GENETIC ASSOCIATION ANALYSIS OF COMPLEX DISEASES THROUGH INFORMATION THEORETIC METRICS AND LINEAR PLEIOTROPY

Dissertation submitted for the degree of
Doctor of Philosophy in Biomedical Engineering

PhD Student: Helena Brunel Montaner
PhD Advisor: Dr. Alexandre Perera i Lluna

Universitat Politècnica de Catalunya
Programa de Doctorat en Enginyeria Biomèdica

Barcelona, 2013

Abstract

The main goal of this thesis was to help in the identification of genetic variants that are responsible for complex traits, combining both linear and nonlinear approaches.

First, two one-locus approaches were proposed. The first one defined and characterized a novel nonlinear test of genetic association, based on the mutual information measure. This test takes into account the genetic structure of the population. It was applied to the GAW17 dataset and compared to the standard linear test of association. Since the solution of the GAW17 simulation model was known, this study served to characterize the performance of the proposed nonlinear methods in comparison to the linear one. The proposed nonlinear test was able to recover the results obtained with linear methods but also detected an additional SNP in a gene related with the phenotype. In addition, the performance of both tests in terms of their accuracy in classification (AUC) was similar. In contrast, the second approach was an exploratory study on the relationship between SNP variability among species and SNP association with disease, at different genetic regions. Two sets of SNPs were compared, one containing deleterious SNPs and the other defined by neutral SNPs. Both sets were stratified depending on the region where the polymorphisms were located, a feature that may have influenced their conservation across species. It was observed that, for most functional regions, SNPs associated to diseases tend to be significantly less variable across species than neutral SNPs.

Second, a novel nonlinear methodology for multiloci genetic association was proposed with the goal of detecting association between combinations of SNPs and a phenotype. The proposed method was based on the mutual information of statistical significance, called MISS. This approach was compared with MLR, the standard linear method used for genetic association based on multiple linear regressions. Both were applied as a relevance criterion of a new multi-solution floating feature selection algorithm (MSSFFS), proposed in the context of multi-loci genetic association for complex diseases. Both were also compared with MECPM, an algorithm for searching predictive multi-loci interactions with a criterion of maximum entropy. The three methods were tested on the SNPs of the F7 gene, and the FVII levels in blood, with the data from the GAIT project. The proposed nonlinear method (MISS) improved the results of traditional genetic association methods, detecting new SNP-SNP interactions. Most of the obtained sets of SNPs were in concordance with the functional results found in the literature where the obtained SNPs have been described as functional elements correlated with the phenotype.

Third, a linear methodological framework for the simultaneous study of several phenotypes was proposed. The methodology consisted in building new phenotypic variables, named metaphenotypes, that capture the joint activity of sets of phenotypes involved in a metabolic pathway. These new variables were used in further association tests with the aim of identifying genetic elements related with the underlying biological process as a whole. As a practical implementation, the methodology was applied to the GAIT project dataset with the aim of identifying genetic markers that could be related to the coagulation process as a whole and thus to thrombosis. Three mathematical models were used for the definition of metaphenotypes, corresponding to one PCA and two ICA models. Using this novel approach, already known associations were retrieved but also new candidates were proposed as regulatory genes with a global effect on the coagulation pathway as a whole.

Acknowledgements

Writing these words is one of the most emotional moments in my life, because it closes an intense and meaningful period. Such is the gratitude that I feel that I would never have enough space to express it, but I'll try to convey it as concise as possible. I hope everyone will feel appreciated as it deserves and if there are some feelings that I have not had time to organize, I hope to get them somehow. I feel extremely lucky that I enjoyed the constant support of many people during these years so that this thesis is a prize that deserves to be shared with all who have been part of it.

Before expressing my gratitude to all those who have offered me help, support and assistance throughout the development of this thesis, let me skip the protocol for a more personal appreciation, without which I do not understand this section. For me, to finish this manuscript is also a proof that I am taking the second chance given to me. And I owe it to three main groups of people. First of all, I do not even know how to express my gratitude and admiration to the infinite generosity of an anonymous family which in a painful situation were able to donate the organs of a deceased relative. I also want to thank and acknowledge the work of the several medical teams that took care of me during the last ten years, In particular, thanks to the cardiologists that treated me before the surgery for keeping me able to carry a normal life for nine years, and to the Cardiac Transplant Unit of the Hospital de Sant Pau which with their excellence have made my recovery as easy and quick as possible. Finally I want to recognize the courage and bravery of my family who have been with me at all times. Their constant support and affection have been essential for me to overcome this process and without them I would never have been able to face it.

Having said that, I can now proceed to thank all the people who have contributed directly or indirectly to the development of this thesis. First of all I would like to express my gratitude to my supervisor, Dr. Alexandre Perera Lluna, for his guidance and support. These last six years working under his supervision have made me grow as a researcher and as a person. It has been a privilege to enjoy his scientific contributions and advices and to translate his exceptional ideas into my work. For me, it has been a great honor but also a big responsibility to be his first PhD student and I sincerely hope not having let him down. I am aware that the circumstances have often been unexpected, unpredictable and even awkward so that I am very grateful to him for his patience and his comprehension at all times. I would also like to thank him for his conviction on the value of this work and for encouraging me and helping me to get over my existential crises and to believe in the culmination of this manuscript. I would also acknowledge him for his concern with my employment and with my future as a researcher. So, thanks to Alex for being my supervisor and, not less important, for being a friend at the same time.

I want to express my gratitude to all the people in the ESAII department at UPC, from the administrative staff to the directors, for their warm welcome and support during this years. I feel proud to have become part of the *SISBIO* work team, composed of people who stand out for their humanity and kindness. A special thank to Prof. Josep Amat for introducing me to the department and to biomedical engineering with the final project of my degree in Mathematics. I also thank Dr. Montserrat Vallverdú for introducing me into the world of biomedical signals, and for teaching me in my first research work. It was a very instructive period for me and I have tried to retain her teachings in my work in this thesis. Finally, I would like to thank Prof. Pere Caminal for his constant support and interest in my work and in my health but above all for conferring me the opportunity of carrying out this thesis at his research group with a grant from the IBEC Institute. The financial support of the IBEC (Institut de Bioenginyeria de Catalunya) has been essential for me to be able to devote time and energy to this thesis.

I would like to thank Dr. José Manuel Soria and all his team at the *Unitat de Genòmica de Malalties Complexes* at Hospital de Sant Pau for his warm welcome at his research group and for his priceless comprehension and patience. I am very grateful to him for waiting for me during my convalescence and I hope I will have time to compensate him with a productive work at his research group. I would also like to thank him and Angel Martínez-Pérez for their straight collaboration in the studies leading to chapter 7 and 8. I would also thank the remaining members of the team for their support in the last months. I specially thank Yorgos for his guidance and support and for correcting parts of this my manuscript.

Thanks also to Dr. Joan-Josep Gallardo-Chacón for his contribution in chapter 6. I also want to thank him for his guidance in biology.

I would like to thank all the remaining colleagues in the *g-sisbio* group. It has been a pleasure to work with all of them and to receive their valuable comments and viewpoints about my work which in many occasions have helped to my work. I specially thank Raimon and Joan for their collaboration in both chapters 6 and 8. I also thank both of them, jointly with Francesc, who I consider my scientific brothers, for their constant emotional support and also for helping me in many computer daily tasks, even when that interrupted their own jobs. Thanks also for their constant predisposition for laughing and joking, sharing worries and dreams. This helped to create the most pleasant work ambiance possible. Thank you also for taking care of me during the different travels to conferences we made together.

I would like to extend this last acknowledgement to my remaining friends at UPC and IBEC and to all the PhD students that have been part of us during these years. Thank to all of you for the great moments we have had together, the original debates at lunchtime, the ping-pong tournaments, the dinners at Esquinica and the different Christmas and birthday parties. As a special person said, " The goal is not as important as the way we take to get it". Thanks to all of you for making my way so pleasant. For me, you are the most valuable present I get from this period of time. Nowadays you are one of my closest friends and I feel really lucky for it.

I would like to thank all the other people around me, mostly family and friends, for their constant interest and curiosity to understand what it means to do a PhD and what is my particular PhD about. They have been an important support during these years, listening to all my worries and insecurities and always trying to understand me and give valuable advices for me to improve and overcome the vicissitudes of life and particularly of the PhD. A special thank to Anna Serraima for her valuable guidance in the last years.

I would like to thank all my teachers from the elementary school to the university. They all had an influence in my academic formation in a way or another but in particular, a special thank to Mr Cassam for arousing my curiosity and my interest in science and more precisely in Mathematics and for giving me an original nickname with a logarithm (lna).

I want to thank my close friends for their encouragement in both good and bad times and for all the good moments and the conversations about life we have shared. I also want to apologize if they have needed me and I have not responded because I was too absorbed in my things.

I would like to thank the entire López and Muñoz family for their affection, and for making me feel like part of the family since the day we met. I would also thank their constant support and willingness and their hospitality and generosity giving to me and to Pedro the opportunity to have our own home during this time.

As I have already mentioned, I am infinitely grateful to my entire family for their unconditional love and support. Without them I would not even have been able to undertake this trip. Thanks to my grandmother for introducing me to Prof. Josep Amat, and for her constant support. Thanks also to my godmother for her goodness and for her valuable advices and jointly with my remaining uncles and aunts for loving me and protecting me as if I were their daughter since I was a kid and for teaching me to watch life with positive eyes and with humor. Thank you also to my cousins for being there, and especially to Anna for the fantastic coffee-breaks at the faculty. Moreover, I would like to specially thank my parents, my sisters and Pedro for standing at the front line and for adapting their lives to my needs.

To my sisters, for always being there. They have always been a reference for me. I admire their "savoir-faire" and I am grateful to learn from them everyday. Their unconditional support and complicity have been essential in the critical moments and helped me to move forward. I want to specially thank Clara for her outstanding work in the design of the cover of this book.

To my father and his wife, for their support and for being present despite the distance that separates us. I would like to thank my father for his generosity and for his constant interest in the progress of my PhD. I also thank him for being an example of dedication and love for the job.

To my mother, who has been always in the lead and who I specially dedicate this thesis because she pushed me to take this option and she always encouraged me and trusted in my possibilities for doing it well. Her total devotion looking for the best for me has brought me here. I am also grateful for all her teachings and admirable qualities that inspire me everyday. She has always been a great role model for me.

The last, but certainly no the least, I want to thank Pedro, for being the rock to steady in. Since we met, my life has take a different meaning. Get to know each other, love each other and build a life together has been and remains being the easiest and most beautiful experience in my life. Knowing that he is standing by my side no matter when or what is priceless. I would like to thank him for being my light in the darkest moments and for believing in me even when I was not able to do it myself. I also thank him for his unlimited patience supporting my mood changes and his willingness for listening at my presentations' testings and reading and correcting my writings. I thing that there are no words to express how much I love him, I admire him and I thank him.

To all of you, I love you and thank you so much for everything. Anyway, I hope you already knew it.

Preface

During the last decade, Genome-Wide Association Studies (GWAS) have become a standard practice in the genetic study of human diseases. Until 2005, the field was mainly a labyrinth, given the bare knowledge on genetic variants and genetic risk factors along the human genome. However, with the completion of the Human Genome Project in 2003 and the International HapMap Project in 2005, researchers now have a set of tools that make it possible to find the genetic contributions to common diseases. The first GWAS was reported in 2005, investigating age-related macular degeneration [151]. It compared 96 patients with 50 healthy controls, and identified 2 SNPs with significantly different allelic frequencies between the two groups. However, the start of the era of GWAS is marked with another landmark publication published in *Nature* in 2007, which has been considered the first major GWAS [322]. This study, carried out by the Wellcome Trust Case Control Consortium, was the first large and well-designed GWAS, testing around 400000 SNPs in 14000 cases of seven common diseases (coronary heart disease, type 1 diabetes, type 2 diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder, and hypertension) and 3000 shared controls. It was successful in uncovering many new genes underlying these diseases. Since then, a large number of GWAS have been published, which are now reported in several GWAS databases. GWAS have rapidly grown in scale and complexity, with studies now looking at over a million genetic markers in cohorts approaching hundreds of thousands of individuals. Nowadays, over 2000 loci have been reported to be significantly and robustly related with disease risk. However, many discoveries have been detected as spurious due to various failures on the experimental design. This has lead to a deep discussion between researchers on the success of GWAS.

One of the main hopes of GWAS was that, as for Mendelian diseases, it would be able to identify genetic variants involved in complex diseases. So far, it hasn't really happened. GWAS represent a great investment which lead to disappointing results. In his review published in the American Journal of Human Genetics in January 2012, Peter Visscher estimates that around 500 thousands chips of SNPs have been necessary for carrying out the GWAS published to date at, on average, 500 dollars per chip, which lead to a total sum of 250 million dollars invested for GWAS research in the past decade [308]. If there are a total of around 2000 discoveries, it indicates that each discovery has cost 125 thousands dollars. Thus it seems that the number of findings is not in correspondence with the investment.

Moreover, the discordance between SNP discoveries obtained from different GWAS for the same disease suggests that false positives exist in such results. Actually, this is one of the major problems of GWAS. This is the reason why, nowadays, the findings from a GWAS are viewed as a preliminary prioritization list of candidate relevant SNPs. This list is available for further analysis using statistical tools that accumulate evidence of genetic association. These secondary analyses are very likely to provide a strongest prioritization of the results. In their review published in 2010 in the American Journal of Human Genetics, Cantor et al. highlight three main strategies for prioritizing SNPs obtained with GWAS [43]. A frequent method for prioritizing GWAS results is to compare several GWAS via a *meta-analysis*. Meta-analysis is a standard and validated approach consisting in the statistical analysis of a collection of analytic results for the purpose of integrating the findings of each study [72]. In the context of GWAS, meta-analyses combine comparable test statistics across independent studies of the same phenotype, weighting them by the confidence in the study-specific results. Note that this practice is computationally demanding since it supposes to repeat the analyses several times. The second approach for SNP prioritization is to search for epistasis within a single GWAS study, in order to find stronger results, only revealed when genes interact. As it is addressed in chapter 7, this strategy consists on revealing combinations of SNPs significantly associated with the phenotype under study, that should not be individually relevant. This prioritization alternative consider that one or more significant interaction provides additional evidence of association. The third prioritization approach takes profit on the information from genetic pathways. This set of methods integrates the results of a GWAS and the genes in a known molecular pathway to test whether the pathway is associated with the disorder. As described in chapter 8, this approach strongly depends on the definition of the pathway selected for providing a biological instrument for enriching GWAS results [120].

One of the most common and strong cause of false positive findings is that a massive number of hypothesis tests are conducted simultaneously. This increases the type I error that occurs when statistical tests are used repeatedly. There has been no clear consensus about how this problem of multiple testing should be dealt with. However, it has been proven that the Bonferroni correction is too much conservative and is not always adequate, since true disease genetic markers with small effects would be hidden under the significance level and then lost within the background noise. As a consequence of this, the genetic variants identified through GWAS only explain a small fraction of the overall genetic variance of disease risk. As McClellan and King pointed out, it is now assumed that common risk variants fail to explain the vast majority of genetic heritability for any human disease [201]. It is suggested that this failure is the result of a wrong initial hypothesis underlying GWAS. GWAS rely heavily on the "common disease, common variant" (CDCV) assumption, which states that the genetic risk for common disease is mostly attributable to a relatively small number of common genetic variants. However this hypothesis was certainly stated for convenience considering that the available catalog of human genetic variation (built up by efforts such as the HapMap project) is largely restricted to common variants. The first and most obvious candidate to explain the 'missing' heritability are rare or low frequency variants, since they are not sufficiently frequent to be captured by current GWA genotyping arrays. In a discussion conducted in the *New England Journal Of Medicine* in 2009 about the success or failure of GWAS, David Goldstein presented a hard criticism to GWAS, suggesting that a non-trivial fraction of the genetic risk of common diseases is the result of rare variants and that current GWAS technologies are not able to unravel the proportion of variation due to rare variants[95]. The fundamental problem is that genotyping chips are not always able to tag rare variation. This could be solved with higher-density SNP chips incorporating variants with lower frequencies. A second candidate to account for a substantial fraction of human genetic variation could be structural variation. In the last five years, widespread, large-scale insertions and deletions of DNA, known as copy number variations (CNVs), have been identified even in healthy genomes. These variants have been shown to play a role in variation in human gene expression and in human evolution. However, the study of CNVs is still in a preliminary stage, since current sequencing technology only detect a small proportion of CNVs. High-resolution arrays, containing millions of probes, can be used to explore CNVs in some areas of the genome. For the complete detection of CNVs from patients and controls, whole-genome sequencing, preferably using methods with much longer read lengths will be required.

Definitely, the solution for both alternatives will be large-scale next generation sequencing projects like the 1000 Genomes Project, which will provide a complete catalog of every variant in the genomes of both patients and controls. The problem will not lie so much on the sequencing itself but on the interpretation. Then, new analytical techniques will be required to convert the data into useful information.

In contrast with the criticisms against GWAS, Visscher makes a determined defense in favor of GWAS advocating that even if they have failed in explaining the genetic variation underlying human diseases in their totality, at least they have lead to new discoveries about genes and pathways involved in complex disorders providing new biological insights [308]. In this direction, Hardy and Singleton support GWAS with a kind comparison. If the genetics of complex diseases is comparable to a jigsaw puzzle, we have put the edges and corners in place thanks to GWAS and now have a framework to perform an extended and deep analysis to decipher the genetic architecture of complex disorders. And if this comparison was real, this thesis would pretend to be a small piece of this puzzle, which major aim is to provide a methodological environment for novel techniques to genetic association studies.

Contents

List of Tables

List of Figures

Part I

Introduction and State of the art

Chapter 1

Introduction

1.1 Thesis Introduction

This thesis is outlined in the field of the genetic study of complex diseases. With the completion of the Human Genome Project in 2003 and the International HapMap Project in 2005, researchers now have a set of research tools that make it possible to find the genetic contributions to common diseases. More recently, even more complete repositories such as the 1000 Genomes project, including rare variants, are becoming very useful for the characterization of genetic variants correlated to diseases. These tools include high-throughput genotyping technologies and computerized databases that contain the reference human genome sequence, a map of human genetic variation, and a set of algorithms for the analysis of whole-genome samples for genetic variations that contribute to the onset of certain diseases.

In particular, genome-wide association studies (GWAS) involve rapidly scanning genetic markers across the whole genome in a population to find genetic variations associated with a particular disease. The results of GWAS are useful for researchers to develop better strategies to detect, treat and prevent the disease. GWAS have been particularly useful in deciphering the genetic architecture of Mendelian disorders, caused by mutations in a single locus. However, complex diseases are not controlled by only one locus but they are influenced by the interaction of multiple loci. In this case, every locus should only have a minor or moderate individual contribution but should need to interact with each other in order to exert their influence. Moreover, complex diseases tend to involve greater difficulties in phenotype definition, so that several traits may be measured for a disorder or its risk factors. Sometimes, some of these phenotypes are correlated between them so that there may be genetic variants affecting several of these traits simultaneously. These pleiotropic effects are also of great importance in the genetic study of complex diseases.

The standard procedure for studying genetic association consists in exploring the linear correlations and interactions between genetic variants and physiological traits. Linear regression models are commonly used for modelling the relationship between predictors (genetic loci) and a clinical outcome (a disease or a related trait). Thus, the behavior of these techniques will strongly depend on the degree of nonlinearity in the mapping between loci and traits. Nonlinearities can arise when different mutations lead to the same trait (locus heterogeneity), when mutations are reached by certain models of dominance, when phenotypes are environmentally determined, without a genetic compound (phenocopy) or when the dependence between genetic loci, environment and traits is not governed by linear patterns. Therefore, exploring the nonlinear correlations between genetic variants and clinical traits may provide useful information for understanding the genetic structure of complex diseases.

1.2 Thesis Objectives

The main goal of this thesis is to help in the identification of genetic variants that are responsible for complex traits such as cardiovascular diseases. In particular this thesis aims to propose a nonlinear methodology for genetic association studies of complex diseases.

In order to achieve this ultimate objective, it can be detached in a set of underlying goals which would serve as a road map for the development of this thesis, as described below:

- The first goal of this thesis is to develop a nonlinear test of association between one genetic variant and one trait or disease. In particular this thesis aims to apply nonlinear correlation measures from information theory. In particular, this includes exploring the properties of information theoretic measures for their suitable application to genetic data.

- An adjacent goal of this thesis is to apply the nonlinear methodology to genetically stratified populations. This implies taking into account the genetic structure o the population when measuring genetic association between genetic loci and diseases.

- Another goal of this thesis is to extend the nonlinear test for multi-loci association, taking into account the effect of the interactions between several genetic variants in the prevalence of diseases. This implies developing algorithms for the search of multi-loci interactions and applying them with a nonlinear association criterion.

- In order to consider all the possible ways of interaction between genetic loci and complex diseases, another purpose of this thesis is to capture the pleiotropic effects of genetic variants on several traits involved in complex diseases. This involves analyzing multiple phenotypes simultaneously and studying the association between genetic variants and their common variability.

- In order to validate the proposed techniques, another goal of this thesis is to compare them with the traditional linear techniques applied on real data.

1.3 Thesis Outline

This thesis is organized into 9 chapters arranged into 3 parts. The aim of the chapters in the first part is to provide an introduction to the subject of this thesis, to describe the state of the art in this field and to introduce the basic concepts necessary for a proper understanding of this document. In particular, chapter 1 is an introduction of this thesis, with the motivation of the work, its main objectives and the description of its outline, in which these words are found. Chapter 2 introduces basic terms and concepts in genetics, setting the biological bases of this work. Chapter 3 offers a review of methods and algorithms for studying the genetic association between genetic variants and diseases. Finally, chapter 4 introduces the concepts and quantities in information theory as well as a review of the application of these measures in Bioinformatics and more specifically to the genetic analysis of complex diseases. The second part of this manuscript contains four chapters concerning the original contributions of this thesis. In particular, chapter 5 contains a descriptive characterization of the datasets used during this thesis. Chapter 6 focus on studying the relationship between a single genetic marker and a phenotype. This chapter contains an exploratory study on sequence variability characteristics of markers related to disease, and proposes a nonlinear methodology for one-locus genetic association based on information theory. Chapter 7 describes a novel methodology for identifying the multi-loci genetic association between several markers and one phenotype also using the mutual information measure. Chapter 8 studies the association between single genetic variants and several phenotypes involved in complex diseases. Finally, the third part of this document contains the conclusions of this thesis, containing a list of the outcome and contributions provided by this thesis, as well as recommendations for future work.

Chapter 2

Background

Genetics is the science that studies variations between individuals and their inheritance. It has attracted the attention of scientists and philosophers since ancient times. Nowadays, it has emerged into the world of medicine trying to provide new tools for diagnosis, monitoring and treatment of diseases. The different disciplines arising from or related to genetics converge to the common goal of identifying genetic variants responsible for diseases. This chapter aims to provide a brief historical overview of the science of genetics as well as to introduce some concepts and terms concerning genetics for a better understanding of this thesis.

2.1 From Hippocrates to the future

Although genetics is a relatively young discipline, the idea which it is based on –inheritance– comes from afar. The Greek philosophers already proposed the first theories about inheritance. Hippocrates theory comes near the later ideas that Darwin called "pangenesis" which describes that each part of the body produce hereditary material called "gemmules" collected by gametes. Aristotle discarded this hypothesis by suggesting that individuals were made by something he called 'the substance', which was found in women, and the 'form', which came from men. Aristotle believed that living things gradually changed from plants through animals, ending in the highest form, humanity. This was the first time the idea of evolution had been recorded.

However, the history of classical genetics started in the late 1800's and the early 1900's. In 1866, Gregor Johann Mendel, an Austrian Augustinian monk, published his study on pea plants, where he showed that the inheritance of certain traits followed particular laws [204]. These are now known as the Mendelian laws of inheritance. Mendel's work was not given any attention in the scientific community until the 20th century, when the bases of genetic science were established. Contemporarily with Mendel, Darwin's theory of evolution by natural selection motivated discussions about modes of inheritance. Even if Darwin's own theory of heredity, pangenesis, was not well accepted, it has served different geneticists around the world to re-discover Mendel's theories about inheritance. In 1900 three European botanists, Hugo De Vries, Carl Correns and Erik Tschermak, published the results of their respective experiments carried out in the 1890s that corroborate Mendel's experimental results and his conclusions. In particular, De Vries asserted that "inheritance of specific traits in organisms comes in particles" and defined this particles with the term "pangenes" [70], what we now know as *genes*. Another important contribution of De Vries' work was the introduction of the term "mutation" in his "Mutation Theory" [71]. He stated that new species arise from the preexisting ones in a single generation by a sudden appearance of marked discontinuous and inheritable variations that he called *mutations*. It was not until 1906, when Bateson introduced the term genetics to define the study of biological inheritance and the science of variation [19]. All the theories about inheritance and biological variation proposed in the first decade of the 20th century presented some contradictions in many aspects. In 1918, Ronald Fisher, well known for his contributions to statistics, initiated a new movement consisting on an unified theory that integrates all these ideas [84]. In particular, Fisher contributed with his statistical knowledge for laying the foundations of population genetics, showing that the observed variation between characters could be the result of the action of several mutations and that natural selection could change gene frequencies in a population.

Simultaneously, research in biochemistry was progressing in such a way that in the second half of the 20th century, genetic research was mainly redirected to what is now called the DNA era. In 1869, the biochemist Friedrich Miescher isolated a new substance from the nuclei of white blood cells. He called it nuclein and now it is known as nucleic acid or more commonly DNA (DeoxyriboNucleic Acid). DNA is found in each cell of an organism and is organized into long structures called *chromosomes*. Diploid organisms (e.g. humans or most mammals) have two homologous copies of each chromosome, one inherited from the mother and one from the father, so that the number of chromosomes is counted in pairs. For example, human cells contain 23 pairs of chromosomes, 22 pairs of autosomes and one pair of sex chromosomes, giving a total of 46 chromosomes per cell. In 1919, Phoebus Levene suggested that DNA basic building blocks are composed by one of the four nucleotides (adenine, guanine, cytosine and thymine) linked to a sugar (deoxyribose). He also suggested that these units are linked through a phosphate group, forming a chain of bases repeated in a fixed order. However, the structure of this chain was not resolved until 1953, when James Watson and Francis Crick proposed a double-helix model for the structure of DNA [320] (Figure 2.1).

In 1958, Francis Crick articulated the central dogma of molecular biology, a framework for the understanding of the relationship between DNA

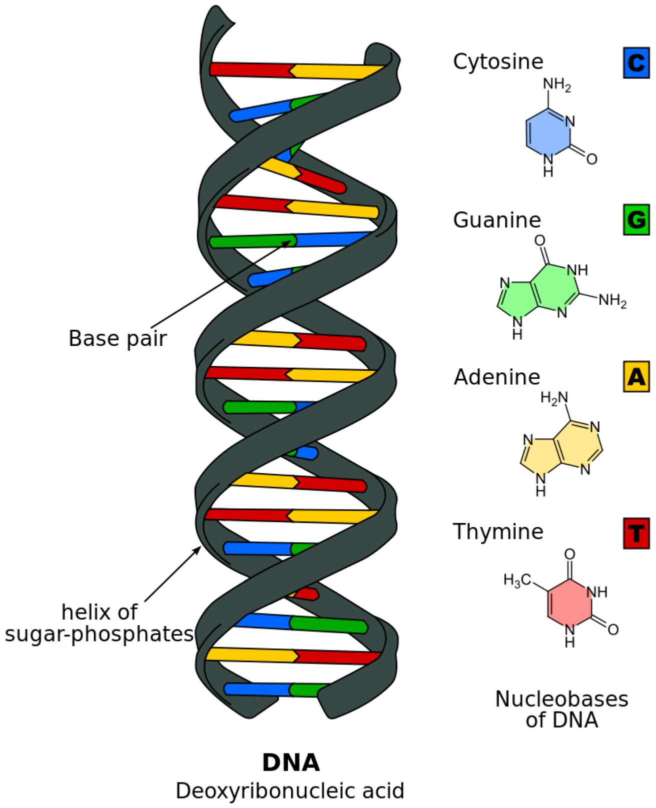


Figure 2.1: The DNA Structure [Image from Wikimedia Commons].

and proteins [64]. In the following years, scientists tried to understand how DNA controls the protein production. This process is divided into two steps: (1) the transcription by which the information contained in the DNA is transferred to a complementary copy of the DNA molecule called messenger RNA (RiboNucleic Acid), and (2) the translation, where the nucleotide sequence of a messenger RNA is translated to an amino acid sequence that forms a protein (Figure 2.2). Nowadays, a gene is defined as the portion of unit of heredity residing on the DNA that codes for a protein or for any RNA chain that has a function in the organism.

The discovery of the DNA structure and function encouraged researchers to decipher the entire DNA sequence (genome) of different species. This discipline is called *genomics* and started in 1972 when Walter Fiers determined the sequence of a gene: the gene for Bacteriophage MS2 coat protein and posteriorly completed the entire genome for this organism [82]. In 1996, the entire DNA sequence of *Saccharomyces Cerevisiae* was the first eukaryote genome sequence to be released. In 2003, after more than a decade of research, the human genome sequencing was completed by an international consortium of research centers around the world led by the U.S. Depart-

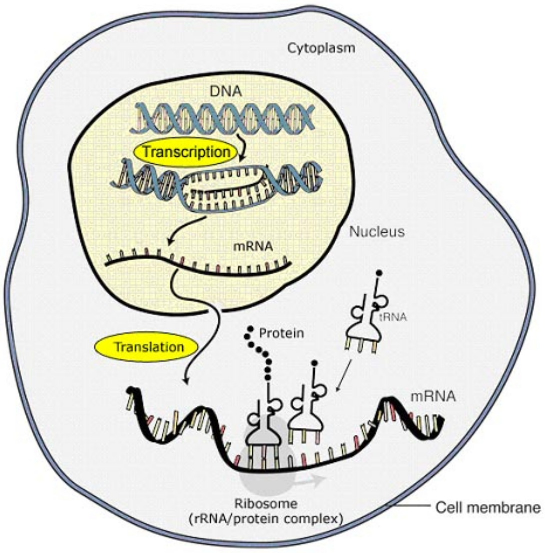


Figure 2.2: The protein synthesis [Image adapted from the National Human Genome Research Institute].

ment of Energy and the National Institute of Health. This partnership was formally initiated in 1990 with the creation of the Human Genome Project (HGP). The project goals were (1) to determine the sequence of the 3 billion bases that constitute the complete human genome, (2) to identify the 20.000 to 25.000 human genes and (3) to make them accessible for further biological studies. The project was scheduled for 15 years, but fast technological advances accelerated the process so that a first draft was published in 2001 [59] and in 2003 the human genome sequence was officially completed. It has been one of the most important milestone for medicine and biology.

