2017 • 2018
FACULTY OF SCIENCES
Master of Statistics

Master thesis

The Impact of Correlated Genetic Markers on Large-scale DNA-based Gene-Gene Interaction Studies

External Supervisors :
Prof. dr. dr. Kristel Van Steen
Dr. Fentaw Abegaz

Internal Supervisor :
Prof. dr. Ziv Shkedy

Marc Joiret

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Bioinformatics.

Abstract

Background

Retrospective case-control designs are widely used in genetic epidemiological studies, especially genome wide association studies (GWAS), for the identification and characterization of susceptibility genes involved in common complex multifactorial human diseases. The basic idea is to compare genotypes of cases and controls. If alleles or genotypes frequencies at different loci are significantly different in cases and controls, these alleles or genotypes are claimed to be associated with the disease status. The evidence of these associations not only contribute to gene mapping programs but also provide insights to possible therapeutic approaches. Complex common diseases have multi-loci (or polygenic) risk factors which marginally only explain a small proportion of the genetic heritability of these diseases. It is suspected that a significant part of genetic risks of complex human diseases is due to interaction effects of two or more genes involved in different forms of epistasis. Among the statistical methods specifically designed to detect gene-gene interaction in genome-wide association studies, the Model-Based Multifactor-Dimensionality Reduction (MB-MDR) algorithm is a non-parametric method of interest. In GWAS, genotyping chips technologies provide between 300.000 and more than 1 million genetic markers (SNPs) information per subject. For 500.000 SNPs, the number of pairwise combinations of these markers is around 125 billion and so is the number of statistical hypotheses tests. This raises a serious multiple testing problem. The multiple testing issue has usually been addressed by different approaches like filtering through prioritization or Linkage Disequilibrium (LD) pruning in addition to family-wise error rate corrections.

Objectives

The performances of MB-MDR in detecting gene-gene interaction in genome-wide association studies is poorly documented in the literature. In this work, we aimed at building different simulated datasets harboring known hidden pairs of epistatic genetic markers. The objective is to use these simulated datasets to measure the sensitivity (power) of MB-MDR in different configurations of the embedded hidden known pair of markers. Two main concerns in building the simulation datasets require to incorporate realistic human genome linkage disequilibrium patterns and to control population stratification because both confound gene-gene interaction effect on the phenotype (disease).

Methods

The work is limited to a binary phenotype (case/control) and is limited to pairs of bi-allelic markers. Real LD patterns from two 250 kbps DNA segments from human chromosome 7 and 8 were downloaded from the HapMap 3 'GBR' subpopulation in order to avoid population substructure. Four configurations for which the position of pairs of 'causal' embedded markers varied within or between LD blocks were set, each with three different effect sizes of purely epistatic interaction of one locus (DSL 1) to the other locus (DSL 2). The three interaction effect sizes were tuned upfront with a logistic regression model. The sensitivity to detect the correct causal SNPs directly or their tagged SNPs indirectly was measured as the number of times out of 100 simulated datasets, the MB-MDR algorithm identified the correct pairs of loci of interests. No prioritization filtering was applied. Only LD pruning was carried out before MB-MDR analysis at four different thresholds of linkage disequilibrium $r^2 = 0.75, 0.60$, 0.50 and 0.20, in addition to no LD pruning at all.

Results

Due to the family wise error rate adopted as a multiple testing correction and to the associated large number of possible false positives, the exact sensitivity was never better than 0.70 without LD pruning and in most of our simulated datasets below 0.50. The signal sensitivity (indirect detection of the pairwise interaction under interest, due to linkage disequilibrium with tag-markers) was always better than 0.70 and in the range $0.70 - 0.95$ with a LD pruning at a threshold of at least $r^2 = 0.50$. In most of our simulated settings, a LD-pruning threshold of 0.20 tends to decrease the indirect sensitivity. It appears that the correlation between SNP markers and the causal loci is useful in detecting gene-gene interaction associated to a complex phenotype. The signal sensitivity, however, does not appear to be much affected by the effect size at least in the case of pure epistasis in our simulated datasets.

As a concrete application of our results, we applied the LD pruning threshold of $r^2 = 0.50$ and MB-MDR on the real life dataset of the ankylosing spondylitis dataset and checked that the already known epistatic functional variants were correctly identified (at least indirectly through their tagged-markers). An inconvenience still remains in the large number of hypothetically false positive encountered results caused by the large number of pairwise combinations leading to a huge number of multiple tests.

Note on softwares and genomic data web repositories

Bioinformatics researchers are advised to get familiar with the UNIX environment and with the Linux operating system. The GIGA multicore cluster platform was used to submit the computer intensive jobs described in this thesis. The main third party bioinformatics softwares of simulation environment or of statistical analysis that were used for this thesis were vcftools, Haploview, PLINK, simuPOP, MB-MDR and R packages.

Used Genomic data web repositories were : NCBI, Ensembl, 1000 Genome Project, UCSC, dbSNP.

Real life dataset

The ankylosing spondylatis dataset was obtained from the WTCCC2.

Keywords

Complex Diseases, GWAS, SNPs, Epistasis, Genetic and Statistical Interactions, Linkage Disequilibrium (LD), Epistasis simulation under LD patterns, Model Based Multi-factor Dimensionality Reduction (MB-MDR), Ankylosing Spondylitis.

Contents

List of figures

List of Tables

Acknowledgements

I am deeply grateful to Professor Dr Dr Kristel van Steen and to the GIGA in Liege for hosting me during the six months duration I spent on this thesis. I value the strong commitment and research interests that Professor Kristel Van Steen engaged in the current and fast moving methods used in genome-wide interactions studies (GWIS) and in detecting epistasis with scientific rigour and statistical power. It took me weeks to fully grasp the extent of the title of this thesis but it was worth the effort and I am sincerely thankful to Kristel for having proposed this advanced GWIS thematic to me at UHasselt as a master thesis topic. It was both a challenge and honour to work on such a vast research field in such a short time span. I very much appreciate her sharing with me the enthusiastic and passionate vision on how to track gene-gene interaction in large scale genome-wide retrospective case-control studies. Professor Kristel van Steen delineated the framework of the core of the thesis with an extraordinary synthetic bird-eye point of view and she introduced me to the latest international literature. She let me do the rest freely on my own. I am thankful for her trust and for her mentoring.

I would like to thank Professor Dr Ziv Shkedy, warmly, for having accepted to co-supervise this thesis, for his interest and extensive knowledge of high dimensionality data. I am indebted to the inspirational content of his lectures and notes in computer intensive methods for bioinformatics and to the assigned projects he gave us that have fueled my interest in applied statistical genetics and for research in biomedicine.

Dr Fentaw Abegaz provided me helpful and supportive comments or answers to my questions about the statistical methods implemented in MB-MDR and advised me as how to simulate datasets for prototyping the MB-MDR analysis workflow. I also want to thank the SEGI for the guest account they created for me and Kris Chaichoompu for his time in introducing me to the UNIX environment and how to *shell script* for submitting jobs on the computer cores cluster dedicated to the GIGA research group.

Numerous technical problems were encountered during the progress of this master thesis, especially related to the diversity of third party simulation softwares that could not be compiled properly due to the evolution of compiler versions or to extinction of support from the original software contributors or due to programming or data format inconsistencies. I want to address a special thank you to Myriam Nemry at the GIGA who often helped me in solving some of these awful issues and who informed me about the simuPOP environment scripted in Python that I finally used for generating 1200 simulation datasets. It must be emphasized that the sound background I benefited from courses using Python in UHasselt were of inestimable support : introduction to algorithms and programming, advanced programming, and computational biology. Professor Frank Neven, Professor Robert Brijder and Dimitri Surinx deserve my sincere appreciation for their courses and lectures as well.

The fulfillment of a master thesis and a Master program in Biostatitics and Bioinformatics required a huge amount of time, many of which had to be drawn from time I haven't spent with my daughters, Emmy, Sarah and Lara ; my partner Laurence, and my father. Their support during the past two years allowed me to pursue my study path that I really enjoy. I dedicate this master thesis to all of them.

January 07th, 2018
Marc Joiret

237-11, rue Walthère Jamar
4430 Ans, Belgium.
marc.joiret@student.uhasselt.be

Introduction

DNA sequences between humans are highly conserved and there is almost no variability within the human species : two *Homo sapiens* individuals taken at random have $99.7 - 99.9\%$ identity in their DNA sequence. The human genome (23 chromosome pairs) is about 3.2 Giga bp in size and it is estimated that the variability roughly concerns $0.1 - 0.3\%$ of the genome, i.e. $3 - 10$ million base-pairs is the order of magnitude for the genotype variability between any two human individuals. The NCBI dbSNP database has more than $\sim$ 70 million validated SNPs in the human species. There are SNP-genotyping chips technologies (Illumina and Affymetrics) that allow us to scan this genome variability and characterize the specific genotype of an individual on a genome-wide basis. The result of one individual genome-wide genotype is a record with a number of variables (SNPs) that ranges between 250.000 to more than 1 million and which provides a fairly good coverage of the human genome variability because of the so-called linkage disequilibrium. If such genotyping records are collected for thousands of individuals in case-control retrospective studies, one immediately sees that the dataset size will be huge and the statistical methods are bound to require computer intensive methods. The number (p) of variables (SNPs) is much larger than the number (n) of observations (individuals) and we have to deal with high dimensionality.

This Master thesis addresses both fields of *Computer intensive methods in Bioinformatics* [1] and *Statistical genetics/genetic epidemiology*.

Genetic epidemiology and statistical genetics have their own jargon. Though it is not the purpose of this master thesis to provide all the definitions, a glossary of the most important terms encountered in this thesis is provided in the appendix and we sum up here the most general concepts. An introductory course in genetic epidemiology was given by Claesen (2016) [2] and in the the book by Laird and Lange (2011) for statistical genetics [3]. A concise introduction to genome-wide association studies can be found in Bush and Moore (2012) [4] and reviews articles on gene-gene interaction can be found in Van Steen (2012) [5] and in Cordell (2009) [6] for the detection of gene-gene interaction that underlie human diseases. An interdisciplinary and consensual approach of the meaning of interaction in genetics is provided by Wang, Elston and Zhu [7]. Because gene-gene interaction is a focus of our work, we will provide shortly below a description of a well documented model that we found helpful in understanding genetic epistasis from a molecular biology perspective.

1.1 Important issues in Human Genetics and in Genomic Study Design

1.1.1 GWAS

Genome-wide association studies (GWAS) are providing a powerful tool for investigating the genetic architecture of common human diseases that have a complex multifactorial etiology. A central goal of human genetics is to identify genetic risk factors for common complex diseases such as inflammatory bowel disease (Crohn's disease), Alzheimer's disease, ankylosing spondylitis, type II diabetes or sporadic breast cancer and for rare Mendelian diseases such as phenylketonuria, Huntington disease, sickle cell anemia or cystic fibrosis. There are different approaches to identifying genetic risk factors. GWAS measures and analyzes DNA sequence variations from across the human genome with the purpose of identifying genetic risk factors for diseases that are common in the population. The ultimate goal is to make predictions about who is at risk and to identify the biological mechanisms of disease susceptibility with the aim of developing new prevention and treatment strategies.

1.1.2 Single Nucleotide Polymorphism (SNPs)

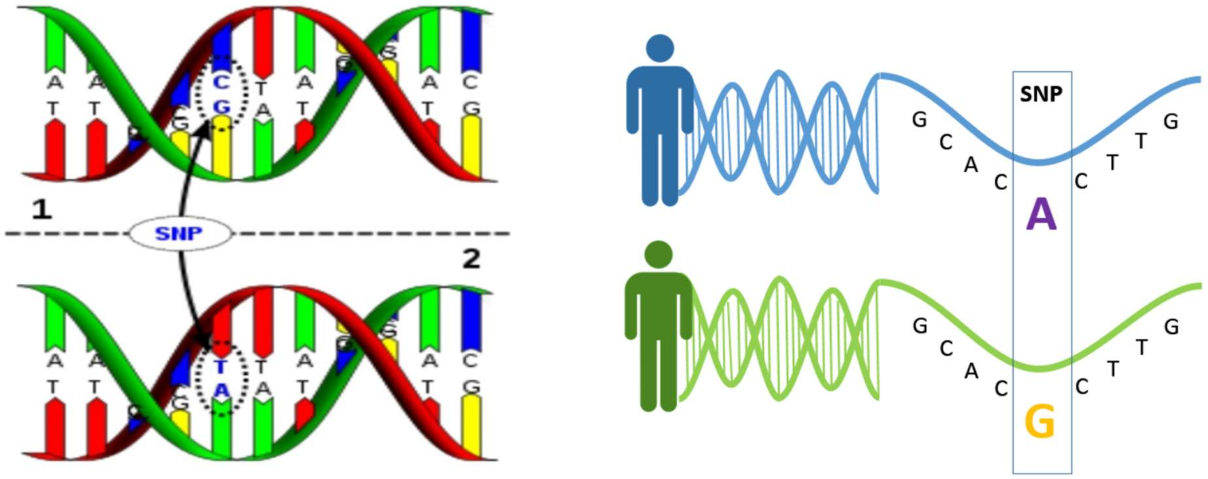


Figure 1.1 – SNP : Single Nucleotide Polymorphism. Left panel : SNP shown as alleles of two homologous chromosomes (heterozygous genotype). Right panel : SNP shown as two variants on two homologous chromosomes from two different subjects.

The unit of genetic variation is the single nucleotide polymorphism (SNP). The term polymorphism is defined simply as a genetic variant at a single location (within a gene or not). The double stranded DNA structure requires that each homologous chromosome has complementary base pairs at each location as displayed on the left panel of Figure 1.1, which shows a SNP at a pair of non-identical, yet homologous, chromosomes of a diploid individual. One chromosomal variant can be labeled 'A', and the other 'a' (bi-allelic variant). The 'A' allele should not be confused with the A (Adenine) base in the DNA sequence ; rather it is just a conventional notation for an allele. Due to the base pairing of the double-stranded structure of DNA (base complementarity), there is redundant information ; and we only need to read one strand. By convention, we read left to right from the $5' \rightarrow 3'$ strand (+ or direct strand, the green strand in the left panel of Figure 1.1). The two alleles in the figure differ only in the fifth base pair, where a C base is substituted for a T. Whether or not this difference is biologically meaningful depends on where it occurs in the sequence and on the nature of the letter change (synonymous or non-synonymous or stop-codon). The right panel of Figure 1.1 shows the SNPs variation between two individuals (only one of the two homologous chromosome is displayed). SNPs are the most abundant type of sequence variants in the genome, occurring approximately once in every 100 to 300 base-pairs [8]. If the SNP is functional (causal), the trait (phenotype or disease) is influenced directly.

Candidate gene studies generally involve multiple SNPs within a single gene. The choice of SNPs depends on defined linkage disequilibrium (LD) blocks and is discussed further below. The underlying motivation is that the SNPs under investigation capture information about the underlying genetic

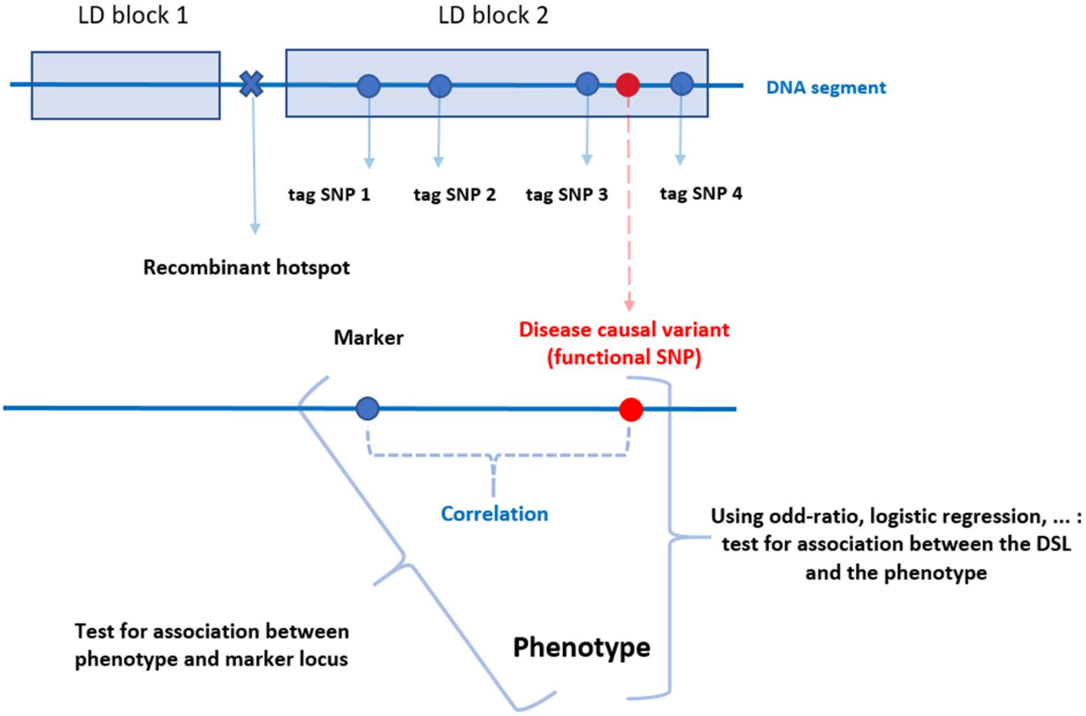


Figure 1.2 – LD blocks with tag-SNPs and causal variant (DSL). Credits : Foulkes [9].

variability of the gene under consideration, though the SNPs (tag-SNPs) may not serve as the true disease causing variants. Suppose we want to investigate the association between a gene and disease. A gene comprises a region of DNA representing a portion of the human genome. This is illustrated in Figure 1.2 inspired from Foulkes [9]. In a simple model, we could assume that a mutation at a single site within the gene region results in the disease. The precise location is generally not known. Instead, multiple SNPs that are presumed close to the functional locus on the genome are measured. The proximate SNPs are commonly referred to as markers since the observed genotype at these locations tends to be associated with the genotype at the true disease-causing locus (causal variant).

1.1.3 Complex diseases-traits

The term "complex trait" refers to any phenotype that does not exhibit classic Mendelian recessive or dominant inheritance attributable to a single gene locus [10]. Complexities arise when the simple correspondance between genotype and phenotype breaks down, either because the same genotype can result in different phenotypes due to effects of chance, environment, or interactions with other genes or different genotypes can result in the same phenotype. It is unlikely to find a genetic marker that shows perfect cosegregation with a complex trait. Complex traits are often characterized by incomplete penetrance and phenocopies, genetic loci heterogeneity and polygenic inheritance. Some individuals who inherit a predisposing allele may not exhibit the trait (incomplete penetrance), whereas others who inherit no predisposing allele may nonetheless get the trait as a result of environmental or random causes (phenocopy). So, the genotype at a given locus may affect the probability of the trait, but not fully determine the outcome. The penetrance function specifies the probability of the trait conditional on the genotype and the penetrance function can also depend on non genetic factors such as age, sex, environment and other genes. For example, the risk of breast cancer by ages 40, 55 and 80 is 37%, 66%, 85% in a woman carrying a mutation at the BRCA1 locus as compared with 0.4%, 3%, 8% in a non carrier. Locus heterogeneity means that mutations in any one of several genes may result in identical phenotypes, such as when the genes are required for a common biochemical pathway. Polygenic inheritance means that some traits may require the simultaneous presence of mutations in multiple genes.

1.1.4 Missing heritability

The basic idea in the study of variation is its partitioning into components attributable to different causes. We want to explain the variation in a phenotype of interest (complex disease) according to biologically plausible nature-nurture models. The components into which the phenotype variance (V_P) is partitioned are basically the genotype (V_G) and the environment (V_E), as detailed in [11]. The genotype or total genetic variance can further be split into additive variance, V_A (additive main effects for single loci, also called breeding value), into dominance variance V_D (interaction between alleles at the same locus) and interaction variance V_I (epistatic effects : interaction between alleles at different loci). There is also the possibility to have correlation between genetic effects and the environment in which case twice the covariance of genotypic values and environment deviations should be added in terms like $2\,\mathrm{Cov}(G, E)$ and interaction between genotypes and environments in which case an extra interaction term, like V_{G*E}, should be added. We take the strong assumption, for now, that genetic effects and environmental effects are independent and we neglect such covariance and interaction terms (because, currently, they often cannot be estimated). Thus :

$$\begin{aligned} V_P &= V_G + V_E && (1.1) \\ &= V_A + V_D + V_I + V_E + \ldots + \underbrace{2\,\mathrm{Cov}(G, E) + V_{G*E}}_{\text{neglected}} + \ldots && (1.2) \end{aligned}$$

The partitioning of the variance into its components allows us to estimate the relative importance of the various determinants of the phenotype. The relative importance of heredity in determining phenotypic values is called the heritability of the character. In quantitative classical genetics, there are two distinctly different meanings of heritability according to whether they refer to the additive component of the genetic variance or the overall component of the genetic variance.

The ratio V_G/V_P is called the heritability in the broad sense and expresses the extent to which individual's phenotype are determined by the genotypes. The ratio V_A/V_P is called heritability in the narrow sense and expresses, in classical quantitative genetics, the extent to which phenotypes are determined by the genes transmitted from parents.

In genome wide association studies of complex traits, one often refers to the "missing heritability". In this context, the missing heritability refers to the fact that the ratio V_A/V_G is surprisingly low for most complex traits.

One suspected possible cause for this missing heritability is the under-representation of complex interplays within and between sets of rare/common variants. Researchers now want to better address the V_I component. This is why the subject of interaction analysis has been investigated in what is called GWAIs (Genome-wide association interaction studies) since the turn of 2001-2003 and in the aftermath of the Human Genome Project and HapMap Projects.

1.1.5 Common disease-common variant (CD/CV) hypothesis

Common diseases have a different underlying genetic architecture than rare disorders. Several susceptibility variants for common diseases have been discovered and show high minor allele frequency (alleles in the apolipoprotein E or APOE gene for Alzheimer's disease are examples). This lead to the common disease/common variant (CD/CV) hypothesis. This hypothesis states that common disorders are likely influenced by genetic variation that is also common in the population. There are important consequences to this hypothesis as developed in [4].

If common genetic variants influence disease, the effect size (or penetrance) for any one variant must be small relative to that found for rare disorders. If a SNP with 40% frequency in the population causes a highly severe phenotype, nearly 40% of the population would have that phenotype. Thus the allele frequency and the population prevalence are completely correlated. If however, that same SNP caused a small change in gene expression that alter risk for disease by some small amount, the prevalence of the disease and the influential allele would be only slightly correlated. As such, common variants cannot have high penetrance.

If common alleles have small genetic effects (low penetrance), but common disorders show heritability (inheritance in families), then multiple common alleles must influence susceptibility. For example, twin studies might estimate the heritability of a common disease to be 40%, that is, 40% of the total variance in disease risk is due to genetic factors. If the allele of a single SNP incurs only a small degree of disease risk, that SNP only explains a small proportion of the total variance due to genetic factors. As such, the total genetic risk due to common genetic variation must be spread across multiple genetic factors.

These two arguments suggest that traditional family based genetic studies are not likely to be successful for complex diseases, prompting a shift toward population-based studies.

1.1.6 The Human Haplotype Map Project

The location and density of commonly occurring SNPs is needed to identify the genomic regions and individual sites that must be examined by genetic studies. Population-specific differences in genetic variation must be cataloged so that studies of phenotypes in different populations can be conducted with the proper design. Finally and most importantly, correlations among genetic variants must be known so that genetic studies do not collect redundant information. The International HapMap Project was designed to identify variation across the genome and to characterize correlations among variants. It was discovered at the turn of 2000-2002 that markers exhibit strong LD in extended regions called blocks that are separated by punctate breakpoints. Given the importance of linkage disequilibrium (LD) patterns in designing association studies (especially genome-wide), there is great interest in evaluating theoretical predictions of the extent and distribution of LD among markers. A full resolution of these issues would require a large-scale project to investigate the nature of LD across the entire genome. The international HapMap project was formally initiated in 2002 including researches from 20 groups in 6 countries with the aim of characterizing millions of DNA sequence variants, their frequencies and the correlations between them in samples from populations with ancestry from Africa, Asia and Europe. Further developments were made in HapMap 2, Hapmap 3 and the 1000 Genomes Project.

HapMap Anonymized samples were selected from four populations : 90 Yoruba (30 parent-parent-offspring trios) from Ibadan, Nigeria (abbreviated YRI) ; 90 individuals (30 trios) of European ancestry from Utah, collected in 1980 by the Centre d'Etude du Polymorphisme Humain (CEU) ; 45 unrelated Han Chinese from Beijing (CHB) ; and 45 unrelated Japanese from Tokyo (JPT). Pilot studies and simulation indicated that complete ascertainment in 45 unrelated individuals would represent 99% of variation with minor allele frequency (MAF) $> 5\%$ in the populations from which the samples were drawn. SNPs were preferentially selected with the aim of generating a map with a common (MAF $> 5\%$) SNP in each population every 5 kilobases. The resulting first generation HapMap [12] yielded 2 important advances : the validation of a large number of common SNPs (1.2 million unique SNPs passed quality control measures) and the opportunity to develop medium and high throughput genotyping technologies which enabled the subsequent wave of genome-wide association studies. The project also verified that the previously observed block-like patterns of LD generalized to the entire genome and was not an artifact of small sample size or lower marker density. Recombination hotspots are ubiquitous in the genome and are the driving force behind the observed LD patterns (with block of high LD corresponding to inter-hotspot intervals and sharp breakdowns of LD mapping to hotspots). The genome-wide LD map made it straightforward to select a subset of common markers that captures nearly all the information contained in the full set. This process of thinning out markers based on r^2, or "SNP tagging", made association studies more comprehensive and efficient.

HapMap 2 A second generation map was published in 2007 [13] with a total of 3.1 million SNPs which still provide better tag SNP selection, more precise estimates of local recombination rates.

HapMap 3 The third phase of HapMap [14] extended to include samples from a number of additional populations and provided a map with a common MAF $> 1\%$ SNP for 11 populations.

1000 Genomes Project The 1000 Genomes project [15] pushes the HapMap paradigm into the analysis of rarer variation, including copy number variation and short insertion/deletion polymorphisms in addition to SNPs. 26 populations around the world make up the database with 2.535 individuals. Some conclusions of this project include the following : More than 79 million variant sites have been validated. The highest rates of variation tended to occur at the HLA encoding region on chromosome 6 and subtelomeric regions. Lowest rates occured in 5 Mb, gene dense region around 3p21. The project is useful to impute SNPs for genome wide association studies. A typical individual human genome harbors more than 10.000 nonsynonymous variants. Each person in the project (each of us in general) has 20-40 variants at conserved sites that are identified as damaging ; 10-20 loss of function variants ; 2-5 damaging mutations ; 1-2 variants previously identified from cancer genome sequencing.

1.1.7 Linkage Disequilibrium

Linkage disequilibrium (LD) is a property of SNPs on a contiguous stretch of genomic sequence that describes the degree to which an allele of one SNPs is inherited or correlated with an allele of another SNP within a population. It is related to the concept of chromosomal linkage, where two markers on a chromosome remain physically joined on a chromosome through generations of a family provided no crossing over occurs between the two markers at meiosis between non-sister chromatids.

