing $\hat{\text{var}}(\epsilon_i)$ by σ^2, and $\hat{\text{cov}}(\epsilon_i, \boldsymbol{H_i\theta} - \bar{y})$ by $-\sigma^2/n$, we obtain

$$R^2 \approx \frac{\sum(\boldsymbol{H_i\theta} - \bar{y})^2}{\sum(\boldsymbol{H_i\theta} - \bar{y})^2 + n\sigma^2 - 2\sigma^2}.$$

This relation between R^2 and σ^2 gives us an expression for σ^2 that depends on R^2,

$$\sigma^2 = \frac{(1 - R^2)\sum(\boldsymbol{H_i\theta} - \bar{y})^2}{R^2(n-2)}.$$

We looked at how much of the coefficient of determination R^2 is explained by main effects, and how much is explained by interaction effects, in order to determine their respective roles in the model.

For a similar reason, when simulating phenotypes, Wang et al. [2014b] examined how much of partial R^2 was due to interaction effects. They selected coefficient values so that 30% of the partial R^2 was explained by interaction effects. Li and Cui [2012] did not use the R^2 directly, but they simulated data assuming different proportions of interaction effects among the total genetic variance. In our study, once the phenotype y had been set for each simulated design matrix, we computed how much of the R^2 could be attributed to interactions and main effects as $p_I = \frac{R_I^2}{R_T^2}$ and $p_M = \frac{R_M^2}{R_T^2}$ respectively, with R_I^2

the R-square value for the model containing only simulated interaction effects, R^2_M the R-square value where there were only simulated main effects, and R^2_T R-square value where there were both simulated main effects and simulated interaction effects. As the contribution of each type of effect may not be distinguishable, the sum of p_I and p_M may not be equal to one.

Binary phenotype

Both models can be adapted to binary phenotypes by using a logit link function,

$$\ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{X}_i^T\boldsymbol{\beta} + \boldsymbol{Z}_i^T\boldsymbol{\gamma},$$

where $\boldsymbol{X}$ is the marginal effect genotype matrix, $\boldsymbol{Z}$ the interaction effect matrix. The phenotype, case = 1 or control = 0, is defined using a Bernouilli distribution with the probability p_i for individual i to be a case.

Using this simulation model, it is not directly possible to calibrate the strength of the association between the covariates and the phenotype as well as the error term does not appear clearly in the model. It is however possible to examine the respective roles of main and interaction effects in the model using a Pseudo R-square value adapted to logistic regression such as the McFadden's R-square R^2_{McF} which is defined as

$$R^2_{McF} = 1 - \frac{\text{Log}(L_m)}{\text{Log}(L_0)}$$

Here, L_m denote the maximized likelihood value from the current fitted model, and L_0 denotes the corresponding value for the null model with only an intercept and no covariate.
Thus, it is possible to compute p_I and p_M using McFadden's R-square, that is $p_I = \frac{R^2_{McF_I}}{R^2_{McF_T}}$ and $p_M = \frac{R^2_{McF_M}}{R^2_{McF_T}}$ and to evaluate the part of R^2_{McF} respectively due to interactions and main effects. As previously, p_I and p_M may not be equal to one because of interdependence of interaction and marginal effects.

4.1.3 Evaluation criteria

For all simulation studies, power and false positive rates for each type of effects will be estimated by:

- Power: $\frac{1}{n}\sum_{i=1}^{n} \frac{\text{TP}_i}{\text{number of simulated effects}}$,
- False Positive Rates: $\frac{1}{n}\sum_{i=1}^{n} \frac{\text{FP}_i}{(\text{TP}_i + \text{FP}_i)}$

with n the total number of iterations, TP_i the number of true discoveries and FP_i the number of false discoveries obtained at the iteration i. The mean of the rates obtained over all iterations are then computed to estimate the false positive rate of each simulation setting.

4.2 Comparison of Group LASSO and univariate models

In the following we will examine the power of the Group LASSO (presented in Section 3.4) to detect interactions in comparison with two univariate estimation models.

The simulation study is conducted with continuous phenotypes and simulated genotypes. Each simulated data set is composed of 600 subjects and 8 genes, each of them containing 6 SNPs. Markers among the same gene are correlated at level $\rho = 0.8$, two markers in each gene are considered as causal with a MAF value of 0.2. Continuous phenotype are obtained using the "Wang Pathway" model (Section 4.1.2) with an identical value, fixed to 2, for coefficients β and γ. Interaction variables, one by couple of genes, are obtained using the G-GEE approach.

When using the Group LASSO, the significance of interaction variables is tested under the following model:

$$\mathbf{y} = \sum_{g=1}^{G} \sum_{j=1}^{p_g} \beta_{g,j} \boldsymbol{X}_j^g + \sum_{r,s} \boldsymbol{\gamma}_{r,s} \boldsymbol{Z}^{r,s} + \boldsymbol{\epsilon}.$$

The coefficients are estimated using Group LASSO with one group defined for each gene g among G and one group defined for each interaction variable of each couple of genes (r, s). The p-values for each group are obtained using an adaptive ridge cleaning as explained in Section 3.4.

In the univariate models, every couple of genes (r, s) and their interactions is tested for an association with the phenotype. We considered two different univariate models:

- LM1:
$$\mathbf{y} = \boldsymbol{\gamma}_{r,s} \boldsymbol{Z}^{r,s} + \boldsymbol{\epsilon},$$
- LM2:
$$\mathbf{y} = \sum_{j=1}^{p_r} \beta_{r,j} \boldsymbol{X}_j^r + \sum_{j=1}^{p_s} \beta_{s,j} \boldsymbol{X}_j^s + \boldsymbol{\gamma}_{r,s} \boldsymbol{Z}^{r,s} + \boldsymbol{\epsilon},$$

In LM1, no main effect is considered, the model is only constituted of the interaction variable $\boldsymbol{Z}^{r,s}$. In LM2, the main effects of each gene are considered with the interaction effect by the presence of all their respective SNPs, a coefficient β is attributed to each marker. As phenotype are continuous, coefficients are estimated using a multivariate regression model. The p-values corresponding to interaction variables are adjusted for multiplicity with Benjamini and Hochberg method.

Four different scenarios are considered:

Id	Main effects	Interaction effects
1		Gene1 x Gene2 Gene3 x Gene4
2	Gene1, Gene2, Gene3, Gene4	Gene1 x Gene2 Gene3 x Gene4
3	Gene1, Gene2, Gene3, Gene4	
4	Gene1, Gene2	Gene3 x Gene4 Gene5 x Gene6

In the first setting, only two interactions are generated, the first one between the two first genes and the second between the third and fourth genes of the data set. The second setting is similar to the first one excepted that the four genes also have main effect. In the third setting, only main effects are generated for the fourth first genes. The last setting combines the presence of main effects (two first genes) and different interaction effects (between the third and fourth genes and between the fifth and sixth genes of the data set).

Results obtained for each model under the four scenarios are presented in Figure 4.1. Overall, the estimations of LM2 model and the Group LASSO are the ones showing the less false discoveries, with a better power to detect the interaction effects for the Group LASSO model. However, the two models have more difficulties to detect interaction effects when main effects are also simulated and especially when the same gene is concerned (scenario 2). The LM1 model detects all interaction variables containing at least one gene that has been simulated as having one main or interacting effect leading to a large number of false discoveries. However, if we look at Figure 4.2, that presents the results obtained with the LM1 model in the first iteration realized in the first scenario, we can see that even if a large number of interaction variables presents a significant p-value, the interaction variables that concern the true couples of interacting genes form a distinct group of very small p-values.

Overall we can see that the Group LASSO is the estimation model that presents the better compromise between good power and small amount of false discover-

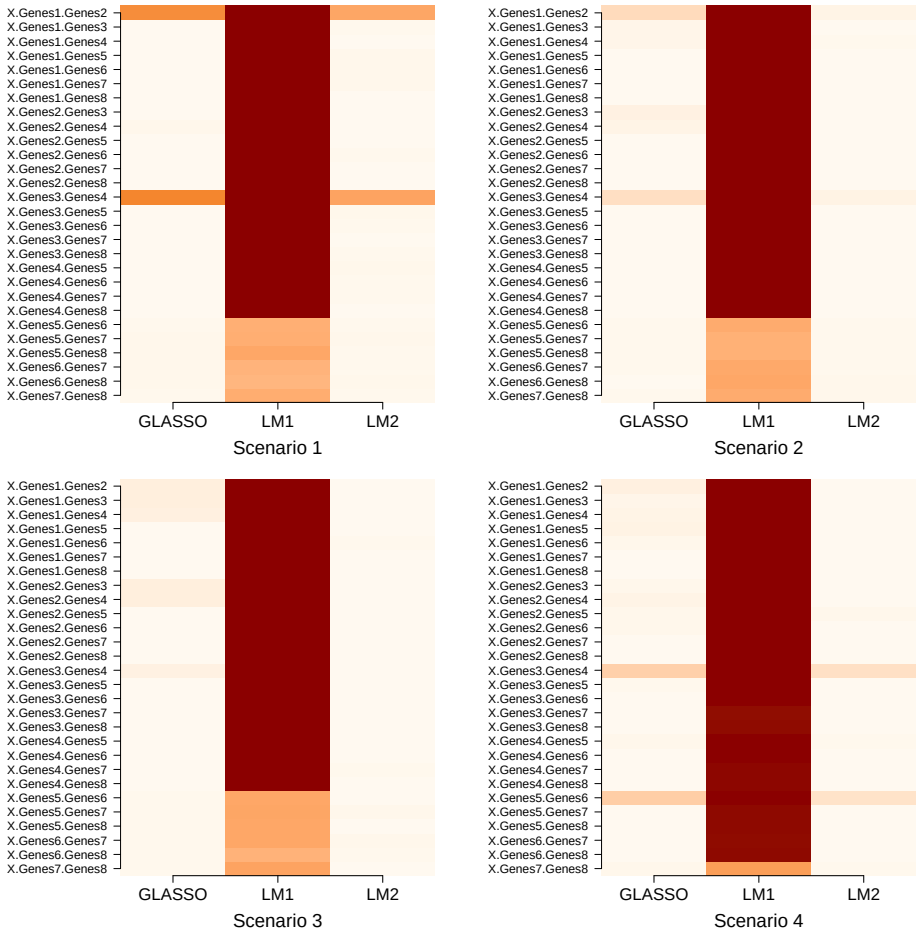


Figure 4.1: Discoveries obtained for each scenario and each model. Each square of each matrix represents the number of times where a variable was detected as significant over the total number of iterations.

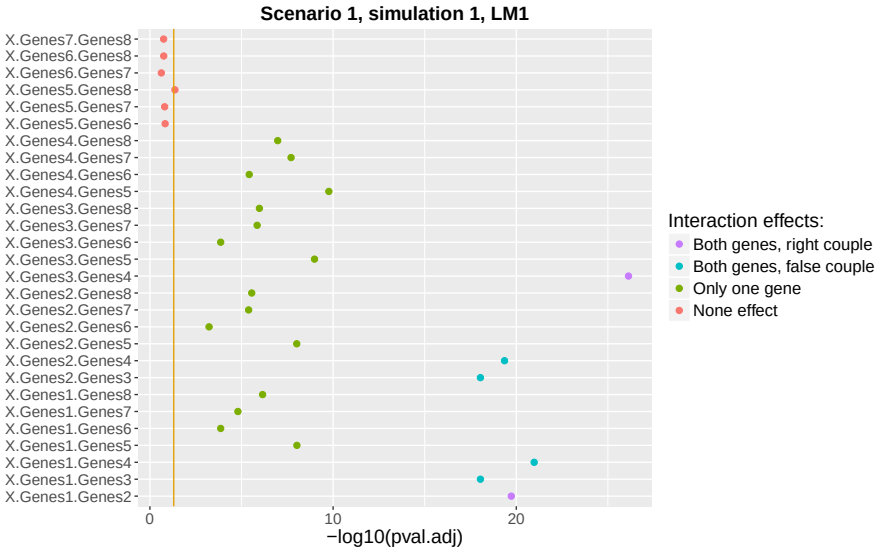


Figure 4.2: Discoveries obtained for the LM1 model under the first simulation of the first scenario. The negative value of the base 10 logarithm of the adjusted p-value is presented for each interaction variable. The threshold of significance (p-values ≤ 0.05) is represented by a orange line. A color is attributed to each dot depending on the type of effect simulate.

	Scenarios			
	1	2	3	4
GL	3h 40m 15s	4h 04m 07s	2h 54m 22s	4h 23m 41s
LM1	23s	18s	13s	17s
LM2	40s	34s	25s	32s

Table 4.1: Execution time for each model under each scenario for the 500 simulations.

ies among the three models and we will keep this model for the next simulation studies. This estimation model is however limited by its computational time as we can see in Table 4.1.

4.3 Comparison of variable interaction modeling

In this section we will compare different strategies to model interactions. The first simulation study will focus on comparing the performance of the PCA, CCA and PLS based approaches and will aim to set the parameters to use for comparison in the other simulation studies. The second section will present the results of the two originals new modeling approaches that we propose to

construct interaction variables in comparison with the previous ones. The most promising approach among the two original ones will be then compared to the PCA and PLS based approaches in a more realistic context. We will finish by presenting the results obtained when representing interactions using machine learning based modeling.

4.3.1 Setting parameters

For each simulation we considered two interaction effects (between the first and second genes and between the third and fourth genes) and we considered two causal SNPs by causal gene. Simulated genotypes are considered and continuous phenotype are obtain using the "Wang Pathway" model (section 4.1.2).

In the following, we used different scenarios in which we chose to vary the following parameters: the correlation level among SNPs belonging to the same gene, the MAF value of causal SNPs, the values for the coefficients β and γ, the number of components to retain to represent the interaction between each couple of causal genes but also the coefficient of determination value r^2, the size of the data set and the presence or absence of main effects. The different values considered for each parameter are presented bellow. Only one parameter is investigated for each scenario the other parameters being fixed to the bold reference values.

- correlation among SNPs: 0.2, **0.8**,
- MAF value of causal SNPs: 0.1, **0.2**, 0.4,
- values of coefficients: $\boldsymbol{\beta = 1}$ or $\beta = 2$ and $\boldsymbol{\gamma = 1}$ or $\gamma = 2$,
- number of components to construct interaction variables: 1, **2**, 3,
- coefficient of determination: $\boldsymbol{R^2 = 0.2}$, $R^2 = 0.05$,
- number of genes: **10**, 64
- number of SNPs in each gene: **6**,
- number of causal SNPs by causal gene: **2**,
- number of subjects: **600**, 300
- type of effects: **4 main effects for 2 interactions** or only 2 interactions,
- number of iterations: **1000**.

The following figures display the power to detect main and interaction effects along with the false positive rates of each type of effects for each simulations

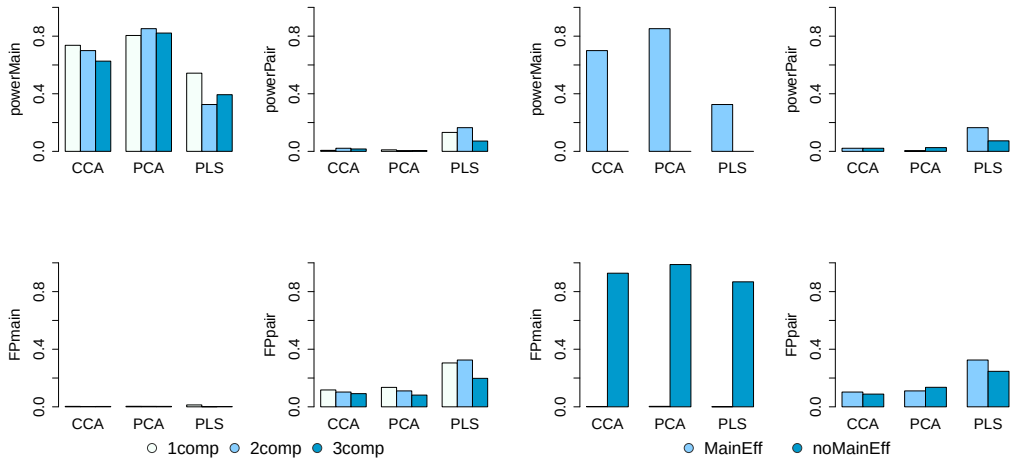


Figure 4.3: On the left: variation of the number of components to construct interaction variables for each couple of genes. On the right: Interactions simulated with or without the presence of their marginal effects.

of each scenario.

Figure 4.3 presents the influence of the number of considered components to represent the interaction of a couple of genes in the model along with the importance for the genes in interaction to also present main effects. As we are not interested in investigating the potential role of the coefficient values with these two scenarios, in all the simulations, the value of the coefficients β and γ will be sampled in the interval $[\log(1.1), \log(1.3)]$. The importance of these latter parameters will be investigated in the next scenarios. Overall, we can see that the PLS based method seems to perform better than PCA or CCA to detect interaction effects but is less efficient to detect the four main simulated effects (if we look at the discoveries matrix of the second scenario in Figure 4.4, we can see that the PLS based method detect only the first gene with a good power). We can also note that the false positive rate among interaction effects is larger for the PLS based method. Regarding the two parameters compared in those simulations, we can see that increasing the number of components considered to define interactions do not seems to have an important effect with the PCA based method apart from a slight reduction of the false positive rate to detect interaction effects. We can observe a difference with CCA but mainly regarding the power to detect main effects. The power of the method diminishing when a larger number of components are considered. Finally, regarding the PLS based method, the power to detect main effect is more important when only one component is considered. When the number of components increases, the ability of the method to detect any type of effects (whether being true or false) decreases. In this scenario, where

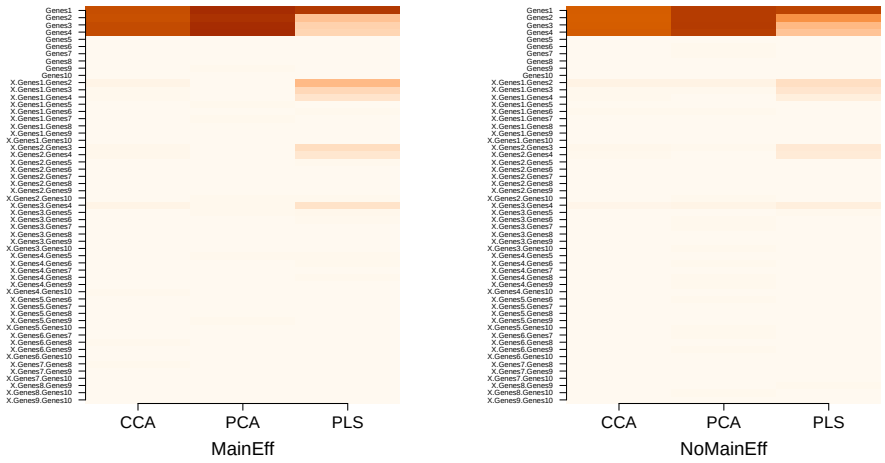


Figure 4.4: Discoveries matrices for the simulations with or without main effects. Each square represents the number of times where a variable was detected as significant over the total number of iterations.

the variation of the number of components is considered, we observe a very low false positive rate for detecting main effects for the three methods and for each simulation. However, in the second scenario, where we are interested in comparing the impact of the presence or absence of the main effects corresponding to the simulated interactions, the false positive rate for detecting main effects varies widely between the simulations. In the simulation where no main effects were simulated we can see that the false positive rate to detect main effects is close to one whereas it is close to zero in the other case. When we look at the discoveries matrix corresponding to this two simulations (Figure 4.4), we can see that these false discoveries correspond to the genes that have been simulated to have interaction effects. We can also note that when the genes simulated as having an interaction do not have a main effect, the power of the PLS based method to detect those interactions is slightly reduced.

We will then focus on the impact the size effects. In the first scenario presented on the left side of the Figure 4.5 we compare two simulations in which the coefficients β, related to the each marginal effects, are alternately smaller or larger than the coefficients γ that are related to the interaction effects of each couple (see the "Wang Pathway" phenotype simulation model in Section 4.1.2). In the first simulation $\beta = 1$ and $\gamma = 2$ whereas in the second one $\beta = 2$ and $\gamma = 1$. Overall we can see that changing the size of the different simulated effects does not seem to have a considerable impact for the three methods. The power to detect main effects seems to be slightly larger when $\beta > \gamma$ for the CCA and PCA based methods. The same observation can be made concerning the false positive rate to detect interactions for the PLS based approach. On the right side of Figure 4.5 we are interested in comparing the impact of the

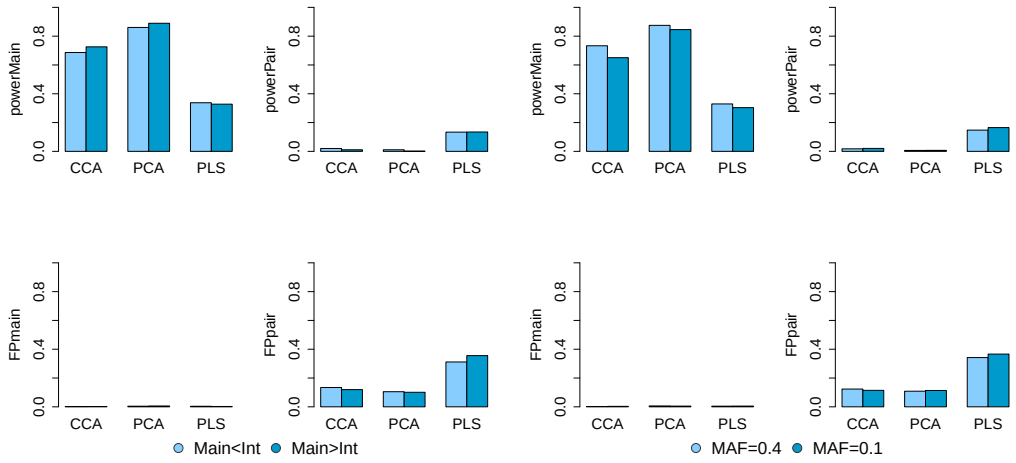


Figure 4.5: On the left: variation of the size of the coefficients β and γ in the simulation of the phenotype. On the right: variation of the MAF of causal SNPs.

MAF value of the causal SNPs. As for the previous scenario, the modification of this parameter does not seem to have a large impact for the three method apart from a slight decrease of the power to detect main effects for the CCA and PCA based methods when the MAF value is low.

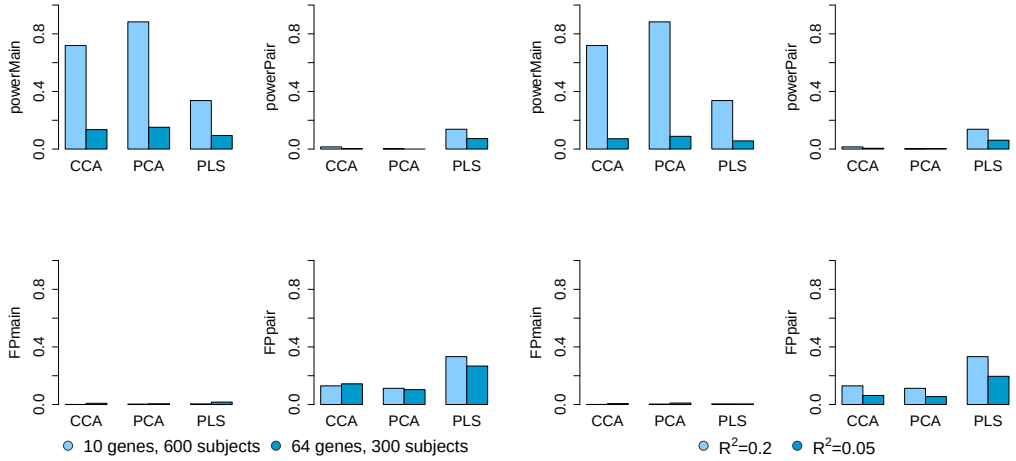


Figure 4.6: On the left: variation of the number of genes and subjects considered. On the right: variation of the R^2 value.

Figure 4.6 presents the impact of the considered data set size along with the R^2 value which parameterizes the difficulty of the simulation model. In both

scenarios we can see that the power of all methods to detect main effect drastically decreases either when the R^2 is low or when the simulations are realized in a higher dimensional context (300 subjects and 64 genes of 6 SNPs corresponding to 384 variables). We can also see that the ability of the PLS based method to detect any kind of interaction effects decreases in those contexts. The power as well as the false positive rates to detect interactions are smaller.

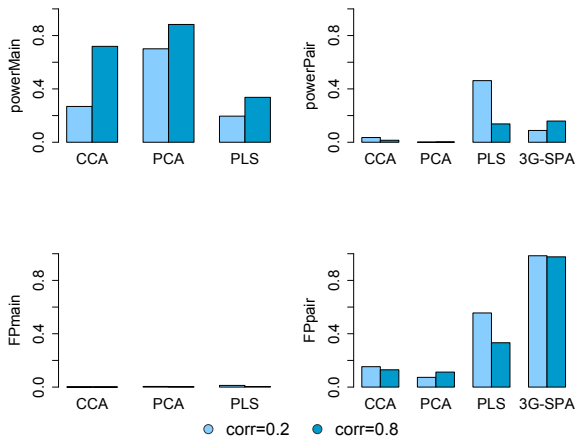


Figure 4.7: Variation of the correlation among genes.

Figure 4.7 presents the impact of changing the level of correlation among SNPs in a gene, and compares the results of the three methods to the 3G-SPA approach proposed by Li and Cui [2012] (more details regarding this method are given in Section 2.3.3). When the correlation between the SNPs is low, the PCA, CCA and PLS based methods are less powerful to detect the main effects. However we can see that the PLS based approach detect more interacting effects in this context as both power and false positive rate to detect interactions increase. Regarding the method 3G-SPA, we can see that the number of false positives among interactions is extremely high compared to the other methods and does not seem adapted to a context close to our simulation setting.

The parameter modifications that present the largest impacts concern the R^2 value, the size of the data set and the presence or absence of the marginal effects of the genes simulated as having an interaction effect. In the next simulation studies we will focus on those parameters in order to compare the performance of the investigated methods. Other parameters will be fixed to the following values:

- correlation among SNPs: 0.8,

- MAF value of causal SNPs: 0.2,
- values of coefficients: fixed to $\beta = 2$ and $\gamma = 2$,
- number of components to construct interaction variables: 2,
- number of genes: 6,
- number of SNPs in each gene: 6,
- number of causal SNPs by causal gene: 2,
- number of subjects: 600,
- number of iterations: 1000.

4.3.2 Two new approaches: G-GEE$_{c1}$ and G-GEE$_{c2}$

In the following, we will compare the PCA, CCA and PLS based approaches with the two new interactions modeling methods, G-GEE$_{c1}$ and G-GEE$_{c2}$, that we described more precisely in the section 3.3.2. We will compare the performances of the five approaches for eight different R^2 values (0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7) and we will consider various types of marginal and interaction effects in order to evaluate more specifically the impact of the presence of main effects. We will also consider scenarios where the size of the data set will change.

For each scenario we will simulate genetic effects using the two different phenotype simulation models presented in Section 4.1.2 and investigate their respective impacts. We will refer to the different simulation settings by using letters as described in Table 4.2.

Settings			Names	
Id	Main effects	Interaction effects	Wang Pathway model	PCA model
1	Genes 1, 2	Genes 1 x 2	A	B
2	Genes 1, 2	Genes 3 x 4	C	D

Table 4.2: Effects simulated in each settings and referring names according to the phenotype simulation model.

Figures 4.8 and 4.9 present the results obtained for the two settings. In Figure 4.8, the first two columns give the estimated power to respectively detect the gene interaction and the marginal effects as a function of the R^2 values. The last two columns show false positive rates among main and interaction effects depending on the R^2 values. Figure 4.9 shows in more details the discoveries obtained for a particular R^2 and for each setting.

Overall we can see that the two methods G-GEE$_{c1}$ and CCA present very low power to detect interaction effects in all simulations of both settings but are more effective to detect marginal effects. Indeed, we can see that for all simulation settings the power to detect the two simulated main effects are close to one as soon as $R^2 > 0.2$ whereas the power to detect the interaction effect is almost always close to zero, with an exception for CCA which tends to slightly detect the interaction effect for large value of R^2 when phenotypes are simulated with the PCA model (Figure 4.8(B, D)). We can also note that, in the second setting, when the phenotypes are simulated using the "Wang Pathway" model (Figure 4.8(C)) both methods detect false positive effects among main effects. Looking at Figure 4.9, we can see that these false positives correspond to the two genes that have been simulated as having an interaction effect only. In the same setting but when the phenotypes are simulated using the PCA model (Figure 4.8(D)), the false positive rates are very low but the genetic effects of both genes simulated with the interaction effect are nearly not detected at all.

Regarding the three other methods, we can see that the PLS based approach is able to detect interaction effects only when the genes involved in the interaction effect have a marginal effect. The power of this method to detect marginal effects is close to 0.5 for all settings as soon as $R^2 > 0.2$. When looking at Figure 4.9 we can see that it is because, in all the settings, only the first gene of the two simulated marginal effects tends to be detected. We can also note that the PLS based approach tends to detect an interacting signal between the genes simulated as having a main effect. Indeed we can see that the false positive rate for interactions is greater in the second setting (Figure 4.8(C, D)) and that the false detecting effects correspond to the interaction between gene 1 and gene 2 (Figure 4.9(C, D)).

The ability of the PCA based method to detect interactions mainly depends on the phenotype simulation model. The PCA based method shows good performance when the phenotypes are simulated using the PCA model (Figure 4.8(B, D)) but this can be explained by the similarity between the phenotype simulation model and the estimation model. When the phenotype is simulated using the "Wang Pathway" model (Figure 4.8(A, C)), the PCA based method presents a good power to detect the two marginal effects but its power to detect the interaction effect sharply drops as the R^2 value declines. For high R^2 values (0.7), and contrary to the PLS based method, the PCA based method seems to detect more easily the interaction effect when the corresponding genes do not have a marginal effect. Like G-GEE$_{c1}$ and CCA, in the second setting, when the phenotypes are simulated using the "Wang Pathway" model (Figure 4.8(C)),

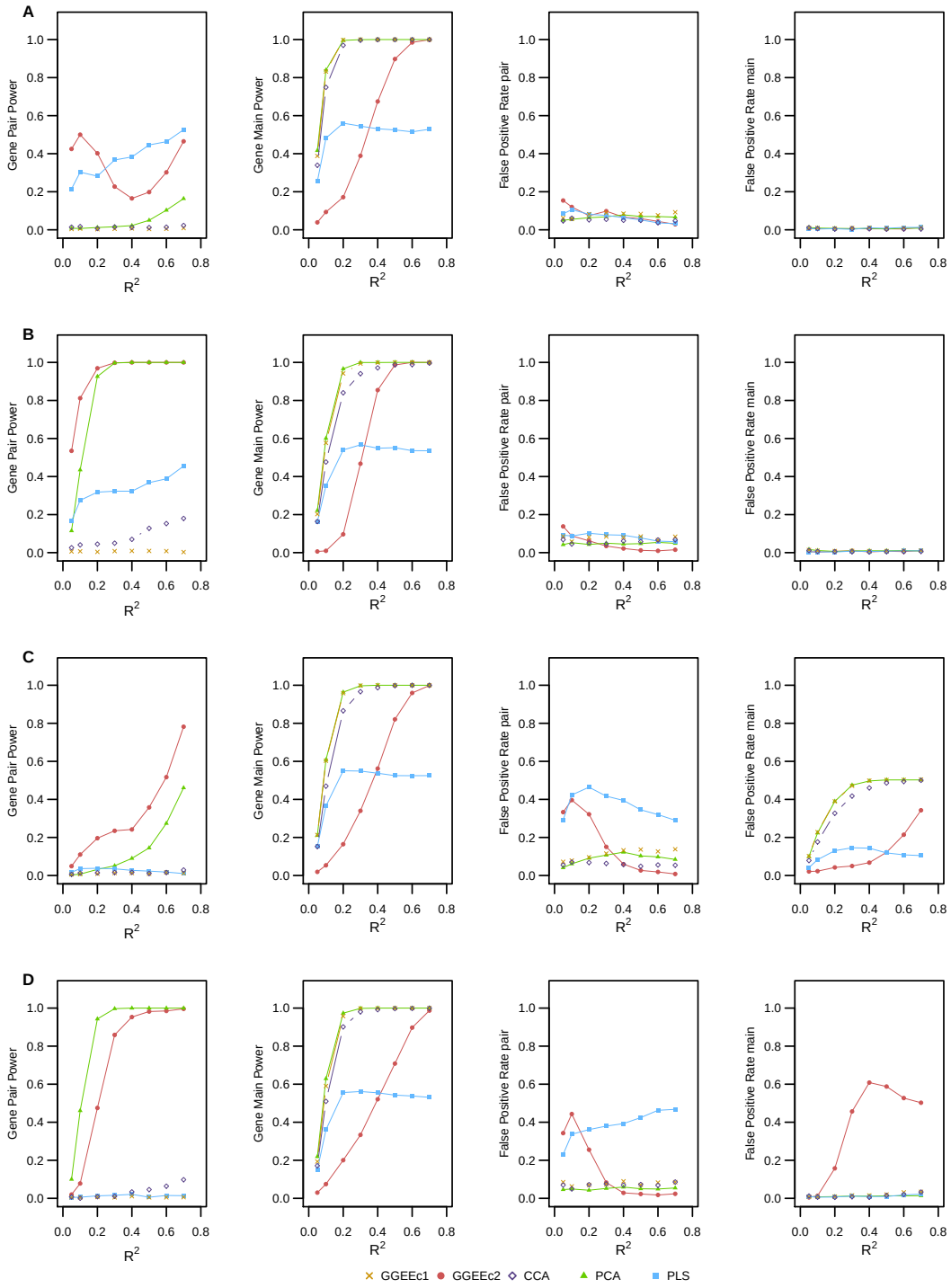


Figure 4.8: **Interaction effect power.** The figures in the two first column show the power to respectively detect interaction and marginal effects of all methods depending on the R^2. The last two columns present the false positive rates of each method depending on the R^2. The panels **A**, **B**, **C** and **D** refer to the different simulation settings described in Table 4.2.

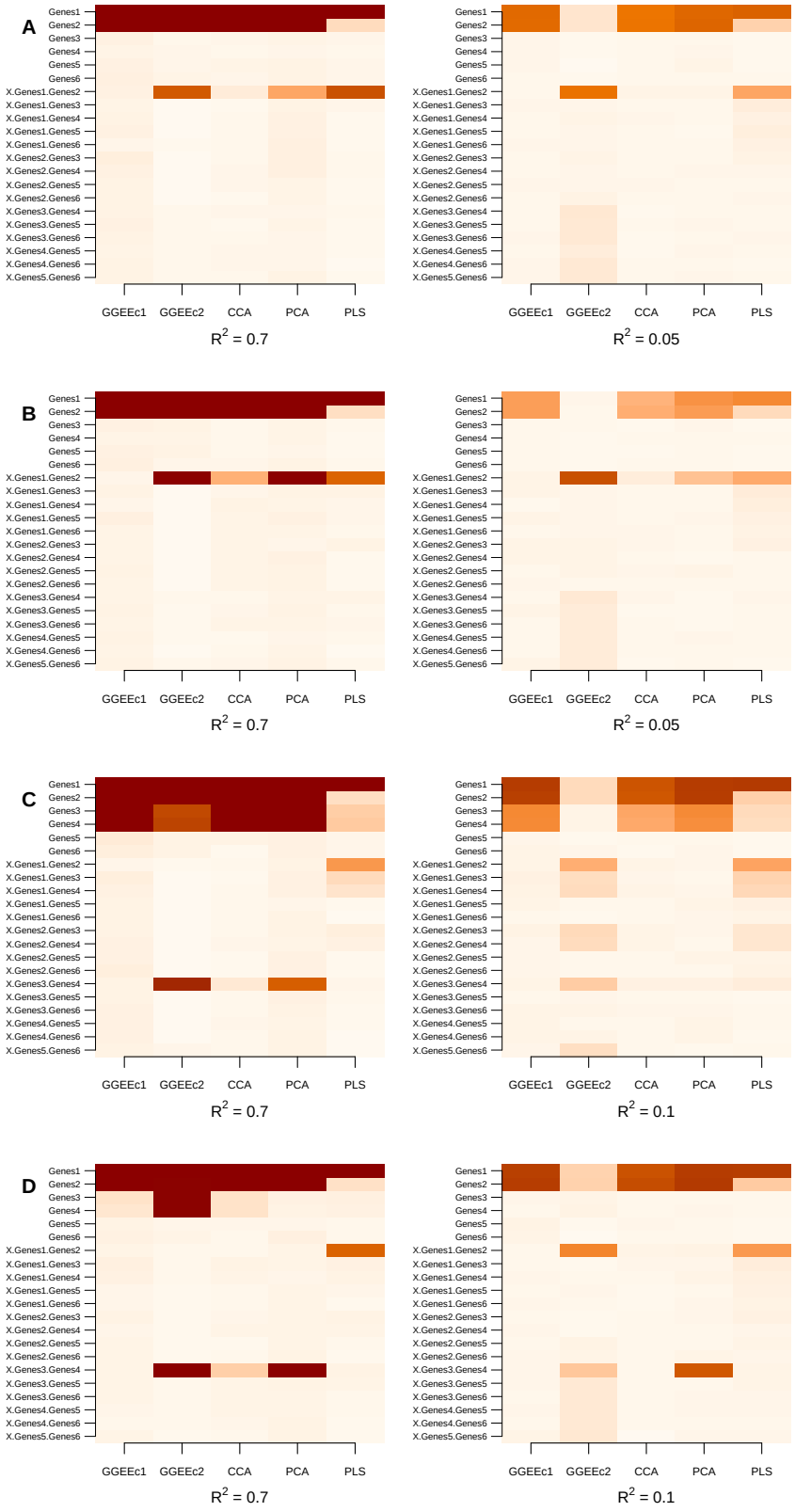


Figure 4.9: **Discoveries matrices.** For each simulation setting **A**, **B**, **C** and **D**, described in Table 4.2, are presented two discoveries matrices for two given values of R^2. Each square of these matrices represents the number of times where a variable was detected as significant over the total number of iterations.