The publication of numerous genome sequences, including that for human, has driven the biosciences into the post-genomic era. In addition to identifying all the genes in genomes, it is crucial to store and distribute the information in databases. Advanced computer-based methods are required for making sense of the mountains of biological data. Bioinformatics (or computational biology) is the field that handles the data. The term *bioinformatics* is used for almost all computer applications in biological sciences, and it was originally coined in the mid-1980s for the analysis of sequence data [15]. One of the first and most important application in Bioinformatics has been the development of the BLAST (Basic Local Alignment Search Tool) program, with the aim of comparing a sequence against all the sequences of a database in a reasonable time [7]. The problem was already of a great interest in the 80's, when databases of sequences were much smaller than now, but computers were slower than now as well.

Since then, biological sequence data is accumulating rapidly. Figure 2.3 illustrates this showing the growth of GenBank, the most important sequence database. It is observed that the period of accelerated growth coincides with the completion of the HGP, setting the bases for the development of high-throughput sequencing technologies [20], such as DNA microarrays.

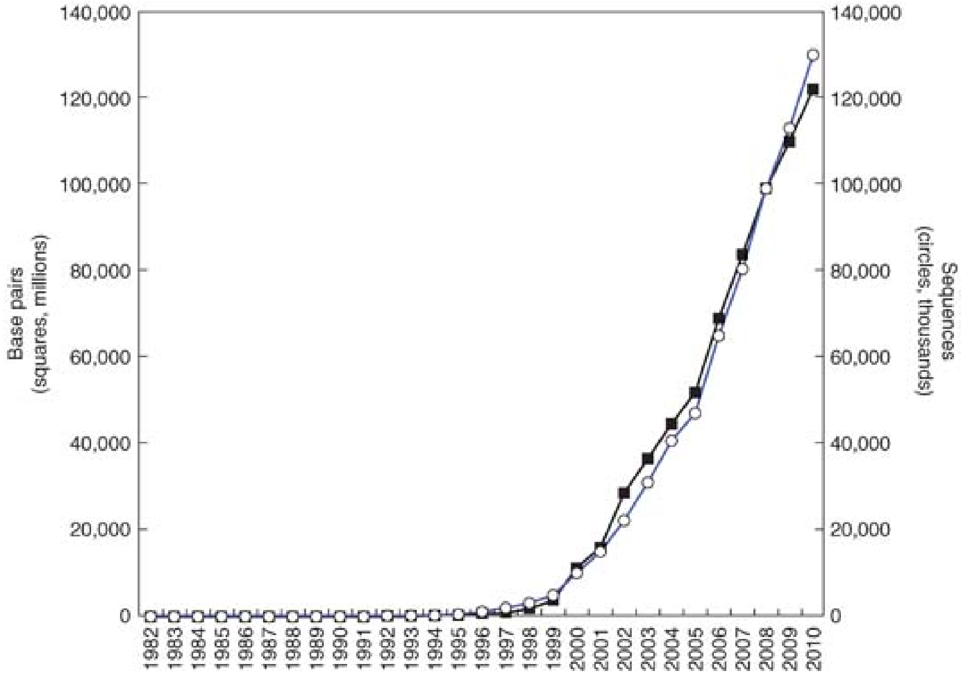


Figure 2.3: Growth of GenBank (from [20]).

DNA microarrays consist in chips with an array of submicroscopic spots containing a specific DNA sequence. The spots are analyzed using techniques based on DNA hybridization with fluorescence microscopy [252]. More recently, Next-Generation Sequencing (NGS) platforms have become an additional alternative to microarrays. These technologies, also referred to as RNA-seq, consist in applying sequencing technologies to sequences of complementary DNA (cDNA) in order to to get information about a sample's RNA content [318]. Thanks to these advances, the cost of sequencing an entire genome has significantly decreased. Although Moore's Law is reserved to computing hardware, predicting that computing power would double every two years, DNA sequencing costs have followed a similar pattern for many years, approximately halving each two years, as shown in Figure 2.4. Nowadays, the cost of a sequencing a genome with NGS is around 6000 dollars and it is predicted that it will presumably be around 1000 dollars per genome in a not too distant future.

These improvements have made possible the development of initiatives

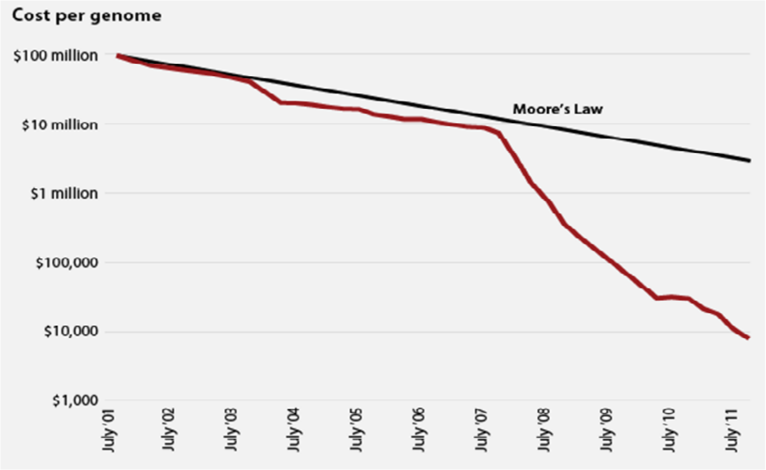


Figure 2.4: Evolution of the cost of sequencing a genome [Image from National Human Genome Research Institute].

such as the 1000 genomes project. The 1000 Genomes Project is an international collaboration to sequence 1000 individuals in an effort to produce the most complete catalog of human genetic variation to date. Building on the International HapMap Project, the 1000 Genomes Project will utilize new sequencing technologies to catalog genetic variants. By creating an important scientific resource, the Project will help to understand the complex relationship between genetic variation and human health and disease. It represents a major step forward on the road to personalized genomic medicine. As pointed out by *Rossbach and Garcia*, the use of next-generation sequencing technologies will improve the quality of life and efficiency of health care delivered to patients [255]. Nowadays, the value of genome-based approaches in personalized medicine is not fully explored. Further research is required to overcome the challenges associated with the translation of genomic knowledge into clinical decision-making in terms of patient data, testing procedures, algorithm development and the use of such information in therapy planning. Only a few personalized-medicine tests have achieved high levels of clinical adoption to date and are mostly in the field of oncology.

2.2 Genotypes

One of the most striking discoveries originated from the HGP is that any two human beings are 99.9% identical in their DNA sequence [81]. The remaining and most interesting 0.1% of the DNA sequence contains the genetic variants responsible for individual characteristics such as physical appearance or susceptibility to disease [250]. The main goal of current genetic research is to find and understand the relationship between the variability found in the genome of an individual (genotypic variability) and the differences observed between individual characteristics (phenotypic variability).

2.2.1 What is a genotype

Each cell of a living being contains the genetic information that characterize him. This information is stored as DNA, a 3 billion-long sequence of nucleotides (A, T, C and G for Adenine, Thymine, Cytosine and Guanine). This sequence is a code that contains all the instructions necessary for the building and maintaining of a creature. Diploid organisms have two copies of any gene, one inherited from each of its parents, that can present different forms, called *alleles* [1]. Examples of genotypes at a particular gene and at a specific locus of a gene are shown in Figure 2.5.

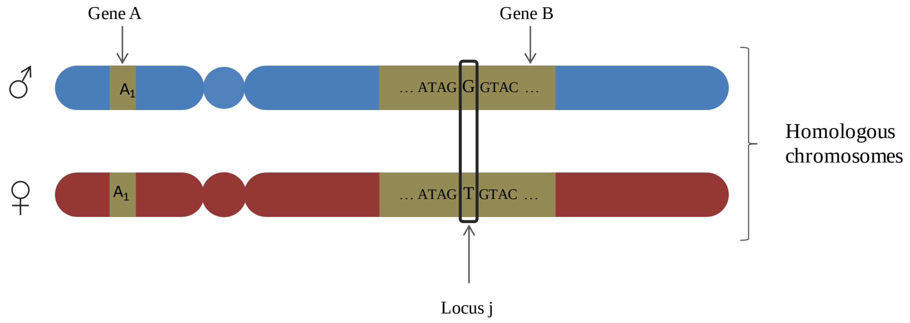


Figure 2.5: Two examples of genotypes. At gene A, the individual shows the genotype A_1/A_1 because he receives the allele A_1 of both of his parents, whereas at locus j of gene B, this individual shows the genotype G/T because he receives the allele G from his father and the allele T from his mother.

2.2.2 Genetic variants

Sometimes mutations occur. Mutations refer to any variant in the genetic sequence among individuals. They may be caused by radiation, viruses as well as errors that occur during DNA replication, the process of copy of the DNA occurring in the synthesis phase of the cell cycle. A portion of all genetic variation is functionally neutral in that they do not produce observable differences between individuals. When mutations cause changes between individuals, these mutations can be passed to offspring. Due to evolutive pressure, individuals with certain variants may survive and reproduce more than individuals with other variants. These mutations are conserved from

[1] With the purpose of simplifying the notations along this document, the two possible alleles of an individual for a particular gene or locus are noted A_1 and A_2. The combination of alleles that an individual carries is called a *genotype* and it is represented with a bi-valued symbol (e.g. A_1/A_1).

generation to generation and when they achieve a frequency greater than 1% in the population, they are called *polymorphisms*. Moreover scientists consider that genetic variants with a frequency higher than a 5% of the population are *common variants* whereas those that appear in less than 5% of the population are called *rare variants* .

There are several types of genetic variants, studied by researchers from the beginning of the HGP. Small-scale mutations are those that affect a small region of one or few nucleotides. Among them, the first type of genetic markers used for studying human diseases were variable numbers of short DNA sequences repeated in tandem, also called DNA satellites. Differences in individual bases are the most common type of genetic variation. These genetic differences are known as Single Nucleotide Polymorphisms, or SNPs. On the other hand, large scale mutations have been described during the last few years [148], such as Copy Number Variants, or CNVs.

There are over 800 databases of human genetic variation of which only a few are most widely used for genetic studies [140]. There is a need to find a reliable, comprehensive, centralized and public resource on genetic variation. dbSNP is a repository of reference for genetic variation [274]. However, the HapMap project has also a central importance. Moreover there is a number of secondary databases aimed at characterizing variation within or across human populations.

Nowadays, dbSNP [274] is the major repository of genetic variants. It was created and hosted by the National Center for Biotechnology Information (Bethesda, USA). dbSNP is a reliable, comprehensive, centralized and public resource on genetic variation and it is integrated to other popular resources such as common genome browsers (NCBI, UCSC, EMBL) [221, 147, 127]. The database provides information about all the variations in the human genome such as their location within or around genes, their functional effects or their population allele frequencies in a variety of populations. Since 2003, researchers are constantly submitting genetic variants to this database that are likely to be related to phenotypic variations. In order to manage this constant increasing of the SNP data, dbSNP releases its content to the public in periodic *builds* that contain the information given by a run of the genome assembly and the annotation process of the set of products generated by that run. Each build is synchronized with a release of new genome assemblies for each organism and with the last build. SNPs uploaded in a build can be divided into two categories: submitted SNPs and reference SNPs. Nowadays (April 2013), the number of SNPs in the build 137 of dbSNP reaches the figure of around 50 million validated SNPs (rs SNPs) and near than 200 million submitted SNPs (ss SNPs).

The International HapMap Project is an organization that originally aimed to develop a haplotype map (HapMap) of the human genome [58]. The HapMap is a tool that allows researchers to find genes and genetic variations that affect health and disease. It began as an effort to survey allele frequencies among common human genetic variants and across worldwide populations, but now, it is a key resource for researchers to find genetic variants affecting health, disease and responses to drugs and environmental factors and it provides a critical platform of information for large-scale genetic association projects. The project has now progressed through 3 phases: Phase I (published in 2005), Phase II (published in 2007), and Phase III (released in 2009). Phase III is the current release of Hapmap. It contains more than one million SNPs genotype data generated for 1115 individuals from 11 worldwide populations and collected using two platforms (Illumina and Affymetrix). This genotype information is available for download or can be viewed through the HapMap or other browsers and within dbSNP records.

The 1000 Genomes Project is an underway initiative based on the successful model of the HapMap Project [307]. It aims to sequence more than one thousand individual human genomes including many HapMap samples [307]. This project began releasing data in 2009 and will provide an even deeper resource on human genetic variation, capturing common variation but also discovering more rare variation than ascertained in earlier HapMap phases.

Among the different types of genetic variants, the most commonly used for the study of human diseases are microsatellites, SNPs and CNVs, described in more detail in next sections.

2.2.2.1 Microsatellites

Satellites are classified as *minisatellites* or *microsatellites* according to their length. Minisatellites, also known as *VNTR* (Variable Number Tandem Repeat), are repeated sequences (tandems) of more than 10 nucleotides, whereas microsatellites, also known as STR (Short Tandem Repeat), are tandems of less than 10 nucleotides [298].

Microsatellites occur abundantly and at random over most eukaryotic genomes [115]. This polymorphism is sufficiently stable to be used in genetic analyses. Microsatellites are therefore ideal markers for constructing high-resolution genetic maps in order to identify susceptible loci involved in common genetic diseases. One of the most remarkable property of microsatellites is their heterozygosity. They are highly mutable markers with often 15 or more alleles in any given population,corresponding to the number of times the given sequence is repeated. They are usually characterized with a numerical representation.

Various microsatellite databases can be found, with different purposes. MICdb contains information on microsatellites occurring in coding and noncoding regions, such as their frequency, size and repeat sequence [287]. SilkSatDb also stores the polymorphism status of different microsatellite loci [239]. Satellog database catalogues triplet repeats associated with human disorders [208]. The database named as EuMicroSatdb (Eukaryotic MicroSatellite database) provides a more generic collection of whole genome eukaryotic microsatellite data. It stores both simple and compound microsatellites from 31 eukaryotic genomes assembled by chromosomes [5].

2.2.2.2 Single Nucleotide Polymorphisms

SNPs are single positions in the DNA where there is variability. Generally, this variability is produced by a mutation during the process of copy of the DNA. These mutations are mostly substitutions of one base for another, but they can also be product of the insertion or deletion of a nucleotide. It is very unlikely that more than one mutation could have occurred at the same locus during the short human evolution. This is the reason why, traditionally, SNPs are assumed to be biallelic i.e. they can take only two forms among the whole population, A_1, which is considered the ancestral allele, and A_2, which is considered the mutated allele [212, 35]. Thus, at a given locus, any individual should have one of the three possible genotypes, A_1/A_1, A_1/A_2, or A_2/A_2. In the example shown in Figure 2.5, at locus j of gene B one should have G/G, G/T or T/T. An individual with two identical alleles (A_1/A_1 or A_2/A_2) is called *homozygous* whereas an individual with both alleles (A_1/A_2) is called *heterozygous*. In Figure 2.5, the individual is homozygous for gene A (A_1/A_1) whereas he is heterozygous at locus j of gene B (G/T).

Traditionally, the standard representation of SNP data was by using symbols from the alphabet {0, 1, 2} [125]. “1” is used to denote the homozygous combination with the major allele (A_1/A_1), and “2” to denote any combination containing the mutated allele (A_1/A_2, and A_2/A_2). Some allele values may be missing due to experimental reasons and these are denoted by “0”. Another codification for SNP data can be {0, 0.5, 1} where 0 and 1 are the homozygous genotypes, and 0.5 the heterozygous [263]. {-1, 0, 1} is also used for coding SNP data where -1 and 1 are the homozygous ancestral and mutated combinations and 0 is the heterozygous genotype [113]. Nowadays, the alphabet {0,1,2,3} is a standard representation of SNP data, where “0” are missing genotypes, “1” and “3” are both homologous genotypes (“1” for the most common allele), and “2” for the heterozygous genotype [16].

An alternative approach is to code the alleles in a numerical alphabet (1, 2, 3, 4) instead of (A, T, C, G), coding a SNP locus with an A/C genotype as 13 [225]. More recently, new methods propose 3-D visualization of SNP data based on the projection of the data in the complex plane [27]. A vector of SNP data $x[n] = (x[0], x[1], \cdots, x[n-1])$ is mapped to a point $F_1(x[n])$ in the complex plane as in equation 2.1.

$$F_1(x[n]) = \sum_{j=0}^{n-1} x[j] e^{-\frac{2\pi i j}{n}} \tag{2.1}$$

where $x[j]$ is a numerical representation of the j-th SNP.

Among a given population each of the two alleles of a SNP has a certain frequency depending on the evolution of the mutation. It is called *allele frequency*. Geneticists commonly use the *Minor Allele Frequency* (*MAF*) for studying the distribution of a mutation among a population. The MAF is the frequency of the less common allele (generally the mutated one A_2). The distribution and allele frequency of mutations as well as their effects on the population are the keys for the evolution of species. This is described by the Hardy-Weinberg (HW) principle, which states that if a large population is free of evolutionary forces, then the allelic frequencies remain constant over time. No such population exists so that allele frequencies are mostly in HW disequilibrium [65].

In addition to be differentiated by their allelic frequencies, SNPs differ from each other by the gene regions that belong to or by their functionality. It is known that less than 2% of the DNA sequence codes for proteins. Part of the remaining regions may have a function on the regulation of the gene expression.

The location of a SNP refers to the region of the genome to which it belongs. SNPs that fall within a coding region in the gene (*exon*) are called cSNPs (c for coding). However, research suggests that most SNPs fall in the noncoding region of the human genome. Promoter regions are the regions that precede the genes and where the transcription of the gene and the posterior translation to a protein is originated. This process regulates the levels of expression of the gene and hence the levels of the protein [277]. SNPs located in these regions are called rSNPs (r for regulatory). There is evidence that rSNPs and cSNPs are most likely to affect disease [158]. However SNPs can also be located in *introns*, regions within a gene that does not code for proteins.

Otherwisem, cSNPs can contribute to complex disorders in two different ways, by either changing the structure of a specific protein, or by changing the abundance of the protein [122]. This is known as the functionality of the SNPs.

When the SNP does not cause a change in the amino-acid sequence of the resulting protein, it is called a *silent* or *synonymous* SNP. Among SNPs that produce a change in the resulting protein, called *nonsynonymous* SNPs, one can distinguish between missense or nonsense SNPs. If during the transcription, the mutation produce a change in the amino-acid sequence of the protein, changing its nature or function, it is called a *missense* mutation. If a *nonsense* mutation is transcribed on a premature STOP codon (a nonsense codon), it will produce a truncated and often nonfunctional protein [248].

Although the HapMap and dbSNP provide a view of worldwide similarities and differences in allele frequency of human variation, there is a number of databases aimed at characterizing variation within or across human populations, including the European SNP database [127], the Japanese SNP database (jSNP) [118], the ThaiSNP database [110], SNP@ethnos [130], the CEPH genotype database [99] and ALFRED [228].

2.2.2.3 Copy-Number Variants

Copy-Number Variants (CNVs) are DNA segments larger than 1kb appearing a variable number of times as copies in a genome. These submicroscopic structural variants can be a result of insertion, deletion or duplication events. Recent studies have characterized that CNVs cover around 12% of the human genome. However, it has been shown that CNVs have a smaller contribution to gene-expression phenotypes than SNPs do [294]. Most CNVs are benign variants that will not directly cause disease. However, there are several instances where CNVs that affect critical developmental genes do cause disease. Recently, genome-wide surveys have demonstrated that rare CNVs altering genes in neurodevelopmental pathways are implicated in autism spectrum disorder and schizophrenia [269].

To increase the value of the data, the Database of Genomic Variants (DGV) was established to house CNVs found in the general population [132]. The Wellcome Trust Sanger Institute (Hinxton, UK) has developed a database of CNVs (called DECIPHER) associated with clinical conditions [83]. Other related CNV databases are the ECARUCA database [79] or the CNV-DB [301].

2.3 Phenotypes

2.3.1 What is a phenotype

Genetic variants described in previous section are responsible for all the observable differences between individuals, such as physical appearance and susceptibility to disease or response to medical treatments. A phenotype refers to any of these observable and measurable traits or characters of an individual that results from a genotype. The phenotype definition is a critical issue in the design of a genetic analysis of a certain disease and it will strongly depend on the disease under study. For instance, a phenotype can be a disease in itself or it may be any biological variable that explain or help to explain human diseases. Diseases that have a genetic compound are called *genetic disorders*. Even if all diseases have a genetic component, whether inherited or resulting from the body's response to biological stresses, genetic disorders refers to illnesses caused by abnormalities in genes or chromosomes. Cancer is a particular disease, due, in part, to a genetic disorder but that can also be caused by environmental factors.

Phenotypes result from the expression of an organism's genes. In eukaryotes, the accessibility of genes corresponding to large regions of DNA can depend on its chromatin structure, which can be altered as a result of histone modifications directed by DNA methylation, ncRNA, or DNA-binding proteins. Hence, these variations may up or down regulate the expression of gene. Certain of these modifications that regulate gene expression are inheritable and are referred to as epigenetic regulation.

This work focus on phenotypes corresponding to gene products, corresponding to the biochemical material, mainly proteins, resulting from the expression of genes. Proteins dictate virtually every reaction in the cells of almost all living things; they serve to regulate, facilitate, or directly cause countless different processes and reactions in most organisms thus are directly responsible for the observable characteristics of an individual.

As they are biological variables, phenotypes may take values at different domains being divided in discrete phenotypes (when taking two or few values) or continuous traits (when taking values in a continuous rank).

2.3.2 Genotype-Phenotype models

Nowadays, genetic studies aim to link genetic loci with specific disease phenotypes in order to identify disease genes or genetic traits associated with human diseases. Those genotype–phenotype associations related with human diseases are being accumulated in such databases as the Online Mendelian Inheritance in Man (OMIM) [106] and the genetic association database (GAD) [23] covering over 12 000 genes. These gene-disease association data should encode intrinsic features of diseases. However, the relationship between genotype and phenotype is not always straightforward. There exist four main models that explain how to relate genotypes and phenotypes (Figure 2.6).

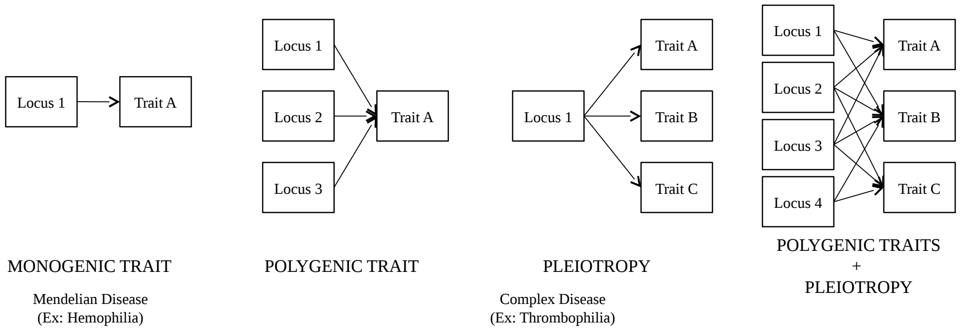


Figure 2.6: Different models for the relationship between genotypes and phenotypes.

Organizing individual disease-gene association data is becoming increasingly complicated and the necessity of a global view of relationships among diseases and genetic components has become essential. In this regard, a conceptual platform to project such associations in its entirety, called the human diseasome, has recently been introduced, which links all disease phenotypic features (human disease phenome) to all known disease genes (human disease genome) [93]. Based on these genetic foundations, human diseases can be divided into two categories: monogenic and polygenic diseases, also known as complex diseases.

2.3.2.1 Simple traits

When a trait is only caused by mutations in a single gene, it is called a *simple* trait or *monogenic* trait. It is also known as a Mendelian trait since all the Mendel's theories were elaborated under the assumption of a monogenic trait. A certain number of human diseases are monogenic, such as Hemophilia B, that is caused by mutations in the *F8* gene that produces a deficiency of FVIII protein levels in blood. Since carrying the mutations is directly related with the disease status of an individual, Mendelian diseases are usually studied using dichotomous variables, or "case-control" phenotypes.

A case-control phenotype is a dichotomous variable that takes two possible values (generally "case" or "control"), indicating whether an individual carries the disease or not. Generally controls are healthy individuals, not affected by the the disease under study. This should imply adding a new source of variability due to secondary traits such as age, gender or environmental factors. An alternative that avoids this problem is to balance both case and control groups in terms of these variables, for example, selecting as many men and women in both groups. This favors the posterior statistical analysis, avoiding other sources of variability between individuals more than the genetic one. However it may involve selecting controls randomly in the global population, assuming the risk to choose both affected and unaffected individuals.

2.3.2.2 Complex traits

Complex traits are caused by the interaction of multiple genes in combination with lifestyle and environmental factors. These are also known as *multifactorial* or *polygenic* disorders. There exist different types of genetic interactions. The most common one, *epistasis*, where the effects of one gene are modified by one or several other genes, which are sometimes called modifier genes. However, a given phenotype can also be the result of the expression of several genes at the same time.

Common traits of physical appearance are polygenic traits. For example, eyes color is determined by multiple genes coding for the different types of pigments. Some of the eye-color genes are *EYCL1* (a green/blue eye-color gene located on chromosome 19), *EYCL2* (a brown eye-color gene at chromosome 15) and *EYCL3* (a brown/blue eye-color gene also at chromosome

15). In this case the phenotype, the eyes-color, is a discrete variable that can take few different values (brown, blue, green, etc).

In the last few years it has been demonstrated that most common human diseases are complex diseases controlled by the interaction of several genes [37]. These diseases often involve a difficult and subjective diagnosis so they attracted the attention of worldwide geneticists and epidemiologists [266, 37]. They tend to involve greater difficulties in phenotype definition. The genetic heterogeneity is often closely associated with "intermediate" phenotypes that index some aspects of disease risk and susceptibility. An intermediate phenotype is a biological variable measured on a continuous quantitative scale such as weight, height, serum cholesterol levels or plasma *FVIII* levels. Furthermore, they are used as signs and symptoms to diagnose disease [284]. For example, thrombosis is a complex disease caused by the suppression of the blood circulation in a vein or artery. This is due to alterations of the coagulation process, where a set of proteins (including *FVIII*) in the blood plasma respond in cascade to form fibrin. These proteins are called coagulation factors and they represent a set of intermediate phenotypes that may be a good starting point for identifying genes involved in disease risk. Other examples of complex diseases are diabetes, Alzheimer's disease or psychiatric diseases among others [205].

In the global burden of complex diseases, cardiovascular diseases represent the majority. According to World Health Organization [189], cardiovascular diseases are one of leading causes of death in the world. Examples of common cardiovascular diseases with a strong genetic compound are coronary disease, stroke, hypertension, hypercholesterolemia, thrombosis or ischemia.

2.3.2.3 Pleiotropy

Pleiotropy occurs when a single gene controls or influences multiple phenotypic traits. Consequently, a mutation in a pleiotropic gene may have an effect on some or all traits simultaneously. One possible underlying mechanism of pleiotropy is when a gene codes for a protein used by various cells or having a signaling function on various targets. The influence of the single gene on different phenotypes can be direct or indirect. An example of a direct influence of a gene on multiple traits is albinism, where single gene mutations have effects on different organ systems, such as the integument system and the eyes, as well as the nervous, hematological, respiratory, and gastrointestinal systems that may occasionally be affected [46]. Besides, secondary or indirect pleiotropy occurs when a single gene might be involved in multiple pathways. For example, the amino acid tyrosine is needed for general protein synthesis, and it is also a precursor for several neurotransmitters (e.g., dopamine, norepinephrine), the hormone thyroxine, and the pigment melanin. Thus, mutations in any one of the genes that affect tyrosine synthesis or metabolism may affect multiple body systems. Related to this, the classic example of pleiotropy in humans is phenylketonuria (PKU). This disease can cause mental retardation and reduced hair and skin pigmentation, and can be caused by any of a large number of mutations in a single gene that codes for the enzyme (phenylalanine hydroxylase), which converts the amino acid phenylalanine to tyrosine [180].

Both direct and secondary types of pleiotropy are not always straightforward. Antagonistic pleiotropy refers to the expression of a gene resulting in multiple competing effects, some beneficial but others detrimental to the organism. The most common example of antagonist pleiotropy is the *p53* gene. This gene helps to avert cancer by preventing cells with DNA damage from dividing, but it can also suppresses the division of stem cells, which allow the body to renew and replace deteriorating tissues during aging [253].

2.3.3 Variability among populations

The phenotypic differences observed between individuals of a same family or population are magnified when they are observed between different populations, different races and also different species. The total number of characteristics in the genetic makeup of a species is called *genetic diversity* and it serves as a way for populations to adapt to changing environments. A population is defined as a set of organisms of the same species, that share the same environment, and that can breed together. The environment exerts a *selective pressure* on the individuals of a population by favoring individuals carrying variations of alleles to be suited for these conditions. Those individuals are candidates to survive and to produce offspring bearing that allele and this will lead to a population that will last for more generations. This is called *natural selection*.

Geographic location is critical for the genetic differences observed between populations, due to differences in selective pressure. However chance has also a role on mutation production due to chromosomal crossover. During meiosis, the alleles of the parents are mixed together to produce the offspring's allele. When the alleles of an individual are a random recombination of the alleles of its progenitor, it is due do chance. *Genetic drift* is the change in the frequency of an allele in a population due to random sampling.