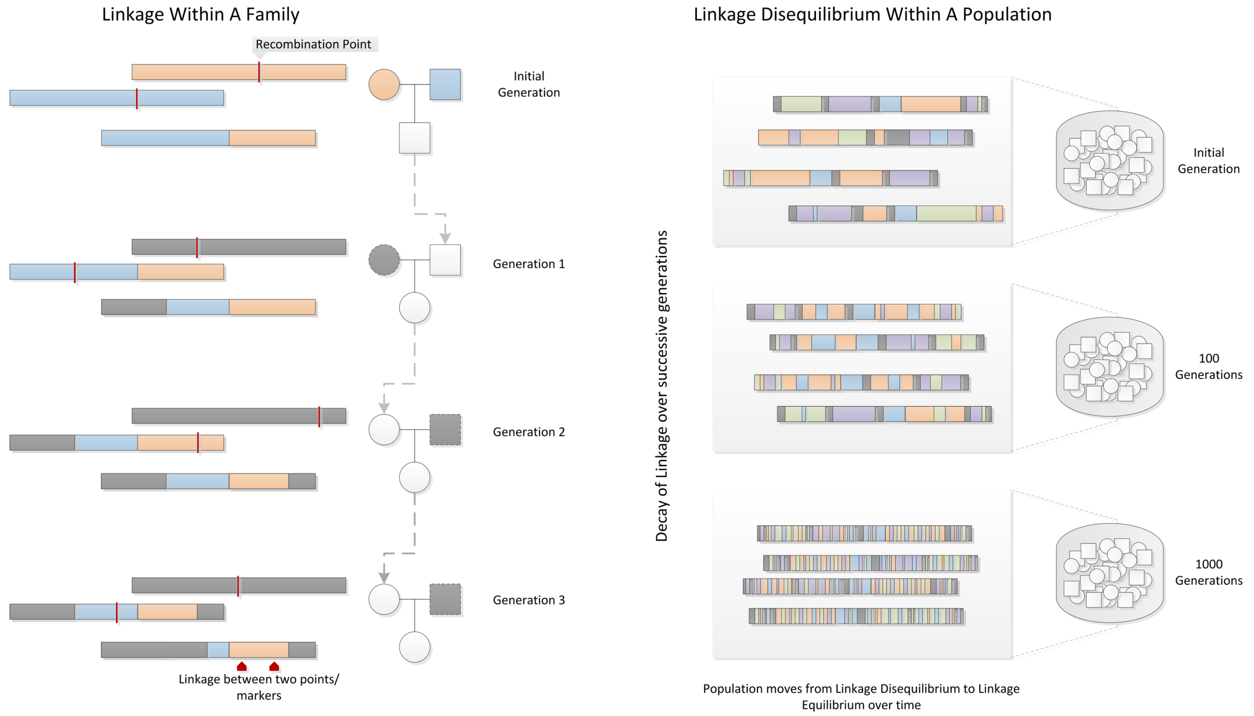

Linkage Within A Family
Recombination Point
Initial Generation
Generation 1
Generation 2
Generation 3
Linkage between two points/ markers
Linkage Disequilibrium Within A Population
Decay of Linkage over successive generations
Initial Generation
100 Generations
1000 Generations
Population moves from Linkage Disequilibrium to Linkage Equilibrium over time

Figure 1.3 – Linkage and Linkage Disequilibrium. Within a family, linkage occurs when two genetic markers (points on a chromosome) remain linked on a chromosome rather than being broken apart by recombination events during meiosis. In a population, contiguous stretches of founder chromosomes from the initial generation are sequentially reduced in size by recombination events. Over time, a pair of markers in the population move from linkage disequilibrium to linkage equilibrium, as recombination events eventually occur between every possible point on the chromosome. Credits : Bush WS and Moore JH [4].

In Figure 1.3, two founder chromosomes are shown (one in blue and one in orange). Recombination events within a family from generation to generation break apart chromosomal segments. This effect is amplified through generations, and in a population of fixed size undergoing random mating, repeated random recombination events will break apart segments of contiguous chromosome that contained linked alleles until eventually all alleles in the population are in linkage equilibrium or are independent. The linkage between markers on a population scale is referred to as linkage disequilibrium (LD). In samples of unrelated subjects, the 'genetic signal' (disease gene or disease susceptibility locus DSL) has a short range in which it can be detected at marker loci, if the genetic variant is old, i.e., it occurred for the first time many, many generations ago. The Figure 1.3 can also illustrate this property. Two 'unrelated', affected subjects whose disease is triggered by the same genetic variant may have had a common ancestor many generations ago in whom the disease mutation initially occurred making them cryptically related. In Figure 1.3, the blue and orange chromosome in the initial generation can be considered as the pair of chromosomes from the common ancestor in whom the disease variant (DSL somewhere in the orange chromosome) occurred for the first time. The last line in the right panel of Figure 1.3 shows a population of 'unrelated' study subjects whose disease chromosomes originated from the common ancestor with the original disease mutation. Due to the numerous recombination events that took place during the meiotic cell divisions between the generations (middle part of Figure), the original chromosomes on the common ancestor have been divided many times and the majority of its parts have been replaced by other copies of the same chromosomal segment. As a consequence the orange area around the original disease mutation that have remained unchanged are much smaller now, naturally reducing the range in which the genetic signal can be detected. The genetic marker loci have now to be very close to the disease mutation in order to identify the disease gene. However, any markers in the orange regions surrounding the DSL allele at the bottom of the left panel in Figure 1.3 have the important property that they share the same ancestral allele. That is, each diseased person shares the ancestral disease allele at the DSL location from the orange chromosome, and they also have the same allele at any marker in the orange area surrounding the DSL allele. In other words, two particular alleles, one from each locus, tend to appear together on the same haplotype in a population. The physical distance between loci is different from the linkage distance (Haldane distance, $L = -\frac{1}{2}\ln(1-\theta)$) but, as a rule of thumb, for the human species, 1 centiMorgan (cM), characterizing a 1% chance of recombination between two loci occurring at meiosis ($L = \theta$ when θ is small), corresponds to 1 million base pairs in physical distance. The rate of LD decay is dependent on multiple factors, including the population size, the number of founding chromosomes in the population, and the number of generations for which the population has existed. As such, different human subpopulations have different degrees and patterns of LD. African-descent populations are the most ancestral and have smaller regions of LD due to the accumulation of more recombination events in that group. European-descent and Asian descent populations were created by founder events (a sampling of chromosomes from the African population), which altered the number of founding chromosomes, the population size, and the generational age of the population. These populations on average have larger regions of LD than African-descent groups.

The two commonly used measures of linkage disequilibrium are D' and r^2 explained below.

Let the alleles at two markers be denoted A,a and B,b. Let the allele frequencies at each marker be p_A, p_a, p_B, p_b and let p_{AB}, p_{Ab}, p_{aB}, p_{ab} denote the frequencies of the four possible haplotypes. Thus p_{AB} denotes the frequency of a randomly selected haplotype from the population with alleles A and B observed at the two loci. Linkage Equilibrium implies that the haplotype frequencies are given by the product of the corresponding allele frequencies. The resulting frequencies are given in Table 1.1 in case of independence. The LE corresponds to our usual notion of independence in a 2×2 table. The haplotype frequency is just the joint probability of A and B being observed on the same haplotype, and the allele frequencies are the marginal frequencies. When LE fails, the number of alleles at two loci is not the product of the individual allele frequencies. The LD coefficient, usually denoted by D, measures the departure from independence :

Table 1.1 – Expected allele distribution under independence

	B locus		
A locus	B	b	Total
A	$p_{AB} = p_A p_B$	$p_{Ab} = p_A p_b$	p_A
a	$p_{aB} = p_a p_B$	$p_{ab} = p_a p_b$	p_a
Column Total	p_B	p_b	

$$D = p_{AB} - p_A p_B \tag{1.3}$$

A substantial difficulty with using D to measure the lack of independence in the 2×2 table shown in Table 1.2 is that D is highly sensitive to marginal values, which makes it difficult to compare LD among many pairs of markers with diverse frequencies. Furthermore, the sign of D depends on an arbitrary coding of the alleles. For this reason, two derived LD statistics are both frequency

Table 1.2 – Observed allele distribution under LD

	B locus		
A locus	B	b	Total
A	$p_{AB} = p_A p_B + D$	$p_{Ab} = p_A p_b - D$	p_A
a	$p_{aB} = p_a p_B - D$	$p_{ab} = p_a p_b + D$	p_a
Column Total	p_B	p_b	

normalized and are in use. We first define minimum and maximum values for D as :

$$\begin{aligned} D_{\min} &= \min(p_A p_B, p_a p_b) && (1.4) \\ D_{\max} &= \max(p_A p_b, p_a p_B) && (1.5) \end{aligned}$$

The first derived LD value is D' :
For D positive :

$$D' = \frac{D}{D_{\max}} \tag{1.6}$$

For D negative :

$$D' = \frac{D}{D_{\min}} \tag{1.7}$$

The second derived LD value is r^2, the squared Pearson correlation coefficient r :

$$r^2 = \frac{D^2}{p_A p_B p_a p_b} \tag{1.8}$$

A D' value of 0 indicates complete linkage equilibrium, which implies frequent recombination between the two markers and statistical independence under assumptions of Hardy-Weinberg equilibrium (random mating, no selection, no mutation, no in or out migration and constant allele frequencies). A $D' = 1$ corresponds to complete LD, indicating no recombination between the two markers within the population. High r^2 values, statistical measure of correlation, indicate that two markers (SNPs) convey similar information. So, only one of the two SNPs needs to be genotyped to capture the allele variation. SNPs that are selected specifically to capture the variation at nearby sites in the genome are called tag SNPs because alleles for these SNPs tag the surrounding stretch of LD. Patterns of LD are population specific.

One often forgotten issue associated with LD measures is that current technology does not allow direct measurement of haplotype frequencies from a sample because each SNP is genotyped independently and the phase or chromosome of origin for each allele is unknown.

We now provide an overview to show how LD between a marker and a DSL will induce asociation between the phenotype and the marker. Consider a case-control study with equal numbers of cases and controls. Let $P(A|\text{case})$ and $P(A|\text{control})$ be the frequency of the disease allele A among the cases and controls, respectively, and let a denote the non-disease alleles. A test of association between the DSL and the disease can be framed as no difference in allele frequency among cases and controls (null hypothesis), or :

$$\begin{aligned} H_0 : \Delta_A &= 0 && (1.9) \\ \text{where}\, \Delta_A &= P(A|\text{case}) - P(A|\text{control}) && (1.10) \end{aligned}$$

Actually, we do not observe the disease locus, but instead a marker with alleles B and b. Then, defining Δ_B as :

$$\Delta_B \quad = \quad P(B|\text{case}) - P(B|\text{control}) \tag{1.11}$$

and assuming that $p(\text{disease})$ does not depend on the marker genotype given the genotype at the DSL, we have [16] :

$$\Delta_B \quad = \quad \Delta_A \cdot (P(B|A) - P(B|a)) \tag{1.12}$$

Note that in the absence of LD, the alleles at the DSL and the marker are independent,

$$P(B|A) \quad = \quad P(B|a) = P(B) \tag{1.13}$$

hence $\Delta_B = 0$. Thus there will be no association between disease and a marker, unless the marker allele is associated with the disease allele. It is shown in [3] that

$$P(B|A) - P(B|a) \quad = \quad \sqrt{p_B p_b} \cdot r \tag{1.14}$$

where r is the allelic correlation between the two loci, and p_B and p_b denote allele frequencies at the marker. Hence we have

$$\Delta_B \quad = \quad \Delta_A \cdot \sqrt{p_B p_b} \cdot r \tag{1.15}$$

which implies that $\Delta_B^2 < \Delta_A^2$. The effect on power of the test depend on allele frequencies at the two loci and on the correlation between the two loci (LD). To achieve (approximately) the same power at the marker locus as would be achieved at the DSL if we knew it, the sample size must be increased by a factor of $1/r^2$ [16].

How far the range of useful LD (meaning the signal from DSL to marker is still strong enough for detection) extends in terms of physical distance in base pairs ? They range between 50 kbp and 300 kbp. The HapMap and 1000 Genome Projects have shown that the relationship between distance and LD is not a smooth one.

Figure 1.4 output from Haploview illustrates the local LD structure of SNPs for the same gene in two different populations. The HBB gene on chromosome 11 (chr11 :5246535-5248462 bp) encodes the β-globin chain for haemoglobin. There are 5 exons and 4 introns. The Yoruba in Ibadan (YRI, Nigeria) and England and Scotland in Great Britain (GBR, UK) populations from the 1000 Genome Project are compared with respect to the LD pattern of the polymorphism related to this gene. The data were downloaded from Ensembl 1000 Genome browser GRCh38 and visualized with Haploview 4.1. There are a dozen SNPs in this 2 kbp region of the human genome. Note that the rs334 variant is present in the YRI population but not in the GBR population. Remember that the rs334 variant is known to cause a substitution from Glutamic acid to Valine aminoacid residue at position 7 in the β-chain HBB protein which is the cause of sickle cell anemia, but can be advantageous to individuals exposed to Malaria. The HBB is a reversed-stranded gene and the majority allele on the forward strand is T at this locus. The rs334 minor allele variant is a A on this SNP locus. Comparing the LD local patterns of the two populations first shows that the polymorphism is higher in the African population than in the British one (there are 14 SNPs in YRI and 10 SNPs in GBR for this 2 kbp region). Second, the LD (r^2) values are higher for LD in GBR ($\sim$ 0.68) than in YRI ($\sim$ 0.58) : correlation between the same pairs of SNPs are higher in the GBR group than in the YRI group meaning more recombination occurred in the YRI group than in the GBR group because of an older ancestry in founder population for the YRI group. This example should remind us that LD patterns may be different across ethnicity. Furthermore, even if there are similar patterns in LD blocks, the measure of the intensity of correlation in LD blocks can be different across ethnic groups and LD patterns change over time (after a large number of generations). This is one of the reasons to control for population substructure in GWAS studies.

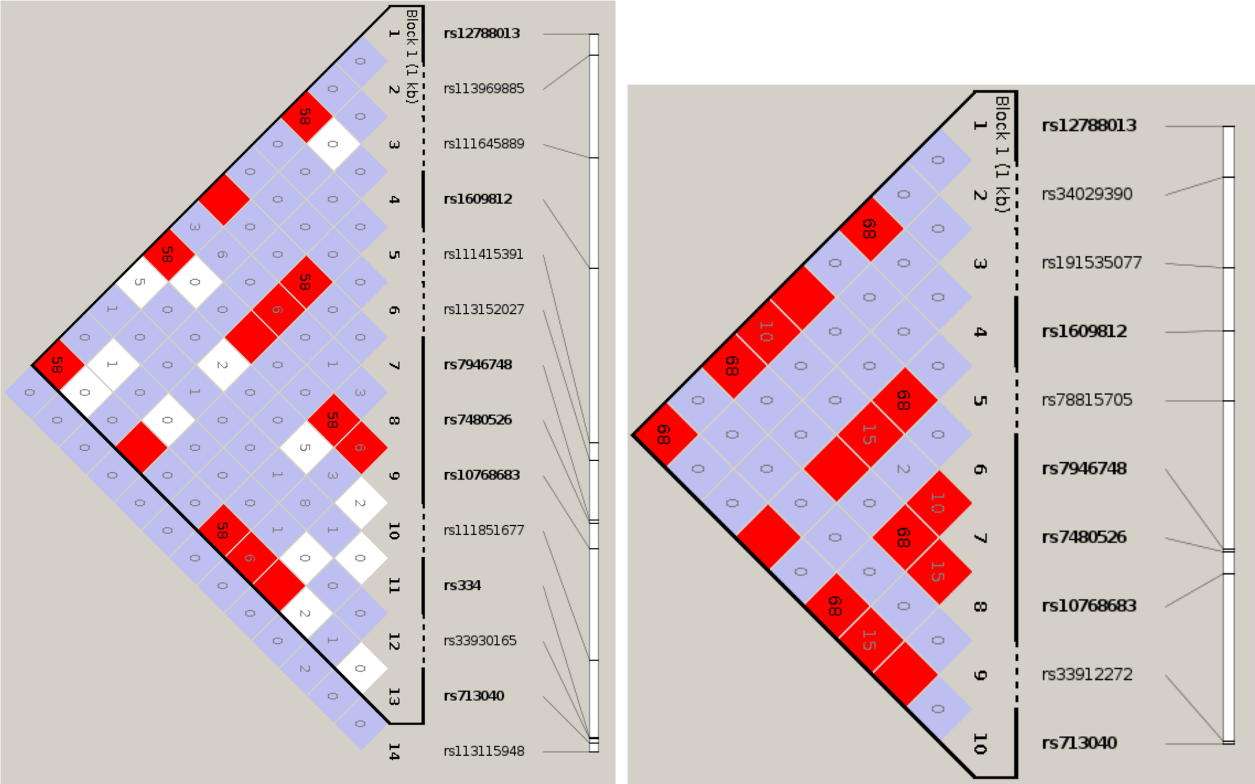


FIGURE 1.4 – Local LD structure in β-globin gene. Red squares showing more intense correlation. The numbers in the squares are the r^2 values between SNPs that correspond to the squares (left panel : YRI Yoruba from Ibadan Nigeria, right panel : GBR England and Scotland from UK).

1.1.8 Epistasis

Epistasis can be loosely defined as the interaction between two or more genes [17]. However, it is not clear what is meant by "interaction" and can vary between biologists and statisticians.

Biological Epistasis

The term epistatic was first coined by William Bateson in 1909 [18]. Epistasis describes a masking effect, whereby a variant or allele at one locus masks the expression of a phenotype at another locus. This definition is analogous to the meaning of dominance from Mendelian genetics, which refers to a situation where an allele masks the expression of other alleles at the same locus. In its broadest sense, epistasis means that the genetic background can determine whether a mutation affects the phenotype of an individual or not. Lehner reviewed different molecular mechanisms of epistasis between genes and within genes [19]. As emphasized by Moore [20], biological epistasis occurs at the level of an individual.

Statistical Epistasis

In the case of quantitative traits, epistasis refers to a deviation from additivity in the effect of alleles at different loci with respect to their contribution to the quantitative phenotype. It describes the situation where the combined effect of two or more loci cannot be predicted from the sum of their individual single-locus effects. This definition of epistasis was first used by Ronald Fisher in 1918 as "epistacy" [21]. The total genetic variance of a quantitative trait can be partitioned into components of variance due to single-locus effects and gene-gene (epistatic) interactions [17].

Moore further insists on the fact that differences in genetical and biological epistasis among individuals in a population give rise to statistical epistasis [20]. It is entirely possible for genetical and biological epistasis to occur in the absence of statistical epistasis. This happens when the DNA sequence variations and biomolecules are the same for every individual sampled from a population. So, genetic and biological variation among individuals is crucial for the statistical detection of epistasis. Statistical epistasis is measured at the level of a population.

An important methodological question is whether statistical evidence of epistasis at the population level can be used to infer biological or genetical epistasis in an individual. Conversely, does biological evidence of epistasis imply that statistical evidence will be found ?

1.1.9 A detailed illustration of epistasis : the color coat of Labrador golden retriever dogs

A classical example of a biological epistatic interaction is the coat color in *Canis lupus familiaris* dogs breed Labrador and Golden retrievers. The coat color of Labrador retrievers can be black, brown or yellow (gold) as displayed in Figure 1.5. The coat color is primarily controlled by two different loci : a black/brown bi-allelic locus (B/b) and a bi-allelic gold locus (E/e) called the extension locus. The black allele (B) is dominant to the brown (b). So, dogs that are heterozygous at this locus will preferentially have a black coat color. However, expression at this locus also depends upon the dog's genotype at the gold (extension) locus. Dogs homozygous for the recessive (e) allele at the gold locus have a golden coat color regardless of their genotype at the black/brown locus. It is said that the extension (gold) locus masks or is epistatic to the effect at the black/brown locus. In the classical Mendelian dihybridism model, the expected ratio at F2 of a mating in case of 2 independent loci would be 9/3/3/1. But here, with epistasis, the expected ratio of a mating between two dogs that are heterozygous at both loci is 9/4/3 (black/golden/brown) as described on Figure 1.6. There is primarily 3 phenotypes and not 4 as in the classical independent dihybridism cross.

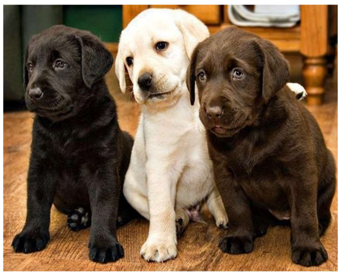
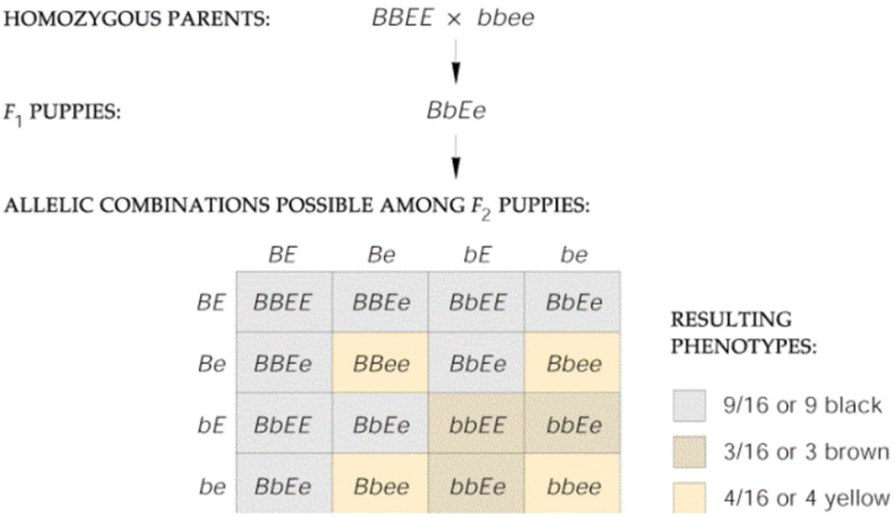


FIGURE 1.5 – Coat color in Labrador retriever puppies.



	BE	Be	bE	be
BE	BBEE	BBEe	BbEE	BbEe
Be	BBEe	BBee	BbEe	Bbee
bE	BbEE	BbEe	bbEE	bbEe
be	BbEe	Bbee	bbEe	bbee

FIGURE 1.6 – Recessive epistasis in dihybridism dogs intercrossing.

Dogs have 39 pairs of chromosomes. It is worth noting that the B/b alleles are basically at a locus on chromosome 11 of the dog, whereas the extension locus (E/e) is on chromosome 5. This an example of an epistatic gene interaction between different chromosomes [1].

It is of pedagogical interest to learn the basics of the coat color biochemistry in dogs as it helps to understand other epistatic effects for other phenotypes in other species. It can give insights for understanding pair wise gene gene interaction underpinning complex diseases.

The coat color of dogs is determined by the relative levels of two pigments : eumelanin, which can either be brown or black and phaeomelanin which is yellow to orange or red. Both pigments are generated by metabolism of tyrosine. These two pigments levels depends on the activity of the enzyme tyrosinase-related protein 1 (TYRP1) whose coding gene is on a locus of chromosome 11 of dogs. This locus has the two alleles (B/b). The B allele is the wild type and dominant. It allows the normal production of black eumelanin. The b allele is recessive : lacking the TYRP1 protein, eumelanin does not undergo the final biochemical conversion, and is milk chocolate brown rather than black. A dog that is homozygous recessive for this variant will show the brown color phenotype. A schematic diagram showing the biochemical pathways producing the pigments is shown on Figure 1.7. In melanocytes, the default pathway produces yellow-red phaeomelanin. Eumelanins are only made if melanocytes receive specific signals from the melanocortin receptor (MC1R) coded by a gene at the E locus situated on chromosome 5 in dogs. As illustrated on Figure 1.8, the MC1R protein is a transmembrane protein located on the plasma membrane of the melanocyte. One of the ligand of the MC1R receptor is the melanocyte stimulating hormone ($\alpha - MSH$). When the hormone $\alpha - MSH$ binds to the MC1R (melanocortin 1 receptor), eumelanin is produced. If the MC1R receptor is defaulted for instance because the protein has not the right configuration due to a premature ending of the translation process, the specific signals are not triggered and eumelanin will not be produced. Only phaeomelanin will be present and the dog will exhibit a golden coat. There are other ligands to the MC1R receptor, like the agouti signaling peptide (ASIP) coded at another locus, but for the sake of simplicity, we do not develop further the agouti signaling effect here.

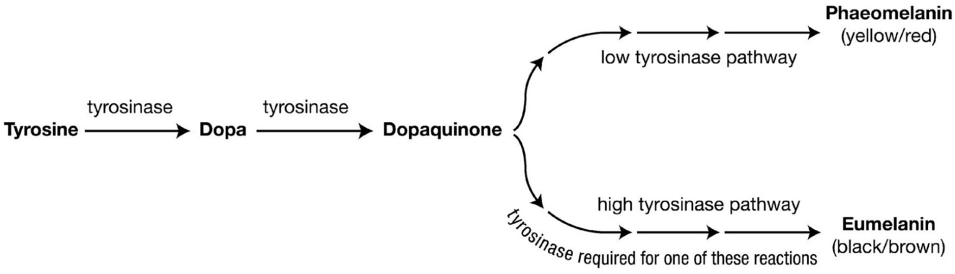


Figure 1.7 – Schematic synthesis pathway for pigments eumelanin and phaeomealin. Credits : Everts *et al.*

From a molecular biology perspective, the (e) allele at the extension locus (E locus) is actually a SNP which results in a stop codon in the gene coding for the melanocortin receptor 1 (MC1R, E locus). This functional SNP identification was elucidated in 2000 by Everts *et al.* at the Faculty of Veterinary Medicine in Utrecht [22] using inverse Polymerase Chain Reaction methods [23]. Comparison of the complete MC1R sequences of a yellow and a black Labrador retriever revealed a single $C \rightarrow T$ mutation at nucleotide position 916 in the yellow dog. This transition changed the codon for

[1] In humans, there is evidence, in ankylosing spondylitis, of an epistatic effect of ERAP1 of chromosome 5 on HLA-B27 of chromosome 6.

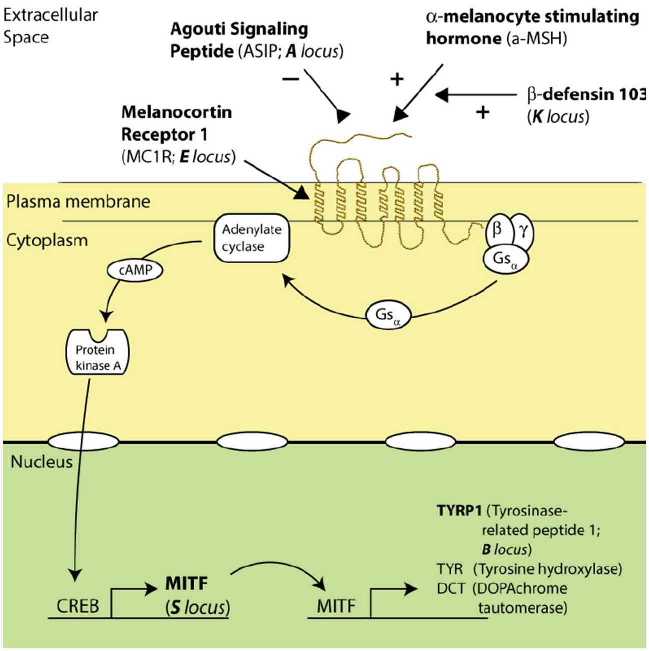


FIGURE 1.8 – Schematic showing how the MC1R regulates melanin synthesis. Credits : Everts *et al.*

arginine at position 305 into a stop codon, resulting in the loss of the evolutionary strongly conserved 12 carboxyterminal amino acid residues. Golden retrievers also appeared to be homozygous for the mutation. The MC1R gene consist of an open reading frame (ORF) of 951 bp and is located in one single exon. Since the 916 $C \rightarrow T$ mutation introduces a premature stop codon in the MC1R gene and thus a truncated protein, a loss of function is the likely effect. Although the deletion of the 12 carboxyterminal amino acids (out of the 317 for the complete functional protein) is relatively small, an adverse effect on the function of the MC1R receptor is nevertheless very likely. In many G protein-coupled receptors (of which MC1R is a family member), one or two cysteines near the end of the protein are attached to the membrane. Lacking these cysteines prevent the protein from properly attaching the melanocyte plasma membrane.

1.1.10 Challenging issues generalizing the dogs coat color illustration

The previous illustration of epistasis through the Labrador golden retriever dogs calls for several comments. First, it shows that a pairwise gene interaction is not necessarily a reciprocal interaction. Indeed, the extension locus (E locus) on chromosome 5 for the MC1R receptor masks the expression of the B/b allele on chromosome 11 but the reciprocal is not true. There is no effect of the alleles B/b (B locus) of chromosome 11 on the expression of the E locus on chromosome 5. This is different from what statisticians generally mean with interactions as they consider interaction to be a reciprocal mutual relation.

Second, the rather simple phenotype of coat color in dogs is underpinned by a rather complex pathway involving more than two genes although the phenotype can be explained simply with only two bi-allelic loci.

Third, the interaction appears from a regulation involving a transmembrane receptor protein. This kind of interaction is usual in molecular biology. There are similar high order interaction/regulation effects in humans in sporadic breast cancer as was described by Ritchie *et al* [24]. There are also similar regulatory pathways involving membrane receptors (MHC assembly) in the complex disease ankylosing spondylitis as is discussed in Evans [25].

Finally, let us assume that we had no prior knowledge of the mechanism underlying the coat color in dogs and that we would set up a case (golden individual dogs)/control (black and brown individual dogs) GWAS analysis. We could perhaps find an association of cases to a marker in LD with the true functional mutation of the MC1R gene on chromosome 5 in a single main effect analysis but would most likely not detect a pair wise interaction of the 2 loci involved in coat color, i.e. a SNP1 in the LD block of TYRP1 on chromosome 11 associated jointly with a SNP2 in the LD block of MC1R on chromosome 5. Indeed, golden color coated individuals (case) and black or brown color coated individuals (control) both have any combination of B and b alleles at the TYRP1 locus and probably any or no linkage at all with possible SNP markers. There is no direct association of the B locus to the cases. Although coat color in dogs is not considered a complex trait, this illustrates the challenge of using statistical methods at the population level to infer genetical or biological epistasis and it rises the question of the biological relevance of the discoveries as was already pinpointed earlier by Moore [20].