PCA based approach tends to detect false positive effects among main effects that correspond to the two genes that have been simulated as having an interaction effect only (Figure 4.9(C)).

Finally, we can note that contrary to G-GEE$_{c1}$, PCA and CCA based approaches, the power of G-GEE$_{c2}$ to detect main effects sharply declines as the R^2 value diminishes. However, this approach presents globally a better power to detect interaction effects than the other methods. When the phenotypes are simulated using the PCA based method, the power of G-GEE$_{c2}$ stays high until the R^2 value reaches 0.3 (Figure 4.8(B, D)). When the phenotypes are simulated with the "Wang Pathway" model, the power to detect interaction decreases quickly as the R^2 value diminishes. However, in the first setting, the power to detect interactions starts to increase for very small R^2 values leading to a U-shaped power curve. This phenomenon can be explained by the fact that G-GEE$_{c2}$ tends to easily assign marginal signals to genetic effects when the R^2 is high and to easily assign interaction signals to genetic effects when the R^2 is low. Indeed we can see in the second setting (Figure 4.8(C, D)) that the false positive rates are more important among the main effects when R^2 values are high and correspond to genes that have been simulated as having interaction effect (Figure 4.9(C, D)). When R^2 values are low the false positive rates are more important among interaction effects and concern couples of genes that have been simulated either to present a marginal effect or an interaction effect but with another gene.
Finally, in the second setting, we can also note a difference between the methods concerning false discoveries for the main effects. G-GEE$_{c2}$ leads to false discoveries when the phenotypes are simulated with the PCA model while the other methods show false discoveries mainly when phenotype are simulated with the "Wang Pathway" model.

To evaluate the performances of the different methods in a more complex context, we also consider a scenario where we vary the number of genes that compose each simulated data sets and where genetic effects are composed from a mixture of independent marginal and interacting effects. In the first simulation setting we will consider 25 genes and in the second 40 genes with the following genetic effects:

Main effects	Interaction effects
Gene1, Gene2	Gene3 x Gene4
	Gene5 x Gene6
	Gene7 x Gene8
	Gene9 x Gene10

In these simulations, interacting genes are different from genes having marginal

effects. Genetic effects are simulated using the "Wang Pathway" model and we only consider the case where $R^2 = 0.7$. The results of these two simulations are presented in Figure 4.10 and they reflect the good performance of the G-GEE$_{c2}$ based method over the other approaches in detecting interactions in a context where several interactions and different main effects are simulated. G-GEE$_{c2}$ is thus the method having the highest power and the lowest false positive rates to detect interactions for both simulations. G-GEE$_{c2}$ is however less powerful in the detection of marginal effects compared to the other methods but shows the lowest false positive rate. Regarding the four other methods, we can see in Figure 4.11 that the false discoveries among main effects concern the genes 3 to 10 that were simulated to only have an interaction effect. We can also see that the PLS based approach shows a good power for only one of the simulated marginal effects and tends to detect a false interaction between the two first genes that are supposed to only have marginal effects. Overall these results are very close to those obtained in the simulation setting C where only data set sizes and numbers of simulated interactions are different. We can however note that, except for G-GEE$_{c2}$, the false positive rates are more important when the number of variables is larger and when more interaction effects are simulated. Regarding G-GEE$_{c2}$, the method presents a larger power to detect interaction effects as the number of variables in the data set increases but also a larger number of false positives interactions.

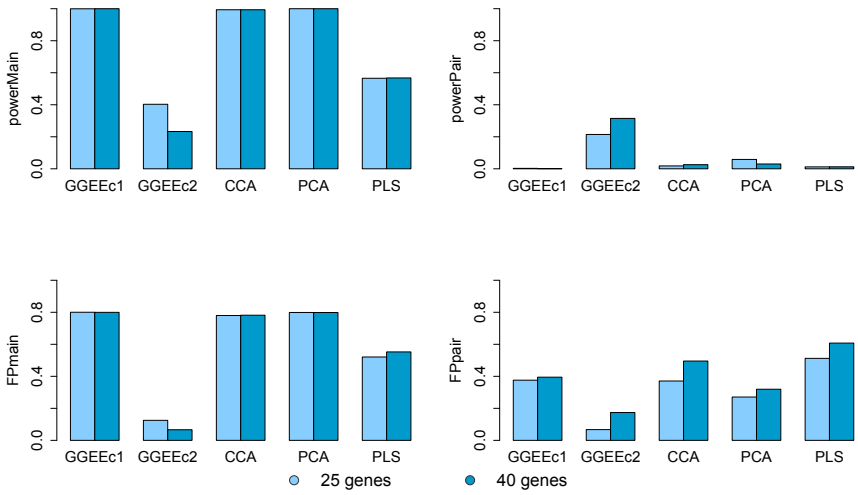


Figure 4.10: Variation of the number of genes in the data set.

From these results, we can make different conclusions. First, regarding the individual performance of each method, G-GEE$_{c1}$ and CCA based methods seem to be inadequate to detect interactions since interaction effects may be missed or confounded with marginal signals. The PCA and PLS based ap-

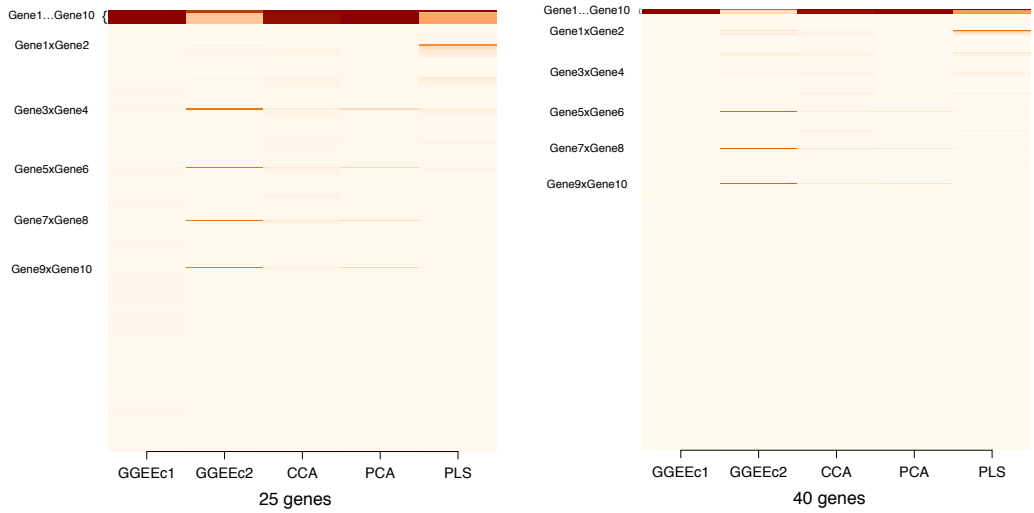


Figure 4.11: Discoveries matrices when $R^2 = 0.7$ for the settings where 25 genes (left) or 40 genes (right) were simulated to constitute the data set. Each square represents the number of times where a variable was detected as significant over the total number of iterations.

proaches present better performance but highly depend on the simulation setting. Indeed, the PLS based method is not able to detect an interaction if the corresponding marginal effects are not present while the PCA based method can only detect interaction effects when the phenotype simulation setting is favorable (either when R^2 values are high or when phenotypes are simulated using the PCA model). G-GEE$_{c2}$ has a better power to detect interaction effects even if this power tends to drop more quickly with the decrease of the R^2 when phenotypes are simulated with the "Wang Patway" model. Second, each method seems to preferentially assign a particular type of signal for any simulated genetic effect. Thus, the PCA based approach seems to have more abilities to detect the genetic effects as marginal; PLS tends to detect main and interaction effects for the genes that have marginal effects and fails to detect interactions for genes without main effect; G-GEE$_{c2}$ tends to more easily detect interaction effects. We can also note that the capacity to detect interactions increases for G-GEE$_{c2}$ when the difficulty grows (for small R^2 values or when the number of variables in the data set increases). Third, it can be noted that according to the phenotype simulation model, the methods perform differently in terms of power and false discovery rate.

4.3.3 Simulations in a realistic context

In the following, we will only focus on comparing the G-GEE$_{c2}$, PCA and PLS based approaches as G-GEE$_{c1}$ and CCA based methods did not show promising results to investigate epistasis. Thereafter, G-GEE$_{c2}$ will be more simply

referred as G-GEE.

The aim of these simulations is to investigate the performance of the three methods in a more realistic context. To this end we will consider genotype derived from a real dataset as described in Section 4.1.1. The parameter values will be the same than the ones fixed at the end of Section 4.3.1. Differences will concern the number of subjects (here 763) and the number of iterations considered for each simulation (currently 500). The different simulation settings that will be investigated are described in Table 4.3 and we will refer to each setting using the corresponding letter. For each simulation setting, we will realize four sets of simulation depending on the following R^2 values: $(0.1, 0.2, 0.3, 0.4)$. We can note that the two first settings are similar to those presented in the Table 4.2 in the previous section.

Settings			Names	
Id	Main effects	Interaction effects	"Wang Pathway" model	PCA model
1	Genes 1, 2	Genes 1 x 2	A	B
2	Genes 1, 2	Genes 3 x 4	C	D
3	-	Genes 1 x 2	E	F
4	Genes 1, 2	-	OME	
5	-	-	NE	

Table 4.3: Effects simulated in each setting and referring names according to the phenotype simulation model.

Figures 4.12, 4.13 and 4.14 show results for the three first settings. Figure 4.12 presents the power to detect interactions and marginal effects along with the false positive rates depending on the R^2, while Figure 4.13 and 4.14 show discoveries matrices for each setting when $R^2 = 0.1$ or $R^2 = 0.4$ respectively. In those figures, the upper row relates to phenotypes simulated using the "Wang Pathway" model, and the lower row to phenotypes simulated using the PCA model.

Overall we can see that the PLS based method favors the detection of marginal effects over interactions. Indeed, we can first see that the power of PLS based method to detect marginal effects is always greater to its power to detect interactions in the two first settings (Figure 4.12(A, B, C, D)). Secondly, in the third setting, where only an interaction effect is simulated, the PLS based method detects the genetic signal predominantly as marginal (Figure 4.13(E, F) and 4.14(E, F)) what explains the high rate of false positives among main effects depicted in the Figure 4.12(E,F). Third, we can also see this inclination for marginal effects in the second setting, where different marginal and interaction effects are present. Indeed, the power to detect interactions is almost null for all R^2 values (Figure 4.12(C, D) and the genetic signal for genes 3

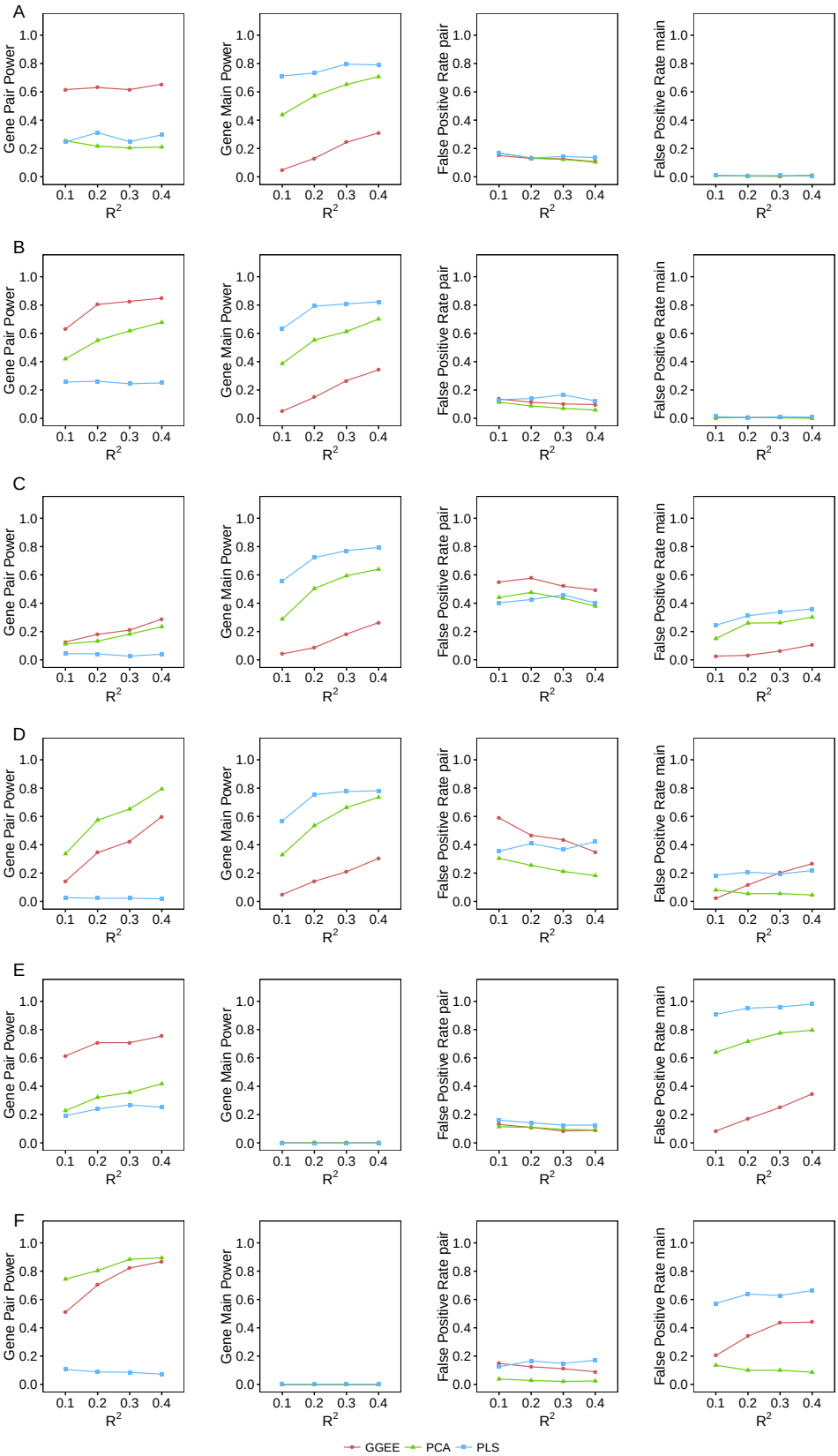


Figure 4.12: The figures show the power to detect marginal or interaction effects along with the false positive rates of the three methods depending on R^2. The panels **A**, **B**, **C**, **D**, **E**, and **F** refer to the different simulation settings described in Table 4.2.

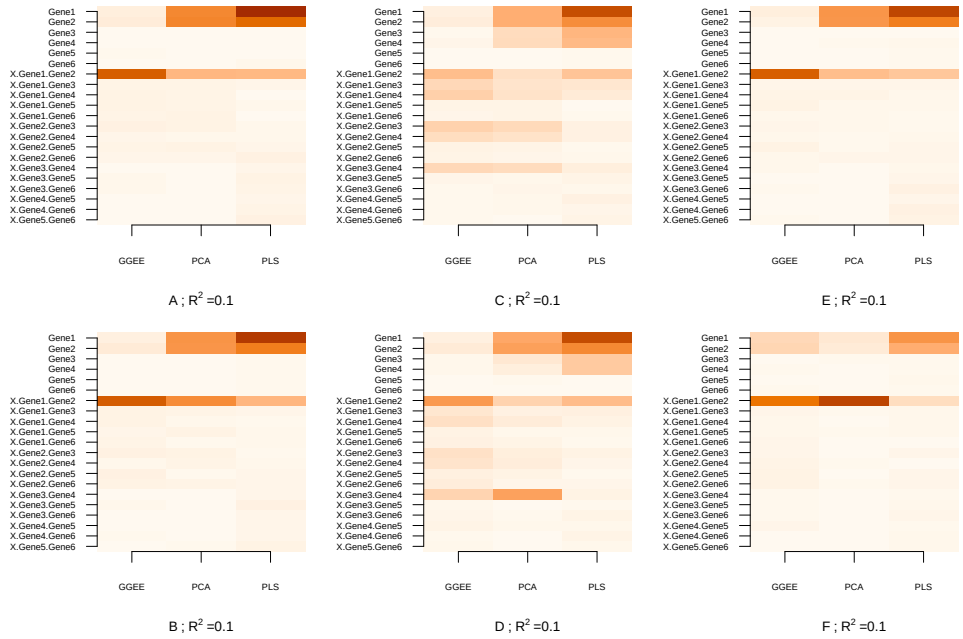


Figure 4.13: Discoveries matrices when $R^2 = 0.1$ for each simulation settings **A**, **B**, **C**, **D**, **E**, and **F** described in Table 4.2. Each square of each matrix represents the number of times where a variable was detected as significant over the total number of iterations.

and 4 that is supposed to be an interaction is falsely detected as two marginal effects (Figure 4.13(C, D) and 4.14(C, D)). However we can also note, as in the previous section, that the PLS based method tends to assign both main and interacting effects for the genes that are simulated with marginal effects. We observe this phenomenon in the second simulation setting where the false positive rate for interactions is close to 0.4 for all R^2 values (Figure 4.12(C, D) and concerns genes that are simulated to only have marginal effects (Figure 4.13(C, D) and 4.14(C, D)). In comparison with the previous section where simulations were realized on simulated genotypes, we can see that the main difference for the PLS based method lies in the attribution of marginal signal. Power to detect main effects and false positive rate for main effects are more important in the current section where genotypes derive from a real data set (Figure 4.8 and Figure 4.12). This better power to detect marginal effects can be explained by the fact that the tendency of PLS to detect only one of the two marginal effects is less pronounced. The higher rate of false positives among main effects may also be explained by the fact that in the second setting of the previous section, PLS was not able to detect any signal for the two genes simulated as having an interaction (gene 3 et gene 4) whereas when the genotypes come from a real data set, the method detects the interaction signal as marginal.

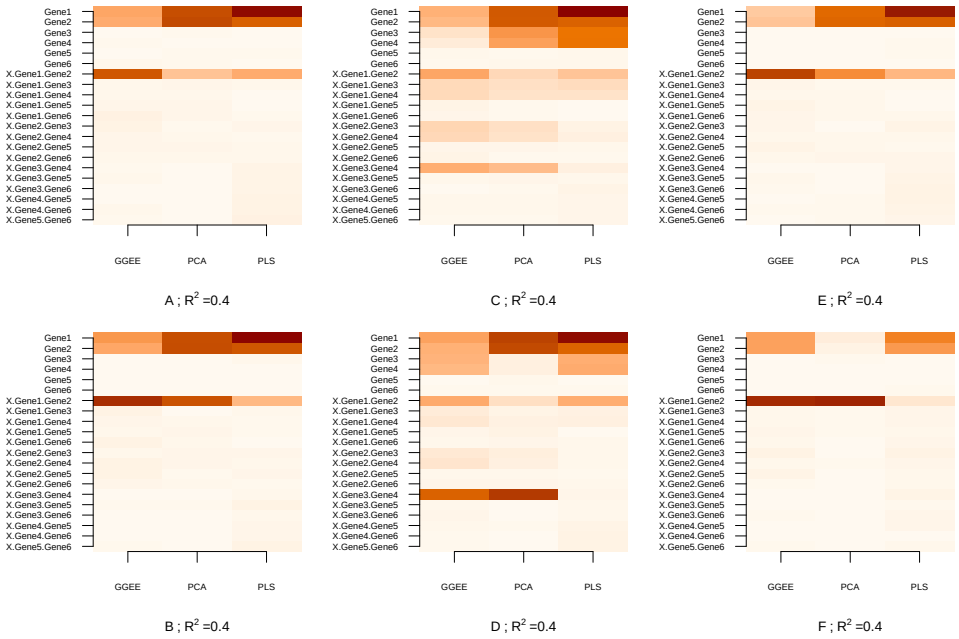


Figure 4.14: Discoveries matrices when $R^2 = 0.4$ for each simulation settings **A**, **B**, **C**, **D**, **E**, and **F** described in Table 4.2. Each square of each matrix represents the number of times where a variable was detected as significant over the total number of iterations.

When phenotypes are simulated using the "Wang Pathway" model, PCA tends to easily detect and assign marginal effects, like PLS based method. Thus, the method has a good power to detect marginal effects but also tends to show false positives among main effects when the simulated interactions do not also present marginal effects (Figure 4.12(C, E)). However, PCA presents a better power to detect interactions than PLS. Indeed, in almost all settings, PCA is able to more effectively detect the interaction effects, even if its power decreases when the R^2 values decrease (Figure 4.12). In the second setting, PCA method also tends to detect false positives among interactions and main effects but contrary to PLS, the false discoveries among interactions concern couples of genes that have been simulated either with a marginal effect or in interaction with another gene and not mainly between gene 1 and gene 2 as PLS tends to detect (Figure 4.13(C) and 4.14(C)). In comparison to the previous section and contrary to the PLS based method, the ability of PCA to detect marginal signal is less important in the current section. The power to detect marginal effects and the false positive rate among main effects are both less important (Figure 4.8 and Figure 4.12). However, when the phenotypes are simulated using the "Wang Pathway" model, the PCA based method tends to more easily attribute interaction signals. Indeed, we observe that power and

false positive rate among interactions in the second setting are slightly higher. The good performance of the PCA based method when phenotypes are simulated using the PCA model are still considerable when using real genotype data because of the similarities between the simulated and estimation model but are however slightly smaller than when using simulated genotypes.

Contrary to PLS and PCA based methods we can see than G-GEE favors the detection of interactions over marginal effects. Indeed the method presents a better power to detect interactions than PLS and PCA based methods in almost all simulations even if its capacity to detect the interactions simulated is less important in the second setting (Figure 4.12). The false positive rates among interaction in the second simulation setting are also more important for G-GEE than for the two other methods. As for PCA the false discoveries among interactions concern couples of genes that have been simulated either to have a marginal effects or to be in interaction with another gene and not mainly between gene 1 and gene 2 as done by PLS (Figure 4.13(C) and 4.14(C)). Regarding marginal effect G-GGE is less efficient than PLS and PCA but also presents the smallest false positive rate among main effects (Figure 4.12). Like the PCA based method, G-GEE tends to detect and assign interaction effects more easily but is less efficient regarding marginal effects when using real genoptypes than when using simulated ones. Indeed, in the first simulation setting the power to detect interactions is higher and does not present a U-shaped form as in the previous section (Figure 4.8(A) and Figure 4.12(A)). The false positive rates among interactions are also more important when the genotypes derive from a real data set in the second simulation setting (Figure 4.8(C) and Figure 4.12(C)). Conversely, and like PCA, the ability of G-GEE to assign marginal effects (either true or false) is less pronounced when real genotypes are used. Finally, the contrast between the type of effects detected by G-GEE according to the values of R^2 (tendency to assign marginal effect for large R^2 values and interaction effects for smaller ones) exists but is less marked when the genotypes come from a real data set.

In the last two settings presented in Table 4.3, we investigate the performance of the three methods when no interaction effect is simulated. In the fourth setting, only two main effects for gene 1 and gene 2 were simulated while in the fifth one, none effect was present. Figure 4.15 shows the results for these two settings. The discoveries matrix on the left corresponds to the setting where only two main effects are simulated. We remark that all methods successfully identify the main effect, PCA and PLS doing so with a higher power. False detections corresponding to the respective interaction effects are observed for G-GEE, and to a lesser extent for PLS. These results concord with the observations already made for the three methods. G-GEE tends to detect the signal of simulated effects as an interaction while PLS tends to both assign main and interaction effects for the two marginal simulated effects. The PLS

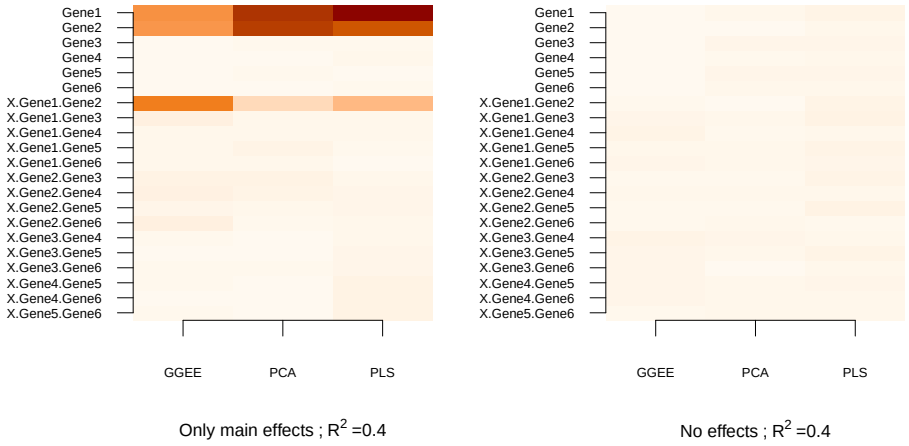


Figure 4.15: Discoveries matrices for $R^2 = 0.4$ when only main effects are simulated for gene 1 and gene 2 (left), and when no effects are simulated (right). Each square represents the number of times where a variable was detected as significant over the total number of iterations.

based method also tends to better detect the first marginal gene over the two. The figure on the right corresponds to the setting where no specific effects are simulated and the result shows that all three methods perform well with very few false detections.

Comparison of execution times required to model interactions and to fit Group LASSO for the five first settings of the realistic simulation study are presented in Figure 4.16. In all settings, estimating the coefficients with the Group LASSO is more computationally expensive than constructing the interaction variables. We can note that the time to compute Group LASSO is particularly high when none effect is simulated. G-GEE and PCA are quite similar in terms of computation time, whereas in some settings PLS has a slightly greater execution time than other methods. Note that the time required by G-GEE for constructing the interaction variables varies according to the number of SNPs that constitute each gene. We can also note that the time to compute Group LASSO tends to increase when the R^2 values decreases and the problem becomes more difficult.

In conclusion we can note that both the PCA and PLS based methods tend to favor the detection of marginal effects over interactions. Both methods have a good power to detect the simulated main effects in all settings but also tend to assign marginal signals to the genes that are simulated with interaction. The two approaches differ however in their capacity to detect interactions. PLS tends to assign interaction signals to the genes that are simulated as marginal (even when the corresponding interaction is not simulated) but the method

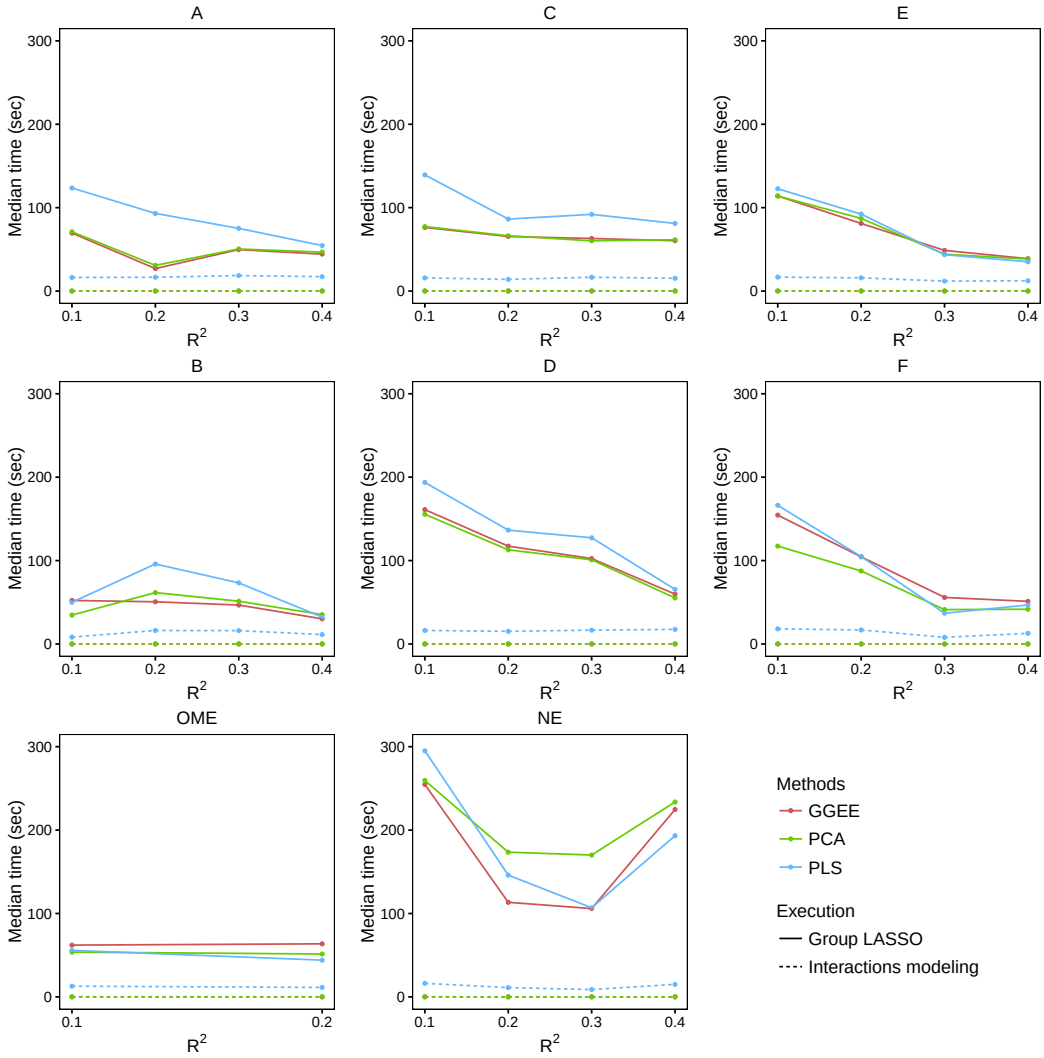


Figure 4.16: **Execution time** Median of the execution time to model interaction and to fit Group LASSO for the three first settings.

is not able to detect pure interactions. PCA seems to be more flexible on this hypothesis and do not assign interaction effects to the genes simulated as purely marginal. Contrary to these two methods, G-GEE tends to favor the detection of interactions over marginal effects. The power of G-GEE to detect interactions is almost always higher than for PLS and PCA but the method also tends to detect marginal simulated effects as interactions.

4.3.4 Simulations with binary phenotypes

In this simulation study the aim is to evaluate our method G-GEE with binary phenotypes. In particular, we will compare the performance of G-GEE with other methods that have specifically been developed to detect interactions between genes in a case control study design. We thus chose to compare G-GEE with some of the methods presented in Section 2.3.3 that have been implemented in the R package GeneGeneInteR (GGI) [Emily et al., 2017] and that are listed below:

- Tests outside a regression framework:
  - CLD: Composite Linkage Disequilibrium method [Rajapakse et al., 2012]
  - Test based on aggregating SNP-SNP level tests p-values:
    - minP: Minimum p-value test [Emily, 2016]
    - GATES procedure [Li et al., 2011b]
    - tProd: Truncated product method [Zaykin et al., 2002]
    - tTS: Truncated tail strength test [Jiang et al., 2011]
  - Co-association test
    - CCA: Canonical Correlation-based U-statistic model [Peng et al., 2010]
    - PLSPM: Partial Least Squares Path Modeling approach [Zhang et al., 2013]
- Regression based approaches
  - PCA: Principal Component Analysis based method [Li et al., 2009]
- Entropy based method
  - GBIGM: Gene Based Information Gain Model [Li et al., 2015]

For all the simulations presented in this section, the value of the parameters will be the same as the ones fixed at the end of Section 4.3.1. The difference will concern the number of iterations considered for each simulation currently fixed to 500.

Settings			Names	
Id	Main effects	Interaction effects	Real data	Simulated genotypes
1	Genes 1, 2	Genes 1 x 2	Ar	As
2	Genes 1, 2	Genes 3 x 4	Cr	Cs
3	-	Genes 1 x 2	Er	Es

Table 4.4: Effects simulated in each setting and referring names according to the type of genotype considered.

We will first present the results of the comparison of G-GEE with the Minimum p-value test (minP) and the PCA based method implemented in GGI. We will consider binary phenotypes simulated using the "Wang Pathway" model (Section 4.1.2) and will compare the performance of the methods according to the type of genotype considered which will be either simulated or derived from a real data set (in this latter case the data set considered is the same that the one presented in the previous Section 4.3.3). The three following scenarios considered are presented in Table 4.4.

Results of these three scenarios are presented in Figure 4.17. Overall we can see than G-GEE tends to perform better to detect interaction than the two others, particularly when genotypes come from real data. We can also note that G-GEE is more able to assign interactions when the genotypes derive from a real data set and to assign marginal effects when genotypes are simulated. We thus notice that in the second setting the discoveries matrix Cr indicates more false positives among interactions than the discoveries matrix Cs. These false discoveries concern couples of genes that have been simulated to present either a marginal effect or an interaction effect with another gene. The power of minP to detect interactions when genotypes are simulated is similar to the power of G-GEE. When simulations are realized using genotypes from a real data set the power of minP is almost null but this is mainly due to the fact that the method was not able to complete and to yield a result for a large number of iterations (400 in the simulation Ar). In those cases, we consider that the methods did not make any discovery and we use the usual estimators for the power and false positive rates over all iterations. The same computational problem occurred with PCA but to a lesser extent.

We can remark that the performance of G-GEE in the current results, where phenotypes are binary, are similar to the results obtained for high values of R^2 when phenotypes were simulated as continuous with either simulated genotypes (Figure 4.9(A, C)) or real genotypes (Figure 4.14(A, C, E)). As a reminder, we noticed in the two previous sections that G-GEE tends to assign marginal signals for large R^2 values and interaction signals for smaller ones. When binary phenotypes are simulated, it is not possible to control the value of the

Figure 4.17: Discoveries matrices for each simulation settings Ar, As, Cr, Cs, Er, and Es described in Table 4.4. Each square of each matrix represents the number of times where a variable was detected as significant over the total number of iterations.

coefficient of determination R^2. However, in our results we can note that the tendency to assign marginal signals to the simulated genetic effects is more pronounced in the actual section when phenotypes are binary.

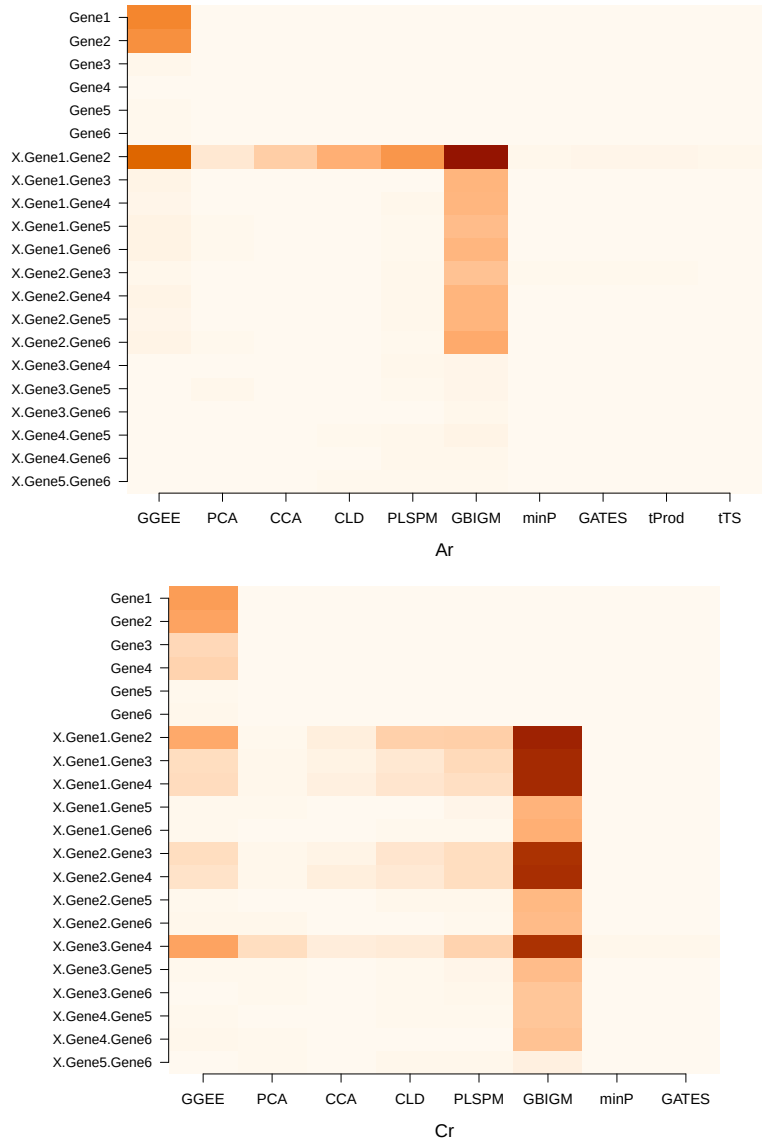


Figure 4.18: Discoveries matrices for simulation settings Ar and Cr described in Table 4.4 for a larger number of methods proposed in the package GeneGeneInteR. Each square of each matrix represents the number of times where a variable was detected as significant over the total number of iterations.