Both genetic drift and environment can cause differences among populations. These differences exist at both genotypic and phenotypic levels. The study that aims understand the function and evolutionary processes producing different species and populations is called *comparative genomics* [107]. Comparative genomics mostly consists on comparing different organisms or individuals of different populations through their sequences (gene sequences or protein sequences). Due to the huge amount of data contained within a single genome and among a large number of organisms' genomes, computational and automatic tools are needed. These tools exploit both similarities and differences in protein, DNA or RNA sequences of different organisms. In 1990, Stephen Altschul and colleagues presented the basic local alignment search tool (BLAST), an application for searching and aligning sequences using a measure of similarity. Given a sequence, a blast algorithm indexes the query sequence and scans it against a database of sequences of a large variety of organisms, selecting those that show similarity scores above a given threshold [7]. In addition, BLAST includes a statistical framework for sequence alignment that provided a conceptual basis for understanding similarity measures, and a method for assessing the statistical significance of a given alignment.

The information obtained from studying the evolutionary conservation of DNA sequences between species has been useful in disease gene discovery studies [256, 12]. Cross-species sequence comparisons have shown that the human genome presents common features to other species [107]. The availability of multiple genomic sequences of different model organisms has allowed finding information about the selective pressure of polymorphisms from an evolutionary point of view [126]. It is assumed that sequence conservation is a good indicator of functionality. Functional sequences tend to evolve slowly, and show more conservation than less relevant sequences [22, 88, 200]. Hence, it is believed that functional genetic variants responsible for diseases are more conserved among populations and species than less functional variants [185].

Chapter 3

State-of-the-art in genetic association data analysis

A genetic association study aims to find statistical associations between genotypes (genetic variants) and phenotypes (traits or disease states) and thus to identify genetic risk factors. Genetic association for complex diseases can be tested either with unrelated people or with family-based designs. Both approaches have advantages and disadvantages. Studies of cases and controls in unrelated individuals are the most commonly used approach since sufficiently large study populations can be easily assembled without the need to enroll also family members of the recruited participants. However, a disadvantage of this approach is the confounding effect due to population admixture. On the other hand, family-based study designs have the advantage that there is a common genetic background among the family members. Thus, the problem of population stratification is bypassed. Moreover, families tend to be more homogeneous regarding the environmental factors possibly associated to the disease etiology. However, large enough samples of well-characterized families are usually more difficult to accumulate so that family-based strategies are less commmonly used for assessing genetic associations of complex diseases.

The first genetic association studies, performed in the 1980's, used a candidate gene approach. Such studies examined a single polymorphism or a set of polymorphisms near a single gene or focused on a candidate region obtained with prior knowledge. In the 1990s, the development of genome-wide genetic maps permitted the widespread application of genome-wide linkage analysis to disease status. However, it has been proved to be largely unsuccessful for complex traits. The last decade has seen revolutionary advances in human genetics. With the completion of the human genome sequence, the identification of large numbers of genetic markers and the development of rapid high-throughput methods to genotype SNPs together with new statistical techniques, now permit comprehensive, large-scale association studies with SNPs to survey genes or regions for variants that contribute to disease susceptibility or other traits of interest. These are known as Genome-Wide Association Studies (GWAS).

This chapter aims to review the different strategies for genetic association studies found in the literature. First of all, methods for unrelated individuals are described, including both one-locus and multi-loci strategies. Besides, family-based approaches for identifying genetic variants related to diseases are also revised. In addition, some statistical issues derived from genetic association studies are also addressed.

3.1 Association Studies for Unrelated individuals

The most commonly used approach for population-based genetic association analysis are the one-locus association measures. They test the association between SNPs and disease, looking marker-by-marker and selecting the best SNP(s) (those that obtain the highest score of association). However, complex diseases are generally dictaed by the interaction of multiple genes jointly with environmental factors so that in some cases multi-loci analyses may be more informative than traditional one-by-one SNP association studies. This section reviews the methods for both one-locus and multi-loci approaches for the genetic association problem.

3.1.1 One-locus measures of association

One-locus genetic association is generally tested using variable ranking strategies. These approaches measure, for each marker, its association with the phenotype establishing a ranking of SNPs. Hence best ranked SNPs are selected. The score attributed to each SNP reveals the significance of its association with the phenotype and it is evaluated using different statistical tools.

Traditionally, genetic association studies focus on population-based case-controls studies. These techniques select SNPs when their allelic frequency among exposed individuals is higher than among unexposed individuals [34]. However, population based methods also exist for traits that show continuous variation.

The most standard practice consists on measuring the correlation between genetic variants and phenotypes and evaluating its significance through a statistical test. Linear statistical models have also been used for measuring the association between a SNP and a phenotype.

3.1.1.1 Correlation-based statistical tests

Most common measures for case-control studies are based on contingency tables [263]. The most typical case corresponds to a genotype with 3 possible values (A_1A_1, A_1A_2 and A_2A_2) and a phenotype with 2 classes (cases and controls). The resulting contingency table is a 3x2 matrix that displays the frequency distribution of the variables as in table 3.1.

Table 3.1: An example of contingency table for case-control genetic association.

	A_1/A_1	A_1/A_2	A_2/A_2
Cases	O_1	O_2	O_3
Controls	O_4	O_5	O_6

Given a contingency table, statistical tools are available to test the association between the genotype and the phenotype. The most commonly used are the odds ratio, the χ^2 test or the log-likelihood test [263].

The odds ratio is a measure of effect size, describing the strength of association or non-independence between two binary data values [343]. The odds ratio (OR) for disease is the ratio of alleles carrying the mutation to non-carrying in cases compared with that in controls as described in equation 3.1.

$$OR = \frac{(O_2 + O_3)O_4}{O_1(O_5 + O_6)} \tag{3.1}$$

The χ^2 statistic is defined as in equation 3.2.

$$\chi^2 \sim \sum_{i=0}^{6} \frac{(O_i - E_i)^2}{E_i} \tag{3.2}$$

where O_i are the observed frequencies and (E_i) the expected value at each cell [170]. The likelihood ratio test is based on the G-statistic defined in equation 3.3.

$$G \sim 2\sum_{i=0}^{6} O_i \ln(\frac{O_i}{E_i}) \tag{3.3}$$

For both χ^2 and G statistics, given a contingency table, the value of the statistic is compared to a χ^2 distribution and a p-value is obtained that determines the significance of the dependence between the two variables in the contingency table (here the genotype and the phenotype).

The Cochran-Armitage test for trend is commonly used as a genetic association test in case-control studies [55]. It is also based on contingency tables and strengthen the χ^2 test by incorporating the ordering in the effects of the categories.

In a similar way, other measures for case-control association have been proposed such as a Hardy-Weinberg Equilibrium (HWE) test that compares the HWE between cases and controls as a means of a measure of disease association [211].

Linear correlation measures can also be applied for measuring the genetic association between genetic variants and both discrete and continuous phenotypes. The most commonly used measures are linear, such as Pearson coefficient or Spearman coefficient.

Pearson's correlation coefficient between a phenotype Y and a SNP S is defined as the covariance of the two variables $cov(S, Y)$ divided by the product of their standard deviations (σ_S and σ_Y respectively) (equation 3.4).

$$\rho(S,Y) = \frac{cov(S,Y)}{\sigma_S \sigma_Y} = \frac{E[(S-\mu_S)(Y-\mu_Y)]}{\sigma_S \sigma_Y} \tag{3.4}$$

Spearman coefficient is a rank coefficient correlation that measures the statistical dependence between two variables. It assesses how well the relationship between two variables can be described using a monotonic function. It is expressed as in equation 3.5.

$$\rho = 1 - \frac{6\sum\limits_{i=1}^{n}(S_i - Y_i)^2}{n(n^2-1)} \tag{3.5}$$

where n is the number of individuals.

Nonlinear correlation measures based on information theory have also been used for genetic association but have not been fully explored. In particular, *Ruiz-Marín et al.* proposed a genetic association test based on the entropy measure[257]. A maximum entropy conditional probability modelling has been proposed for finding interactions between SNPs and disease [206]. The genetic association between genetic markers and disease has also been measured with the mutual information [263, 299]. Measures of information theory are further described in chapter 4.

3.1.1.2 Linear statistical models

Regression is an approach to modelling the relationship between an outcome (here the phenotype Y) and a variable (here a SNP S). In linear regression, data are modeled using linear functions, and unknown model parameters are estimated from the data. Linear models try to express the relationship between Y and S by an affine function as described in equation 3.6.

$$Y \sim \alpha + \beta S + \epsilon \tag{3.6}$$

where β is called the regression coefficient and it is estimated from the data, and ϵ is the error term of the model. A statistical test based on the t-Student statistic evaluates the significance of β, which represents the significance of the association between Y and S. Statistically, this method is equivalent to the tests defined previously.

For dichotomous phenotypes, regression models are not efficient so that nonlinear transformations of the output of the linear regression model have been proposed. The most commonly applied for genetic association is the logistic regression [18], that consists on applying a logit function on the result of the linear model. The logistic regression model is defined in equation 3.7.

$$Y = \frac{1}{1 + e^{-z}} \tag{3.7}$$

where $z = \alpha + \beta S + \epsilon$.

Regression models are often used as explanatory models for genetic association with continuous phenotypes [18]. By contrast, predictive models are aimed at predicting the effect of the genetic variant on the disease. Moreover, applying predictive models to genetic association is especially interesting for a large number of SNPs, when the methods described previously become computationally expensive [10].

Predictive modelling is the process by which a model is created or chosen to try to best predict the probability of an outcome (here a phenotype). The main aim of predictive models is to estimate the outcome (the phenotype Y) from the predictor variables (here a SNP S). The best prediction $\hat{Y}$ of Y is the one that minimizes the prediction error $\epsilon = |\hat{Y} - Y|$. They can be applied to both discrete and continuous phenotypes.

Another predictive modelling approach already used for genetic association is the random forest analysis, with the aim of identifying SNPs that may increase the susceptibility to disease [39]. A random forest is a collection of trees generated by a modified tree-growing algorithm [33]. The class (here case or control) of an observation (here an individual) is predicted by assigning this observation to a terminal node based on its predictive values (here the genotype). A SNP that differentiates between cases and controls is found by quantifying how much it contributes to the predictive accuracy of a random forest.

3.1.1.3 Genetic recombination and Linkage Disequilibrium

Linkage is the tendency of genes or other DNA sequences at specific loci to be inherited together as a consequence of their physical proximity on a single chromosome. Genetic recombination occurs when two homologous chromosomes exchange parts of their DNA. This often happens in gametes so that new combinations of alleles can be passed on to the next generation. Recombination occurs at random. A haplotype is a combination of alleles, generally at adjacent locations (loci) on a chromosome that are transmitted together. A haplotype may be one locus, several loci, or an entire chromosome depending on the number of recombination events that have occurred between a given set of loci.

Two SNP loci may share a certain amount of correlation, that is, they are linked. Sometimes, it is possible to predict the allele at one SNP position based on the allele at the other locus. This happens more frequently when positions are close to each other and so are inherited together, without being separated by recombination. If the prediction accuracy is 100%, these two SNPs are said to be fully linked. More often, the allele knowledge at one locus gives some partial information about the allele at the other, and then they are said to be in linkage disequilibrium (LD). Non-random associations between polymorphisms at different loci are measured by the degree of LD.

The basic component of LD metrics is the difference between the observed and the expected frequencies of a haplotype assuming no statistical association. If A and B are two loci with two alleles A_1, A_2 and B_1, B_2 respectively, the linkage disequilibrium between A and B is D is measured as defined in equation 3.8 [172].

$$D(A, B) = p_{A_1B_1} - p_{A_1}p_{B_1} \tag{3.8}$$

where p_{A_1} is the frequency of allele A_1 at locus A, p_{B_1} is the frequency of allele B_1 at locus B, and $p_{A_1B_1}$ is the frequency of the A_1B_1 combination. However, the most commonly used measures for describing LD are D' and r^2 (equations 3.9 and 3.11) [171, 87].

$$D' = \frac{|D|}{D_{max}} \tag{3.9}$$

where

$$D_{max} = \begin{cases} \min\left(p_{A_1}p_{B_2}, p_{A_2}p_{B_1}\right) & \text{if } D \geq 0 \\ \min\left(p_{A_1}p_{B_1}, p_{A_2}p_{B_2}\right) & \text{if } D < 0 \end{cases} \tag{3.10}$$

$$r^2 = \frac{D}{p_{A_1}p_{B_1}p_{A_2}p_{B_2}} \tag{3.11}$$

In a set of SNPs with elevated LD, there is redundant information so that it is possible to select a representative SNP and use it to infer remaining SNPs. This SNP is called a tag-SNP and the methodology is called tag-SNP selection [296]. Current approaches for tag-SNP selection can be classified as 'block-based' and 'block-free' methods [317]. Block-based algorithms initially define haplotype blocks at distinct chromosomal regions of elevated LD and subsequently select the corresponding tag- SNPs [337]. Block-free methods use flexible networks of SNPs and exploit the inter-marker dependencies within these networks [104]. Tag-SNP selection based on LD measures can be seen as a filtering procedure in the sense that it allows to reduce the dimensionality of SNP data. Given that recently microarrays chips can genotype the order of 1 million SNP, these techniques may be useful as a preprocessing step in a genetic association study.

Note that most of the LD measures described above can also be used as correlation measures in a genetic association test, as described in section 3.1.1.1. It has also been suggested that the information in LD is also useful for association studies since genetic association can be detected by comparing LD patterns between cases and controls [332]. Actually this strategy is often considered as a multi-loci approach for genetic association as it analyses multiple loci simultaneously. However, the more general goal of multi-loci association studies is to analyze the interactions existing between several loci, being those in LD or not.

3.1.2 Feature selection methods for multi-loci association

As described in section 2.3.2.2, genetic diseases may be polygenic, that is they are caused by the interaction of several mutations at different loci. These multiple and combinatorial interactions are difficult to detect with traditional statistical methods. This is the reason why the development of computational and statistical methods to face this problem is of clinical interest. As many other fields in bioinformatics, genetic association and linkage studies require to use techniques from other engineering sciences. Genetic association can be approached from a pattern recognition point of view. Actually, finding association between several genetic markers and a phenotype can be seen as a feature selection (FS) procedure, in the sense of selecting genetic variants associated to the phenotype. These methods can be split in two basic aspects: the relevance criterion that determines how well a set of SNPs represents the observed variability in the phenotype and the search method used in the selection algorithm [105].

This section aim to be a review on Feature Selection algorithms and its application in bioinformatics. There exist in the literature several considerations to characterize FS algorithms [136, 210]. Among them, three criteria are used for this characterization: the general scheme of the algorithm, the evaluation measure or the relevance criterion, and finally the search organization.

3.1.2.1 General Scheme of a Feature Selection Algorithm

The FS problem is an inductive machine learning process and it is widely used for pattern recognition. Given a set S of n candidate features, a feature selection algorithm (FSA) selects the subset S^* of features that performs better under a certain FS criterion measure Φ [136]. The optimal solution to the FS problem requires to make an exhaustive search, looking at all the possible combinations of features, which is computationally unfeasible.

The relationship between a FSA and the inductive method used for evaluating the usefulness of the obtained subset of features can take three different forms: filter, wrapper and embedded.

A filter approach does not depend on the evaluation measure used for determining the optimality of features subsets. The relevance of features is assessed by looking only at the intrinsic properties of the data. Generally, these methods take place before of the FSA itself. In most cases, a relevance score is calculated for each feature and low-scoring features are removed. Las Vegas Filter (LVF) [178] is an algorithm that generates random subsets of features (S) and evaluates its relevance by measuring the consistency of the features in S through a specific criterion. This procedure is repeated until the minimum consistent set of features (S^*) is found [178]. LVF is the first in a family of algorithms called Las Vegas algorithms that are new versions of this algorithm by improving some of its characteristics [210]. RelieF is another family of filter algorithms that avoid an heuristic search [149]. The key idea of RelieF algorithms is to estimate the quality of attributes according to how well they distinguish between instances that are near to each other. Given a feature S_i chosen randomly, the algorithm searches for its two nearest neighbors: one from the same class, called nearest-hit (H_i), and the other from a different class, called nearest-miss (M_i). Then an attribute able to separate S_i and H_i has a low relevance whereas an attribute that separates S_i and M_i has a high relevance. RelieF algorithms return a weighted version of the original feature set S [309].

Wrapper FS methods use the learning algorithm as a subroutine. The relevance of features sets is calculated according to the evaluation measure that characterize itself the FSA. Wrapper FSA are the most commonly used for the problem of genetic association and they will be described in section 3.1.2.3.

For embedded FSA, the evaluation method has its own FSA. Traditional machine learning tools such as decision trees or Artificial Neural Networks are included in this scheme [209]. In particular, Neural Networks have been used for multi-loci interaction detection [206].

3.1.2.2 Evaluation method

In the context of a genetic association study, the evaluation measure of a FSA should look at the relationship between genetic variants and a phenotype. Some of these measures have already been described in section 3.1.1.1. In particular, relevance criteria of a FSA may evaluate the statistical significance of the association between features (SNPs) and the response variable (the phenotype), which is generally measured through a correlation measure. This relevance criterion helps the search organization method to decide, at each step, the optimality of a feature subset.

3.1.2.3 Search organization

A FSA can be single-solution or multi-solution. The former type of FSA only find a single solution, the best suboptimal subset of features, whereas multi-solution FSA store a queue of possible suboptimal sets of features. Besides of this consideration, FS strategies can be classified depending on their search organization that can be sequential or random.

Sequential algorithms are the most commonly used. They work in an iterative manner, adding or removing features to the selection set by using forward and backward steps respectively. The most common variants are called Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) [89]. These are single-solution FSA that start with an initial feature subset (the empty one in SFS or the total one in SBS) and iteratively add (for SFS) or remove (for SBS) features until the optimal solution or the termination criterion are met. This strategy does not take into account the correlations between features and may produce the effect of finding redundant sets of features. In order to avoid this problem, algorithms that combine forward and backward steps have been proposed. Plus r - take away l algorithm combines r SFS and l SBS [289]. Afterwards, floating algorithms were introduced (SFFS: Sequential Forward Floating Selection; SBFS: Sequential Backward Floating Selection) [281]. Floating algorithms do not depend on the r and l parameters but combine forward and backward steps dynamically.

Randomized methods depend on a random element that could produce different sets on every run. Random search or genetic algorithms are examples of randomized wrapper approaches also widely used for feature selection. Genetic algorithms (GA) are based on the mechanics of biological evolution such as inheritance, mutation, natural selection, and recombination (or crossover). In a GA approach, a given feature subset is represented as a binary string (a "chromosome") of length n with a zero or one in position i denoting the absence or presence of feature i in the set. Initially many possible solutions (chromosomes) are randomly generated to form an initial population. A proportion of this population is selected to breed a new generation. Each chromosome is selected through a fitness-based process where fitter solutions are more likely to be selected. New generations are generated iteratively using genetic operators: crossover or recombination and mutation. This generational process is repeated until a termination condition has been reached that can be the optimality of the solution, a fixed number of generations reached or any other criteria [276].

Moreover, noteworthy on its own is the Branch and Bound algorithm (BB), the only "optimal" FS algorithm that avoids an exhaustive search [218]. The algorithm is very efficient because it avoids exhaustive enumeration by rejecting suboptimal subsets without direct evaluation and guarantees that the selected subset yields the globally best value of any criterion function Φ. The optimality of this algorithm is guaranteed only when ϕ is monotone. The search space is structured as a tree, called a search tree, which is dynamically constructed top-down during the running of the BB algorithm. The search process begins at the root node, the complete feature set S, and continues by eliminating one feature each time to produce its successors (smaller subsets S_i). This step is called branching since its recursive application defines a tree structure. The bounding step consists in computing upper and lower bounds for the optimal value of $\Phi(x)$ within a given subset S_i. Whenever a node's evaluation value is found to be less than or equal to the bound, the sub-tree rooted from this node will be pruned, because there is no chance that a better target subset could exist in the sub-tree. Several versions of the BB algorithm constitute a family of methods for feature selection [67].

One of the main characteristics of feature selection algorithms is their high computational cost. Table 3.2 compares the time complexity of the algorithms described above [159]. It is observed that suboptimal algorithms (SFFS, SFBS and BB) present an exponential time complexity. GA improves the performance of such algorithms by not exploring all the feature space.

Table 3.2: Time complexity of Feature Selection algorithms (From [159]).

Algorithm	Time Complexity
SFS, SBS	$\Theta(n^2)$
SFFS, SBFS	$O(2^n)$
BB	$O(2^n)$
GA	$\Theta(1)(\Theta(n))$

3.1.2.4 Feature Selection in Bioinformatics

In bioinformatics, FS has been applied with several purposes. As it will be specifically described in section 3.1.3, FS has been applied for genetic association. Moreover, FSA have been applied to many modelling tasks in bioinformatics going from sequence analysis over microarray analysis to literature mining [262].

In the sequence analysis domain, FS has been applied for the prediction of subsequences (coding potential prediction), for the prediction of proteins from their sequences, for the recognition of promoter regions [57] and for the recognition of certain signals or motifs in the DNA sequence, such as binding sites for proteins and also in the context of gene prediction, for example for splice site prediction [146].

FS has also been used for microarray analysis. Because of the high dimensionality of most microarray analyses, fast and efficient FS techniques such as univariate filter methods have attracted most attention [166]. Univariate selection methods have certain restrictions such as not taking into account gene-gene interactions. Thus, researchers have proposed multivariate techniques that try to capture these correlations between genes [31].

However, wrapper methods have also been used in gene selection such as sequential search [134, 224], or GA [227, 139].

Text and literature mining is an emerging field in bioinformatics. The extraction and interpretation of biological results are not always easy to deal with. Recently, researchers have found that mining certain keywords in the literature may help for carrying out this task. Actually the application of FS algorithms is common in the field of text classification [85]. More particularly, in bioinformatics, FS has also been used in order to search relevant publications for manual database annotation [74].

3.1.3 Algorithms for Genetic Association Studies

Until the advent of technologies that allow the genotyping of hundred of thousands of genetic markers spread along the whole genome, genetic studies where limited to a small number of genes. Nowadays, high-throughput technologies enable individual genotyping of more than 10^6 SNPs using arrays. Effective storage, handling and analysis of this amount of data represent a challenge to modern computational and statistical genetics.

Traditionally, most Genome Wide Association (GWA) algorithms available in the scientific community have been based on a one-locus strategy which is computationally feasible for genome wide data. They differ on the evaluation method used for the SNP selection strategy. Linear regressions are widely used. For example, GenABEL is a commonly used R package that facilitates data quality control and rapid single-SNP GWA analysis using linear regressions [16]. Plink is a GWA software that uses either linear or logistic regression, depending on whether the phenotype is a quantitative or binary trait [244]. Merlin can test for association between a SNP and one or more quantitative traits by using a LOD score test or a likelihood-ratio test [1]. The Helix Tree module from Golden Helix also proposes different strategies for GWAS, such as the χ^2 test, linear or logistic regressions and others [94]. SNPassoc is also an R package for whole genome genetic association studies using linear or logistic regressions [96, 280].

Since it is computationally unfeasible to explore all the possible combinations of SNPs in a GWA scheme, filtering of SNPs has become a standard in genome wide tools. Most one-locus methods use or can be used as a preliminary step of a GWAS by filtering irrelevant SNPs and reducing the number of candidate SNPs. Nevertheless, computational methods from machine learning have been applied for genome-wide filtering of SNPs [214]. Filter FS algorithms such as RelieF have been used for this purpose [234, 182, 213]. Classification or decision trees are widely used for modelling the correlation between one or more features (here SNPs) and a discrete response variable (here a case-control phenotype). For example, random forests have been used for detecting gene-gene interactions at a genome-wide scale [39], as described in section 3.1.1.2. Furthermore, tag-SNP selection is also an alternative for reducing the number of candidate SNPs. These methods use a clustering of SNPs in terms of their LD and select a representative SNP for each cluster [41, 225]. Feature selection has been used for tag-SNP selection using LD [45, 105], Multiple Linear Regression [114] or sequential search [234].

Multifactor Dimensionality Reduction (MDR) has been used to identify potential interacting loci in several phenotypes [62]. This technique consists on finding combinations of SNPs that are related with a case-control phenotype. MDR reduces the dimensionality by converting a multivariate multiloci model to a one-dimensional model, by classifying genotypical classes as either high risk or low risk according to the ratio of cases and controls in each class. *Brinza et al.* proposed a FS algorithm based on a combinatorial search for finding disease-related multi-SNP combinations [34]. Evolutionary algorithms have been also used for the selection of SNPs [128, 270]. Filter algorithms, relieF, a tuned version of relieF (tuRF) and a combination of relieF with random forests have been used for genome wide analysis [62]. PCA-BCIT, a PCA-based bootstrap algorithm, has also been proposed in the context of finding gene-disease association [231]. Moreover a sequential FS method based on relevance chains has also been proposed for finding multi-SNPs sets related with a phenotype [68]. Exhaustive search has also been applied in the context of multi-loci genetic association [311].

3.2 Family based studies

First genetic association studies were designed with familiar data and for a long time they have been the standard practice in genetic studies of human diseases [162]. In particular, first and simplest approaches were based on trios, consisting of one offspring and its two parents [286]. Later, sib-pairs analysis was introduced [232]. Sib-pairs, pairs of brothers or sisters, tend to present more homogeneity of age and environment than other pairs of relatives, and they are relatively easy to ascertain [157]. However these methods have been extended to small pedigrees (nuclear families) and more recently to large pedigrees (extended families) [29]. Family-based studies take into account the dependence of the genetic information between relatives and use it in order to explore both within-family and between-family information. An important advantage of family-based studies is that they are robust against population admixture and stratification [247]. This section will review different strategies for dealing with familiar data in the study of genetic association between genetic variants and diseases. The two existing approaches are genetic linkage analysis and association studies.

3.2.1 Genetic similarity between individuals

Most family-based methods come from the idea of estimating the genetic similarity between individuals at a specific marker. All measures of relatedness are based on the concept of identity by descent (IBD), a key concept in quantitative genetics. Two alleles are said to be identical by descent if they are identical copies of the same ancestral allele. The probability that two alleles of two relatives are IBD is called the kinship coefficient. It is a measure of the degree of genetic relatedness of two individuals. For example, the kinship coefficient between identical twins is 0.5 it is 0.25 between father and son. The matrix that contains these probabilities for a given locus and for all the pairs of individuals is called the kinship matrix or also IBD matrix. These matrices express the genetic relatedness among individuals at a particular locus [157]. There exist several methods in the literature for the estimation of IBD coefficients ($\pi_{i,j}$) but the simplest one is defined in equation refeq:ibd.

$$\pi_{i,j} = k_1/2 + k_2 \tag{3.12}$$

where k_1 and k_2 are the probabilities that individuals i and j share 1 and 2 alleles IBD respectively [315].

In practice, the number of alleles shared by IBD at a given locus is difficult to be ascertained because the allelic measurements of the parents are not always available. The genotypic distance between relatives can also been inferred using Identity-By-State (IBS) probabilities, that is the probabilities that two individuals share an allele, regardless of its ancestral origin [29]. Figure 3.1 shows a particular pedigree where , individuals 4 and 5 share

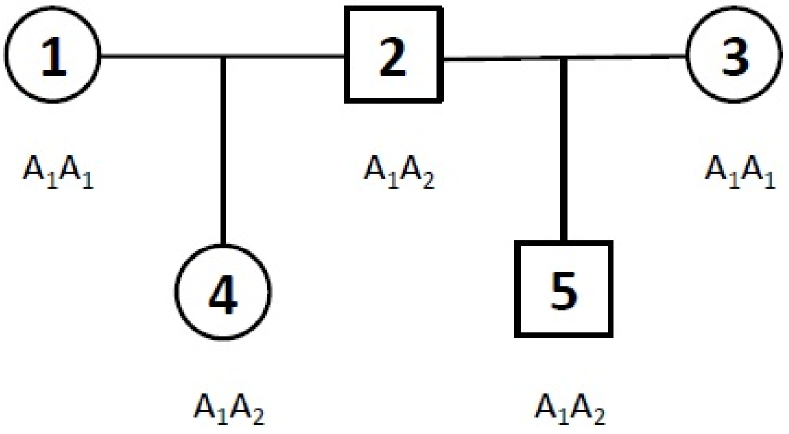


Figure 3.1: An example of pedigree.

allele A_2 IBD because they are copies of the allele of their common father. However, they share the allele A_1 IBS but not IBD because they are identical but provide from a different mother.

Sometimes, one can establish the genetic similarity between individuals directly from counting the number of alleles shared IBD or IBS between individuals and avoiding the corresponding probability estimation. The computation of probability estimates is usually costly, specially for high dimensional IBD matrices, so that different methods for the estimation of IBD or IBS matrices have been proposed [26]. Deterministic approaches based on regression methods have been proposed [26, 238] whereas LOKI is a stochastic method based on Markov-Chains Monte-Carlo simulations [116].

3.2.2 Linkage methods

Linkage studies consist in evaluating the statistical evidence of the co-segregation of a marker loci with a trait in a family. Relatives who share a particular trait will also share alleles at markers surrounding the gene(s) influencing this phenotype, and vice versa.

The most commonly used linkage methods are the Hasseman and Elston regression and the Variance Components Analysis (VCA).

Haseman and Elston [109] were the first to describe a widely used method to map human quantitative trait loci (QTLs). This method is based on a regression model between the differences for the trait between two relatives and the estimated proportion of alleles shared IBD at a marker (equation 3.13).

$$E(\Delta_j^2) = \alpha + \beta\Pi_j \tag{3.13}$$

where Δ_j^2 is the squared phenotypic difference between j-th pair of relatives, and Π_j is the proportion of marker alleles shared IBD for this pair. If the slope β is significantly negative, there is a QTL linked to the marker.

Variance Components (VC) linkage analysis is an extension of the Haseman and Elston regression for extended families. In VC, the cosegregation of the linked markers with a trait locus is used to decompose inter-individual variability into linked and unlinked components of variance. In particular, the variance of the phenotype (Y) can be expressed as a linear polygenic model by means of the variance of the effect loci (QTLs) and the variance of the environmental conditions [9]. A quantitative phenotype Y can be expressed as a linear function of the overall mean μ, the effect of n QTL's q_i and the environmental deviation e as in equation 3.14.

$$Y = \mu + \sum_{i=1}^{n} q_i + e \tag{3.14}$$

Assuming that q_i and e are uncorrelated random variables, the variance of Y is expressed in equation 3.15.

$$\sigma_Y^2 = \sum_{i=1}^{n} \sigma_{q_j}^2 + \sigma_e^2 = \sigma_G^2 + \sigma_e^2 \tag{3.15}$$

where σ_G^2 is the genetic variance. Hence, the covariance between two individuals' phenotypes (Y_1 and Y_2) is defined in equation 3.16.

$$Cov(Y_1, Y_2) = \sum_{i=1}^{n} \pi_{12i}\sigma_{q_i}^2 \tag{3.16}$$

where π_{12i} is the proportion of alleles shared IBD between the the two individuals at locus i. This covariance can be approximated by equation 3.17.

$$Cov(Y_1, Y_2) \approx 2\phi\sigma_G^2 \tag{3.17}$$

where $2\phi = E[\pi_{12i}]$ is the expected kinship coefficient.