Finally, in case-control studies the phenotype is correctly or wrongly assumed to be dichotomized (binary trait). It could be that the phenotype would actually harbor a multinomial outcome (trichotomized trait as in the golden retriever coat color example), with a third extra category being blinded from the method of characterisation. The particular experimental setting should carefully examine all possible hidden outcome and investigators should make sure that the outcome they want to infer from genotype information is truly binary in a case-control retrospective study.

1.2 Scope and contribution of the thesis

The discovery of genetic and biological epistasis via statistical methods is a big challenge, especially in the absence of prior hypothesis, as well as the biological relevance of the findings. The reproducibility of experimental settings, statistical methods and tools for epistasis analysis should be warranted. This motivates the need of simulation data to test the performances and relevance of the statistical methods and bioinformatics tools that are used. The genomic data have a special structure which has to be incorporated so that the simulations are as realistic as possible in order to be eventually useful for inference purposes on real life datasets. Genetic variants are often in linkage disequilibrium with one another or with the functional allele of interest - the Disease Susceptibility Locus (DSL). This so called linkage disequilibrium is of paramount importance in the genomic data structure and biostatisticians must not overlook this characteristic of the whole genome data structure.

This master thesis addresses the six following research questions.

1. How do you detect a gene-gene interaction related to a complex trait in a retrospective case-control genome-wide association study ?
2. How the size effect of a gene-gene interaction on a disease trait (or phenotype) in a non-fully penetrant genetic model impacts the sensitivity of the detection ?
3. What is the genetic variability for the case and control population supposed to be homogeneous (and in Hardy-Weinberg equilibrium) ?
4. How linkage disequilibrium impacts the power of the method ?
5. How is the multiple testing problem taken into account ?

6. What is the estimated power of the experimental setting in association studies for a given sample size of cases and controls and for a fixed family wise error rate (FWER) controlling the multiple testing issue ?

In this Master thesis, we will examine the impact of this inherent correlation between genetic markers on the inferences that can be made from the statistical analysis conducted in large scale DNA based gene-gene interaction studies. Assessing quantitatively the impact of linkage disequilibrium on the power (while controlling for the type I error rate and for a given sample size) of a genome wide association study, designed to unravel a pairwise gene interaction, is desirable to secure the reliability of conclusions drawn from such a study.

Different methods exist to detect epistatic effects but, in this thesis, we will focus on Model-based Multifactor Dimensionality Reduction (MB-MDR). The performances of MB-MDR are poorly documented in the literature. It is the core of this thesis to provide a performance analysis of MB-MDR based on simulated data where true causal SNPs are known and hidden in simulated datasets in order to challenge the MB-MDR method and check if the algorithm will pinpoint the hidden interacting SNPs. It is a fundamental pre-requisite pathway for this method to be used in real GWAS dataset applications.

We will construct 1200 simulated datasets which reflects the statistical properties of human genomic data with as much realism as possible, reproducing linkage disequilibrium patterns, haplotype blocks, minor allele frequencies ranges of real complex diseases and epistatic effects which are important for evaluating the performance of the former developed MB-MDR algorithm.

The method will eventually be applied on a real life dataset from the Wellcome Trust Case Control Consortium on ankylosing spondylitis.

Methods

2.1 Multilocus analysis

In this chapter, the algorithms and methods are presented that are used to detect epistasis with MB-MDR. How the simulated datasets were prepared and the workflow to estimate the sensitivity of MB-MDR is presented. As touched upon in the introduction, the "missing heritability" in GWAS is expected to be explained, at least partially, by nonlinear interactive effects of multiple SNPs, namely epistasis. Therefore, genome-wide interaction analysis (GWIA) focusing on epistasis detection is our assigned task now. In this thesis, we focus only on the MB-MDR method to detect SNPs related to epistatic loci of order two (pairwise interactions).

2.2 MB-MDR Model-based Multifactor Dimensionality Reduction

MB-MDR (Model based Multifactor Dimensionality Reduction) has continuously been developed from 2011 onwards [5] [26] and its recent ancestor method – MDR (Multifactor Dimensionality Reduction), was introduced by Ritchie *et al.* in 2001 [24]. Both are methods for reducing the dimensionality of multilocus genotype information to improve the identification of polymorphism (SNPs) combinations associated with disease risk. MBR is nonparametric (no hypothesis about the value of a statistical parameter is made), is model-free (it assumes no particular inheritance genetic model) and is directly applicable to case-control studies. MB-MDR improved MDR in four ways :

1. MB-MDR can deal with binary trait, continuous trait or survival data (censored data).
2. MB-MDR breaks with cross-validation of MDR and dedicates computing time in permutation-based multilocus significance assessments of the appropriate association test dependng on the data at hand (for binary trait, the association test is the χ^2 test of independence).
3. MB-MDR handles covariates adjustments (possibly correcting for main effects and population stratification) in addition to the interaction effect of interest, while controlling type I error and false positives.
4. MB-MDR recently implemented fast multiple testing correction algorithms to control the family-wise error rate (FWER). Since version 4.2.2 of the software (`MBMDR-4.2.2`), the memory usage was made independent of the size of the dataset and the number of permutations and most importantly, the `gammaMAXT` algorithm implementation contributed by Van Lishout *et al.* speeds up the computing time, still controlling the FWER and with similar power as the previous versions [27]. With a 256-core computer cluster, a dataset composed of one million SNPs and 1000 cases and controls can be analysed in less than 24 hours although $\sim 5\,10^{11}$ pairwise combinations are worked out [27].

2.2.1 MB-MDR algorithm description for binary traits

Global epistasis test

A set of n genetic factors is selected and their possible multi-factor classes or cells are represented in a n dimensional space. For $n = 2$ bi-allelelic loci, there are 9 possible genotype cells in a 3×3 table. The ratio of the number of cases to the number of controls is estimated in each cell and the cell is labelled as either high risk (H) if the case-control reaches or exceeds a predetermined threshold (for example ≥ 1, in MDR) and low risk (L) if it does not reach this threshold. This reduces the original $n-$dimensional model to a one dimensional model, i.e. one variable with two classes : high risk and low risk. The procedure is repeated for each possible n-factor combination and the combinations that maximizes the case-control ratio of the high risk group are ranked. In MB-MDR, a statistical χ^2 test is carried out to compare the proportion of cases and controls in each cell and a third class is considered if there are less than 10 observations in a cell or if the null hypothesis of identity of cases and controls in the cell cannot be rejected. In this case, in MB-MDR, the cell is labelled as null (0) instead of H or L as illustrated in Figure 2.9 reproduced from Van Lishout PhD thesis [26]. The next step is to determine the t_j- statistic for the particular pair of SNPs $(SNP_{rj} \times SNP_{lj})$.

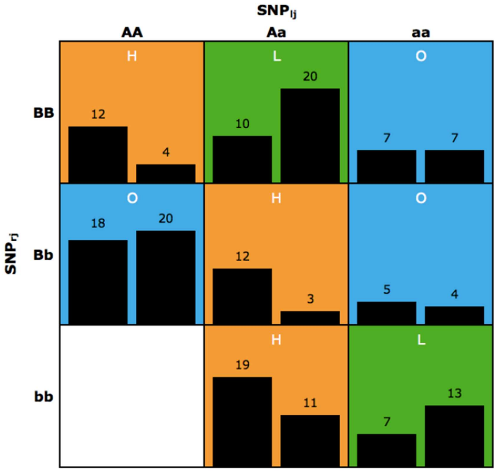


FIGURE 2.9 – MB-MDR 3×3 table with risk classification (H/L/O) according to the relative proportion of cases (left bars in boxes) and controls (right bars in boxes). r and l indices refer to any SNP taken from the M total number of SNPs and j refers to the j^{th} best ranked pair of SNPs. Credits : Van Lishout PhD thesis [26].

The 3×3 table of figure 2.9 is collapsed in a 2×3 table where all the three classes are summed up by cases and controls to obtain a table like Table 2.3.

TABLE 2.3 – 3×2 table of disease risk classification (High, Low and 0) by cases and controls for a particular pair of SNPs.

	H	L	0
Cases	a	b	c
Controls	d	e	f

The 3×2 table is further collapsed in order to conduct two χ^2 association tests of the disease status with the risk classification for each of the two 2×2 tables to test High risk versus $L + 0$ on the one hand (χ_1^2) and Low risk versus $H+0$ on the other hand (χ_2^2) as tabulated on Tables 2.4 and 2.5. An option of the algorithm also allows to test High risk (H) versus Low (L) on the one hand and Low risk versus High, excluding the 0 labelled cells. We actually used this last option in our data analysis (as seen in `Code 2a-2` in appendix).

Table 2.4 – 2 × 2 table of disease risk classification (High versus Low and 0) by cases and controls for a particular pair of SNPs.

	H	L + 0
Cases	a	b + c
Controls	d	e + f

Table 2.5 – 2 × 2 table of disease risk classification (Low versus High and 0) by cases and controls for a particular pair of SNPs.

	L	H + 0
Cases	b	a + c
Controls	e	d + f

The resulting t_j test statistic is the maximum of the two χ^2 tests :

$$t_j = \max(\chi_1^2, \chi_2^2) \tag{2.16}$$

Epistasis test with correction for main effects

In the previous section, the basic algorithm was presented to perform a global epistasis test. In the full extent of MB-MDR, a targeted epistasis test can be done by adjusting for lower-order effects or by adjusting for population substructure. This is the origin of the model based (MB) component in the name of the MB-MDR method. Correcting for main effects prevent false epistasis effects. Two coding schemes are possible : either the additive model or the codominant model. If A is the major allele and G the minor allele at the locus of interest, the wild type homozygote (AA) is coded 0, the heterozygote (AG) is coded 1 and the mutant homozygote (GG) is coded 2. In the additive model, the allele dosage (number of variant alleles in the genotype) increases the disease risk in an additive way. For the codominant coding scheme for bi-allelic locus, two indicator variables are used : the first indicator variable is set to 1 for the (AG) genotype, 0 otherwise and the second indicator variable is set to 1 for the (GG) genotype and 0 otherwise. The codominant coding for the two indicators is respectively 00 for homozygous wild type (AA), 10 for heterozygous (AG) and 01 for homozygous variant genotype (GG). The sum of the two indicator variables is always 1 when there is one or two minor alleles in the genotype.

For lower order corrections, a model is fit that adjust the outcome by implementing a generalized linear regression as detailed in Van Lishout [26].

MB-MDR output

The output that the MB-MDR algorithm produces is a ranked table like Table 2.6. To compute the p-values, the statistical distribution of the test statistic t_j must be known or at least approximated by the empirical distribution obtained by permutations, considering that under the null, all SNPs

Table 2.6 – Ranked table of epistatic SNPs pairs in MB-MDR.

Pair	MB-MDR statistic	p-value
$SNP_{l1} \times SNP_{r1}$	t_1	p_1
...	...	...
$SNP_{lj} \times SNP_{rj}$	t_j	p_j
...	...	...
$SNP_{ln} \times SNP_{rn}$	t_n	p_n

pairs are independent and could be permuted with any other SNPs pair. Each line in the MB-MDR output can be analysed naively by a single hypothesis test.

Actually, a huge number of tests are carried out and the multiple testing issue has to be addressed. With 500.000, SNPs there are 125 billion pairwise combinations if we examine all pairs exhaustively.

2.2.2 Multiple testing correction procedure in MB-MDR

We aim at testing for the presence of associations between the trait and a huge number of potent multi-locus interactions. Testing multiple hypotheses can result in an inflation of the type I error rate (false positive - rejecting the null given that it is true). Recall that, in practice, if $\alpha = 0.05$ is the type I error for one test, if m *independent* tests are conducted, the probability that no error is made is $(1 - \alpha)^m$. Hence, the probability that at least one hypothesis will be rejected wrongly is $1 - (1 - \alpha)^m$. Suppose we conduct 100 tests, then the probability to reject wrongly at least one test out of the 100 is 0.994 or 0.634 (for $\alpha = 0.05$ or 0.01 respectively). We see that when the number of tests (supposed here to be mutually independent) is huge, it is almost certain that we will reject wrongly a great number of tests and will declare a large number of false positive. In the Bonferroni multiple testing adjustment, α is lowered so as to minimize $1 - (1 - \alpha)^m$. For instance, with 100 tests, α must be set to 0.0005 so that the multiple type I error be less than 0.05.

In GWIS, the number of tests is in the order of million or billion. Of course, because of linkage disequilibrium, the tests are actually not all independent as tag SNPs are correlated to causal variants and are mutually correlated among themselves.

The Bonferroni correction for multiple testing is not fully exact because the independence of tests does not hold and the Bonferroni procedure is too conservative.

We should also keep in mind that minimizing type I error increases type II error (β) : The sensitivity or power ($= 1-\beta$) to reject correctly a null that is really not true decreases if type I error α is set to lower values. So, being too conservative in type I error will reduce the power to detect an effect, given that there is truly an effect.

Generally, there are different methods for adjusting for multiple testing to control two error rates : the famililly-wise error rate (FWER) and the false discovery rate (FDR). Detailed discussions of FWER and FDR can be found in Westfall and Young [28]. These two types of error measures are :

Family-wise error rate : probability of at least one type I error across m hypothesis tests.

False-discovery rate : probability of type I errors among the rejected null hypothesis when m hypothesis tests are conducted.

FDR is less conservative than FWER and we always have $FDR \leq FWER$. For that reason, the MB-MDR algorithm was designed at minimizing FWER or controlling FWER at 5%.

Free step-down resampling method MaxT

The free step-down resampling method (FSDR), as given by Westfall and Young [28], is a resampling based method that offers the advantage to *account for underlying unknown correlation structure among multiple hypotheses.* This comes however with a shortcoming of another strong assumption, called the *subset pivotability*, which may or may not be appropriate in the settings of interest. It must be mentioned that there is a method, called *null unrestricted bootstrap*, proposed by Pollard and Van der Laan in 2004, relaxing the subset pivotability condition [29]. But this algorithm is currently not implemented in MB-MDR. The subset pivotability condition states that the distribution of test statistics for a subset of true null hypotheses is the same regardless of whether just this subset is true or the complete null is true.

Under this approach, the individual threshold for α stays at 0.05 but the p-values are adjusted to account for multiple-testing. Let's denote $\mathcal{M}_0$ the unknown subset of true null hypotheses and m_0 the size of this subset. Under the complete null, all hypothesis are truly null and we have $m_0 = m$. Recall that for 1751 SNPs, the number of pairwise combinations is $m = \binom{1751}{2} = 1.532.125$. For a dataset containing 10^6 SNPs, there are $m \simeq 5 \cdot 10^{11}$ pairs.

The `maxT` algorithm starts by computing the test statistics for all pairs of SNPs and sorting them in decreasing order. The sorted statistics are denoted $t_1 \geq t_2 \geq t_3 \geq \ldots \geq t_m$ and refer to the corresponding pair of SNPs of interest (SNP_{l1}, SNP_{r1}), …, (SNP_{lm}, SNP_{rm}). The highest observed test statistic is t_1 and we would like to estimate the probability of observing a maximum value that is at least as extreme as t_1, just by chance, under the null that no pair of SNP is associated to the trait. Under the complete null, if we permute the affection status among the subjects, recompute the m test statistics, find the maximum denoted t_b for permutation b, conduct B such permutations ($b \in [1, \cdots, B]$), we can get the empirical distribution of t_{max}. Then, we determine what is the probability to have a t statistic equal or more extreme than t_1 from this empirical distribution that we obtained by bootstrapping under the complete null.

In Van Lishout's implementation of maxT [27], the test statistics of all m pairs are calculated but the adjusted p-values of only the n best pairs are computed, i.e the ones with the n lowest adjusted p-values. The Van Lishout's implementation of MaxT is summed up hereafter [27] :

1. Compute the test-statistics for all m pairs, but only store the n highest tests values. The result is a data vector where $T_{0,1} \geq T_{0,2} \geq \ldots \geq T_{0,n}$.
2. Initialize a vector p of size n with 1's.
3. Perform B bootstrap replicates $(i = 1, \ldots, B)$:
   (a) Generate a random permutation of the trait column.
   (b) Compute $T_{i,1}, \ldots, T_{i,n}$ and store them in a Permutation-i vector.
   (c) Compute the maximum M_i of the test-statistics values $T_{i,n+1}, \ldots, T_{i,m}$.
   (d) Replace $T_{i,n}$ by M_i if $T_{i,n} < M_i$.
   (e) Force the monotonicity of the Permutation-i, vector : for $j = n-1, \ldots, 1$ replace $T_{i,j}$ by $T_{i,j+1}$ if $T_{i,j} < T_{i,j+1}$.
   (f) For each $j = 1, \ldots, n$, if $T_{i,j} \geq T_{0,j}$ increment p_j by one.
4. Divide all values of vector p by $B + 1$ to obtain the adjusted p-values vector. Force monotonicity : for $j = 1, \ldots, n-1$, replace p_{j+1} by p_j if $p_{j+1} < p_j$.

Gamma-MAXT

In the previous MaxT algorithm which computes the adjusted p-values correcting for multiple testing, the most penalizing algorithmic complexity arises from the step $3(c)$ which is of the order

$O(B\,m)$. A novel algorithm reducing drastically the computer burden was implemented in MB-MDR-4.2.2 by Van Lishout [27]. In the previous step $3(c)$, we are interested in the distribution of a number of test values overall several SNP-pairs, from which to derive the maximum value M_i. The novel idea was to note that the test values in $[T_{i,n+1}, \ldots, T_{i,m}]$ with $i > 0$, follow a mixture distribution of a shifted gamma distribution and a Dirac distribution at zero (zero test values arise in situations for which the MB-MDR test statistics cannot be computed due to the undecisive "O"-labelled category mentioned before). Instead of searching the maximum in the set $[t_{i,n+1}, \ldots, t_{i,m}]$ directly, it is possible to predict it from the fitted mixed distribution. A sample of size $\sim 10^6$ of non-zero values suffices to predict the maximum of the test values in the above set, as it is a tradeoff between computing time and precision of the prediction. The probability density function to fit is :

$$T_i \in [t_{i,n+1}, \ldots, t_{i,m}] \sim \chi_i \tag{2.17}$$

$$f_{\chi_i}(x) = (1-\pi)\delta(x) + \pi\, g_{\chi_i}(x) I_{[x>0]} \tag{2.18}$$

where χ_i is a random variable returning a value from the set $[t_{i,n+1}, \ldots, t_{i,m}]$, $I_{[x>0]}$ is the characteristic function equals to 1 if x is positive, 0 otherwise and $g_{\chi_i}(x)$ is approximately a shifted gamma distribution. The main goal is predicting a maximum, we are not interested in fitting the distribution of $g_{\chi_i}(x)$ on the entire set of strictly positive values. Fitting the tail of the distribution is sufficient as fully motivated in Van Lishout PhD thesis [26]. The shifted gamma distribution has three parameters to be estimated from the top 10% of strictly observed positive values taken from $[t_{i,n+1}, \ldots, t_{i,m}]$. The outcome of the random variable $\mathcal{Y}_i$ from the set of the strictly positive values has the cumulative distribution function (CDF) :

$$F_{\mathcal{Y}_i}(y) = \frac{\gamma(k, \frac{y-y_0}{\theta})}{\Gamma(k)} \tag{2.19}$$

where γ is the incomplete lower gamma function $\gamma(k, y) = \int_0^y t^{k-1} e^{-t} dt$. The three parameters y_0 (location parameter of the shifted gamma), k (shape parameter) and θ (scale parameter) are estimated from a sample of size 10^6 of the strictly positive values in $[t_{i,n+1}, \ldots, t_{i,m}]$. This allows the estimation of the maximum M_i instead of computing it directly. This estimation of M_i replaces the step $3(c)$ of the Max-T algorithm and provides faster computing performance for MB-MDR.

2.3 Need of benchmark data

The performances of MB-MDR are poorly documented in the literature. It is the core of this thesis to carry out a performance analysis of MB-MDR based on simulated data where true causal SNPS are known and hidden in simulated datasets in order to challenge the MB-MDR method and check if the algorithm will pinpoint the hidden interacting SNPs. It is a fundamental pre-requisite pathway for this method to be used in real GWAS dataset applications.

We will construct simulated datasets which reflects the statistical properties of human genomic data in as much realism as possible, reproducing linkage disequilibrium patterns, haplotype blocks, minor allele frequencies ranges of real complex diseases and epistatic effects which are important for evaluating the performance of the methods.

2.4 Simulation methods and algorithms

This section describes a simulation method which belongs to the forward-time simulation family. There are other simulation methods families falling into four categories, namely : *coalescent*, *forward-time*, *resampling* and *Markov chain* simulators that are described elsewhere [30]. We want to simulate genetic samples with realistic patterns of linkage disequilibrium and haplotype blocks. To retain the complex genetic structure of human populations, the forward-time simulation we implemented has four steps :

Step 1 First, we chose an initial population of selected markers from a real sample. We selected two segments of two chromosomes from the 91 unrelated individuals of the GBR population of HapMap 3. The data were downloaded from the Ensembl repository of the 1000 Genome project. A unique HapMap subpopulation was chosen to guard from population substructure or stratification issues. The GBR population of HapMap3 was chosen because we will later discuss the ankylosing spondylitis dataset from the WTCCC2 case-control study that was mainly composed of individuals with British ancestry. The GRCh37.p13 assembly was taken and the genotyped data were extracted for the two arbitrarily chosen following segments with their starting and ending physical positions on the human genome :

- (chr 7 :110200000-110450000) which spans a 250 kbps region with 964 markers (SNPs) at an average marker distance of 260 bps.
- (chr 8 :91525000-91775000) which spans a 250 kbps region with 787 markers (SNPS) at an average marker distance of 318 bps.

One SNP (rs28568272 on chr8 at locus position 91652958) had to be removed because it was not bi-allelic. We restrict ourselves only to bi-allelic markers for reasons of simplicity and compatibility with PLINK software and with the analysis methods. A total of 1751 bi-allelic markers with a real LD pattern from this single homogeneous population of 91 individuals from GBR ancestry (England and Scotland) is now the *prepared founder population.* The LD pattern of these two juxtaposed DNA segments is displayed on Figure B.1 in the appendix. The LD pattern shows interesting features of separate LD blocks of different sizes and LD intensities. We will hide 2 Disease Susceptibility Loci (DSL1 and DSL2) in 4 different configurations into these two selected segments, as will be explained soon.

Step 2 The algorithm then *evolves* this population *forward in time*, subject to possible mutations, recombinations, natural selection forces and population expansion. The process uses a trajectory simulation method to control the frequency of the disease predisposing alleles (DPA of the DSL).

Step 3 This step involves a postprocessing *rejection-sampling algorithm* useful to simulate case-control samples.

Step 4 The dataset outputs from the previous steps must be formatted appropriately for further visualization with `Haploview` or analysis with other softwares like `PLINK` and `MB-MDR`, e.g. convenient `.PED` and `.MAP` file format are prepared.

We repeat the first three steps until we have a number of datasets, 100 in this study, that fulfill our criteria such as the number of cases-control from a population of cases with given chosen disease prevalence, specified disease predisposing alleles frequency and user provided penetrance functions for a binary trait given the genotypes. Multiloci (pairwise) epistasis is also implemented.

A detailed description of the forward-time simulation implementation is given in Peng [31] and is practically carried out with Python scripts from the `simuPOP` simulation environment developed by Peng *et al.* [32] from 2004 onwards and used by an increasing number of users for population genetics studies. The `simuPOP` Python scripting environment is briefly described in the appendix.

The forward-time simulation algorithm was used in a pioneering work conducted by Grady *et al.* in 2011 [33] in producing datasets for a simulation study to test the sensitivity of MDR according to different levels of LD. In their work, Grady *et al.* did not use real LD patterns from former HapMap projects. The HapMap3 data were not yet available in 2011. They simulated their "own" LD patterns instead. The distribution of LD levels in the HapMap data is more complex than in the *in silico*-made data. They used a software called `genomeSIMLA` that is not supported anymore and fails to compile with current versions of C++ compilers. Other softwares exist which implement real LD pattern samples, like `Hapgen2`, but cannot implement directly epistatic interaction between multi-loci. A R package must be used in complement to `Hapgen2`. There is also `epiSIM` (epistasis simulator with a Markov chain) enabling implementation of epistasis model but, again, independent of real outside datasource with real LD patterns [30].

We wrote a Python script and used the existing classes, attributes, versatile and flexible operators and methods of the `simuPOP` environment to produce all the simulated datasets. The relevant codes (code 1a to 1f) are provided in appendix B.2. In the following subsections, we describe the implemented algorithms to serve our objective of generating the simulated datasets (case-control population) starting from the downloaded initial founder population with realistic patterns of linkage disequilibrium.

2.4.1 Generating simulated datasets of case-control populations with realistic pattern of linkage disequilibrium

Demographic model

The initial chosen founding population is small (here 91 unrelated individuals), is considered isolated and belongs to a single homogeneous subpopulation (extracted from HapMap3 GBR subpopulation) before expansion of a typical human population (around 10.000-15.000 years of 500-750 generations if we assume 20 years per generation). We will expand this population linearly to a larger population by adding the same number of individuals each year, possibly subject to mutation, recombination and natural selection. The expanded population size to be reached has been fixed to 10.000 individuals.

Evolving the founder population

During this evolution-expansion, all SNP markers could be mutated according to a symmetric bi-allelic mutation model with a mutation rate of 10^{-8} per base pair per generation. At each generation, parents are chosen at random (random mating) and pass their genotypes to offspring according to Mendelian laws. Parental chromosomes can also be recombined according to the fine-scale genetic map estimated from the HapMap dataset before one of the recombinants is passed to an offspring. If a selection model is specified, parents are chosen with probabilities that are proportional to their relative fitness values. In our work, we defaulted the mutation rate to zero to make sure that all alleles stayed bi-allelic. We did not use recombination either, as the Haldane genetic distances of the SNPs were all set to zero. The selection model only affected the two hidden DSL in order to control the final allele frequency of the last generation of the expanded population. Hence, the last generation of the expanded population can be considered in Hardy-Weinberg equilibrium for all the SNPs given that the mating was made completely at random.

Control of disease allele frequency

To simulate a genetic disease, we control the frequencies of the disease predisposing allele (DPA) at the DSL using presimulated allele frequency trajectories. Either a forward-time approach or a backward-time approach can be applied. If it is assumed that a DPA existed before population expansion, we simulate the frequency of DPA forward in time until it reaches the present generation. The simulation starts from the frequency of DPA in the initial population and is restarted if the allele frequency at the present generation falls out of the desired range. The simulated trajectory forward in time over 500 generations is displayed at Figure 2.10 for the 2 loci DSL 1 and DSL 2A that we choose as functional SNPs to hide in the simulated datasets for the first configuration of LD block positions (setting A).

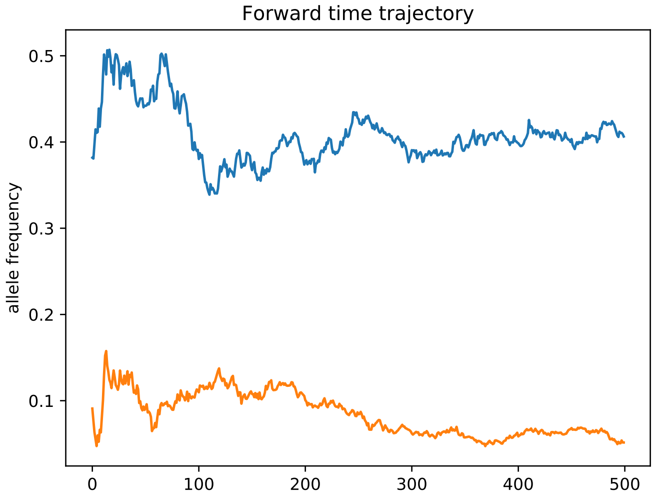


FIGURE 2.10 – Simulated forward-time trajectory of allele frequency over 500 generations. The blue line is the trajectory for DSL 2A moving from 0.42 to 0.40 in 500 generations. The orange line is the trajectory for DSL 1 moving from 0.088 to 0.05. The allele frequencies are for minor alleles at both loci.

If the mutant is recent (appeared within the last 500 generations), we can simulate from the frequency of DPA at the current generation backward in time until the allele gets lost. This was not used in our simulation work.

After the allele frequency trajectories of DPA are simulated, we use a special random mating scheme to evolve the population forward in time while following the simulated trajectories at the loci of interest. The allele frequencies of our four selected DSL are tabulated in Table 2.7 in the initial founder population from the 91 unrelated individuals and at the current generation of the 10.000 individuals expanded population. The current frequencies of the 4 alleles were fixed considering a realistic common variant-common complex disease assumption and the latter frequencies are used for calculation of the disease prevalence we want to simulate in the current population with the use of a penetrance table. The penetrance table will be calculated under a particular genetic model that will produce the case-control samples that are drawn from the expanded population conditioned on the genotype. Hence, the right genotype will be statistically associated to the affection status.

Sample generation

For the final postprocessing step of the simulation, there are two approaches according to whether we deal with a common disease with enough affected individuals in the simulated population or with a rare disease or if the requested sample size is large and if it is difficult to draw enough cases in the simulated population.