We also investigate the performance of other methods proposed in the package GeneGeneInteR for the two first settings with real genotypes. The results are presented in the Figure 4.18. Overall we can see that G-GEE presents a bet-

ter power to detect the simulated interaction in both settings over the other approaches. The lack of power of the other methods may be explained by the fact that for several iterations, some of the methods were not able to complete a final result.

Finally, we also consider a fourth scenario where binary phenotypes were simulated following the model used by Emily [2016]:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{rs}\sum_{(j,k)\in\Omega} \gamma_{jk}^{rs} X_{ij}^{r} X_{ik}^{s}, \tag{4.3}$$

where Ω is a set of selected SNP pairs and γ_{jk}^{rs} the interaction effect size for the SNPs (j,k) that respectively belong to the genes (r,s). The coefficient associated to each pair is defined as follows:

$$\gamma_{jk}^{rs} = \frac{\log(\tau)}{16} |\log_{10}(\text{MAF}(\mathbf{X}_j^r) \times \text{MAF}(\mathbf{X}_k^s))|, \tag{4.4}$$

where $\text{MAF}(\mathbf{X}_j^r)$ is the minor allele frequency for the SNP j of gene r and $\text{MAF}(\mathbf{X}_k^s)$ for the SNP k of gene s.

In this scenario we simulated one interaction effect between gene 1 and gene 2 with five causal SNP pairs: $\mathbf{X}_1^1\mathbf{X}_1^2$, $\mathbf{X}_2^1\mathbf{X}_2^2$, ... , $\mathbf{X}_5^1\mathbf{X}_5^2$. We set $\tau = 2.5$ and defined a MAF=0.1 for each causal SNP.

Figure 4.19 shows the lack of ability of all methods to detect the simulated interaction.

4.3.5 Percentages of R^2 attributable to interaction and main effects respectively

In order to better interpret the results obtained in the three previous sections (4.3.2, 4.3.3 and 4.3.4), notably regarding the distinction of main and interaction signals, we investigate the relative role of each type of effect in the variability of simulated phenotypes. Thus, for each simulated data set, we compute the proportion of R^2 (or R^2_{McF} in the case of binary phenotype) that could be attributed to interaction (p_I) and main effects (p_M) as described in section 4.1.2. In the tables 4.5, 4.6, 4.7 and 4.8 we present the average p_I and p_M obtained for the simulations presented in the discoveries matrices of the three previous sections.

For most settings, the p_I depends on the number of simulated effects. With one interaction and two main effects the R^2 part attributable to interaction effects is around 34% (Table 4.5(B, C, D) for simulated genotypes and Table 4.5(C, D) for real genotypes). For both settings wherein a larger number of

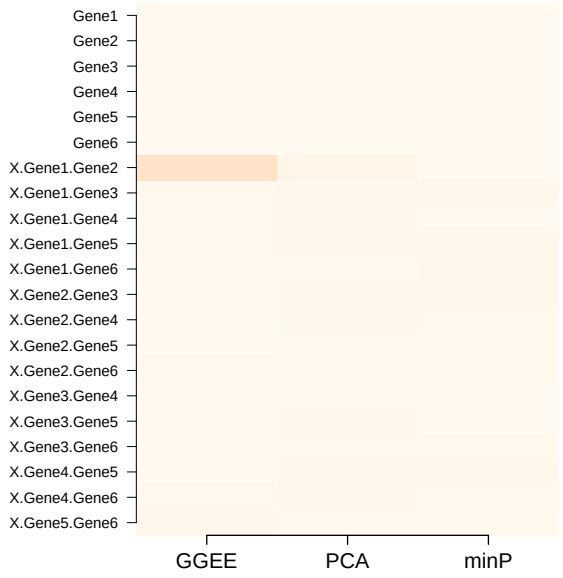


Figure 4.19: Discoveries matrix when phenotype are simulated following the model of Emily [2016]. Each square of each matrix represents the number of times where a variable was detected as significant over the total number of iterations.

Genotypes \ Settings		A		B		C		D	
Simulated	R^2	0.7	0.05	0.7	0.05	0.7	0.1	0.7	0.1
	p_I	0.98	0.92	0.33	0.33	0.34	0.34	0.34	0.33
	p_M	0.99	0.96	0.66	0.67	0.67	0.66	0.67	0.67
Real data	R^2	0.4	0.1	0.4	0.1	0.4	0.1	0.4	0.1
	p_I	0.94	0.93	0.52	0.50	0.33	0.34	0.33	0.34
	p_M	0.99	0.98	0.78	0.78	0.67	0.66	0.67	0.66

Table 4.5: Average proportion of R^2 attributable to interactions and main effects, by setting when genotypes are either completely simulated or derived from a real data set (simulations presented in section 4.3.2 and in section 4.3.3) A, B, C, D refer to the different simulation settings described in Table 4.2 or Table 4.3.

Genotypes \ Settings		25 G	40 G
Simulated	R^2	0.7	0.7
	p_I	0.67	0.67
	p_M	0.34	0.34

Table 4.6: Average proportion of R^2 attributable to interactions and main effects, when genotypes are completely simulated and when we vary the number of genes (25 or 40 genes) considered in the simulated data set. (simulations presented in section 4.3.2).

Genotypes \ Settings		E		F		OME
Real data	R^2	0.4	0.1	0.4	0.1	0.4
	p_I	1	1	1	1	0
	p_M	0	0	0	0	1

Table 4.7: Average proportion of R^2 attributable to interactions and main effects, by setting when genotypes come from a real data set (simulations presented in section 4.3.3). E, F and OME refer to the different simulation settings described in Table 4.3.

variables is considered in the data sets (Table 4.6), the average p_I is 67% because we consider four interaction effects for only two main effects. Finally, as expected, when only interaction effects are simulated, the average p_I is 100% (Table 4.7(E, F)) and 0% when only main effects are simulated (Table 4.7 (OME)). However, the R^2 distribution between main and interaction effects is not distinguishable in the setting where the phenotype is simulated using the "Wang Pathway" model with the same main and interaction effects. The p_I and p_M values are all above 90% (Table 4.5(A)). When real genotypes are considered in the simulations, the R^2 distribution is also not well divided between main and interaction effects when the phenotype is simulated under the PCA model, though p_M is still higher than p_I (Table 4.5(B) for real data).

Genotypes \ Settings		A	C	E
Simulated	p_I	0.97	0.23	1
	p_M	0.99	0.60	0
Real data	p_I	0.94	0.27	1
	p_M	0.98	0.55	0

Table 4.8: Average proportion of R^2 attributable to interactions and main effects, by setting when phenotypes are binary (simulations presented in Section 4.3.4). A, C, E refer to the different simulation settings described in Table 4.4.

Table 4.8, presents the average p_I and p_M obtained for each simulation when binary phenotypes are considered using the McFadden's R-square R^2_{McF}. The conclusions of these results are similar to those obtained with continuous phenotypes. We can see that the proportion of the R^2 relative to marginal and interaction effects cannot be distinguished when the same genes are considered for both effects: all the proportions are above 90%. When main and interaction effects concern distinct genes we can see that p_M is superior to p_I as two main effects are simulated for only one interaction and that $p_I = 100\%$ when only interaction effects are simulated.

4.3.6 Investigation of the use of non linear methods to represent interactions

In the previous section we have seen the good performance of G-GEE over various other modeling approaches. However, one limit of this method lies in the size of the considered genes. Indeed, G-GEE needs the calculation of the matrix $\boldsymbol{F}^{rs}$, that contains all pairwise products of SNPs for each couple of genes, what could be computationally challenging for very large genes. In order to bypass this problem we chose to investigate new manners to construct interaction variables by using non linear approaches. In this last section, we present the results obtained when comparing G-GEE, PCA and PLS with these new approaches.

We use genotypes derived from a real data set (corresponding to the one presented in Section 4.3.3). We consider the same parameters values than the ones fixed at the end of Section 4.3.1 except for the number of subjects (763), the number of iterations (500) and will consider only one value for the coefficient of determination $R^2 = 0.4$.

We consider the two following scenarios:

Settings			Names
Id	Main effects	Interaction effects	Real data
1	Genes 1, 2	Genes 1 x 2	A
2	Genes 1, 2	Genes 3 x 4	C

Table 4.9: Effects simulated in the two settings

In those simulations we split the subjects of the data set in two equal parts. The first part of the data set is used to construct interactions variables and to estimate the coefficients with the Group LASSO. The second part is used in the cleaning Ridge to compute permuted p-values for each group. We investigated the use of Random Forest (RF), Boosting (BOOST), Support Vector Machine

(SVM) and Neural Network (NN). For each couple of genes we model an interaction variable by the predictions obtained using each method. The details of the construction of interaction variables using these non-linear methods are presented in Section 3.3.2.

Regarding Random Forest, we construct two types of interaction variables, the first one using the predictions obtained when constructing a Random Forest on the SNPs belonging to both genes and the phenotype $(\mathbf{X}^r, \mathbf{X}^s, \mathbf{y})$ and the second one using the predictions obtained when constructing a Random Forest using the phenotype and the matrix of all pairwise products $(\boldsymbol{F}^{rs}, \mathbf{y})$. We will refer to the first strategy as RF and the second one as RF_F. For Support Vector Machine, we consider the data $(\mathbf{X}^r, \mathbf{X}^s, \mathbf{y})$ and use four different kernels, linear kernel (SVMlin), polynomial kernel of degree $d = 3$ (SVMpol) or $d = 5$ (SVMpol5) and radial basis (SVMrad).

Results of the two settings are presented in the Figure 4.20. We can see that the methods give different results in the detection of simulated effects, but do not seem to outperform G-GEE overall. We can first note that Random Forest is not convincing for both settings. When using the predictions calculated on the couple of genes $(\mathbf{X}^r, \mathbf{X}^s)$ no interaction was detected and all the signal was detected as marginal. When the predictions are calculated on the matrix $\boldsymbol{F}^{rs}$, RF based method is able to assign interaction signals but with a large number of false positives. Regarding SVM based methods, we observe a good performance to detect the interactions when using the linear kernel in the setting A. This good performance is however nuanced by the fact that the method generates an important amount of false discoveries among interactions when marginal effects are simulated without the corresponding interaction. We can also point out that the method is not able to detect any marginal signal. The simulated main effects are detected as interactions and with a higher frequency than the real simulated interaction. Radial basis kernel seems to act in a similar way to linear kernel but with less power. When using a polynomial kernel (degree $d = 3$ or $d = 5$), the simulated interaction can be detected when the corresponding main effects are also present but with a not negligible amount of false discoveries. Overall the use of polynomial kernels tends to favor the detection of marginal effects over interactions. We can remark that Boosting tends to present a low power to detect marginal or interaction effects and tends to generate false discoveries when different genes are simulated for both types of effects. Finally, we were able to detect interaction effects when using Neural Network but with a certain number of false discoveries. As when using SVM with linear kernel, the simulated main effects are detected as interactions and with a higher frequency than the real simulated interaction.

Figure 4.21 presents the execution time of each method to compute interaction variable and to estimate coefficients with Group LASSO. Overall we can

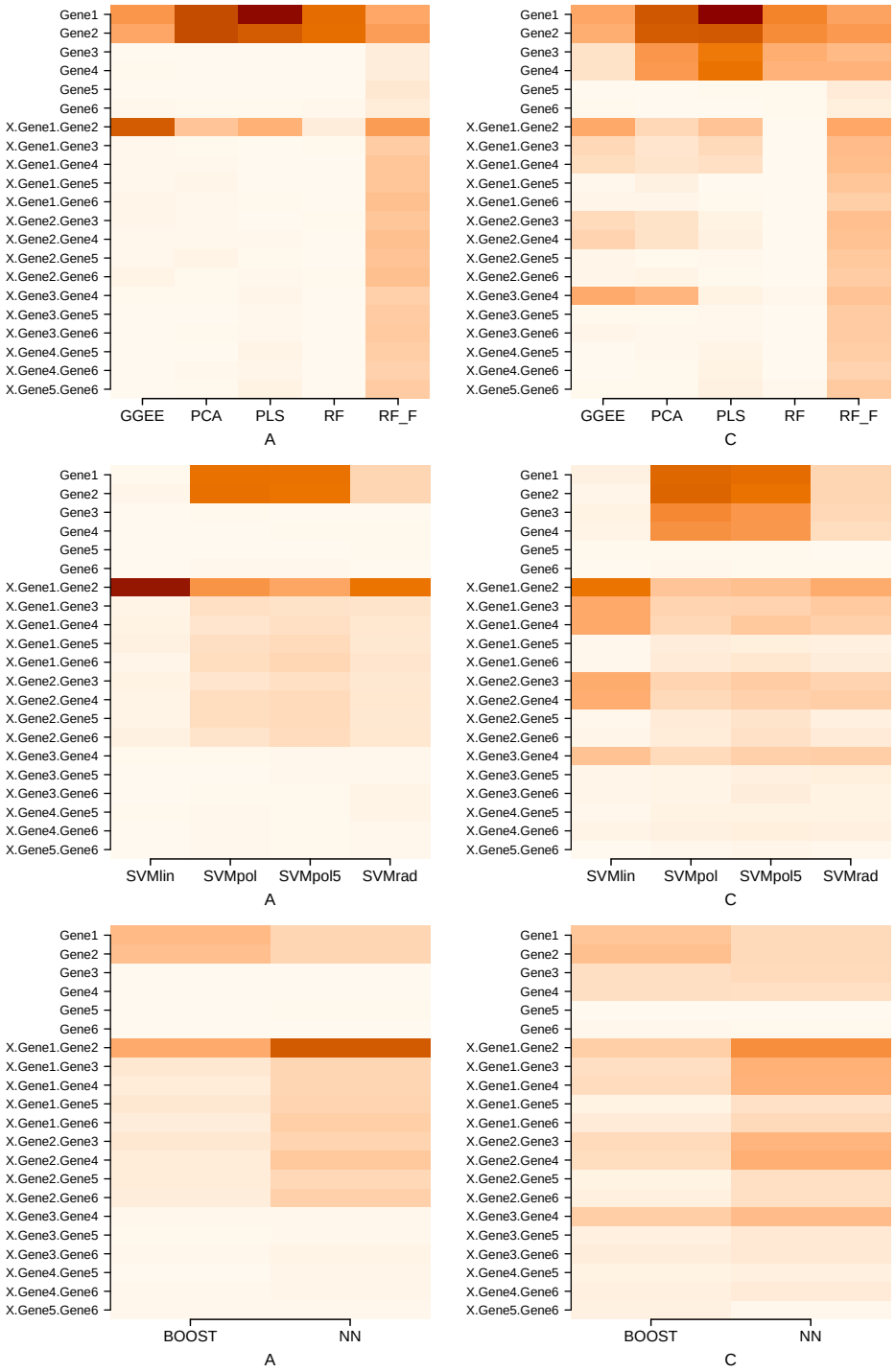


Figure 4.20: Discoveries matrices for the two simulation settings A and C. Each square of each matrix represents the number of times where a variable was detected as significant over the total number of iterations.

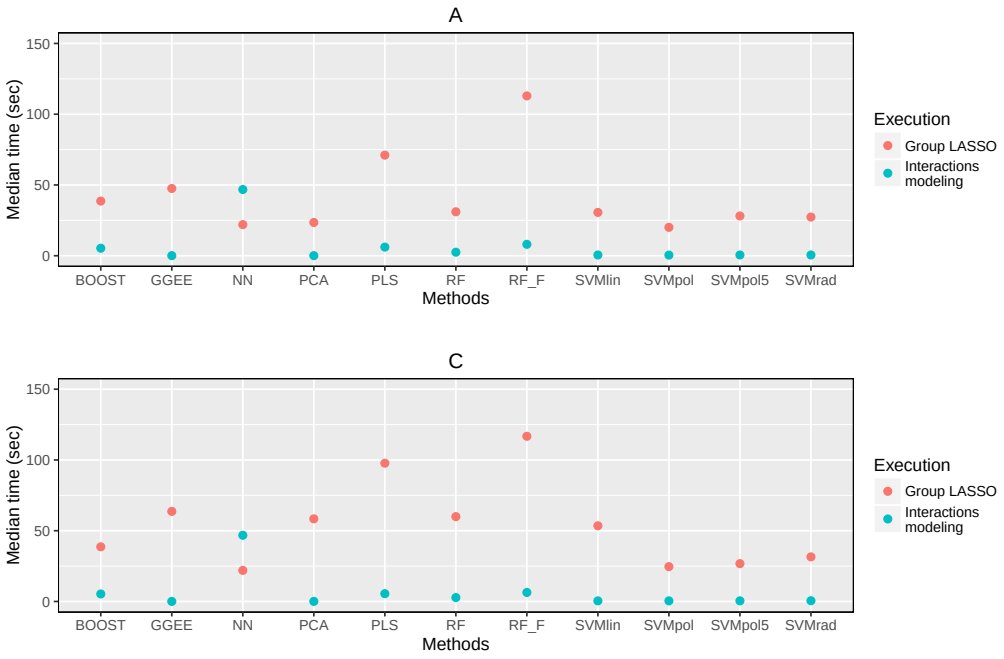


Figure 4.21: Median of the execution time to model interaction and to fit Group LASSO for the two settings.

see that the estimation of the coefficients is the longest step in the procedure except for NN based method. We can also note that the time to estimate coefficients is particularly time-consuming for RF_F and PLS and increases in the second setting for PLS, G-GEE, PCA, RF and RF_F.

In a last simulation we investigated the performance of G-GEE, PCA and NN based methods when the part of the phenotype explained by the marginal effects is first removed from the phenotype variable using the two same settings.

The model to estimate becomes the following:

$$\mathbf{y} - \hat{\mathbf{y}}_{ME} = \boldsymbol{X}^T\boldsymbol{\beta} + \boldsymbol{Z}^T\boldsymbol{\gamma} + \epsilon.$$

Here $\hat{\mathbf{y}}_{ME}$ corresponds to the prediction obtained using the Group LASSO on the matrix of all SNPs $\boldsymbol{X}$:

$$\hat{\mathbf{y}}_{ME} = \boldsymbol{X}^T\hat{\boldsymbol{\beta}}_{GL}.$$

Figure 4.22 shows that G-GEE and PCA based methods detect the interaction for only a small number of iterations. We can conclude that a part of the interaction was captured by the model with only marginal effects.

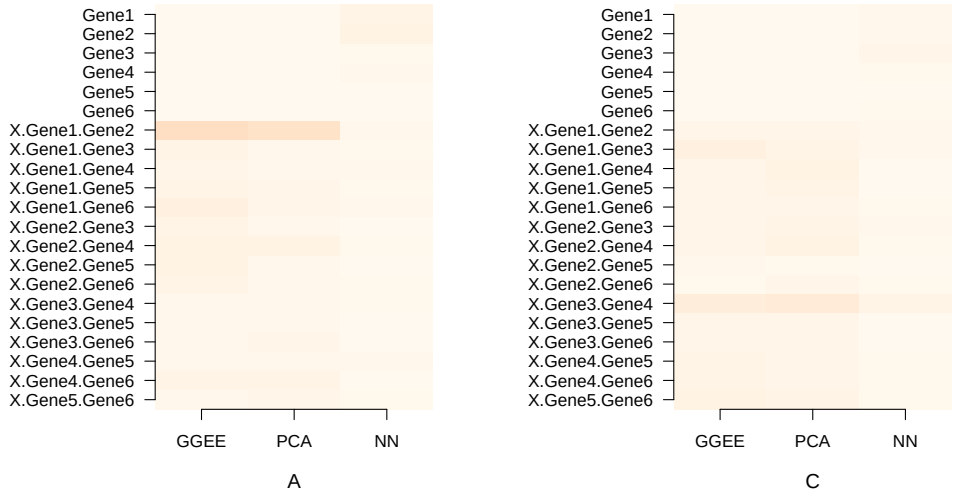


Figure 4.22: Discoveries matrices for the two simulation settings A and C. Each square of each matrix represents the number of times where a variable was detected as significant over the total number of iterations.

4.4 Investigation of two pathologies

To illustrate the use of our proposed framework, we applied the G-GEE, PLS and PCA based methods on two datasets related to ankylosing spondylitis and Crohn's disease.

4.4.1 Quality controls and filtering

For each data set, we first realized a data quality control (see Section 1.2.4) and excluded the markers and subjects based on the following criteria:

- Markers filtering:
  - SNPs call rate $\leq 95\%$,
  - MAF $\leq 5\%$,
  - Deviation from the Hardy Weinberg Equilibrium in unaffected samples ($p < 1 \times 10^{-5}$) ,
  - Marker duplicates. If the duplicates are not equal, one of the markers is randomly chosen.
  - SNPs not belonging to one and unique gene.
- Subjects filtering:
  - Sample call rate $\leq 93\%$,
  - Verification of the presence of duplicates.

For the two data sets, missing values were replaced by the median value of the corresponding markers on all subjects. Gene affiliation of each marker was obtained either using the NCBI2R R package [Melville, 2012] or via the application programming interface for NCBI databases using Biopython. We considered only SNPs located within a single gene in order to form groups of genes without overlap.

We investigated the presence of epistasis using the three approaches presenting the best performances in the simulations studies namely, G-GEE$_{c2}$ (referee as G-GEE), PLS and PCA based approaches. When using PCA and PLS we used two components to define interaction variables. For each analysis, subjects were randomly split in two groups of equal size and we used the same partition for the three methods. The first half of the data was used to construct the interaction variables and to estimate the Group LASSO coefficients. The second half of the sample was used for the cleaning stage to compute permuted p-values for each group. The number of permutations B to estimate p-values for each group was fixed to 1000.

4.4.2 Ankylosing spondylitis

Presentation

Ankylosing spondylitis (AS) is a common form of inflammatory arthritis predominantly affecting the spine and pelvis. It occurs with a prevalence of 0.1% to 1.4% depending on the considered population [Tsui et al., 2014]. Genetic factors account for more than 90% of the risk of susceptibility to AS. Human leukocyte antigen (HLA) class I molecule HLA B27, was the first genetic risk factor identified as associated with ankylosing spondylitis in the 1970's [Schlosstein et al., 1973; Woodrow and Eastmond, 1978] and remains the most important risk locus for this pathology. Despite the strong association only a small portion of HLA-B27 carriers develop the disease. Furthermore, studies in families indicate that less than 50% of the overall genetic risk is due to HLA-B27, what suggests that other genetic factors are involved [Thomas and Brown, 2010]. HLAB-27 is a gene belonging to the Major Histocompatibility Complex (MHC) region which is one of the most dense region of the genome. The MHC is located on the short arm of chromosome 6 and controls a major part of the immune system. A number of updated reviews on AS genetics identified new MHC associated genes as well as genes outside of the MHC region [Tsui et al., 2014; Reveille et al., 2010; Brown et al., 2016]. Additionally, several studies showed that ERAP1 variants influence risk of ankylosing spondylitis in HLA-B27 positive but not negative individuals, suggesting an epistatic effect of both genes [Cortes et al., 2015; Evans et al., 2011].

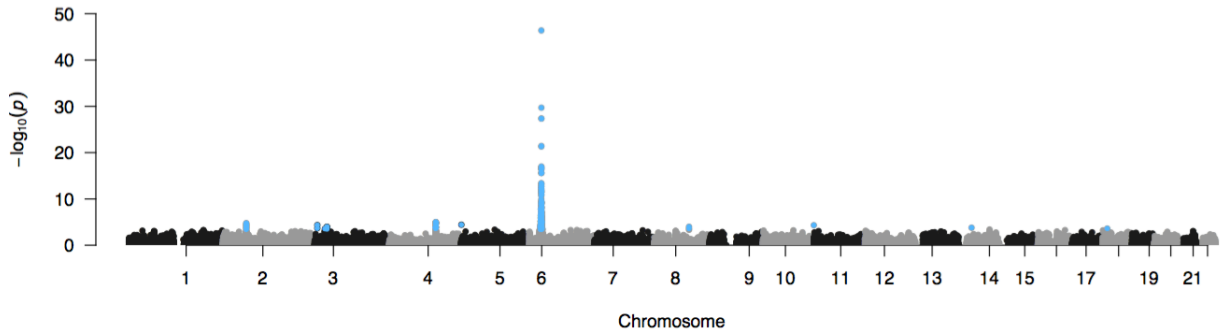


Figure 4.23: Manhattan plot of the GWAS realized for ankylosing spondylitis. Significant SNPs after Benjamini-Hochberg correction are highlighted in blue.

Data set

The dataset regarding ankylosing spondylitis consists of the French subset of a large study of the International Genetics of Ankylosing Spondylitis (IGAS) study [Cortes et al., 2013]. For this subset, unrelated cases were recruited through the Rheumatology clinic of Ambroise Paré Hospital (Boulogne-Billancourt, France) or through the national self-help patients' association: "Association Française des Spondylarthritiques". Population-matched unrelated controls were obtained from the "Centre d'Etude du Polymorphisme Humain", or were recruited as healthy spouses of cases. The protocol was reviewed and approved by the Ethics committee of the Ambroise Paré hospital. All participants gave their informed consent to the study.

The data contain 763 individuals (357 cases, 401 controls and 5 without assigned phenotype) genotyped for 118,826 SNPs with the Immunochip technology of Illumina. The Immunochip microarray is a microarray that has been specifically designed for the study of autoimmune diseases such as psoriasis or ankylosing spondylitis [Cortes and Brown, 2011]. It focuses on particular genome regions thought to be related to this type of diseases.

After pre-processing, 51,287 SNPs belonging to 6,611 genes are retained in the analysis for 758 individuals (357 cases, 401 control).

Analysis

We first realized a GWAS in order to identify markers marginally associated with ankylosing spondylitis. These univariate analyses allowed us to identify 637 significantly associated SNPs, a large number of them being located on chromosome 6 (Figure 4.23). Among those 637 SNPs, 188 belong to a single gene for a total number of 62 genes. We also decided to consider 29 other genes

RUNX3	Runt-related transcription factor 3
IL23R	Interleukin 23 receptor
IL12Rβ2	Interleukin 12 receptor, β2
GRP25	G-protein-coupled receptor 25
KIF21B	Kinesin family member 21B
PTGER4	Prostaglandin E receptor 4 (subtype EP_4)
ERAP1	Endoplasmic reticulum aminopeptidase 1
ERAP2	Endoplasmic reticulum aminopeptidase 2
LNPEP	Leucyl/cystinyl aminopeptidase
IL12B	Interleukin 12B
CARD9	Caspase recruitment-domain family member 9
LTβR	Lymphotoxin β-receptor (TNFR superfamily, member 3)
TNFRSF1A	Tumor-necrosis factor-receptor superfamily member 1A
NPEPPS	Aminopeptidase puromycin-sensitive
TBκBP1	TNFR-associated factor family member-associated nuclear factor-κB-binding kinase 1-binding protein
TBX21	T-box 21
IL6R	Interleukin 6 receptor
FCGR2A	Fc fragment of immunoglobulin G, low-affinity IIa, receptor (CD32)
UBE2E3	Ubiquitin-conjugating enzyme E2E 3
GPR35	G-protein-coupled receptor 35
NKX2-3	NK2 homeobox 3
ZMIZ1	Zinc finger, MIZ type-containing 1
SH2B3	Src homology 2B adaptor protein 3
GPR65	G-protein-coupled receptor 65
IL27	Interleukin 27
SULT1A1	Sulfotransferase family cytosolic 1A
TYK2	Tyrosine kinase 2
ICOSLG	Inducible T-cell costimulator ligand
EOMES	Eomesodermin
IL7R	Interleukin 7 receptor
BACH2	BTB and CNC homology 1, basic leucine-zipper transcription-factor 2

Abbreviation: CD, classification determinant.

Figure 4.24: Summary of ankylosing spondylitis-susceptibility genes identified by GWAS from Tsui et al. [2014]

that were identified as having a main effect in previously published GWAS (Figure 4.24). Epistasis was investigated among those 91 genes and we detected the following effects:

Method	Variables
G-GEE	NKX2-3 x HCG27
PCA	HLA-B
	HCP5
	HLA-B x HCG27
PLS	HLA-B
	EOMES x HCP5
	IL1R2 x MICB
	ZFP57 x LOC101929772
	TRIM31 x HCG26

We can note that the three methods identified interaction effects but for different couples. In addition, none of them confirmed the previously identified interaction between ERAP1 and HLA-B. Only the PCA and the PLS based approaches detect marginal effects, both identify HLA-B and PCA additionally detect HCP5, a MHC gene. Among the genes showing interaction effects NKX2-3, EOMES and IL1R2 correspond to genes already identified as associated with ankylosing spondylitis while HCG27, HCG26, HLA-B, HCP5, MICB, ZFP57 and TRIM31 are genes located in the MHC which is an important region for ankylosing spondylitis.

4.4.3 Crohn's disease

Presentation

Crohn's disease (CD) is one of the subtypes of Inflammatory Bowel Disease (IBD) with ulcerative colitis (UC). Both pathologies present a similar incidence (10-30 per 100,000 in Europe and North America) [Pascal et al., 2017] and are characterized by a chronic inflammation of the gastrointestinal tract. However, they differ in the location and type of lesions. CD can affect any part of the intestine (from the mouth to the anus) and is characterized by chronic and discontinuous transmural lesions of the gut wall. In UC, inflammation is confined to the colon and rectum, and lesions are continuous and superficial [Mathew and Lewis, 2004]. Although the etiology of Inflammatory Bowel Disease is not completely understood, there is recognition that IBD occurs at the intersection of three distinct spheres, namely genetic, external environment and gut microbiota with no factor sufficient in itself for the development of the disease [Ananthakrishnan, 2015]. Various external environment influences are suspected to be related to IBD and have also been suggested by the increase of IBD incidence along with the industrialization of societies [Molodecky et al.,

2012]. Microbiome analysis has demonstrated the key role of the interface between the immune response and the gut microbiota and a recent study showed that CD and UC are two distinct subtypes of IBD at the microbiome level [Pascal et al., 2017].

Concerning genetic factors, the first Crohn's disease associated gene, NOD2 (also known as CARD15), was described in 2001 [Hugot et al., 2001; Ogura et al., 2001] and was followed by IBD5 the same year [Rioux et al., 2001]. A large number of associated loci have subsequently been identified in various studies [Jostins et al., 2012; Franke et al., 2010; Wellcome Trust Case Control Consortium, 2007; Barrett et al., 2008]. Considering that the identified genes associated with Crohn's disease present only modest individual effects, the existence of potential gene-gene interactions has also started to be investigated [Achkar and Fiocchi, 2009]. Possible interactions between NOD2 and TLR proteins have been suggested in some studies [van Heel et al., 2005; Török et al., 2009; Martinez-Chamorro et al., 2016]. Török et al. [2009] thus provide evidence for genetic interactions between polymorphisms in TLR9 and NOD2 but also with other CD-associated variants as IL23R and DLG5 while Martinez-Chamorro et al. [2016] detected an epistatic interaction between the genes NOD2 and TLR4. In 2011 Abad et al. [2011] used the MDR method to study the interaction between NOD2 and TLR10 but the interaction analysis revealed no statistical epistasis, suggesting different signaling pathways for both genes. Possible interactions between CTLA4 with variants in NOD2 and IL23R genes were also observed [Hradsky et al., 2010] while interactions between IBD5 and NOD2 [Mathew and Lewis, 2004] or between IBD5, ATGL16L1 and IL23R [Okazaki et al., 2008] have also been suggested but require further investigations.

Data set

We applied our approach to the Wellcome Trust Case-Control Consortium genome-wide association dataset for Crohn disease. The data contains 1949 case for 461,896 SNPs genotyped with the Affymetrix 500K SNP chip. The control group was constituted of 1500 individuals. After quality contol step, 140,487 SNPs belonging to 17,304 genes were retained for 1500 controls and 1938 cases.

Analysis

Figure 4.25 shows that the 67 SNPs significantly associated with Crohn's disease are located all along the genome with some peaks on chromosomes 1, 2, 5, and 16. Among those 67 SNPs, 60 belong to a single gene for a total number of 32 genes. To look for interactions, we also considered 72 genes identified as associated with Crohn's disease in published studies [Jostins et al., 2012; Franke et al., 2010; Barrett et al., 2008] as well as 6 of the genes suspected of

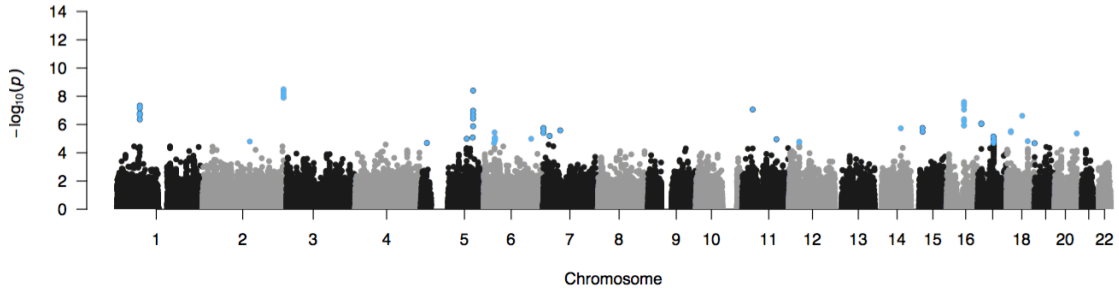


Figure 4.25: Manhattan plot of the GWAS realized for Crohn's disease. Significant SNPs after Benjamini and Hochberg correction are highlighted in blue.

presenting interactions in the introduction. We end up with a list of 110 genes to investigate. The results are presented in the following Table:

Method	Variables
G-GEE	LOC105369715 x STAT1
	STAT1 x CD6
PLS	IFNGR1 x SBNO2
	IRGM x NOD2
PCA	IRGM
	LOC101929544 x TLR4
	BATF x IL10

G-GEE detected two interactions, both of them involving the gene STAT1; PLS detected two interactions including different genes; PCA detected two interactions and one marginal effect. Most of the effects identified by the three methods are attributed to genes already known or suspected to be related to Crohn's disease such as NOD2, STAT1, CD6, IRGM, IL10, SBNO2 and TLR4. These genes all play a role in the immunity. For example, STAT1 is involved in multiple immune system functions, such as the body defense against the Candida fungus while IRGM provides instructions for making a protein that helps trigger autophagy in cells infected with certain kinds of bacteria. PCA was able to detect an interacting effect for the gene TLR4 but not with NOD2, like it has recently been suggested by Martinez-Chamorro et al. [2016].

In conclusion, the three methods G-GEE, PLS and PCA were able to detect interactions for both ankylosing spondylitis and Crohn's disease, but with different genes. Further examinations will be needed to evaluate the real effect of these genes.

Discussion

Our primary concern in this thesis was the detection of interaction effects in the context of the generalized linear model. Through this work we proposed an original approach with several alternatives to model interactions. We compared our proposals with other gene epistasis detecting approaches from literature. We were able to identify possible interaction effects for ankylosing spondylitis and Crohn's disease. Our developed approach G-GEE$_{c2}$, has been implemented in an R package available on github: https://github.com/vstanislas/GGEE.

Contributions and limitations

The first part of our contribution consists in using a Group LASSO for the estimation of the parameters. In comparison to SNP-SNP interaction approaches, the group-scale dimension means that considerably fewer interaction variables need to be considered within a genetic region. Furthermore, resorting to a penalized regression based framework to estimate the model coefficients allows to handle a true multivariate approach over a larger number of genes. This reduction in problem size makes the detection of interactions between different genetic regions possible. This penalized framework also extends initial PCA and PLS proposed gene-scale regression approaches for investigating epistasis.

The second contribution of our work concerns the definition of new types of interaction variables. Among our two propositions, G-GEE$_{c2}$ shows good performance to detect interaction effects over all the methods investigated through various simulation studies. We also note that G-GEE$_{c2}$ tends to detect and assign interaction effects more easily when simulations were realized using real genotypes but is less efficient regarding marginal effects. Since the simulation study based on real genotypes is meant to mimic real data structure, we conclude that G-GEE$_{c2}$ will be better to detect interaction effects than main effects.