The covariance matrix (Ω) of a general pedigree can be estimated by generalizing the bivariate covariance as in equation 3.18.

$$\Omega \approx \sum_{i=1}^{n} \Pi_i\sigma_{q_i}^2 + 2\Phi\sigma_G^2 + I\sigma_e^2 \tag{3.18}$$

where Π_i is the IBD matrix for the locus q_i, 2Φ is the expected kinship matrix, and I is the identity matrix.

The null hypothesis in VC methods for mapping QTL's is that the additive genetic variance due to the QTL is zero. This hypothesis is contrasted using a statistical test. The most commonly used is the likelihood ratio statistic or LOD Score test [6]. This test compares the likelihood of this model with that of a model in which the variance due to the i-th QTL is estimated. When multiple QTLs are jointly considered, the resulting likelihood-ratio test statistic has a more complex asymptotic distribution that is still a mixture of χ^2 distributions.

Several softwares for family-based genetic association are based on a linkage approach both with Hasseman and Elston regressions and with VC analysis. This is the case of Merlin [1]. SOLAR (Sequential Oligogenic Linkage Analysis Routines) is a software package that uses a multipoint VC model for general pedigrees [6]. LAMP also handles familiar data using linkage analysis [173]. LOKI also performs linkage analyses based on oligogenic models [116].

Linkage studies generally focus on microsatellites since they examine large number of families and see when the alleles of specific markers are inherited together with a phenotype in more cases than not. Microsatellites are good markers for studies of genetic linkage because they have a high heterozygosity. This means that allelic identity-by-descent can be readily established (unlike with bi-allelic SNPs) and linkage can be easily determined [336]. In contrast, family-based association tests generally use SNPs.

3.2.3 Association tests

The simplest family-based association design the Transmission Disequilibrium Test (TDT) [286]. TDT is an association test for the presence of linkage between a genetic marker and a trait. It was first used with genotype data from trios, which consist of an affected offspring and his or her two parents for the detection of genetic variants related to Mendelian diseases [162]. The principle of TDT is to determine which marker alleles are transmitted to the affected offspring. The TDT compares the observed number of alleles that are transmitted with those expected to be transmitted assuming Mendelian laws. Given n affected offsprings and their $2n$ parents and 2 alleles of a genetic locus (A_1 and A_2), a 2x2 contingency table is established as in Table 3.3.

Table 3.3: The TDT contingency table.

	Non-Transmitted allele		
Transmitted allele	A_1	A_2	Total
A_1	w	x	w+x
A_2	y	z	y+z
total	w+y	x+z	2n

Thus, a statistic test is defined under the hypothesis that two heterozygous parents (case x and y) are independent. The TDT statistic compares the proportions $x/(x+y)$ and $y/(x+y)$ and is adjusted to a χ^2 distribution as expressed in equation 3.19.

$$\chi^2 \sim \frac{(x-y)^2}{(x+y)^2} \tag{3.19}$$

The FBAT approach is a generalization of the TDT for general pedigrees [247]. If X and P denote the genotypes of the offspring and its parents respectively and T the offspring's trait, the covariance statistic used in the FBAT test is defined as in equation 3.20.

$$U = \sum_{ij} T_{ij} \cdot [X_{ij} - E(X_{ij}|P_i)] \tag{3.20}$$

where i indexes the pedigree and j indexes the offsprings, and where $E(X|P)$ is the expected value of X under the null hipothesis.

Other extensions of the TDT have been proposed for applying it to complex diseases schemes. For example, the Pedigree Disequilibrium Test (PDT) has been developed for the analysis of linkage disequilibrium in general pedigrees [194]. The Quantitative Pedigree Disequilibrium Test (QPDT) is a generalization of PDT for quantitative traits [339]. ParenTDT is an extension of TDT that also incorporates parental phenotype information. Many family based association software integrate the TDT or any of these extensions, such as Plink [244] or UNPHASED [75].

In addition, there are family-based association tests directly based on the polygenic model proposed in equation 3.14 [163, 51].

Finally, combined linkage and association strategies have been proposed [230]. These applications require that family relationships and linkage be appropriately accounted for in the association test. MERLIN, ILINK and LAMP include such strategies [173].

3.3 Genetic association studies in multiphenotypic schemes

In many cases, complex diseases are not dictated by a single trait but several symptoms appear at the same time to determine the syndrome or disease. All these symptoms that describe the disease are part of a collection of biological variables. These can be from physiological traits to certain proteins' levels in blood going through the expression levels of the genes that code for these proteins.

To understand the genetic basis of such diseases, each trait is often separately tested for association with one or more markers. However, if a locus affects two or more traits, a single-trait study may lose the power to detect a pleiotropic effect. In the past decade, both genetic association and linkage researches have focused on statistical and computational techniques for extending the traditional study of genetic association between one genotype and one phenotype to polygenic and multiphenotypic schemes [202]. In particular, the simultaneous analysis of multiple traits in the context of linkage mapping of quantitative trait loci (QTL) has attracted much attention. Several strategies have been followed and commonly applied for the analysis of multiple traits. The first consists on combining several univariate tests, one for each trait [326, 329]. Generalizations of the Maximum Likelihood (ML) method commonly used for linkage analyses have been proposed [6, 138, 154]. The most common strategy, which is integrated in the majority of genetic association softwares is the multivariate regression (MR) [153]. It generalizes the common genetic association test based on linear regressions [16, 101, 42, 155]. However, both generalizations (ML and MR) present a common drawback. Although they can be applied to multiple traits, a large number of correlated traits requires the simultaneous estimation of too many parameters, restraining its practical use.

An alternative approach is based on the transformation of the original traits to a reduced number of canonical variables [191, 324]. This category of techniques is included in a larger family of methods of phenotypic dimensionality reduction [203]. This approach is often implemented in two steps. First, new canonical phenotypic variables are generated. Next, a classical single trait genetic analysis is used to test the association between candidate loci and the canonical variable. The most common technique for the canonical transformation of multiple traits is the Principal Component Analysis (PCA) [323, 163, 150].

This is especially interesting when studying the genetic compounds of a collection of phenotypes related between them. This occurs when the disease under study results from a metabolic process, where a set of proteins respond in cascade within a given pathway until they cause a physiological change that correspond to the syndrome. For example, *Mathias et al.* propose to study several platelet function phenotypes, tightly involved in coronary artery disease [197]. In such cases, it has already been proposed to define new canonical variables in representation of sets of related phenotypes [52, 197, 47]. These new variables should meet two main requirements. The first one is that it should explain the largest possible proportion of the covariance between the phenotypes. Second, the canonical variable should be able to capture the common activity of the metabolic pathway under study and so recovering the dimensions of the entire syndrome [278, 190]. Using PCA allows reaching these conditions since it consists on a canonical transformation on the data where the new variables are the eigenvectors of the covariance matrix and so they capture a certain amount of common variance among the different variables. In particular, *Mathias et al.*, propose to use PCA for determining a set of independent factors explaining the phenotypic covariance between phenotypes biologically related to the platelet function. These new variables are used in further genetic analyses as new phenotypes in both genetic association and linkage analyses in order to identify genetic loci susceptible to be related to the common activity of the collection of original phenotypes.

3.4 Statistical issues in genetic association studies

GWAS are an important advance in discovering genetic variants influencing disease but also have important limitations, including their potential for false-positive and false-negative results and for biases related to the selection of study participants and genotyping errors. Due to the availability of many SNPs, GWAS must utilize very large sample sizes, which at the same time highlights their statistical limitations, leading to false positive findings. This is the most important drawback of GWAS and it is common to all the methods proposed previously. The main sources of spurious findings in GWAS are multiple testing problems and population stratification. Another important pitfall of GWAS is their lack of replication or validation. This can be dealt using some enrichment tools such as prioritization criteria. In this section, statistical methods to address these issues are described.

3.4.1 Multiple testing

Nowadays, thanks to recent technological advances, investigators can test the association between up to a million SNPs and disease phenotypes. In addition, some analyses study multiple and often correlated phenotypes and use multiple methods of statistical analysis, including different models or models with different covariates. When a genetic association study includes multiple genetic loci, multiple phenotypes, and multiple methods of evaluating associations between genotype and phenotype, it involves the testing of an enormous number of hypotheses, which implies a risk of inflation of the type I error rate (i.e. the probability of falsely claiming a positive association when it is not true) [279]. It becomes even more evident in Next Generation Sequencing.

The existing methods for the correction of multiple testing can be divided into two main categories, those limiting the number of tests to be performed and those adjusting the results for the number of tests. Since all the existing methods for correcting for multiple testing have strengths and weaknesses, there is not a clear consensus about how it should be dealt with [187].

The simplest correction for multiple testing is the Bonferroni adjustment, consisting on multiplying the nominal p-values by the total number of tests performed. This method assumes that the set of tests are independent. A refinement of the Bonferroni procedure has been proposed for the case where SNPs are in linkage disequilibrium [226]. It consists on estimating the effective number of independent SNPs, those that are not in linkage disequilibrium, and multiplying it by the nominal p-values. Because the effective number of independent SNPs is always less than or equal to the total number of SNPs tested, this method is less conservative than the Bonferroni procedure.

Permutation testing is an alternative way to adjust for multiple testing in genetic association studies. The basic principle of this is to permute the phenotype(s) with respect to the genotype(s) among observations, removing any association between phenotypes and genotypes but retaining the correlation among phenotypes and among genotypes, resulting from LD, within an individual. The minimum p-value of the original association tests is compared with a distribution of the p-values obtained from repeating this process thousands of times [53]. However this procedure is computationally intensive since it requires generating about 10000 permutations for achieving a significance threshold of 0.01 and about 1000 permutations for achieving a threshold of 0.05, which is the less restrictive alternative.

As an alternative, False Discovery Rate (FDR) control is a statistical method used in multiple hypothesis testing to correct for multiple comparisons [24]. Given a list of statistically significant findings, FDR procedures are designed to control the expected proportion of incorrectly rejected null hypotheses ("false discoveries"). FDR controlling procedures exert less conservative control over false discovery compared to the previously described family-wise error rate procedures (such as the Bonferroni correction). This increases power at the cost of increasing the rate of type I errors. FDR strategies have been used for multiple testing correction in GWAS [260, 76].

With related individuals, the dependence among relatives' genotypes can also contribute to the correlation between tests. Methods for multiple testing correction in these cases have been proposed [319].

3.4.2 Population stratification

In genetic studies, associations between genotypes and phenotypes may be confounded by unrecognized population structure and/or admixture. Studies have shown that even in European populations, which are thought to be relatively homogeneous, population stratification exists and can affect the validity of association studies [179].

When no evolutionary agents are affecting a population, the population is in Hardy-Weinberg equilibrium, that is haplotype probabilities and consequently allelic frequencies are constantly distributed within the population and are not changing from generation to generation. When, by contrast, the population is stratified, the HWE is violated and individuals may present different allelic frequencies depending on the subpopulation they belong to. In the context of genetic association studies, the differences found between the allele frequencies of cases and controls may be due to the population structure rather than a true association of the locus with the disease, introducing a statistical bias to the test [174]. This has become one of the most important pitfalls in genetic association studies.

There exist two main strategies to deal with population stratification, the early approach of the genomic control and the structured association tests. One one hand, genomic control is a family of methods for detecting and/or correcting for stratification based on the genome-wide inflation of association statistics with an inflation factor obtained from a set of random markers that are not associated with the phenotypes of interest. On the other hand, structured association methods explicitly infer genetic ancestry providing an effective correction for population stratification.

When the population structure is not known, the easiest way to correct for stratification would be to adjust the association model for any covariate that may be related to the population structure, such as ethnicity or geographic location. However the most standard procedure is to ascertain the genetic structure of the population by looking at the genetic correlations between individuals. This should be done by computing the kinship matrix, that measures the pairwise genetic distance between individuals. This matrix corresponds to the IBD matrix defined in section 3.2.1. Actually, when using familiar data, the effect of population stratification is easily controlled since they already contemplate the genetic population structure.

Once the kinship matrix is obtained, most methods use these correlations for assigning them to discrete clusters (subpopulations) and then segment the association analyses within each subpopulation. This method requires a computational effort that depends strongly on the number of clusters in which the individuals are assigned. It has been shown that in an association study with a sufficiently large number of markers and individuals, subjects can be partitioned into genetic clusters that match with the major geographic areas, where individuals from intermediate geographic locations have a mixed membership in the clusters that correspond to neighboring regions [254].

The most standard and traditional practice for detecting and correcting genetic association studies for the population structure is to use STRUCTURE, a model-based algorithm specifically designed for population structure inference [243]. The algorithm assumes a model in which there are K populations (where K may be unknown), each of which is characterized by a set of allele frequencies at each locus. Individuals in the sample are assigned (probabilistically) to populations using a Bayesian approach.

However improved methods have been proposed that are commonly based on principal-component analysis (PCA) method and multidimensional-scaling (MDS) method [313]. The PCA method identifies principal components that represent the population structure based on genetic correlations among individuals. The MDS method detects genetic similarities among individuals by obtaining the optimal dimension, lower than the original dimension of the data, so that the distance is preserved [174]. The MDS method has been combined with a clustering algorithm for grouping subjects into a variable number of clusters applying a clustering method over the subject coordinates in the new vectorial space [175].

Once the population structure (P) is ascertained, the standard procedure is to introduce it in the association test as a covariate. For instance in linear regression tests of association such as the proposed in the software GenABEL, P is introduced as a covariate in a mixed linear regression strategy [331]. The simple linear regression model described in section 3.1.1.2 is adapted as a mixed effects model as described in equation 3.21. Even if it remains an unanswered question, researchers suggest that population effects should be treated as fixed effects rather than random effects, since they are the same for all samples [240].

$$Y \sim \alpha + \beta S + \gamma P + u + \epsilon \tag{3.21}$$

where Y is the phenotypes matrix, S is the genotype matrix, α, β and γ the regression coefficients of the fixed effects (genotype and population effects) and u represent the random effects, assumed to be normally distributed with $var(u) = KV_g$ where K is the kinship matrix and V_g is the genetic variance.

The use of these techniques may help to avoid or at least reduce the number of false positive findings. In order to ascertain the true positive among the remaining positives, enrichment tools for GWAS are useful such as prioritization criteria.

3.4.2.1 Enrichment tools for Genome-Wide Association Studies

In addition to the false positive findings, an important drawback of current genetic association studies is their difficulty of replication and validation. Thus, prioritization criteria are needed to be established in order to rank candidate SNPs [28]. Most of these criteria can be used both as prioritization techniques and as a tool for analyzing and interpreting GWAS results. As, an example, meta-analysis is one of the most commonly used method for validating GWAS results.

Prioritization of SNPs may also entail adding biological knowledge to the system and use it for discarding irrelevant SNPs [199]. Biological information about the SNP. This information is found at SNP databases such as dbSNP [274], ensemble attributes of SNPS that can be used for prioritize them can be found in the genotypic data, the genetic context of the SNP, or previous established associations [235]. In the genetic context of a SNP, several features may be relevant such as the SNP location (the gene it belongs to), the chromosomal region, the SNP functional class (its functional location within a gene) or the overlap of the SNP with a Transcription Factor Binding Site (TFBS) or a splice site. The SNP evolutionary conservation can also be used as a prioritization criterion [261].

Recently, information from Protein-Protein Interaction (PPI) networks have been used to prioritize candidate genes [165, 176]. There usually are several biochemical pathways that are believed to play an important role in the development of certain disease. Selecting genes in these pathways using gene-gene interaction analysis may reduce the number of interactions tests to be performed. The prioritization is based on the idea that a gene that codes for a protein that interacts with other proteins might be a good candidate for interacting with other genes [214]. Thus one can only look for interactions among SNPs in genes with many PPI. Moreover, this information is also useful for scoring SNPs by looking both at the number of connections of the protein coded by the gene under study and at the strength of the PPI network. PPI networks are usually built based on pathway knowledge in public databases. The most commonly used is the Gene Ontology (GO), which provides a controlled vocabulary that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells [11]. Several studies have recognized similarity in GO annotation as one of the strongest predictors for protein interaction. GO annotation-driven interaction inference is based on the observation that proteins localized to the same cellular compartment or that that share a common biological process are more likely to interact and then to be predictive for PPI. Nowadays, prioritization tools based on PPI networks integrate KEGG, BioCarta, or other pathway databases in order to more efficiently examine the genes in a network context [314]. In particular, the Kyoto Encycopledia of Genes and Genomes (KEGG) offers a large database of pathways that have been curated by hand, and annotated to the KEGG orthologies, which are similar to the semantic ontologies proposed by the Gene Ontology Consortium [143].

Chapter 4

Information Theory

4.1 Historical Background

The most fundamental quantity in information theory is the entropy, the measure of information. The entropy of a physical system is the minimum number of bits needed to fully describe the detailed state of the system. It was firstly introduced in 1865 by Rudolf Clausius in the field of thermodynamics [54]. According to the second law of thermodynamics, the entropy is the degree of randomness in any system and it always increases or remain constant. Acccording to Clausius, the entropy change (ΔS) of a thermodynamic system absorbing a quantity of heat (ΔQ) at an absolute temperature T is simply the ratio between the two as described in equation 4.1.

$$\Delta S = \frac{\Delta Q}{T} \tag{4.1}$$

Between 1872 and 1875, Ludwig Boltzmann suggested that this quantity corresponds to the number of molecular degrees of freedom of the system. Boltzmann was able to show that the number of degrees of freedom of a physical system can be easily linked to the number of microstates Ω of that system. And it comes with a relatively simple expression from a mathematical point of view [32] (equation 4.2).

$$S = k \cdot \log\Omega \tag{4.2}$$

where $k = 1.38062 \cdot 10^{-23}$ joule/kelvin is the Boltzmann constant and Ω the amount of states the system has.

The information theory is a branch of mathematics originated by the publication of Claude E. Shannon's classic paper "A Mathematical Theory of Communication" in 1948 [272]. The main question motivating Shannon's work was how to design communication systems to carry the maximum amount of information and how to correct for distortions on the lines. In his revolutionary paper, Shannon proposed a qualitative and quantitative model of communication. He introduced the concept of a channel, consisting of a sender (a source of information), a transmission medium (with noise and distortion), and a receiver (whose goal is to reconstruct the sender's messages). In order to quantitatively analyze the transmission through the channel he also introduced a measure of the amount of information in a message. To Shannon, the amount of information was closely related to the chance of a message to be transmitted. A message is very informative if the chance of its occurrence is small. If, in contrast, a message is very predictable, then it has a small amount of information. This measure of information, that he called entropy, only depends on the statistical properties of the information source, independently of the kind of information it transmits. In his work, Shannon managed to mathematically quantify the concept of "information", defining the measure of entropy, as it is known nowadays. Shannon's entropy of a system represents the amount of uncertainty one particular observer has about the state of this system. Moreover, Shannon's mathematical formulation of the entropy was inherited as well by the measure of uncertainty proposed by Hartley in 1928 [108].

Even if Shannon is considered the father of information theory, previous works proposed measures that were later associated to this discipline. It is the case of the Kullback-Leibler (KL) divergence, introduced by Solomon Kullback and Richard Leibler in 1951 [160]. The KL divergence is nowadays the stem of information theoretic measures.

While Shannon never referred to his work as *information theory*, this new branch of mathematics attracted the attention of many scientists with theoretical and applied viewpoints. It is noteworthy to mention that Rényi proposed a framework that generalizes Shannon's and Kullback-Leibler's quantities [251].

Information theory constituted the theoretical fundamentals of digital communications systems. Later it has been exported to other branches of engineering as well as to physics, statistics or economics among others. In biomedical engineering, measures of information theory have been used, among others, in medical imaging [236] or in the study of heart rate dynamics [305, 124]. Information theory has also been applied in bioinformatics, as it will be described in section 4.4.

4.2 Measures of information Theory

Central quantities of information are the entropy (the information in a random variable) and the mutual information (the amount of information in common between two random variables). Both measures are obtained with a logarithmic expression depending on the probability mass functions of the random variables. The choice of logarithmic base determines the unit of information theoretic (IT) measures. The usual unit of information is the *bit*, based on the binary logarithm. However the natural logarithm is increasingly used for computational reasons, and in this case, the unit of information measures is the *nat*. The unit of measures based on the common logarithm is the *ban*.

Let introduce some notations that applies for the following sections. X and Y denote random variables. $p(X)$ and $p(Y)$ are their marginal probability mass function and $p(X, Y)$ is the joint probability mass function.

4.2.1 Entropy

The entropy is a measure of uncertainty associated with a random variable X. It quantifies the expected value of the information contained in a specific realization of the random variable as expresssed in equation 4.3.

$$H(X) = E(I(X)) = \sum_{i=1}^{N} p(x_i)I(x_i) = -\sum_{i=1}^{N} p(x_i)\text{log}p(x_i) \tag{4.3}$$

where $I(x_i) = -\log p(x_i)$ is the information content of the realization x_i.

The expression $p \log p$, tend to be equal to zero whenever $p = 0$ (equation 4.4).

$$\lim_{p \to 0^+} p \log p = 0 \tag{4.4}$$

Figure 4.1 shows the behavior of the entropy as a function of the probability in the case of a binary variable.

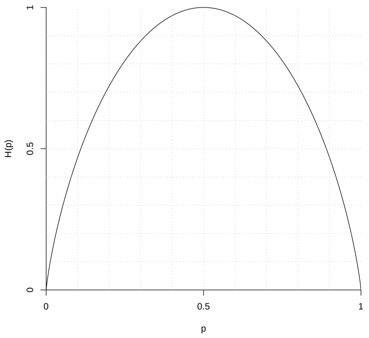


Figure 4.1: $H(p)$ as a function of p.

Figure 4.1 shows a concave curve that equals 0 when the random variable is deterministic ($p = 0$ or $p = 1$) and that takes its maximal value when $p = \frac{1}{2}$, which corresponds to the maximal uncertainty or randomness.

Rényi's entropy is a parametric generalization of the Shannon's entropy as defined in equation 4.5.

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \sum_{i=1}^{N} p(x_i)^\alpha \tag{4.5}$$

where α is the Rényi parameter. When α tends to 1, Rényi's entropy corresponds to the Shannon entropy.

If, in contrast, X is a continuous random variable, its *differential entropy* is defined replacing the sum by an integral as in equation 4.6.

$$H(X) = E(I(X)) = \int_\chi p(x) \log p(x) dx \tag{4.6}$$

where $p(x)$ is the probability density function defined on the domain χ.

The *joint entropy* of two discrete random variables measures the information they share. If X and Y are two random variables, their joint entropy is defined as the entropy of their pairings (X, Y) as in equation 4.7.

$$H(X,Y) = -\sum_{i=1}^{N} p(x_i, y_i) \text{log} p(x_i, y_i) \tag{4.7}$$

The joint entropy of two random variables is lower than the sum of their entropies.

$$H(X,Y) \leq H(X) + H(Y) \tag{4.8}$$

When the variables are independent(i. e., $p(X,Y) = p(X)p(Y)$), the joint entropy reaches it upper bound , being equal to the sum of the entropies.

When a variable X is conditioned by another variable Y, the conditional entropy is expressed as in equation 4.9.

$$H(X|Y) = -\sum_{i=1}^{N} p(x_i, y_i) \log \frac{p(x_i, y_i)}{p(y_i)} \tag{4.9}$$

The conditional entropy can also be expressed as in equation 4.10.

$$H(X|Y) = H(X,Y) - H(Y) \tag{4.10}$$

4.2.2 Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence is a central measure in information theory since it sets the basics for the definition of almost all the IT quantities.

The KL divergence measures the similarities between two probability distributions p and q. It is generally used for comparing a “true” distribution of data or theoretical distribution (p) with a model or approximation of p (q).

The KL divergence between two distributions $p(X)$ and $q(X)$ of a discrete random variable X is defined in equation 4.11) whereas the KL divergence between $p(X)$ and $q(X)$ when X is a continuous random variable is defined in equation 4.12.

$$D_{KL}(p(X), q(X)) = \sum_{i=1}^{N} p(x_i) \log \frac{p(x_i)}{q(x_i)} \tag{4.11}$$

$$D_{KL}(p(X), q(X)) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx \tag{4.12}$$

One of the most important properties of the KL divergence is its non-negativity as described in equations 4.13 and 4.14.

$$D_{KL}(p(X), q(X) \geq 0 \tag{4.13}$$

$$D_{KL}(p(X), q(X) = 0 \iff p = q \tag{4.14}$$

Actually, the KL divergence is often intuited as a metric but it does not fulfill the symmetry condition (the KL divergence between p and q is generally not the same as between q and p). Moreover it does not fulfill the triangle inequality.

The KL divergence can be decomposed as follows:

$$D_{KL}(p(X), q(X)) = \sum_{i=1}^{N} p(x_i) \log p(x_i) - \sum_{i=1}^{N} p(x_i) \log q(x_i) \tag{4.15}$$

where the first term is the opposite of the entropy of X.

4.2.3 Mutual information

The mutual information between two discrete random variables X and Y was defined by Shannon as in equation 4.16.

$$I(X;Y) = \sum_{x} \sum_{y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{4.16}$$

where p(x,y) is the joint probability distribution function and p(x) and p(y) the marginal distributions of X and Y respectively.

Mutual information can be also expressed as the Kullback-Leibler divergence between the joint distribution $p(X,Y)$ and the product distribution $p(X)p(Y)$ as in equation 4.17.

$$I(X;Y) = D_K L(p(X,Y), p(X)p(Y)) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{4.17}$$

For continuous random variables, the mutual information is defined as in equation 4.18.

$$I(X;Y) = \int\int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdx \tag{4.18}$$

The mutual information can be also interpreted as the reduction of the entropy due to conditioning as expressed in equation 4.19.

$$I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y) \tag{4.19}$$

Figure 4.2 shows a Venn diagram that relates information theoretic measures between them.

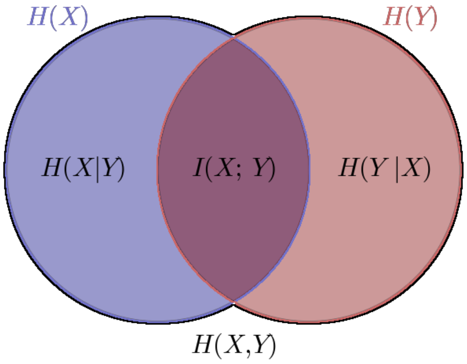


Figure 4.2: A Venn diagram relating IT quantities.

The non-negativity (equations 4.20 and 4.21) and the symmetry (equation 4.22) are remarkable properties of the mutual information.

$$I(X;Y) \geq 0 \tag{4.20}$$

being

$$I(X;Y) = 0 \iff X\text{and }Y\text{are independent} \tag{4.21}$$

$$I(X;Y) = I(Y;X) \tag{4.22}$$

As for the KL divergence, the mutual information cannot be considered a distance in the mathematical sense since it does not fulfill the triangle inequality. However both measures are commonly used as measures of similarity between variables.

4.3 Estimation of information theoretic measures

The computation of IT measures encounters two major problems.

On one hand, in the case of continuous variables, one has to deal with the computation of the integral. Computationally, integrals are estimated by partitioning its domain in several intervals and applying Riemann sums to each interval. This produces an error in the resulting IT quantities. The discretization of continuous variables for applying the discrete IT measures is an alternative, even if it also produces an error.

On the other hand, the most critical issue concerns the sample size. In real applications of IT measures, researchers work with data of finite sample size. Applying IT measures to finite data involves estimating the function of a probability distribution from a finite set of samples. Since IT measures are defined from real distributions, the use of probability distributions estimated from a finite sample of data lead to erroneous values for IT quantities.

Schneider et al. proposed an estimation for the sampling error in the entropy measurement [265]. Given a DNA position, if n_a, n_c, n_g and n_t are the numbers of A's, C's, G's and T's in this site and P_a, P_c, P_g and P_t are the frequencies of each base in the genome, then the probability of obtaining a particular combination of n_a to n_t (called n_b) is estimated as in equation 4.23.

$$P_{nb} = \frac{n!}{n_a!n_c!n_g!n_t!} P_a^{n_a} P_c^{n_c} P_g^{n_g} P_t^{n_t} \tag{4.23}$$

where $n = n_a + n_c + n_g + n_t$. P_{nb} is the probability of obtaining the entropy H_{nb} defined as in equation 4.24.

$$H_{nb} = -\sum (\frac{n_b}{n}) \log(\frac{n_b}{n}) \tag{4.24}$$

Finally the sampling error is defined in equation 4.25.

$$SE = \sum_{n_b} P_{nb} H_{nb} \tag{4.25}$$

Thus IT measures are needed to be corrected for this error. An alternative to this is to estimate IT measures by taking to account the bias produced by the finite sample size.

This section aims to review the different strategies for estimating IT measures. In particular this review focus on three main strategies to tackle this problem, Bayesian estimators for the entropy measure, Taylor-based estimations for the mutual information measure and Information Theoretic Learning (ITL), a nonparametric framework for the estimation of information theoretic measures.

4.3.1 Bayesian estimators

Traditionally, probability estimators are based on empirical estimations of the frequencies by counting observations. However, several methods for correcting the bias that it supposes are proposed. Among others, Bayesian estimators will be reviewed in this section, as they are the most important and most commonly used [111].

The empirical estimation of the frequencies $p_i = p(x_i)$ is also called maximum likelihood (ML), "plug-in" or "naive" estimation and it corresponds to the classical inference of the frequencies, done by frequentist statisticians from the countings of the observations 4.26.

$$\hat{p}_i^{ML} = \frac{n_i}{n} \tag{4.26}$$

where n_i is the number of counted observations and n is the total number of counts.