To simulate a common disease with enough affected individuals in the simulated population, we can draw samples directly from the population after the affection status of each individual has been determined using a penetrance table that yields the probability that an individual is affected with

Table 2.7 – Allele frequencies of DSL in founder and expanded populations. The first allele in each pair is the minor allele.

		Minor allele frequencies p	
		Founder population	Expanded population
		91 individuals	10000 individuals
causal SNP	alleles		
DSL 1 (rs17644404)	A/T	0.09	0.05
DSL 2 A (rs10956767)	C/A	0.42	0.40
DSL 2 B (rs2073640)	T/C	0.33	0.40
DSL 2 C (rs1476427)	T/C	0.35	0.40
DSL 2 D (rs112698197)	T/C	0.19	0.40

a disease according to his or her genotype. The penetrance table is tabulated in Table 2.8, once the genetic model of the disease has been fixed (see next section).

For a rare disease, a rejection-sampling algorithm can be used to draw case-control samples : we choose parents from the simulated population and produce offspring repeatedly, apply the penetrance function to determine the affection status of each offspring and continue until enough samples are collected with the desired number of cases and controls.

The previous steps conserved the LD patterns as can be seen by comparing Figure B.3 in the appendix after population expansion from the founder population and case-control sample drawing. The LD pattern displayed with Haploview on Figure B.3 for one of the the 1000 cases-1000 controls sample is similar to the LD pattern of the initial founder population.

2.4.2 Genetic disease with epistatic loci : DSL1 and DSL2

A genetic disease with two interacting loci is hidden in all simulated case-control datasets following the methodology described herafter.

Embedding 2 hidden true causal loci in different LD blocks in four configurations with their own real LD pattern

We fixed 4 settings of gene-gene interaction depending on the position of the epistatic locus (DSL 2) relative to the other locus (DSL 1). In setting A, both loci belong to a common LD block on chromosome 8 and are 56 kbps apart. In setting B, the second locus (DSL 2 B) is in a different LD block and 90 kbps appart from DSL 1 but in the middle of a LD block. In setting C, the second locus (DSL 2 C) is still in another LD block, 132 kbps further apart from DSL 1 but DSL 2 C is lying at the edge of an LD block. And finally, in setting D, both loci are on different chromosomes : DSL 2 D is on chromosome 7. All four settings belong to a different LD pattern framework, yet completely compliant to a real human LD pattern.

2.4.3 Changing the effect size : 3 different odds ratios

The model incorporating a gene-gene interaction effect is implemented via logistic regression. The model parameters control the effect size of the epistatic effect of DSL 1 on DSL 2 in addition of possible main effects of either of the two loci. Let Y be the binary outcome indicating the disease status (affected or unaffected) of an individual drawn from the current generation of the expanded population. This outcome is a Bernoulli random variable and if π denotes the probability for an individual to be affected, the model writes :

$$Y \sim \text{Bernoulli}(\pi) \tag{2.20}$$

$$\begin{aligned} \pi &= \Pr(Y = 1 \,|\, g_1, g_2) && (2.21) \\ logit(\pi) &= \beta_0 + \beta_1 \cdot g_1 + \beta_2 \cdot g_2 + \beta_3 \cdot g_1 \cdot g_2 && (2.22) \end{aligned}$$

The β_3 term accounts for departure from additive main effects and measures the intensity of the interaction term. This last term implements epistasis in the model. Depending on the genotype coding scheme, a recessive epistasis or a codominant or an additive or a multiplicative allele dosing epistasis can be modeled.

In ankylosing spondylitis, `HLA-B*27` (DSL 1) is epistatic recessive on `ERAP1` (DSL 2) and both loci are bi-allelic causal SNPs. It has been shown that the major allele dosage of DSL 2 is protective of `HLA-B*27` positive subjects [25]. The odds ratio for being affected is $2.5 - 3$ times lower for homozygous major allele subjects on DSL 2 (`ERAP1`) than for homozygous minor allele on DSL 2 but only for `HLA-B*27` positive subjects. In our simulation datasets, we mimick this fact inspired from ankylosing spondylitis but with a fine tuning on the effect size of this epistatic interaction.

In all our simulation datasets, we implemented the worst case scenario of a purely epistatic effect, i.e. without main effects of the two contributing functional loci ($\beta_1 = 0$, $\beta_2 = 0$ in equation (2.22)). Such a scenario is never detected in classical GWAS where only single loci analysis are conducted.

Our setting models a recessive epistatic effect of DSL 1 (=rs17644404) on DSL 2 (=rs10956767 in DSL 2A) : the major allele T of DSL 1 only masks the effect of DSL 2 if DSL 1 locus is homozygous TT. Besides, the minor allele A of DSL 2A has a multiplicative effect on the odds ratio of affection status as compared to the baseline which is set for the genotype DSL1/DSL 2A = TT/CC. Each increase in A allele dosage of DSL 2A multiplies the odds of affection status by a factor $\exp(\beta_3) =$ 1.65, 2.12, 2.46 in cases where β_3 are 0.50, 0.75, 0.90 respectively, if and only if there is at least one copy of allele A on DSL 1 locus. To implement this model setting, the indicator variables $g1$ and $g2$ of equation (2.22) must obey the following rules :

indicator variables :

$$\begin{aligned} g1 &= \begin{cases} 1 & \texttt{if DSL 2A} = \texttt{(CC)} \\ 2 & \texttt{if DSL 2A} = \texttt{(CA)} \\ 3 & \texttt{if DSL 2A} = \texttt{(AA)} \end{cases} \\ g2 &= \begin{cases} 0 & \texttt{if DSL 1} = \texttt{(TT)} \\ 1 & \quad \texttt{otherwise} \end{cases} \end{aligned}$$

We have built these three effect sizes in each of the four LD position configurations, while simultaneously considering the DSL allele frequencies values in the current generation, in a way to mimick a real human disease prevalence similar to the known one of ankylosing spondylitis which is around $p(D) = K = 0.5\%1.0\%(= 0.005 - 0.010)$. This constraints the value of $\beta_0 = -5$ to set approximatively the correct disease prevalence in the global population.

Table 2.8 shows the penetrance table built from the selected logistic regression parameters that are used to tune the epistatic effect size. In Table 2.8, the values corresponding to a pure epistatic pair $DSL1 \times DSL2A$ when $\beta_3 = 0.90$ are tabulated. The three epistatic effect sizes that are implemented in each LD position scenario were provided with the three β_3 values : $\beta_3 \in [0.50,\, 0.75,\, 0.90]$

Four penetrance tables (times three effect sizes), similar to Table 2.8, were built corresponding to the four LD position configurations described on Figures B.1 and B.2 displayed in appendix.

Table 2.8 – Imposed genotype penetrance table and disease prevalence calculation in the general population with allele frequencies under assumption of Hardy-Weinberg equilibrium. In all settings, the minor allele frequency for DSL1 is $p = 0.05$ and for DSL 2 is $p = 0.40$. In black : probabilities of disease given the genotype, values for simulated datasets in setting A (DSL 1 and DSL 2A) with epistasis effect size $\beta_3 = 0.90$ (see text). In blue : odds ratio with major homozygous (TT) as baseline in setting A with epistasis effect size $\beta_3 = 0.90$. Note that the prevalence in the general population with this setting is around 1%.

Genotype		Penetrance of genotype			Marginal penetrance
		AA	Aa	aa	
		$(1-p)^2$	$2p(1-p)$	p^2	
BB	$(1-p)^2$	$p(D \mid G_1)$	$p(D \mid G_2)$	$p(D \mid G_3)$	$M_x(x=1)$
Bb	$2p(1-p)$	$p(D \mid G_4)$	$p(D \mid G_5)$	$p(D \mid G_6)$	$M_x(x=2)$
bb	p^2	$p(D \mid G_7)$	$p(D \mid G_8)$	$p(D \mid G_9)$	$M_x(x=3)$
Marginal penetrance		$M_y(y=1)$	$M_y(y=2)$	$M_y(y=3)$	$p(D) = K$
	DSL 1	AA=TT	Aa=TA	aa=AA	
DSL 2A		0.9025	0.095	0.0025	
BB=AA	0.36	0.0067	0.0911	0.0911	0.015
Bb=CA	0.48	0.0067	0.0392	0.0392	0.010
bb=CC	0.16	0.0067	0.0163	0.0163	0.008
Marginal penetrance		0.0067	0.054	0.054	$\mathbf{p(D) = 0.0113}$
		Odds ratio as compared to double homozygous CC/TT as baseline			
		AA=TT	Aa=TA	aa=AA	
BB=AA		1.00	14.88	14.88	
Bb=CA		1.00	6.05	6.05	
bb=CC		1.00	2.46	2.46	

The three effect sizes that are implemented in all the simulated case-control datasets designed to assess the power of MB-MDR are displayed on Figure 2.11 instantiated for the first LD position scenario, i.e. $DSL1 \times DSL2A$. The odds ratio to be affected by the disease are graphically represented conditioning on the genotype composition of the two simulated causal DSL that are hidden for the construction of the case-control datasets. The interpretation of the epistatic effect is explained again in the caption note of Figure 2.11.

2.4.4 Heritabilities associated to the 3 different effect sizes

As recalled from the introduction, heritability h^2 is the ratio of the genetic variance to the phenotypic variance and expresses the extent to which phenotypes are determined by the genes. Heritability is computed according to the following equation and by using the values arrayed in Table 2.8 for a given effect size. Because our model incorporates pure epistatic interaction between

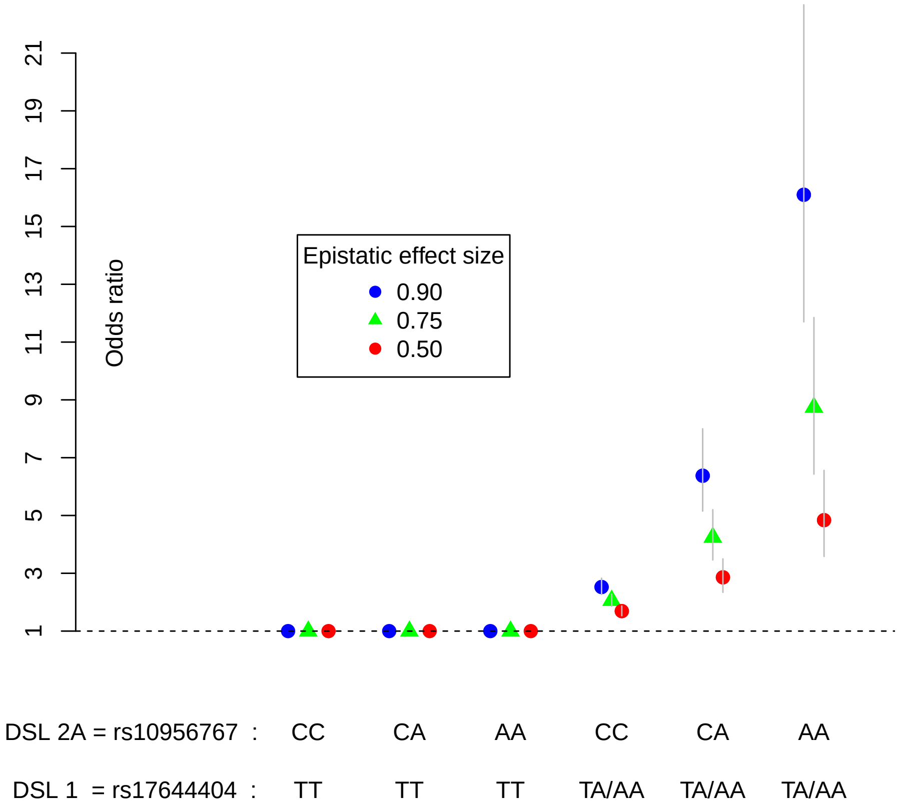


FIGURE 2.11 – Odds ratio effect sizes conditioned on pure epistatic pairs of loci for affection status association in the simulated case-control datasets. Causal effects for DSL 1 and DSL 2A are conditioned on allele A for DSL 1. Note that DSL 2A risk allele A only increases risk for individuals carrying at least one copy of the DSL 1 risk allele (DSL 1 is epistatic to DSL 2A). The low risk CC/TT genotype was set as the baseline ($OR = 1$). The other genotype combinations are coded according to 2 indicator variables $g1$, $g2$ and their product $g1 \times g2$. Odds ratio are obtained by exponentiating the β_3 coefficient of the interaction term from the logistic regression. Error bars : 95% confidence intervals of possible odds ratio that are obtained in different simulated case-control samples.

our chosen pair of functional loci, the heritability is meant broad sense here.

$$h^2 \quad = \quad \frac{\sum_i^9 [p(D|G_i) \cdot p(G_i) - p(D)]^2}{p(D) \ \cdot \ (1 - p(D))} \tag{2.23}$$

In our simulation settings, the tree effect sizes determine the three different penetrance tables like Table 2.8 from which the associated broad sense heritabilities are easily calculated as displayed in Table 2.9. In their pioneering work, Grady *et al.* [33] used higher simulated values for broad sense

TABLE 2.9 – Heritabilities associated to effect sizes for the epistatic interaction in all simulated datasets.

Simulated setting	Interaction β_3	Heritability h^2
Effect size 1	$\beta_3 = 0.90$	$h^2 = 0.083$
Effect size 2	$\beta_3 = 0.75$	$h^2 = 0.071$
Effect size 3	$\beta_3 = 0.50$	$h^2 = 0.059$

heritability ($h^2 = 0.05$, 0.15, 0.25) of their pure epistatic disease model. In Grady *et al.* work, the modeled disease prevalence was not indicated.

2.5 Datasets pre-processing : LD pruning

Before conducting the dataset analysis through the MB-MDR algorithm to detect pairwise SNPs interaction, a pre-processing may or may not be conducted. The purpose of this pre-processing step is to reduce the huge number of pairwise SNP combinations. Different approaches are commonly used in the literature. A first approach, known as prioritization, makes use of prior knowledge and select only SNPs that are supposed to be biologically relevant from prior studies or from prior gene ontology information. `Biofilter` is a software tool developed by Ritchie *et al.* [34] which uses a list of public biological databases to generate pairwise gene-gene interaction models and allows $\sim$ 10-fold reduction of the original marker set without using disease-specific information.

Another approach advocating a complete unbiased investigation only relies on the data at hand and will just remove the redundant tag SNPs. This is where LD pruning comes into play. The correlation between SNPs (LD disequilibrium) makes the tag SNPs redundant. LD pruning will filter out this redundancy. Linkage disequilibrium based SNP pruning (or LD pruning) will generate a (pruned) smaller subset of SNPs that are more independent from one another. The objective is to keep just enough SNPs still tagging the causal SNPs but without all the redundancy of all sets of SNPs that tag the very same causal SNP. A r^2 threshold is arbitrarily fixed to remove some of the SNPs that are correlated. The redundancy reduction gained by LD pruning comes at a price of possibly loosing true causal SNPs.

The LD pruning removes SNPs that are pairwisely correlated within a sliding window of given length (50 contiguous SNPs or SNPS in a contiguous DNA length of 50 kbps for instance). To give a concrete example, a LD pruning threshold of $r^2 \leq 0.75$ at a window width of 50 SNPs with a step of 5 would a) consider a window of 50 SNPs, b) calculate LD between each pair of SNPs within this window, c) remove one of the element of the pair if the LD(r^2) is larger than 0.75, d) shift the window 5 SNPs forward and repeat the procedure until all the original SNP list has been scanned.

In our work, we carried out five levels of pruning to check the LD pruning impact on the sensitivity of the MB-MDR analysis :

- No pruning
- LD pruning at a 0.75 threshold
- LD pruning at a 0.60 threshold
- LD pruning at a 0.50 threshold
- LD pruning at a 0.20 threshold

The window width we chose, after preliminary investigations, is 10 SNPs with a shift step of 2 SNPs.

A side objective of the thesis is to try to recommend a LD pruning threshold that could be used as a general rule or at least to provide a practical range for such a LD pruning threshold, applicable in similar settings for homogeneous sub-populations or genome stretches with similar LD patterns.

2.6 Performance analysis criteria

2.6.1 Sensitivity or power

We are primarily interested in assessing the sensitivity (power) of MB-MDR in detecting the correct interacting causal pair of SNPs in the different LD block position contexts and for the different effect sizes that were hidden in the simulated datasets.

Because of the correlation of markers between them and with the causal SNPs, it is expected that the markers (tag SNPs) are able to predict disease status as well as the functional loci ; thus there is a high probability that any pair of tag-SNPs that are correlated with the functional SNPs will be selected by MB-MDR as significant. This should result in the blurring of the exact pair in a larger set of detected significant pairs composed of tag-SNPs. Each element of a pair of tag-SNP is correlated, i.e. is in LD, with each of the 2 elements of the functional variants. Hence, we will use two different operational sensitivity definitions as performance analysis criteria :

Exact sensitivity : the number of times (number of simulated datasets) out of 100 where the true causal pair of SNPs is detected significant with MB-MDR at an adjusted p-value ≤ 0.05.

Signal sensitivity : the number of times (number of simulated datasets) out of 100 where any of the pairs of tag-SNP is detected significant with MB-MDR at an adjusted p-value ≤ 0.05.

The second definition of sensitivity, i.e. *signal sensitivity*, requires to know the tag-SNPs list of each of the causal SNPs. The tag SNP list also depends on another threshold, the $LD(r^2)$ threshold we fix. The particular results may depend on this threshold and we arbitrarily fix the level of $LD(r^2) = 0.20$ to determine the tag SNPs list for each causal variant. To assess the impact of this arbitrary threshold, we also calculated the signal sensitivity with a $LD(r^2)$ threshold of 0.45 to determine the tag SNPs list. As can be seen from table 2.10, this change in threshold will mainly affect the number of tag-SNPs related to DSL 1 : only 2 SNPs are correlated to the causal locus DSL 1 above $LD(r^2) = 0.45$, while there were 60 of them at $LD(r^2) = 0.20$.

The tag SNP list of each causal DSP that are hidden in the different scenario for the datasets are stored in specific files obtained from the `show-tags` PLINK command. We sum up in Table 2.10 the number of tag-SNP of each variant at different values of $LD(r^2)$.

The algorithm to compute both the exact sensitivity and the signal sensitivity is implemented in a customized Python program that we have written and that is further embedded in a job script to scan the 100 simulated datasets automatically in a particular setting. The specifications, functions and code for this program, called `Sensitivity.py` , are detailed in appendix B.3.

TABLE 2.10 – Tag SNPs number associated to causal variants for different LD(r2) values.

Causal	Number of tag SNP at LD(r^2) value :				
SNP	$r^2 = 0.20$	$r^2 = 0.45$	$r^2 = 0.55$	$r^2 = 0.65$	$r^2 = 0.75$
DSL 1	60	2	2	1	1
DSL 2 A	115	114	114	111	98
DSL 2 B	110	110	109	107	107
DSL 2 C	81	80	80	78	78
DSL 2 D	76	48	31	31	24

2.6.2 Family wise error rate and estimation of the minimal proportion of signal detected SNPs pairs in the ranked MB-MDR output shortlist

The MB-MDR analysis settings were set to control the family wise error rate (FWER) at 5%. Hence we expect that the MB-MDR output file will give 5% of false positive SNPs pairs. Considering that our simulated case-control datasets had 1751 SNPs, the total number of pairwise combinations is :

$$\binom{1751}{2} = 1.532.125 \tag{2.24}$$

With a FWER of 5%, an estimation of the number of potent false positive SNPs pairs is at least $1.532.125 \times 0,05 = 76.606$. By default, the first 1000 best pairs of SNPs are ranked by decreasing test statistics or increasing adjusted p-values. In the case of setting A, with causal variants DSL 1 and DSL 2A, we showed that there were 2 and 114 associated tag SNPs to each of the causal variants respectively, at an $LDr^2 = 0.45$. The number of pairwise combinations of these tag SNPs from each of the two sets of tag SNPs (Table 2.10) for DSL 1 and DSL 2 A is $2 \times 114 = 228$. So, it can be estimated that the expected proportion of true positive signal detection among the false positive is $\frac{228}{76606} = 3\,10^{-3}$, meaning that for the 1000 top ranked list, we would detect around 3 pairs of tag SNPs for the case of setting A by mere chance. With $LDr^2 = 0.20$, the number of pairs would be : $\frac{60 \times 115}{76606} = 0.09$, meaning that we would expect to get 90 pairs of tag SNPs blurred in a set of false positive detected pairs. This shows that even with strict FWER control, the very large number of pairwise combination of SNPs will raise a very large number of false positive detected pairs and it will be difficult to distinguish the true and false positive results.

Simulation Results and Discussions

3.1 Sensitivity results for the simulated data

The results for the signal sensitivities of MB-MDR to detect the simulated two-locus pure epistatic interaction in the different settings, for three implemented effect sizes and for the 5 LD pruning levels are displayed graphically on Figure 3.12 for the subset of tag SNPs that are correlated to the two epistatic causal SNPs with LD $r^2 \geq 0.45$ and on Figure 3.13 for the subset of tag SNPs that are correlated to the two epistatic causal SNPs with LD $r^2 \geq 0.20$.

The exact sensitivities are displayed on the lower panels on both Figures (they do not depend on tag-SNPs conditions). The corresponding tabulated results are given in Table C.1 in appendix C.

3.2 Discussions on the sensitivity results of MB-MDR on the simulated data

The sample size in each setting was 1000 cases and 1000 controls.

The family wise error rate (FWER) adopted by default in the MB-MDR algorithm was 0.05.

The exact sensitivity was computed as the number of times out of the 100 simulated datasets where the true causal pair of SNPs was detected significant with MB-MDR at an adjusted p-value ≤ 0.05. The signal sensitivity was computed as the number of times out of the 100 simulated datasets where any of the pairs of tag-SNPs was detected significant with MB-MDR at an adjusted p-value ≤ 0.05.

For signal sensitivities, two subsets of tag-SNPs were used : the first subset contains all tag-SNPs with a LD $r^2 \geq 0.45$ correlation to the causal loci, the second subset contains all tag-SNPs with a LD $r^2 \geq 0.20$ correlation to the causal loci. In the second subset (LD $r^2 \geq 0.20$), the number of tag-SNPs is larger and it is expected the signal sensitivities will be higher than with the first tag-SNPs subset.

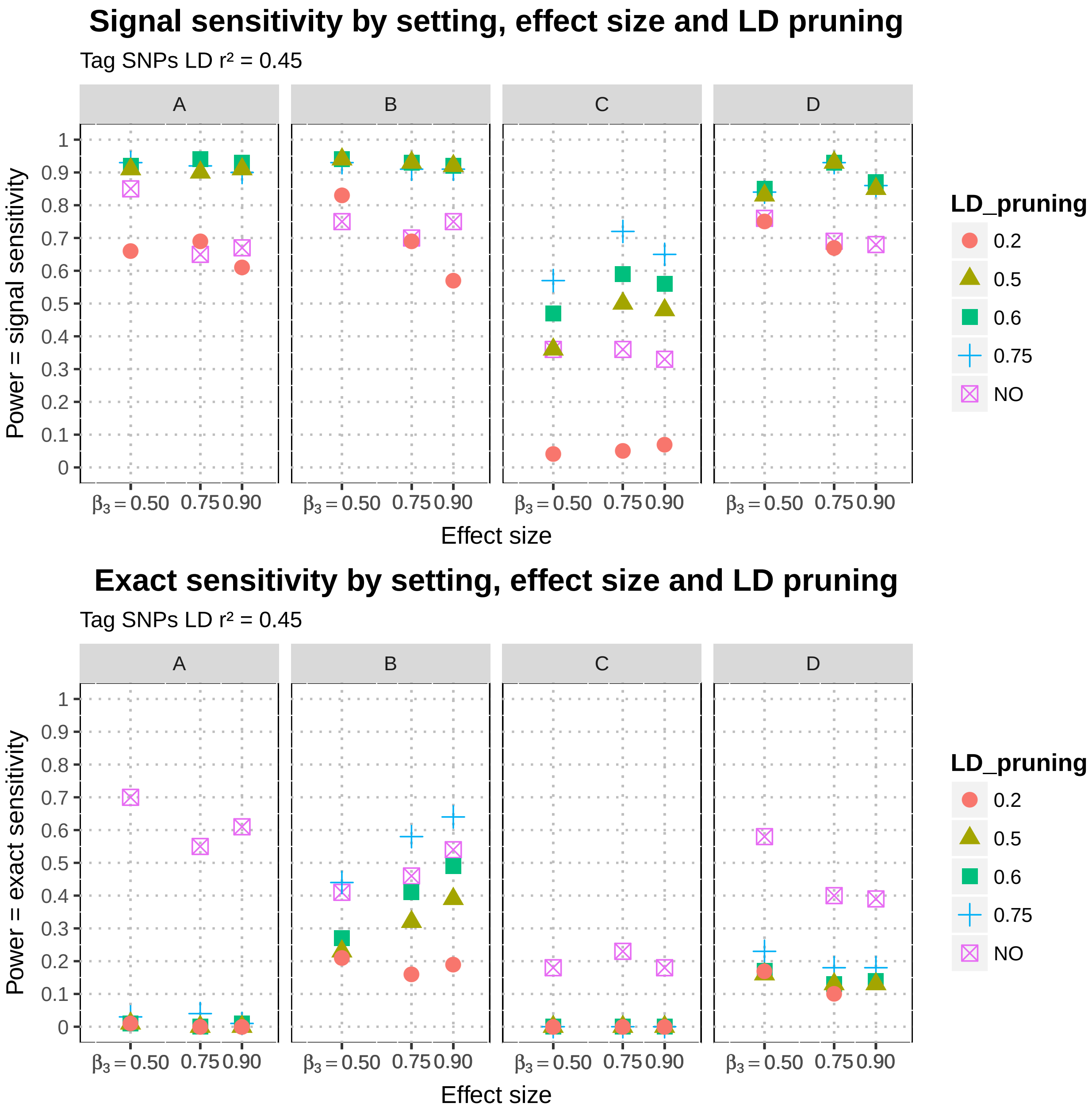


FIGURE 3.12 – Sensitivities of MB-MDR to detect two-loci pure epistatic interaction in 4 settings at three effect sizes and with different LD pruning levels. Signal sensitivities determined with tag-SNP subsets at LD $r^2 \geq 0.45$ with causal SNPs. Signal sensitivities (upper panel) and exact sensitivities (lower panel) are displayed at different LD pruning thresholds (unpruned data or LD pruning at 0.75, 0.60, 0.50 and 0.20).

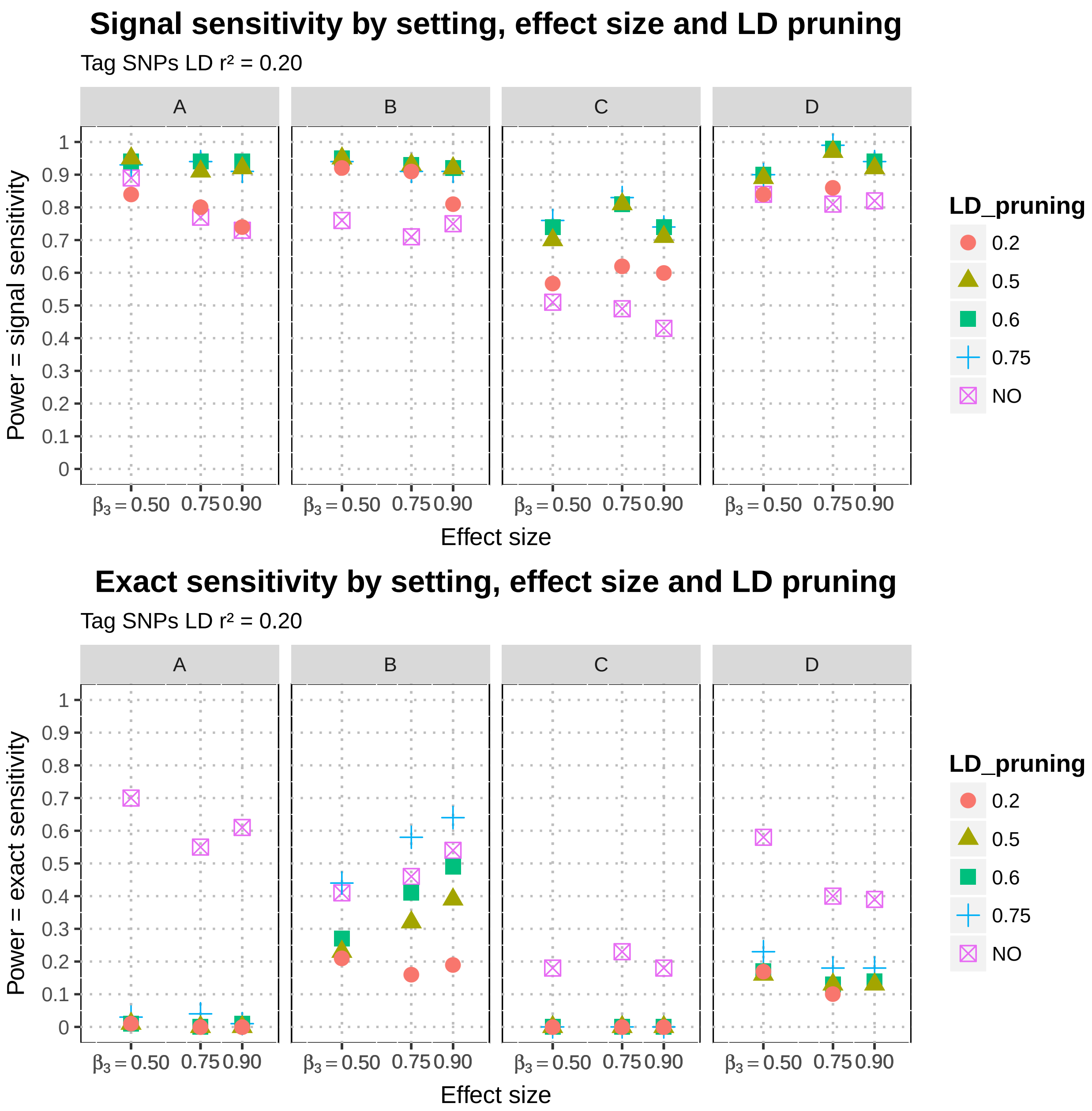


FIGURE 3.13 – Sensitivities of MB-MDR to detect two-loci pure epistatic interaction in 4 settings at three effect sizes and with different LD pruning levels. Signal sensitivities determined with tag-SNP subsets at LD $r^2 \geq 0.20$ with causal SNPs. Signal sensitivities (upper panel) and exact sensitivities (lower panel) are displayed at different LD pruning thresholds (unpruned data or LD pruning at 0.75, 0.60, 0.50 and 0.20).