A limitation of our proposed approach lies in the size of the considered genes in terms of their number of representative SNPs. Computing the gene Eigen-Epistasis vector for two genes of size p_r and p_s requires an $n \times (p_r p_s)$ matrix to be computed. To bypass this limit we defined new type of interaction variables that did not resort on the computation of the matrix $\mathbf{F}^{rs}$ by the use of non

parametric approaches but we did not obtain results that outperformed the performance of G-GEE$_{c2}$. Another limitation of our framework is the computation cost for estimating the coefficients using Group LASSO, which compares poorly with other methods such as univariate approaches as shown in Section 4.2. Concerning the application of our approach on real data set, as the G-GEE$_{c2}$ method is not able to consider all human genes at the same time, it is necessary to reduce the search space by specifying a limited list of genes. Given that its power to detect main effects is low, it will be safer to use previously acquired knowledge of the genetic effects, or to use a pre-processing method for the detection of main effects. It is however worth noting that our approach is sensitive to the group definition. Some effects may be not detected anymore when the list of genes to investigate is modified.

Finally, focusing at a group scale, our proposed approach might be useful as an initial step but the identified gene interactions will need to be investigated in more details. A first idea will be to resort to SNP-SNP interaction detection methods in a second step to provide more accurate information inside the identified interacting groups. Other possible investigations will be to explore the biological relations of the identified couples in order to see if the identified statistical epistasis correspond to interactions at the biological level.

Perspectives

Several methodological improvements of G-GEE$_{c2}$ could be considered. For example, we could explore new interaction functions to be plugged into our framework in order to bypass the gene size limit of G-GEE$_{c2}$. Another aspect could be to focus on optimizing the computational cost of the matrix $\mathbf{F}^{rs}$ calculation. We also could work on other penalized regressions than the Group LASSO. Then, an interesting direction could consist in constructing interaction variables and estimating the model parameters at the same time. This could be done in devising a more complex optimization problem.

Through our simulations studies, we noticed a confusion phenomenon for all interaction variable modeling approaches when active genes were not simulated with both main and interaction effects. In that case interaction effects tended to be detected as marginal while interacting signals tended to be assigned to single genes, leading to an augmentation of false detections. This phenomenon reveals the difficulty that all methods encounter in clearly distinguishing the different types of effects. Future developments could thus aim at resolving the confusion between interaction and marginal signals.

Other methodological developments may be necessary to better understand the lack of performance of several approaches tested in the simulation studies. In particular, it would be interesting to investigate in more details the use of

machine learning based methods to represent interactions. We could also consider an alternative use of these methods without using the linear model layer which is a more classical approach.

When using G-GEE$_{c2}$ it is necessary to specify a list of genes to be considered for potential interactions. In this thesis, we selected groups of genes to investigate based on the results from univariate analyses or using genes either already identified or suspected to present a link in the literature with the considered disease. As presented in Section 2.3.1, a variety of methods can be used to filters genetic markers to be considered for possible interactions. One promising idea could be to use biological knowledge and select genes belonging to a same pathway or encoding for particular proteins involved in relevant interactions but also to use data mining filtering as TuRF or SURF [Moore and White, 2007; Greene et al., 2009] and thus limit biological bias.

Furthermore, other grouping strategies could be investigated. Through this work we presented results when groups of SNPs correspond to genes, but we could also group SNPs in LD blocks using the clustering approach proposed by Dehman et al. [2015]. The use of LD blocks may be an interesting perspective in order to explore the presence of interactions in larger genomic regions since LD blocks can comprise several genes. For example, for the investigated datasets, many genes were only composed of a small number of genetic markers (less than 5 SNPs). Considering LD blocks would allow to group correlated genes in a unique set and to take into account SNPs that are located in intergenic regions.

Contribution

Publications

Stanislas, V., Dalmasso, C., and Ambroise, C. (2017). Eigen-Epistasis for detecting gene-gene interactions. BMC Bioinformatics, 18(1):54.

Presentations

Jul. 2017 Detecting interactions in GWAS with the Gene-Gene Eigen-Epistasis approach. 38th Annual Conference of the International Society for Clinical Biostatistics (ISCB), Vigo - Spain.

Jun. 2016 Eigen-Epistasis for detecting Gene-Gene interactions in GWAS. Open Days in Biology, Computer Science and Mathematics (JOBIM), Lyon.

Jun. 2016 Eigen-Epistasis for detecting Gene-Gene interactions in GWAS. 48th Days of Statistics - French Society of Statistics (SFdS), Montpellier.

Sep. 2015 Searching gene-gene interactions in GWAS using a Group Lasso approach. Seminar, Statistics Mathematics and Applications, Fréjus.

Posters

Jan. 2017 Detecting interactions in GWAS with the Gene-Gene Eigen-Epistasis approach. SMPGD 2017 - Statistical Methods for Post Genomic Data, London - England.

Mar. 2016 Eigen-Epistasis for detecting Gene-Gene interactions in GWAS. DAGStat 2016 - Statistics under one umbrella, Gottingen - Germany.

Nov. 2015 Detecting gene-gene interactions in GWAS using a Group Lasso approach. Statistical Analysis of Massive Genomic Data, Évry.

Software

Jan. 2016 **Package GGEE**, This package implements the group lasso Gene-Gene Eigen Epistasis (G-GEE) method for detecting epistasis at the gene level. The proposed functions allow to compute various interaction variables for each pair of genes and to fit a general model with a Group Lasso penalty. The package allows to generate gene structured genotype data and continuous phenotype. https://github.com/vstanislas/GGEE

GGEE Vignette

Virginie Stanislas

2016-03-03

1) Introduction

The `GGEE` package implements the "Gene-Gene Eigen Epistasis" approach to detect epistasis at the gene level in genome wide association studies (GWAS). This approach compute interaction variables for each gene pair then uses a penalized regression method based on group lasso to select the significant main or interaction effects.

The two main functions of this package are `BuiltEpiVar` and `GLmodel`. `BuiltEpiVar` allows to calculate interaction variables under four different interaction variable modeling approaches. The Eigen-epistasis approach find for each gene pair a component defined as the linear combination of gene markers (SNPs) having the highest correlation with the phenotype. The three other modeling approaches are inspired by previous literature proposals, they compute interaction variable using: Principal Component Analysis (PCA), Partial Least squares (PLS) or Canonical-Correlation Analysis (CCA). `GLmodel` fits a group lasso model on the genetic data set enhanced by interaction variables then uses a screen and clean procedure in order to compute p-values for each group. A group is either made with the SNPs from a given gene or of interaction terms relative to a given gene-pair interaction.

Additionally, the package allows to generate genotype and phenotype data under two phenotypic models.

2) Generating genotype and phenotype data

The `GGEE` package allows to generate gene structured data and associated continuous phenotype according to the model :

$$\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{\beta} + \boldsymbol{Z}^T\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

Where $\boldsymbol{y} \in \mathbb{R}^n$ denotes the vector of trait values for n individuals, $\mathbf{X} \in \{1,2,3\}^{n\times p}$ represents the SNP matrix, $\mathbf{Z}$ the matrix gathering interaction variables and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ a gaussian error term. The columns of $\mathbf{X}$ are structured on G non overlapping genes. Each gene is described by a given number of SNPs p_g where $\sum_g p_g = p$. The matrix of interaction $\mathbf{Z}$ is structured into $G(G-1)/2$ submatrices each submatrice being the group of interaction variables for a specific pair of genes.

The two functions `simGeno` and `simPheno` allows to respectively simulate genotype and phenotype data.

```
library(GGEE)
sizeGenesMain <- rep(6,2) # 2 genes with 6 SNPs
sizeGenesPair <- rep(6,2) # 2 genes with 6 SNPs
sizeGenesRemain <- rep(6,4) # 4 genes with 6 SNPs
SameMainPair <- FALSE  # Specify that genes with interaction effects will not have main effects
N<- 600
causalSNPnb <- 2
corr <- 0.8
MAFcausalSNP=0.2


Geno <- simGeno(N=N, corr=corr, sizeGenesMain=sizeGenesMain, sizeGenesPair=sizeGenesPair,
     sizeGenesRemain=sizeGenesRemain, SameMainPair=SameMainPair, MAFcausalSNP=MAFcausalSNP,
     causalSNPnb=causalSNPnb)
```

With these parameters the function `simGeno` simulate a data set of 6 genes, each one composed of 6 SNPs, for 600 individuals. The 2 first genes are considered to have main effects and the gene 3 and gene 4 to have an interaction effect. For the four causal genes their 2 first SNPs are considered as causal variants. Rather than a defined number of causal SNPs by causal gene, it is possible to use a portion of causal SNPs with the option `causalSNPportion`. In this case the option `causalSNPnb` has to be `NULL`. In both cases, the SNPs considered as causal are the first listed in the gene. The MAF of each SNP is randomly set between the values `minMAF` and `maxMaf` (by befault `minMAF=0.05` and `maxMaf=0.5`). For the causal SNPs the MAF value correspond to `MAFcausalSNP`. The correlation between SNPs belonging to the same gene is set to `corr=0.8`.

The output of `Geno` contain the following elements :

- The 600×48 genotype matrix X
- The list `listGenesSNP` that indicate the names of the SNPs composing each gene
- The vectors `MainEff` and `GenePair` which give the names of genes having respectively main or interaction effects. The size of the vector `GenePair` is an even number, the pairs being defined with genes successively taken two by two along the vector.
- The vector `MAF` which give the minor allele frequency observed for each simulated SNP

```
Geno$X[1:5,1:8]
```

```
##      Gene.1.SNP.1 Gene.1.SNP.2 Gene.1.SNP.3 Gene.1.SNP.4 Gene.1.SNP.5
## [1,]            1            1            1            1            1
## [2,]            1            1            1            1            1
## [3,]            1            1            1            1            1
## [4,]            1            1            1            1            1
## [5,]            1            1            1            1            1
##      Gene.1.SNP.6 Gene.2.SNP.1 Gene.2.SNP.2
## [1,]            2            1            1
## [2,]            1            1            2
## [3,]            1            1            1
## [4,]            2            2            1
## [5,]            1            1            1
```

```
Geno$listGenesSNP
```

```
## $Genes1
## [1] "Gene.1.SNP.1" "Gene.1.SNP.2" "Gene.1.SNP.3" "Gene.1.SNP.4"
## [5] "Gene.1.SNP.5" "Gene.1.SNP.6"
##
## $Genes2
## [1] "Gene.2.SNP.1" "Gene.2.SNP.2" "Gene.2.SNP.3" "Gene.2.SNP.4"
## [5] "Gene.2.SNP.5" "Gene.2.SNP.6"
##
## $Genes3
## [1] "Gene.3.SNP.1" "Gene.3.SNP.2" "Gene.3.SNP.3" "Gene.3.SNP.4"
## [5] "Gene.3.SNP.5" "Gene.3.SNP.6"
##
## $Genes4
## [1] "Gene.4.SNP.1" "Gene.4.SNP.2" "Gene.4.SNP.3" "Gene.4.SNP.4"
## [5] "Gene.4.SNP.5" "Gene.4.SNP.6"
##
## $Genes5
## [1] "Gene.5.SNP.1" "Gene.5.SNP.2" "Gene.5.SNP.3" "Gene.5.SNP.4"
```

```
## [5] "Gene.5.SNP.5" "Gene.5.SNP.6"
##
## $Genes6
## [1] "Gene.6.SNP.1" "Gene.6.SNP.2" "Gene.6.SNP.3" "Gene.6.SNP.4"
## [5] "Gene.6.SNP.5" "Gene.6.SNP.6"
##
## $Genes7
## [1] "Gene.7.SNP.1" "Gene.7.SNP.2" "Gene.7.SNP.3" "Gene.7.SNP.4"
## [5] "Gene.7.SNP.5" "Gene.7.SNP.6"
##
## $Genes8
## [1] "Gene.8.SNP.1" "Gene.8.SNP.2" "Gene.8.SNP.3" "Gene.8.SNP.4"
## [5] "Gene.8.SNP.5" "Gene.8.SNP.6"
```

```
Geno$MainEff
```

```
## [1] "Genes1" "Genes2"
```

```
Geno$GenePair
```

```
## [1] "Genes3" "Genes4"
```

```
Geno$MAF
```

```
## Gene.1.SNP.1 Gene.1.SNP.2 Gene.1.SNP.3 Gene.1.SNP.4 Gene.1.SNP.5
##   0.19666667   0.19916667   0.17416667   0.26500000   0.05416667
## Gene.1.SNP.6 Gene.2.SNP.1 Gene.2.SNP.2 Gene.2.SNP.3 Gene.2.SNP.4
##   0.26166667   0.20583333   0.19500000   0.18583333   0.22500000
## Gene.2.SNP.5 Gene.2.SNP.6 Gene.3.SNP.1 Gene.3.SNP.2 Gene.3.SNP.3
##           NA   0.47250000   0.19750000   0.21083333   0.29333333
## Gene.3.SNP.4 Gene.3.SNP.5 Gene.3.SNP.6 Gene.4.SNP.1 Gene.4.SNP.2
##   0.47750000   0.35666667   0.13916667   0.19750000   0.21583333
## Gene.4.SNP.3 Gene.4.SNP.4 Gene.4.SNP.5 Gene.4.SNP.6 Gene.5.SNP.1
##   0.35000000   0.40250000   0.27916667   0.20416667   0.20333333
## Gene.5.SNP.2 Gene.5.SNP.3 Gene.5.SNP.4 Gene.5.SNP.5 Gene.5.SNP.6
##   0.20083333   0.14750000   0.11583333   0.24916667   0.45000000
## Gene.6.SNP.1 Gene.6.SNP.2 Gene.6.SNP.3 Gene.6.SNP.4 Gene.6.SNP.5
##   0.19333333   0.19333333   0.45083333   0.09500000   0.49666667
## Gene.6.SNP.6 Gene.7.SNP.1 Gene.7.SNP.2 Gene.7.SNP.3 Gene.7.SNP.4
##   0.40750000   0.19416667   0.20500000   0.44333333   0.26083333
## Gene.7.SNP.5 Gene.7.SNP.6 Gene.8.SNP.1 Gene.8.SNP.2 Gene.8.SNP.3
##   0.45416667   0.42583333   0.19916667   0.19500000   0.41750000
## Gene.8.SNP.4 Gene.8.SNP.5 Gene.8.SNP.6
##   0.18416667   0.40666667   0.43333333
```

Once the genotype matrix obtained, phenotype values can be simulated through the function `simPheno`. The function takes as parameters:

- the outputs of the `simGeno` function,
- two vectors of possible values for coefficients β and γ,
- the number or portion of causal SNPs to consider by gene (It has to be the same value than the one chosen for `simGeno`),

- a r^2 value that calibrate the difficulty of the problem,
- a value for the intercept β_0 (default $\beta_0 = 0$),
- the model to consider to simulate interaction effects:
  - `"SNPproduct"` : $Y_i = \beta_0 + \sum_g \beta_g \left(\sum_{k \in \mathcal{C}} X_{ik}^g\right) + \sum_{rs} \gamma_{rs} \left(\sum_{(j,k) \in \mathcal{C}^2} X_{ij}^r X_{ik}^s\right) + \epsilon_i$
  - `"PCproduct"`: $Y_i = \beta_0 + \sum_g \beta_g \left(\sum_{k \in \mathcal{C}} X_{ik}^g\right) + \sum_{rs} \gamma_{rs} C_{i1}^r C_{i1}^s + \epsilon_i$

where $\mathcal{C}$ and $\mathcal{C}^2$ are respectively the set of causal SNPs and causal interactions, and ϵ_i a random Gaussian variable. For each causal gene g a coefficient β_g is assigned to the standardized sum of the causal SNPs. for the interactions, in the first model `"SNPproduct"`, all the causal SNPs from a causal pair (r, s) are pairwise multiplied and the interaction of the causal pair is represented by the standardized sum of the products. In the second model `"PCproduct"`, the interaction is represented by the standardized product of the first PCA component $\boldsymbol{C}_{.1}^r$ of gene r and the first PCA component $\boldsymbol{C}_{.1}^s$ of gene s. The computation of PCA components is realized on the whole gene and not only on the causal SNPs.

```
# possible values for coef Beta or Gamma
pvBeta <- c(2,2)
pvGamma <-  c(2,2)
r2 <- 0.4

Pheno <- simPheno(X=Geno$X, listGenes=Geno$listGenesSNP, MainEff=Geno$MainEff, GenePair=Geno$GenePair,
        model="SNPproduct", pvBeta=pvBeta, pvGamma=pvGamma, r2=r2, causalSNPnb=causalSNPnb)
```

The outputs of the function `simPheno` includes

- the vector of phenotype continuous values y,
- the matrix G of the simulated main effects, each column represent one causal gene and correspond to the standardized sum of its causal SNPs,
- the matrix GG of the simulated interaction effects, each column represent one causal interaction defined depending of the selected model,
- values for the coefficient of determination R^2 when considering the model containing only simulated interaction effects `R2I` or only simulated main effects `R2S` or both simulated main and interaction effects `R2T`,
- a list `caract` with the characteristic of the simulation. The information about the part of the coefficient of determination R^2 hat can be attributed to either interaction effects $p_{R_I^2} = \dfrac{R_I^2}{R_T^2}$ or main effects $p_{R_M^2} = \dfrac{R_M^2}{R_T^2}$ is given.

```
head(Pheno$y)
```

```
##           [,1]
## V1 -6.2029755
## V2  4.8071671
## V3  3.1516677
## V4  3.0961886
## V5  3.0868815
## V6  0.1770048
```

```
head(Pheno$G)
```

```
##        Genes1     Genes2
## V1 -0.7851669 -0.7974191
## V2 -0.7851669  0.1972825
## V3 -0.7851669 -0.7974191
## V4 -0.7851669  0.1972825
## V5 -0.7851669 -0.7974191
## V6 -0.7851669 -0.7974191
```

```
head(Pheno$GG)
```

```
##      X.Genes3.Genes4
## [1,]    -0.454578817
## [2,]     2.266092409
## [3,]     0.905756796
## [4,]    -0.908024022
## [5,]    -0.908024022
## [6,]    -0.001133613
```

```
Pheno[c("R2T","R2I","R2S")]
```

```
## $R2T
## [1] 0.4288422
##
## $R2I
## [1] 0.1334254
##
## $R2S
## [1] 0.3008971
```

```
Pheno$caract
```

```
## $MainEff
## [1] "Genes1" "Genes2"
##
## $nbSNPbyMainEff
## Genes1 Genes2
##      6      6
##
## $Coef_MainEff
## Genes1 Genes2
##      2      2
##
## $causalSNPMainEff
##      Genes1         Genes2
## [1,] "Gene.1.SNP.1" "Gene.2.SNP.1"
## [2,] "Gene.1.SNP.2" "Gene.2.SNP.2"
##
## $GenePair
## [1] "Genes3" "Genes4"
##
## $nbSNPbyInterGene
## Genes3 Genes4
```

```
##      6      6
##
## $Coef_GenePair
## X.Genes3.Genes4
##               2
##
## $causalSNPInter
##      Genes3         Genes4
## [1,] "Gene.3.SNP.1" "Gene.4.SNP.1"
## [2,] "Gene.3.SNP.2" "Gene.4.SNP.2"
##
## $beta0
## [1] 0
##
## $r2
## [1] 0.4
##
## $causalSNPportion
## NULL
##
## $causalSNPnb
## [1] 2
##
## $R2T
## [1] 0.4288422
##
## $PartR2I
## [1] 31.11293
##
## $PartR2S
## [1] 70.16498
```

3) The G-GEE method

Once genotype and phenotype data are obtained we can apply the G-GEE approach to seek for interaction effects. The first step is to create interaction variables from each gene couple, the second is to test for potential main or interaction effects.

Interaction variable modeling can be done with the function `BuiltEpiVar`. The function takes as parameters the matrix of genotype X, the vector of phenotypic traits y, a list `listGenesSNP` that indicate the names of the SNPs composing each gene and `nbcomp` the number of components to consider to compute interaction variables. Four different methods can be use to create interaction variables :

- `"GGEE"` which find for each gene pair its Eigen-epistasis Component that maximize the correlation between all possible SNP-SNP interactions and the phenotype.
- `"PCA"` which first compute PCA on each gene of the pair and represent the interaction with component products.
- `"PLS"` interaction variables are defined by components that maximize the covariance between the two genes and the phenotype.
- `"CCA"` interaction variables are here represent by the product of pairwise components obtained by a canonical correlation analysis on the gene pair.

Here we show an example using `"GGEE"`option. As this method can compute only one interaction by gene couple, the parameter `nbcomp` doesn't need to be used.

```
Int <- BuiltEpiVar(Geno$X, Pheno$y, method="GGEE", listGenesSNP=Geno$listGenesSNP)
```

```
## [1] "X.Genes1.Genes2"
## [1] "X.Genes1.Genes3"
## [1] "X.Genes1.Genes4"
## [1] "X.Genes1.Genes5"
## [1] "X.Genes1.Genes6"
## [1] "X.Genes1.Genes7"
## [1] "X.Genes1.Genes8"
## [1] "X.Genes2.Genes3"
## [1] "X.Genes2.Genes4"
## [1] "X.Genes2.Genes5"
## [1] "X.Genes2.Genes6"
## [1] "X.Genes2.Genes7"
## [1] "X.Genes2.Genes8"
## [1] "X.Genes3.Genes4"
## [1] "X.Genes3.Genes5"
## [1] "X.Genes3.Genes6"
## [1] "X.Genes3.Genes7"
## [1] "X.Genes3.Genes8"
## [1] "X.Genes4.Genes5"
## [1] "X.Genes4.Genes6"
## [1] "X.Genes4.Genes7"
## [1] "X.Genes4.Genes8"
## [1] "X.Genes5.Genes6"
## [1] "X.Genes5.Genes7"
## [1] "X.Genes5.Genes8"
## [1] "X.Genes6.Genes7"
## [1] "X.Genes6.Genes8"
## [1] "X.Genes7.Genes8"
```

`Int` is a list composed of the interaction variable matrix `XBet` and a vector `interLength` indicating the number of interaction variables for each couple.

```
head(Int$XBet)
```

```
##      X.Genes1.Genes2 X.Genes1.Genes3 X.Genes1.Genes4 X.Genes1.Genes5
## [1,]     -0.37483330     -0.94670703     -0.38919401     -0.08003515
## [2,]     -0.39592814      0.91481607     -0.45494578     -0.92748635
## [3,]     -1.13993875     -0.62887479     -0.08627101     -0.92748635
## [4,]      0.07089144     -0.77947346     -0.44337368     -0.19737197
## [5,]     -0.65541149     -1.09900054     -0.98329433     -1.06206221
## [6,]     -0.98275928     -0.09209224     -1.12114191     -0.92748635
##      X.Genes1.Genes6 X.Genes1.Genes7 X.Genes1.Genes8 X.Genes2.Genes3
## [1,]   -0.9840520014     -1.04211751      0.82969437     -0.74267166
## [2,]   -1.1604971995     -0.09719768     -1.14936788      2.63455984
## [3,]   -0.5770498185     -1.03724033     -1.14936788     -0.70647238
## [4,]   -0.4921236289     -1.04211751     -0.08545953     -0.20605273
## [5,]   -0.0006838417     -1.19063609     -0.86384571     -0.77196252
## [6,]   -0.7251952842     -0.59579760     -0.42786058      0.07351239
##      X.Genes2.Genes4 X.Genes2.Genes5 X.Genes2.Genes6 X.Genes2.Genes7
## [1,]     -0.09313049       0.2318984      -0.7985800      -0.8502372
```

```
## [2,]      0.45555707      -0.2437522      -0.5602747       1.0355134
## [3,]     -0.17521410      -0.9946540      -0.6498773      -1.0609108
## [4,]      0.28202240       0.6780865       0.2360542      -0.5501733
## [5,]     -0.60399652      -0.7052613       0.7138565      -0.8725055
## [6,]     -1.04522407      -0.8485122      -0.6000253      -0.4724529
##      X.Genes2.Genes8 X.Genes3.Genes4 X.Genes3.Genes5 X.Genes3.Genes6
## [1,]       1.2861504      -0.7459509     -0.53651266     -1.22455631
## [2,]      -0.6156480       2.2484680      1.12889685      0.57016583
## [3,]      -1.2127405       0.5246863     -0.53661195     -0.04100884
## [4,]       0.8990921      -0.6381722     -0.41782982     -0.68456483
## [5,]      -0.4856079      -1.0574103     -1.12197499     -0.21699927
## [6,]      -0.2747193      -0.2948713      0.05735485      0.45018384
##      X.Genes3.Genes7 X.Genes3.Genes8 X.Genes4.Genes5 X.Genes4.Genes6
## [1,]      -1.2289787       0.2062986      0.21256666     -0.82009749
## [2,]       3.2952219       0.5603280     -0.33551056     -0.68389022
## [3,]      -0.6615142      -0.8208560      0.07837627      0.77142967
## [4,]      -1.1144847      -0.2816493     -0.02322449     -0.41441027
## [5,]      -1.2289787      -0.9767993     -1.03239192      0.04376543
## [6,]       0.6031412       0.9669974     -1.03038818     -0.81726379
##      X.Genes4.Genes7 X.Genes4.Genes8 X.Genes5.Genes6 X.Genes5.Genes7
## [1,]      -0.8343211      1.27464199     -0.62192832    -0.540364461
## [2,]       0.8602911     -0.67289634     -1.16637670    -0.002284382
## [3,]      -0.1070488     -0.30912262     -0.59863632    -0.910825826
## [4,]      -0.9240035      0.03565863      0.23633828    -0.504340543
## [5,]      -1.1552503     -0.84450730      0.01109224    -1.125938435
## [6,]      -0.7294656     -0.56757659     -0.64043302    -0.562121158
##      X.Genes5.Genes8 X.Genes6.Genes7 X.Genes6.Genes8 X.Genes7.Genes8
## [1,]       1.6829651       1.3868208     -0.12956983      0.11597046
## [2,]      -1.0196164       0.2845163      1.23380165     -0.08384636
## [3,]      -1.0196164       0.7923786      0.73089992      0.08244763
## [4,]       0.6340071       1.1216686      0.01762371     -0.20021082
## [5,]      -0.7815674       0.3919226     -0.30010144     -0.26012786
## [6,]      -0.3694684      -0.2546694     -0.17668121     -1.08647434
```

```
Int$interLength
```

```
## X.Genes1.Genes2 X.Genes1.Genes3 X.Genes1.Genes4 X.Genes1.Genes5
##               1               1               1               1
## X.Genes1.Genes6 X.Genes1.Genes7 X.Genes1.Genes8 X.Genes2.Genes3
##               1               1               1               1
## X.Genes2.Genes4 X.Genes2.Genes5 X.Genes2.Genes6 X.Genes2.Genes7
##               1               1               1               1
## X.Genes2.Genes8 X.Genes3.Genes4 X.Genes3.Genes5 X.Genes3.Genes6
##               1               1               1               1
## X.Genes3.Genes7 X.Genes3.Genes8 X.Genes4.Genes5 X.Genes4.Genes6
##               1               1               1               1
## X.Genes4.Genes7 X.Genes4.Genes8 X.Genes5.Genes6 X.Genes5.Genes7
##               1               1               1               1
## X.Genes5.Genes8 X.Genes6.Genes7 X.Genes6.Genes8 X.Genes7.Genes8
##               1               1               1               1
```

Test for potential main or interaction effects is done with the function `GLmodel`. Parameters include `nlambda`, the length of the grid of possible lambda values, `limitLambda` the number of the largest lambda values among

the grid to consider for the cross validation and `lambda.cri` the criteria for lambda selection (minimum or oneSE value).

```
res <- GLmodel(Geno$X, Pheno$y, Int$XBet, interLength=Int$interLength,
     listGenesSNP=Geno$listGenesSNP, nlambda=100, limitLambda=25, lambda.cri="min")
```

```
##
## LinReg()
## Lambda = 1388.49 1100.355 872.0123 691.0549 547.6492 434.0026 343.9396 272.5663 216.0041 171.1796 135
```

The outputs of `GLmodel` contain:

- the matrix `res_GL.min` giving for each SNP an interaction variable the group lasso coefficient values at the optimal lambda level,
- the matrix `pval.adj` that give adjusted pvalues of each variable with nonzero group lasso coefficient.

```
res
```

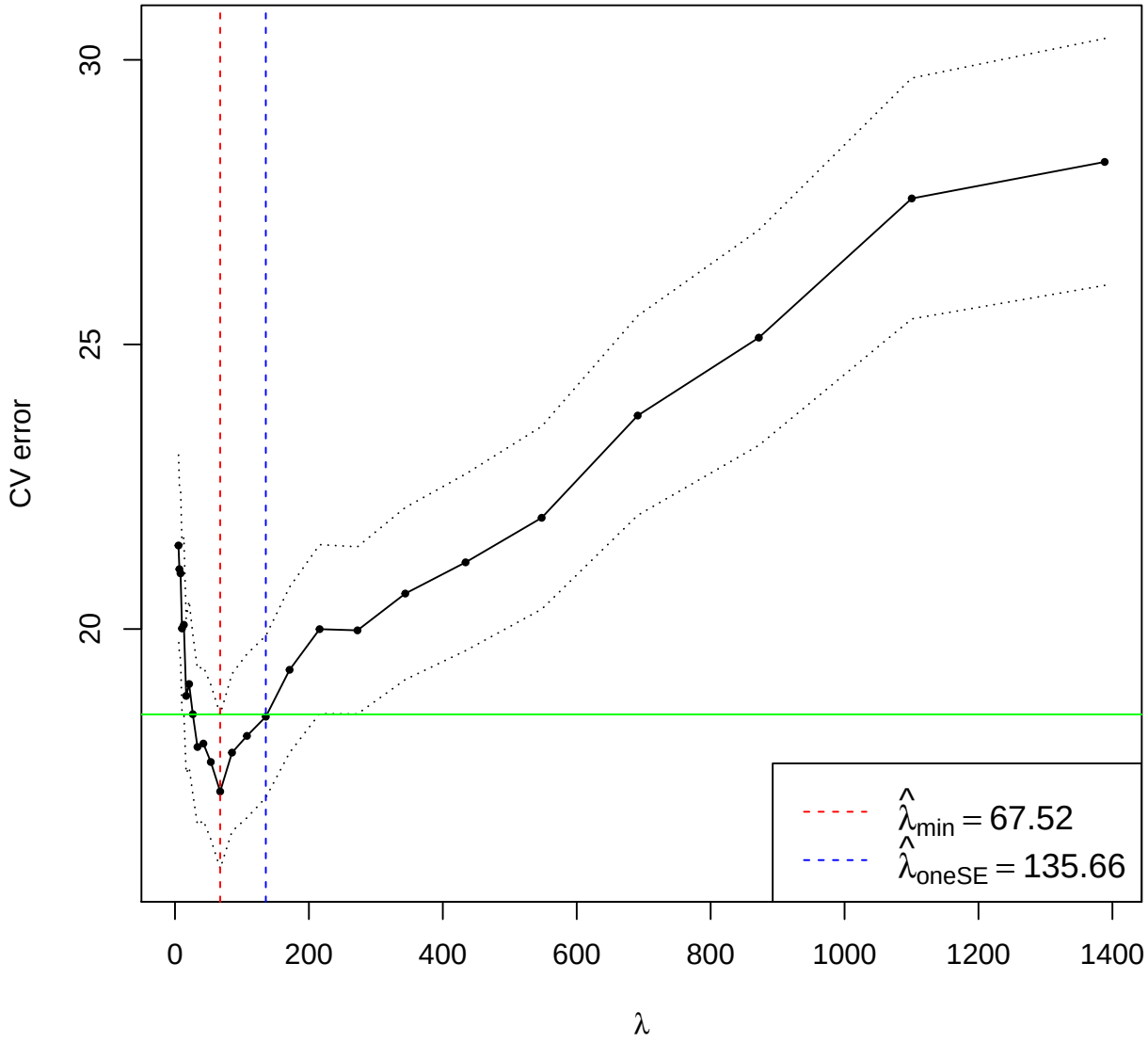
```
## $res_GL.min
##                       Coefs
## Genes1           0.774115652
## Genes1           1.088838532
## Genes1          -0.212205483
## Genes1          -0.359715714
## Genes1           0.251117271
## Genes1           0.077745993
## Genes2           1.034953657
## Genes2           0.751272211
## Genes2          -0.227326819
## Genes2           0.007677615
## Genes2          -0.186737098
## Genes2           0.397955300
## Genes3           0.097299058
## Genes3           0.161389619
## Genes3           0.016012206
## Genes3          -0.036971423
## Genes3          -0.043448945
## Genes3           0.021385553
## Genes4           0.757663956
## Genes4           0.325003095
## Genes4          -0.139607623
## Genes4          -0.194381959
## Genes4          -0.254329496
## Genes4           0.136103406
## Genes5           0.115314233
## Genes5          -0.136702380
## Genes5           0.158719438
## Genes5          -0.128245444
## Genes5          -0.069903366
## Genes5           0.031734515
## Genes6           0.000000000
## Genes6           0.000000000
## Genes6           0.000000000
```

```
## Genes6           0.000000000
## Genes6           0.000000000
## Genes6           0.000000000
## Genes7           0.000000000
## Genes7           0.000000000
## Genes7           0.000000000
## Genes7           0.000000000
## Genes7           0.000000000
## Genes7           0.000000000
## Genes8           0.000000000
## Genes8           0.000000000
## Genes8           0.000000000
## Genes8           0.000000000
## Genes8           0.000000000
## Genes8           0.000000000
## X.Genes1.Genes2  0.000000000
## X.Genes1.Genes3  0.243086593
## X.Genes1.Genes4  0.000000000
## X.Genes1.Genes5  0.401950190
## X.Genes1.Genes6  0.000000000
## X.Genes1.Genes7  0.000000000
## X.Genes1.Genes8  0.000000000
## X.Genes2.Genes3  0.000000000
## X.Genes2.Genes4  0.000000000
## X.Genes2.Genes5  0.235994954
## X.Genes2.Genes6  0.000000000
## X.Genes2.Genes7  0.000000000
## X.Genes2.Genes8  0.000000000
## X.Genes3.Genes4  1.026008870
## X.Genes3.Genes5  0.000000000
## X.Genes3.Genes6  0.000000000
## X.Genes3.Genes7  0.000000000
## X.Genes3.Genes8  0.000000000
## X.Genes4.Genes5  0.000000000
## X.Genes4.Genes6  0.000000000
## X.Genes4.Genes7  0.140106570
## X.Genes4.Genes8  0.000000000
## X.Genes5.Genes6  0.000000000
## X.Genes5.Genes7  0.000000000
## X.Genes5.Genes8  0.000000000
## X.Genes6.Genes7  0.218978412
## X.Genes6.Genes8  0.000000000
## X.Genes7.Genes8  0.620258051
##
## $pval.adj
##                  pval.adj
## Genes1             0.0060
## Genes2             0.0000
## Genes3             0.7680
## Genes4             0.7680
## Genes5             0.7656
## X.Genes1.Genes3    1.0000
## X.Genes1.Genes5    0.9330
## X.Genes2.Genes5    1.0000
```

```
## X.Genes3.Genes4   0.0440
## X.Genes4.Genes7   1.0000
## X.Genes6.Genes7   0.4980
## X.Genes7.Genes8   1.0000
##
## $vc
## $vc$cv.error
##           [,1]     [,2]     [,3]
##  [1,] 26.03778 28.20617 30.37455
##  [2,] 25.44842 27.56409 29.67975
##  [3,] 23.22780 25.11934 27.01087
##  [4,] 21.99849 23.75046 25.50242
##  [5,] 20.34819 21.95416 23.56013
##  [6,] 19.61963 21.17145 22.72326
##  [7,] 19.11130 20.62290 22.13450
##  [8,] 18.50820 19.97642 21.44464
##  [9,] 18.51268 19.99756 21.48243
## [10,] 17.82899 19.28354 20.73808
## [11,] 17.03897 18.46031 19.88165
## [12,] 16.68447 18.12269 19.56090
## [13,] 16.44592 17.82879 19.21167
## [14,] 15.79100 17.14569 18.50038
## [15,] 16.30621 17.66392 19.02162
## [16,] 16.63534 17.98745 19.33957
## [17,] 16.55811 17.92767 19.29724
## [18,] 17.10219 18.50330 19.90440
## [19,] 17.57334 19.03364 20.49394
## [20,] 17.46342 18.82543 20.18744
## [21,] 18.53222 20.07354 21.61485
## [22,] 18.54618 20.01007 21.47396
## [23,] 19.47074 20.97730 22.48385
## [24,] 19.57690 21.05210 22.52729
## [25,] 19.84027 21.46739 23.09450
##
## $vc$lambda
##  [1] 1388.490254 1100.354780  872.012346  691.054871  547.649167
##  [6]  434.002599  343.939638  272.566281  216.004116  171.179568
## [11]  135.656880  107.505758   85.196474   67.516749   53.505869
## [16]   42.402486   33.603245   26.629997   21.103817   16.724414
## [21]   13.253812   10.503420    8.323780    6.596454    5.227577
##
## $vc$lambda.min
## [1] 67.51675
##
## $vc$lambda.oneSE
## [1] 135.6569
##
## $vc$id.lambda.min
## [1] 14
##
##
## attr(,"class")
## [1] "GLmodel"
```

The `GGEE` package contains a plot function `plot.GLmodel`. The function takes as parameter a `GLmodel` object and provides a representation of cross validation results. It depicts the value of the cross validation error for each lambda considered and thus allows to identify the optimal lambda values depending of the criteria of interest (minimal or oneSE). This plot allows to verify that enough lambda values was considered for the cross validation. If the curve doesn't show a clear minimal value the parameter `limitLambda` of the `GLmodel` has to be enlarged.

```
plotGLmodel(res)
```



Stanislas *et al. BMC Bioinformatics* (2017) 18:54
DOI 10.1186/s12859-017-1488-0

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Eigen-Epistasis for detecting gene-gene interactions

Virginie Stanislas*, Cyril Dalmasso and Christophe Ambroise

Abstract

Background: A large amount of research has been devoted to the detection and investigation of epistatic interactions in genome-wide association studies (GWASs). Most of the literature focuses on low-order interactions between single-nucleotide polymorphisms (SNPs) with significant main effects.