Given this estimation of the frequencies, the maximum likelihood estimator of the entropy and mutual information is expressed in equations 4.27 and 4.28.

$$\hat{H}^{ML} = -\sum_{i=1}^{n} \hat{p}_i^{ML} log \hat{p}_i^{ML} \tag{4.27}$$

$$\hat{I}^{ML}(X,Y) = \hat{H}^{ML}(X) + \hat{H}^{ML}(X) + \hat{H}^{ML}(X,Y) \tag{4.28}$$

Even if they are unbiased, the estimations of the frequencies $\hat{p}_i^{ML}$ downwardly biases the estimate $\hat{H}^{ML}$ of the entropy, because $\hat{H}^{ML}$ is defined by a concave-downward function, so that the average estimate derived from a range of estimates of the frequencies $\hat{p}_i^{ML}$ is less than the value of $\hat{H}^{ML}$ given by equation 4.27. This is the reason why other estimators such as the Miller-Madow (MM) estimator use bias correction as in equation 4.29 [207].

$$\hat{H}^{MM} = \hat{H}^{ML} + \frac{\hat{m}-1}{2n} \tag{4.29}$$

where $\frac{\hat{m}-1}{2n}$ is a first order bias correction and where $\hat{m}$ is the number of bins with nonzero probability.

The Chao-Shen (CS) entropy estimator [49] is another entropy estimator, based on the Good Turing bias correction for the empirical estimation of the frequencies described in equation 4.30.

$$\hat{p}_i^{GT} = (1 - \frac{m_1}{n})\hat{p}_i^{ML} \tag{4.30}$$

where m_1 is the number of bins with n_i=1. This correction is used for estimating the entropy as in equation 4.31.

$$\hat{H}^{CS} = -\sum_{i=1}^{n} \frac{\hat{p}_i^{GT} log \hat{p}_i^{GT}}{(1-(1-\hat{p}_i^{GT})^n)} \tag{4.31}$$

The Bayesian approach for the entropy estimation consist on a Bayesian regularization of frequencies. The premise of Bayesian statistics is to incorporate prior knowledge, along with a given set of current observations, in order to make statistical inferences. The prior information could come from operational or observational data, from previous comparable experiments or from engineering knowledge. Using the Dirichlet distribution with parameters a_i as a prior, the resulting posterior distribution is also Dirichlet as described in equation 4.32.

$$\hat{p}_i^{Bayes} = \frac{n_i + a_i}{n + A} \tag{4.32}$$

where

$$A = \sum_{i=1}^{n} a_i \tag{4.33}$$

The consequent entropy estimator is defined in equation 4.34.

$$\hat{H}^{Bayes} = -\sum_{i=1}^{n} \hat{p}_i^{Bayes} log \hat{p}_i^{Bayes} \tag{4.34}$$

Depending on the choice of a_i, the Bayesian entropy estimator receives a different name as described in table 4.1 [111]. Note that when a_i=0, it corresponds to the empirical estimator. The most commonly used estimator is the Schürmann and Grassberger estimator, when $a_i = 1/n$, [267].

Table 4.1: Bayesian estimators and their prior definition.

a_i	Entropy estimator
0	Empirical (ML)
$\frac{1}{2}$	Krichevsky-Trofimov
1	Laplace's prior estimation
$\frac{1}{n}$	Schürmann-Grassberger
$\frac{\sqrt{n}}{n}$	minimax prior estimation

In addition, the NSB and James-Stein estimators are entropy estimators that can be seen as special Bayesian estimators. The NSB estimator proposed by Nemenman, Shafee and Bialek uses a prior that is a Dirichlet mixture with infinite components [222]. The James-Stein (JS) entropy estimator is based on a plug-in shrinkage type of estimation of the frequencies [111]. With certain shrinkage conditions, this estimation corresponds to a Bayesian estimator.

4.3.2 Taylor-based estimation

Information theoretic estimates are distributed according their probability density functions, as any other random variable. Occasionally, the knowledge of the entire distribution of information-theoretic measures is desirable, even for the statistical assessment of the estimates. Actually, finding an exact expression for the distribution of a mutual information estimator is a non-trivial problem because of its nonlinearity. An alternative for doing this is to approximate the expression for the mutual information by a second-order Taylor series [92]. Consider the mutual information as defined in 4.16. After expanding it into a second order Taylor series, the mutual information $I(X;Y)$ can be approximated as in equation 4.35.

$$\hat{I}(X;Y) = \frac{1}{2\text{ln}2}\sum_{x}\sum_{y}\frac{(p(x,y)-p(x)p(y))^2}{p(x)p(y)} \tag{4.35}$$

Note that this expression is similar to the χ^2 statistic variable described in equation 4.36.

$$X^2 = \sum_{x\in X}\sum_{y\in Y}\frac{(n_{ij}-(n_{i\cdot}n_{\cdot j})/N)^2}{(n_{i\cdot}n_{\cdot j})/N} \tag{4.36}$$

X^2 follows a χ^2 distribution with $(|X|-1)(|Y|-1)$ degrees of freedom when X and Y are independent.

Equations 4.36 and 4.35 lead to establish the relationship between $\hat{I}(X;Y)$ and X^2 as in equation 4.37.

$$X^2 = 2N\text{ln}2\hat{I}(X;Y) \tag{4.37}$$

Based on these assumptions, *Goebel et al.* demonstrated that the estimator of the mutual information between two independent or conditionally independent random variables ($\hat{I}(X;Y)$) follows a gamma distribution, whereas for dependent variables it follows a non-central gamma distribution [92].

If X and Y are independent random variables with $|X|$ and $|Y|$ realizations respectively, $\hat{I}(X;Y)$ follows a gamma distribution as in equation 4.38.

$$\hat{I}(X;Y) \sim \Gamma(\frac{1}{2}(|X|-1)(|Y|-1), \frac{1}{N\text{ln}2}) \tag{4.38}$$

with mean and variance given by equations 4.39 and 4.40 respectively.

$$E[\hat{I}(X,Y)] = \frac{(|X|-1)(|Y|-1)}{2N\text{ln}2} \tag{4.39}$$

$$E[\hat{I}^2(X,Y)] - E^2[\hat{I}(X,Y)] = \frac{(|X|-1)(|Y|-1)}{2N^2\text{ln}2^2} \tag{4.40}$$

If X, Y, Z are random variables with X and Y conditionally independent given Z, $\hat{I}(X;Y|Z)$ also follows a gamma distribution as in equation 4.41.

$$\hat{I}(X;Y|Z) \sim \Gamma(\frac{|Z|}{2}(|X|-1)(|Y|-1), \frac{1}{N\text{ln}2}) \tag{4.41}$$

If, in contrast, X and Y are statistically dependent, $\hat{I}(X;Y)$ follows a non-central gamma distribution as in equation 4.42.

$$\hat{I}(X;Y) \sim \gamma(\frac{1}{2}(|X|-1)(|Y|-1), \frac{1}{N\text{ln}2}, \lambda) \tag{4.42}$$

where $\lambda = I(X,Y)$ is the non-centrality parameter.

As mentioned, this approximation is useful when one needs to assess the independence of two random variables through the statistical significance of the mutual information between them. Given a significance level α (usually α=0.05), the significance of an observed value of the mutual information between two variables X and Y ($\hat{I}(X;Y)$) is obtained by comparing $\hat{I}(X;Y)$ with the corresponding quantile of the statistical distribution (here, the Gamma distribution). This comparison lead to a p-value that determines the significance as in equation 4.43.

$$p = \Gamma_{1-\alpha}(\frac{1}{2}(|X|-1)(|Y|-1), \frac{1}{N\text{ln}2}) \tag{4.43}$$

4.3.3 Nonparametric estimation and Information Theoretic Learning

Information Theoretic Learning (ITL) was introduced in 1999 in the context of adaptative filtering in machine learning [241]. This theory is based on nonparametric approximations of the Rényi's entropy and mutual information for solving problems of dimensionality reduction or feature extraction, among others.

In ITL, the probability densities are estimated using Parzen windowing with Gaussian kernels.

Let $H_\alpha(X)$ be the Rényi's entropy, which can also be written with an expectation operator as in equation 4.45.

$$\begin{aligned} H_\alpha(X) &= \frac{1}{1-\alpha}\text{log}\int_{-\infty}^{\infty} p_X^\alpha(x)dx \\ &= \frac{1}{1-\alpha}\text{log}E_X[p_X^{\alpha-1}(X)] \\ &\approx \frac{1}{1-\alpha}\text{log}\frac{1}{N}\sum_{j=1}^{N} p_X^{\alpha-1}(x_j) \end{aligned} \tag{4.44}$$

Given the Parzen estimate of the probability density function $p_X(x)$, defined in equation 4.45, the resulting Rényi's entropy is expressed as in 4.46.

$$\hat{p}_X(x) = \frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(x - x_i) \tag{4.45}$$

$$\begin{aligned}\hat{H}_\alpha(X) &= \frac{1}{1-\alpha}\log\frac{1}{N}\sum_{j=1}^{N}(\frac{1}{N}\sum_{j=i}^{N}\kappa_\sigma(x_j - x_i))^{\alpha-1} \\ &= \frac{1}{1-\alpha}\log\frac{1}{N^\alpha}\sum_{j=1}^{N}(\sum_{j=i}^{N}\kappa_\sigma(x_j - x_i))^{\alpha-1}\end{aligned} \tag{4.46}$$

The Rényi's entropy can also be expressed as in equation 4.48 as a function of the information potential defined in equation 4.47.

$$V_\alpha(X) = \int_{-\infty}^{\infty} p_X^\alpha(x)dx = \frac{1}{N^\alpha}\sum_{j=1}^{N}(\sum_{j=i}^{N}\kappa_\sigma(x_j - x_i))^{\alpha-1} \tag{4.47}$$

$$H_\alpha(X) = \frac{1}{1-\alpha}\log V_\alpha(X) \tag{4.48}$$

Principe et al. extend this nonparametric approximation for estimating the entropy to the estimation of quadratic mutual information [241].

This kernel-based framework for IT measures estimation allows the learning machine to learn not only directly from the data but also making use of information contained in the probability density function. This approximation is specially useful in the context of optimization. In particular, one of the most powerful optimization criteria is the maximum entropy principle (MaxEnt) enunciated by Jaynes [137].

4.4 Information Theory in Bioinformatics

Information theoretic estimations have been widely used with a large range of applications in statistical mechanics, physical sciences, economics and engineering, and more particularly in Bioinformatics [144]. Surprisingly, Shannon himself initiated his scientific research applying mathematics to study how different trait combinations propagate through several generations in his PhD thesis [271]. Later he shifted his focus towards the area of digital communications where he developed his theory of information. In molecular biology, the concept of information also plays a central role. Biologists frequently speak about information in different contexts, including information content in heritability, information conservation among the members of species, information content of the gene (exons, introns,...). In a certain manner, information is the subject of the so called Central Dogma of Molecular Biology, which states that the biologic information flows from DNA towards proteins. This section aims to review the applications of IT measures to bioinformatics. For an organization purpose, four main areas of research have been revised separately, such as DNA sequence analysis, gene mapping and proteins sequence analysis.

4.4.1 DNA sequence analysis

One of the earliest applications of IT measures to genetics has been the characterization of the information content along a DNA sequence or across several sequences. The information content in genomes is related with the randomness, or conversely, the regularity, of the sequence [2]. The entropy provides a quantitative, additive and conservative measure of information describing the statistical properties of the sequences [330]. The entropy is obtained by ascertaining the probabilities with which symbols are found on a sequence and it determines the average information per symbol for a given kind of DNA [91].

The information content of DNA sequences has been calculated for various organisms. The result of aligning several sequences is represented by a *consensus* sequence that contains the most common nucleotide or amino acid at each position within the sequence. *Schneider and Stephens* proposed a graphical representation of the information content in a sequence after its alignment with multiple sequences called *sequence logos* [264]. This representation is based on the measure of redundancy, that normalizes the decrease of uncertainty (or entropy) at a given site as in equation 4.49.

$$R = 1 - \frac{H}{H_{max}} \tag{4.49}$$

where H_{max} is the maximum entropy for the same number of states.

In a sequence logo, the total height at a given position is proportional to the information of the site, represented by the redundancy (or information) of the site, whereas the height of each nucleotide letter is proportional to its frequency (or conservation across sequences). Figure 4.3 shows an example of a sequence logo.

Studying the information content of genomic sequences has been useful for motif finding and especially in the context of the characterization of binding sites (BS), sites where transcription factors (TF) bind [292]. *Schneider et al.* first introduced the information content and more specifically the redundancy measure to the problem of TF-binding site recognition [265]. This constituted a theoretical framework followed by other researchers in this area of study [198, 233, 156].

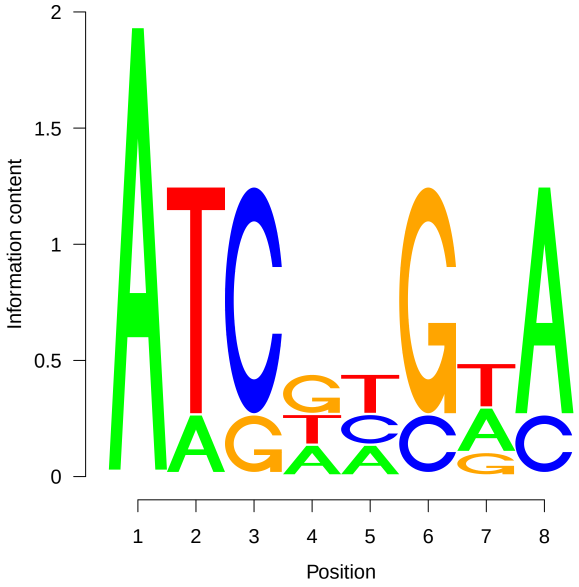


Figure 4.3: An example of a sequence logo.

Another problem related to DNA content recognition is the differentiation of coding and noncoding regions, i. e. the content recognition of coding regions. Several techniques were developed to extract information about the coding or noncoding status of the DNA. These techniques are generally based on finding differences in some statistical patterns between both [77]. It has been demonstrated that the mutual information in coding and noncoding sequences shows different patterns, being higher in coding regions than in noncoding regions [98]. This is an encouraging observation for the use of mutual information in DNA content recognition. However, this task involves using a distance measure for determining the similarity between a content sequence (e. g. an exon of a known genome) and an unknown sequence. Although it does not fulfill the conditions to be a metric, the mutual information can be converted to a bounded distance by normalizing it and subtracting it from one [102]. This approximation showed positive results obtaining a correct content recognition performance [69].

The problem of DNA classification is highly related to the DNA content recognition. Both of them rely on compression algorithms that themselves are based on distance measures. DNA classification aims to predict structural characteristcs of a test sequence of DNA given only the sequence of nucleotides, and no other information about the sequence [181]. DNA classification algorithms take advantage on content recognition techniques for identifying structural DNA properties from the comparison with other sequences. Traditionally this task has been approached using inductive learning methods such as artificial neural networks. These techniques attempt to form models from a collection of training data that can be used to predict future data. The majority of these methods constructs the classifier treating each nucleotide within a sequence separately. However, in DNA classification it is important to take into account the ordering of the nucleotides and the repetitions of subsequences. This is the reason why compression-based algorithms are known to be efficient. Using DNA compression, common subsequences are detected and replaced by shorter codes. Applying IT measures for compression purposes has been widely explored [344, 102, 69]. In this case the distance measure must be a metric so that other normalizations of the mutual information are proposed that guarantee the fulfillment of the triangle inequality [332, 102, 176, 333].

Most of these distance measures have been also used in other contexts such as the clustering of gene expression patterns [145]. Gene expression refers to the process by which information from a gene is use d in the synthesis of a functional gene product. The detection of associations between gene expression patterns and genotypes is a major challenge of current genetic research. It belongs to the area of study called *gene mapping.*

4.4.2 Gene mapping

The term gene mapping often refers to the detection of genes related to disease, or more specifically to the identification of genotypes (genes) related to phenotypes (traits). Gene mapping may be approached from different directions. This section reviews the different ways to discover candidate genes using IT measures.

Mutual information-based distances such as those used for DNA classification have been used for differentiating gene expression profiles [245, 152]. Moreover, there are also been used in clustering algorithms applied to gene expression data in order to identify gene expression patterns and associate them to some genetic variant [145, 242, 73].

IT measures have also been used in the context of analyzing gene-gene interactions or gene-environment interaction [327, 297]. This is specially interesting for the study of complex disorders, where it is known that the combination of several genes and environmental factors are responsible for disease. In particular, a novel measure of information, called k-way interaction information (KWII) is defined based on the entropy measure [48].

Finally, IT measures have also been used in association gene mapping. In particular measures of linkage disequilibrium (LD) have been proposed based on the mutual information measure [338, 334]. Moreover, the mutual information measure has been also used in genetic association, for measuring the direct correlation between a genetic factor (a genotype or gene expression levels) and a trait or a disease [68, 224, 299]. In particular, *Dawy et al.* use the mutual information for detecting direct association between genotypes and phenotypes [68].

4.4.3 Amino-acid sequence analysis

As well as for genetic sequences, IT measures have been applied to proteins' sequences of amino-acids. Mutual information based similarity measures have been applied to proteins for measuring common structures of different protein sequences or even common patterns within the same sequence. For instance, BLOSUM (BLOcks of Amino Acid SUbstitution Matrix) matrices are built using a scoring measure very similar to the mutual information [117, 78].

An IT framework has also been proposed for protein structure prediction, called the *GOR* method [90]. The GOR method studies the correlations between amino-acids within the same sequence, which is useful to detect the interactions between amino-acids that defines the secondary structure of the proteins [2].

Measures from information theory have also been applied in studies of protein-protein interaction networks, through semantic similarity measures. Semantic similarities between proteins are obtained using biomedical ontologies, namely the Gene Ontology (GO). They are mainly used to compare genes and proteins based on the similarity of their functions rather than on their sequence similarity [186, 314, 195, 196].

Part II

Materials and Methods

Chapter 5

Datasets

Before entering in detail with the original contributions of this thesis, the present chapter is disposed to describe the datasets involved in the development of the proposed methodology. In particular, the research presented in this dissertation has been carried out with two main datasets of different dimensions and characteristics. On one hand, the dataset provided in the 17th Genetic Analysis Workshop has been used for the characterization of a non-linear methodology for detecting genetic association, as described in chapter 6. The high dimensionality of this dataset has been exploited for comparing multiple association tests at a genome wide scale. On the other hand, the GAIT (Genetic Analysis of Idiopatic Thrombophilia) project dataset, provided by the Hospital de la Santa Creu i Sant Pau de Barcelona, has been used in the rest of the research. The most interesting characteristic of this dataset is its specific design for the study of thrombosis, including a large and varied collection of both genotypic and phenotypic data related to the disease. This make it suitable for the study of both polygenic and multiphenotypic scenarios, as it will be described in chapters 7 and 8 respectively.

5.1 The 17th Genetic Analysis Workshop dataset

5.1.1 The 1000 Genomes Project

The Genetic Analysis Workshops (GAWs) are a collaborative effort among researchers worldwide to evaluate and compare statistical genetic methods and relevant to current analytical problems in genetic epidemiology and statistical genetics. For each GAW, topics are chosen that are relevant to current analytical problems in genetic epidemiology, and sets of real or computer-simulated data are distributed to investigators worldwide. Results of analyses are discussed and compared at meetings held in even-numbered years. In particular, the 17th edition of the GAW was yielded in 2010. The data distributed for GAW17 is a “mini-exome” scan, using real sequence data for several hundred genes donated by the 1000 Genomes Project and simulated phenotypes.

The genomes of approximately 2 thousands individuals of each of the five major population groups (populations in or with ancestry from Europe, East Asia, South Asia, West Africa and the Americas) from different worldwide populations were sequenced. In particular the GAW17 mini exome is a selection of sequence variants designated as ‘functional’ and phenotypes simulated to produce a disease trait and related quantitative risk factors influenced by multiple genes with a variety of underlying genetic models. Although exome scans are becoming increasingly popular in complex disease genetics, GAW17 is many analysts’ first encounter with large scale exon sequence data and it provides opportunities to develop and test analytical tools and approaches that could shape the standards for analysis of the upcoming wave of exome data set.

Table 5.1: Geographical populations of the GAW17 dataset.

Ancestry	Population	Number of individuals	Total
European	Utah residents (CEPH - 1)	45	156
	CEPH - 2	45	
	Tuscan	62	
	Tuscan - 2	4	
Asian	Denver Chinese	87	321
	Denver Chinese - 2	20	
	Han Chinese - 1	25	
	Han Chinese - 2	36	
	Han Chinese - 3	48	
	Japanese - 1	31	
	Japanese - 2	41	
	Japanese - 3	33	
African	Luhya - 1	90	220
	Luhya - 2	18	
	Yoruba - 1	40	
	Yoruba - 2	47	
	Yoruba - 3	25	
TOTAL			697

5.1.2 Sample description

The GAW17 data contains two data sets. One contains the genotypes and phenotypes of 697 unrelated individuals, selected from the 1000 Genomes Project. In this dataset, there are 327 males and 370 females with ages from 16 to 91. The second dataset also consists of 697 individuals but organized in 8 extended families (351 males and 346 females with the same rank of ages that the first dataset). The 202 founders in the family dataset were chosen at random from the set of unrelated individuals.

The individuals of the GAW17 data belong to 13 geographical populations as described in table 5.1.

5.1.3 Genotypes

SNP genotypes were obtained from the sequence alignment files provided by the third pilot study of the 1000 Genomes Project [307]. The alignments were done using as the reference genome sequence a male human genome, for both male and female sequences. Some genotypes were missing due to incomplete sequence coverage in some individuals, because the 1000 Genomes Project genotypes were not phased. There is a total of 24487 SNPs, all of which are autosomal, located in 3205 genes. Many of the SNPs are rare variants, i.e. mutations with a low Minor Allele Frequency (only one mutated individual or equivalently a MAF of 0.07%).

5.1.4 Phenotypes

The phenotypes simulation model was build using 3 quantitative risk factors (Q1, Q2 and Q4), simulated as normally distributed variables. Disease affection was simulated using a liability model described in equation 5.1 with a 30% of affected people in the distribution. All SNP effects are additive on each of the four phenotypes. The design of the simulation model was based on the knowledge of biological pathways and statistical predictions regarding the potential deleteriousness of coding variants. A collection of genes, which sequence data was available in the 1000 Genomes Project, and that belong to particular pathways were selected. However, the phenotype simulations were done independently of the population origin of the 1000 Genome Project participants.

Genes influencing Q1 come primary from the Vascular Endothelial Growth Factor (VEGF) pathway, those influencing Q2 are primarily related to cardiovascular disease risk and inflammation, and those influencing latent disease liability also come primarily from VEGF (a different section of the pathway from the genes selected for Q1). Information on predicted deleteriousness was used to select functional variants. The functional variants include both rare and common alleles and a range of effect sizes, with most having small effects but a few having large effects that should be reliably detectable in most replicates of the data set. Some genes contain a single functional variant and others contain many. Environment was taken into account in the simulation model. There are environmental correlations between Q1 and Q2 and latent liability. Values of Q1 are higher in smokers. Q2 is not influenced by age, sex, or smoking. Q4 is lower in smokers, decreases with age, and is lower in females.

Whereas Q1 and Q2 are correlated with the latent liability to disease, Q4 is protective. A normally distributed latent liability trait was simulated using the model described in equation 5.1. This latent liability trait is also higher in smokers and increases with age.

$$\text{Liability to disease} = \text{latent liability} + Q1 + Q2 - Q4. \tag{5.1}$$

The genes and SNPs related to the quantitative traits, Q1 and Q2, and to the liability to disease are described in tables A.1, A.2 and A.3, shown in Appendix A.

The quantitative trait Q1 is influenced by 39 SNPs in 9 genes (table A.1). It can be observed that the Minor Allele Frequencies (MAFs) of the SNPs range from 0.07% (only one copy of the minor allele) to 16.5%.

13 genes containing 72 SNPs are influencing Q2 (table A.2). SNPs' MAF range from 0.07% to 17.07%. None of the genetic variants influencing Q4 are included in the GAW17 dataset. The liability to disease is influenced by 51 SNPs in 15 genes, with MAFs from 0.07% to 25.8% (table A.3).

The trait simulation was carried out 200 times, generating a total of 200 replicates for both data sets. The four simulated traits, Q1, Q2, Q4, and the affected as well as the smoking status varies across the replicates.

5.2 The Genetic Analysis of Idiopatic Thrombophilia Project

5.2.1 Thrombosis

The Genetic Analysis of Idiopatic Thrombophilia (GAIT) project is a modern family-based genetic study started in 1995 in the Hospital de la Santa Creu i Sant Pau, with the goal of discovering the genetic factors underlying thrombosis risk. Thrombosis is a one of the most morbid cardiovascular diseases. It is a common cause of mortality or morbidity in industrialized countries. It is known that the causes of thrombosis include environmental influences such as smoking or oral contraceptive treatment, as well as multiple genes with varying effects involved in determining the susceptibility to thrombosis ([285]). Ischemia and venous or arterial thromboses are complex diseases caused by a blood clot or an obstruction that blocks the blood circulation in a vessel (vein or an artery).

5.2.2 Sample description

The GAIT Project sample was recruited between 1995 and 1997 by the *Unitat d'Hemostàsia i Trombosis* of the Hospital de la Santa Creu i Sant Pau in Barcelona, Spain. The sample consists of 398 individuals in 21 extended spanish families. The inclusion criterion required the families to have more than 10 living individuals in at least 3 generations. 12 families were selected with a person affected of Idiopatic Thrombophilia. The remaining families were selected without regarding the phenotypes. Among the 398 subjects, 97 individuals were founders (subjects without parents in the study). The sample is approximately balanced in gender (46% males and 54% females). The age of the subjects ranges from less than one year to 88 years, with a mean of 37.7 years.

5.2.3 Genotypes

The GAIT sample was genotyped for both SNPs and microsatellites. On one hand, SNPs genotyping was performed using the Illumina Infinium 317k Beadchip. A total of 318 104 SNPs genotypes were measured. SNPs with a call rate lower than 90%, a minor allele frequency (MAF) lower than 2.5% and a false discovery rate when checking for Hardy-Weinberg Equilibrium of more than 20% were filtered out. After quality control, 266966 SNPs remained for analysis [38].

On the other hand, microsatellites were obtained using the ABI-Prism genotyping set MD-10, spaced at a density of 9.5 cMs. A total of 363 highly informative microsatellite DNA markers were typed. The PCR products were analyzed on PE 310, PE 377, and PE 3700 automated sequencers and were genotyped using the PE Genotyper software. The average heterozygosity of these markers was 0.79 [282].

There exist several approaches for dealing with missing values in genotype data. Most imputation methods proposed in the literature are based on using haplotypes for inferring missing values at a given locus.[66, 341]. Since it was not the main scope of this thesis and given that haplotypic information is not always available in SNP datasets, the chosen option in this thesis was another common strategy for dealing with missing values that consists on omitting the missing observation, assuming the cost of increasing the sampling error.

5.2.4 Phenotypes

During the coagulation process, a set of proteins in the blood plasma respond in cascade to form fibrin clots. These proteins are referred to as coagulation factors. The physiological cascade that underlies the pathological endpoint of thrombosis is complex, can be divided in different pathways, the coagulation and fibrinolysis pathways as described in figure 5.1. In particular, the coagulation cascade is divided itself in three pathways, the intrinsic pathway or contact activation pathway, the extrinsic pathway or tissue factor pathway, and the final common pathway. The fibrinolysis pathway acts in

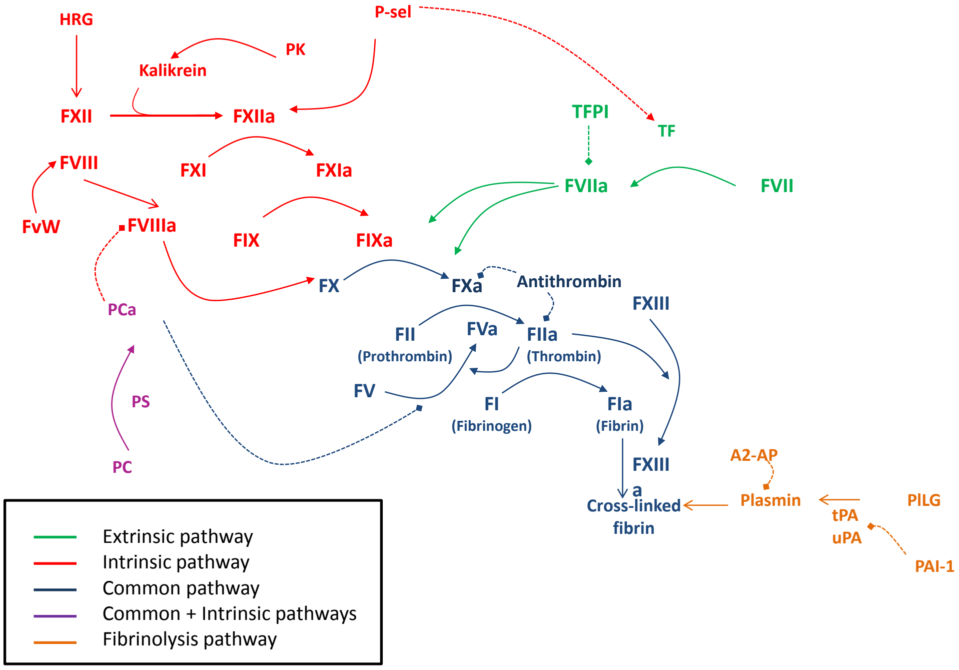


Figure 5.1: The coagulation cascade.

parallel of the coagulation pathway, and is also involved in the fibrin formation.

In addition of these phenotypes, other measures were recorded for each individual in the GAIT study including, among others, biochemical variables of the homocysteine metabolism, measures related to lipids, iron metabolism phenotypes or proteins of the complement system. In total, 85 phenotypes were available for each individual in the GAIT sample.