The sensitivity results call for the following comments and discussions :

1. In all settings, the signal sensitivity is always higher than the exact sensitivity. This is obviously expected and is a benefit of linkage disequilibrium on epistasis detection.
2. The exact sensitivities are in the range 0.2 - 0.7 meaning that the probability to detect an epistatic effect between the two true causal SNPs is generally small. A direct detection will most often prove to be disappointing.
3. The higher power in detecting the true causal SNPs obviously supposes no LD pruning pre-processing of the data.
4. The exact sensitivities appear to be highly dependent on the LD-patterns in which the causal loci are hidden. In setting C (DSL 1 × DSL 2C), the DSL 2C causal SNP is at the edge of an LD block (see Figure B.2 in appendix B) and this setting appears to be more difficult to detect. Even the signal sensitivities in this setting C are lower than for the other LD patterns.
5. The exact sensitivities are largely decreased when LD pruning is applied on the data. The true causal SNP pairs are most often removed from the data by pruning the data prior to the analysis with MB-MDR. It should be noticed however, that the exact sensitivity is not reduced to zero after LD pruning in the case where the LD blocks are on separate chromosomes (setting D, see Figure B.1 in appendix B) or when the causal loci are in the middle of separate large LD blocks (setting B, see Figure B.2).
6. Except for setting B (DSL 1 × DSL 2 B), the exact sensitivities are not conclusively dependent on the effect sizes of the pure epistatic interaction which were fixed in the simulated datasets. In setting B, the power to detect the exact causal epistatic variants increases with the effect size from 0.41 to 0.54 when β_3 moves from 0.50 to 0.90 for the unpruned data. A moderate LD pruning at LD $r^2 = 0.75$ increases further the exact sensitivity from 0.44 to 0.64 when β_3 moves from 0.50 to 0.90.
7. The signal sensitivities are not conclusively dependent on the effect sizes. A small pure epistatic effect is detected indirectly as well as a higher effect. The tag-SNP markers correlated to the true causal loci are helpful for the interaction detection. Again a benefit of linkage disequilibrium.
8. In all settings, LD pruning increases the signal sensitivity as compared to no pruning at all. A proposed explanation is that the SNP redundancy due to LD is partially removed and this removes a higher proportion of false positive blurring the tagging-SNPs detected pairs that are genuinely positive.
9. For all settings (A, B, C and D) and for the signal sensitivity calculated with the largest tag-SNP subset (retaining tag-SNPs correlated to the causal SNPs above LD $r^2 = 0.20$), there is not much difference in signal sensitivity achieved after pruning, whether the LD pruning is done at $r^2 = 0.75$, 0.60 or 0.50. The signal sensitivities were quite good and always larger than 0.70 in setting C (with DSL 2 C at the edge of an LD block) and even larger than 0.90 in settings A, B and D.
10. Pre-processing the data at a very low LD pruning of 0.20, decreases the signal sensitivities as compared to more conservative LD-pruning thresholds in the range $0.50 - 0.75$.
11. Comparing Figure 3.13 with Figure 3.12 shows that the tag-SNPs LD conditions interfere with the LD pruning data pre-processing conditions. The tag-SNP conditions used to build the subset of tag-SNPs for indirect (signal) detection of the two causal loci interaction should be kept below the LD pruning conditions. Figure 3.12 shows that a tag $r^2 \geq 0.45$ and a LD pruning at $r^2 = 0.20$ threshold decreases the signal sensitivities as compared to Figure 3.13, where the tag $r^2 \geq 0.20$ (see the red dots for signal sensitivity panels on both Figures). Obviously, pruning the data at a lower threshold than the one used for the subset of tag-SNPs is not recommended.

Real life data : Ankylosing spondylitis

4.1 Ankylosing spondylitis

Ankylosing spondylitis (AS) is a common form of inflammatory arthritis predominantly affecting the spine and the pelvis that occurs in approximately 5 out of 1.000 adults of European descent [25]. Men are affected 2-3 times more frequently than women. AS starts by inflammation involving the attachments of tendons and ligaments to bone, followed by bone formation, leading to fusion (ankylosing) of the affected joints. AS is considered a dysimmune disease, meaning that it could be triggered by a previous viral or bacterial infection of the digestive tract (e.g. *Shigella sp.*) and that antigen molecular mimicry[2] eventually occurs that misleads the immune response against specific tissues. The current therapeutic treatments for AS include anti-TNF-a, anti-IL-17, anti-IL-12/23 (antagonist agents) and non-steroid-anti-inflammatory drugs [35].

HLAs, human leucocyte antigens, are human major histocompatibility complex (MHC) proteins (see [36]). The human MHC proteins are called human leucocyte antigens (HLA) because they were discovered as antigens of leucocytes that could be identified with specific antibodies. The MHC locus contains two sets of highly polymorphic genes, called the class I and class II MHC genes, on chromosome 6, encoding the class I and class II MHC molecules that display antigenic peptides to T cell (involved in cell mediated immune response). The total number of HLA alleles in the population is estimated to be more than 5000, making the MHC genes the most polymorphic of all genes in mammals.

HLA-B (as well as HLA-A and HLA-C) alleles belong to the MHC class I coding region. Class I MHC molecules are part of complex protein assemblies which specifically bind to a single antigen peptide (made of 5 to 11 amino-acid residues) and which can be recognized by CD-8+ T lymphocytes (CTL cytotoxic T lymphocytes) by a binding mechanism that involves simultaneously two T-cell receptors : one receptor for the conserved chain of the MHC class I molecule (recognized only by CD8-T cells) and one receptor for the specific antigen peptide bound on the groove (cleft) of the MHC class I protein assembly. The MHC molecule biological function is to display the peptides derived from protein antigens for recognition by T cells. The MHC class I molecules along with its antigen peptide and its CD-8 recognition site are expressed on the outer surface of antigen presenting cells (APC). All nucleated cells can express MHC class I proteins ; while only dendritic cells, macrophages and B-cells can express MHC class II proteins.

The assembly of newly synthesized MHC class I molecule with the antigen processed peptide is built in the endoplasmic reticulum (ER) in the APC maturing cell by a sophisticated process fully detailed in Abbas *et al.*, chap. 3, p.64 [36].

Ankylosing spondylitis is strongly associated with a variant called `HLA-B*27`. Only $1 - 5\%$ of `HLA-B*27` positive individuals develop ankylosing spondylitis ; it is clearly not sufficient alone to

[2] Molecular mimicry : cross-reactions between microbial and self-antigens.

cause disease [37]. HLA-B*27 allele is present in 95% of Caucasian patients but only in 50% of African-American patients, showing strong variation with ethnicity [38].

The HLA-B*27 alleles are coding for MHC class I molecular variants in the floor groove (cleft) binding region of the antigen peptide (peptide with 5-11 amino-acid residues derived from protein antigens).

In 2011, Evans *et al.* [25] performed a genome-wide association study, as part of the Wellcome Trust Case Control Consortium 2 (WTCCC2), of 1.788 British and Australian affected individuals (cases) and 4.799 controls of European ancestry.

As part of the results of this and other similar studies, it was found that variants of the gene ERAP1 interact with HLA-B*27 to affect disease susceptibility, one of the first confirmed example of gene-gene interaction seen in humans. For individuals who carry HLA-B*27, their risk of developing ankylosing spondylitis decreases by a factor of four if they are homozygous for the protective variants of ERAP1 which lies on chromosome 5. Endoplasmic reticulum aminopeptidase 1 (ERAP1)'s main function is to trim peptides in the endoplasmic reticulum (ER) to optimal length for binding to MHC class I molecules on antigen-presenting cells for subsequent interaction with CD8+ T cells. This association of ERAP1 to AS has so far uniquely been found in HLA-B*27-positive subjects. These findings that ERAP1 variants influence risk of disease in HLA-B*27 positive, but not negative individuals, strongly support the notion that both these molecules act in the same biological pathway (MHC class I antigen presentation pathway) to affect disease susceptibility. The suggested mechanism to explain the genetic epistasis effect observed with AS is that a ERAP1 loss of function is protecting against HLA-B*27 associated AS [37].

From epidemiological data, it has been shown that susceptibility to AS is affected by several other genes within and outside the MHC : 26 risk loci outside the MHC have been identified by genome-wide association studies so far [37].

4.2 Dataset obtained from the Wellcome Trust Case Control Consortium

4.2.1 Dataset description

The WTCCC2 dataset on ankylosing spondylitis (AS) included a total of 6.587 unrelated individuals (1.788 AS affected cases and 4.799 matched controls) and 487.780 genotyped SNPs available on 22 autosomal chromosomes. Information about sex chromosomes were not available. The SNP genotyping chips used were Illumina 660W-quad chips for the cases and Illumina Human 1.2M-Duo chips for the controls. Although according to the project description, imputation and basic quality controls have been done beforehand, we performed additional basic quality controls to ensure the validity of important assumptions regarding the data (like missingness filtering, minor allele frequency checks, conformance with Hardy-Weinberg equilibrium, and absence of population substructure). The phenotype is the disease status (binary trait : unaffected=1 or affected=2, missing=0 have been filtered out beforehand). No covariate were available except for gender.

Dataset structure

The GIGA approved access to WTCCC2 dataset [39] was composed of two files :

AS_NBS_58C_CH1_to_22_NatureGen.ped : PLINK formatted .PED file with $n = 6.587$ rows (individuals) and $p = 6 + 487.780 \times 2$ columns (SNPs). Note that we have $p \gg n$ featuring very high dimensionality. Column 1 = Family ID, Column 2 = Individual ID, Column 3 and 4 father and mother (not relevant here), Column 5 = Gender (0 = unspecified, 1 = male, 2 = female), Column 6 = Phenotype (0 = missing, 1 = unaffected, 2 = affected), Column 7+8 = genotype pair at SNP1 (one column for each allele : 1 codes for the minor allele, 2 codes for the major allele), and all such genotype pairs for all the 487.780 SNPs.

AS_NBS_58C_CH1_to_22_NatureGen.map : PLINK formatted `.MAP` file with genomic information about the SNPs : column 1 = chromosome, column 2 = SNP id, column 3 = genetic distance (in Morgan units), column 4 = physical base pair position (in bp units).

The only available covariate is `GENDER`. The 2 x 2 table case/control by gender in the setting is tabulated in Table 4.11 :

Table 4.11 – WTCCC2 AS study : case-control by gender

	Gender			
	Males	Females	Unspecified	Total
Cases	976 $(66,2\%)$	498 $(33,8\%)$	314	1.788
Controls	2.433 $(50,7\%)$	2.366 $(49,3\%)$		4.799
Column Total	3.409	2.864	314	6.587

4.3 Statistical analysis methods and strategy

The PLINK and MB-MDR codes implementing the analysis are given in appendix B.5.

4.3.1 Quality control

Missingness filtering

No individuals in the dataset had missing phenotype. But 314 individuals have unspecified gender. We checked for the genotyping success rate of each SNP at a level of 90% : if more than 10% of missingness was observed for a particular SNP, the SNP was removed from the analysis. We also made sure that SNPs with minor allele frequency (MAF) less than 1% were excluded and this was already the case in the provided dataset.

Hardy-Weinberg test and filtering

In addition, we checked for SNPs that failed the Hardy-Weinberg test at $P \leq 5 \cdot 10^{-15}$, departing from the Hardy-Weinberg equilibrium. No SNP were found in Hardy-Weinberg disequilibrium in the dataset.

Population substructure

We examined the data for population substructure by multidimensional scaling (MDS) or principal component analysis (PCA). Both methods provide visual means of identifying population substructure. The aim of PCA is to identify $k\,(k < p)$ linear combinations of the data, called the principal components, that capture as much overall variability as possible, where p is the number of variables (the SNPs in our setting). Multidimensional scaling (MDS) fits the data into a lower dimensional space such that the pairwise distances between individuals are similar to the original distances in the higher dimensional space. MDS is mathematically equivalent to PCA if the distance is defined as Euclidian in MDS. For a given individual, each SNP is represented by the number of variant (minor allele number) for the 2 homologs at the given SNP locus. The similarity between two individuals is defined as the Euclidean distance between the two respective vectors of data. From the similarity matrix, the PCA can be calculated and/or the MDS. PLINK was used for MDS and cases versus controls were plotted with R, as well as females versus males. In PLINK the distance can be measured either as the proportion of IBS (identical by state) shared alleles between any two subjects or as a distance ($1 - IBS =$ proportion of unshared alleles). The following approach has been followed to investigate the potent population substructure :

Genome-wide SNPs subset approach All the SNPs in the dataset were LD-pruned at $r^2 = 0.03$ meaning that all pairs of SNPs in LD with a $r^2 \geq 0.03$ were removed for further analysis. This very strict LD-pruning leave only a subset of SNPs that are almost stochastically independent from one another. From the 487780 SNPs before pruning, only 22861 were left in the subset for the analysis. These left SNPs are lying all across the genome on the 22 autosomes ; that is why it is still considered a genome-wide subset of SNPs. MDS is conducted on this reduced SNPs subset ($p = 22861$). The graphs of the first 2 PC (C1 and C2) calculated from the distance covariance matrix are presented on figure D.1 in appendix D.

The visual inspection of the graph of Figure D.1 for the principal components does not indicate any obvious substructure pattern.

The Genomic inflation factor (GIF) calculated for the genomic control analysis, and the mean chi-squared statistic (that should be 1 under the null that there is no stratification) were $GIF = 1.04585$ and $\chi^2 = 1.06254$ respectively. These values suggest that no very strong stratification exists. So, the absence of population substructure is reasonable in the WTCCC2 dataset on ankylosing splondylitis.

4.3.2 Single Loci Association Analysis

As ankylosing spondylitis affection status has been known to be strongly associated to the `HLA-B*27` genotype since 1974, we carried out allelic association tests for polymorphic nucleotides (SNPs) in the 3 Mbps region of chromosome 6 around the `HLA-B` locus which spans a 3316 bps region between position 31,353,872 and 31,357,187 bp (UCSC GRCh38/h38 web repository) [3]. Chromosome 6 is 171 Mbps in length ; so the investigated region is only 1.75% of the total length of chr 6. The investigated 3 Mbps region was not LD pruned and contains 946 SNPs (SNPs density of about 1 SNP per 3150 bps) in the AS dataset.

The allele test was conducted on each of the 946 SNPs (single locus analysis) of this genomic region and the Benjamini-Hochberg adjustment was used to control for false discovery rate due to multiple testing. The adjusted p-values (i.e. minus their decimal logarithms) are plotted against the SNP bp position in a Manhattan plot displayed on Figure D.2 as seen in appendix D. The 30 most

[3] In the AS dataset, the bps positions of chromosome 6 are shifted 76005 bps to the right of the positions referenced in GRCh38/h38 UCSC human genome repository.

significant genetic markers are highlighted as red triangles. The red and blue marks refer to the SNPs showing linkage disequilibrium at LD $r^2 \geq 0.50$ with the 30 most significants markers and the green bars refer to the markers that are in LD $r^2 \geq 0.20$ with the red labelled most significant markers associated to the affection status.

The single polymorphic nucleotide rs2523608 is in the HLA-B coding region. The rs2523554 is 7kb centromeric to the HLA-B region. The Manhattan plot shows that at least 30 SNPs are most significantly strongly associated to ankylosing spondylitis affection status through linkage disequilibrium with possible unknown causal variant(s) around or in the HLA-B locus. Almost all these 30 most significant SNPs are in physical positions close to the HLA-B locus, at a distance shorter than ±300 kb around the HLA-B locus. The Manhattan plot shows that the genomic region with the SNPs associated with the AS affection status and in linkage disequilibrium together spans around 600 kbps. This region is interestingly centered on the HLA-B locus.

The list of these 30 SNPs is given in Table D.3 in appendix D along with their physical positions, allele test statistics and their FDR Benjamini-Hochberg adjusted p-values for multiple testing.

4.3.3 Multiple loci Association Analysis

HLA-B*27 of chromosome 6 is known to be epistatic on ERAP1 gene on chromosome 5 [25]. The questions we now address here are :

- Does MB-MDR detect any interaction between markers from the HLA-B region of chromosome 6 and markers from chromosome 5 ?
- What is the best set of marker candidates of chr 5 found to be in interaction with the markers significantly associated to AS of chromosome 6 in the region of HLA-B*27 locus ?
- Is this detection or subset list of markers influenced by the LD pruning levels consistent with the results of the simulation results we obtained in the previous chapter ?
- Are the detected markers (if any) in linkage disequilibrium at physical positions close to the ERAP1 locus of chromosome 5 ?

The multiple loci analysis is limited to all the SNPs of chromosome 5 (30 723 SNPs) and to the 3 Mbps portion of chromosome 6 centered on the HLA-B locus (946 SNPs). Those two DNA region were merged in a single file before analysis (31.669 SNPs). The subset of the best 30 SNPs and in LD $r^2 \geq 0.50$ around the HLA-B locus of chromosome 6 that are associated to AS affection status has been determined and is recorded (best tag SNP list for HLA-B causal variant to the affection status on chr 6). A subset of SNPs that are in linkage disequilibrium with the ERAP1 locus (rs30187, bp=96150086) at a LD $r^2 \geq 0.20$ is also recorded and contains 22 SNPs (tag SNP list for ERAP1 on chr 5).

Four levels of LD pruning were carried out on the merged file of chromosome 5 and HLA-B 3Mbps region of chromosome 6 :

- no pruning
- pruning at LD$r^2 = 0.75$
- pruning at LD$r^2 = 0.50$
- pruning at LD$r^2 = 0.20$

The numbers of the remaining SNPs after the different levels of pruning are given in table D.4 in the appendix.

The MB-MDR algorithm was processed on each of the LD pruned file.

The detected SNPs pairs (candidate epistatic pairs) were screened for the belonging of one of their marker to the HLA-B tag SNPs'subset along with (or not) the belonging of the second marker to the ERAP1 tag SNPs'subset. The results are graphically displayed on Figure D.3 in the appendix D.3 for the LD$r^2 = 0.50$ pruned file (similar graphs with other pruning levels were also obtained but not shown here). Figure D.3 displays all the SNPs on chromosome 5 reordered by increasing base pair position. The blue triangles show the SNPs on chromosome 5 that were found as positive candidate markers interacting with the best HLA-B tag SNPs of chromosome 6. The red triangles are the SNPs on chromosome 5 that are found positive candidates marker interacting with the best HLA-B tag SNPs causal variant of chromosome 6 while simultaneously being in LD $r^2 \geq 0.20$ with ERAP1 marker. The position of ERAP1 locus is shown on the Figure. From the 30.723 SNPs of chromosome 5 before pruning, 13.798 SNPs remain after LD-pruning at $r^2 = 0.50$, of which 4.541 are found in interacting pairs by MB-MDR with the 30 best HLA-B tagging markers of chromosome 6 associated to affection status and 3 of them are tagging ERAP1 locus of chromosome 5.

We see that the SNPs obtained as interacting candidates contains 3 markers around the ERAP1 locus : SNPs rs2042381, rs149313, and rs34753. The chromosome 6 paired HLA-B tagged SNPs were rs3868542, rs1841, rs1841 respectively. Interestingly, the highest obtained chi-squared statistic corresponds to the SNP rs2042381 tagging the ERAP1 locus. But this result is drown in a very large number of marker candidates that are uniformly distributed across chromosome 5 and are hypothetically false positive.

Without LD-pruning, two SNPs were found as HLA-B locus interacting candidates tagging the ERAP1 locus : rs152468, rs 469783. The chromosome 6 paired HLA-B tagged SNP was rs4495304. 3.502 hypothetically interacting candidates were also found all across chromosome 5.

At a LD-pruning $r^2 = 0.75$, four SNPs were found as HLA-B locus interacting candidates tagging the ERAP1 locus : rs149313, rs28081, rs469783, rs152468. The chromosome 6 paired HLA-B tagged SNP was rs2233984. 4.519 hypothetically interacting candidates were also found all across chromosome 5.

At a LD-pruning $r^2 = 0.20$, three SNPs were found as HLA-B locus interacting candidates tagging the ERAP1 locus : rs41135, rs28096, rs11748795. The chromosome 6 paired HLA-B tagged SNP was rs38685542. 4.863 hypothetically interacting candidates were also found all across chromosome 5.

There is no way to exclude the risk of false positive results in the detected interaction. This is the drawback of the huge number of pairwise markers combination contributing to a very large number of multiple testing. Controlling the FWER even more tightly, still gives an absolute value for the expected number of false positive which amounts to the thousands.

The questions stated in the beginning of the section are addressed and the summary answers are :

— MB-MDR detect from 3502 to 4813 markers of chromosome 5 that are hypothetically interacting with HLA-B locus tagging SNPs of chromosome 6 best associated to the affection status of ankylosing spondylitis.

— The subset list of markers increases with the level of LD-pruning consistent with the previous chapter results : with no LD-pruning, the number of positive findings were 3.502 and increases to 4.519, 4.541, 4.863 with increasing the LD-pruning level from LD $r^2 = 0.75$, $r^2 = 0.50$ and $r^2 = 0.20$ respectively. But whatever the pruning level, only from 2 to 4 pairs of SNPs are detected that are both tagging the ERAP1 and HLA-B locus which are the true causal interacting pair known from [25]. The detected epistatic pairs are not the same across the LD-pruning levels.

— There is a risk of false positive results in the detected interacting pairs.

Discussions and Conclusions

In this master thesis we studied genetic and biological epistasis via a non-parametric and model free statistical method, MB-MDR (Model based multi-factor dimensionality reduction), applied on case-control genome wide association studies. This method advocates a complete unbiased approach as no prior hypothesis on the molecular biological interaction of epistasis is taken and no prior selection of the genetic loci involved is retained. The only required assumptions were that all loci are bi-allelic and that the interactions are of order 2 (pairwise interaction).

Biological genetic epistasis is the situation in which an allele at a locus masks the effect of another allele at another locus. Biological epistasis occurs at the level of an individual. Statistical epistasis is a pairwise loci association to an affection status measured at a population level in a retrospective case-control study. Linkage disequilibrium (the non random association of an allele at a locus to another allele at another close locus) is confounding gene-gene interaction.

The primary objectives of our study was to investigate the performances of MB-MDR in detecting gene-gene interaction on a genome wide basis and the impact of correlated markers on these performances.

In our study, we built 1200 simulated case-control datasets, with sample size 1000 cases and 1000 controls, from a homogeneous population (without population substructure) in which 2 common variants causal loci were hidden in 4 linkage disequilibrium (LD) block settings, mimicking real human LD patterns, associated to a common disease similar to the real life ankylosing spondylitis disease. Penetrance tables were tuned upfront by a logistic regression model with 3 effect sizes for a pure epistatic interaction between the 2 known hidden loci with chosen minor allele frequencies of 0.40 for DSL2 and 0.05 for DSL1.

It was our main contribution to show from the 1200 simulated case-control datasets that linkage disequilibrium pattern influences both the direct and signal sensitivity of the detection by MB-MDR. The worst case scenario being when one of the causal locus is at the edge of a LD block (setting C). The exact and signal sensitivities of MB-MDR were not found conclusively dependent on the effect sizes in our simulated settings. LD-pruning at a level between LD $r^2 = 0.50 - 0.75$ was found to have a positive effect on the signal sensitivity but obviously not on the exact sensitivity of MB-MDR.

Controlling for FWER at 0.05, MB-MDR analysis of SNP marker pairs in epistasis for both the simulated datasets and the real life ankylosing spondylitis dataset gave a very large number of hypothetically false positive results due to the multiple testing issue arising from the huge number of pairwise combination of marker epistatic candidates.

From a fundamental point of view, the statement by Moore [4] still holds as a conclusion of our work : biological evidence of genetic epistasis does not imply that statistical evidence will be easily found. Although the computational burden had been solved in previous works by efficient algorithms like `gammaMAX` in MB-MDR [27], the main hurdle to be encountered remains the multiple testing issue that is amplified by the huge number of pairwise combination of the markers and giving rise to a very large expected number of false positive results, blurring true epistatic signals.

Appendix A

Softwares

A.1 UNIX/Linux Environment and Multiple Core Resources

It is an advice for any researchers planning on performing many large-scale analysis to look into adopting a Linux environment. In the framework of this Master thesis, I did the conversion from Windows to Linux myself. Unix/Linux environment is a hub for shell scripting and submitting jobs to multiple core resources. Processing 1200 datasets at 4 different LD pruning levels and working out MB-MDR on these datasets or on the ankylosing spondylitis dataset is computer intensive and results in a large computer burden. Most of the UNIX written jobs were submitted on the SEGI multiple core platform. On this platform, 8 nodes were available (4 nodes with a 512 GB RAM/node with 24 cores/node and 4 nodes with 40 cores/node (4GB/core). The allocated maximum CPU time for running the jobs was between 4 hours and 24 hours.

A.2 Haploview 4.1

Haploview is a software package that provides computation of linkage disequilibrium statistics and population haplotype patterns from primary genotype data in a visually appealing and interactive interface.

Haploview was developed by Barrett *et al.* in 2005 [40], is written entirely in Java, which means it is usable on any platform with a Java compiler. URL : http ://www.broad.mit.edu/mpg/haploview

Haploview accepts input in a variety of formats, e.g. .PED and .MAP files interfaced from PLINK. It can also accept datasets from the .vcf format (variant calling format) after conversion to PLINK format with the vcftools package. This allows to download 1000 Genome Project genotype data and visualize them using Haploview. Haploview can only process with bi-allelic variants. Thus, files with multi-allelic variant have to be filtered first with vcftools. Haploview generates marker quality statistics, LD information, haplotype blocks, population haplotype frequency and single marker association statistics, check for conformance with Hardy-Weinberg equilibrium, percentage of individuals successfully genotyped for that marker. The graphical displayed information can be exported to a PNG for use in publications.

A.3 Hapgen2 2.0.2

HAPGEN2 is a free, open-source simulation toolset, based on resampling algorithm of known haplotypes that can produce patterns of linkage desiquilibrium (LD), which mimic those in real data, for simulating multiple disease SNPs on the same chromosome.

HAPGEN2 was developed by Su Z, Marchini J and Donelly P in 2011 [41], is written in C++, can run as a stand alone tool from the command line or via scripting in UNIX/Linux environment. URL : http ://www.stats.ox.ac.uk/ marchini/software/gwas/gwas.html.

HAPGEN2 helps to prototype new methods for statistical analysis and to examine the power of different experimental designs. The traditional approach of simulating a population forwards (like GenomeSimla, see [33]) or backwards in time ignore the large amount of observed genetic data that are now available (1000 Genome Project) which can provide real LD patterns. Given a reference panel of haplotypes, the method produces a sample of haplotypes with patterns of LD similar to those in the reference panel. The user can specify the risk allele and heterozygote and homozygote relative risks. In HAPGEN2 multiple-associated loci can be simulated on the same chromosome. HAPGEN2 can only simulate independent disease SNPs. However, the function simulateDiscretePhenotypes in the R package SimulatePhenotypes can simulate phenotype data under a multiple-SNP interaction disease model. Thus, one can run HAPGEN2 under the null (by setting the effects sizes to 1.0 for all SNPs), then load the simulated data into R and pass it to the function to simulate the phenotype data.