Results: In this paper we propose an original approach for detecting epistasis at the gene level, without systematically filtering on significant genes. We first compute interaction variables for each gene pair by finding its Eigen-Epistasis component, defined as the linear combination of Gene SNPs having the highest correlation with the phenotype. The selection of significant effects is done using a penalized regression method based on Group Lasso controlling the False Discovery Rate.

Conclusion: The method is tested against two recent alternative proposals from the literature using synthetic data, and shows good performances in different settings. We demonstrate the power of our approach by detecting new gene-gene interactions on three genome-wide association studies.

Keywords: Genome-wide association study, Gene-gene interactions, Epistasis, Group Lasso

Background

Genome Wide Association Studies (GWASs) look for genetic markers linked to a phenotype of interest. Typically, hundreds of thousands of single nucleotide polymorphisms (SNPs) are studied for a limited number of individuals using high-density genotyping arrays. Usually the association between each SNP and the phenotype is tested using single-marker methods. Multiple markers may also be considered, but these are typically selected using simple forward-selection methods. GWASs are a powerful tool for investigating the genetic architecture of complex diseases and have been successful in identifying hundreds of variants. However, they have been able to explain only a small proportion of the phenotypic variations expected from classical family studies [1]. A number of explanations for this missing heritability have been put forward. For example, it has been suggested that shared environments among relatives are not adequately taken into account. Another suggestion is that much larger numbers of variants with small effects remain to be identified. Rare variants, which are difficult to find using existing genotyping arrays [1], seem to be important causal factors, and so do structural variations. But complex diseases may also be caused, at least in part, by complex genetic structures with multiple interactions between markers (a phenomenon termed *epistasis*). Whereas in pedigree studies the genetic effect on phenotype is seen as part of the additive genetic variance, in GWASs it is seen as an unmeasured interaction between genes [2]. For example, Zuk et al. proposed a model that takes into account epistatic interaction in relation to Crohn's disease [3]. They found that 80% of the missing heritability could be due to genetic interactions.

In recent years a number of methods for studying epistasis have been proposed and reported in various reviews [4–6]. They vary in terms of their data analysis (genome-wide or filtering) and their statistical methodology (Bayesian, frequentist, machine learning or data mining). Most of them focus on single-locus interactions, but considering interactions at the gene level can have several advantages. First, given that genes are the

*Correspondence: virginie.stanislas@math.cnrs.fr
Laboratoire de Mathématiques et Modélisation d'Evry (LaMME), Université d'Evry Val d'Essonne, UMR CNRS 8071, ENSIIE, USC INRA, 23 bvd de France, 91 037, Evry Cedex, Paris, France

functional unit of the genome, results may be more biologically interpretable. Second, genetic effects are more easily detected when SNP effects are aggregated together. Third, gene-based analysis simplifies the multiple testing problem by reducing the number of variables. Several gene-gene methods have been proposed. These are based on a summarizing step which is used to obtain information at the gene level. In more recent methods, filters or penalized models are used to make the method applicable to a large number of genes, while older methods are only applicable to two or a very limited number of genes. For the summarizing step, most methods resort to a principal components (PC) approach, but each method has its specific characteristics. We describe some of these below.

Chatterjee et al. harnessed Tukey's one-degree-of-freedom method to investigate interaction between two genes [7]. Their method is based on the assumption that the SNPs included in each gene region act as surrogates for an underlying biological phenotype. The genotypic information for the gene region is extracted as a single component by a weighted sum of all SNPs. The weights are determined according to the SNP's correlation with the trait. The product of the two sums is then introduced as the gene-gene interaction term into a logistic model, where marginal effects are represented by the respective sums. Building on this idea, Wang et al. compared two different interaction tests [8]. On the one hand, they used *Principal Component Analysis* (PCA) to summarize SNP information within a gene, and on the other hand they used *Partial Least Squares* (PLS) to extract components that summarize, first, the information among SNPs in a gene and, second, the correlation between SNPs and the outcome of interest. They then proposed an interaction test based on either the first PC or the first PLS component for each gene, and were able to show that the PCA and PLS methods often outperformed Tukey's one-degree-of-freedom method. But it is worth noting that the main objective of these three methods was improving the detection of associations in the presence of gene-gene interactions, rather than identifying the interactions themselves. Other approaches based on principal component analysis have since been proposed for epistasis detection. Li et al. proposed selecting, as the gene representation, PCs that are able to explain at least 80% of the variation [9].

Genotypic data are characterized by the high correlation among markers resulting from so-called linkage disequilibrium (LD). Procedures that take LD information into account have been developed for epistasis detection. For example, He et al. proposed an approach using LD information to weight genotype scores which are then aggregated using principal components [10]. Rajapakse et al. developed a gene-based test of interactions for case-control studies which compares LD patterns between cases and controls [11]. Using the same idea, Peng et al. used a canonical correlation-based U-statistic model (CCU) to detect co-association in case-control studies [12]. The idea is to test for two given genes the difference between canonical correlation coefficient computed by Canonical Correlation Analysis (CCA) among cases and among controls. Their work was subsequently extended to include kernel [13, 14].

However most of these methods can be applied only to a reduced number of genes. Computational constraints mean that it is not feasible to model all gene-gene interactions directly. One way of overcoming this is to reduce the gene-gene search space by eliminating unimportant genes, and to this end two-step procedures have been developed that first filter out specific genes or SNPs through a genome-wide search before testing for interactions. One example of this is the model-based kernel machine method (3G-SPA) proposed by Li and Cui, which first performs a search for gene pairs contributing to the overall phenotypic variations [15]. Significant pairs are then tested for interaction effects. Another attractive alternative is offered by penalized regression methods that select a subset of important predictors out of a large number of potential predictors. These methods operate by shrinking the size of the coefficients. The coefficients of predictors with little or no apparent effect are force to be set to zero, reducing the effective degrees of freedom and in many cases making model selection possible. A few approaches using penalized models have been proposed. D'Angelo et al. combined principal component analysis and lasso penalized regression [16]. Wang et al. used a principal component analysis combined with an L1 penalty, with adaptive weights based on gene size, pathway support and effect size [17].

Here we propose a Group Lasso approach [18] that takes into account the group structure of each gene in order to detect epistasis. We introduce *Gene-Gene Eigen-Epistasis* (G-GEE) as a new approach for computing the gene-gene interaction part of the model, and we compare G-GEE with two different interaction variable modeling approaches inspired by previous proposals in the literature, namely PCA and PLS. An adaptive ridge-cleaning approach [19] is then used in order to compute p-values for each group.

In the next section, we detail each model and outline the design of the simulation studies performed to compare the performance of the different approaches. In the Results section, the findings of the simulation studies are shown, and we illustrate our approach on three real datasets relating to ankylosing spondylitis, thyroid carcinomas and inflammatory bowel disease. The different approaches and the results are discussed in the last section.

Methods

We consider n individuals where $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ denotes the vector of trait values. For each individual, genetic variants among G genes are considered. Each gene is described by a given number of SNPs p_g where $\sum_g p_g = p$. The SNP matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ considers an additive coding scheme in which the genotype value of each SNP j from individual i is denoted $X_{ij} \in \{1, 2, 3\}$. $\mathbf{X}_i$ is a p-dimensional vector of covariates for observation i and for $j \in \{1, \ldots, p\}$. $\boldsymbol{X}^g$ denotes the submatrix of $\boldsymbol{X}$ whose columns are the p_g SNPs of gene g. A generalized linear model is generally assumed for GWAS, where the phenotype is considered as a random variable y_i whose conditional expectation can be written as a function of the covariates $\boldsymbol{X}_i$ and their interactions $\boldsymbol{Z}_i$,

$$g\left(E\left[y_i|X\right]\right) = \boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \boldsymbol{\gamma},$$

where

$$\boldsymbol{\beta} = \left(\underbrace{\beta_{1,1}, \beta_{1,2}, \ldots, \beta_{1,p_1}}_{gene_1}, \ldots, \underbrace{\beta_{G,1}, \ldots, \beta_{G,p_G}}_{gene_G} \right)^T,$$

and $\boldsymbol{Z}_i$ is the ith line of the matrix of interactions and $\boldsymbol{\gamma}$ a parameter vector of appropriate dimension. When the phenotype is binary (case control study), it is usual to assume a logistic model where $g()$ is the logit and Y is assumed to follow a binomial distribution. Below we will consider only quantitative phenotypes using a classical linear model. In this case $g()$ is the identity and the residuals are assumed to be Gaussian.

The main effect of each gene is modeled through the sum of the effects of all its SNPs. Concerning interaction effects, we compute new variables representing interactions between two specific genes and define as a group all the interaction variables related to a given pair of genes. The matrix of interaction is thus structured into $G(G-1)/2$ submatrices:

$$\boldsymbol{Z} = \left[\boldsymbol{Z}^{11} \cdots \boldsymbol{Z}^{rs} \cdots \boldsymbol{Z}^{G(G-1)/2} \right]$$

where $\boldsymbol{Z}^{rs}$ describes the interactions between the two genes r and s. The parameter vector $\boldsymbol{\gamma}$ is accordingly structured into sub-vectors $\boldsymbol{\gamma}^{rs}$. We will now present and compare three different approaches for modeling gene-gene interactions.

Modeling gene-gene interactions

Let us consider two genes r and s described respectively by p_r and p_s SNPs. A possible interaction term describing the epistasis between the two genes is

$$\boldsymbol{Z}_i^{rs\,T} \boldsymbol{\gamma}^{rs} = \sum_{j=1}^{p_r} \sum_{k=1}^{p_s} \gamma_{jk}^{rs} X_{ij}^r X_{ik}^s. \quad (1)$$

We hereafter set $\boldsymbol{W}^{rs} = \{X_{ij}^r X_{ik}^s\}_{i=1\cdots n}^{j=1,\cdots,p_r; k=1,\cdots,p_s}$. In this case the submatrix of interactions is $\boldsymbol{Z}^{rs} = \boldsymbol{W}^{rs}$ and $\boldsymbol{\gamma}^{rs} = \{\gamma_{jk}^{rs}\}$ is a vector of size $p_r p_s$. The number of parameters in such a model is obviously too large to be reliably estimated. For this reason a number of papers in the literature consider reducing the dimension of $\boldsymbol{\gamma}$.

In this paper we will consider three different methods for reducing the dimension reduction, namely Principal Component Analysis (PCA), Partial Least Squares (PLS), and our proposed Gene-Gene Eigen-Epistasis approach that we have termed G-GEE.

Principal component analysis

Principal Component Analysis (PCA) can reduce the number of variables describing each gene r from p_r to $q_r < p_r$. Considering gene r described by p_r SNPs, we compute the matrix of the first q principal components

$$C^r = \boldsymbol{X}^r U^r,$$

where U^r is the matrix of the first q_r principal axes. Using C^r and C^s instead of $\boldsymbol{X}^r$ and $\boldsymbol{X}^s$ in the computation of the interaction allows the number of parameters relative to each interaction to be controlled. This control is achieved by choosing the number of principal components q. The PCA model that we describe draws upon ideas in [20]. The interaction term takes the form

$$\boldsymbol{Z}_i^{rs\,T} \boldsymbol{\gamma}^{rs} = \sum_{j=1}^{q} \sum_{k=1}^{q} \gamma_{jk}^{rs} C_{ij}^r C_{ik}^s.$$

Relating this expression to the general form of the interaction term $\boldsymbol{W}_i^{rs}$ described above, we can see that performing PCA prior to computing the interactions is a means of constraining the linear interaction term of Eq. 1.

The submatrix of interactions is $\boldsymbol{Z}^{rs} = \{C_{ij}^r C_{ik}^s\}_{i=1\cdots n}^{j=1,\cdots,q; k=1,\cdots,q}$, and $\boldsymbol{\gamma}^{rs} = \{\gamma_{jk}^{rs}\}$ is a vector of size q^2 describing the interaction between genes r and s. In particular, if a single principal component is chosen, there will be only one parameter to estimate per interaction.

Partial least squares

Wang et al. proposed an alternative method for integrating interactions using a PLS approach [8]. Let $(\boldsymbol{X}^r, \boldsymbol{X}^s)$ be the genotypic matrix for the given pair of genes (r, s). Their approach computes the components that maximize $cov^2(\boldsymbol{X}^r \boldsymbol{u}, \boldsymbol{T}\boldsymbol{v})$, with $\boldsymbol{T} = (\boldsymbol{y}, \boldsymbol{X}^s)$ and $(\boldsymbol{u}, \boldsymbol{v})$ the weight vectors. The interaction of a couple of genes (r, s) is then represented by the first q components:

$$\boldsymbol{Z}_i^{rs\,T} \boldsymbol{\gamma}^{rs} = \sum_{j=1}^{q} \gamma_j^{rs} T_{ij}^{rs}.$$

In this approach phenotypic information is retained when the interaction variables are constructed.

Gene-gene Eigen-Epistasis

We propose an original approach for modeling interactions. The general idea is to consider the interaction variable between the two genes r and s as a function $f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s)$ parameterized by $\boldsymbol{u}$. One way to estimate $\boldsymbol{u}$ is to maximize the correlation between the interaction function and the phenotype:

$$\hat{\boldsymbol{u}} = arg \max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} cov^2 \left(\boldsymbol{y}, f_{\boldsymbol{u}} \left(\boldsymbol{X}^r, \boldsymbol{X}^s\right)\right).$$

If we consider the function f to be linear, our problem becomes easily tractable and has only one solution. Setting

$$\boldsymbol{Z}^{rs} = f_{\boldsymbol{u}} \left(\boldsymbol{X}^r, \boldsymbol{X}^s\right) = \boldsymbol{W}^{rs}\boldsymbol{u},$$

where $\boldsymbol{W}^{rs} = \{X^r_{ij} X^s_{ik}\}_{i=1\cdots n}^{j=1\cdots,p_r;k=1,\cdots,p_s}$ and $\boldsymbol{u} \in \mathbb{R}^{p_r p_s}$ we obtain the following problem:

$$\begin{aligned} \max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} \left\| \hat{cov}\left[\boldsymbol{W}^{rs}\boldsymbol{u}, \boldsymbol{y}\right]\right\|^2 &= \max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} \left\|\boldsymbol{u}^T \boldsymbol{W}^{rs\,T}\boldsymbol{y}\right\|^2 \\ &= \max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} \boldsymbol{u}^T \boldsymbol{W}^{rs\,T}\boldsymbol{y}\boldsymbol{y}^T \boldsymbol{W}^{rs}\boldsymbol{u}. \end{aligned} \tag{2}$$

The solution $\boldsymbol{u}$ is the eigenvector corresponding to the largest eigenvalue of the matrix $\boldsymbol{W}^{rs\,T}\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{W}^{rs}$, which is the vector $\boldsymbol{W}^{rs\,T}\boldsymbol{y}$. The complexity of computing $\boldsymbol{u}$ is therefore in $O(np_rp_s)$. We then use the projection of the matrix $\boldsymbol{W}^{rs}$ on $\boldsymbol{u}$ as the interaction variable. The resulting Eigen-Epistasis vector $\boldsymbol{Z}$ is the linear combination of all the SNP-SNP interactions being the most correlated with the phenotype. In its construction, G-GEE has similarities with PLS. The main difference lies in the original design matrix. PLS searches for components that maximize $cov^2(\boldsymbol{X}^r\boldsymbol{u}, \boldsymbol{y}\boldsymbol{X}^s v)$, whereas G-GEE retains the component that maximizes $cov^2(\boldsymbol{y}, \boldsymbol{W}^{rs}\boldsymbol{u})$, with $\boldsymbol{W}^{rs}$ the matrix of all pairwise interaction between the two genes r and s. Like PLS, G-GEE takes phenotypic information into account in the construction of the interaction variables. Other methods as such as CCU [12] and the kernel versions of CCU [13, 14] that we referred to in the introduction also consider the phenotype in their construction, but these methods can be applied only to case-control problems.

Estimation of coefficients

We propose a Group Lasso approach [18] for estimating the parameters of linear or logistic (case control) regression. A group comprises either the SNPs of a given gene, or interaction terms relative to a given gene-pair interaction. In the particular case of linear regression, the model parameters are estimated by:

$$\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right) = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\operatorname{argmin}} \left(\sum_i (y_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{\gamma})^2 + \lambda \left[\sum_g \sqrt{p_g}\|\boldsymbol{\beta}^g\|_2 + \sum_{rs} \sqrt{p_r p_s}\|\boldsymbol{\gamma}^{rs}\|_2 \right] \right),$$

The parameter λ is selected by cross-validation.

In order to improve estimation accuracy and to obtain p-values for each of the selected groups, we use the adaptive ridge cleaning approach proposed by Bécu et al. [19]. This screen and clean procedure is a two-stage method. The group lasso model is first fitted on half of the data. The coefficient of the candidate groups selected by the model are then introduced into a ridge regression model fitted on the second half of the data with a specific penalty that allows the group structure to be taken into account. For each group the significance of the regression coefficients is estimated using permutation tests.

Simulation design

To evaluate the performance of the proposed approach, we conducted two simulation studies, the first using simulated data and the second using a real dataset relating to ankylosing spondylitis. In each case we compared the proposed G-GEE model to the two other interaction variable modeling approaches. The first simulation corresponds to a simplified context where all parameters were controlled and external interference limited, while the second simulation corresponds to a realistic context with a realistic pattern of minor allele frequency (MAF) and LD.

Design

Genotypes Our first (simplified) simulation study was adapted from the model used in [21] with an extension to control the MAF of each SNP. The n lines of the genotype matrix are an i.i.d. sample from a multivariate random vector $\boldsymbol{X}_i \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. The correlation matrix $\boldsymbol{\Sigma}$ is block diagonal, each block corresponding to a gene. Two variables belonging to the same gene are correlated at level $\rho = 0.8$ while all other correlations are null. Each SNP (column of the genotype matrix) is randomly assigned an MAF p from a uniform distribution between 0.05 and 0.5. An MAF value of 0.2 is assigned to all causal SNPs. The genotype frequencies derived from the Hardy-Weinberg equation are then used to discretize X_{ik} values to 0, 1 or 2. In practice, X_{ik} is set to 1 if $X_{ik} < q_{p^2;N(0,1)}$, X_{ik} is set to 3 if $X_{ik} < q_{(1-p)^2;N(0,1)}$ and X_{ik} is set to 2 otherwise.

In the second (realistic) simulation study using a real ankylosing spondylitis dataset, genes are randomly selected. The number of SNPs composing each genes varies according to the selection.

Phenotypes For both simulation studies, we generated phenotype vectors using two different schemes. Our first scheme corresponds to the model proposed by Wang et al. [17] (which, for the sake of brevity, we will refer to hereafter as the "Wang Pathway" model):

$$Y_i = \beta_0 + \sum_g \beta_g \left(\sum_{k\in C} X^g_{ik}\right) + \sum_{rs} \gamma_{rs} \left(\sum_{(j,k)\in C^2} X^r_{ij} X^s_{ik}\right) + \epsilon_i, \tag{3}$$

where $\mathcal{C}$ and $\mathcal{C}^2$ are respectively the set of causal SNPs and causal interactions, and ϵ_i a random Gaussian variable. For each causal gene g, we consider two causal SNPs and a coefficient β_g is assigned to the standardized sum of these causal SNPs. In the same way, for the interactions, all the causal SNPs from a causal pair (r,s) are pairwise multiplied and a coefficient γ_{rs} is assigned to the standardized sum of the product.

Our second scheme for simulating phenotypes is based on the following model:

$$Y_i = \beta_0 + \sum_g \beta_g \left(\sum_{k \in \mathcal{C}} X_{ik}^g \right) + \sum_{rs} \gamma_{rs} \left(\sum_{(j,k) \in \mathcal{C}^2} C_{ij}^r C_{ik}^s \right) + \epsilon_i. \tag{4}$$

The difference with the first model concerns the simulation of the interaction effect. In the second model the interaction effect for a causal couple (r,s) is defined as the product of the first PCA component $\boldsymbol{C}_{.1}^r$ of gene r and the first PCA component $\boldsymbol{C}_{.1}^s$ of gene s.

In both models, β_0 is set to 0, and ϵ_i are generated independently from a $\mathcal{N}(0,\sigma^2)$, with σ^2 determined from the coefficient of determination R^2 that calibrates the strength of the association. Both simulation models can be written as $y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \boldsymbol{\gamma} + \epsilon_i$ where $\boldsymbol{X}$ the marginal effect genotype matrix and $\boldsymbol{Z}$ the interaction effect matrix.

Let us denote $\boldsymbol{Q\phi} = [\boldsymbol{X}, \boldsymbol{Z}] \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix}$ and

$$\begin{aligned} R^2 &= \frac{\sum (\boldsymbol{Q_i\phi} - \bar{y})^2}{\sum (\boldsymbol{Q_i\phi} + \epsilon_i - \bar{y})^2} \\ &= \frac{\sum (\boldsymbol{Q_i\phi} - \bar{y})^2}{\sum (\boldsymbol{Q_i\phi} - \bar{y})^2 + \sum \epsilon_i^2 + \sum 2\,(\epsilon_i\,(\boldsymbol{Q_i\phi} - \bar{y}))} \\ &= \frac{\sum (\boldsymbol{Q_i\phi} - \bar{y})^2}{\sum (\boldsymbol{Q_i\phi} - \bar{y})^2 + \mathrm{n}\,\hat{\mathrm{var}}(\epsilon_i) + 2\mathrm{n}\hat{\mathrm{cov}}\,(\epsilon_i, \boldsymbol{Q_i\phi} - \bar{y})}. \end{aligned}$$

We remark that:

$$\begin{aligned} 2\mathrm{ncov}\,(\epsilon_i, \boldsymbol{Q_i\phi} - \bar{y}) &= 2\mathrm{ncov}\left(\epsilon_i, \boldsymbol{Q_i\phi} - \frac{\sum_j y_j}{n}\right) \\ &= 2\mathrm{ncov}\,(\epsilon_i, \boldsymbol{Q_i\phi}) - \sum_j \frac{2\mathrm{n}}{\mathrm{n}} \mathrm{cov}\left(\epsilon_i, y_j\right) \\ &= 0 - 2\mathrm{cov}\,(\epsilon_i, \epsilon_i) = -2\sigma^2 \end{aligned}$$

Thus, replacing $\hat{\mathrm{var}}(\epsilon_i)$ by σ^2, and $\hat{\mathrm{cov}}(\epsilon_i, \boldsymbol{Q_i\phi} - \bar{y})$ by $-\sigma^2/n$, we obtain $R^2 \approx \frac{\sum(\boldsymbol{Q_i\phi} - \bar{y})^2}{\sum(\boldsymbol{Q_i\phi} - \bar{y})^2 + \mathrm{n}\sigma^2 - 2\sigma^2}$. This relation between R^2 and σ^2 gives us an expression for σ^2 that depends on R^2, $\sigma^2 = \frac{(R^2-1)\sum(\boldsymbol{Q_i\phi} - \bar{y})^2}{R^2(2-n)}$.

We looked at how much of the coefficient of determination R^2 is explained by main effects, and how much is explained by interaction effects, in order to determine their respective roles in the model.

For a similar reason, when simulating phenotypes, Wang et al. [17] examined how much of partial R^2 was due to interaction effects. They selected coefficient values so that 30% of the partial R^2 was explained by interaction effects. Li and Cui [15] did not use the R^2 directly, but they simulated data assuming different proportions of interaction effects among the total genetic variance. In our study, once the phenotype y had been set for each simulated design matrix, we computed how much of the R^2 could be attributed to interaction and main effects as $p_I = \frac{R_I^2}{R_T^2}$ and $p_M = \frac{R_M^2}{R_T^2}$ respectively, with R_I^2 the R-square value for the model containing only simulated interaction effects, R_M^2 the R-square value where there were only simulated main effects, and R_T^2 R-square value where there were both simulated main effects and simulated interaction effects.

Scenarios In the first (simplified) simulation study, genotypes are simulated as described in the design. We considered six genes, each composed of six SNPs and for 600 subjects. We define one causal interaction between genes and two causal genes with main effects, and the simulation takes place using two alternative simulation settings:

- (1) one interaction and two main effects involving the same genes
- (2) one interaction and two different main effects

For these two settings, different coefficients of determination, from 0.05 to 0.7, are considered and 1000 iterations are performed.

In the second (realistic) simulation study, genotypes come from a real dataset comprising 763 individuals. At each iteration we randomly select six genes of various size (from 1 to 1119 with a median of 2 SNPs) in the dataset. We consider the five following settings:

- (1) one interaction and two main effects involving the same genes
- (2) one interaction and two different main effects
- (3) one interaction effect only
- (4) two main effects only
- (5) no effects

For each setting, coefficients of determination, from 0.1 to 0.4, are considered and 500 iterations are performed.

For both simulation studies, main effects and interaction effects are weighted with the same coefficient values ($\beta_g = \gamma_{rs} = 2, \forall g, r, s$). For each interaction, the power is estimated as the proportion of detected interactions over the total number of simulations.

Real data illustration

To illustrate our approach we applied the proposed method on three real datasets related to ankylosing

spondylitis, thyroid carcinomas and inflammatory bowel disease.

The dataset regarding ankylosing spondylitis consists of the French subset of the large study of the International Genetics of Ankylosing Spondylitis (IGAS) study [22]. For this subset, unrelated cases were recruited through the Rheumatology clinic of Ambroise Paré Hospital (Boulogne-Billancourt, France) or through the national self-help patients' association: "Association Française des Spondylarthritiques". Population-matched unrelated controls were obtained from the "Centre d'Etude du Polymorphisme Humain", or were recruited as healthy spouses of cases. The protocol was reviewed and approved by the Ethics committee of the Ambroise Paré hospital. All participants gave their informed consent to the study. The application on thyroid carcinomas was carried out on a public dataset that came from the study of Luzón-Toro et al. on identification of epistatic interactions in two different types of thyroid carcinomas [23]. Finally, we used the Wellcome Trust Case-Control Consortium genome-wide association dataset to study Inflammatory Bowel Disease.

Results

Simulation studies

In the following, we will refer to the different simulation settings by using letters as described in Table 1.

Results from the simplified simulation study

Figure 1 shows results obtained for the two settings. The first column gives the estimated power to detect the gene interaction as a function of the R^2 values. The last two columns show heatmap matrices reflecting the proportion of significant values for each variable and each method over the 1000 simulations for different R^2 values.

In the first setting (Fig. 1a, b), we consider genes 1 and 2, both having main and interaction effects. When the phenotype is simulated using the Wang Pathway model, the G-GEE and PLS methods have a higher power to detect the interaction effect than PCA method, which tends to identify only the two main effects of the two genes (Fig. 1a). Whereas for PCA and PLS the power is nondecreasing with R^2, for G-GEE we observe a U-shaped curve. For the smallest R^2 values, which correspond to the most difficult cases, the power of G-GEE to detect the interaction tends to decrease. When R^2 values reach 0.4, G-GEE's power to detect the interaction starts to increase. The situation is different for the main effects, since G-GEE's power to detect these increases continuously with R^2 [see Additional file 1]. For PLS, the power to detect the interaction effect is continuously nondecreasing. Note, however, that for this method one of the two main effects (here gene 1) is detected to the detriment of the second, regardless of the value of R^2. In the PCA phenotype simulation model (Fig. 1b), G-GEE has a higher power than the other methods to detect interaction effects while retaining a good specificity, whatever the value of R^2. The reasonably high power of the PCA method can be explained by the similarity between the phenotype simulation model and the estimation model. It is worth noting that in this first setting, only a few variables are falsely significant, which reflects a good specificity for all methods (the worst being for the gene 3 $\times$ gene 4 interaction variable in the case of the Wang Pathway model and $r^2 = 0.1$, where the false discovery rate is 0.068).

In the second setting (Fig. 1c, d), genes 1 and 2 have only main effects, and genes 3 and 4 have only an interaction effect. When the phenotype is simulated using the Wang Pathway model, the interaction power of G-GEE is uniformly higher than that of the other methods (Fig. 1c). For all values of R^2, PCA tends to detect false main effects for genes 3 and 4, but not to detect interaction effects. In the PCA phenotype simulation (Fig. 1d), PCA has a good power to detect interaction effects, but once again these good performances can be explained by the similarity between the simulation model and the estimation model. The interaction power for G-GEE is lower, but still good. With this model, only G-GEE tends to attribute a false main effect to genes 3 and 4. In this second setting, whatever the phenotype simulation model, the power of the PLS method is almost null. PLS identifies only the first gene as having a main effect, while the effects of genes 3 and 4 are not detected, whether as main or as interaction effects. Moreover, PLS tends to attribute a false interaction effect between genes 1 and 2.

To evaluate the performances of the different methods in a more complex context, we also consider a setting where we simulate 25 genes with four causal interactions between genes, and two genes with causal main effects. In these simulations, interaction genes are different from main effect genes, and we only consider the case where $R^2 = 0.7$. The results of this setting reflect the good performance of the G-GEE method over PCA and PLS in detecting interaction in a context where further interactions and different main effects are simulated [see Additional file 2].

Table 1 Effects simulated in each settings and referring names according to the phenotype simulation model

	Settings		Names	
Id	Main effects	Interaction effects	Wang Pathway model	PCA model
1	Genes 1 & 2	Genes 1 x 2	A	B
2	Genes 1 & 2	Genes 3 x 4	C	D
3	-	Genes 1 x 2	E	F
4	Genes 1 & 2	-	OME	
5	-	-	NE	

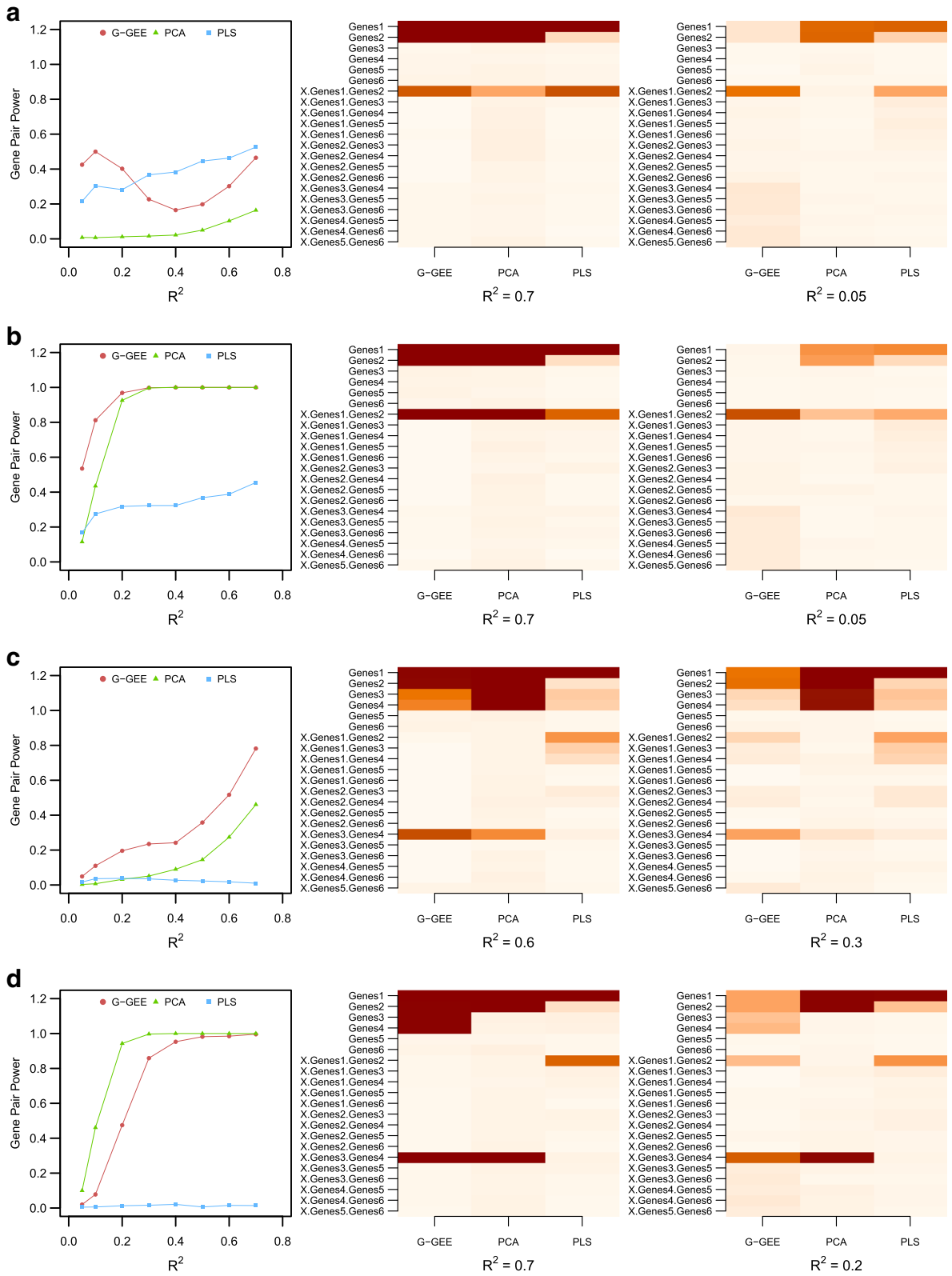


Fig. 1 Power and discoveries under a simplified context. The figures in the first column shows the power to detect interaction effects of the three methods depending on the R^2. The last two columns show the ratio of the number of times where each variable was significant to the total number of simulations for a given R^2. The panels **a**, **b**, **c** and **d** refer to the different simulation settings described in Table 1

Results from the realistic simulation study

Figures 2 and 3 show results for the first three settings. Fig. 2 shows the power to detect gene interaction depending on the R^2, and Fig. 3 shows heatmaps of significant effects when $R^2 = 0.2$. In both figures, the upper row relates to phenotypes simulated using the Wang Pathway model, and the lower row to phenotypes simulated using the PCA model.

In the first setting (Fig. 2a, b, Fig. 3a, b), the same two genes (genes 1 and 2) are simulated with main and interaction effects. In this setting G-GEE has the best power to detect interaction effects for all R^2 values. The interaction power of PLS remains close to 0.3. The power of PCA depends on the phenotype simulation model. When the phenotype is simulated using the Wang Pathway model, the power is similar to PLS. When it is simulated using PCA it increases continuously, because of the similarity between the phenotype simulation model and the estimation model. Looking at the heatmaps (Fig. 3a, b) we can see that only a few variables are falsely significant. We also observe that unlike G-GEE, PCA and PLS can detect the two simulated main effects, with a preference for gene 1 in the case of PLS. These results are obtained when $R^2 = 0.2$, but other R^2 values give similar results [see Additional file 3].

In the second setting (Fig. 2c, d, Fig. 3c, d), main effects are simulated for genes 1 and 2, and one interaction is simulated between genes 3 and 4. In this setting the power of PLS to detect interaction effects is almost null, while the respective powers of PCA and G-GEE are different, according to which phenotype simulation model is used. Both methods have a higher power when the phenotype is simulated using the PCA model. Regarding the detection of main effects, the results are similar to the first setting, with G-GEE less successful than PCA and PLS (Fig. 3c, d). But unlike in the first setting, here some variables are falsely significant. False detections among interaction variables are more pronounced for G-GEE and concern genes that have been simulated to have only main effects. False detections among main effects are more pronounced for PCA and PLS when the phenotype is simulated using the Wang Pathway model and concern genes that have been simulated to have an interaction effect. Under the PCA phenotype model, false detections among main effects are more pronounced for PLS and G-GEE when R^2 values are higher [see Additional file 3].

In the third setting, where only one interaction is simulated between genes 1 and 2, G-GEE has a higher power to detect interaction than PLS and PCA when the phenotype is simulated using the Wang Pathway model (Fig. 2e).