5.2.5 The coagulation factor VII

Levels of coagulation factors in blood represent a set of intermediate phenotypes that may be a good starting point for identifying genes involved in disease risk for thromboses and ischemia. It has been demonstrated that some of the coagulation factors have a genetic compound and show significant heritability ([285]). For example, Factor V Leiden is a variant of factor V produced by a mutation on the gene that codes this protein ($F5$). Factor V Leiden is the most common hereditary disorder of the coagulation process ([290]). It has also been published that coagulation Factor VII (FVII) has a genetic effect on disorders of hemostasis ([283]). The genetic variability in $F7$ gene is the most responsible for observed phenotypic variations in FVII levels. This is the reason why a part of this thesis has been developed focusing on the coagulation factor VII and the gene that codes for it, the *F7* gene (chapter 7). Extended genetic studies on the *F7* gene serve as a reference to validate the methodology proposed in this work. The *F7* gene is located at the segment 13q34, on the chromosome 13 of the human genome. It is about 13000 bases long among which around 50 polymorphisms have been identified as described in figure 5.2.

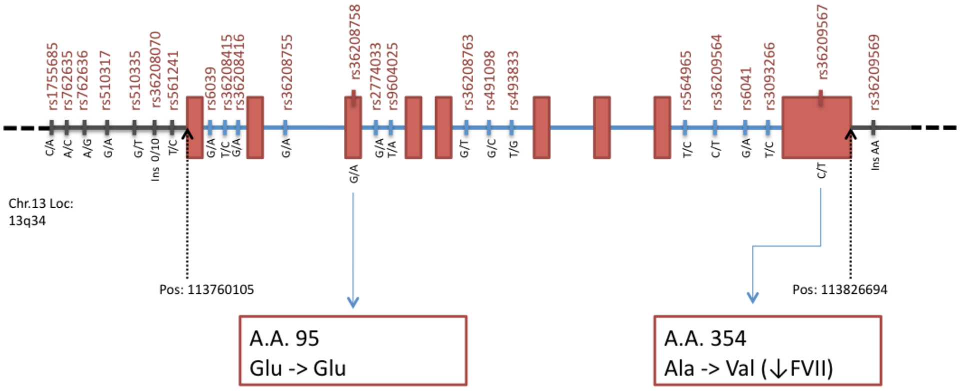


Figure 5.2: The *F7* gene polymorphisms.

One part of the study has been developed using the founders. For this sample, levels of the FVII phenotype follow a normal distribution with mean 121.6 and standard deviation 29. On the other hand, a sib-pairs analysis has been performed. The 345 pairs of sibs from the GAIT study were used in this part of the study.

Chapter 6

Single Point Genetic Association

As the name suggests, the main goal of genetic association tests is to detect association between genetic polymorphisms and traits or diseases. However, as it has been described in section 2.3.2, the causal-effect relationship between genotypes and phenotypes is not always straightforward. Actually in the majority of complex diseases, other two situations are often observed: multiple SNPs can be related with a same phenotype or a single locus can be related with several phenotypes. In this thesis, these three scenarios have been taken into account that correspond to the next three chapters.

This chapter focuses on studying the genetic study between polymorphisms and diseases looking at single point mutations, one-by-one. It has been approached from two different points of view. On one hand, a nonlinear test for one-locus genetic association studies was proposed, based on the mutual information measure. On the other hand, an exploratory study was carried out on the relationship between SNP variability among species and SNP association with disease, at different genetic regions.

6.1 A nonlinear test for genetic association

6.1.1 Linear versus nonlinear association measures

Genetic association studies consist on identifying genetic polymorphisms that are related with susceptibility to disease. A common procedure for genetic association is to apply correlation measures to relate the variability among individuals at a given SNP site with the phenotypic variability. Traditionally, this has been done using linear measures as it has been described in section 3.1.1.1.

However, these measures are only sensitive to linear relationships between two variables, so that they are not always able to capture all the possible types of correlation between a SNP and a phenotype.

When a genotype is correlated with a phenotype, the correlation is not necessarily linear, especially when the model of dominance is not the most common one, where one allele is dominant over the other. Figure 6.1 shows two examples of linear (a) and nonlinear (b) dependences between a SNP and a phenotype, corresponding to two different models of dominance (dominant and additive models respectively). The figure was built with real data corresponding to two SNPs of the F7 gene and the phenotype corresponding to the *FVII* levels in blood. In order to graphically represent the phenotypic variability given a genotypic status, box-plots were used to describe the dispersion of *FVII* levels for each allele combination. It is observed in Subfigure 6.1.a that *FVII* levels increase for individuals carrying the mutation. This follows a common model of dominance, where the mutation effect is dominant over the ancestral allele one. In this case a clearly linear tendency is observed between the two variables. In contrast, in Subfigure 6.1.b, it is observed that homozygous individuals for this particular SNP show higher levels of *FVII* in blood than heterozygous subjects. It corresponds to a particular additive model, where both alleles are co-dominant.

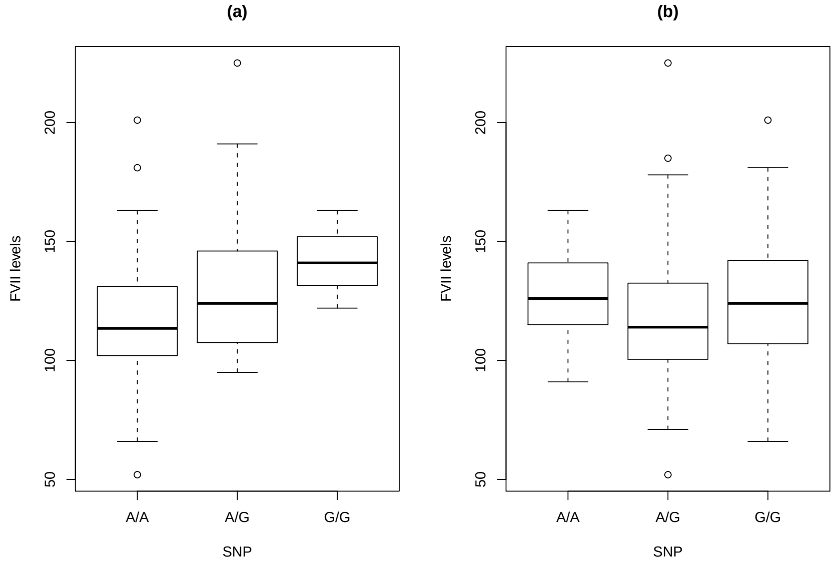


Figure 6.1: Examples of possible dependencies between a SNP and a continuous phenotype.

Linear regression models are defined under certain assumptions requiring the errors to be independent and identically distributed (i.i.d.), homoscedastic and normally distributed. When applying linear regression to genetic association, the most compromising requirement is the normality of the errors' distribution and their i.i.d.. Violation of these conditions often arise either because (a) the distributions of the dependent and/or independent variables are themselves significantly non-normal, and/or (b) the variables are not linearly correlated.

Linear regression is also sensible to extreme allele frequencies in genotypic data, corresponding SNPs where both alleles have nearly the same frequency or to to rare variants (those with a very low MAF). In particular it has been demonstrated that rare variants result in much false discovery rate in traditional genetic association studies than common variants [300]. When only one individual in the population carries one allele, it can be seen as an outlier and it can bias the results of association, making it significant while it is actually not, as illustrated in figure 6.2.

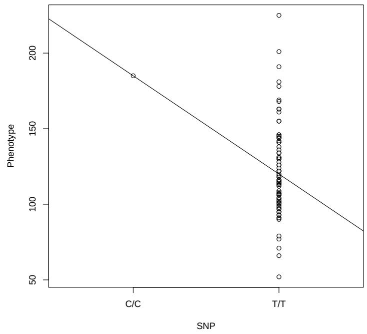


Figure 6.2: The impact of extreme MAFs in linear regression.

The interest on applying nonlinear measures to genetic association is aroused by the need of solving the aforementioned constraints of linear methods. One option to deal with these problems is to use the mutual information measure, a nonlinear correlation measure from information theory.

The main advantage of applying mutual information is that it detects both linear and nonlinear correlations between genotypes and phenotypes, without making any assumptions on the data. Figure 6.3 illustrates how mutual information captures both linear and nonlinear correlations between generic variables built in a synthetic problem, in comparison to linear correlation. It is observed that, when the two variables are linearly related both Pearson's correlation and mutual information measures detect the correlation (Subfigures 6.3.a and 6.3.b), whereas only the mutual information detects nonlinear relationships between two variables (Subfigures 6.3.c and 6.3.d). In Subfigures 6.3.e and 6.3.f, none of the measures detects any correlation.

It is known that most genetic association are very sensitive to outliers [300]. However, it has been shown that the sensitiveness of the mutual information measure to outliers is lower than for standard measures of dependencies between variables [242].

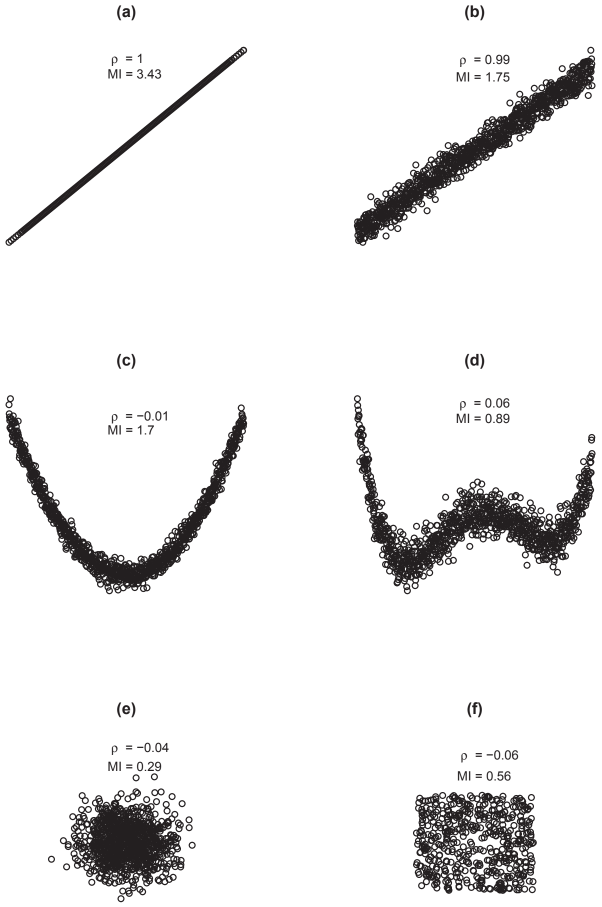


Figure 6.3: The Pearson correlation coefficient(ρ) and the mutual information (MI) in different cases. (a) and (b) reflects situations where the two variables are linearly correlated, (c) and (d) exemplify nonlinear correlations between the two variables and (e) and (f) reflect randomly related variables.

Another advantage of applying mutual information to genetic association is that it doesn't involve making any numerical representation of genotypic symbols, since it only needs to estimate the frequency of each item, independently of its nature.

6.1.2 The mutual information as a measure of genetic association

In the context of genetic association, the mutual information has been applied to measure the general correlation between a SNP and a phenotype.

The mutual information $I(S, Y)$ between a SNP S and a phenotype Y is defined in equation 6.1.

$$I(S, Y) = \sum_{S} \sum_{Y} p(S, Y) \log \frac{p(S, Y)}{p(S)p(Y)} = H(S) + H(Y) - H(S, Y) \quad (6.1)$$

Traditional genetic association studies have been developed with binary phenotypes corresponding to a disease affection state (affected/non affected). However for the study of complex diseases, one have to deal with intermediate phenotypes that are usually continuous variables. Since the mutual information was originally defined for discrete variables, its application to genetic association with continuous phenotypes implies estimating the expression defined in equation 6.2.

$$I(S, Y) = \int_{Y} \sum_{S} p(S, Y) \text{log} \frac{p(S, Y)}{p(S)p(Y)} dy \quad (6.2)$$

The most common approach for estimating $I(S, Y)$ is to partition the continuous values into discrete bins [291]. In any case, the discretization process overestimates the mutual information values. The different techniques of discretization can be divided into two main strategies. “Range discretization” methods distribute the samples into bins of equal width. In contrast, in “quantile discretization” each bin receives equal number of samples. The *bin-width* varies according to the data values it contains. In this case, the resulting histogram shows a rectangular shape, corresponding to a uniform distribution which may be away from the true distribution of the data. This equiprobability scenario maximizes the entropy values. In order not to overestimate $I(S, Y)$, this option was discarded. The most important aspect of range discretization is the definition of the number of bins of the resulting histogram or equivalently its bin width. Ideally, it should be chosen so that the histogram displays the essential structure of the data. The choice of that parameter is of difficult validation [86]. Several methods have been proposed for determining the optimal value of this parameter, such as the Sturges' rule [295] or the Scott's Rule [268]. However it has been suggested that even if these rules lead to a good performance of the histogram, they are not consistent to large samples. In order to overcome this inconsistency, an extension of the Scott's rule has been proposed, based on kernel density estimates [312].

Given $X = \{X_1, X_2, ..., X_n\}$ a random variable and $f(X)$ its density function, Scott's formula states the following rule for determining the optimal bin-width of an histogram $\hat{h}$ (equation 6.3).

$$\hat{h} = \frac{3.49\hat{\sigma}}{n^{\frac{1}{3}}} \quad (6.3)$$

where $\hat{\sigma}$ is an estimate of the standard deviation. *Wand* presented an estimation of the optimal bin-width, the one which maximizes its asymptotic performance, with an optimal Mean Integrated Square Error (MISE). For doing this, the Scott's formula described in equation 6.3 is expressed as in equation 6.4.

$$\hat{h} = (\frac{6}{-\psi_2 n})^{\frac{1}{3}} \tag{6.4}$$

where

$$\psi_r \equiv E\{f^{(r)}(X)\} = \int_{-\infty}^{\infty} f^{(r)}(x)f(x)dx \tag{6.5}$$

Then, a kernel density estimation of ψ_r is carried out as in equation 6.6.

$$\hat{\psi_r}(g) = n^{-2}g^{-r-1}\sum_{i=1}^{n}\sum_{j=1}^{n} K^{(r)}\{(X_i - X_j)/g\} \tag{6.6}$$

where K is an $r-th$ order kernel function and g the bandwidth of the kernel, which is estimated using a normal scale estimator of ψ_r [273].

Generally, this kernel-based method discretizes the data in a quite high number of bins that guarantee obtaining a good approximation of the real structure of the data. However, $I(S,Y)$ is sensitive to this number of bins. The highest is the number of bins, the most the mutual information is overestimated [100].

6.1.2.1 The problem of finite samples

In addition to the overestimation of $I(S,Y)$ due to the phenotype discretization, the mutual information is also sensitive to finite sample effects. In particular, when the sample size is finite, the probability density functions need to be estimated, adding a positive bias on the information theoretic quantities. This is a critical issue when applying the mutual information measure to real data. Several approaches have been proposed for solving or correcting this effect, as described in chapter 4. In this section a more detailed description of the strategy used in this thesis is exposed. $I(S,Y)$ was implemented using three estimation methods. Among the several approaches for the estimation of $I(S,Y)$ described in section 4.3.1, few of them were selected for a comparative analysis. In particular the chosen procedures were the empirical probability estimation, the Miller-Madow estimator, which uses an asymptotic correction of first order bias, and the Schurmann-Grassberger estimator, a Bayesian estimator of information theoretic measures using a Dirichlet probability distribution for the estimation of the frequencies as a prior.

The first approximation used for estimating $I(S,Y)$ was the empirical estimation of the frequencies $p_i = p(x_i)$, which corresponds to the classical inference of the frequencies from the countings of the observations. This is

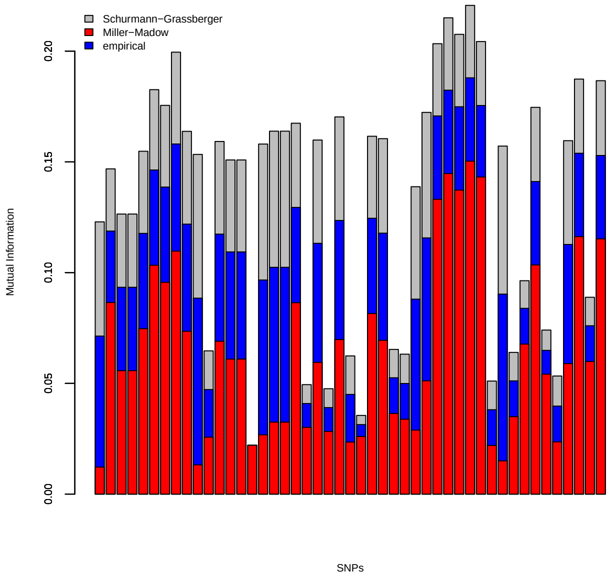


Figure 6.4: The mutual information of polymorphisms in the F7 gene, against the phenotype (*FVII* levels in blood) with different estimation techniques.

the most standard procedure and the simplest as well. Note that it does not apply any correction for solving the finite sample size problem.

The empirical estimator of the mutual information $(\hat{I}^{ML}(S, Y)$ is obtained as described in equation 4.28. The Miller-Madow estimator for the entropy measure is defined as in equation 4.29. The Schurmann-Grassberger (SG) estimator was selected among the family of Bayesian estimators for information-theoretic measures described in section 4.3.1, since it is the most commonly used. Among the Dirichlet priors enumerated in table 4.1, the estimation of the frequencies as defined in equation 4.32 for the SG estimator results as in equation 6.7.

$$\hat{p}_i^{SG} = \frac{n_i + 1/n}{n + A} \tag{6.7}$$

The resulting estimator of the entropy is expressed in equation 6.8.

$$\hat{H}^{SG} = -\sum_{i=1}^{n} \hat{p}_i^{SG} log \hat{p}_i^{SG} \tag{6.8}$$

Figure 6.4 compares the mutual information of the SNPs of the F7 gene against the *FVII* levels in blood, for the three selected estimations. It is observed that for almost every SNP, the highest value of mutual information values is obtained with the SG estimator whereas the lowest one corresponds to the MM estimator, where the empirical estimation falls in-between the two former estimators. This suggests that the SG is the most sensitive to overestimations, whereas the MM estimator reduces these effects substantially, even underestimating MI values.

In any case, a statistical test is necessary to determine if a positive mutual information value denotes a true association or it is only due to finite sample size effects or to the discretization process. This statistical test assumes a null hypothesis as true, in this case the hypothesis of non association between a SNP and a phenotype. A test statistic is then computed in function of the data. The sampling distribution of this statistic is called the null distribution and allows to compute p-values, which indicate if the null hypothesis should be accepted or rejected. A p-value lower than the significance threshold (generally 0.05) indicates that the null hypothesis should be rejected.

Often the most difficult part of applying a statistical test is to determine the null distribution. Most parametric tests are easily describable mathematically. However, one of the main drawbacks of the proposed nonlinear statistical tests is that the null distribution is unknown. The best analytical approximation for the null distribution (D_0) of $I(S,Y)$ has been proposed in *Dawy et al.* as expressed in equation 6.9 [68].

$$D_0 \sim \Gamma(k,\theta) \tag{6.9}$$

where $k = \frac{1}{2}(|S|-1)(|Y|-1)$ is the shape parameter of the gamma distribution and $\theta = \frac{1}{N \ln 2}$ is its scale parameter and $|S|$ and $|Y|$ are the number of symbols of S and Y respectively.

As it is described in equation 6.9, this distribution is modulated by the number of genotypic symbols which is straightly related to the MAF. The number of genotypic symbols decreases when the MAF is low, specially for finite samples. However, most of the heritability not explained in genetic association studies is caused by rare variants, so that SNPs with very low MAFs are of great importance and should be treated carefully [177]. The analytical approximation proposed by *Dawy et al.* does not contemplate the case of rare variants, so that this approximation is not always suitable in the context of genetic association.

Szymczak et al. propose a permutation-based procedure to generate an empirical null distribution, which seems to be preferable [299]. This approach consists on building an empirical distribution from permutations of the original data and obtaining a significance level associated to the original mutual information value comparing it to this distribution. The resulting p-value indicates how significantly positive the mutual information is [86].

Since the permutation-based approach is computationally expensive, a synthetic experiment was developed in order to adjust D_0 to the fittest analytical probability distribution. D_0 was obtained by generating random copies of one of the two variables. In order to preserve the allelic frequencies of the random sample, those were obtained by surrogating the SNPs, destroying its individual order so that the allelic frequencies are respected. For this experiment, D_0 was obtained using the empirical estimation of $I(S,Y)$ ($\hat{I}^{ML}(S,Y)$).

In order to adjust an analytical expression of D_0, two parameters were taken into account in the experiment, the allelic frequencies and the number of bins in the discretization process. Random SNPs with different allele frequencies going from 0.1% to 50% were generated, as well as a random gaussian phenotype that was discretized with different numbers of bins going from 5 to 15.

Figure 6.5 shows a particular case of D_0 generated with permutations of the samples using surrogate data. The best fitted gamma distribution is also shown (dashed line). In this case, it is observed that the gamma distribution is a good adjustment of the empirical null distribution of the mutual information. This figure was generated using a synthetic SNP with a MAF of 0.005 and a phenotype discretized to 8 bins.

However, the null distribution is strongly dependent on both the MAF of the SNP and the phenotype discretization number of bins. Figure 6.6 shows the dependence of the mutual information null distribution to the Minor Allelic Frequencies of the SNPs. The gamma distributions fitted to

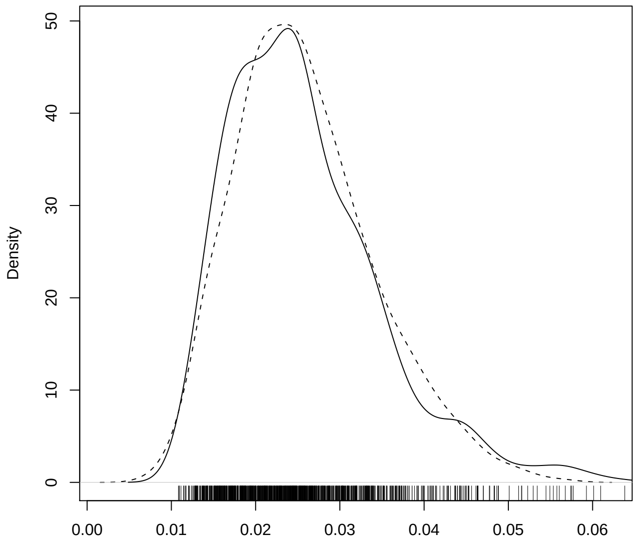


Figure 6.5: The mutual information null distribution. The solid line represents the empirical null distribution and the dashed line corresponds to an adjusted gamma distribution. Here the shape and scale of the gamma distribution were fitted to $k = 9$ and $\theta = 346$ respectively.

the empirical null distributions corresponding to SNPs with different allelic frequencies were generated.

It is observed that the distribution varies in function of the MAF. However, as shown in figure 6.7, the relationship between the parameters of the gamma distribution and the allelic frequencies does not follow any clear pattern. Figure 6.7 shows that for low MAFs, both the shape and the scale parameters are out of the range of values obtained for the remaining MAFs. This indicates that this approximation is not suitable to this problem.

On the other hand, the dependence of the mutual information null distribution to the number of bins in the discretization of the phenotype was studied. Null distributions were obtained for phenotype discretization with a number of bins going from 5 to 15. It is observed in figures 6.8 and 6.9 that in a similar manner than with the allelic frequencies, the mutual information null distribution depends on the number of bins in the discretization of the phenotype. On one hand, it is observed that the number of bins has a lower impact on the appearance of the null distribution than the MAF, but in the same manner than for the MAF, this dependence can not be explained with simple models.

It has been shown that it is not possible to characterize an analytical expression for the null distribution of the mutual information, especially for SNPs with low MAFs. Thus, it is still more convenient to use the empirical

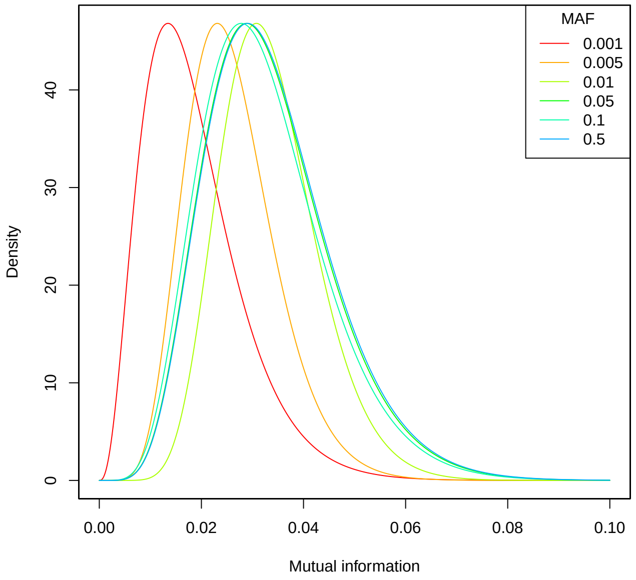


Figure 6.6: The mutual information null distribution for different allelic frequencies.

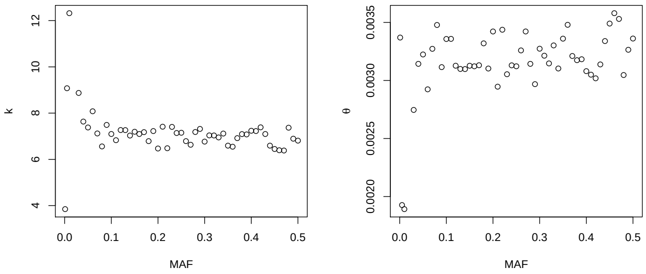


Figure 6.7: The relationship between the shape k and scale θ parameters of the distribution of mutual information against the Minor Allelic Frequencies (MAFs).

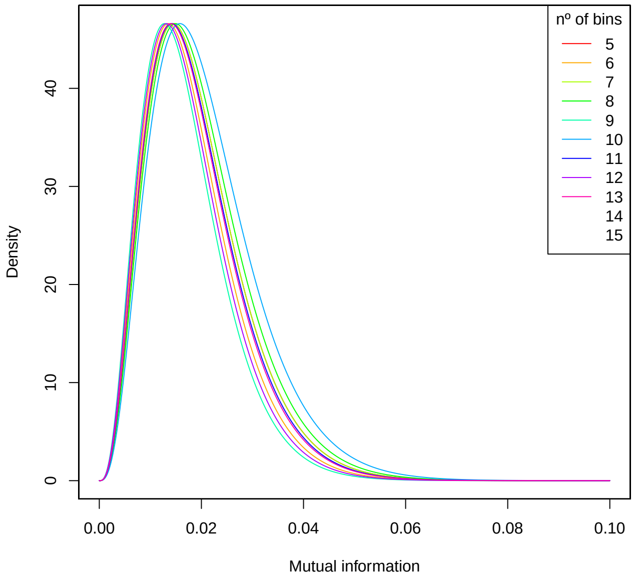


Figure 6.8: The mutual information null distribution for different binnings of the phenotype.

null distribution of the mutual information by using surrogate data at the cost of increasing the computing time.

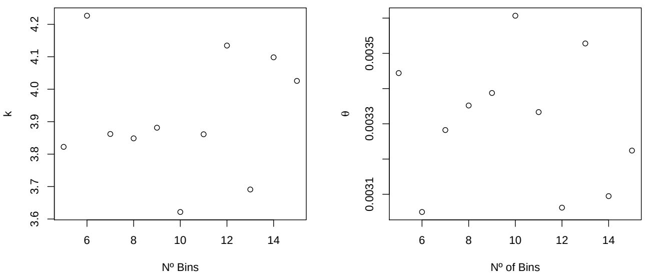


Figure 6.9: The relationship between the shape k and scale θ parameters of the distribution of mutual information against the number of bins in the discretization of the phenotype.

6.1.3 A mutual information-based test of genetic association for stratified populations

As described in section 3.4.2, the population structure may be taken into account in the genetic association test. In the same manner that the population structure P is introduced in the linear regression tests of association, as described in equation 3.21, a mutual information-based test is proposed that takes P into account. This test is based on the conditional mutual information measure.

The conditional mutual information $I(S,Y|P)$ measures the association between the genotypes S and the phenotypes Y, at each level of P as expressed in equation 6.10.

$$\begin{aligned} I(S,Y|P) &= \textstyle\sum_P p(P) \sum_S \sum_Y p(S,Y|P) log \frac{p(S,Y|P)}{p(S|P)p(Y|P)} \\ &= H(S,P) + H(Y,P) - H(S,Y,P) - H(P) \end{aligned} \tag{6.10}$$

This stratification of $I(S,Y)$ allows to detect the association between the genotype S and phenotype Y even if they are both conditioned by a third variable, in this case the population structure P. Computing the conditional mutual information as defined in equation 6.10 is equivalent to make a pondered sum of $I(S,Y)$ within each subpopulation as in equation 6.11.

$$I(S,Y|P) = \sum_{i=1}^{n_p} p(P_i) I(S,Y|P = P_i) \tag{6.11}$$

where n_p is the number of subpopulations. $I(S,Y|P)$ is also affected by a bias due to finite sample size, introduced as many times as $I(S,Y)$ is computed (here n_p times). Thus, the use of a statistical test for ensuring the veracity of an association is even more justified in the case of the conditional mutual information. In order to assess the statistical significance of the association measured by $I(S, Y|P)$, the following statistical test based on surrogate data is proposed.

For each candidate SNP S_j, the test of genetic association with the phenotype Y consists on computing $I_j = I(S_j, Y|P)$ and comparing it to the mutual information null distribution $D_0 = \{I(S_r; Y|P) : r = 1..N_c\}$ given by the mutual information of N_c random copies S_r of the S_j obtained using surrogate data. A kernel density estimator is applied on this vector of mutual information values in order to estimate D_0. The resulting p-value indicates if the association measured with the mutual information is statistically significant. The p-value is obtained by integrating the resulting D_0 density function from the I_j value and comparing it with the total area under the D_0 density function curve. For reaching a significance level of 0.05, the length of the D_0 has been set to $N_c = 1000$.