A.4 simuPOP 1.1.8.3

simuPOP is a free, open source, forward-time population genetics simulation environment.

simuPOP was developed by Peng B *et.al* starting in 2005 [32]. URL : http ://simupop.sourceforge.net, distributed under GPL licence. A very complete user guide of simuPOP has been regularly updated from 2005 ownwards. The last version was completed in January 2016.

The core of simuPOP is a scripting language written in Python that provides a large number of objects and functions to manipulate populations, and a mechanism to evolve populations forward in time. A large number (70) of built-in scripts are provided that perform simulations ranging from the implementation of basic population genetics models to generating datasets under complex scenario. The produced datasets can further be processed with any other tool to examine the power of methods of interest for a particular statistical analysis.

simuPOP can incorporate constraints on the final evolved populations such as allele frequencies or penetrance functions for a phenotype given the genotype or prevalence of a given phenotype in the current population. It can implement simulations forward in time of realistic samples incorporating real linkage disequilibrium patterns. Disease susceptibility multi-loci interactions can be included within or between LD blocks.

A.5 PLINK 1.07

PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.

PLINK was developed by Purcell et al. in 2007 [42], is written in C/C++, can run as a stand alone tool from the command line or via scripting in UNIX/Linux environment.
URL : http ://zzz.bwh.harvard.edu/plink/

The focus of PLINK is purely on analysis of genotype/phenotype data and addresses data management, summary statistics for quality control (genotyping rate, Hardy-Weinberg equilibrium test, minor allele frequency calculation), population stratification detection, basic association testing for case/control (allele test, Fisher's exact test, Cochran-Armitage trend test), linkage disequilibrium calculation, haplotype tests, epistasis assessment based on logistic regression model to cite but a few of the toolset.

A.6 MB-MDR 4.4.1 or 4.4.2

`MB-MDR 4.4.1` is a free program designed to perform model-based multi-factor dimensionality reduction in order to detect multiple sets of significant gene-gene and or gene-environment associations in relation to a trait of interest, while efficiently controlling type I error rates. The trait can be expressed either on binary or continuous scale, or as a censored trait. As its `MBR` ancestor algorithm introduced by Ritchie [24] in 2001, `MB-MDR` is non-parametric (no statistical parameter is estimated) and model-free regarding the genetic model for the gene-gene interaction.

`MB-MDR 4.4.1` was conceived by Kristel Van Steen and further developed by Van Lishout *et al.* between 2011 and 2015 [26]. `MB-MDR` is written in C/C++ and can run as stand alone from the command line or via scripting in UNIX/Linux environment.

Binary versions of the program are available at :
URL : http ://bio3.giga.ulg.ac.be/index.php/software/MB-MDR/

From its software ancestry predecessors, `MB-MDR` can adjust for main effects or for possible population substructure (model based prefix in the name of the software), and was improved for multiple testing correction computational burden. Since `MB-MDR 4.2.2`, the latest versions implement the `gammaMAXT` fast multiple-testing correction algorithm which shows power comparable to the MaxT algorithm but requires much less computational resources and time [27].

Appendix B

Simulation workflow and analytical procedures

B.1 Simulation workflow

B.1.1 Founder population LD pattern

The LD patterns from Haploview from the HapMap 3 data in 500 kb region on chromosome 7 and 8 for 91 unrelated GBR (England and Scotland) individuals are plotted on Figure B.1. A total of 1751 bi-allelic SNPs are genotyped and are LD characterised. Figure B.2 displays the LD pattern detailed view from Haploview for HapMap3 data for GBR subpopulation of 91 unrelated individuals and in a 250 kb region spanning from 91.525 kb to 91775 kb of chromosome 8. In both figures, the arrows refer to the true simulated imposed causal variants with epistatic effect that have been hidden as disease causing phenotype with known penetrance functions to control for effect size in the four different simulation settings.

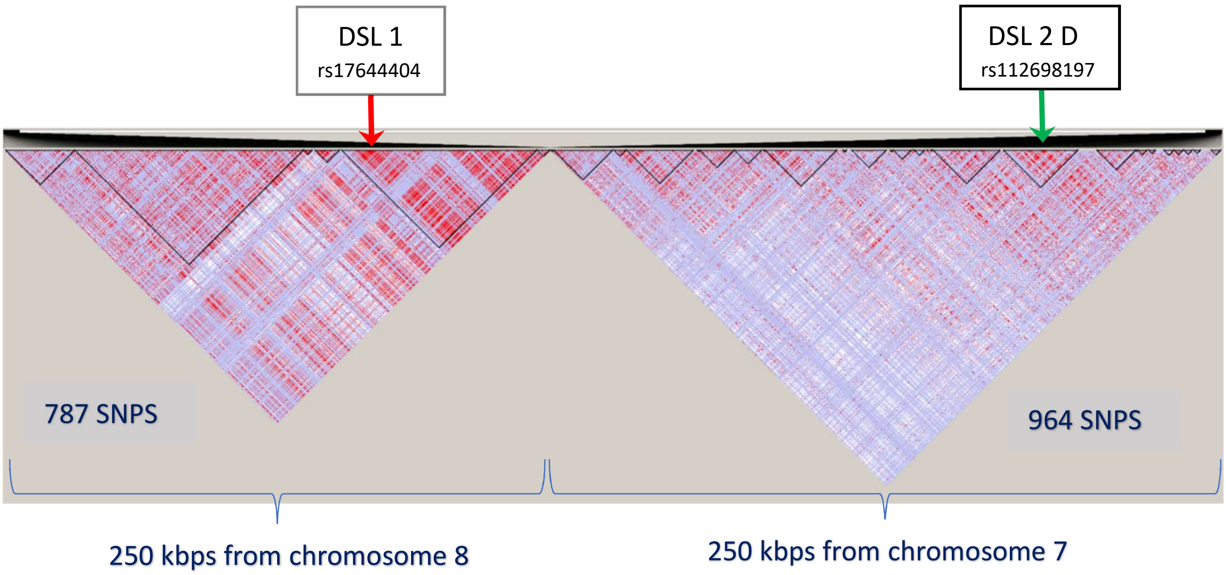


Figure B.1 – LD plots of HapMap3 GBR subpopulation of 91 unrelated individuals. Chromosome 8 on the left with 787 SNPs, chromosome 7 on the right with 964 SNPs. Arrows : two simulated causal DSL in epistatic relation on two different chromosomes.

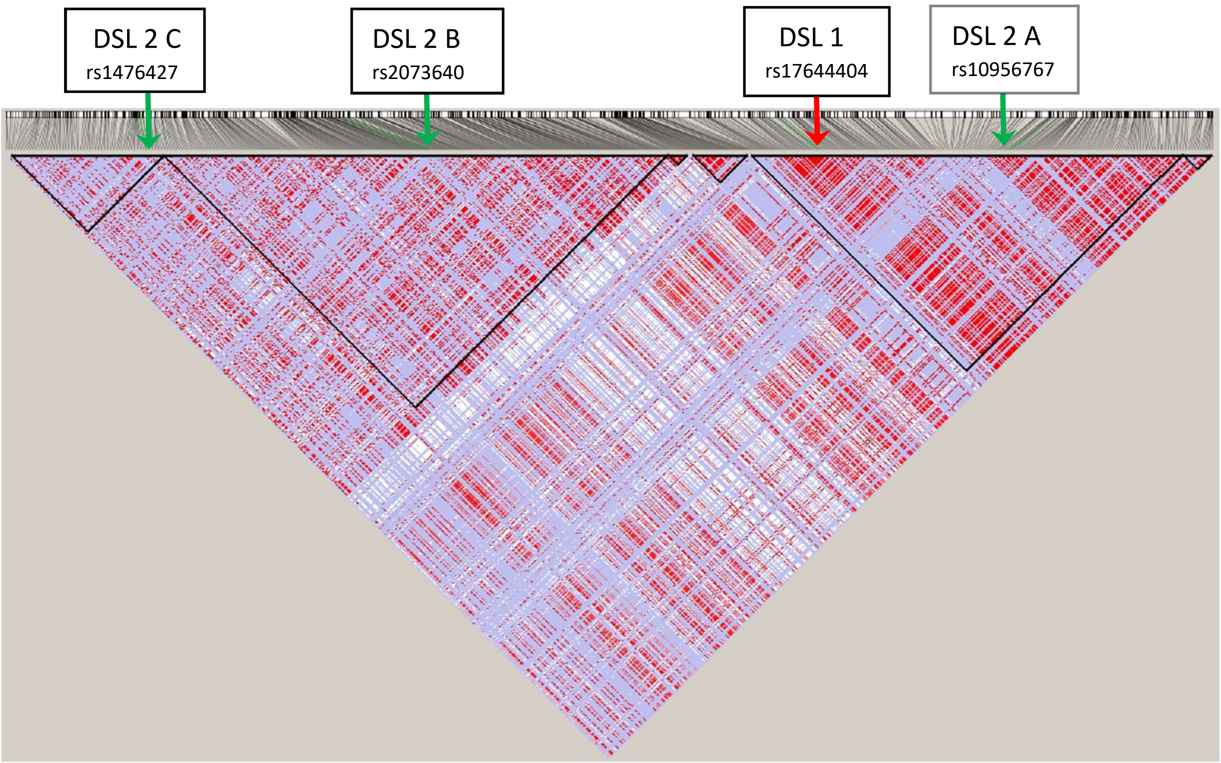


Figure B.2 – LD plots for HapMap3 GBR subpopulation of 91 unrelated individuals. Chromosome 8 with configuration of 3 other pairs of simulated causal DSL in epistatic relation on the same chromosome.

B.1.2 Simulated case-control sample LD pattern

After the expansion of founder population and the drawing the case-control sample following the imposed penetrance table, the LD pattern of a simulated sample with 1000 cases and controls is shown on Figure B.3.

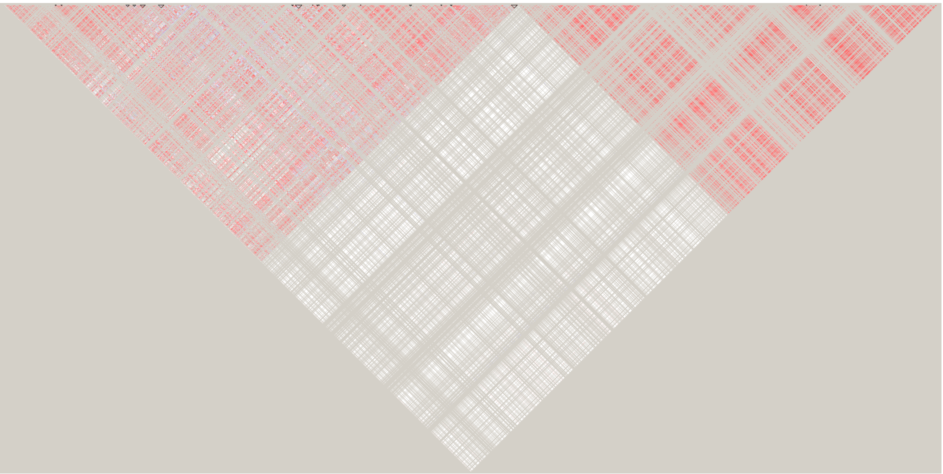


Figure B.3 – LD pattern of a simulated case-control sample drawn from the expanded population with imposed penetrance function. The resulting LD pattern is similar to the LD pattern of the founding population. Note that chromosome 8 is here on the right of chromosome 7.

B.2 Codes for building the simulation datasets

The set of codes for the production of the 1200 simulated datasets is composed of :

1. A bash script named `PyGen` parsing 100 iterated names for the dataset output files to be generated by a `PyGenCC` main program for a given setting and a given effect size.
2. A main program `PyGenCC.py` that requires to instantiate the two DSL 1 and DSL 2 loci for the particular setting, and the effect size for the interaction to be provided in the penetrance table. This main program calls four Python written functions from the simuPOP environment listed hereafter.
3. Two simuPOP functions `LinearExpansion` and `simuGWAS` which expand linearly forward in time the founder population of 91 unrelated individuals to 10.000 individuals by random mating while controlling the allele frequencies of DSL 1 and DSL 2.
4. A simuPOP function `Penetrance` which implements the epistatic interaction between the two loci and provides a probability for the individual to be affected by the disease following a specific logistic regression model with a given effect size in the interaction term.
5. A simuPOP function `genCaseControlSample` implementing the rejection-sampling algorithm which draws from the previously expanded population n=1000 control unaffected individuals and n=1000 case affected individuals by producing offspring from population repeatedly until enough cases and controls are collected.

Generating 100 datasets following this procedure with a single computer core (laptop) takes approximately 300 minutes. Running 12 bash scripts on a multicore platform will produce the 1200 datasets in less than 24 hours while 5 full days are necessary on a single laptop.

The produced simulated datasets have the `.ped` format and can be further processed with PLINK or MB-MDR.

Code # 1a : Produce 100 simulated datasets for a particular setting and effect size `PyGen` (UNIX-Linux bash script) : code instance for setting B, with effect size = 0,75.

```
#!/usr/bin/bash
for numb in {1..100};
do
    python PyGenCC.py datbor2_${numb}.ped;
done
exit 0
```

Code # 1b : Evolving the founder population `simuGWAS` and `LinearExpnasion` (Python - simuPOP functions) :

```
#---------------------------------------------------------------------------
# DEMOGRAPHIC MODEL and GWAS simulation functions :
#---------------------------------------------------------------------------
def linearExpansion(N0, N1, G):
    # Return a linear population expansion demographic function that expands
    # a population from size N0 to N1 linearly in G generations. N0 and N1
    # should be a list of subpopulation sizes.
    step = [float(x - y) / G for x, y in zip(N1, N0)]
    def func(gen):
        if gen == G - 1:
            return N1
        return [int(x + (gen + 1) * y) for x, y in zip(N0, step)]
    return func

def simuGWAS(pop, mutaRate=1e-8, recIntensity=1e-8, migrRate=0.0001,
             expandGen=500, expandSize=[10000], DPL=[], curFreq=[], fitness=[1, 1, 1],
             scale=1, logger=None):
    # Handling scaling if scaling is appropriate...
    mutaRate *= scale
    recIntensity *= scale
    migrRate *= scale
    expandGen = int(expandGen / scale)
    fitness = [1 + (x-1) * scale for x in fitness]
    pop.dvars().scale = scale
    # Demographic function
    demoFunc = linearExpansion(pop.subPopSizes(), expandSize, expandGen)
    # define a trajectory function
    trajFunc = None
    introOps = []
    if len(DPL) > 0:
        sim.stat(pop, alleleFreq=DPL, vars='alleleFreq_sp')
        currentFreq = []
        for sp in range(pop.numSubPop()):
            for loc in pop.lociByNames(DPL):
                # marc added :
                if loc == 1619:
                    allelecode = 1  # allele with minor frequency is C for this locus
                if loc == 1494:
                    allelecode = 0    # allele with minor frequency is A for this locus
                if loc == 1238:
                    allelecode = 1  # allele with minor frequency is C for this locus
                if loc == 1055:
                    allelecode = 3  # allele with minor frequency is T for this locus
                if loc == 709:
                    allelecode = 3  # allele with minor frequency is T for this locus
                # end marc added
                currentFreq.append(pop.dvars(sp).alleleFreq[loc][allelecode])
    print('pop.dvars() = ', pop.dvars())
    print('curFreq = ', curFreq)
    print('currentFreq = ', currentFreq, 'sum of =', sum(currentFreq))
    print('nLoci = ', len(DPL))
    if sum(currentFreq) != 0.:
        endFreq = [(x - min(0.01, x/5.), x + min(0.01, x/5., (1-x)/5.)) for x in curFreq]
        traj = simulateForwardTrajectory(N=demoFunc, beginGen=0, endGen=expandGen,
            beginFreq=currentFreq, endFreq=endFreq, nLoci=len(DPL),
            fitness=fitness, maxAttempts=1500, logger=logger)
        introOps=[]
    else:
        traj=simulateBackwardTrajectory(N=demoFunc, endGen=expandGen, endFreq=curFreq,
             nLoci=len(DPL), fitness=fitness, minMutAge=1, maxMutAge=expandGen,
             logger=logger)
        introOps = traj.mutators(loci=DPL)
    if traj is None:
        raise SystemError('Failed to generate trajectory after 1500 attempts.')
    trajFunc=traj.func()

if pop.numSubPop() > 1:
    pop.addInfoFields('migrate_to')
pop.dvars().scale = scale

# Evolving the founder population
pop.evolve(
    initOps=sim.InitSex(),
    preOps=[
        sim.SNPMutator(u=mutaRate, v=mutaRate),
        sim.IfElse(pop.numSubPop() > 1,
            sim.Migrator(rate=migrSteppingStoneRates(migrRate, pop.numSubPop()))),
        ] + introOps,
    matingScheme = sim.ControlledRandomMating(loci=DPL, alleles=[1, 0],
        freqFunc=trajFunc, ops=sim.Recombinator(intensity=recIntensity),
        subPopSize=demoFunc),
    postOps = [
        sim.Stat(popSize = True, structure=range(pop.totNumLoci())),
        sim.Stat(alleleFreq=[1238, 1494], vars='alleleFreq'),
        sim.PyEval(r"'At the end of generation %3d : allele Freq is: %.3f\n' % (gen,
        alleleFreq[1238][1])", at= [-1]),
        sim.PyEval(r"'and other allele Freq is: %.3f\n' % alleleFreq[1494][0]", at=[-1]),
        sim.IfElse(pop.numSubPop() > 1,
                   sim.PyEval(r"'F_st = %.3f\n' % F_st", step=10), step=10)
    ],
    finalOps = sim.SavePopulation(output='pop2.pop', at=[499]),
        # save the last population at the last generation
    gen = expandGen
)
return pop
```

Code # 1c : Implement the penetrance table with logistic regression for the probability of an individual to be affected conditional on its genotye `Penetrance` (Python - simuPOP) :

```
#-----------------------------------------------------------------
# Function penetrance returns a logistic
# regression model implementing interaction term (epistasis)
#-----------------------------------------------------------------
def penetrance(beta0, beta1, beta2, beta3):
    def func(geno):
        #g1 = geno[0] + geno[1]
        #g2 = geno[2] + geno[3]

        #-----------------------------------------------------------------------------
        # for rs2073640 considered functional DSL2B and playing the role of ERAP1 in
        # ankylosing spondylitis :
        # minor allele = C(1), major = T(3)
        # for rs17644404 considered functional DSL1 and playing the role of HLA-B*27 :
        # minor allele = A(0), major = T(3)
        # with allele coding convention (0, 1, 2, 3) = (A, C, G, T)
        #-----------------------------------------------------------------------------
        # g1 is the dosage of the major allele of the locus DSL2B in our penetrance model
        if int(geno[0] == 1) + int(geno[1] == 1) == 2: # homozygous minor
            g1 = 1
        elif int(geno[0] == 3) + int(geno[1] == 3) == 2: # homozygous major
            g1 = 3
        elif geno[0] != geno[1]: # heterozygous
            g1 = 2
        # g2 is 0 or 1. It is 1 if at least one minor allele is present at the second
        # locus (DSL1). g2 is 0 if the second locus is homozygous major.
        if int(geno[2] == 3) + int(geno[3] == 3) == 2: # homozygous major
            g2 = 0
        else:
            g2 = 1

        # logistic regression model to implement pure epistasis :
        logit = beta0 + beta1 * g1 + beta2 * g2 + beta3*g1*g2
        prob = 1 / (1. + math.exp(-logit))
        print('probability for disease with previous genotype = ', prob)
        return 1 / (1. + math.exp(-logit))   # this gives the probability pi
    return func
```

Code # 1d : Rejection-sampling algorithm to draw cases and controls `genCaseControlSample` (Python - simuPOP) :

```
#---------------------------------------------------------------------
# Generating Case-Control Samples with rejection/sampling algorithm :
#---------------------------------------------------------------------
def genCaseControlSample(pop, nCase, nControl, penetrance):
    # Draw nControl unaffected individuals and nCase affected individuals by
    # producing offspring from pop repeatedly until enough cases and controls are collected.
    # A penetrance operator is needed to assign affection status to each offspring.

    sample = pop.clone()
    sample.setVirtualSplitter(sim.ProductSplitter([
        sim.AffectionSplitter(),
        sim.RangeSplitter([[0, nCase], [nCase, nCase + nControl]])]))
    sample.evolve(
        matingScheme=sim.RandomMating(ops=[
          sim.MendelianGenoTransmitter(),
            penetrance,
            sim.DiscardIf(True, subPops=[(0, 0), (0, 3)])],
            subPopSize=nCase + nControl
        ),
        gen=1
    )
    return sample
```

Code # 1f : Main program `PyGenCC.py` (Python - simuPOP) :

```
# ---------------------------------------------------------------------------
#                 -----  M A I N   ----
#----------------------------------------------------------------------------
#!/usr/bin/python
# This version of the file receives iterated filenames as argument 1,
# run a simuPOP population expansion than generate a case-control
# sample based on a specified penetrance table and finally export
# the sample as formatted .ped file with a name given in argument 1 of the
# calling bash script.

# sys is required for the argument parsing :
import sys
import time
import simuOpt
import matplotlib.pyplot as plt
simuOpt.setOptions(optimized=False, alleleType='long', debug='DBG_WARNING')
import simuPOP as sim
from simuPOP.utils import export
from simuPOP.utils import *
import logging
import random, math
#-------------------------------------------------------------------------------
# Check the number of arguments that were parsed if there are the one expected:
#-------------------------------------------------------------------------------
if len(sys.argv) < 2:
    print('Usage : python ' + sys.argv[0] + ' filename')
    exit()
if len(sys.argv) > 2:
    print('Error : too many arguments added after the python program.')
    exit()
# ---------------------------------------------------------------------------
# Get the name of the file to be exported at the end of the process :
# ---------------------------------------------------------------------------
exportedFileName = sys.argv[1]
# This will only be used at the end of the main program...
t0 = time.clock()
#----------------------------------------------------------------------------
# Read the input data files with their own correct format expected :
#----------------------------------------------------------------------------
# Read the files with their own format expected :
snps_loci_list, chr_ct_list, bps_list = importSNPs('ch78work.map')
pop1 = importData('ch78work.ped', snps_loci_list, chr_ct_list, bps_list)
sim.dump(pop1)
```

```
#---------------------------------------------------------------------------
# Save the uploaded population that was read from the file into a .pop
# format for later upload :
#---------------------------------------------------------------------------
#pop1.save('pop1saved.pop')
# Then, retrieve this population with 'load ' under a new name :
pop_retrieved = sim.loadPopulation('pop1saved.pop')
# -------------------------
# Allele coding convention :
# -------------------------
# The allele coding convention of the last file is : (0, 1, 2, 3) = (A, C, G, T).
# We keep the allele coding convention as : (0, 1, 2, 3) = (A, C, G, T).
# coding convention of this population :
for i in range(4):
    print('allele name = ',i,' : ', pop_retrieved.alleleNames()[i])
print(pop_retrieved.alleleNames())

loc = [1619, 1494, 1238, 1055, 709]
sim.stat(pop_retrieved, alleleFreq=loc, vars='alleleFreq')

#-------------------------------------------------------
# Compute LD statistics for pairs of SNPs of interest :
#-------------------------------------------------------
# Retrieve marker's name from loci position :
bps = []
for loc in [710, 1056, 1239, 1495, 1620]:
    bps.append(bps_list[loc-1])
    print('locus :', loc, 'SNP name = ', snps_loci_list[loc-1], 'bps = ', bps_list[loc-1])

# Compute r2 between pairs of interest :
# Compute D prime between pairs of interest
# simuPOP User's guide p.125 :
sim.stat(pop1, LD=[[1494, 1619], [1238, 1619], [1055, 1619], [709, 1619]],
         vars=['LD', 'LD_prime', 'R2'])
from pprint import pprint
pprint(pop1.vars())
```

```
pop1_rec = pop_retrieved.clone()
# Compute allele frequencies for the markers of interest and their names:
sim.stat(pop1_rec, alleleFreq=[709, 1055, 1238, 1494, 1619])
for loc in [709, 1055, 1238, 1494, 1619]:
    freq = pop1_rec.dvars().alleleFreq[loc]
    freqA = pop1_rec.dvars().alleleFreq[loc][0]
    freqC = pop1_rec.dvars().alleleFreq[loc][1]
    freqG = pop1_rec.dvars().alleleFreq[loc][2]
    freqT = pop1_rec.dvars().alleleFreq[loc][3]

    print('allele freq = ', freq, 'SNP name = ', snps_loci_list[loc])
    print('allele freq for A = ', freqA, 'SNP name = ', snps_loci_list[loc])
    print('allele freq for C = ', freqC, 'SNP name = ', snps_loci_list[loc])
    print('allele freq for G = ', freqG, 'SNP name = ', snps_loci_list[loc])
    print('allele freq for T = ', freqT, 'SNP name = ', snps_loci_list[loc])
# -----------------------------------------------------------------------------
# Expand the founder population (pop1_rec) to a size of 10.000 in 500 generations
# under the linear demographic model.
# Evolve with mutation, recombination, natural selection
# This gives population 'pop2'
#------------------------------------------------------------------------------
# Choose the DSL pair :
# setting A : DSL 1 ='rs17644404', DSL 2 A = 'rs10956767'
# setting B : DSL 1 ='rs17644404', DSL 2 B = 'rs2073640'
# setting C : DSL 1 ='rs17644404', DSL 2 C = 'rs1476427'
# setting D : DSL 1 ='rs17644404', DSL 2 D = 'rs112698197'

DSLA = ['rs10956767', 'rs17644404'] # freq = 0.09
DSLB = ['rs2073640', 'rs17644404'] # freq = 0.33
DSLC = ['rs1476427', 'rs17644404'] # freq = 0.35
DSLD =['rs112698197', 'rs17644404'] # freq 0.19

DSL = ['rs10956767', 'rs17644404', 'rs2073640', 'rs1476427', 'rs112698197']
hapmapAlleleFreq = [0.33, 0.08791]  # list of HapMap allele frequency
# The ending target allele frequencies are : 0.40 and 0.05 respectively :
presentFreq = [0.40, 0.05]
fitnessLIST =[1.0, 1.001, 1.002, 1.0, 0.999, 0.999]

pop_expanded = simuGWAS(pop1_rec, mutaRate=0.0, recIntensity=1e-8, migrRate=0.0001,
                expandGen=500, expandSize=[10000],
                DPL=DSLB, curFreq=presentFreq, fitness=fitnessLIST, scale=1, logger=None)
loc = [1619, 1494, 1238, 1055, 709]
print('pop_expanded : -------------++++++++++++++++++++++++++++++++++++++++++++++++++++++++++')
print('pop_expanded size : ', pop_expanded.popSize())
```

```
#-------------------------------------------------------------------------------
# From the previous population-pop_expanded : Generate a dataset sample of 1000 case
# and control where two loci have been selected as DSL with an epistatic effect.
# The epistasis is implemented via logistic regression (penetrance function
# depending on beta0, beta1, beta2, beta3).
# The result is a dataset sample that you can further process for other analysis.
# The result is affected to ’dsSample’.
# This dataset sample is eventually saved as a .ped file.
#
# ------------------------------------------------------------------------------
sample = genCaseControlSample(pop_expanded, 1000, 1000,
         sim.PyPenetrance(func=penetrance(-5, 0, 0, 0.75),
         loci=[’rs2073640’, ’rs17644404’]))
print(’stat association :’, sim.stat(sample, association=sim.ALL_AVAIL))
#------------------------
# Export to .PED format :
#------------------------
# We use the exported file name variable here in the next line :
export(sample, format=’ped’, output=exportedFileName)
#--------------------------------------------------------------------------------------------
t1 = time.clock()
print("elapsed time ({0}), ({0:.6f}secs)".format(t1-t0))
print("This is for beta 3 = 0.75.")
```

B.3 Jobs and codes for LD pruning with PLINK and for GWIS with MB-MDR

Code # 2a-1 : bash script `Jobs` (Unix/Linux) : processing 100 simulation files (e.g aor1_#inMBmdr.txt)

`aor1` refers to the setting A with effect size 1 (odds ratio with $\beta_3 = 0.90$) and # is a file number $\in [1, \cdots, 100]$.