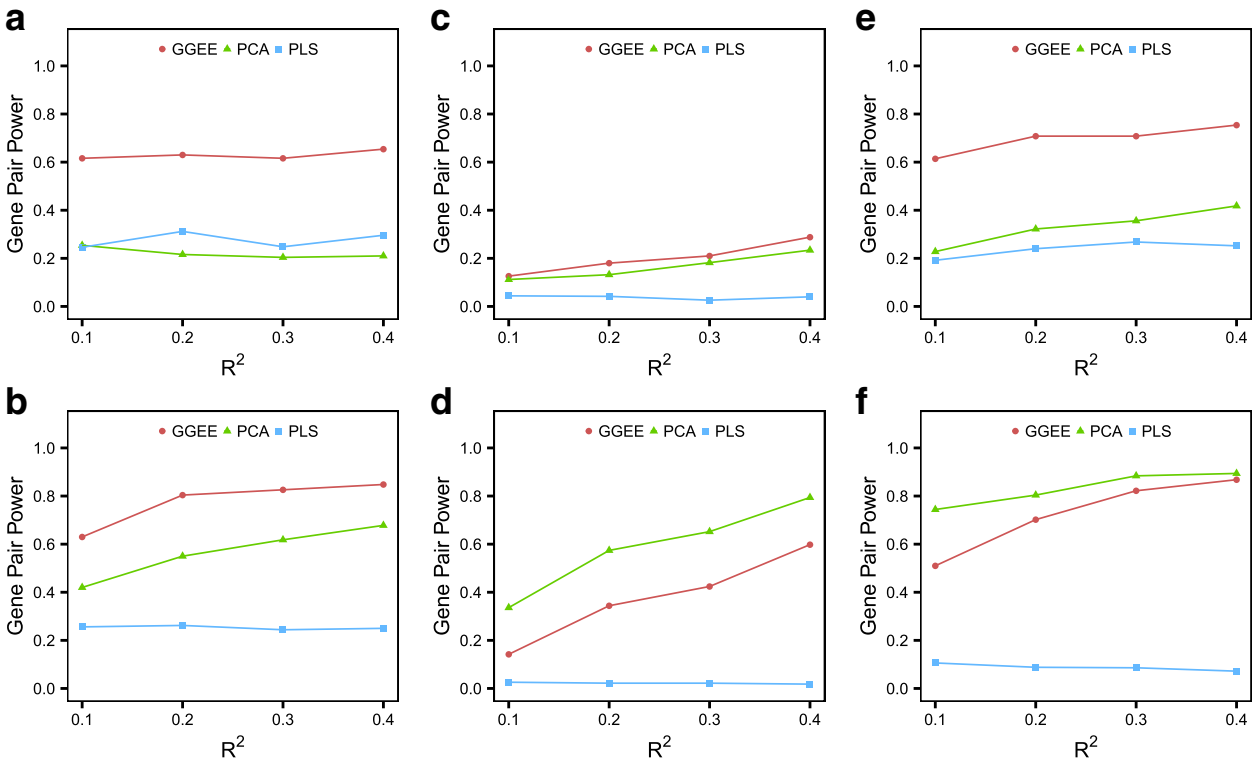


Fig. 2 Power under a realistic context. The figures show the power to detect interaction effects of the three methods depending on R^2. The panels **a**, **b**, **c**, **d**, **e** and **f** refer to the different simulation settings described in Table 1

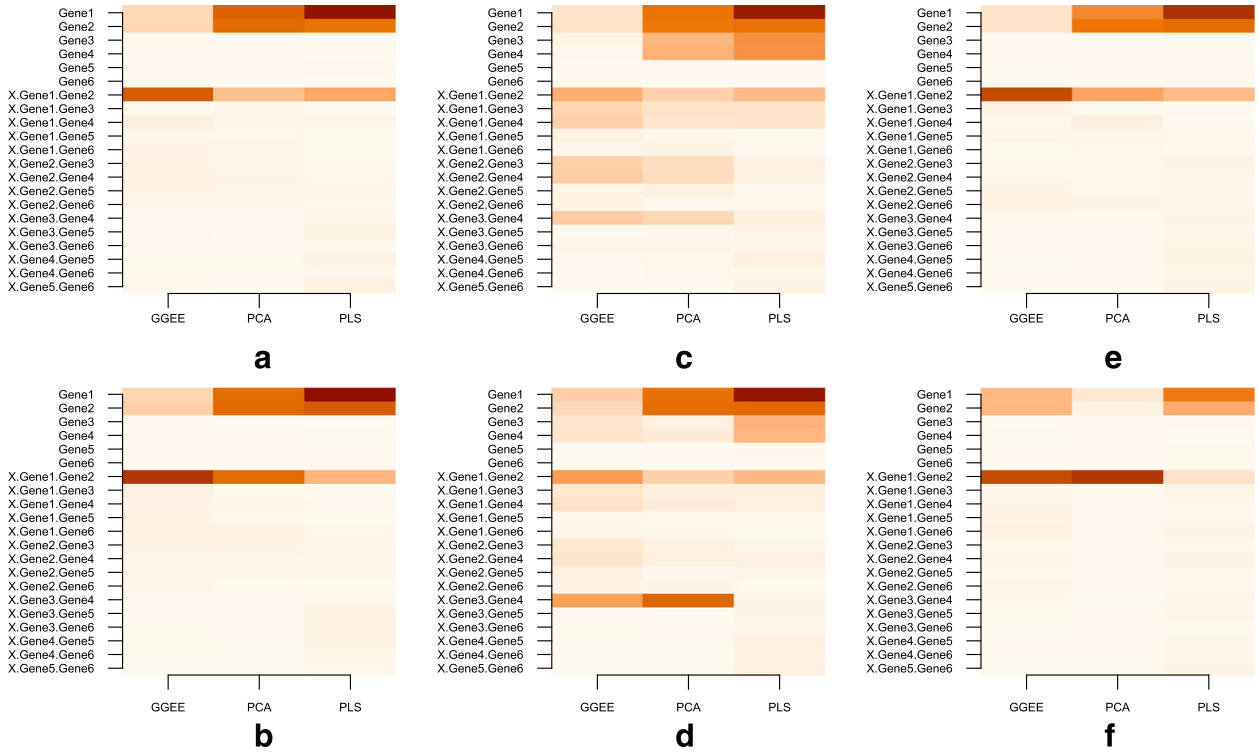


Fig. 3 Discoveries under a realistic context. Heatmaps of the ratio of the number of times where each variable was significant to the total number of simulations for $R^2 = 0.2$. The panels **a**, **b**, **c**, **d**, **e** and **f** refer to the different simulation settings described in Table 1

The power of PCA is higher in the PCA phenotype simulation model because of its similarity to the estimation model, whereas the power of PLS is almost null (Fig. 2f). In the Wang Pathway phenotype simulation model, PCA and PLS both falsely detect main effects. In the PCA phenotype simulation model, the false detections are made by PLS and G-GEE (Fig. 3e, f). In all cases these false detections concern genes that are simulated to have an interaction effect.

Figure 4 shows the results for the fourth and fifth settings. The heatmap on the left corresponds to the fourth setting, where only two main effects are simulated. We remark that all methods successfully identify the main effect, PCA and PLS doing so with a higher power. False

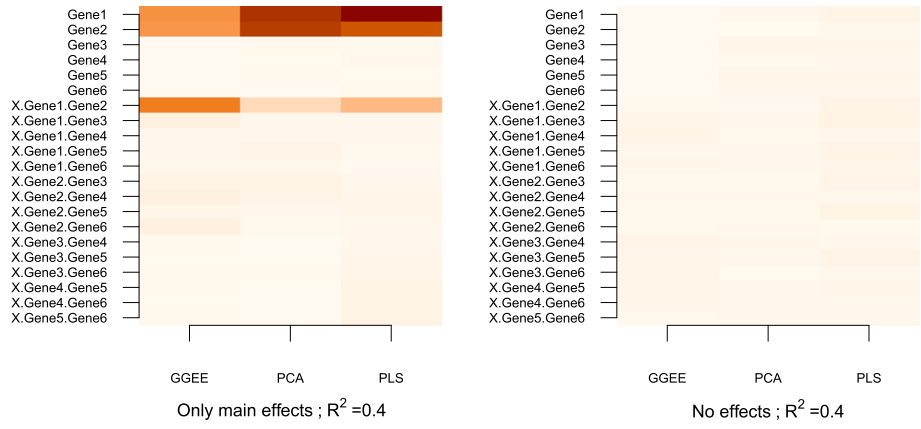


Fig. 4 Discoveries for the fifth and the sixth settings. Heatmaps of the ratio of the number of times where each variable was significant to the total number of simulations for $R^2 = 0.4$ when only main effects are simulated for gene 1 and gene 2 (*left*), and when no effects are simulated (*right*)

detections corresponding to the respective interaction effects are observed for G-GEE, and to a lesser extent for PLS. The figure on the right corresponds to the fifth setting, where no specific effects are simulated and the result shows that all three methods perform well with very few false detections.

In all settings, estimating the coefficients with the group lasso is more computationally expensive than constructing the interaction variables. G-GEE and PCA are quite similar in terms of computation time, whereas in some settings PLS has a slightly greater execution time than other methods. Note that the time required by G-GEE for constructing the interaction variables varies according to the number of SNPs that constitute each gene [see Additional file 4].

Percentage of R^2 attributable to interaction and main effects respectively

Using each setting in both simulation studies, we determine the p_I and p_M average values that correspond to the proportion of the R^2 attributed to interaction and main effects, respectively. For most settings, the p_I depends on the number of simulated effects. With one interaction and two main effects simulated the R^2 part attributable to interaction effects is around 33% (Table 2 (B, C, D), Table 3 (C, D)). For the setting with numerous effects [see Additional file 2], the average p_I is 67% because we consider four interaction effects for only two main effects. Finally, as expected, when only interaction effects are simulated, the average p_I is 100% (Table 3 (E, F)) and 0% when only main effects are simulated (Table 3 (OME)). However, the R^2 distribution between main and interaction effects is not distinguishable in the setting where the phenotype is simulated using the Wang Pathway model with the same main and interaction effects. The p_I and p_M values are all above 90% (Table 2 (A), Table 3 (A)). In the second simulation study, the R^2 distribution is also not well divided between main and interaction effects when the phenotype is simulated under the PCA model, though p_M is still higher than p_I (Table 3 (B)).

Real data illustrations

Ankylosing spondylitis

Ankylosing spondylitis (AS) is a common form of inflammatory arthritis predominantly affecting the spine and pelvis. It occurs with a prevalence of 0.1% to 1.4% depending on the considered population [24]. Genetic factors account for more than 90% of the risk of susceptibility to AS. Human leukocyte antigen (HLA) class I molecule HLA B27, belonging to the Major Histocompatibility Complex (MHC) region, was the first genetic risk factor identified as associated with ankylosing spondylitis in the 1970's [25, 26] and remains the most important risk locus for this pathology. Despite the strong association only a small portion of HLA-B27 carriers develop the disease. Furthermore, studies in families suggest that less than 50% of the overall genetic risk is due to HLA-B27, which suggests that other genetic factors are involved [27]. A number of updated reviews on AS genetics, including genome-wide association study (GWAS) results, identified new ankylosing spondylitis-susceptibility genes outside of the MHC region [28, 29].

We applied all the methods described above to the AS dataset. The data contain 408 cases and 358 controls, and each individual was genotyped for 116, 513 SNPs with Immunochip technology. For each SNP we obtained detailed genetic information, such as gene affiliation, with the NCBI2R package [30] which annotates lists of SNPs with current information from NCBI. We considered only SNPs located within a single gene in order to form gene groups without overlap. We focused our analysis on a list of 29 genes previously identified as having a main effect in GWAS.

The three methods tested yield different results, and only the PLS and G-GEE methods identify interactions. PCA detects only the main effect HLA-B and identifies no interactions. PLS detects the main effect HLA-B, but also identifies one interaction effect between the genes EOMES and BACH2. Our method G-GEE does not detect any main effects, but it shows two significant interactions, the first between the genes HLA-B and SULT1A1 and the second between IL23R and ERAP2.

Thyroid carcinomas

Thyroid cancers are thought to be related to a number of environmental and genetic predisposing factors and can be classified in various types and subtypes. Most association studies have focused on main effects but only a limited number of genes were identified. Recently, some papers focus on the detection of epistatic interactions [23, 31]. We applied our proposed approach on the two data sets used in Luzón-Toro et al. [23] regarding

Table 2 Average percentage of R^2 attributable to interaction and main effects, by setting, in the first simulation study

	A		B		C		D	
	R^2=0.7	R^2=0.05	R^2=0.7	R^2=0.05	R^2=0.6	R^2=0.3	R^2=0.7	R^2=0.2
p_I	97.73	92.08	33.11	32.80	33.32	33.47	33.51	33.57
p_M	98.84	95.57	66.42	66.97	66.60	66.57	66.70	66.56

Table 3 Average percentage of R^2 attributable to interaction and main effects, by setting, in the second simulation study

	A R^2=0.2	B R^2=0.2	C R^2=0.2	D R^2=0.2	E R^2=0.2	F R^2=0.2	OME R^2=0.4
p_I	94.01	52.04	33.36	33.27	100	100	0
p_M	99.08	78.62	66.60	66.87	0	0	100

two rare tumours, sporadic medullary thyroid carcinoma (sMTC) and juvenile papillary thyroid carcinoma (jPTC). Affymetrix Genome-Wide Human SNP 6.0 arrays were used to hybridized DNA. The data set related to sMTC contains 66 cases and the jPTC data set 30 cases. The same 125 healthy controls and 232, 607 SNPs were used for both studies. As for the ankylosing spondylitis dataset, we obtained gene affiliation for each SNP with the NCBI2R package and considered only SNPs located within a single gene. We focused the analysis of the sMTC data set on a list of 10 genes, 3 of these genes (CHFR, AC016582.2 and C8orf37) were chosen following the conclusions of Luzón-Toro et al., the others because they contained markers that were susceptible to be associated with the disease from univariate analysis. The analysis of the jPTC data set was realized on a list of 20 genes among them we can cite DIO, RP11-648K4.2, LOXL1, DMGDH, PAX8 and STK17B from which epistatic interaction were already detected (even if the interaction between PAX8 and STK17B was identified in a study concerning papillary thyroid and not the juvenile form). The 14 others genes contained susceptible associated individual markers from univariate analysis. Regarding the sMTC study, G-GEE identifies one interaction between genes NCK1 and TRIQK. PCA detects only one main effect for the gene TRIQK whereas none effects were identified with PLS. Concerning the jPTC data set, 3 interactions were identified by G-GEE (NCAM1 and MNDA, MNDA and STK17B, LOC105370481 and STX3). PLS identifies 2 interactions (LOC105370236 and LOC105370481, LOC105370236 and PIKFYVE) and PCA detects only one main effect for the gene LOC105370481. We note that the effects detected with our approach concerned different genes that the ones identified in the presented previous studies (except for the gene STK17B). More analysese are needed to better understand these differences.

Inflammatory bowel disease

Although the etiology of Inflammatory Bowel Disease (IBD) is not completely understood, previous studies have underlined the contribution of an important genetic susceptibility. Recently, Martinez-Chamorro et al. [32] detected an epistatic interaction between the genes NOD2 and TLR4. We applied our approach to the Wellcome Trust Case-Control Consortium genome-wide association dataset for Inflammatory Bowel Disease. The data contains 1949 case for 159 960 SNPs genotyped by Affymetrix. The control group was constituted of 1972 individuals from the Wellcome Trust Case-Control Consortium genome-wide association dataset for hypertension. As for the two previous real data analysis, we obtained gene affiliation for each SNP with the NCBI2R package and considered only SNPs located within a single gene. The analysis was realized on a list of 22 genes that contain SNPs that are suspected to be associated with IBD from an univariate analysis. The two genes NOD2 and TLR4 were added to the list as they were previously detected as having an epistatic interaction. G-GEE identifies one interaction between the genes LOC105376008 and CACNB2 whereas PCA detects 9 main effects (IL23R, PODN, ATG16L1, C5orf56, DNAH11, LOC105378282, HSD17B12, LINC00558, ADCY4) but none interaction. Finally PLS identifies 3 main effects for the genes IL23R, PODN and DNAH11 as well as 2 interactions the first one between the genes PODN and FCRLA, the second one between PVT1 and NOD2.

Discussion

The results obtained in both simulation studies point to a certain confusion between main and interaction effects. When simulated interaction and main effects involve different genes, the methods tend to detect as interaction effects the pairs of genes simulated to have main effects and, conversely, to detect as main effects the genes simulated to having interaction effects.

Overall, G-GEE tends to detect more false interactions than false main effect whereas PLS and PCA tend to detect more false main effects though PLS tends to attribute a false interaction effect between genes 1 and 2. This type of confusion may explain the U-shaped power curve for G-GEE observed in the first simulation study (Fig. 1a). As the problem becomes harder, the genetic effects of both genes are preferentially assigned to the interaction effect, implying a better power to detect interaction where R^2 values are small. Finally, we remark that for G-GEE false detections of main effects are more frequent when the PCA phenotype simulation model is used, whereas for the PLS and PCA methods, where the number of false detections for main effects is higher when the Wang Pathway phenotype simulation model is used.

Other observations regarding the power of the different methods can be made with these simulation results. PLS has more trouble than PCA and G-GEE in detecting interaction effects, and has a tendency to detect the first main

effect with a higher power than the second main effect when two main effects are simulated. For all methods, the power to detect interactions increases more slowly with respect to R^2 when simulations are performed using real data genotypes than with fully simulated genotypes, but we observe that in the first setting the curve representing the interaction power of G-GEE is detached from the others, reflecting the superior performance of G-GEE over PLS and PCA. Note that the power of G-GEE to detect main effects is always less than that of PCA and PLS when $R^2 < 0.4$ [see Additional file 1 and 5]. In short, G-GEE performs better when detecting interactions than when detecting main effects.

Conclusions

In this paper we compared different approaches for modelling gene-gene epistasis in a penalized regression framework. Our primary concern was the detection of interaction effects, and for this purpose we defined a general model and tested different interaction terms. We focused our analysis at the gene scale and compared three ways to design the interaction term. Some methods were inspired by previous proposed approaches based on dimensional reduction methods including Principal Component Analysis (PCA) and Partial Least Square analysis (PLS). We additionally proposed a new interaction modeling approach that we called Gene-Gene Eigen-Epistasis (G-GEE), where one interaction variable is built for each couple. The interaction variable was defined based on a criterion that maximizes the covariance between the phenotype and the pairwise SNP product matrix of the two genes. The interaction components were then introduced in a Group Lasso penalized regression model that takes the gene structure into account and is capable of handling a large number of genes simultaneously.

A power study of the different methods based on two different simulation schemes (simplified and realistic) provided us with a rich body of information. Across various papers in the literature we find comparisons of similar methods that use different phenotype simulation settings. In the present work we compared two simulation models. Our first model was from a previous study [17] that simulated the interaction component of each couple in an SNP pairwise product fashion. Our second model defined the interaction component as a pairwise product of representative variables of each gene. Overall the G-GEE method performed well in detecting interactions in all the settings that were tested, although it was not always able to do so in the settings where main and interaction effects involved different genes. The power of the PCA method is highly dependent on the phenotype simulation model, because of the similarity between the second phenotype simulation model and the estimation model of the PCA method. The PLS method is characterized by a lack of power in detecting interactions. PLS performs well only when the related main effects are also present. When the simulated main and interaction effects do not concern the same genes, the detection capability of the PLS approach collapses dramatically.

For all methods we observed a confusion phenomenon when active genes are not simulated with both main and interaction effects. False detections of interactions concern genes that were simulated to have main effects, and false detections of main effects concern genes simulated to have interaction effects. This phenomenon reveals the difficulty that all methods encounter in clearly distinguishing the different types of effects. There are more false main detections when using methods such as PCA and PLS that are better at detecting main effects (except when the phenotype is simulated using the PCA model). As for interaction effects, the G-GEE methods make more false interaction detections than PCA and PLS.

When genotypes are fully simulated in the simplified simulation study, the G-GEE and PCA approaches performed better when the PCA phenotype simulation model was used, whereas the PLS method was not very sensitive to the choice of phenotype simulation model. Unlike PCA and PLS, G-GEE is better at detecting interaction effects than at detecting main effects when simulations use a real data set. Since the simulation study using realistic data is meant to mimic real genotype data structure, we conclude that in a real context G-GEE will be better at detecting interaction effects than main effects.

In comparison to SNP-SNP interaction approaches, the gene-scale dimension of our proposed method means that considerably fewer interaction variables need to be considered within a genetic region. This reduction in problem size allows larger problems to be handled. Moreover, a penalized regression method allows a true multivariate approach over a larger number of genes. It also extends other proposed gene-scale approaches, such as that presented by Wang et al. [8]. The ability to handle a relatively large number of genes simultaneously makes the detection of interactions between different genetic regions possible. This might be useful as an initial step, prior to using SNP-SNP interaction methods that may provide more accurate information.

As the G-GEE method is not able yet to consider all human genes at the same time, it is necessary to specify a list of genes to be explored for potential interactions. Given that its power to detect main effects is low, for the detection of main effects it will be safer to use previously acquired knowledge of the genetic effects, or to use a preprocessing method. Another limitation of the method is gene size. Computing the gene Eigen-Epistasis vector for two genes of size p_r and p_s requires an $n \times (p_r p_s)$ matrix to be computed.

Prospects are improving the G-GEE method's performance by optimizing the computational cost and exploring new interaction functions to be plugged into the G-GEE criterion.

Additional files

Additional file 1: Figure S1. Comparison of power to detect main effects in the simplified simulation study. Power under a simplified context. The figures show the power to detect main effects of the three methods depending on R^2. (PDF 8 kb)

Additional file 2: Figure S2. Simulation on 25 genes with fully simulated data and various simulated effects. Discoveries for the setting with numerous effects. Heatmap of the ratio of the number of times where each variable was significant to the total number of simulations for $R^2 = 0.7$ using the Wang Pathway model for the phenotype simulation with fully simulated data. We consider 25 genes with two main effects for genes 1 and 2, and four interaction effects between genes 3 and 4, genes 5 and 6, genes 7 and 8, and genes 9 and 10. (PDF 30 kb)

Additional file 3: Figure S3. Discoveries in the realistic simulation study with $R^2 = 0.4$. Discoveries under a realistic context. Heatmaps of the ratio of the number of times where each variable was significant to the total number of simulations for $R^2 = 0.4$. (PDF 9 kb)

Additional file 4: Figure S4. Comparison of execution time required to model interaction and to fit Group Lasso for the five first settings of the realistic simulation study. Execution time. Median of the execution time to model interaction and to fit Group Lasso for the five first settings of the realistic simulation study. (PDF 8 kb)

Additional file 5: Figure S5. Comparison of power to detect main effects in the realistic simulation study. Power under a realistic context. The figures show the power to detect main effects of the three methods depending on R^2. (PDF 7 kb)

Abbreviations

3G-SPA: model-based kernel machine method; AS: Ankylosing spondylitis; CCA: Canonical correlation analysis; CCU: Canonical correlation-based U-statistic model; GGEE: Gene-gene Eigen-Epistasis; GWAS: Genome-wide association studies; HLA: Human leukocyte antigen; LD: Linkage disequilibrium; MAF: Minor allele frequency; MHC: Major histocompatibility complex; PC: Principal components; PCA: Principal component analysis; PLS: Partial least squares; SNP: Single-nucleotide polymorphism

Acknowledgements

None

Funding

Université d'Evry Val d'Essonne, CNRS 8071, ENSIIE, INRA.

Availability of data and material

Our proposed G-GEE method has been implemented in an R package which is available on github: https://github.com/vstanislas/GGEE

Authors' contributions

The three authors developed the approach and the design. VS conducted the simulation study, applied the approach to the real data and drafted the manuscript. CD and CA supervised the implementation process and critically read and edit the manuscript. The work of VS fulfills part of the requirements of her PhD. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable

Ethics approval and consent to participate

The dataset regarding ankylosing spondylitis consists of the French subset of the large study of the International Genetics of Ankylosing Spondylitis (IGAS) study [22]. For this subset, unrelated cases were recruited through the Rheumatology clinic of Ambroise Paré Hospital (Boulogne-Billancourt, France) or through the national self-help patients' association: "Association Française des Spondylarthritiques". Population-matched unrelated controls were obtained from the "Centre d'Etude du Polymorphisme Humain", or were recruited as healthy spouses of cases. The protocol was reviewed and approved by the Ethics committee of the Ambroise Paré hospital. All participants gave their informed consent to the study.

The application on thyroid carcinomas was carried out on a public dataset that came from the study of Luzón-Toro et al. on identification of epistatic interactions in two different types of thyroid carcinomas [23]. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction. All subjects underwent peripheral blood extraction for genomic DNA isolation using MagNA Pure LC system (Roche, Indianapolis, IN) according to the manufacturer's instructions. A written informed consent was obtained from all the participants for clinical and molecular genetic studies. The study was approved by the Ethics Committee for clinical research in the University Hospital Virgen del Rocío (Seville, Spain) and complies with the tenets of the declaration of Helsinki.

Finally, the study on Inflammatory Bowel Diseases makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk.

Received: 23 June 2016 Accepted: 17 January 2017
Published online: 23 January 2017

References

1. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the Missing Heritability of Complex Diseases. Nature. 2009;461(7265):747–53.
2. Haig D. Does Heritability Hide in Epistasis between Linked SNPs?. Eur J Hum Genet. 2011;19(2):123.
3. Zuk O, Hechter E, Sunyaev SR, Lander ES. The Mystery of Missing Heritability: Genetic Interactions Create Phantom Heritability. PNAS. 2012;109(4):1193–8.
4. Niel C, Sinoquet C, Dina C, Rocheleau G. A survey about methods dedicated to epistasis detection. Front Genet. 2015;6:285.
5. Wei WH, Hemani G, Haley CS. Detecting Epistasis in Human Complex Traits. Nat Rev Genet. 2014;15(11):722–33.
6. Steen KV. Travelling the World of Gene-Gene Interactions. Brief Bioinformatics. 2012;13(1):1–19.
7. Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. Powerful Multilocus Tests of Genetic Association in the Presence of Gene-Gene and Gene-Environment Interactions. Am J Hum Genet. 2006;79(6):1002–16. Turkey's 1-df.
8. Wang T, Ho G, Ye K, Strickler H, Elston RC. A Partial Least-Square Approach for Modeling Gene-Gene and Gene-Environment Interactions When Multiple Markers Are Genotyped. Genet Epidemiol. 2009;33(1):PLS approach.
9. Li J, Tang R, Biernacka JM, de Andrade M. Identification of Gene-Gene Interaction Using Principal Components. BMC Proceedings. 2009;3(Suppl 7):S78. PC.
10. He J, Wang K, Edmondson AC, Rader DJ, Li C, Li M. Gene-Based Interaction Analysis by Incorporating External Linkage Disequilibrium Information. Eur J Hum Genet. 2011;19(2):164–72. PC Framework.
11. Rajapakse I, Perlman MD, Martin PJ, Hansen JA, Kooperberg C. Multivariate Detection of Gene-Gene Interactions. Genet Epidemiol. 2012;36(6):622–30. CLD.
12. Peng Q, Zhao J, Xue F. A Gene-Based Method for Detecting Genegene Co-Association in a Case–control Association Study. Eur J Hum Genet. 2010;18(5):582–7. CCU.
13. Larson NB, Jenkins GD, Larson MC, Vierkant RA, Sellers TA, Phelan CM, et al. Kernel Canonical Correlation Analysis for Assessing Gene-Gene Interactions and Application to Ovarian Cancer. Eur J Hum Genet. 2014;22(1):126–31. KCCA.
14. Yuan Z, Gao Q, He Y, Zhang X, Li F, Zhao J, et al. Detection for Gene-Gene Co-Association via Kernel Canonical Correlation Analysis. BMC Genet. 2012;13:83. KCCU.

15. Li S, Cui Y. Gene-Centric Gene–gene Interaction: A Model-Based Kernel Machine Method. Ann Appl Stat. 2012;6(3):1134–61. 3G-SPA.
16. D'Angelo GM, Rao D, Gu CC. Combining Least Absolute Shrinkage and Selection Operator (LASSO) and Principal-Components Analysis for Detection of Gene-Gene Interactions in Genome-Wide Association Studies. BMC Proc. 2009;3(Suppl 7):PCA-LASSO.
17. Wang X, Zhang D, Tzeng JY. Pathway-Guided Identification of Gene-Gene Interactions. Ann Hum Genet. 2014;78(6):Pathway guided.
18. Yuan M, Lin Y. Model Selection and Estimation in Regression with Grouped Variables. J R Stat Soc Series B. 2006;68:49–67.
19. Bécu JM, Grandvalet Y, Ambroise C, Dalmasso C. Beyond support in two-stage variable selection. Statistics and Computing. 2017;27:169-179.
20. Zhang F, Wagener D. An Approach to Incorporate Linkage Disequilibrium Structure into Genomic Association Analysis. J Genet Genomics. 2008;35(6):381–385. PC-LR.
21. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-Wide Association Analysis by Lasso Penalized Logistic Regression. Bioinformatics. 2009;25(6):714–21. Lasso penalized logistic regression.
22. Cortes A, Hadler J, Pointon JP, Robinson PC, Karaderi T, Leo P, et al. Identification of Multiple Risk Variants for Ankylosing Spondylitis through High-Density Genotyping of Immune-Related Loci. Nat Genet. 2013;45(7):730–8.
23. Luzón-Toro B, Bleda M, Navarro E, García-Alonso L, Ruiz-Ferrer M, Medina I, et al. Identification of Epistatic Interactions through Genome-Wide Association Studies in Sporadic Medullary and Juvenile Papillary Thyroid Carcinomas. BMC Med Genomics. 2015;8(1):83.
24. Sieper J, Braun J, Rudwaleit M, Boonen A, Zink A. Ankylosing Spondylitis: An Overview. Ann Rheum Dis. 2002;61(Suppl 3):. iii8.
25. Schlosstein L, Terasaki PI, Bluestone R, Pearson CM. High Association of an HL-A Antigen, W27, with Ankylosing Spondylitis. N Engl J Med. 1973;288(14):704–6.
26. Woodrow JC, Eastmond CJ. HLA B27 and the Genetics of Ankylosing Spondylitis. Ann Rheum Dis. 1978;37(6):504–9.
27. Thomas GP, Brown MA. Genetics and Genomics of Ankylosing Spondylitis. Immunol Rev. 2010;233(1):162–80.
28. Tsui FW, Tsui HW, Akram A, Haroon N, Inman RD. The genetic basis of ankylosing spondylitis: New insights into disease pathogenesis. Appl Clin Genet. 2014;7:105–15.
29. Reveille JD, Sims AM, Danoy P, Evans DM, Leo P, Pointon JJ, et al. Genome-Wide Association Study of Ankylosing Spondylitis Identifies Non-MHC Susceptibility Loci. Nat Genet. 2010;42(2):123–7.
30. Melville S, Melville MS. Package 'NCBI2R'. 2012. Available online at: https://cran.r-project.org/src/contrib/Archive/NCBI2R/. Accessed 16 Jan 2017.
31. Landa I, Boullosa C, Inglada-Pérez L, Sastre-Perona A, Pastor S, Velázquez A, et al. An Epistatic Interaction between the PAX8 and STK17B Genes in Papillary Thyroid Cancer Susceptibility. PLoS ONE. 2013;8(9):e74765.
32. Martinez-Chamorro A, Moreno A, Gómez-García M, Cabello MJ, Martin J, Lopez-Nevot MÁ. Epistatic Interaction between TLR4 and NOD2 in Patients with Crohn's Disease: Relation with Risk and Phenotype in a Spanish Cohort. Immunobiology. 2016;221(9):927–33.

Bibliography

Abad, C., González-Escribano, M. F., Diaz-Gallo, L. M., Lucena-Soto, J. M., Márquez, J. L., Leo, E., Crivell, C., Gómez-García, M., Martín, J., Núñez-Roldán, A., and García-Lozano, J. R. (2011). Association of Toll-like receptor 10 and susceptibility to Crohn's disease independent of NOD2. *Genes Immun.*, 12(8):635–642. 126

Achkar, J.-P. and Fiocchi, C. (2009). Editorial: Gene-Gene Interactions in Inflammatory Bowel Disease: Biological and Clinical Implications. *Am. J. Gastroenterol.*, 104(7):1734–1736. 14, 126

Ackermann, M. and Beyer, A. (2012). Systematic Detection of Epistatic Interactions Based on Allele Pair Frequencies. *PLoS Genet*, 8(2):e1002463. ImAP. 58

Alam, M. A., Komori, O., Calhoun, V., and Wang, Y.-P. (2016). Robust Kernel Canonical Correlation Analysis to Detect Gene-Gene Interaction for Imaging Genetics Data. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '16, pages 279–288, New York, NY, USA. ACM. robust KCCU. 67

Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic Mapping in Human Disease. *Science*, 322(5903):881–888. 15

An, P., Mukherjee, O., Chanda, P., Yao, L., Engelman, C. D., Huang, C.-H., Zheng, T., Kovac, I. P., Dubé, M.-P., Liang, X., Li, J., de Andrade, M., Culverhouse, R., Malzahn, D., Manning, A. K., Clarke, G. M., Jung, J., and Province, M. A. (2009). The Challenge of Detecting Epistasis (GxG Interactions): Genetic Analysis Workshop 16. *Genet Epidemiol*, 33(0 1):S58–S67. 56

Ananthakrishnan, A. N. (2015). Epidemiology and risk factors for IBD. *Nat Rev Gastroenterol Hepatol*, 12(4):205–217. 125

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nat. Protocols*, 5(9):1564–1573. 20, 21, 22

Austin, S. R., Dialsingh, I., and Altman, N. S. (2014). Multiple Hypothesis Testing: A Review. *Journal of the Indian Society of Agricultural Statistics*, 68(2):303–314. 40

Ayers, K. L. and Cordell, H. J. (2010). SNP Selection in Genome-Wide and Candidate Gene Studies via Penalized Logistic Regression. *Genet Epidemiol*, 34(8):879–891. 46

Aylor, D. L. and Zeng, Z.-B. (2008). From Classical Genetics to Quantitative Genetics to Systems Biology: Modeling Epistasis. *PLOS Genetics*, 4(3):e1000029. 14

Bakker, P. I., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nat Genet*, 37. 41

Baron, M. (2001). The search for complex disease genes: Fault by linkage or fault by association? *Molecular Psychiatry*, 6(2):143–149. 15, 16

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2):263–265. 42

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., Brant, S. R., Silverberg, M. S., Taylor, K. D., Barmada, M. M., Bitton, A., Dassopoulos, T., Datta, L. W., Green, T., Griffiths, A. M., Kistner, E. O., Murtha, M. T., Regueiro, M. D., Rotter, J. I., Schumm, L. P., Steinhart, A. H., Targan, S. R., Xavier, R. J., Libioulle, C., Sandor, C., Lathrop, M., Belaiche, J., Dewit, O., Gut, I., Heath, S., Laukens, D., Mni, M., Rutgeerts, P., Van Gossum, A., Zelenika, D., Franchimont, D., Hugot, J.-P., de Vos, M., Vermeire, S., Louis, E., Cardon, L. R., Anderson, C. A., Drummond, H., Nimmo, E., Ahmad, T., Prescott, N. J., Onnie, C. M., Fisher, S. A., Marchini, J., Ghori, J., Bumpstead, S., Gwilliam, R., Tremelling, M., Deloukas, P., Mansfield, J., Jewell, D., Satsangi, J., Mathew, C. G., Parkes, M., Georges, M., and Daly, M. J. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet*, 40(8):955–962. 126

Bateson, W. (1909). *Mendel's Principles of Heredity.* Cambridge University Press,, Cambridge. 11, 12

Bécu, J.-M., Grandvalet, Y., Ambroise, C., and Dalmasso, C. (2017). Beyond support in two-stage variable selection. *Statistics and Computing*, 27:169–79. 78, 79

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300. 41

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Ann. Statist.*, 41(3):1111–1141. hiernet. 58

Bland, J. M. and Altman, D. G. (1995). Multiple significance tests: The Bonferroni method. *BMJ*, 310(6973):170. 40

Botta, V., Louppe, G., Geurts, P., and Wehenkel, L. (2014). Exploiting SNP Correlations within Random Forest for Genome-Wide Association Studies. *PLoS One*, 9(4). T-Trees.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 24(2):123–140. 54

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32. 54

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA. 53

Brinza, D., Schultz, M., Tesler, G., and Bafna, V. (2010). RAPID detection of gene–gene interactions in genome-wide association studies. *Bioinformatics*, 26(22):2856–2862. RAPID.