6.1.4 Methodology

The nonlinear test proposed previously was applied to the GAW17 dataset. Validating a genetic association test is an intricate task, since it is difficult to ascertain if an association really exists or it has been found by chance. This is called the jackpot effect [97]. Actually, the only way to know if a positive result obtained with a new association technique is reliable is through a functional analysis in the laboratory. In order to know if a test correctly detects a polymorphism with an effect on a phenotype, an alternative is to make use of synthetic models that emulate the biological relationships between genes and phenotypes. This is the main goal of using the GAW17 dataset, since both the simulation model and the SNPs related to the phenotypes are known. In particular, the proposed association test was applied to the $Q1$ phenotype, since it is the phenotype with more associated SNPs. The list of the SNPs related with the phenotype $Q1$, included in the simulation model of the GAW17 dataset is described in table A.1 (appendix A).

6.1.4.1 Pre-processing

This study was carried out using the first replicate of the GAW17 dataset containing unrelated individuals. In order to ascertain the population structure (P), the kinship matrix of genetic distances between individuals was obtained using Identity-By-State methods [29]. Multidimensional-scaling was applied to the similarity matrix defined by IBS distances in order to obtain a low-dimensional representation of the individuals. In this case, the number of dimensions was set to 8. In the new bi-dimensional vectorial space, the individuals were represented by vectors of principal coordinates. Then, a k-means clustering algorithm was applied to group subjects into separate clusters. The number of clusters that best represent the structure of the data was obtained by looking at three clustering indexes, the Dunn index, the Hubert's gamma coefficient and the WB ratio [103]. The Dunn index attempts to identify compact and well separated clusters by maximizing the ratio between the minimum separation and the maximum diameter of the clusters. The Hubert's gamma coefficient measures the correlation between distances. It takes values from 0 to 1, where 0 means that clusters are not separated and 1 means that clusters are well-separated. Finally, the WB ratio measures the quotient between the similarity average within a cluster and the similarity average between clusters.

Figure 6.10 shows these indexes in function of the number of clusters used in the k-means algorithm. It is observed that the Dunn's index was maximized and the WB was minimized for 6 clusters, whereas the Hubert's gamma coefficient was maximized for 3 clusters. According to other works published in the proceedings of the GAW17, the balance was decanted to use 3 clusters [112, 21].

Figure 6.11 represents the individuals in the new vectorial space, organized in 3 clusters. Figure 6.11 was obtained with a bi-dimensional PCA for visualizing the clustering of the MDS of 3 dimensions. It is observed that this configuration allowed to separate the subjects depending on their ancestry.

The resulting vector that indicates the subpopulation (P) to which the individuals belong was included in the association analyses, considering it

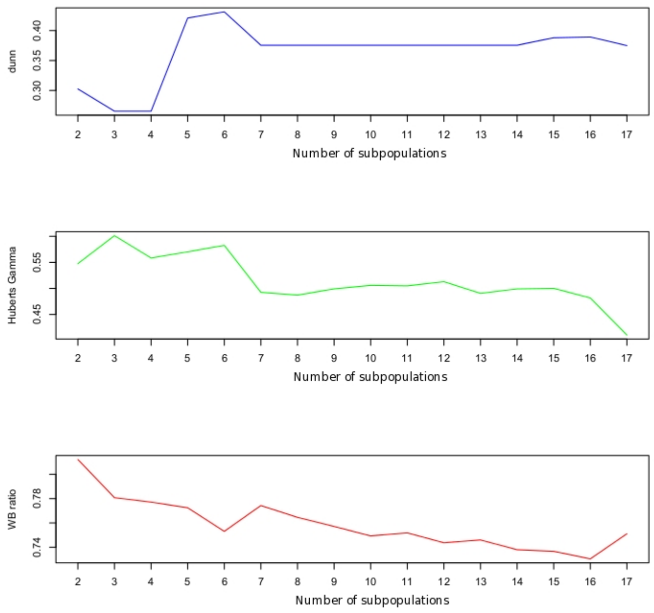


Figure 6.10: Clustering statistics for different number of subpopulations.

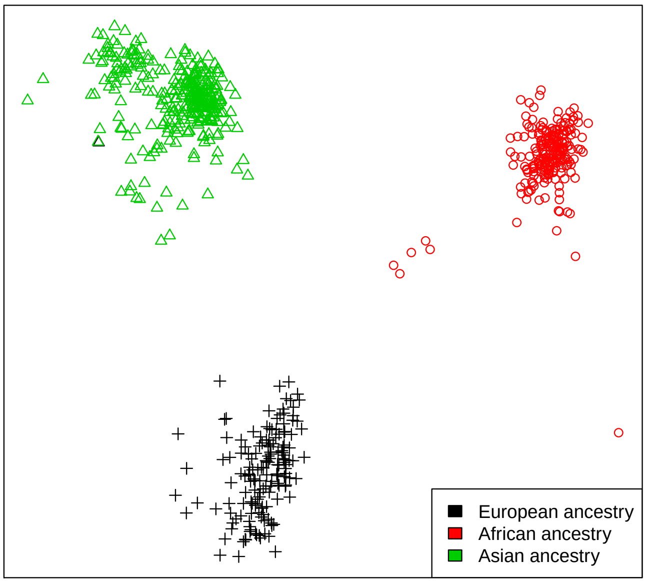


Figure 6.11: Graphical representation of the individuals and their membership to the 3 clusters.

as a covariate in the linear regression model, and as a stratification variable using the conditional mutual information measure $I(S, Y|P)$.

In addition, the data was subjected to a quality control procedure, excluding individuals with individuals with too high autosomal heterozygosity (FDR rate threshold of 0.01) and with too high IBS (IBS threshold of 0.99. SNPs that were out of Hardy-Weinberg Equilibrium were also discarded (FDR threshold of 0.1). Finally, a total of 684 individuals and 19915 SNPs were used for the study.

6.1.4.2 Genetic association tests

The nonlinear mutual information-based genetic association test proposed in section 6.1.2 was applied using the three aforementioned estimation methods for $\hat{I}(S, Y|P)$, the empirical estimation (EMP), the Miller-Madow (MM) asymptotic correction of the empirical estimation and the Schurmann-Grassberger (SG) estimator. The resulting three nonlinear tests were compared with the traditional linear test contemplating population stratification based on mixed linear regression models proposed in the GenABEL software (GA) as described in equation 3.21 [16]. This association test returns a p-value corresponding to the significance of the linear regression between the genotype and the phenotype, taking into account the population stratification.

In summary, four tests of association (three nonlinear tests and the conventional linear test) were applied to the GAW17 dataset and in particular for testing the association of SNPs with the phenotype $Q1$. The $Q1$ phenotype was deprived of the variability due to age, gender and smoking status. through a linear regression model (equation).

The residuals from the linear regression model described in equation 6.12 were used as the phenotype in the following association scans, since theyonly contain the variability due to the $Q1$ phenotype. The $Q1$ phenotype was discretized with the optimal number of bins obtained using kernel-based methods described in section 6.1.2, corresponding to 16 bins.

$$Q1 \sim \text{sex} + \text{SMOKE} + \text{age} \tag{6.12}$$

6.1.5 Results and discussion

The results obtained with the four proposed tests of association for the $Q1$ phenotype of the GAW17 were contrasted with the corresponding answers found in table A.1. Since the answers were already known, whenever the test detected an association it was possible to determine if this association was correctly detected or not. Table 6.1 lists the SNPs correctly associated with the $Q1$ phenotype, for each of the four methods.

Table 6.1: SNPs correctly detected by the four test.

Method	SNPs	Gene	SNP type	Chromosome	MAF	p-value
EMP	C4S1884	KDR	Nonsynonymous	4	0.020803	$3.48 \cdot 10^{-2}$
	C4S4935	VEGFC	Nonsynonymous	4	0.000717	$4.7 \cdot 10^{-4}$
	C13S522	FLT1	Nonsynonymous	13	0.027977	$1.12 \cdot 10^{-17}$
	C13S523	FLT1	Nonsynonymous	13	0.066714	$1.03 \cdot 10^{-17}$
	C14S1734	HIF1A	Nonsynonymous	14	0.012195	$2.13 \cdot 10^{-2}$
MM	C13S399	FLT1	Nonsynonymous	13	0.000717	$4.63 \cdot 10^{-2}$
SG	C4S4935	VEGFC	Nonsynonymous	4	0.000717	$4.7 \cdot 10^{-4}$
	C13S522	FLT1	Nonsynonymous	13	0.027977	$1.56 \cdot 10^{-17}$
	C13S523	FLT1	Nonsynonymous	13	0.066714	$8.07 \cdot 10^{-19}$
	C14S1734	HIF1A	Nonsynonymous	14	0.012195	$1.6 \cdot 10^{-2}$
GA	C4S1884	KDR	Nonsynonymous	4	0.020803	$5.7 \cdot 10^{-3}$
	C4S4935	VEGFC	Nonsynonymous	4	0.000717	$1.4 \cdot 10^{-4}$
	C13S522	FLT1	Nonsynonymous	13	0.027977	0
	C13S523	FLT1	Nonsynonymous	13	0.066714	0
	C14S1734	HIF1A	Nonsynonymous	14	0.012195	$3.2 \cdot 10^{-3}$

It is observed that the best results were obtained when using the nonlinear test based on the empirical estimation of the mutual information (EMP) and the GenABEL software (GA). Both tests detect association for exactly the same SNPs, confirming that the proposed nonlinear method is able to replicate the results obtained with standard linear regressions. In both cases, only 5 of the 39 SNPs associated with the phenotype $Q1$ were detected. Figure 6.12 shows a qualitative description of the p-values of all the positives SNPs for both the EMP and GA tests as well as their MAF.

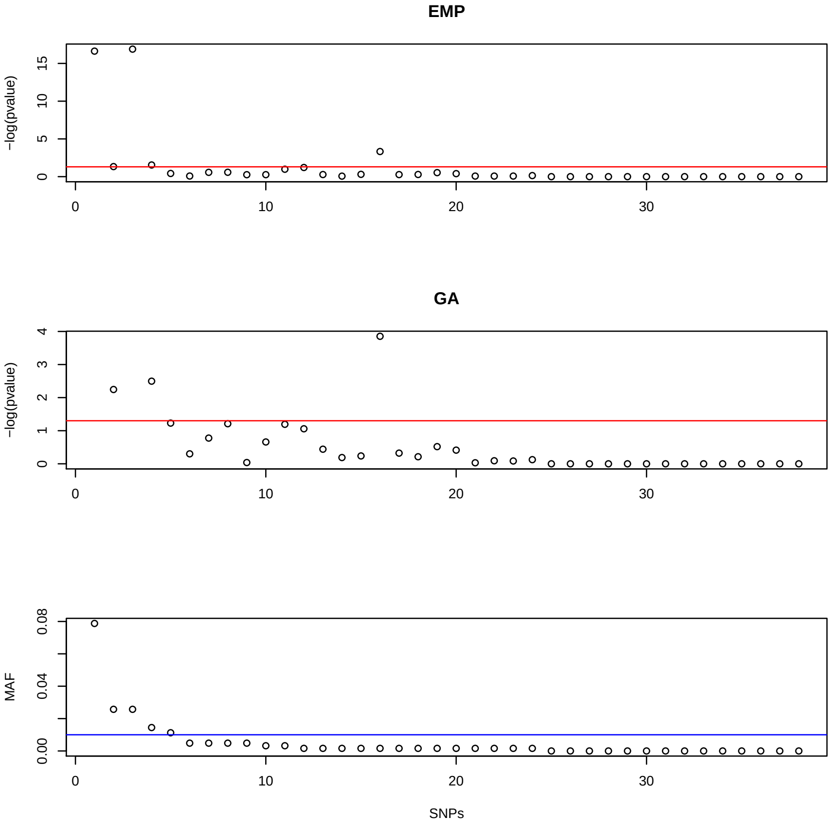


Figure 6.12: p-value and MAF distribution of the SNPs associated with the phenotype $Q1$ for the proposed nonlinear test with the empirical estimation of the mutual information measure (EMP) and for the conventional linear association test (GA).

From top to bottom, the first two figures represent the negative logarithm of the p-values for each test respectively, as well as the threshold corresponding to a p-value of 0.05, represented by the red line. The third figure shows the MAF of each SNP, with a threshold line set to 0.01, the value that discriminate current polymorphisms with rare variants. The SNPs were ordered according to their MAF.

Figure 6.12 shows a clear correspondence between the MAF of the SNPs and their p-value with both methods. In particular, it is observed that significant SNPs correspond to SNPs with a MAFs higher than 1% and conversely, SNPs presenting extremely low MAFs are not detected by any of the tests.

SNP $C4S4935$ is an exception, since it has been detected, even with a very low MAF (0.000717). This variant belongs to the $VEGFC$, being the

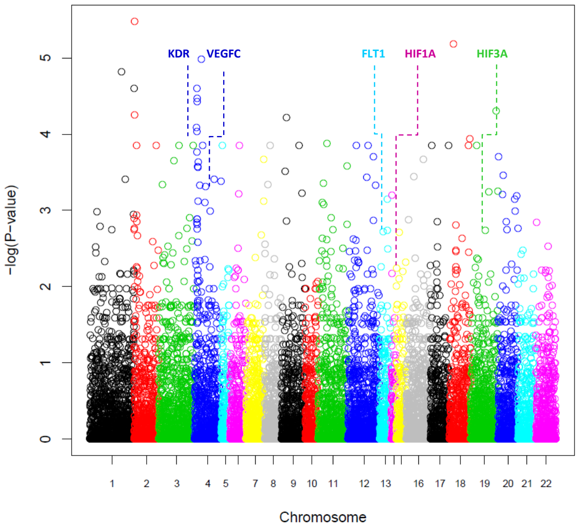


Figure 6.13: Manhattan plot for the nonlinear genetic association test based on the empirical estimation of the mutual information.

only variant found in this gene that had an important role in the simulation model for the $Q1$ phenotype of the GAW17 dataset. In particular, the $Q1$ phenotype model was build with genes of the Vascular Endothelial Growth Factor (VEGF) pathway. Thus, this variant should have a clear relation with the phenotype.

It is also observed that the significance levels were similar for both tests for the correctly detected SNPs.

Figure 6.13 shows the results obtained with the proposed nonlinear test using the empirical estimation of the mutual information. Peaks of significance were found in chromosomes 4, 13 and 14 corresponding to the true positives presented in table 6.1, concretely the genes *KDR*, *VEGFC*, *FLT1* and *HIF1A*. It is worth mentioning that the nonlinear test also detects significance in chromosome 19. In particular, SNP $C19S4840$ (MAF 0.0007) was detected, located in the *HIF3A* gene. Even if this particular SNP is not associated with the $Q1$ trait, the *HIF3A* gene appears in table A.1, suggesting that SNP $C19S4840$ could be in LD with one of the SNPs belonging to *HIF3A* appearing in table A.1. In addition, the *HIF3A* gene belongs to the VEGF pathway.

An advantage of using the GAW17 dataset is that the characterization of the nonlinear test can be seen as a classification problem since it is done by counting how many positives and negatives are correctly detected and

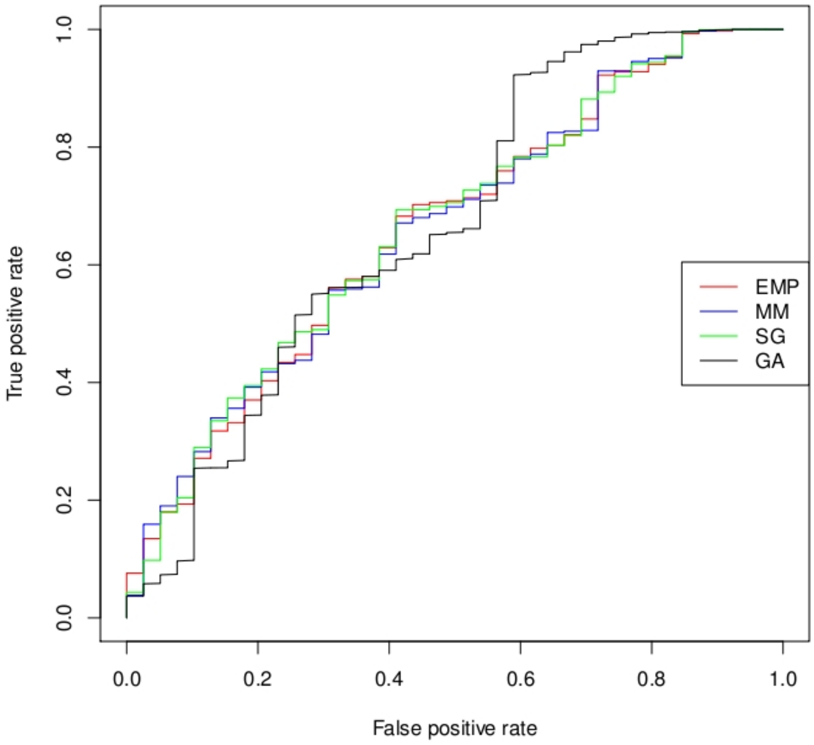


Figure 6.14: ROC curve of the four association tests.

discarded respectively. In particular, the true positives correctly detected are called true positives (TP), whereas the positives not detected (those classified as negatives) are called false negatives (FN). In contrast the negatives detected as positives are called false positives (FP) and the negatives correctly detected are called true negatives (TN). Based on these countings, the concepts of *sensitivity* and *specificity* are often used for evaluating the performance of the classification procedure. The sensitivity, also known as the true positive rate (TPR), is the proportion of detected positive SNPs (TP) with all the real positive SNPs (TP+FN). It can be seen as the probability that the test is positive given that the SNP is associated with the phenotype. The relationship between sensitivity and specificity, as well as the performance of the classifier, can be visualized with a Receiver operating Characteristic (ROC) curve, a graphical plot of the sensitivity, or true positive rate versus the false positive rate (1 - specificity), for a binary classifier system as its discrimination threshold is varied.

A more general analysis of the performance of the four association tests

Table 6.2: Performance in classification of the four methods. TP and (TP+FP) were counted for a significance threshold of 0.05.

Method	TP	TP+FP	AUC
EMP	5	724	0.6592
MM	1	736	0.6599
SG	4	725	0.6630
GA	5	839	0.6661

was carried out. The ROC curves of each test were generated. The accuracy of each method was measured by the area under the ROC curve. An area of 1 represents a perfect test, whereas an area of 0.5 represents a worthless test. Figure 6.14 shows the ROC curves of the four tests, while table 6.2 shows the number of true positives, the number of SNPs detected as positives (TP +FP) found for each method for a significance threshold of 0.05, as well as the AUC of each method.

It is observed that for a significance level of 0.05, the method that obtained less false positives was the proposed nonlinear method with the empirical approach for the estimation of the mutual information measure. However, it is observed that, in terms of the AUC, none of the tests presented a good performance, obtaining in the four cases similar results ($AUC \sim 0.66$).

6.2 Effect of genetic regions on the correlation between single point mutation variability and morbidity

6.2.1 The importance of SNP location

The functional class of a polymorphism, the genomic regions where it occurs or the comparison with other species may provide useful information for characterizing its influence on physiological affections [28, 135, 249]. In particular, *Adie et al.* propose using sequence based features for prioritizing SNPs to study cross-species sequence similarities for identifying relevant SNPs [3]. The availability of multiple genomic sequences of different model organisms has made it possible to ascertain information about the selective pressure of polymorphisms [126, 36]. For example, it has been demonstrated that functional regions of the genome are preserved in different organisms throughout evolution and thus present a low variability across species [40, 342, 185, 183]. Nowadays, the evolutionary conservation of genetic sequences has been incorporated into prioritization tools [341, 223].

Even though the sequence variability across different organisms has been studied widely and correlated with functionality, studies based on morbidity are rare [40, 165, 30]. Note that SNP functionality refers to the altering effect the mutation has on the resulting protein, whereas SNP morbidity refers to its association to disease. The standard practice for studying morbidity consists of comparing two groups of SNPs, generally deleterious SNPs (SNPs related to disease) against neutral SNPs (SNPs for which no associations to disease are known).

Most studies on cross-species variability focus on SNP functionality and, more specifically, on functional SNPs [3, 126, 40, 129, 4]. SNP functionality could bias the results if it is not taken into account (i. e., a positive correlation could be due to functionality instead of morbidity).

However, it is widely assumed that deleterious genes are conserved across species more than neutral ones are [184, 161].

Functional SNPs are mainly located in exons (regions that code for proteins). However, there are different types of SNPs in exonic regions. Synonymous SNPs, those that do not change the resulting protein, are not functional. In contrast, nonsynonymous SNPs, those that modify the amino acid sequence of the resulting protein, are clearly functional. Among nonsynonymous SNPs, one may distinguish between nonsense and missense mutations. A nonsense mutation produces a premature stop codon and the resulting protein is consequently truncated. Missense mutations produce a change in an amino acid of the protein sequence, with a variable effect on its function depending on the region of the protein affected and on the characteristics of the new amino acid. Noncoding sequences are found either in introns or in regulatory regions (such as promoters or other near-gene regions). It is known that mutations in regulatory regions may also have a certain functionality, given that they affect protein regulation and consequently its structure or abundance. Moreover, the effect of these changes on the nucleotide sequences could vary considerably, depending on the regulation mechanism. These mutations are also functional, so they have also been the object of study in comparative genomics research [193, 183]. Such studies generally focus on only one functional class, either coding or regulatory SNPs, but they do not take the two categories into account, even separately [165]. Moreover, SNPs located at other genetic regions (neither coding nor regulatory regions) should also have a certain effect on genetic disorders, even if they are not functional [25, 119].

Studies on sequence variability are usually applied to sequences that span entire genes or in the sequence of amino acids of the resulting protein [129, 3, 4]. Other works have already been carried out on SNP resolution [342, 12, 335, 44]. In particular, *Asthana et al.* proposed to differentiate variability patterns at different regions by looking at nucleotide sequence variability across species [12].

On the other hand, tools for scoring the cross-species variability of a set of homologous sequences from different organisms are based on similarity measures. Traditionally, cross-species sequence similarities have been measured through the rate of evolution [121]. Nowadays, standard measures of sequence variability across species are based on this measure. Genomic Evolutionary Rate Profiling (GERP) is the current reference for measuring evolutionary similarity between sequences of different organisms [60]. This measure depends on an estimation of the sequence similarities using a maximum likelihood-based method. This estimation relies on the phylogenetic tree that represents the inferred evolutionary relationships between the different species or other entities. Furthermore, information-theoretic measures have also been used for evaluating the variability between sequences of different species without depending on the phylogenetic dependencies between these organisms [304, 316]. In particular, the Shannon Entropy [272] is a measure of variability that has already been applied in comparative genomics [142, 316], usually for measuring the variability of amino acid sequences of different species.

6.2.2 An overview of the study

The main goal of this study was to observe the different patterns of variability of SNP sequences and their morbidity, for SNPs in different genetic regions. Specifically, a set of deleterious SNPs were compared with a set of neutral SNPs to statistically differentiate their patterns of variability across species and identify a common pattern of variability for disease-related SNPs. It is clear that functionality will impact on sequence variability. In order to explore this effect, this study proposed considering separately the SNPs belonging to different genetic regions and with different functional effects. Instead of differentiating the SNP cross-species variability patterns between different regions or functionalities, one of our study goals was to analyze the differences of sequence variability between deleterious and neutral SNPs for each functional category, one by one. Cross-species comparisons were carried out at nucleotide resolution, concretely, using reduced-length sequences of nucleotides located around SNPs.

The methodology followed for carrying out this study is divided in four specific steps. First of all, SNP data was collected. The second step consists on finding homologous sequences, given the local sequence of a SNP. Afterwards, the sequence variability for each SNP was measured using Shannon's entropy. Finally, a statistical analysis was applied for comparing the entropy values between deleterious SNPs and neutral SNPs.

6.2.3 Data Collection

SNP data for Homo sapiens was acquired from the dbSNP database [274] in its Build 130. This database is the main public repository for genetic variation within and across species. It contains several features on SNPs that can be obtained using the Entrez SNP search tool. A variety of queries can be used for searching SNPs by ID, gene name, organism, genetic region, func-

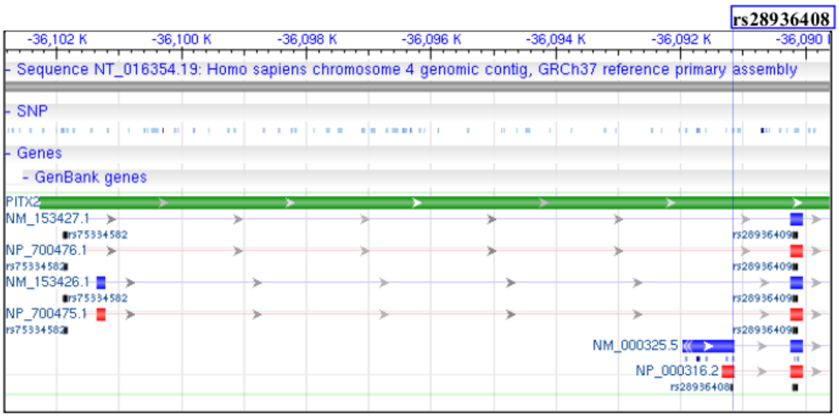


Figure 6.15: SNP rs28936408 is an example of class "Intron + Missense". It is positioned in chromosome 4 sequence (data and figure obtained from NCBI Sequence Viewer).

tion class, or even annotations to clinical databases. In particular, dbSNP includes annotations to the Online Mendelian Inheritance in Man (OMIM) database, which is a catalog of all the known diseases with a genetic component as well as the corresponding relevant genes in the human genome. Looking at OMIM-related genetic markers has become a standard practice for comparing behaviors between deleterious and neutral SNPs [40, 165, 30].

Our deleterious SNPs were obtained by making a query at dbSNP for all the human SNP variants annotated in the OMIM database. The result of this search produced a sample containing 3658 SNPs located at different genetic regions and having different functional effects, as shown in table 6.3.

The dbSNP database contains SNPs of an unknown functional class and SNPs associated with more than one functional class. The former correspond to SNPs that have not been classified in any functional category in the dbSNP database. The different annotations a SNP has in terms of functional classes could correspond to different open reading frames or could be due to alternative splicing. Figure 6.15 illustrates this phenomenon, showing a SNP cataloged in 2 categories (Intron and Missense). The polymorphism has been included in this class because, on the one hand, it is located in the intronic region of the "paired-like homeodomain 2" gene (GeneID: 5308, PITX2) for the sequences corresponding to isoforms a (NM_153427.1/NP_700476.1) and b (NM_153426.1/NP_700475.1) of the protein; and, on the other hand, it belongs to the sequence of isoform c at the first intron of the (NM_000325.5/NP_000316.2) gene. For this protein, sequence changes imply changes in the amino acid sequence (L→Q).

Both deleterious and neutral SNP samples were stratified in a balanced manner to avoid introducing an additional bias in the statistical test. The stratification by genetic region and functionality of the deleterious sample is described in table 6.3. To avoid small sample issues, only categories with more than 15 SNPs were considered. The SNPs corresponding to categories with fewer than 15 SNPs were removed from the sample, yielding a final set containing 3568 SNPs.

Table 6.3: Number of deleterious SNPs for each genetic region or functional category.

Genetic region and functional class	n
Intron	648
Nonsense	176
3´ UTR	31
Coding-Synonymous	114
5´ UTR	15
Missense	2204
Near gene 5´	64
Intron + Missense	45
Near gene 5´+ Missense	40
Near gene 5´ + Intron	21
Unknown	210

The neutral sample contained dbSNP SNPs with no OMIM annotations. These SNPs were selected randomly among all the dbSNP SNPs not annotated in OMIM yet maintaining the same functional stratification as the one described in table 6.3 for SNPs from the deleterious sample. No limitations on the allele frequencies were used.

The sequence associated with a given SNP spanned a region of variable length around the locus. Both the sequence and the SNP position within this sequence were stored for each SNP.

R software [246] (version 2.8.1) was used for data acquisition, by making automatic queries at dbSNP through eUtils, a set of tools for accessing NCBI databases remotely [219].

6.2.4 Search of homologies

The second stage of the methodology applied a BLAST algorithm for searching homologous sequences. Given a SNP sequence acquired in the previous step, this stage consisted of searching a set of homologous sequences against the non-redundant nucleotide database (*nr/nt*), the largest database available through BLAST. The species considered in this database are attached as supplementary material. This was accomplished with version 3.7 BLAST [8]. A local *blastn* algorithm was applied, under its version 2.2.21. The expectation value threshold was set to a lower significant value of $E = 0.05$.

The variability between a set of homologous sequences was analyzed in further steps to compare both samples.

The sequences of nucleotides extracted from the NCBI database varied in length. As it was assumed that SNPs occurred in every 200 to 1000 base pairs, the length of the SNP sequence had to oscillate within this interval [169]. However, the a priori chance of finding homologies between sequences is also known to be proportional to sequence length [8]. Consequently, the SNP sequences were cut to a common length of 300 nucleotides in order to guarantee finding enough homologous sequences for measuring variability and to maintain the statistical power around the SNP. Polymorphic sites were codified following the IUPAC code [63] so as not to influence the search for homologies with any of the possible alleles.

For each SNP sequence in both samples, the *blastn* algorithm returned an alignment of homologous sequences corresponding to nucleotide sequences of different organisms.

6.2.5 Measuring the SNP variability among species

The cross-species sequence variability of each SNP was computed by applying the entropy measure to the columns of the matrix of homologous sequences. Let S be the column of the matrix corresponding to the position of the SNP within the original sequence; the entropy H of S is given by (6.13).

$$H = -\sum_{i=1}^{N} p(S_i) \cdot log_2 p(S_i) \tag{6.13}$$

where N is the number of possible symbols for the SNP S (here $N = 4$ for A, T, C or G) and where $p(S_i)$ is the probability of having the symbol i at S [272]. The entropy measures the disorder found at the SNP position. High entropy values correspond to SNPs with a high variability across species, whereas SNPs totally preserved along evolution will present null entropy values.