The following code calls 100 times the code detailed in code 2a-2. Each time with a different input filename.

```
#!/bin/bash
WORKDIR=/home/u/f043139/simuAor1
sbatch_script_0=${WORKDIR}/sbatch_script_0.sh
#############
# Execution #
#############
for i in {1..100};
do
   sbatch --job-name=marc${i} --output=${WORKDIR}/aor1_${i}outMBmdr.log --partition=all_5hrs ${sbatch_script_0} ${i} ${WORKDIR}
done
```

Code # 2a-2 : MB-MDR steps on each of the simulation dataset - no LD pruning, MB-MDR :

The following code carry out the MB-MDR algorithm on unpruned dataset. The MB-MDR options are :

- -d 2D : analyse interaction
- -a NONE : no correction for main effect
- HvsL : the category "0" (undecided) is excluded from the test comparing H vs L or L vs H in MDR
- -pb NONE : no progress bar to be print (would otherwise cause a failure in the computer cluster)
- -v SHORT : verbose shortly

```
#!/bin/bash
#${1} = i
#${2} = WORKDIR
WORKDIR=/home/u/f043139/simuAor1
SOFTWDIR2=/home/u/f043139/DataMarc
executable=${SOFTWDIR2}/mbmdr-4.4.2.out
options="--binary -d 2D -a NONE -h HVSL -pb NONE -v SHORT"
INPUTDIR=/home/u/f043139/simuAor1
WORKDIR= $2
#############
#############
# Execution #
#############
infile=${INPUTDIR}/aor1_${1}inMBmdr.txt
${executable} ${options}  -o ${WORKDIR}/aor1_${1}outMBmdr.txt ${infile}
```

Code # 2b-1 : bash script `Jobs` (Unix/Linux) : processing 100 simulation files (e.g aor1_#inMBmdr.txt) for LD pruning with PLINK and convert the output file with proper format for MB-MDR later input :

```
#!/bin/bash
WORKDIR=/home/u/f043139/simuAor1
sbatch_script_LD_convert=${WORKDIR}/sbatch_script_LD_convert.sh
#############
# Execution #
#############
for i in {1..100};
do
    sbatch --job-name=marc${i} --output=${WORKDIR}/Pru20_aor1_${i}_in_MBmdr.log --partition=all_5hrs ${sbatch_script_LD_convert} ${i} ${WORKDIR}
done
```

Code # 2b-2 : PLINK steps and MB-MDR steps - LD pruning at $r^2 \geq 0.20$ and MB-MDR format conversion :

```
#!/bin/bash
#${1} = i
#${2} = WORKDIR

WORKDIR=/home/u/f043139/simuAor1
SOFTWDIR=/home/u/f043139/testfolder
module load plink
executable1=plink
executable2=${SOFTWDIR}/mbmdr-4.4.1-linux-64bits.out
options="--binary -d 2D"
INPUTDIR=/home/u/f043139/simuAor1
#############
# Execution #
#############
# LD Pruning with PLINK at a specified r2 level in 2 steps :
# LD pruning step (a) : get a subset of more independent SNPs with lower redundancy in tagging-SNP :
${executable1} --noweb --ped ${WORKDIR}/dataor1_${1}.ped --map ${WORKDIR}/ch78work.map --indep-pairwise 10 2 0.20 --out ${WORKDIR}/dataor1_${1}_pru20
# LD pruning step (b) : performs the pruning :
${executable1} --noweb --ped ${WORKDIR}/dataor1_${1}.ped --map ${WORKDIR}/ch78work.map --extract ${WORKDIR}/dataor1_${1}_pru20.prune.in
--make-bed --out ${WORKDIR}/pru20aor1_${1}
# Reload the .ped file :
${executable1} --noweb --bfile ${WORKDIR}/pru20aor1_${1} --recode --out pru20aor1_${1}
# .ped and .map are produced by plink....

# .map management step : the purpose is to add a dummy header before conversion to mb-mdr format :
cp ${WORKDIR}/pru20aor1_${1}.map ${WORKDIR}/pru20aor1_${1}_ORIG.map
echo dummy header line > ${WORKDIR}/dummyHeader20_${1}.txt
cat ${WORKDIR}/pru20aor1_${1}.map >> ${WORKDIR}/dummyHeader20_${1}.txt
mv ${WORKDIR}/dummyHeader20_${1}.txt ${WORKDIR}/pru20aor1_${1}.map

# final conversion from .ped to input format for mb-mdr
${executable2} --plink2mbmdr --binary -ped ${WORKDIR}/pru20aor1_${1}.ped -map ${WORKDIR}/pru20aor1_${1}.map
-o ${WORKDIR}/pru20aor1_${1}_in_MBmdr.txt -tr ${WORKDIR}/tr_datAor
```

Code # 2c-1 : bash script `Jobs` (Unix/Linux) : processing 100 pruned simulation files (e.g pru20aor1_#_in_MBmdr.txt) after LD pruning with PLINK and after format conversion suitable for MB-MDR input, the batch code called will carry out MB-MDR algorithm on each of the 100 LD pruned datafiles :

```
#!/bin/bash
WORKDIR=/home/u/f043139/simuAor1
sbatch_script_0=${WORKDIR}/sbatcha_script_75.sh
#############
# Execution #
#############
for i in {1..100};
do
   sbatch --job-name=marc${i} --output=${WORKDIR}/Pru20aor1_${i}outMBmdr.log --partition=all_5hrs ${sbatch_script_0} ${i} ${WORKDIR}
done
```

Code # 2c-2 : MB-MDR analysis of LD pruned dataset :

The MB-MDR options are the same as for the unpruned dataset detailed before.

```
#!/bin/bash
#${1} = i
#${2} = WORKDIR
WORKDIR=/home/u/f043139/simuAor1
SOFTWDIR2=/home/u/f043139/DataMarc
executable=${SOFTWDIR2}/mbmdr-4.4.2.out
options="--binary -d 2D -a NONE -h HVSL -pb NONE -v SHORT"
INPUTDIR=/home/u/f043139/simuAor1
WORKDIR= $2
#############
# Execution #
#############
infile=${INPUTDIR}/pru20aor1_${1}_in_MBmdr.txt
${executable} ${options}  -o ${WORKDIR}/pru20aor1_${1}outMBmdr.txt ${infile}
```

B.4 Codes for sensitivity analysis of MB-MDR

The specifications and functions of the program `Sensitivity.py` are the following. The program receives three filenames in arguments :

1. `Aor1_1outMBmdr.txt` is a file name instance of the MB-MDR output file containing the ranked list of the detected marker pairs, the chi-square statistic and the FWER adjusted p-values (computed from the gammaMAXT algorithm). This file has three header lines with 4 columns :

   ```
   First marker    Second marker    chi-square  p-value.
   ```

2. `DSL1snp20.tags` is the file name of the tag SNP list associated to the first known causal marker of the interacting pair of interest. This file has no header and two colums and was produced by PLINK :

   ```
   tag marker  yes/no
   rs1234        0
   rs1235        1
   ```

   etc...

3. `DSL2snp20.tags` is the file name of the tag SNP list associated to the second known causal marker of the interacting pair of interest. This file has no header and two columns and was produced by PLINK :

   ```
   tag marker  yes/no
   rs3456        0
   rs3457        1
   ```

   etc..

The `Sensitivity.py` program executes the following :

1. Counts the total number of detected SNP pairs (lines) of the MB-MDR output file : $C1$.
2. Counts the number of detected SNP pairs that are significant at a p-value level ≤ 0.05. This is $C2$.
3. Returns a causal flag $= 1$ if the MB-MDR output file contains both causal variants that are known to the program through proper initialization (causal variant 1 variable initialized to 'rs1223' and causal variant 2 variable initialized to 'rs5678' for instance if these particular markers are the known causal interacting SNPs) : $C3$. If $C3$ is 1, the corresponding adjusted p-value is given in $C7$. The flag $C3$ is set to 1 if and only if the p-value is ≤ 0.05.
4. Returns the number of the MB-MDR output file where both causal variants were found (or equivalently where a causal flag $= 1$ was returned) : $C4$.
5. Counts the number of significant detected pairs that are composed of two marker SNPs that belong each to each of the two tag-SNP files : a counter is incremented when marker 1 $\in$ tagSNP1.txt and marker 2 $\in$ tagSNP2.txt (or the other way around, i.e, marker 1 $\in$ tagSNP2.txt and marker 2 $\in$ tagSNP1.txt) or when a causal variant matches a tag-SNP : $C5$.
6. Calculate the proportion of the significant tag SNPs pairs in the total detected significant SNP pairs : $C6$ (= C5/C2) with 3 digits. If $C2 = 0$, then $C6 = NA$.
7. Outputs all the previous information in a file that has only one line (with no header) with the following tab delimited columns : $C1 \;\; C2 \;\; C3 \;\; C4 \;\; C5 \;\; C6 \;\; C7$.
8. Append the previous file to an existing pre-formatted file with a header for each of the seven columns. In the case of the 100 datasets `Aor1_${i}outMBmdr.txt` ($i \in [1, \ldots, 100]$), the preformatted file is named `Aor1_sensitivity.txt`.

The `Sensitivity.py` program is wrapped in a bash script called `SensiBatch` to process 100 datasets with a for loop on the filename number in order to produce the final appended file that will contain 100 lines and from which both the exact sensitivity can be computed (sum of $C3$ column) and the signal sensitivity can be determined (as well as a quantile of column $C6$) for the given setting (particular known pair of causal variants, known effect size, known level of pruning).

The 3 data files should have their directory path indicated in the bash wrapper script.

Code # 3a : bash script `SensiBatch` (Unix/Linux) :

```
#!/usr/bin/bash
# Define paths for the input files requested by the Python program:
INPUTDIR1=/home/marc/GIGA/SIMUResults/notpruned_HvsL
INPUTDIR2=/home/marc/GIGA/SIMUResults
for fileNbID in {1..100};
do
# The Python program is named Sensitivity.py and is called from this bash script.
python Sensitivity.py ${INPUTDIR1}/Aor1_${fileNbID}outMBmdr.txt ${INPUTDIR2}/DSL1snp20.tags
    ${INPUTDIR2}/DSL2Asnp20.tags
done
# exit properly :
exit 0
```

Code # 3b : code of `Sensitivity.py` (Python) :

```
#!/usr/bin/python
# This program receives iterated filenames (Aor1_{$i}outMBmdr.txt) as argument 1 from
# a calling bash script, a DSL1snp20.tags file as argument 2 where the tag SNPs are listed
# from causal variant DSL 1 with LD(r2) = 0.20, a DSL2snp20.tags file as argument 3 where
# the tag SNPs are listed from causal variant DSL 2 with LD(r2) = 0.20.
# DSL 1 and DSL 2 are in pure epistatic interaction.
# The program process the output file of MB-MDR and outputs information that will be used
# to determine both exact sensitivity (power) and signal sensitivity as defined in
# the thesis.
# The paths for the input files are :
# /home/marc/GIGA/SIMUResults ... for both DSLsnp20.tags files
# /home/marc/GIGA/SIMUResults/notpruned_HvsL ... for the MBmdr results output txt file.
# These paths are used in the bash script to parse the right filename with the right
# argument.

# sys is required for the argument parsing :
import sys
#------------------------------------------------------------------------------
# Check the number of arguments that were parsed if there are the ones expected:
#------------------------------------------------------------------------------
if len(sys.argv) < 4:
    print(’Usage : python ’ + sys.argv[0] + ’ filenames’)
    exit()
if len(sys.argv) > 4:
    print(’Error : too many arguments added after the python program.’)
    exit()
```

```
# --------------------------------------------------------------------------
# Get the name of the 3 files to be imported as the required inputs :
# --------------------------------------------------------------------------
infile1 = sys.argv[1]
infile2 = sys.argv[2]
infile3 = sys.argv[3]
#----------------------------------------------------------------------------
# Import and read files to retrieve data of interest from the input files
# The 3 requested files paths should be known from this program.
#----------------------------------------------------------------------------
def importPAIRS(filename1):
    # Read the SNPs pairs from "filename1" : marker1, marker2,
    # their chi-square statistic, and the adjusted p-value.
    MBmdrResultsFile = open(filename1, 'r')
    marker1 = list()  # or = []
    marker2 = list()
    chi_2 = list()
    p_val = list()

    Tot_number_of_significant_pairs = 0

    # There are normally 3 header lines in this file
    # skip the first three header lines :
    header_line_1 = MBmdrResultsFile.readline()
    header_line_2 = MBmdrResultsFile.readline()
    header_line_3 = MBmdrResultsFile.readline()
    # Read in all the lines that are left till the end of file:
    line = MBmdrResultsFile.readline()
    line_count = 0
    while line != ’’:
        # get all the snps’names:
        col = line.split()
        FirstMarker = col[0]
        SecondMarker = col[1]
        statistic = float(col[2])
        p = float(col[3])
        if p <= 0.05:
            Tot_number_of_significant_pairs += 1

        marker1.append(FirstMarker)
        marker2.append(SecondMarker)
        chi_2.append(statistic)
        p_val.append(p)

        line_count += 1
        line = MBmdrResultsFile.readline()

    print(line_count, ’ lines have been read from the MB-MDR output file.’)  # C1
    print(Tot_number_of_significant_pairs, ’ significant SNP pairs have been detected.’)
    # C2

    # close the file :
    MBmdrResultsFile.close()

    return marker1, marker2, chi_2, p_val, line_count, Tot_number_of_significant_pairs

def importTAGsnp(filenameForTagSNPs):
    # There is normally no header for the tag SNP file.

    # open the file of interest :
    TAGsnpFile = open(filenameForTagSNPs, ’r’)

    # initialize tagSNPlist :
    tagSNPlist = list()

    # read the file and select only the tag-SNPs based on second column flag(0/1) :
    for line in TAGsnpFile.readlines():
        col = line.split()
        if int(col[1]) == 1:
            tagSNPlist.append(col[0])

    # close the file :
    TAGsnpFile.close()

    return tagSNPlist
```

```
#-----------------------------------------------------------------------------------
#    MAIN
#-----------------------------------------------------------------------------------
# Get tag-SNPs for causal variant DSL 1 :
tagSNP1 = importTAGsnp(infile2)
# the length of tagSNPlist is the number of tag-SNPs for the causal variant of interest :
print(len(tagSNP1), ’ tag SNPs are associated to DSL1 causal marker.’)
print(’tag-SNPs list :’)
for snp in tagSNP1[0:]:
    print(snp)

# Get tag-SNPs for causal variant DSL 2 :
tagSNP2 = importTAGsnp(infile3)
# the length of tagSNPlist is the number of tag-SNPs for the causal variant of interest :
print(len(tagSNP2), ’ tag SNPs are associated to DSL2 causal marker.’)
print(’tag-SNPs list :’)
for snp in tagSNP2[0:]:
    print(snp)

# Get information from current MB-mdr output file of interest :
FirstMarker, SecondMarker, chi_square, p_value, C1, C2 = importPAIRS(infile1)

# Initialize the current pair of causal loci of interest:
DSL1 = ’rs17644404’
DSL2 = ’rs10956767’ # DSL2 A
# DSL2 = ’rs112698197’ # DSL2 D
# DSL2 = ’rs1476427’ # DSL2 C
# DSL2 = ’rs2073640’ # DSL2 B

# Task 1 and Task 2 are the C1 and C2 information retrieved from previous function call.

# Task 3: set a flag to 1 (otherwise 0) if the current MB-MDR file contains the true
# causal pair :
# and if and only if p-value is significant for the pair:
flag = 0
for i in range(C1):
    if DSL1 == FirstMarker[i] and DSL2 == SecondMarker[i] and p_value[i] <= 0.05:
        flag = 1
        idx = i
    elif DSL2 == FirstMarker[i] and DSL1 == SecondMarker[i] and p_value[i] <= 0.05:
        flag = 1
        idx = i

if flag == 1:
    C3 = 1
    C7 = p_value[idx]
else:
    C3 = 0
    C7 = ’NA’
```

```
# Task 4 : identifies the number ID of the current MB-MDR output file where
# a pair of causal SNPs were found significant
print('filename = ', infile1)
C4 = 'NA'
if C3 == 1:
    pos = infile1.find('out')
    if infile1[pos-2].isdigit():
        C4 = infile1[pos-2:pos]
    else:
        C4 = infile1[pos-1]

# Task 5 : counts the number of signal detection (detection of both tag SNPs of the 2
# causal variants):
signal_count = 0
for i in range(C1):
    if FirstMarker[i] in tagSNP1 and SecondMarker[i] in tagSNP2 and p_value[i] <= 0.05:
        signal_count += 1
        print(FirstMarker[i], ' X ', SecondMarker[i], ' is an indirect detection pair.')
    elif FirstMarker[i] in tagSNP2 and SecondMarker[i] in tagSNP1 and p_value[i] <= 0.05:
        signal_count += 1
C5 = signal_count

# Task 6 : Calculate the proportion of indirect detection among the pairs detected
# significant:
if C2 > 0:
    C6 = format(float(C5/C2), '5.3f')
else:
    C6 = 'NA'

# Task 7 : Outputs all collected information regarding the current file in a line saved
# in a single lined output file.
detection_output = open('detectionResultLine.txt', 'w')
single_output_line = str(C1)+'\t'+str(C2)+'\t'+str(C3)+'\t'+str(C4)+'\t'+str(C5)+
'\t'+str(C6)+'\t'+str(C7)
detection_output.write(single_output_line)
detection_output.close()

# Task 8 : append the current single_output_line to an existing pre-formatted file with
# proper header for the 7 columns. The existing file is like 'Aor1_sensitivity.txt'.
outfile = open('Aor1_sensitivity.txt', 'a')
outfile.write(single_output_line +'\n')
outfile.close()

print(C1, ' lines were read in the current MB-MDR output file.')
print(C2, ' SNP pairs were detected significant at 0.05 level.')
print(C3, ' 1 yes the causal pair was found significant : direct hit. 0: otherwise.')
if C3 == 1:
    print(FirstMarker[idx], ' X ', SecondMarker[idx],
    ' is the pair of causal variant found significant.')
print(C4, ' ID number of the file where a direct hit of causal pair was detected.')
print(C5, ' indirect detections were found in the file.')
print(C6, ' is the proportion of indirect signal detections among the significant
detections.')
print(C7, ' is the p-value for the direct detection of the causal pair of SNPs.')
```

B.5 Statistical analysis for the real dataset : ankylosing spondylitis dataset

The jobs and PLINK commands for quality control and statistical analysis of the ankylosing spondylitis dataset are the following.

Code # 4a : PLINK command for phenotype missingness filtering :

```
#!/bin/bash
#SBATCH -p all_5hrs
#SBATCH --job-name=MJOpca
#SBATCH --output=sim-%A.log
module load plink
# ----------------------------------------------------------------
# Step PLINK
# ----------------------------------------------------------------
echo "=============================================================="
echo "      MEMO: Plink 1.07 :                                      "
echo "=============================================================="
# Making a binary PED file :
plink --noweb --file AS_NBS_58C_CH1_to_22_NatureGen --make-bed --out ASfile1

# Remove individuals with missing phenotype :
plink --noweb --bfile ASfile1 --prune --mind 0.02 --make-bed --out ASfile2
```

Code # 4b : PLINK command to remove SNPs with genotype missing rate $\geq 10\%$ and SNPs with minor allele frequencies (MAF) less than 1% and Hardy-Weinberg equilibrium significance test with p-value $5.0e-15$. :

```
# Remove the SNPs with genotype missing rate > 10%, MAF < 1% and HWE p-value < 5.0E-15 :
plink --noweb --bfile ASfile2 --geno 0.1 --maf 0.01 --hwe 5.0E-15 --make-bed --out ASfile_Cleaned
```

Code # 4 c : PLINK command for multidimensional scaling (MDS) (or equivalently PCA) to generate a file for further graphical display in R. MDS on whole genome after LD pruning at $r^2 \leq 0.03$. Genomic Inflation factor and mean chi-squared statistic for stratification test :

```
# Generate files for PCA (Multidimensional Scaling) (check for substructure) :
# Step (a) prune SNPs to get a subset of more independent ones :
plink --noweb --bfile ASfile_Cleaned --maf 0.01 --indep-pairwise 50 5 .03 --out ASfile.prune

## Pruning proper :
plink --noweb --bfile ASfile_Cleaned --extract ASfile.prune.prune.in --make-bed --out ASfileindep

# Check for absence of population substructure :
# Genomic inflation factor in the log file :
plink --noweb --bfile ASfileindep --assoc --adjust --out as_pop_sub1
plink --noweb --bfile ASfileindep --assoc --adjust --gc --out as_pop_sub2gc

# Step (b) generate principal components after pruning can be visualized in R later :
plink --noweb --bfile ASfileindep --genome --out PCAstep1
plink --noweb --bfile ASfileindep --read-genome PCAstep1.genome --cluster --mds-plot 4 --out ibs_view1
```

Code # 4 f : PLINK command for chromosome 6 analysis of 3 Mbps region around `HLA-B` locus and association test to ankylosing spondylitis affection status, LD analysis of best SNP list at r2=0.50 and r2 = 0.20. :

```
# Making a binary PED file after extracting only the snps in region of HLA coding genes
# in chr 6 (30Mb to 33Mb) from 30000000 bp position to 33000000 bp position as inquired on UCSC Genome web browser :
plink --noweb --bfile ASfile_Cleaned --chr 6 --from-bp 30000000 --to-bp 33000000 --make-bed --out ch6HLA_snps

# Basic association analysis of affection status with single SNPs from chr 6 in HLA region :
plink --noweb --bfile ch6HLA_snps --assoc --out hla_as1

# Basic association analysis of affection status with single SNPs from chr 6 in HLA region
and multiple testing correction :
plink --noweb --bfile ch6HLA_snps --assoc --adjust --out hla_as2

# Obtaining LD values within 1Mb for all SNPs contained in the 30 SNPs best associated with AS :
plink --noweb --bfile ch6HLA_snps --r2 --ld-snp-list MySNP_hla_as.txt --ld-window-kb 1000 --ld-window 99999 --ld-window-r2 0 --out hlaLDallSNPset

# Find tag-SNPs of each SNPs best subset:
# with LD r2 > 0.50
plink --noweb --bfile ch6HLA_snps --show-tags MySNP_hla_as.txt --tag-r2 0.50 --tag-kb 1000 --out hla_tagSNP50

# Find tag-SNPs of each SNPs best subset:
# with LD r2 > 0.20
plink --noweb --bfile ch6HLA_snps --show-tags MySNP_hla_as.txt --tag-r2 0.20 --tag-kb 1000 --out hla_tagSNP20
```

Code # 4 g : PLINK command for merging chromosome 5 and chromosome 6 - 3 Mbps region around `HLA-B` locus :

```
# Making a binary PED file after extracting only chromosome 5 :
plink --noweb --bfile ASfile_Cleaned --chr 5 --make-bed --out chr5
# Merge chromosome 5 (with all snps) with chromosome 6 (with only 3Mb region around HLA : 946 snps).
# Note the --recode option to generate .ped and .map format for the merged output file.
# These formats .ped and .map are necessary for later processing by MB-MDR.
plink --noweb --bfile chr5 --bmerge ch6HLA_snps.bed ch6HLA_snps.bim ch6HLA_snps.fam --recode --out chr56HLA
```

Code # 4 h : PLINK command for LD pruning the merged chromosome 5 and chromosome 6 - 3 Mbps region around `HLA-B` locus :

```
#!/bin/bash
#${1} = i
#${2} = WORKDIR
WORKDIR=/home/u/f043139/Projet1
SOFTWDIR=/home/u/f043139/testfolder
module load plink
executable1=plink
options="--binary -d 2D"
INPUTDIR=/home/u/f043139/simuAor1
#WORKDIR= ${2}
#############
# Execution #
#############
# LD Pruning with PLINK at a specified r2 level in 2 steps :
# LD pruning step (a) : get a subset of more independent SNPs with lower redundancy in tagging-SNP :
${executable1} --noweb --ped ${WORKDIR}/chr56HLA.ped --map ${WORKDIR}/chr56HLA.map --indep-pairwise 50 5 0.50 --out ${WORKDIR}/chr56HLA_${1}_pru50
# LD pruning step (b) : performs the pruning :
${executable1} --noweb --ped ${WORKDIR}/chr56HLA.ped --map ${WORKDIR}/chr56HLA.map --extract ${WORKDIR}/chr56HLA_${1}_pru50.prune.in --make-bed --out ${WORKDIR}/pru5
# Reload the .ped file :
${executable1} --noweb --bfile ${WORKDIR}/pru50chr56HLA_${1} --recode --out pru50chr56HLA1_${1}
# .ped and .map are produced by plink....

# .map management step : the purpose is to add a dummy header before conversion to mb-mdr format :
cp ${WORKDIR}/pru50chr56HLA1_${1}.map ${WORKDIR}/pru50chr56HLA1_${1}_ORIG.map
echo dummy header line > ${WORKDIR}/dummyHeader50_${1}.txt
cat ${WORKDIR}/pru50chr56HLA1_${1}.map >> ${WORKDIR}/dummyHeader50_${1}.txt
mv ${WORKDIR}/dummyHeader50_${1}.txt ${WORKDIR}/pru50chr56HLA1_${1}.map

# final conversion from .ped to input format for mb-mdr
${executable2} --plink2mbmdr --binary -ped ${WORKDIR}/pru50chr56HLA1_${1}.ped -map ${WORKDIR}/pru50chr56HLA1_${1}.map -o ${WORKDIR}/pru50chr56HLA_${1}_in_MBmdr.txt -
```

Code # 4 i : MB-MDR analysis of the LD-pruned file :

```
#!/bin/bash
WORKDIR=/home/u/f043139/Projet1
sbatch_script_0=${WORKDIR}/asbatch_script_50.sh
#############
# Execution #
#############
for i in {1..1};
do
   sbatch --job-name=maxN0${i} --output=${WORKDIR}/Pru50as_${i}outMBmdr.log --partition=urtgen_24hrs ${sbatch_script_0} ${i} ${WORKDIR}
done
############## calls for :
#!/bin/bash
#${1} = i
#${2} = WORKDIR

WORKDIR=/home/u/f043139/Projet1
SOFTWDIR2=/home/u/f043139/DataMarc
executable=${SOFTWDIR2}/mbmdr-4.4.2.out
# no correction for main effects and HvsL version with gammaMAX :
options="--binary -d 2D -a NONE -h HVSL -n 5000 -pb NONE -v SHORT"
# H vs L and 0 (undecided) version with gammaMAX :
#options="--binary -d 2D -a NONE -pb NONE -v SHORT"
# correction for main effects ADDITIVE or CODOMINANT :
#options="--binary -d 2D -a ADDITIVE -n 5000 -pb NONE -v SHORT"
#options="--binary -d 2D -a CODOMINANT -n 5000 -pb NONE -v SHORT"
# with MAXT version (too long) :
#options="--binary -d 2D -mt MAXT -a NONE -h HVSL -n 5000 -pb NONE -v SHORT"
#options="--binary -d 2D -mt MAXT -a CODOMINANT -n 5000 -pb NONE -v SHORT"

INPUTDIR=/home/u/f043139/Projet1
WORKDIR= $2
#############
#############
# Execution #
#############
infile=${INPUTDIR}/pru50chr56HLA_${1}_in_MBmdr.txt
${executable} ${options}  -o ${WORKDIR}/pru50k5chr56HLA_${1}outMBmdr.txt ${infile}
```

Appendix C

Simulation Results

The results for the sensitivities (exact and signal sensitivities) of MB-MDR to detect our two-locus pure epistatic interaction in the different settings are tabulated in Table C.1 and were graphically displayed at chapter 3. Table C.1 shows sensitivity results of MB-MDR for the detection of two epistatic loci in the same LD block (setting A), two epistatic loci in the middle of separate LD blocks (seeting B), two epistatic loci in separate LD blocks but with one locus at an edge (setting C), and two epistatic loci in separate LD blocks on different chromosomes (setting D), for three implemented effect sizes and for five LD pruning levels. The sensitivities were calculated as the number of times the epistatic loci were detected out of the 100 simulated datasets that were constructed following the procedures exposed in Methods.

Table C.1 – Sensitivity results of MB-MDR to detect two locus model of pure epistatic interaction in 1200 simulated datasets with real human genome LD patterns, for 3 effect sizes and after 5 LD pruning levels.

LD block setting	LD pruning	Effect Size	Exact Sensitivity	Signal Sensitivity	
				tag-SNP condition LD $r^2 \geq 0.45$	*tag-SNP condition* LD $r^2 \geq 0.20$
A Two SNPs in same LD block	unpruned	$\beta_3 = 0.90$	0.61	0.67	0.73
		$\beta_3 = 0.75$	0.55	0.65	0.77
		$\beta_3 = 0.50$	0.70	0.85	0.89
	LD $r^2 \leq 0.75$	$\beta_3 = 0.90$	0.01	0.90	0.91
		$\beta_3 = 0.75$	0.04	0.92	0.94
		$\beta_3 = 0.50$	0.03	0.93	0.93
	LD $r^2 \leq 0.60$	$\beta_3 = 0.90$	0.01	0.93	0.94
		$\beta_3 = 0.75$	0.00	0.94	0.94
		$\beta_3 = 0.50$	0.01	0.92	0.94
	LD $r^2 \leq 0.50$	$\beta_3 = 0.90$	0.00	0.91	0.92
		$\beta_3 = 0.75$	0.00	0.90	0.91
		$\beta_3 = 0.50$	0.01	0.91	0.95
	LD $r^2 \leq 0.20$	$\beta_3 = 0.90$	0.00	0.61	0.74
		$\beta_3 = 0.75$	0.00	0.69	0.80
		$\beta_3 = 0.50$	0.01	0.66	0.84

Table C.1 – Sensitivity results of MB-MDR to detect two locus model of pure epistatic interaction in 1200 simulated datasets with real human genome LD patterns, for 3 effect sizes and after 5 LD pruning levels.