Brown, M. A., Kenna, T., and Wordsworth, B. P. (2016). Genetics of ankylosing spondylitis—insights into pathogenesis. *Nat Rev Rheumatol*, 12(2):81–91. 122

Bush, W. S., Dudek, S. M., and Ritchie, M. D. (2006). Parallel Multifactor Dimensionality Reduction: A tool for the large scale analysis of gene-gene interactions. *Bioinformatics*, 22(17):2173–2174. Parallel MDR. 64

Bush, W. S., Dudek, S. M., and Ritchie, M. D. (2009). Biofilter: A knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac Symp Biocomput*, pages 368–379. 58, 64

Bush, W. S. and Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol*, 8(12):e1002822. 10, 16, 17, 19, 20, 23, 25

Cantor, R. M., Lange, K., and Sinsheimer, J. S. (2010). Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. *Am J Hum Genet*, 86(1):6–22. 29, 32

Carlborg, O. and Haley, C. S. (2004). Epistasis: Too often neglected in complex trait studies? *Nat Rev Genet*, 5(8):618–625. 31

Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., and Nickerson, D. A. (2004). Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium. *Am J Hum Genet*, 74(1):106–120. 43

Chanda, P., Sucheston, L., Zhang, A., Brazeau, D., Freudenheim, J. L., Ambrosone, C., and Ramanathan, M. (2008). AMBIENCE: A Novel Approach and Efficient Algorithm for Identifying Informative Genetic and Environmental Associations With Complex Phenotypes. *Genetics*, 180(2):1191–1210. AMBIENCE.

Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G., Hirschhorn, J. N., Abecasis, G., Altshuler, D., Bailey-Wilson, J. E., Brooks, L. D., Cardon, L. R., Daly, M., Donnelly, P., Fraumeni, J. F., Freimer, N. B., Gerhard, D. S., Gunter, C., Guttmacher, A. E., Guyer, M. S., Harris, E. L., Hoh, J., Hoover, R., Kong, C. A., Merikangas, K. R., Morton, C. C., Palmer, L. J., Phimister, E. G., Rice, J. P., Roberts, J., Rotimi, C., Tucker, M. A., Vogan, K. J., Wacholder, S., Wijsman, E. M., Winn, D. M., and Collins, F. S. (2007). Replicating genotype–phenotype associations. *Nature*, 447(7145):655–660. 25

Chatr-Aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., Reguly, T., Nixon, J., Ramage, L., Winter, A., Sellam, A., Chang, C., Hirschman, J., Theesfeld, C., Rust, J., Livstone, M. S., Dolinski, K., and Tyers, M. (2015). The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, 43(Database issue):D470–478. 58

Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U., and Wacholder, S. (2006). Powerful Multilocus Tests of Genetic Association in the Presence of Gene-Gene and Gene-Environment Interactions. *Am J Hum Genet*, 79(6):1002–1016. Turkey's 1DF. 67

Chen, L. S., Hutter, C. M., Potter, J. D., Liu, Y., Prentice, R. L., Peters, U., and Hsu, L. (2010). Insights into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data. *The American Journal of Human Genetics*, 86(6):860–871. 44

Chen, S.-H., Sun, J., Dimitrov, L., Turner, A. R., Adams, T. S., Meyers, D. A., Chang, B.-L., Zheng, S. L., Grönberg, H., Xu, J., and Hsu, F.-C. (2008). A support vector machine approach for detecting gene-gene interaction. *Genet. Epidemiol.*, 32(2):152–167. 64

Chen, X., Liu, C.-T., Zhang, M., and Zhang, H. (2007). A forest-based approach to identifying gene and gene–gene interactions. *Proc Natl Acad Sci U S A*, 104(49):19199–19203. HapForest. 69

Chipman, H. (1996). Bayesian variable selection with related predictors. *Can J Statistics*, 24(1):17–36. 58

Clancy, S. (2008). Genetic Recombination. *Nature Education*, 1(1):40. 7

Cook, N. R., Zee, R. Y. L., and Ridker, P. M. (2004). Tree and spline based association analysis of gene–gene interaction models for ischemic stroke. *Statist. Med.*, 23(9):1439–1453. MARS and CART.

Cordell, H. J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.*, 11(20):2463–2468. 11, 12, 14, 30

Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature reviews. Genetics*, 10(6):392–404. 56

Cortes, A. and Brown, M. A. (2011). Promise and pitfalls of the Immunochip. *Arthritis Res Ther*, 13(1):101. 123

Cortes, A., Hadler, J., Pointon, J. P., Robinson, P. C., Karaderi, T., Leo, P., Cremin, K., Pryce, K., Harris, J., lee, S., Joo, K. B., Shim, S.-C., Weisman, M., Ward, M., Zhou, X., Garchon, H.-J., Chiocchia, G., Nossent, J., Lie, B. A., Førre, Ø., Tuomilehto, J., Laiho, K., Jiang, L., Liu, Y., Wu, X., Bradbury, L. A., Elewaut, D., Burgos-Vargas, R., Stebbings, S., Appleton, L., Farrah, C., Lau, J., Kenna, T. J., Haroon, N., Ferreira, M. A., Yang, J., Mulero, J., Fernandez-Sueiro, J. L., Gonzalez-Gay, M. A., lopez-Larrea, C., Deloukas, P., Donnelly, P., Bowness, P., Gafney, K., Gaston, H., Gladman, D. D., Rahman, P., Maksymowych, W. P., Xu, H., Crusius, J. B. A., van der Horst-Bruinsma, I. E., Chou, C.-T., Valle-Oñate, R., Romero-Sánchez, C., Hansen, I. M., Pimentel-Santos, F. M., Inman, R. D., Videm, V., Martin, J., Breban, M., Reveille, J. D., Evans, D. M., Kim, T.-H., Wordsworth, B. P., and Brown, M. A. (2013). Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nat Genet*, 45(7):730–738. 83, 123

Cortes, A., Pulit, S. L., Leo, P. J., Pointon, J. J., Robinson, P. C., Weisman, M. H., Ward, M., Gensler, L. S., Zhou, X., Garchon, H.-J., Chiocchia, G., Nossent, J., Lie, B. A., Førre, Ø., Tuomilehto, J., Laiho, K., Bradbury, L. A., Elewaut, D., Burgos-Vargas, R., Stebbings, S., Appleton, L., Farrah, C., Lau, J., Haroon, N., Mulero, J., Blanco, F. J., Gonzalez-Gay, M. A., Lopez-Larrea, C., Bowness, P., Gaffney, K., Gaston, H., Gladman, D. D., Rahman, P., Maksymowych, W. P., Crusius, J. B. A., van der Horst-Bruinsma, I. E., Valle-Oñate, R., Romero-Sánchez, C., Hansen, I. M., Pimentel-Santos, F. M., Inman, R. D., Martin, J., Breban, M., Wordsworth, B. P., Reveille, J. D., Evans, D. M., de Bakker, P. I. W., and Brown, M. A. (2015). Major histocompatibility complex associations of ankylosing spondylitis are complex and involve further epistasis with *ERAP1*. *Nature Communications*, 6:ncomms8146. 122

Dandine-Roulland, C. and Perdry, H. (2015). The Use of the Linear Mixed Model in Human Genetics. *Hum. Hered.*, 80(4):196–206. 31

D'Angelo, G. M., Rao, D., and Gu, C. C. (2009). Combining least absolute shrinkage and selection operator (LASSO) and principal-components analysis for detection of gene-gene interactions in genome-wide association studies. *BMC Proc*, 3 Suppl 7:S62. PCA LASSO. 69

Dasgupta, A., Sun, Y. V., König, I. R., Bailey-Wilson, J. E., and Malley, J. D. (2011). Brief review of regression-based and machine learning methods in genetic epidemiology: The Genetic Analysis Workshop 17 experience. *Genet. Epidemiol.*, 35 Suppl 1:S5–11. 31

Dehman, A., Ambroise, C., and Neuvial, P. (2015). Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinformatics*, 16:148. 42, 73, 131

Deonier, R. C., Tavaré, S., and Waterman, M. (2005). *Computational Genome Analysis - An Introduction*. Springer-Verlag New York. 8, 9

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, 55(4):997–1004. 22

DiStefano, J. K. and Taverna, D. M. (2011). Technological Issues and Experimental Design of Gene Association Studies. In *Disease Gene Identification*, Methods in Molecular Biology, pages 3–16. Humana Press, Totowa, NJ. 19

Dong, C., Chu, X., Wang, Y., Wang, Y., Jin, L., Shi, T., Huang, W., and Li, Y. (2007). Exploration of gene–gene interaction effects using entropy-based methods. *Eur J Hum Genet*, 16(2):229–235. Entropy based method.

Du, Y., Xie, J., Chang, W., Han, Y., and Cao, G. (2012). Genome-wide association studies: Inherent limitations and future challenges. *Front. Med.*, 6(4):444–450. 28, 30

Duerr, R. H., Taylor, K. D., Brant, S. R., Rioux, J. D., Silverberg, M. S., Daly, M. J., Steinhart, A. H., Abraham, C., Regueiro, M., Griffiths, A., Dassopoulos, T., Bitton, A., Yang, H., Targan, S., Datta, L. W., Kistner, E. O., Schumm, L. P., Lee, A. T., Gregersen, P. K., Barmada, M. M., Rotter, J. I., Nicolae, D. L., and Cho, J. H. (2006). A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. *Science*, 314(5804):1461–1463. 27

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D. P., Thompson, D., Ballinger, D. G., Struewing, J. P., Morrison, J., Field, H., Luben, R., Wareham, N., Ahmed, S., Healey, C. S., Bowman, R., SEARCH collaborators, Meyer, K. B., Haiman, C. A., Kolonel, L. K., Henderson, B. E., Le Marchand, L., Brennan, P., Sangrajrang, S., Gaborieau, V., Odefrey, F., Shen, C.-Y., Wu, P.-E., Wang, H.-C., Eccles, D., Evans, D. G., Peto, J.,

Fletcher, O., Johnson, N., Seal, S., Stratton, M. R., Rahman, N., Chenevix-Trench, G., Bojesen, S. E., Nordestgaard, B. G., Axelsson, C. K., Garcia-Closas, M., Brinton, L., Chanock, S., Lissowska, J., Peplonska, B., Nevanlinna, H., Fagerholm, R., Eerola, H., Kang, D., Yoo, K.-Y., Noh, D.-Y., Ahn, S.-H., Hunter, D. J., Hankinson, S. E., Cox, D. G., Hall, P., Wedren, S., Liu, J., Low, Y.-L., Bogdanova, N., Schürmann, P., Dörk, T., Tollenaar, R. A. E. M., Jacobi, C. E., Devilee, P., Klijn, J. G. M., Sigurdson, A. J., Doody, M. M., Alexander, B. H., Zhang, J., Cox, A., Brock, I. W., MacPherson, G., Reed, M. W. R., Couch, F. J., Goode, E. L., Olson, J. E., Meijers-Heijboer, H., van den Ouweland, A., Uitterlinden, A., Rivadeneira, F., Milne, R. L., Ribas, G., Gonzalez-Neira, A., Benitez, J., Hopper, J. L., McCredie, M., Southey, M., Giles, G. G., Schroen, C., Justenhoven, C., Brauch, H., Hamann, U., Ko, Y.-D., Spurdle, A. B., Beesley, J., Chen, X., kConFab, AOCS Management Group, Mannermaa, A., Kosma, V.-M., Kataja, V., Hartikainen, J., Day, N. E., Cox, D. R., and Ponder, B. A. J. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447(7148):1087–1093. 27

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.*, 7(1):1–26. 54

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, 32(2):407–499. 47

Elith, J., Leathwick, J. R., and Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813. 55

Emily, M. (2012). IndOR: A new statistical procedure to test for SNP–SNP epistasis in genome-wide association studies. *Statist. Med.*, 31(21):2359–2373. IndOR. 63

Emily, M. (2016). AGGrEGATOr: A Gene-based GEne-Gene interActTiOn test for case-control association studies. *Stat Appl Genet Mol Biol*, 15(2):151–171. AGGrEGATOr. 66, 110, 114, 115

Emily, M., Houée-Bigot, M., Kroell, F., and Sounac, N. (2017). GeneGeneInteR: Tools for Testing Gene-Gene Interaction at the Gene Level. R package version 0.99.0. 110

Emily, M., Mailund, T., Hein, J., Schauser, L., and Schierup, M. H. (2009). Using biological networks to search for interacting loci in genome-wide association studies. *Eur J Hum Genet*, 17(10):1231–1240. 59

Epstein, M. P. and Satten, G. A. (2003). Inference on Haplotype Effects in Case-Control Studies Using Unphased Genotype Data. *Am J Hum Genet*, 73(6):1316–1329. 42

Evans, D. M., Marchini, J., Morris, A. P., and Cardon, L. R. (2006). Two-Stage Two-Locus Models in Genome-Wide Association. *PLOS Genetics*, 2(9):e157. 58

Evans, D. M., Spencer, C. C. A., Pointon, J. J., Su, Z., Harvey, D., Kochan, G., Oppermann, U., Opperman, U., Dilthey, A., Pirinen, M., Stone, M. A., Appleton, L., Moutsianas, L., Moutsianis, L., Leslie, S., Wordsworth, T., Kenna, T. J., Karaderi, T., Thomas, G. P., Ward, M. M., Weisman, M. H., Farrar, C., Bradbury, L. A., Danoy, P., Inman, R. D., Maksymowych, W., Gladman, D., Rahman, P., Spondyloarthritis Research Consortium of Canada (SPARCC), Morgan, A., Marzo-Ortega, H., Bowness, P., Gaffney, K., Gaston, J. S. H., Smith, M., Bruges-Armas, J., Couto, A.-R., Sorrentino, R., Paladini, F., Ferreira, M. A., Xu, H., Liu, Y., Jiang, L., Lopez-Larrea, C., Díaz-Peña, R., López-Vázquez, A., Zayats, T., Band, G., Bellenguez, C., Blackburn, H., Blackwell, J. M., Bramon, E., Bumpstead, S. J., Casas, J. P., Corvin, A., Craddock, N., Deloukas, P., Dronov, S., Duncanson, A., Edkins, S., Freeman, C., Gillman, M., Gray, E., Gwilliam, R., Hammond, N., Hunt, S. E., Jankowski, J., Jayakumar, A., Langford, C., Liddle, J., Markus, H. S., Mathew, C. G., McCann, O. T., McCarthy, M. I., Palmer, C. N. A., Peltonen, L., Plomin, R., Potter, S. C., Rautanen, A., Ravindrarajah, R., Ricketts, M., Samani, N., Sawcer, S. J., Strange, A., Trembath, R. C., Viswanathan, A. C., Waller, M., Weston, P., Whittaker, P., Widaa, S., Wood, N. W., McVean, G., Reveille, J. D., Wordsworth, B. P., Brown, M. A., Donnelly, P., Australo-Anglo-American Spondyloarthritis Consortium (TASC), and Wellcome Trust Case Control Consortium 2 (WTCCC2) (2011). Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat. Genet.*, 43(8):761–767. 122

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., Matthews, L., May, B., Milacic, M., Rothfels, K., Shamovsky, V., Webber, M., Weiser, J., Williams, M., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P. (2016). The Reactome pathway Knowledgebase. *Nucleic Acids Res.*, 44(D1):D481–487. 58

Fan, R., Zhong, M., Wang, S., Zhang, Y., Andrew, A., Karagas, M., Chen, H., Amos, C. I., Xiong, M., and Moore, J. H. (2011). Entropy-based information gain approaches to detect and to characterize gene-gene and gene-environment interactions/correlations of complex diseases. *Genet. Epidemiol.*, 35(7):706–721.

Fang, Y.-H. and Chiu, Y.-F. (2012). SVM-based generalized multifactor dimensionality reduction approaches for detecting gene-gene interactions in family studies. *Genet. Epidemiol.*, 36(2):88–98. PGMDR.

Ferrario, P. G. and König, I. R. (2016). Transferring entropy to the realm of GxG interactions. *Brief Bioinform.* 65

Ferreira, T. and Marchini, J. (2011). Modeling interactions with known risk loci—a Bayesian model averaging approach. *Annals of Human Genetics*, 75(1):1–9. SNPTEST.

Fiocchi, C. (2009). Susceptibility Genes and Overall Pathogenesis of Inflammatory Bowel Disease: Where Do We Stand? *Dig Dis*, 27(3):226–235. 1

Fisher, R. A. (1918). The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society Edinburgh*, 52(2):399–433. 12

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L. J. (2013). STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, 41(Database issue):D808–815. 58

Franke, A., McGovern, D. P. B., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., Anderson, C. A., Bis, J. C., Bumpstead, S., Ellinghaus, D., Festen, E. M., Georges, M., Green, T., Haritunians, T., Jostins, L., Latiano, A., Mathew, C. G., Montgomery, G. W., Prescott, N. J., Raychaudhuri, S., Rotter, J. I., Schumm, P., Sharma, Y., Simms, L. A., Taylor, K. D., Whiteman, D., Wijmenga, C., Baldassano, R. N., Barclay, M., Bayless, T. M., Brand, S., Büning, C., Cohen, A., Colombel, J.-F., Cottone, M., Stronati, L., Denson, T., De Vos, M., D'Inca, R., Dubinsky, M., Edwards, C., Florin, T., Franchimont, D., Gearry, R., Glas, J., Van Gossum, A., Guthery, S. L., Halfvarson, J., Verspaget, H. W., Hugot, J.-P., Karban, A., Laukens, D., Lawrance, I., Lemann, M., Levine, A., Libioulle, C., Louis, E., Mowat, C., Newman, W., Panés, J., Phillips, A., Proctor, D. D., Regueiro, M., Russell, R., Rutgeerts, P., Sanderson, J., Sans, M., Seibold, F., Steinhart, A. H., Stokkers, P. C. F., Torkvist, L., Kullak-Ublick, G., Wilson, D., Walters, T., Targan, S. R., Brant, S. R., Rioux, J. D., D'Amato, M., Weersma, R. K., Kugathasan, S., Griffiths, A. M., Mansfield, J. C., Vermeire, S., Duerr, R. H., Silverberg, M. S., Satsangi, J., Schreiber, S., Cho, J. H., Annese, V., Hakonarson, H., Daly, M. J., and Parkes, M. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.*, 42(12):1118–1125. 126

Freund, Y. (1995). Boosting a Weak Learning Algorithm by Majority. *Information and Computation*, 121(2):256–285. 54

Fridley, B. L. and Biernacka, J. M. (2011). Gene set analysis of SNP data: Benefits, challenges, and future directions. *Eur. J. Hum. Genet.*, 19(8):837–843. 42, 43

Gao, H., Wu, Y., Li, J., Li, H., Li, J., and Yang, R. (2013). Forward LASSO analysis for high-order interactions in genome-wide association study. *Brief Bioinform*, page bbt037. Forward LASSO. 58, 62

Gauderman, W. J., Murcray, C., Gilliland, F., and Conti, D. V. (2007). Testing association between disease and multiple SNPs in a candidate gene. *Genet. Epidemiol.*, 31(5):383–395. 44

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., Tam, P. K.-H., Tsui, L.-C., Waye, M. M. Y., Wong, J. T.-F., Zeng, C., Zhang, Q., Chee, M. S., Galver, L. M., Kruglyak, S., Murray, S. S., Oliphant, A. R., Montpetit, A., Hudson, T. J., Chagnon, F., Ferretti, V., Leboeuf, M., Phillips, M. S., Verner, A., Kwok, P.-Y., Duan, S., Lind, D. L., Miller, R. D., Rice, J. P., Saccone, N. L., Taillon-Miller, P., Xiao, M., Nakamura, Y., Sekine, A., Sorimachi, K., Tanaka, T., Tanaka, Y., Tsunoda, T., Yoshino, E., Bentley, D. R., Deloukas, P., Hunt, S., Powell, D., Altshuler, D., Gabriel, S. B., Zhang, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Aniagwu, T., Marshall, P. A., Matthew, O., Nkwodimmah, C., Royal, C. D. M., Leppert, M. F., Dixon, M., Stein, L. D., Cunningham, F., Kanani, A., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Donnelly, P., Marchini, J., McVean, G. A. T., Myers, S. R., Cardon, L. R., Abecasis, G. R., Morris, A., Weir, B. S., Mullikin, J. C., Sherry, S. T., Feolo, M., Daly, M. J., Schaffner, S. F., Qiu, R., Kent, A., Dunston, G. M., Kato, K., Niikawa, N., Knoppers, B. M., Foster, M. W., Clayton, E. W., Wang, V. O., Watkin, J., Sodergren, E., Weinstock, G. M., Wilson, R. K., Fulton, L. L., Rogers, J., Birren, B. W., Han, H., Wang, H., Godbout, M., Wallenburg, J. C., L'Archevêque, P., Bellemare, G., Todani, K., Fujita, T., Tanaka, S., Holden, A. L., Lai, E. H., Collins, F. S., Brooks, L. D., McEwen, J. E., Guyer, M. S., Jordan, E., Peterson, J. L., Spiegel, J., Sung, L. M., Zacharia, L. F., Kennedy, K., Dunn, M. G., Seabrook, R., Shillito, M., Skene, B., Stewart, J. G., (chair), D. L. V., (co Chair), E. W. C., (co Chair), L. B. J., Cho, M. K., Duster, T., Jasperse, M., Licinio, J., Long, J. C., Ossorio, P. N., Spallone, P., Terry, S. F., (chair), E. S. L., (co Chair), E. H. L., (co Chair), D. A. N., Boehnke, M., Douglas, J. A., Hudson, R. R., Kruglyak, L., and Nussbaum, R. L. (2003). The International HapMap Project. *Nature*, 426(6968):789–796. 17

Gibson, G. (2012). Rare and Common Variants: Twenty arguments. *Nat Rev Genet*, 13(2):135–145. 29

Gola, D., Mahachie John, J. M., van Steen, K., and König, I. R. (2016). A roadmap to multifactor dimensionality reduction methods. *Brief Bioinform*, 17(2):293–308. 64

Greene, C. S., Penrod, N. M., Kiralis, J., and Moore, J. H. (2009). Spatially

Uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions. *BioData Mining*, 2:5. 59, 131

Greene, C. S., Sinnott-Armstrong, N. A., Himmelstein, D. S., Park, P. J., Moore, J. H., and Harris, B. T. (2010). Multifactor dimensionality reduction for graphics processing units enables genome-wide testing of epistasis in sporadic ALS. *Bioinformatics*, 26(5):694–695. MDRGPU. 64

Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V., Zusmanovich, P., Sulem, P., Thorlacius, S., Gylfason, A., Steinberg, S., Helgadottir, A., Ingason, A., Steinthorsdottir, V., Olafsdottir, E. J., Olafsdottir, G. H., Jonsson, T., Borch-Johnsen, K., Hansen, T., Andersen, G., Jorgensen, T., Pedersen, O., Aben, K. K., Witjes, J. A., Swinkels, D. W., Heijer, M. D., Franke, B., Verbeek, A. L. M., Becker, D. M., Yanek, L. R., Becker, L. C., Tryggvadottir, L., Rafnar, T., Gulcher, J., Kiemeney, L. A., Kong, A., Thorsteinsdottir, U., and Stefansson, K. (2008). Many sequence variants affecting diversity of adult human height. *Nature genetics, Nature genetics.*, 40(5):609–615. 28

Gui, J., Andrew, A. S., Andrews, P., Nelson, H. M., Kelsey, K. T., Karagas, M. R., and Moore, J. H. (2011). A robust multifactor dimensionality reduction method for detecting gene-gene interactions with application to the genetic analysis of bladder cancer susceptibility. *Ann. Hum. Genet.*, 75(1):20–28. RMDR.

Gui, J., Moore, J. H., Williams, S. M., Andrews, P., Hillege, H. L., van der Harst, P., Navis, G., Gilst, W. H. V., Asselbergs, F. W., and Gilbert-Diamond, D. (2013). A Simple and Computationally Efficient Approach to Multifactor Dimensionality Reduction Analysis of Gene-Gene Interactions for Quantitative Traits. *PLOS ONE*, 8(6):e66545. QMDR. 64

Günther, F., Wawro, N., and Bammann, K. (2009). Neural networks for modeling gene-gene interactions in association studies. *BMC Genetics*, 10:87.

Gyenesei, A., Moody, J., Laiho, A., Semple, C. A., Haley, C. S., and Wei, W.-H. (2012). BiForce Toolbox: Powerful high-throughput computational analysis of gene–gene interactions in genome-wide association studies. *Nucleic Acids Res*, 40(Web Server issue):W628–W632. BiForce. 62

Ha, N.-T., Freytag, S., and Bickeboeller, H. (2014). Coverage and efficiency in current SNP chips. *Eur J Hum Genet*, 22(9):1124–1130. 19

Haig, D. (2011). Does heritability hide in epistasis between linked SNPs? *Eur J Hum Genet*, 19(2):123. 30

Han, B., Chen, X.-w., Talebizadeh, Z., and Xu, H. (2012). Genetic studies of complex human diseases: Characterizing SNP-disease associations using Bayesian networks. *BMC Systems Biology*, 6(3):S14. EpiBN.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC. 46

He, J., Wang, K., Edmondson, A. C., Rader, D. J., Li, C., and Li, M. (2011). Gene-based interaction analysis by incorporating external linkage disequilibrium information. *Eur. J. Hum. Genet.*, 19(2):164–172. PC Framework. 68

Hemani, G., Theocharidis, A., Wei, W., and Haley, C. (2011). EpiGPU: Exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics*, 27(11):1462–1465. EpiGPU.

Herold, C., Steffens, M., Brockschmidt, F. F., Baur, M. P., and Becker, T. (2009). INTERSNP: Genome-wide interaction analysis guided by a priori information. *Bioinformatics*, 25(24):3275–3281. INTERSNP.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., and Collins, F. S. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, 106. 28

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*, 6(2):95–108. 16

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67. 46

Hradsky, O., Dusatkova, P., Lenicek, M., Bronsky, J., Nevoral, J., Vitek, L., Lukas, M., Zeniskova, I., and Cinek, O. (2010). The CTLA4 variants may interact with the IL23R- and NOD2-conferred risk in development of Crohn's disease. *BMC Med Genet*, 11:91. 126

Hu, J. K., Wang, X., and Wang, P. (2014). Testing gene-gene interactions in genome wide association studies. *Genet. Epidemiol.*, 38(2):123–134. 56

Hu, X., Liu, Q., Zhang, Z., Li, Z., Wang, S., He, L., and Shi, Y. (2010). SHEsisEpi, a GPU-enhanced genome-wide SNP-SNP interaction scanning algorithm, efficiently reveals the risk genetic epistasis in bipolar disorder. *Cell Res*, 20(7):854–857. SHEsisEpi.

Huang, B. E. and Lin, D. Y. (2007). Efficient Association Mapping of Quantitative Trait Loci with Selective Genotyping. *Am J Hum Genet*, 80(3):567–576. 17

Huang, J., Breheny, P., and Ma, S. (2012). A Selective Review of Group Selection in High-Dimensional Models. *Statist. Sci.*, 27(4):481–499. 49

Hugot, J. P., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J. P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C. A., Gassull, M., Binder, V., Finkel, Y., Cortot, A., Modigliani, R., Laurent-Puig, P., Gower-Rousseau, C., Macry, J., Colombel, J. F., Sahbatou, M., and Thomas, G. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, 411(6837):599–603. 126

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., Wang, J., Yu, K., Chatterjee, N., Orr, N., Willett, W. C., Colditz, G. A., Ziegler, R. G., Berg, C. D., Buys, S. S., McCarty, C. A., Feigelson, H. S., Calle, E. E., Thun, M. J., Hayes, R. B., Tucker, M., Gerhard, D. S., Fraumeni, J. F., Hoover, R. N., Thomas, G., and Chanock, S. J. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.*, 39(7):870–874. 27

Jiang, B., Zhang, X., Zuo, Y., and Kang, G. (2011). A powerful truncated tail strength method for testing multiple null hypotheses in one dataset. *Journal of Theoretical Biology*, 277(1):67–73. 66, 110

Jing, P.-J. and Shen, H.-B. (2015). MACOED: A multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies. *Bioinformatics*, 31(5):634–641. MACOED.

Jorgenson, E. and Witte, J. S. (2006). A gene-centric approach to genome-wide association studies. *Nat Rev Genet*, 7(11):885–891. 42, 43

Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., Lee, J. C., Philip Schumm, L., Sharma, Y., Anderson, C. A., Essers, J., Mitrovic, M., Ning, K., Cleynen, I., Theatre, E., Spain, S. L., Raychaudhuri, S., Goyette, P., Wei, Z., Abraham, C., Achkar, J.-P., Ahmad, T., Amininejad, L., Ananthakrishnan, A. N., Andersen, V., Andrews, J. M., Baidoo, L., Balschun, T., Bampton, P. A., Bitton, A., Boucher, G., Brand, S., Büning, C., Cohain, A., Cichon, S., D'Amato, M., De Jong, D., Devaney, K. L., Dubinsky, M., Edwards, C., Ellinghaus, D., Ferguson, L. R., Franchimont, D., Fransen, K., Gearry, R., Georges, M., Gieger, C., Glas, J., Haritunians, T., Hart, A., Hawkey, C., Hedl, M., Hu, X., Karlsen, T. H., Kupcinskas, L., Kugathasan, S., Latiano, A., Laukens, D., Lawrance, I. C., Lees, C. W., Louis, E., Mahy, G., Mansfield, J., Morgan, A. R., Mowat, C., Newman, W., Palmieri, O., Ponsioen, C. Y., Potocnik, U., Prescott, N. J., Regueiro, M., Rotter, J. I., Russell, R. K., Sanderson, J. D., Sans, M., Satsangi, J., Schreiber, S., Simms, L. A., Sventoraityte, J., Targan, S. R., Taylor, K. D., Tremelling, M., Verspaget, H. W., De Vos, M., Wijmenga, C., Wilson, D. C., Winkelmann, J., Xavier, R. J., Zeissig, S., Zhang, B., Zhang, C. K., Zhao, H., The International IBD Genetics Consortium (iibdgc), Silverberg, M. S., Annese, V., Hakonarson, H., Brant, S. R., Radford-Smith, G.,

Mathew, C. G., Rioux, J. D., Schadt, E. E., Daly, M. J., Franke, A., Parkes, M., Vermeire, S., Barrett, J. C., and Cho, J. H. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124. 126

Kam-Thong, T., Czamara, D., Tsuda, K., Borgwardt, K., Lewis, C. M., Erhardt-Lehmann, A., Hemmer, B., Rieckmann, P., Daake, M., Weber, F., Wolf, C., Ziegler, A., Pütz, B., Holsboer, F., Schölkopf, B., and Müller-Myhsok, B. (2011). EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. *Eur. J. Hum. Genet.*, 19(4):465–471. EPIBLASTER.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, 40(Database issue):D109–114. 58

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*, 42(4):348–354. 31

Karki, R., Pandya, D., Elston, R. C., and Ferlini, C. (2015). Defining "mutation" and "polymorphism" in the era of personal genomics. *BMC Medical Genomics*, 8:37. 8

Keedwell, E. and Narayanan, A. (2005). Discovering gene networks with a neural-genetic hybrid. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(3):231–242.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., and Hermjakob, H. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, 40(Database issue):D841–846. 58

Kevans, D., Silverberg, M. S., Borowski, K., Griffiths, A., Xu, W., Onay, V., Paterson, A. D., Knight, J., and Croitoru, K. (2016). IBD Genetic Risk Profile in Healthy First-Degree Relatives of Crohn's Disease Patients. *J Crohns Colitis*, 10(2):209–215. 22

Kira, K. and Rendell, L. A. (1992). The Feature Selection Problem: Traditional Methods and a New Algorithm. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, AAAI'92, pages 129–134, San Jose, California. AAAI Press. 59

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh., J. (2005). Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, 308(5720):385–389. 27

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. In *Machine Learning: ECML-94*, Lecture Notes in Computer Science, pages 171–182. Springer, Berlin, Heidelberg. 59

Koo, C. L., Liew, M. J., Mohamad, M. S., and Mohamed Salleh, A. H. (2013). A Review for Detecting Gene-Gene Interactions Using Machine Learning Methods in Genetic Epidemiology. *BioMed Research International*, 2013:e432375. 63

Kooperberg, C. and LeBlanc, M. (2008). Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genet. Epidemiol.*, 32(3):255–263. 58, 59

Kooperberg, C. and Ruczinski, I. (2005). Identifying interacting SNPs using Monte Carlo logic regression. *Genet. Epidemiol.*, 28(2):157–170. MCMC logic regression. 62

Kwee, L. C., Liu, D., Lin, X., Ghosh, D., and Epstein, M. P. (2008). A Powerful and Flexible Multilocus Association Test for Quantitative Traits. *Am J Hum Genet*, 82(2):386–397. 44

Kwon, M.-S., Park, M., and Park, T. (2014). IGENT: Efficient entropy based algorithm for genome-wide gene-gene interaction analysis. *BMC Medical Genomics*, 7(1):S6. IGENT.

Laird, N. M. and Lange, C. (2011). *The Fundamentals of Modern Statistical Genetics*. Springer-Verlag New York. 4, 6, 16, 17, 19, 20, 22, 29, 31

Lamy, P., Grove, J., and Wiuf, C. (2011). A review of software for microarray genotyping. *Hum Genomics*, 5(4):304–309. 19

Larson, N. B., Jenkins, G. D., Larson, M. C., Vierkant, R. A., Sellers, T. A., Phelan, C. M., Schildkraut, J. M., Sutphen, R., Pharoah, P. P. D., Gayther, S. A., Wentzensen, N., Ovarian Cancer Association Consortium, Goode, E. L., and Fridley, B. L. (2014). Kernel canonical correlation analysis for assessing gene-gene interactions and application to ovarian cancer. *Eur. J. Hum. Genet.*, 22(1):126–131. KCCA algorithm. 67, 75

Larson, N. B. and Schaid, D. J. (2013). A Kernel Regression Approach to Gene-Gene Interaction Detection for Case-Control Studies. *Genet Epidemiol*, 37(7):695–703. 69

Lee, S. Y., Chung, Y., Elston, R. C., Kim, Y., and Park, T. (2007). Log-linear model-based multifactor dimensionality reduction method to detect gene gene interactions. *Bioinformatics*, 23(19):2589–2595. LM MDR.

Lehne, B., Lewis, C. M., and Schlitt, T. (2011). From SNPs to Genes: Disease Association at the Gene Level. *PLOS ONE*, 6(6):e20133. 43

Lettre, G., Jackson, A. U., Gieger, C., Schumacher, F. R., Berndt, S. I., Sanna, S., Eyheramendy, S., Voight, B. F., Butler, J. L., Guiducci, C., Illig, T., Hackett, R., Heid, I. M., Jacobs, K. B., Lyssenko, V., Uda, M., Boehnke, M., Chanock, S. J., Groop, L. C., Hu, F. B., Isomaa, B., Kraft, P., Peltonen, L., Salomaa, V., Schlessinger, D., Hunter, D. J., Hayes, R. B., Abecasis, G. R., Wichmann, H.-E., Mohlke, K. L., and Hirschhorn, J. N. (2008). Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet*, 40(5):584–591. 28

Lettre, G., Lange, C., and Hirschhorn, J. N. (2007). Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet. Epidemiol.*, 31(4):358–362. 24

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., MacDonald, J. R., Pang, A. W. C., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., Busam, D. A., Beeson, K. Y., McIntosh, T. C., Remington, K. A., Abril, J. F., Gill, J., Borman, J., Rogers, Y.-H., Frazier, M. E., Scherer, S. W., Strausberg, R. L., and Venter, J. C. (2007). The Diploid Genome Sequence of an Individual Human. *PLOS Biology*, 5(10):e254. 6

Lewis, C. M. (2002). Genetic association studies: Design, analysis and interpretation. *Brief. Bioinformatics*, 3(2):146–153. 18

Li, J., Huang, D., Guo, M., Liu, X., Wang, C., Teng, Z., Zhang, R., Jiang, Y., Lv, H., and Wang, L. (2015). A gene-based information gain method for detecting gene–gene interactions in case–control studies. *Eur J Hum Genet*, 23(11):1566–1572. GBIGM. 66, 110

Li, J., Malley, J. D., Andrew, A. S., Karagas, M. R., and Moore, J. H. (2016). Detecting gene-gene interactions using a permutation-based random forest method. *BioData Min*, 9:14. pRF. 65

Li, J., Tang, R., Biernacka, J. M., and de Andrade, M. (2009). Identification of gene-gene interaction using principal components. *BMC Proceedings*, 3(Suppl 7):S78. PC. 68, 110

Li, J., Zhang, K., and Yi, N. (2011a). A Bayesian Hierarchical Model for Detecting Haplotype-Haplotype and Haplotype-Environment Interactions in Genetic Association Studies. *Hum Hered*, 71(3):148–160. BayesGLM. 69

Li, J., Zhong, W., Li, R., and Wu, R. (2014). A fast algorithm for detecting gene–gene interactions in genome-wide association studies. *Ann. Appl. Stat.*, 8(4):2292–2318. TS-SIS. 58, 62

Li, M., Romero, R., Fu, W. J., and Cui, Y. (2010). Mapping Haplotype-haplotype Interactions with Adaptive LASSO. *BMC Genetics*, 11(1):79. HAPAL. 69

Li, M. J., Wang, P., Liu, X., Lim, E. L., Wang, Z., Yeager, M., Wong, M. P., Sham, P. C., Chanock, S. J., and Wang, J. (2012). GWASdb: A database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, 40(Database issue):D1047–1054. 17

Li, M.-X., Gui, H.-S., Kwan, J. S., and Sham, P. C. (2011b). GATES: A Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure. *Am J Hum Genet*, 88(3):283–293. 66, 110

Li, S. and Cui, Y. (2012). Gene-centric gene–gene interaction: A model-based kernel machine method. *Ann. Appl. Stat.*, 6(3):1134–1161. 3G-SPA. 68, 85, 95

Lily Wang, P. J. (2011). Gene set analysis of genome-wide association studies: Methodological issues and perspectives. *Genomics*, 98(1):1–8. 43

Lin, H.-Y., Chen, Y. A., Tsai, Y.-Y., Qu, X., Tseng, T.-S., and Park, J. Y. (2012). TRM: A Powerful Two-stage Machine Learning Approach for Identifying SNP-SNP Interactions. *Ann Hum Genet*, 76(1):53–62. TRM.