LD block setting	LD pruning	Effect Size	Exact Sensitivity	Signal Sensitivity	
				tag-SNP condition LD $r^2 \geq 0.45$	*tag-SNP condition* LD $r^2 \geq 0.20$
B Two SNPs in middle of two separate LD blocks	unpruned	$\beta_3 = 0.90$	0.54	0.75	0.75
		$\beta_3 = 0.75$	0.46	0.70	0.71
		$\beta_3 = 0.50$	0.41	0.75	0.76
	LD $r^2 \leq 0.75$	$\beta_3 = 0.90$	0.64	0.91	0.91
		$\beta_3 = 0.75$	0.58	0.91	0.91
		$\beta_3 = 0.50$	0.44	0.93	0.94
	LD $r^2 \leq 0.60$	$\beta_3 = 0.90$	0.49	0.92	0.92
		$\beta_3 = 0.75$	0.41	0.93	0.93
		$\beta_3 = 0.50$	0.27	0.94	0.95
	LD $r^2 \leq 0.50$	$\beta_3 = 0.90$	0.39	0.92	0.92
		$\beta_3 = 0.75$	0.32	0.93	0.93
		$\beta_3 = 0.50$	0.23	0.94	0.95
	LD $r^2 \leq 0.20$	$\beta_3 = 0.90$	0.19	0.57	0.81
		$\beta_3 = 0.75$	0.16	0.69	0.91
		$\beta_3 = 0.50$	0.21	0.83	0.92
C One SNP in a block and one in the edge of a separate LD block	unpruned	$\beta_3 = 0.90$	0.18	0.33	0.43
		$\beta_3 = 0.75$	0.23	0.36	0.49
		$\beta_3 = 0.50$	0.18	0.36	0.51
	LD $r^2 \leq 0.75$	$\beta_3 = 0.90$	0.0	0.65	0.74
		$\beta_3 = 0.75$	0.0	0.72	0.83
		$\beta_3 = 0.50$	0.0	0.57	0.76
	LD $r^2 \leq 0.60$	$\beta_3 = 0.90$	0.0	0.56	0.74
		$\beta_3 = 0.75$	0.0	0.59	0.81
		$\beta_3 = 0.50$	0.0	0.47	0.74
	LD $r^2 \leq 0.50$	$\beta_3 = 0.90$	0.0	0.48	0.71
		$\beta_3 = 0.75$	0.0	0.50	0.81
		$\beta_3 = 0.50$	0.0	0.36	0.70
	LD $r^2 \leq 0.20$	$\beta_3 = 0.90$	0.0	0.07	0.60
		$\beta_3 = 0.75$	0.0	0.05	0.62
		$\beta_3 = 0.50$	0.0	0.04	0.57

Table C.1 – Sensitivity results of MB-MDR to detect two locus model of pure epistatic interaction in 1200 simulated datasets with real human genome LD patterns, for 3 effect sizes and after 5 LD pruning levels.

LD block setting	LD pruning	Effect Size	Exact Sensitivity	Signal Sensitivity	
				tag-SNP condition LD $r^2 \geq 0.45$	*tag-SNP condition* LD $r^2 \geq 0.20$
D Two SNPs on LD blocks of separate chromosomes	unpruned	$\beta_3 = 0.90$	0.39	0.68	0.82
		$\beta_3 = 0.75$	0.40	0.69	0.81
		$\beta_3 = 0.50$	0.58	0.76	0.84
	LD $r^2 \leq 0.75$	$\beta_3 = 0.90$	0.18	0.86	0.94
		$\beta_3 = 0.75$	0.18	0.93	0.99
		$\beta_3 = 0.50$	0.23	0.84	0.90
	LD $r^2 \leq 0.60$	$\beta_3 = 0.90$	0.14	0.87	0.94
		$\beta_3 = 0.75$	0.13	0.93	0.98
		$\beta_3 = 0.50$	0.17	0.85	0.90
	LD $r^2 \leq 0.50$	$\beta_3 = 0.90$	0.13	0.85	0.92
		$\beta_3 = 0.75$	0.13	0.93	0.97
		$\beta_3 = 0.50$	0.16	0.83	0.89
	LD $r^2 \leq 0.20$	$\beta_3 = 0.90$	NA	NA	NA
		$\beta_3 = 0.75$	0.10	0.67	0.86
		$\beta_3 = 0.50$	0.17	0.75	0.84

Appendix D

Real life dataset results on ankylosing spondylitis

D.1 Population substructure analysis results

The visual inspection of the graphs of Figure D.1 for the principal components does not indicate any obvious substructure pattern. So, the absence of population substructure is reasonable in the WTCCC2 dataset on ankylosing splondylitis.

D.2 Single loci association analysis

Table D.1 – Genotype table of rs2523608 SNP for affection status.

	Genotype			
	AA	Aa	aa	total
	'TT'	'TC'	'CC'	
CASES	845	603	26	1474
CONTROL	1689	2288	807	4784
total	2534	2891	833	6258

As an example of single locus association analysis, we tabulate each genotype (homozygote AA, heterozygote Aa and homozygote aa) against case and control. We have a 2 rows by 3 columns, so the test will have $(2-1)(3-1) = 2$ degrees of freedom, for each SNP. This basic allele test makes no assumptions about the genetic model.

The genotype of the `rs2523608` SNP in the HLA-B region of chromosome 6 is in Table D.1 (the 314 cases with unspecified gender were removed from the calculation and 15 controls with incomplete genotypes were also removed).

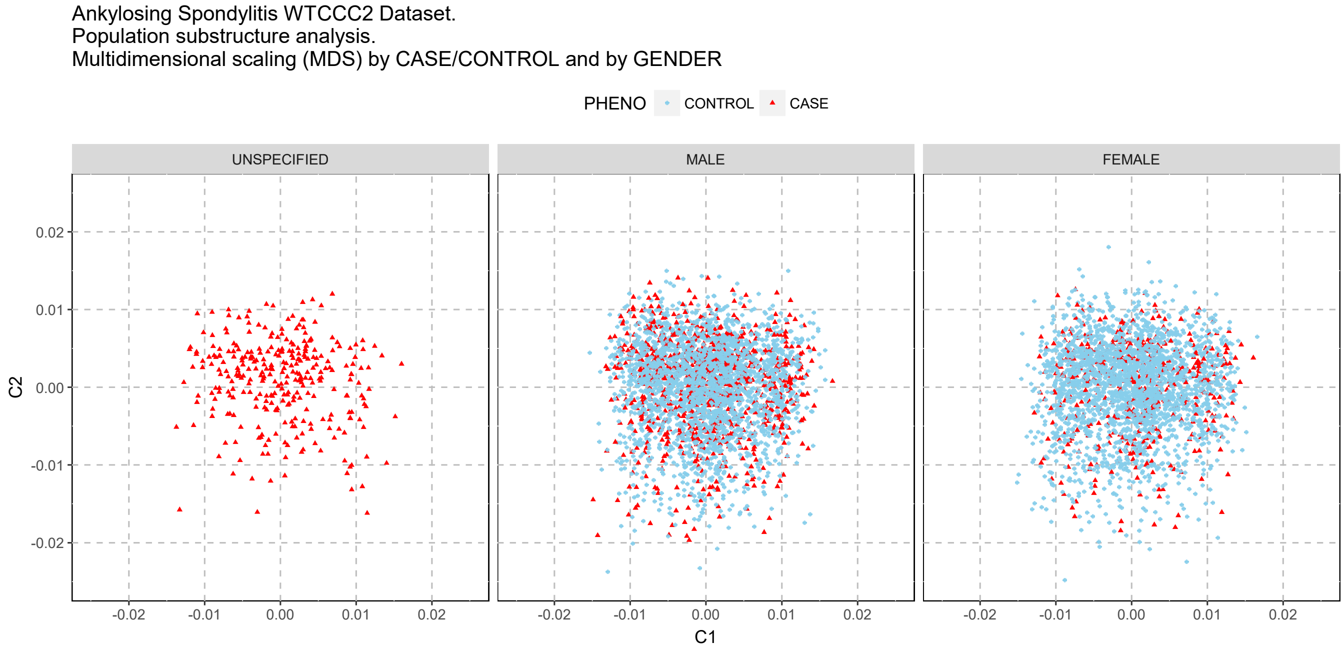


FIGURE D.1 – Applying multidimensional scaling (MDS) for population substructure analysis in the ankylosing spondylitis WTCCC2 dataset : controls (blue points) and cases (red triangles) by gender. Features space : subset of 22861 stochastically independent SNPs on 22 autosomal chromosomes.

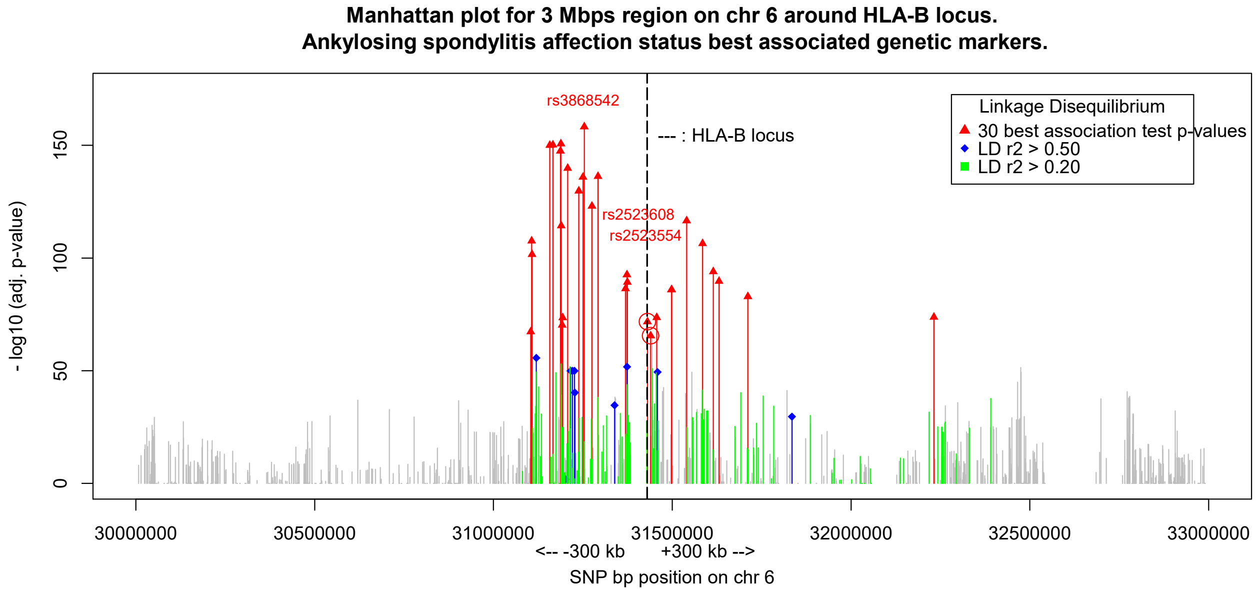


FIGURE D.2 – Manhattan plot of ankylosing spondylitis affection status association test for 946 SNPs genotypes in the chr 6 region around HLA-B locus. Vertical hyphenated black line : HLA-B locus position. Red triangles : 30 most significant SNPs results. Circled red triangles : rs2523608 in HLA-B locus and rs2523554 7 kbs centromeric to HLA-B locus. All red and blue markers are in LD r2 > 0.50 altogether. Green bars are SNPs in LD r2 > 0.20 with 30 most significant SNPs results. Grey bars are SNPs at LD r2 < 0.20 with red SNPs.

The allele test for rs2523608 SNP then yields Table D.2.

Table D.2 – Allele association test of rs2523608 SNP to affection status.

	Alleles		
	Major allele A 'T'	Minor allele a 'C'	total
CASES	2293	655	2948
CONTROL	5666	3902	9568
total	7959	4557	12516

This polymorphic nucleotide (SNP rs2523608) has a='C' as minor allele and A='T' has major allele. The minor allele frequency in the general population is MAF= 40% for 'C' ('T' is 60%). The allele frequency is significantly different for the cases : 22.2% for a='C' minor allele as compared to 40.78% for the 'C' nucleotide for the the controls (similar to the general population). The allele test for association formally tests for the null hypothesis that the allele proportions are the same for cases and controls. The null hypothesis is rejected if the $\chi^2_{obs.} = \sum_{i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ is too large. The p-value is obtained as the probability to have such a large or a more extreme χ^2, under the null, than this observed $\chi^2_{obs.}$ value. Here, we have $\chi^2_{obs.} = 335.4$ with an associated p-value for 2 degrees of freedom, equals to $6.34 \cdot 10^{-75}$. Hence, for this single test, the null hypothesis is clearly rejected that the allele proportion are the same between cases and controls.

When 946 SNPs are tested, we adjust the p-values for controlling the false discovery rate (FDR) correcting for multiple testing by the Benjamini-Hochberg (FDR-BH) procedure.

The Manhattan plot showing the AS affection status association allele test FDR-BH adjusted p-values as a function of base pair position of 946 polymorphic nucleotides (SNPs) in HLA-B neighbouring region of chromosome 6 is displayed on Figure D.2

The 30 nominally most significant SNPs of this chromosome 6 region are outlined in Table D.3 and sorted by decreasing FDR-BH adjusted p-values (-log10 scale).

D.3 Multiple loci association analysis

The remaining SNPs after the LD-pruning are outlined in Table D.4 by the LD-pruning r^2 levels :

Table D.3 – Allele test best associated SNPs of chr 6 HLA-B region to affection status.

	SNP	BP	-log10(FDR-BH adj. p-value)
30	rs2523554	31439808	65.53
29	rs12665700	31104111	67.29
28	rs1042127	31192149	70.26
27	rs2523608	31430537	71.78
26	rs3094212	31193749	73.58
25	rs9266689	31456559	73.64
24	rs2269425	32231617	73.79
23	rs2261033	31711570	82.93
22	rs2844511	31497763	85.94
21	rs2516448	31498389	85.94
20	rs2243868	31369255	86.45
19	rs2524089	31374501	89.31
18	rs6929796	31630648	89.74
17	rs2246954	31373241	92.61
16	rs2071596	31614670	93.97
15	rs10947121	31107976	101.60
14	rs2844498	31584833	106.47
13	rs6457300	31106721	107.60
12	rs1265048	31189388	114.33
11	rs2284178	31540104	116.54
10	rs9295961	31275477	123.03
9	rs1841	31238739	129.73
8	rs1265156	31250276	135.90
7	rs9501522	31292404	136.25
6	rs4959053	31207556	139.83
5	rs2233984	31187243	147.45
4	rs9380215	31157634	150.03
3	rs4947296	31166157	150.20
2	rs4495304	31188697	150.67
1	rs3868542	31253818	158.27

Table D.4 – Remaining SNPs after LD-pruning in the AS merged chromosome 5 and 6 (3Mbps region around HLA-B) dataset.

LD pruning r^2 level	SNPs on chr 5	SNPs on chr 6	Total SNPs
no pruning	30.723	946	31.669
LD $r^2 \geq 0.75$	20.450	545	20.995
LD $r^2 \geq 0.50$	13.798	335	14.133
LD $r^2 \geq 0.20$	5930	110	6.040

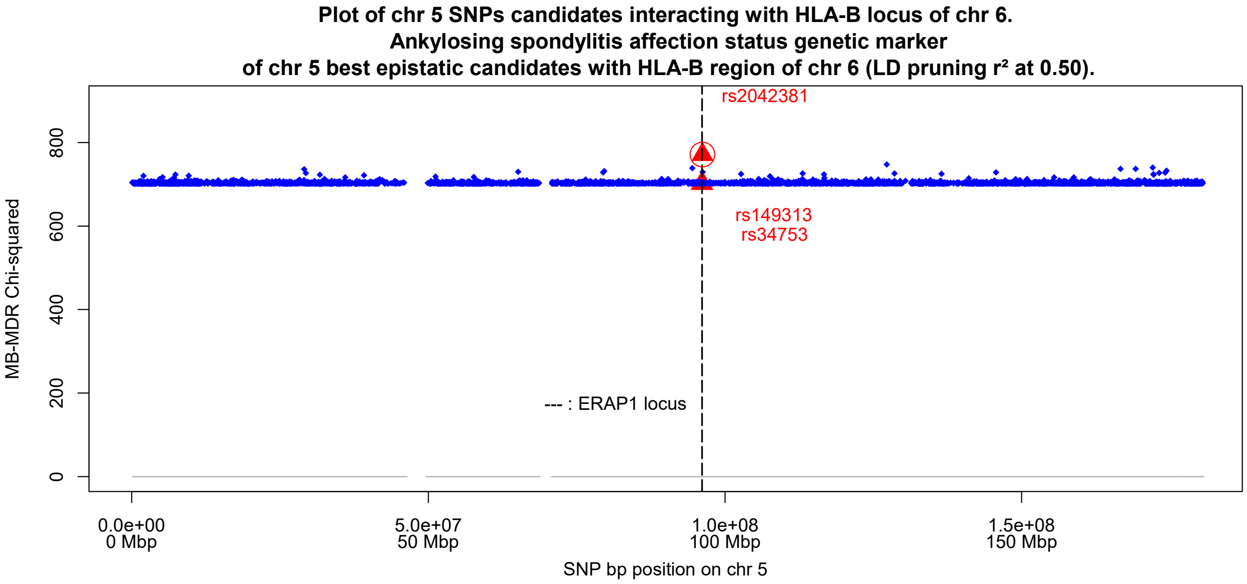


Figure D.3 – Plot of chromosome 5 SNPs candidates interacting with HLA-B locus of chromosome 6. Ankylosing spondylitis affection status genetic marker of chromosome 5 best epistatic candidates with HLA-B region of chromosome 6 (LD-pruning r^2 at 0.50). Vertical hyphenated black line : ERAP1 locus position. Red triangles : 3 SNPs rs2042381, rs149313 and rs34753 in LD $r^2 \geq 0.20$ with ERAP1 locus. Red and blue markers are members of MB-MDR detected pure epistatic pairs with HLA-B locus 30 best tag SNPs.

Glossary

Admixed population — population in which mating occurs between subgroups with different allelic distributions, or more loosely a population in which multiple subgroups are present.

Allelic phase — alignment of nucleotides on a single homolog.

Ascertainment — the act of ascertaining, the process of determining which individuals are sampled and included in the study (case finding) ; or what are the characteristics, status, or events in a population or study group, e.g. exposure ascertainment. Ascertainment bias : systematic failure to represent equally all classes of cases or persons supposed to be represented in a sample.

Case-control design — type of retrospective study design widely used in epidemiological studies, especially genome-wide association studies. People with a specific disease (cases) are chosen with people who do not have the disease (controls). The basic idea is to compare genotypes of cases and controls. If alleles or genotypes at a locus are significantly different in cases and controls, theses alleles or genotypes are claimed to be associated with the disease status. Because disease outcome might be influenced by other characteristics such as sex, age or ethnicity, cases and controls are matched so that these characteristics are similar in these two groups.

Compositional epistasis — the blocking of one allelic effect by an allele at another locus.

Confounding — phenomenon whereby the measure of association between two variables is distorted because other variables, associated with both variables of interest, are not controlled for in the analysis.

Epistasis — Biological epistasis describes a masking effect, whereby a variant or allele at one locus masks the expression of a phenotype at another locus (biological gene-gene interaction). Statistical epistasis describes the situation where the combined effect of two or more loci cannot be predicted from the sum of their individual single-locus effect (statistical gene-gene interaction). In its broadest sense, epistasis refers to the the dependence of the outcome of a mutation on the genetic background.

Gametic Phase Disequilibrium — the correlation between genes at different loci in the same gamete or equivalently the non-random association of alleles within gametes. Unlinked loci (even on different chromosomes) can be associated and are said to be in gametic phase disequilibrium ; abbreviated GPD. There is complete confounding between interaction and GPD.

Genotype — observed pair of DNA bases, one inherited from each parent, at a site on the genome (locus), represented by a categorical variable that takes on values from a predefined set of discrete characters.

GWAS — exploratory investigation of genotype-trait association that involves characterization of a large segment (500-1000 kbp region) of DNA or a whole genome ($\sim$ 3.200 Mbp) ; Genome-wide Association Study, abbreviated GWAS.

Haplotype — refers to the specific combination of alleles that are in alignment on a single homolog, defined as one of the two homologous chromosomes in humans.

Haplotype tagging SNP — sites on the genome that capture overall variability within the gene under consideration and are potentially associated with disease causing variants.

Hardy-Weinberg equilibrium — state in which allele frequencies are constant within a population over generations, or equivalently independence of alleles at a single site between two homologous chromosomes, also referred to as random mating ; abbreviated HWE. Hardy-Weinberg Disequilibrium is a measure of allelic association between two homologous chromosomes at a single site.

Heritability — in the broad sense, it is the ratio of the genetic variance to the phenotypic variance. In the narrow sense, it is the ratio of the additive components of the genetic variance to the phenotypic variance.

Homolog — one member of a pair of homologous chromosomes.

IBD — Identity by descent : 2 alleles at the same genetic marker locus, from 2 individuals (offspring), are called identical by descent (IBD) if these 2 alleles are copies of the identical allele carried by a recent common ancestor. In the case of siblings, this means that the allele shared IBD is from the same parental chromosome, assuming no inbreeding.

IBS — Identity by status : 2 alleles at the same genetic marker locus, from 2 individuals, are called identical by state (IBS) if their DNA sequence is physically identical, i.e., both alleles are A or a, for instance.

Linkage disequilibrium — the measure of allelic association between two different sites on the genome or the non-random association of alleles due to linked loci (in this latter case it is a special case of GPD) ; abbreviated LD.

Locus — portion of the genome that encodes a single gene or the location of a single nucleotide on the genome.

Marker — proximate SNP at which the genotype tends to be associated with the genotype at the true disease-causing locus.

Minor allele — less common allele in a population ; used interchangeably with variant allele.

Penetrance — a measure of the extent to which the presence of a disease allele results in the disease phenotype. The penetrance function describes the conditional probabilities of exhibiting the disease phenotype given the genotype variants under study : $P(Y|G)$ where $Y = 1$ for a case $Y = 0$ for a control and $G = dd$ or $G = Dd$ or $G = DD$ for a bi-allelic genotype. In Mendelian model, the penetrance is equal to one (fully penetrant) for a given value of the genotype while it is less than one in complex traits (reduced penetrance).

Phenocopy — characteristic of an individual who exhibits the disease phenotype but does not carry the disease allele under study.

Population substructure or stratification — presence of multiple subgroups between which there is minimal mating or gene transfer ; also referred to as population stratification.

Proband — an affected individual who is identified independently of everyone else.

SNP — describes a single base pair change that is variable across the general population at a frequency of at least 1% ; Single Nucleotide Polymorphism.

References

[1] Shkedy Z. Lectures Notes in Computational Intensive Methods in Bioinformatics - Second Master in Biostatistics. Hasselt University, 2016.

[2] Claesen J. Lectures Notes in Genetic Epidemiology - Second Master in Biostatistics. Hasselt University, 2016.

[3] Laird NM, Lange C. *The Fundamentals of Modern Statistical Genetics.* Springer, New York, 2011.

[4] Bush WS, Moore JH. Chapter 11 : Genome-wide association studies. *PLoS Comput Biol*, 8(12) :1–11, 2012.

[5] Van Steen K. Travelling the world of gene-gene interactions. *Briefings in Bioinformatics*, 13(1) :1–19, 2012.

[6] Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10 :392–404, 2009.

[7] Wang X, Elston RC, Zhu X. The meaning of interaction. *Human Heredity*, 70 :269–277, 2010.

[8] Sham PC, Cherry SS. *Analysis of Complex Disease Association Studies—A Practical Guide edited by Zeggini E and Morris A. Chapter 1 : Genetic Architecture of Complex Disease.* AP Elsevier, London, 2011.

[9] Foulkes AS. *Applied Statistical Genetics with R—For Population-based Association Studies.* Springer, New York, 2009.

[10] Lander ES, Schork NJ. Genetic dissection of complex traits. *Science*, 265 :2037–2048, 1994.

[11] Falkoner DS and Mackay TFC. *Introduction to Quantitative Genetics.* Pearson, London, 1996.

[12] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063) :1299–1320, 2005.

[13] International HapMap Consortium. A second generation of human haplotype map of over 3.1 million snps. *Nature*, 449 :851–861, 2007.

[14] International HapMap Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311) :52–58, 2010.

[15] 1000 Genomes Project Consortium. An integrated map of genetic variation from 1092 human genomes. *Nature*, 491(7422) :56–65, 2012.

[16] Pritchard JK, Przeworski M. Linkage disequilibrium in humans : Models and data. *Am. J. Hum. Genet*, 69 :1–14, 2001.

[17] Evans DM. *Analysis of Complex Disease Association Studies—A Practical Guide edited by Zeggini E and Morris A. Chapter 12 : Gene-Gene Interaction and Epistasis.* AP Elsevier, London, 2011.

[18] Bateson W. *Mendel's Principles of Heredity.* Cambridge University Press, Cambridge, 1909.

[19] Lehner B. Molecular mechanisms of epistasis within and between genes. *Trends in Genetics*, 27(8) :323–331, 2011.

[20] Moore JH. A global view of epistasis. *Nature Genetics*, 37(1) :13–14, 2005.

[21] Fisher RA. The correlation between relatives on the supposition of mendelian inheritance. *Philosophical Transactions of the Royal Society of Edinburg*, 52 :399–433, 1918.

[22] Everts RE, Rothuizen J, van Oost BA. Identification of a premature stop codon in the melanocyte stimulating hormone receptor gene (MC1R) in Labrador and Golden retrievers with yellow coat colour. *Anim. Genet*, 31 :194–199, 2000.

[23] Ochman H, Gerber AS, Hartl DL. Genetic application of an inverse polymerase chain reaction. *Genetics*, 120 :621–623, 1988.

[24] Ritchie MD, Hahn LW, Roodi N, Bailey R, Dupont WD, Parl FF, Moore JH. Multifator-dimensionality reduction reveals high order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet*, 69(1) :138–147, 2001.

[25] Evans *et al.* 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, 2011.

[26] Van Lishout F. An efficient and flexible software tool for genome wide association interactions studies. PhD thesis, Liege University, 2016.

[27] Van Lishout F, Gadaleta F, Moore JH, Wehenkel L, Van Steen K. gammaMAXT : a fast multiple-testing correction algorithm. *BioData Mining*, 8(36) :1–15, 2015.

[28] Westfall P, Young S. *Resampling-based Multiple Testing : Examples and Methods for P-value Adjustment.* John Wiley & Sons, 1993.

[29] Pollard K, Van der Laan M. Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference*, 125 :85–100, 2004.

[30] Shang J, Zhang J, Lei X, Zhao W, Dong y. EpiSIM : simulation of multiple epistasis, linkage disequilibrium patterns and haplotype blocks for genome-wide interaction analysis. *Genes Genom*, 35 :305–316, 2013.

[31] Peng B, Kimmel M, Amos CI. *Forward-time population genetics simulations—Methods, implementation, and applications.* Wiley-Blackwell, 2012.

[32] Peng B, Kimmel M. simuPOP : a forward-time population genetics simulation environment. *Bioinformatics*, 21(18) :3686–3687, 2005.

[33] Grady BJ, Torstenson ES, Ritchie MD. The effects of linkage disequilibrium in large scale datasets for MDR. *BioData Mining*, 4(1) :1–13, 2011.

[34] Bush WS, Dudek SM, Ritchie MD. Biofilter : a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac Symp Biocomput*, pages 368–379, 2009.

[35] Malaise M. Lectures Notes in Introduction to Rumatology - Medical School First Master. Liege University, 2016.

[36] Abbas AK, Lichtman AH, Pillai S. *Basic Immunology—Functions and Disorders of the Immune System.* Elsevier Saunders 4th edition, 2014.

[37] Cortes *et al.* Major histocompatibility complex associations of ankylosing spondylitis are complex and involve further epistasis with ERAP1. *Nat. Commun.*, 6 :7146 doi :10.1038/ncomms8146, 2015.

[38] Delves PJ, Martin SJ, Burton DR, Roitt IM. *Roitt's Essential Immunology.* Wiley-Blackwell 12th ed, London, 2012.

[39] Bessonov K, Gusareva ES, Van Steen K. A cautionary note on the impact of protocol changes for genome-wide association SNP x SNP interaction studies : an example on ankylosing spondylitis. *Hum. Genet*, 134 :761–773, 2015.

[40] Barrett JC, Fry B, Maller J, Daly MJ. Haploview : analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2) :263–265, 2005.

[41] Su Z, Marchini J, Donelly P. HAPGEN2 : simulation of multiple disease SNPs. *Bioinformatics*, 27(16) :2304–2305, 2011.

[42] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK : A tool Set for Whole-Genome Association and Population-Based Linkage Analysis. *Am. J. Hum. Genet*, 81 :559–575, 2007.