Liu, D., Ghosh, D., and Lin, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, 9:292. 44

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. *Biometrics*, 63(4):1079–1088. 44

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., Hayward, N. K., Montgomery, G. W., Visscher, P. M., Martin, N. G., and Macgregor, S. (2010). A Versatile Gene-Based Test for Genome-wide Association Studies. *Am J Hum Genet*, 87(1):139–145. 66

Liu, Y., Xu, H., Chen, S., Chen, X., Zhang, Z., Zhu, Z., Qin, X., Hu, L., Zhu, J., Zhao, G.-P., and Kong, X. (2011). Genome-Wide Interaction-Based Association Analysis Identified Multiple New Susceptibility Loci for Common Diseases. *PLoS Genet*, 7(3):e1001338. PIAM.

Lou, X.-Y., Chen, G.-B., Yan, L., Ma, J. Z., Zhu, J., Elston, R. C., and Li, M. D. (2007). A Generalized Combinatorial Approach for Detecting Gene-by-Gene and Gene-by-Environment Interactions with Application to Nicotine Dependence. *Am J Hum Genet*, 80(6):1125–1137. GMDR.

Luzón-Toro, B., Bleda, M., Navarro, E., García-Alonso, L., Ruiz-Ferrer, M., Medina, I., Martín-Sánchez, M., Gonzalez, C. Y., Fernández, R. M., Torroglosa, A., Antiñolo, G., Dopazo, J., and Borrego, S. (2015). Identification of epistatic interactions through genome-wide association studies in sporadic medullary and juvenile papillary thyroid carcinomas. *BMC Medical Genomics*, 8(1):83. 21

Ma, L., Brautbar, A., Boerwinkle, E., Sing, C. F., Clark, A. G., and Keinan, A. (2012). Knowledge-Driven Analysis Identifies a Gene–Gene Interaction Affecting High-Density Lipoprotein Cholesterol Levels in Multi-Ethnic Populations. *PLOS Genetics*, 8(5):e1002714. 58

Ma, L., Clark, A. G., and Keinan, A. (2013). Gene-Based Testing of Interactions in Association Studies of Quantitative Traits. *PLoS Genet*, 9(2):e1003321. GGG tests. 66

Ma, S., Song, X., and Huang, J. (2007). Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics*, 8:60. SGLasso. 49

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature News*, 456(7218):18–21. 29

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753. 28, 29

Marchini, J., Donnelly, P., and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*, 37(4):413–417. Log lin reg. 58

Martinez-Chamorro, A., Moreno, A., Gómez-García, M., Cabello, M. J., Martin, J., and Lopez-Nevot, M. Á. (2016). Epistatic interaction between TLR4 and NOD2 in patients with Crohn's Disease: Relation with risk and phenotype in a Spanish cohort. *Immunobiology*, 221(9):927–933. 126, 127

Matchenko-Shimko, N. and Dube, M. P. (2007). Gene-Gene Interaction Tests Using SVM and Neural Network Modeling. In *2007 IEEE Symposium on*

Computational Intelligence and Bioinformatics and Computational Biology, pages 90–97.

Mathew, C. G. and Lewis, C. M. (2004). Genetics of inflammatory bowel disease: Progress and prospects. *Hum Mol Genet*, 13(suppl_1):R161–R168. 125, 126

McKinney, B. A., Reif, D. M., Ritchie, M. D., and Moore, J. H. (2006). Machine Learning for Detecting Gene-Gene Interactions. *Appl Bioinformatics*, 5(2):77–88. 63

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71. 48

Melville, S. (2012). *Package 'NCBI2R'*. Available online at: https://cran.r-project.org/src/contrib/Archive/NCBI2R/. Accessed 16 Jan 2017. 84, 122

Millstein, J., Conti, D. V., Gilliland, F. D., and Gauderman, W. J. (2006). A Testing Framework for Identifying Susceptibility Genes in the Presence of Epistasis. *Am J Hum Genet*, 78(1):15–27. FITF.

Molodecky, N. A., Soon, I. S., Rabi, D. M., Ghali, W. A., Ferris, M., Chernoff, G., Benchimol, E. I., Panaccione, R., Ghosh, S., Barkema, H. W., and Kaplan, G. G. (2012). Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology*, 142(1):46–54.e42; quiz e30. 125

Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered.*, 56(1-3):73–82. 15

Moore, J. H. and Ritchie, M. D. (2004). The Challenges of Whole-Genome Approaches to Common Diseases. *JAMA*, 291(13):1642–1643. 30, 32

Moore, J. H. and White, B. C. (2007). Tuning ReliefF for Genome-Wide Genetic Analysis. In *Evolutionary Computation,Machine Learning and Data Mining in Bioinformatics*, Lecture Notes in Computer Science, pages 166–175. Springer, Berlin, Heidelberg. 59, 131

Moore, J. H. and Williams, S. M. (2005). Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis. *Bioessays*, 27(6):637–646. 12, 13, 14

Moore, J. H. and Williams, S. M. (2009). Epistasis and its implications for personal genetics. *Am. J. Hum. Genet.*, 85(3):309–320. 15, 30, 31

Morris, A. and Cardon, L. R. (2003). Whole Genome Association. In *Handbook of Statistical Genetics*. Balding DJ, Bishop M, Cannings C, Wiley, New York. 16

Motsinger-Reif, A. A., Reif, D. M., Fanelli, T. J., and Ritchie, M. D. (2008). A comparison of analytical methods for genetic association studies. *Genet. Epidemiol.*, 32(8):767–778. 64

Mueller, J. C. (2004). Linkage disequilibrium for different scales and applications. *Brief Bioinform*, 5(4):355–364. 10

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. 30, 42, 43

Neuman, R. J. and Rice, J. P. (1992). Two-locus models of disease. *Genet. Epidemiol.*, 9(5):347–365. 12

Niel, C., Sinoquet, C., Dina, C., and Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. *Front. Genet*, 6:285. 33, 56

Nishimura, D. (2001). BioCarta. *Biotech Software & Internet Report*, 2(3):117–120. 58

Niu, T. (2004). Algorithms for inferring haplotypes. *Genet. Epidemiol.*, 27(4):334–347. 42

O'Connor, C. (2008). Meiosis, genetic recombination, and sexual reproduction. *Nature Education*, 1(1):174. 7

Ogura, Y., Bonen, D. K., Inohara, N., Nicolae, D. L., Chen, F. F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R. H., Achkar, J. P., Brant, S. R., Bayless, T. M., Kirschner, B. S., Hanauer, S. B., Nuñez, G., and Cho, J. H. (2001). A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature*, 411(6837):603–606. 126

Oh, S., Lee, J., Kwon, M.-S., Weir, B., Ha, K., and Park, T. (2012). A novel method to identify high order gene-gene interactions in genome-wide association studies: Gene-based MDR. *BMC Bioinformatics*, 13(Suppl 9):S5. gene-based MDR. 66

Okazaki, T., Wang, M.-H., Rawsthorne, P., Sargent, M., Datta, L. W., Shugart, Y. Y., Bernstein, C. N., and Brant, S. R. (2008). Contributions of IBD5, IL23R, ATG16L1, and NOD2 to Crohn's Disease Risk in a Population-Based Case-Control Study: Evidence of Gene–Gene Interactions. *Inflamm Bowel Dis*, 14(11):1528–1541. 126

Park, M. Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostat*, 9(1):30–50.

Pascal, V., Pozuelo, M., Borruel, N., Casellas, F., Campos, D., Santiago, A., Martinez, X., Varela, E., Sarrabayrouse, G., Machiels, K., Vermeire, S.,

Sokol, H., Guarner, F., and Manichanh, C. (2017). A microbial signature for Crohn's disease. *Gut*, pages gutjnl–2016–313235. 125, 126

Peng, Q., Zhao, J., and Xue, F. (2010). A gene-based method for detecting gene–gene co-association in a case–control association study. *Eur J Hum Genet*, 18(5):582–587. CCU. 66, 75, 110

Phillips, P. C. (2008). Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*, 9(11):855–867. 14

Piriyapongsa, J., Ngamphiw, C., Intarapanich, A., Kulawonganunchai, S., Assawamakin, A., Bootchai, C., Shaw, P. J., and Tongsima, S. (2012). iLOCi: A SNP interaction prioritization technique for detecting epistasis in genome-wide association studies. *BMC Genomics*, 13(Suppl 7):S2. iLOCI. 63

Prabhu, S. and Pe'er, I. (2012). Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. *Genome Res.*, 22(11):2230–2240. SIXPAC. 63

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. 22

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959. 22

Pulst, S. M. (1999). Genetic Linkage Analysis. *Arch Neurol*, 56(6):667–672. 16

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575. 57, 62, 63

Qin, Z. S., Gopalakrishnan, S., and Abecasis, G. R. (2006). An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria. *Bioinformatics*, 22(2):220–225. 43

Rajapakse, I., Perlman, M. D., Martin, P. J., Hansen, J. A., and Kooperberg, C. (2012). Multivariate detection of gene-gene interactions. *Genet. Epidemiol.*, 36(6):622–630. CLD. 66, 110

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J.,

Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., and Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, 444(7118):444–454. 29

Reveille, J. D., Sims, A.-M., Danoy, P., Evans, D. M., Leo, P., Pointon, J. J., Jin, R., Zhou, X., Bradbury, L. A., Appleton, L. H., Davis, J. C., Diekman, L., Doan, T., Dowling, A., Duan, R., Duncan, E. L., Farrar, C., Hadler, J., Harvey, D., Karaderi, T., Mogg, R., Pomeroy, E., Pryce, K., Taylor, J., Savage, L., Deloukas, P., Kumanduri, V., Peltonen, L., Ring, S. M., Whittaker, P., Glazov, E., Thomas, G. P., Maksymowych, W. P., Inman, R. D., Ward, M. M., Stone, M. A., Weisman, M. H., Wordsworth, B. P., and Brown, M. A. (2010). Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. *Nat Genet*, 42(2):123–127. 122

Richards, J. E. and Hawley, R. S. (2010). Chapter 12 - The Human Genome: How the Sequence Enables Genome-wide Studies. In *The Human Genome : A User's Guide*, pages 405–452. Academic Press/Elsevier. 5

Rioux, J. D., Daly, M. J., Silverberg, M. S., Lindblad, K., Steinhart, H., Cohen, Z., Delmonte, T., Kocher, K., Miller, K., Guschwan, S., Kulbokas, E. J., O'Leary, S., Winchester, E., Dewar, K., Green, T., Stone, V., Chow, C., Cohen, A., Langelier, D., Lapointe, G., Gaudet, D., Faith, J., Branco, N., Bull, S. B., McLeod, R. S., Griffiths, A. M., Bitton, A., Greenberg, G. R., Lander, E. S., Siminovitch, K. A., and Hudson, T. J. (2001). Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat. Genet.*, 29(2):223–228. 126

Ritchie, M. D. (2011). Using Biological Knowledge to Uncover the Mystery in the Search for Epistasis in Genome-Wide Association Studies. *Ann Hum Genet*, 75(1):172–182. 59

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. MDR. 63

Ritchie, M. D., White, B. C., Parker, J. S., Hahn, L. W., and Moore, J. H. (2003). Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. *BMC Bioinformatics*, 4:28. BPNN/GPNN. 64

Saxena, R., Voight, B. F., Lyssenko, V., Burtt, N. P., de Bakker, P. I. W., Chen, H., Roix, J. J., Kathiresan, S., Hirschhorn, J. N., Daly, M. J., Hughes, T. E., Groop, L., Altshuler, D., Almgren, P., Florez, J. C., Meyer, J., Ardlie, K., Bengtsson, K., Isomaa, B., Lettre, G., Lindblad, U., Lyon, H. N., Melander,

O., Newton-Cheh, C., Nilsson, P., Orho-Melander, M., Råstam, L., Speliotes, E. K., Taskinen, M.-R., Tuomi, T., Guiducci, C., Berglund, A., Carlson, J., Gianniny, L., Hackett, R., Hall, L., Holmkvist, J., Laurila, E., Sjögren, M., Sterner, M., Surti, A., Svensson, M., Svensson, M., Tewhey, R., Blumenstiel, B., Parkin, M., DeFelice, M., Barry, R., Brodeur, W., Camarata, J., Chia, N., Fava, M., Gibbons, J., Handsaker, B., Healy, C., Nguyen, K., Gates, C., Sougnez, C., Gage, D., Nizzari, M., Gabriel, S. B., Chirn, G.-W., Ma, Q., Parikh, H., Richardson, D., Ricke, D., and Purcell, S. (2007). Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels. *Science*. 27

Schaid, D. J. (2004a). Evaluating associations of haplotypes with traits. *Genet. Epidemiol.*, 27(4):348–364. 42

Schaid, D. J. (2004b). Genetic epidemiology and haplotypes. *Genet. Epidemiol.*, 27(4):317–320. 42

Schaid, D. J., Sinnwell, J. P., Jenkins, G. D., McDonnell, S. K., Ingle, J. N., Kubo, M., Goss, P. E., Costantino, J. P., Wickerham, D. L., and Weinshilboum, R. M. (2012). Using the Gene Ontology to Scan Multi-Level Gene Sets for Associations in Genome Wide Association Studies. *Genet Epidemiol*, 36(1):3–16. 43

Schlosstein, L., Terasaki, P. I., Bluestone, R., and Pearson, C. M. (1973). High Association of an HL-A Antigen, W27, with Ankylosing Spondylitis. *New England Journal of Medicine*, 288(14):704–706. 122

Schüpbach, T., Xenarios, I., Bergmann, S., and Kapur, K. (2010). FastEpistasis: A high performance computing solution for quantitative trait epistasis. *Bioinformatics*, 26(11):1468–1469. FastEpistasis.

Schwarz, D. F., König, I. R., and Ziegler, A. (2010). On safari to Random Jungle: A fast implementation of Random Forests for high-dimensional data. *Bioinformatics*, 26(14):1752–1758. Random Jungle. 65

Schwender, H. and Ickstadt, K. (2008). Identification of SNP interactions using logic regression. *Biostatistics*, 9(1):187–198. logicFS. 62

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., Erdos, M. R., Stringham, H. M., Chines, P. S., Jackson, A. U., Prokunina-Olsson, L., Ding, C.-J., Swift, A. J., Narisu, N., Hu, T., Pruim, R., Xiao, R., Li, X.-Y., Conneely, K. N., Riebow, N. L., Sprau, A. G., Tong, M., White, P. P., Hetrick, K. N., Barnhart, M. W., Bark, C. W., Goldstein, J. L., Watkins, L., Xiang, F., Saramies, J., Buchanan, T. A., Watanabe, R. M., Valle, T. T., Kinnunen, L., Abecasis, G. R., Pugh, E. W., Doheny, K. F., Bergman, R. N., Tuomilehto, J., Collins, F. S., and Boehnke, M. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, 316(5829):1341–1345. 27

Seng, K. C. and Seng, C. K. (2008). The success of the genome-wide association approach: A brief story of a long struggle. *Eur J Hum Genet*, 16(5):554–564. 27

Sengupta Chattopadhyay, A., Hsiao, C.-L., Chang, C. C., Lian, I.-B., and Fann, C. S. J. (2014). Summarizing techniques that combine three non-parametric scores to detect disease-associated 2-way SNP-SNP interactions. *Gene*, 533(1):304–312. PCS/ZSS.

Shang, J., Zhang, J., Sun, Y., Liu, D., Ye, D., and Yin, Y. (2011). Performance analysis of novel methods for detecting epistasis. *BMC Bioinformatics*, 12(1):475. 56

Shao, J. (1993). Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association*, 88(422):486–494. 47

Shen, Y., Liu, Z., and Ott, J. (2010). Detecting gene-gene interactions using support vector machines with L1 penalty. In *2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, pages 309–311. 64

Stanislas, V., Dalmasso, C., and Ambroise, C. (2017). Eigen-Epistasis for detecting gene-gene interactions. *BMC Bioinformatics*, 18(1):54. 32

Steen, K. V. (2012). Travelling the world of gene-gene interactions. *Brief. Bioinformatics*, 13(1):1–19. 33, 56

Stephens, M. and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nat Rev Genet*, 10(10):681–690. 31

Stram, D. O. (2004). Tag SNP selection for association studies. *Genet. Epidemiol.*, 27(4):365–374. 41

Su, S.-Y., Balding, D. J., and Coin, L. J. M. (2008). Disease association tests by inferring ancestral haplotypes using a hidden markov model. *Bioinformatics*, 24(7):972–978. 42

Sun, X., Lu, Q., Mukherjee, S., Crane, P. K., Elston, R., and Ritchie, M. D. (2014). Analysis pipeline for the epistasis search – statistical versus biological filtering. *Front Genet*, 5. 33, 58

Szymczak, S., Biernacka, J. M., Cordell, H. J., González-Recio, O., König, I. R., Zhang, H., and Sun, Y. V. (2009). Machine learning in genome-wide association studies. *Genet. Epidemiol.*, 33 Suppl 1:S51–57. 31

Tang, W., Wu, X., Jiang, R., and Li, Y. (2009). Epistatic Module Detection for Case-Control Studies: A Bayesian Model with a Gibbs Sampling Strategy. *PLoS Genet*, 5(5). epiMODE.

Tenesa, A. and Haley, C. S. (2013). The heritability of human disease: Estimation, uses and abuses. *Nat. Rev. Genet.*, 14(2):139–149. 27

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320. 17

Thomas, D. (2010). Gene-Environment-Wide Association Studies: Emerging Approaches. *Nat Rev Genet*, 11(4):259–272. 29

Thomas, D. C. (2004). *Statistical Methods in Genetic Epidemiology*. Oxford University Press, Oxford, New York. 15, 23

Thomas, G., Jacobs, K. B., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N., Welch, R., Hutchinson, A., Crenshaw, A., Cancel-Tassin, G., Staats, B. J., Wang, Z., Gonzalez-Bosquet, J., Fang, J., Deng, X., Berndt, S. I., Calle, E. E., Feigelson, H. S., Thun, M. J., Rodriguez, C., Albanes, D., Virtamo, J., Weinstein, S., Schumacher, F. R., Giovannucci, E., Willett, W. C., Cussenot, O., Valeri, A., Andriole, G. L., Crawford, E. D., Tucker, M., Gerhard, D. S., Fraumeni, J. F., Hoover, R., Hayes, R. B., Hunter, D. J., and Chanock, S. J. (2008). Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.*, 40(3):310–315. 27

Thomas, G. P. and Brown, M. A. (2010). Genetics and genomics of ankylosing spondylitis. *Immunol. Rev.*, 233(1):162–180. 122

Thompson, E. (2003). Linkage Analysis. In *Handbook of Statistical Genetics*. Balding DJ, Bishop M, Cannings C, Wiley, New York. 15

Thompson, W. D. (1991). Effect modification and the limits of biological inference from epidemiologic data. *Journal of Clinical Epidemiology*, 44(3):221–232. 14

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288. 46

Tomita, Y., Tomida, S., Hasegawa, Y., Suzuki, Y., Shirakawa, T., Kobayashi, T., and Honda, H. (2004). Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. *BMC Bioinformatics*, 5:120.

Török, H. P., Glas, J., Endres, I., Tonenchi, L., Teshome, M. Y., Wetzke, M., Klein, W., Lohse, P., Ochsenkühn, T., Folwaczny, M., Göke, B., Folwaczny, C., Müller-Myhsok, B., and Brand, S. (2009). Epistasis between Toll-like receptor-9 polymorphisms and variants in NOD2 and IL23R modulates susceptibility to Crohn's disease. *Am. J. Gastroenterol.*, 104(7):1723–1733. 126

Trinh, T. T. and Rioux, J. D. (2005). The promise and perils of interpreting genetic associations in Crohn's disease. *Gut*, 54(10):1354–1357. 14

Tsui, F., Tsui, H. W., Akram, A., Haroon, N., and Inman, R. (2014). The genetic basis of ankylosing spondylitis: New insights into disease pathogenesis. *The Application of Clinical Genetics*, 7:105–15. 122, 124

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., de Andrade, M., Doheny, K. F., Haines, J. L., Hayes, G., Jarvik, G., Jiang, L., Kullo, I. J., Li, R., Ling, H., Manolio, T. A., Matsumoto, M., McCarty, C. A., McDavid, A. N., Mirel, D. B., Paschall, J. E., Pugh, E. W., Rasmussen, L. V., Wilke, R. A., Zuvich, R. L., and Ritchie, M. D. (2011a). Quality Control Procedures for Genome Wide Association Studies. *Curr Protoc Hum Genet*, CHAPTER:Unit1.19. 20, 21

Turner, S. D., Berg, R. L., Linneman, J. G., Peissig, P. L., Crawford, D. C., Denny, J. C., Roden, D. M., McCarty, C. A., Ritchie, M. D., and Wilke, R. A. (2011b). Knowledge-Driven Multi-Locus Analysis Reveals Gene-Gene Interactions Influencing HDL Cholesterol Level in Two Independent EMR-Linked Biobanks. *PLOS ONE*, 6(5):e19586. 58

Turner, S. D., Dudek, S. M., and Ritchie, M. D. (2010). ATHENA: A knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait Loci. *BioData Min*, 3:5. ATHENA. 64

Ueki, M. and Cordell, H. J. (2012). Improved Statistics for Genome-Wide Interaction Analysis. *PLoS Genet*, 8(4):e1002625. 63

Ueki, M. and Tamiya, G. (2012). Ultrahigh-dimensional variable selection method for whole-genome gene-gene interaction analysis. *BMC Bioinformatics*, 13(1):72. EPISIS.

Upstill-Goddard, R., Eccles, D., Fliege, J., and Collins, A. (2012). Machine learning approaches for the discovery of gene–gene interactions in disease data. *Brief Bioinform*, page bbs024. 63

van Heel, D. A., Ghosh, S., Hunt, K. A., Mathew, C. G., Forbes, A., Jewell, D. P., and Playford, R. J. (2005). Synergy between TLR9 and NOD2 innate immune responses is lost in genetic Crohn's disease. *Gut*, 54(11):1553–1557. 126

Vapnik, V. (1979). *Estimation of Dependences Based on Empirical Data*. Moscow: Nauka. 55

Vinkhuyzen, A. A. E., Wray, N. R., Yang, J., Goddard, M. E., and Visscher, P. M. (2013). Estimation and Partition of Heritability in Human Populations Using Whole-Genome Analysis Methods. *Annual Review of Genetics*, 47(1):75–95. 26

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of GWAS discovery. *Am. J. Hum. Genet.*, 90(1):7–24. 20, 27

Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the genomics era — concepts and misconceptions. *Nat Rev Genet*, 9(4):255–266. 26

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L., and Yu, W. (2010a). BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies. *Am J Hum Genet*, 87(3):325–340. BOOST. 62

Wan, X., Yang, C., Yang, Q., Xue, H., Tang, N. L., and Yu, W. (2009). MegaSNPHunter: A learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study. *BMC Bioinformatics*, 10(1):13. MegaSNPHunter.

Wan, X., Yang, C., Yang, Q., Xue, H., Tang, N. L. S., and Yu, W. (2010b). Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics*, 26(1):30–37. SNPRuler.

Wang, K. and Abbott, D. (2008). A principal components regression approach to multilocus genetic association studies. *Genet. Epidemiol.*, 32(2):108–118. PCReg approach. 44

Wang, T., Ho, G., Ye, K., Strickler, H., and Elston, R. C. (2009). A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. *Genet. Epidemiol.*, 33(1):6–15. PLS approach. 67, 74

Wang, X., Elston, R. C., and Zhu, X. (2011a). The Meaning of Interaction. *Hum Hered*, 70(4):269–277. 11

Wang, X., Epstein, M. P., and Tzeng, J.-Y. (2014a). Analysis of gene-gene interactions using gene-trait similarity regression. *Hum. Hered.*, 78(1):17–26. SimReg. 68

Wang, X., Zhang, D., and Tzeng, J.-Y. (2014b). Pathway-Guided Identification of Gene-Gene Interactions. *Annals of Human Genetics*, 78(6):478–491. Pathway guided. 69, 84, 85

Wang, Y., Liu, X., Robbins, K., and Rekaya, R. (2010). AntEpiSeeker: Detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. *BMC Research Notes*, 3(1):117. AntEpiSeeker.

Wang, Z., Wang, Y., Tan, K.-L., Wong, L., and Agrawal, D. (2011b). eCEO: An efficient Cloud Epistasis cOmputing model in genome-wide association study. *Bioinformatics*, 27(8):1045–1051. eCEO.

Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., Freathy, R. M., Perry, J. R. B., Stevens, S., Hall, A. S., Samani, N. J., Shields, B., Prokopenko, I., Farrall, M., Dominiczak, A., Johnson, T., Bergmann, S., Beckmann, J. S., Vollenweider, P., Waterworth, D. M., Mooser, V., Palmer, C. N. A., Morris, A. D., Ouwehand, W. H., Caulfield, M., Munroe, P. B., Hattersley, A. T., McCarthy, M. I., Frayling, T. M., Zhao, J. H., Li, S., Loos, R. J. F., Barroso, I., Deloukas, P., Sandhu, M. S., Wheeler, E., Soranzo, N., Inouye, M., and Wareham, N. J. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet*, 40(5):575–583. 28

Wei, C. and Lu, Q. (2014). GWGGI: Software for genome-wide gene-gene interaction analysis. *BMC Genetics*, 15:101. GWGGI.

Wei, W.-H., Hemani, G., and Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nat Rev Genet*, 15(11):722–733. 33, 56, 57

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678. 20, 126

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*, 42(Database issue):D1001–D1006. 17

Winham, S. J., Colby, C. L., Freimuth, R. R., Wang, X., de Andrade, M., Huebner, M., and Biernacka, J. M. (2012). SNP interaction detection with Random Forests in high-dimensional genetic data. *BMC Bioinformatics*, 13:164. 65

Witte, J. S. (2010). Genome-Wide Association Studies and Beyond. *Annu Rev Public Health*, 31:9–20. 18, 28, 29, 30

Wittke-Thompson, J. K., Pluzhnikov, A., and Cox, N. J. (2005). Rational Inferences about Departures from Hardy-Weinberg Equilibrium. *Am J Hum Genet*, 76(6):967–986. 22

Woodrow, J. C. and Eastmond, C. J. (1978). HLA B27 and the genetics of ankylosing spondylitis. *Ann Rheum Dis*, 37(6):504–509. 122

Wu, C. and Cui, Y. (2014). Boosting signals in gene-based association studies via efficient SNP selection. *Brief. Bioinformatics*, 15(2):279–291. diSNPselection. 43

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. Screen and Clean.

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. logistic kernel machine based SNP set analysis. 44

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am J Hum Genet*, 89(1):82–93. 31, 44

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721. 58, 83

Wu, X., Dong, H., Luo, L., Zhu, Y., Peng, G., Reveille, J. D., and Xiong, M. (2010c). A Novel Statistic for Genome-Wide Interaction Analysis. *PLoS Genet*, 6(9). T1H. 63

Wu, X., Jin, L., and Xiong, M. (2008). Composite measure of linkage disequilibrium for testing interaction between unlinked loci. *Eur J Hum Genet*, 16(5):644–651. Composite LD based test. 63

Xu, J., Yuan, Z., Ji, J., Zhang, X., Li, H., Wu, X., Xue, F., and Liu, Y. (2016). A powerful score-based test statistic for detecting gene-gene co-association. *BMC Genet*, 17(1):1–8. SBS. 67

Yang, C., He, Z., Wan, X., Yang, Q., Xue, H., and Yu, W. (2009a). SNPHarvester: A filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics*, 25(4):504–511. SNP Harvester. 62

Yang, C., Wan, X., Yang, Q., Xue, H., and Yu, W. (2010a). Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group Lasso. *BMC Bioinformatics*, 11(Suppl 1):S18. AGL.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. (2010b). Common SNPs explain a large proportion of heritability for human height. *Nat Genet*, 42(7):565–569. 29

Yang, Y., He, C., and Ott, J. (2009b). Testing association with interactions by partitioning chi-squares. *Ann. Hum. Genet.*, 73(1):109–117. Chi-Square Test. 63

Yeager, M., Orr, N., Hayes, R. B., Jacobs, K. B., Kraft, P., Wacholder, S., Minichiello, M. J., Fearnhead, P., Yu, K., Chatterjee, N., Wang, Z., Welch, R., Staats, B. J., Calle, E. E., Feigelson, H. S., Thun, M. J., Rodriguez, C., Albanes, D., Virtamo, J., Weinstein, S., Schumacher, F. R., Giovannucci, E., Willett, W. C., Cancel-Tassin, G., Cussenot, O., Valeri, A., Andriole,

G. L., Gelmann, E. P., Tucker, M., Gerhard, D. S., Fraumeni, J. F., Hoover, R., Hunter, D. J., Chanock, S. J., and Thomas, G. (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.*, 39(5):645–649. 27

Yee, J., Kwon, M.-S., Park, T., and Park, M. (2013). A Modified Entropy-Based Approach for Identifying Gene-Gene Interactions in Case-Control Study. *PLOS ONE*, 8(7):e69321. RIG.

Yi, N., Kaklamani, V. G., and Pasche, B. (2011). Bayesian analysis of genetic interactions in case-control studies, with application to adiponectin genes and colorectal cancer risk. *Ann. Hum. Genet.*, 75(1):90–104. Bayesian GLM.

Yoshida, M. and Koike, A. (2011). SNPInterForest: A new method for detecting epistatic interactions. *BMC Bioinformatics*, 12(1):469. SNPInterForest.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67. 46, 48, 78

Yuan, Z., Gao, Q., He, Y., Zhang, X., Li, F., Zhao, J., and Xue, F. (2012). Detection for gene-gene co-association via kernel canonical correlation analysis. *BMC Genet*, 13:83. KCCU. 67, 75

Yuan, Z., Liu, H., Zhang, X., Li, F., Zhao, J., Zhang, F., and Xue, F. (2013). From Interaction to Co-Association —A Fisher r -To- z Transformation-Based Simple Statistic for Real World Genome-Wide Association Study. *PLOS ONE*, 8(7):e70774. Fisher r to z. 67

Yung, L. S., Yang, C., Wan, X., and Yu, W. (2011). GBOOST: A GPU-based tool for detecting gene–gene interactions in genome–wide case control studies. *Bioinformatics*, 27(9):1309–1310. GBOOST.

Zaykin, D., Zhivotovsky, L. A., Westfall, P., and Weir, B. (2002). Truncated product method for combining P-values. *Genet. Epidemiol.*, 22(2):170–185. 66, 110

Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., Timpson, N. J., Perry, J. R. B., Rayner, N. W., Freathy, R. M., Barrett, J. C., Shields, B., Morris, A. P., Ellard, S., Groves, C. J., Harries, L. W., Marchini, J. L., Owen, K. R., Knight, B., Cardon, L. R., Walker, M., Hitman, G. A., Morris, A. D., Doney, A. S. F., Wellcome Trust Case Control Consortium (WTCCC), McCarthy, M. I., and Hattersley, A. T. (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, 316(5829):1336–1341. 27

Zhang, F. and Wagener, D. (2008). An approach to incorporate linkage disequilibrium structure into genomic association analysis. *Journal of Genetics and Genomics*, 35(6):381–385. PC-LR. 44, 74

Zhang, F., Xie, D., Liang, M., and Xiong, M. (2016). Functional Regression Models for Epistasis Analysis of Multiple Quantitative Traits. *PLoS Genet*, 12(4). MFRG. 68

Zhang, L., Liu, R., Wang, Z., Culver, D. A., and Wu, R. (2012). Modeling haplotype-haplotype interactions in case-control genetic association studies. *Front Genet*, 3:2. Haplo-Haplo. 69

Zhang, X., Huang, S., Zou, F., and Wang, W. (2010). TEAM: Efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*, 26(12):i217–i227. TEAM.

Zhang, X., Yang, X., Yuan, Z., Liu, Y., Li, F., Peng, B., Zhu, D., Zhao, J., and Xue, F. (2013). A PLSPM-Based Test Statistic for Detecting Gene-Gene Co-Association in Genome-Wide Association Study with Case-Control Design. *PLoS ONE*, 8(4):e62129. PLSPM Based Test. 67, 110

Zhang, Y. and Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nat Genet*, 39(9):1167–1173. BEAM.

Zhao, B., Zhu, Y., Wang, W., Cui, H.-m., Wang, Y.-p., and Lai, J.-h. (2013). Analysis of Variations in the Glutamate Receptor, N-Methyl D-Aspartate 2A (GRIN2A) Gene Reveals Their Relative Importance as Genetic Susceptibility Factors for Heroin Addiction. *PLoS One*, 8(8). 11

Zhao, J., Jin, L., and Xiong, M. (2006). Test for Interaction between Two Unlinked Loci. *Am J Hum Genet*, 79(5):831–845. LD-based test. 63

Zhao, J., Zhu, Y., and Xiong, M. (2016). Genome-wide gene-gene interaction analysis for next-generation sequencing. *Eur. J. Hum. Genet.*, 24(3):421–428. FLR. 68

Zheng, G., Yang, Y., Zhu, X., and Elston, R. (2012). *Analysis of Genetic Association Studies*. 22, 23, 24

Zhu, R., Zhao, H., and Ma, S. (2014). Identifying gene-environment and gene-gene interactions using a progressive penalization approach. *Genet. Epidemiol.*, 38(4):353–368. 62

Zhu, Z., Tong, X., Zhu, Z., Liang, M., Cui, W., Su, K., Li, M. D., and Zhu, J. (2013). Development of GMDR-GPU for gene-gene interaction analysis and its application to WTCCC GWAS data for type 2 diabetes. *PLoS ONE*, 8(4):e61943. GMDR-GPU. 64

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *PNAS*, 109(4):1193–1198. 2, 31

université PARIS-SACLAY
ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Approches statistiques pour la détection d'épistasie dans les études d'associations pangénomiques

Mots clefs : épistasie, études d'associations pangénomiques, régression en grande dimension

Résumé : De nombreux travaux de recherche portent sur la détection et l'étude des interactions dans les études d'association pangénomique (GWAS). La plupart des méthodes proposées se concentrent principalement sur les interactions entre polymorphismes simples de l'ADN (SNPs), mais des stratégies de regroupement peuvent également être envisagées.
Dans cette thèse, nous développons une approche originale pour la détection des interactions à l'échelle des gènes. De nouvelles variables représentant les interactions entre deux gènes sont définies à l'aide de méthodes de réduction de dimension. Ainsi, toutes les informations apportées par les marqueurs génétiques sont résumées au niveau du gène. Ces nouvelles variables d'interaction sont ensuite introduites dans un modèle de régression. La sélection des effets significatifs est réalisée à l'aide d'une méthode de régression pénalisée basée sur le Group LASSO avec contrôle du taux de fausse découvertes.
Nous comparons les différentes méthodes de modélisation des variables d'interaction à travers des études de simulations afin de montrer les bonnes performances de notre approche. Enfin, nous illustrons son utilisation pratique pour identifier des interactions entre gènes en analysant deux jeux de données réelles.

Title : Statistical approaches to detect epistasis in genome wide association studies.

Keywords : epistasis, GWAS, high dimensional regression

Abstract : A large amount of research has been devoted to the detection and investigation of epistatic interactions in Genome-Wide Association Studies (GWAS). Most of the literature focuses on interactions between single-nucleotide polymorphisms (SNPs), but grouping strategies can also be considered.
In this thesis, we develop an original approach for the detection of interactions at the gene level. New variables representing the interactions between two genes are defined using dimensionality reduction methods. Thus, all information brought from genetic markers is summarized at the gene level. These new interaction variables are then introduced into a regression model. The selection of significant effects is done using a penalized regression method based on Group LASSO controlling the False Discovery Rate.
We compare the different methods of modeling interaction variables through simulations in order to show the good performance of our proposed approach. Finally, we illustrate its practical use for identifying gene-gene interactions by analyzing two real data sets.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France