Every matrix of sequences had a finite sample size corresponding to the number of homologous sequences. This was translated to an error in the probability estimation error using the nucleotide frequency that led to an error in the entropy measurement as defined in equation 4.25. If this sampling error (SE) is greater than the difference of entropies between sets, the statistical comparison between samples will be unreliable. To avoid this bias, the SE on the computation of the entropy was estimated for different sample sizes, corresponding to the number of sequences, using the methodology proposed in *Schneider et al.* [265]. Given that the expected number of homologous sequences is low, the exact approach of the SE estimation was applied. Figure 6.16 shows the SE jointly with the mean difference (MD) of the entropies. Figure 6.16 was obtained using all SNPs from both samples, but the decision on the optimal number of sequences needed to satisfy the required conditions for the statistical test was determined for each class separately, as follows. The optimal number of sequences corresponds to the one that reaches a maximum MD and the maximum gap between the MD and the SE. It can be observed in igure 6.16 that, starting from 50 homologous sequences, the SE was lower than the MD. In this example, this value also corresponded to the threshold of statistical significance in the statistical test comparing both means (p-value < 0.05). The optimal number of sequences was determined for each functional class to guarantee test accuracy, as shown in table 6.4. In the example shown in figure 6.16 the optimal number of sequences was 54, the value that reached a compromise between a minimum SE and a maximum MD. In this case, this optimal value also guarantees maximum significance in the statistical test described as follows.

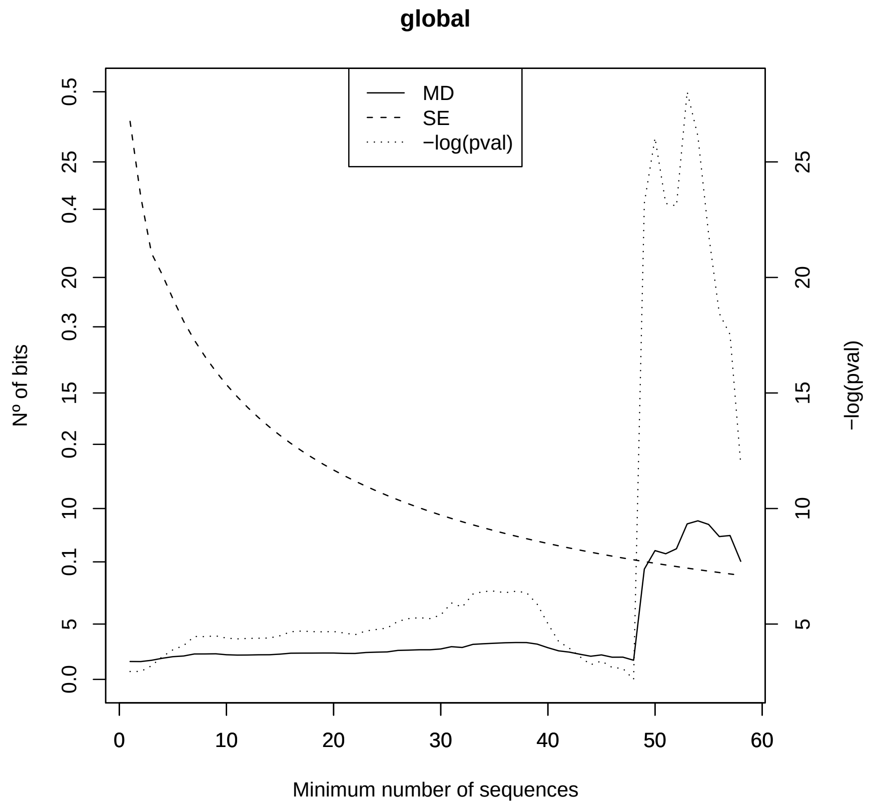


Figure 6.16: Mean Differences (MD) of the entropies between the two sets, Sampling Error (SE) in the entropy computation, estimated as in [265] and - log (p-value) in the statistical test comparing both means, for different sample sizes (number of homologous sequences).

6.2.6 Statistical analysis

A Mann-Whitney U-test is applied to the entropy values for comparing deleterious and neutral SNP samples, where the null hypothesis was that the probability distributions of the two samples were equal. The alternative hypothesis was that the distribution of one sample was greater than the other. It required the two samples to be independent, and the observations to be ordinal or continuous measurements. In this case, a one-sided test was performed, where the null hypothesis stated no difference between the means of the entropy of the two samples and the alternative hypothesis stated that the mean of the entropy for the neutral SNPs was greater than that for the deleterious sample, meaning that neutral SNPs showed a higher variability across species than SNPs related to disease. The Mann-Whitney test was applied for each of the 11 groups defined previously. An additional test was applied to each of the two samples globally to observe the general tendencies of the samples. Specifically, an analysis of variance was carried out through a Kruskal-Wallis test.

6.2.7 Results

The study on the cross-species variability of disease-related SNPs compared with neutral SNPs showed significant differences between the entropy values for both sets. Without taking into account functional stratification, the morbid set was found less variable than the control set (p-value=1.8×10^{-10}, under a Kruskal-Wallis test). It can be observed in figure 6.17 that, on average, the entropy of the neutral SNPs was higher ($\mu = 0.68$) than the entropy of deleterious SNPs ($\mu = 0.57$). This reflects the fact that disease-related SNPs are better preserved across species than SNPs selected randomly, as expected. The results obtained for each of the categories are described qualitatively in figure 6.17 and quantitatively in table 6.4.

Near-gene 5´ SNPs were observed not to present significant differences between the two samples. However, the mean of the entropies of neutral SNPs tended to be higher than the mean of deleterious SNPs, referring to a lower variability for disease-related SNPs (Figure 6.17). The near-gene 5´ region corresponded to a region of variable length located within 2 kb 5´ of a gene but not in the transcript for the gene [220]. Hence, this region may include variations of uncertain functionality. This functional diversity of SNPs at 5´ UTR regions may imply difficulties in statistical differentiation of SNPs according to their association with disease.

SNPs in 5´ UTR presented significant differences between samples. Deleterious SNPs presented a mean entropy lower than that of neutral SNPs (p-value = 5×10^{-3}). Consequently, one can establish that disease-related SNPs in the 5´ UTR region show lower cross-species variability than neutral SNPs in the same region. The five prime untranslated region (5´ UTR) set

Table 6.4: Results obtained for the different functional classes. For each class, the optimal number of homologous sequences selected is shown, as well as the p-value of the comparison between the two samples and the descriptive statistics within each sample.

category	Number of sequences	p-value	Deleterious SNPs		Neutral SNPs	
			# SNPs	$\mu \pm sd$	# SNPs	$\mu \pm sd$
NEAR GENE 5´	27	0.08	38	0.36 ± 0.48	36	0.49 ± 0.49
5´ UTR	13	5×10^{-3}	15	0.71 ± 0.38	15	1.04 ± 0.27
CODING SYNONYMOUS	55	0.61	9	0.95 ± 0.65	34	0.92 ± 0.46
NONSENSE	56	0.09	7	0.55 ± 0.47	17	0.82 ± 0.49
MISSENSE	57	1.1×10^{-9}	783	0.57 ± 0.45	479	0.758 ± 0.42
INTRON	52	6.4×10^{-3}	134	0.45 ± 0.48	178	0.57 ± 0.48
INTRON + MISSENSE	55	0.23	24	0.82 ± 0.37	7	1 ± 0.42
NEAR GENE 5´ + MISSENSE	22	0.04	38	0.32 ± 0.43	32	0.51 ± 0.52
NEAR GENE 5´ + INTRON	51	0.8	5	0.48 ± 0.26	5	0.35 ± 0.26
3´ UTR	33	0.79	16	0.79 ± 0.58	14	0.58 ± 0.49
UNKNOWN	44	8×10^{-4}	102	0.59 ± 0.47	178	0.75 ± 0.39
ALL	54	1.8×10^{-10}	1217	0.57 ± 0.44	1217	0.68 ± 0.41

contains sequences that may be a hundred or more nucleotides long preceding the gene. It is a specific section of messenger RNA (mRNA) resulting from transcription but not translated as proteins. It usually contains regulatory sequences such as binding sites. It has already been demonstrated that 5´ UTR are sequences preserved across different organisms and that they contain functional SNPs [217]. In this study, morbid SNPs in 5´ UTR were also found to be less variable across species than neutral SNPs.

Synonymous SNPs located in coding regions did not show significant differences between sets. Moreover, the mean entropy of deleterious SNPs tended to be higher than that of neutral SNPs, denoting that disease-related SNPs located at these regions show higher cross-species variability than neutral SNPs (Figure 6.17), although results for this set may have lost significance due to the small sample size. It is known that coding regions are preserved throughout evolution. Given that they do not affect the resulting protein amino acid chain, silent mutations found in these regions presumably show high variability across species regardless of their association with disease.

For nonsense mutations, the results show that neutral SNPs tended to have, on average, higher entropy than deleterious SNPs, indicating that disease-related SNPs were less variable across species than neutral SNPs (Figure 6.17), although this difference was not statistically significant. Here, the small sample size affected the results obtained by the statistical test, decreasing its statistical power. Moreover, the results showed that the entropy of neutral SNPs was, on average, higher than that of deleterious SNPs, indicating that disease-related SNPs are less variable across species than neutral SNPs (Figure 6.17). As well as silent mutations, nonsense mutations are likely to be preserved whatever their association to disease.

It can be seen in tables 6.3 and 6.4 that the majority of disease-related

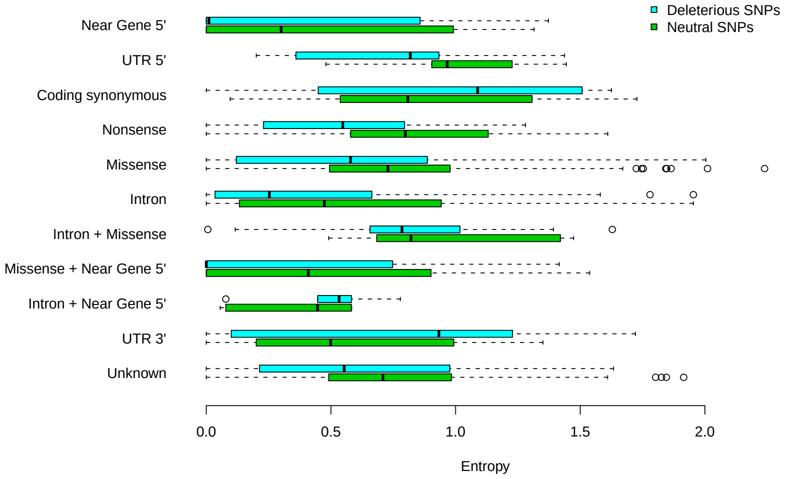


Figure 6.17: Comparison of the statistical distribution of the sequence variability (entropy) for SNPs in both samples for each functional class.

SNPs belonged to the missense category. For this functional class, deleterious SNPs were significantly less variable across species than neutral SNPs (p-value $= 1.1 \times 10^{-9}$). This category unbalanced the global sample and was the factor most responsible for the statistical difference obtained when comparing both samples globally. Missense SNPs correspond to the most functional sort of mutation, as they modify the amino acid chain of the resulting protein. It has already been demonstrated that missense SNPs show significantly lower variability across species [193] but, in this study, this low variability was also associated with the deleterious condition of missense SNPs.

The intron category constituted the second most abundant functional class for disease-related SNPs. For SNPs in intronic regions, significant differences were found between neutral and deleterious samples, the former showing a lower entropy, corresponding to lower variability across species. Genetic disorder studies usually focus on coding regions; however, it is known that even if intronic SNPs are not as studied as missense mutations, intronic SNPs may also be related to disease. It was shown in this study that low entropy values for these SNPs may characterize their association with disease.

For SNPs labeled at both intron and missense regions, Figure 6.17 shows that disease-related SNPs tended to be less variable across species than neutral SNPs. However, these differences were not statistically significant, but results for this set may have lost significance due to small sample size.

SNPs located at both near-gene 5´ and missense regions related to disease were significantly less variable than neutral SNPs, showing lower entropy. Actually, this category showed the same pattern as missense SNPs, which indicates that the additional annotation to the near-gene 5´ region does not affect the results.

SNPs located at both near-gene 5´ and intron regions did not present significant differences in the entropy comparison due to the limited sample size. This category was not representative because it contained just 5 SNPs.

SNPs at the 3´ UTR region did not present significant differences between samples. However, it was observed that neutral SNPs tended to show lower entropy and were thus less variable across species than deleterious SNPs (Figure 6.17). Just as with the 5´ UTR, a 3´ UTR is a specific section of messenger RNA (mRNA) resulting from transcription but not translated as proteins. It is located right after the last codon of the gene. These regions are believed to contain regulatory elements such as binding sites and they have been demonstrated to be deeply preserved in vertebrates [50]. Surprisingly, the results obtained in this study go in the opposite direction, perhaps due to the small sample size.

SNPs of unknown functionality presented significant differences between samples. The data in table 6.4 reveal that disease-related SNPs presented a significantly lower entropy ($\mu = 0.59$) than neutral SNPs ($\mu = 0.75$), the former being less variable across species than the latter. This observation takes on added importance, as it links morbidity with low sequence variability across species for any SNPs.

As expected, a clear tendency was observed showing that disease-related SNPs showed lower entropy values than neutral SNPs. For functional classes such as missense SNPs or SNPs in 5´ UTR, as well as for SNPs of unknown provenance and SNPs at intron regions, the entropy differences were statistically significant. For some classes, our suboptimal sample size did not make it possible to draw any conclusions. For SNPs of clearly or ambiguous functionality, their morbidity was significantly associated with low cross-species sequence variability.

6.3 Concluding remarks

In this chapter, two approaches to the one-locus genetic analysis of human diseases were studied. On one hand a nonlinear one-locus test of genetic association was proposed. On the other hand, an exploratory study on the sequence properties of SNPs related to diseases was carried out for SNPs located at different genomic regions.

In section 6.1, a one-locus nonlinear test of genetic association was proposed based on the mutual information measure that takes into account the genetic structure of the population. This nonlinear test was applied using three approximations for the estimation of the mutual information measure, the empirical one (EMP), the Miller-Madow estimator (MM) and the Schurmann-Grassberger estimator (SG). The resulting three nonlinear tests were compared to traditional linear models proposed in the GenABEL software (GA) and were applied to the GAW17 dataset. Since the solution of the GAW17 simulation model is known, this study served to characterize the performance of the proposed nonlinear test in comparison to the standard genetic association tests based on linear regressions.

It was first demonstrated with synthetic data that the mutual information captures both linear and nonlinear correlations between variables. In its application to the GAW17 dataset, the proposed nonlinear test not only replicated the results found with linear regressions but also detected a SNP in a gene correlated with the phenotype, suggesting that it could be in linkage disequilibrium with a true positive SNP within the same region. It was also shown that the performance of the nonlinear tests in terms of their accuracy in classification (AUC) was similar for the four tests. The low performance results obtained with the four tests were caused by a high false positive rate, which decreases the performance (AUC) of the association test. This is a generic drawback in all the association tests. The high false positive rate may depend on different aspects of the study design. In this case, it could possibly be originated from the high number of rare variants, or by the direct application of the association test on SNP data, without taking into account any enrichment tool, neither prioritization criteria. In order to improve this aspect, it is emplaced as future work to extend the definition of the nonlinear test in terms of Renyi's divergences with the aim of better detecting rare variants. Moreover, in order to differentiate between markers that are truly related with disease and those found by chance, the search of additional criteria for prioritizing the biological relevance of SNPs has become a common practice in genetic research. Prioritization criteria may look at other polymorphisms' features that could determine their functionality.

In particular, the study proposed in the section 6.2 revealed that exploring the SNP sequence variability among species may discriminates deleterious SNPs from neutral SNPs. In this study, a methodology was developed for studying SNP cross-species sequence variability in order to evaluate the polymorphism region effect on the possible correlation between SNP sequence variability and the SNP association with disease. The methodology was based on the definition of two sets, one containing deleterious SNPs and the other defined by neutral SNPs. Both sets were stratified depending on the region where the polymorphism is located and its functionality, a feature that may have influenced the evolution of species. It was observedthat deleterious SNPs tend to be less variable across species than neutral SNPs. For most functional classes, these differences were shown to be statistically significant. This study was only a preliminary exploratory analysis of the SNP variability among species. It indicates that SNP sequence variability among species may help to discriminate deleterious SNPs from neutral SNPs at different genetic regions. The characterization of a prioritization criterion based on this observation to support genetic association studies is proposed as future work.

Chapter 7

Multi-Loci Genetic Association

7.1 Combinatorial effects of SNPs

The technique described in chapter 6 facilitated the identification of several human genes where there are mutations that lead to Mendelian diseases. However, they are not always useful for studying complex diseases that are product of the interaction of many loci and the environment. For such diseases, multi-loci analysis is expected to be more powerful than the traditional locus-by-locus SNP association studies ([34]), detecting even more interactions. Sometimes, the combination of two of more-loci may provides more information about the phenotype than only one single SNP.

This is illustrated in figure 7.1, where the statistical significance of the correlation between pairs of SNPs of the *F7* gene and the FVII levels in blood is plotted as well as the individual correlation of each SNP with the phenotype. This figure was generated using the nonlinear method based on information theory described in section 6.1.2. The central matrix represents the significance of the mutual information of each pair of SNPs and the phenotype, where significant combinations are shown with a black square. On the left side of the matrix, an histogram with the statistical significance of the mutual information between each SNP and the phenotype is represented. Significant SNPs are also represented with a black square. On the right side of the plot, an LD matrix is represented in order to evaluate if SNPs are in linkage disequilibrium with each other.

It can be observed that SNPs $rs9604025$ and $rs510335$ are not individually significantly correlated with the phenotype, whereas the combination of these two SNPs is significantly related with the phenotype (red square). Furthermore, SNP $rs510335$ is a relevant SNP cited in many works related to the factor VII. The rare T allele is associated with lower plasma concentrations of FVII protein and fully activated FVII molecules [306]. The LD

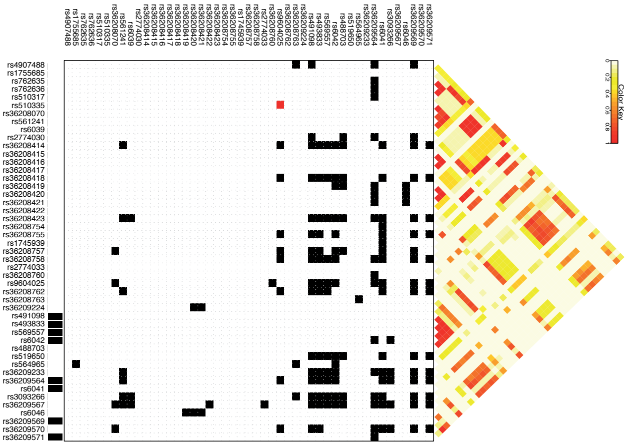


Figure 7.1: Two-loci interactions between SNPs. The histogram on the left represents the significance of the correlation of each individual SNP with the phenotype. The matrix at the center shows the statistical significance of adding the SNPs in files to the SNPs in columns (black squares represent significant combinations of SNPs) and LD figure on the right represents the r^2 LD measure between each pair of SNPs.

plot shows that these SNPs do not present a significant correlation.

Figure 7.1 shows an example where a SNP related to a phenotype is not detected through traditional one-locus techniques but it can be detected in combination with another polymorphism, which helps to manifest its association with the phenotype. Since it is observed with SNPs in the same gene, it can be deduced that this may presumably occur recurrently, especially when looking at SNPs in different genes, where epistatic effects may favor this situation [215]. Epistasis, the interaction between genes, has been widely undervalued in the context of genetic association studies. Traditional association algorithms most often look for individual genes with large impacts on a single phenotype. However it frequently results in spurious and irreproducible results because network interactions are not taken into account [303]. Thus, statistical methods to incorporate SNP-SNP interactions in an association study are needed. Various approaches have been proposed and shown to be superior to SNP-by-SNP association analysis. Most of them are based on multiple regression models [61]. However, other approaches have also been proposed using different learning algorithms[340, 311, 206].

The aim of this chapter is to describe a novel multi-loci genetic association method, based on information theory. The methodology proposed in chapter 6 was adapted for building a nonlinear multi-loci association test.

7.2 A nonlinear multi-loci association test

7.2.1 A feature selection problem

As mentioned in section 3.1.2, finding association between sets of genetic variants and a phenotype can be seen as a feature selection (*FS*) procedure, in the sense of selecting genetic variants with a relevance criterion of association with the phenotype.

In order to perform a multivariate SNP subset selection, wrapper and embedded feature selection methods are the most adequate. As described in section 7.1, the “best” set of SNPs (the most significantly related to the phenotype) does not always contain the best individual SNPs. Thus, a multi-solution approach seems to be appropriate allowing to obtain all the possible combinations of SNPs associated with phenotypes and not to omit possible combinations of SNPs associated with diseases. Sequential feature selection algorithms have been considered and compared to an embedded algorithm based on a greedy search proposed by *Miller et al.* [206].

The principal disadvantage of applying a sequential forward or backward selection algorithms (SFS and SBS) is that they do not take into account the correlations between features and may produce the effect of finding redundant sets of features. When two SNPs with a similar variability among individuals are selected together, the selection set contains twice the same information about the phenotype. This occurs generally when SNPs are inherited together in the same haplotype. In order to avoid this problem, algorithms that combine forward and backward steps have been proposed. Floating variants of SFS and SBS were introduced in order to combine forward and backward steps dynamical. These are sequential forward floating selection (SFFS) and sequential backward floating selection (SBFS). The difference between these two algorithms is that SFFS starts with an empty set and first applies a forward selection step, adding features to the selection set, whereas SBFS starts with the total set of features and first applies a backward step, eliminating features from the selection set. However, as it is explained in section 7.2.2.2.1, a few number of SNPs are enough for explaining the information of the phenotype. Thus, if starting with the total set of SNPs, any of the SNPs in the set should be removed by a backward selection procedure. This implies starting as many new searches as SNPs in the selection set at each step, which is computationally demanding. Thus, it makes more sense to start with an empty set and keep adding features than starting with the total set (47 SNPs) and remove features, since it guarantees reaching the optimal sets with fewer search steps.

Here, a multi-solution version of the *SFFS* algorithm (*MSSFFS*) was proposed. The multi-solution strategy consisted in starting a new search for each significant SNP. The *MSSFFS* algorithm returned multiple sets of SNPs that are able to represent the information of the phenotype. Floating algorithms combine forward and backward steps as described in figure 7.2. Forward steps add relevant features whereas backward steps allow to deflate the selection set, removing the redundant SNPs. The *MSSFFS* algorithm starts with a forward step applied to the empty set.

The algorithm returns sets of SNPs that, together, show a significant correlation with the phenotype, but that do not share redundant information one with each other. Thus, SNPs appearing in the same set may have a different variability among individuals, and so they are expected to belong to different haplotypic regions.

The main structure of the developed *MSSFFS* algorithm is described in figure 7.2.

1. Initialization of the set $S = \{\}$.
2. **Forward step**: For each available SNP S_i, the p-value associated to the gain of information produced when adding this sNPs to the feature set is computed according to a given relevance criterion.
3. For each significant SNP, a new forward search (2) is started from the new set $S = S + \{S_i\}$. The forward step (2) is repeated whereas there are significant SNPs.
4. **Backward step**: For each SNP S_i in S, the p-value associated to the loss of information when removing S_i from S is computed.
5. For each non-significant SNP, a new backward search (4) is started from the new set $S = S - \{S_i\}$. The backward step (4) is repeated whereas there are nonsignificant SNPs and the set S has more than one SNP.
6. Go to step 2.
7. If there are neither significant SNPs in step 3 nor non-significant SNPs in step 5, the search is stopped.

Figure 7.2: The *MSSFFS* algorithm for genetic association.

The methodology described above was compared to the *MECPM* methodology ([206]). *MECPM* is an available algorithm based on the maximum entropy principle. This algorithm applies a greedy search based on a Maximum Entropy model induction. *MECPM*'s key features are: (i) interactions are added one at a time; (ii) coding models (dominant, recessive) are considered for each of the SNPs in a set; (iii) for each set, candidates are evaluated by increasing order; (iv) only small features subsets are considered for achieving a low-complexity method (first or second order interactions).

7.2.2 Relevance criteria

The *MSSFFS* was applied with two relevance criteria. Both of them aim to determine if the gain of information produced by one SNP on a set of previously selected SNPs is significant. The first one corresponds to the standard practice for multi-loci association based on multiple linear regression models [114]. Besides a novel multivariate nonlinear method based on information theory was proposed for multi-loci genetic association.

7.2.2.1 Linear method

The linear method was chosen since it is a reference method for selecting multiple SNPs correlated with the phenotype. It consists on a multiple linear regression model (*MLR*) model that measures the linear dependencies between variables based on multi-linear regressions [114]. *MLR* tries to fit a model that represents the linear relations existing between a set of independent variables $S = \{S_i\}$ (here S_i are SNPs), and an observed variable (for instance the phenotype Y) as in (7.1).

$$Y = \beta_0 + \beta_1 \cdot S_1 + \ldots + \beta_n \cdot S_n + \epsilon \tag{7.1}$$

where β_i are the regression coefficients and ϵ is the error of the model. The method estimates the values of β_i that minimize ϵ. Each coefficient (β_i) represents the individual contribution of a SNP (S_i) for the prediction of Y.

7.2.2.1.1 Individual SNP correlation with the phenotype

Given a set of SNPs S, the gain of information provided by a SNP S_i about the phenotype Y is represented by the regression coefficient β_i. The statistical significance of this correlation is determined by a Student's t statistical test. The null hypothesis supposes the nullity of the corresponding regression coefficient ($\beta_i = 0$). Given a set of SNPs S and a SNP S_i, a t-Student test over the regression coefficient β_i was used to determine if S_i, individually, adds information about the phenotype Y respect to S.

7.2.2.1.2 SNPs set correlation with the phenotype

The significance of the correlation of the total set ($S + S_i$) was obtained using a Fisher hypothesis test (F-test). In this case, the null hypothesis supposes the nullity of the slope of the regression line, i.e. all the regression coefficients at the same time, ($\{\beta_i\} = 0$). The resulting p-value was used to determine if all the SNPs ($\{S_i\}_{i=1\cdots n}$), jointly have a significant predictive linear capacity over the phenotype Y. Once a SNP set was obtained, the F-test was used to determine if it is significantly linearly correlated, as a set, with the phenotype Y.

7.2.2.2 The mutual information-based method

7.2.2.2.1 Multivariate mutual information measures

The purpose of the multivariate mutual information measure is to determine the amount of information shared by several variables. The mutual information between two SNPs S_i and S_j and a phenotype Y can be defined as in equation 7.2.

$$\begin{aligned} I(S_i, S_j; Y) &= H(S_i, S_j) - H(S_i, S_j|Y) \\ &= H(S_i, S_j) + H(Y) - H(S_i, S_j, Y) \end{aligned} \tag{7.2}$$

It can be generalized to the case of n SNPs $S_1, \ldots, S_n$ as in equation 7.3.

$$\begin{aligned} I(S_1, \ldots, S_n; Y) &= H(S_1, \ldots, S_n) - H(S_1, \ldots, S_n|Y) \\ &= H(S_1, \ldots, S_n) + H(Y) - H(S_1, \ldots, S_n, Y) \end{aligned} \tag{7.3}$$

with

$$H(S_1, \ldots, S_n) = \sum_{i=1}^{n} H(S_1, \ldots, S_{i-1}, S_{i+1} \ldots, S_n) \tag{7.4}$$

Equation 7.4, indicates that the multivariate entropy measure is computationally hard to implement. In order to avoid this problem, it is possible to use that considers the set of several SNPs as a single random variable with symbolic values that are the concatenation of the symbols of each SNP. This approach does not affect the entropy measures. The joint probability density function of n SNPs, is not affected because each combination of symbols maintains its frequency. The mutual information between a set of SNPs $S = S_1, \ldots, S_n$ and a phenotype Y is described in equation 7.5

$$\begin{aligned} I(S, Y) &= \sum \sum p(S, Y) log_2 \left(\frac{p(S, Y)}{p(S)p(Y)} \right) \\ &= H(S) + H(Y) - H(S, Y) \end{aligned} \tag{7.5}$$

An important property of the mutual information sets that no other variable can contain more information about another variable than itself (equation 7.6).

$$I(S, Y) \leq max\{H(S), H(Y)\} \tag{7.6}$$

Moreover, in a similar manner than with univariate information theoretic measures, the finite sample size affects the multivariate mutual information measure.

The mutual information between a set of SNPs and a phenotype a monotonous function, being strictly monotonous when the sample size is finite (equation 7.7).

$$I(S + \{S_i\}, Y) \geq I(S, Y) \tag{7.7}$$

or, equivalently, the increase of information ΔI supposed by adding a SNP to a set is always positive due the finite sample size effects.

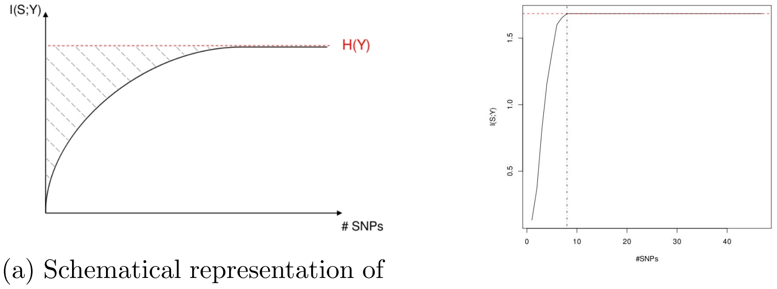
$$\Delta I = I(S + \{S_i\}, Y) - I(S, Y) \geq 0 \tag{7.8}$$

The properties of the multiple information presented in equations 7.6 and 7.8 reveal that the mutual information between multiple SNPs and a phenotype will grow as the number of SNPs increases, reaching and never exceeding the information (entropy) of the phenotype, as shown in figure 7.3a. Consequently, the information of the phenotype will be recovered with only few SNPs, even random, especially when the sample size is small.

Figure 7.3 shows the evolution of the $I(S, Y)$ as the number of SNPs in S increases. On one hand, figure 7.3a shows schematically the evolution of $I(S, Y)$ when using totally random SNPs, whereas figure 7.3b shows the real behaviour of $I(S, Y)$ when adding randomly the SNPs of the *F7* gene, where Y represents the FVII levels in blood. Figure 7.3 evidences that adding a SNP S_i to the selection set S, even if S_i is totally random, always suppose a positive gain of information as described in equation 7.8, due to finite sample size effects. The curve represent the amount of information $I(S, Y)$ about the phenotype explained by sets of SNPs S of different sizes. Moreover it will never exceed the entropy the phenotype ($H(Y)$). It is observed in figures 7.3a and 7.3b that sooner or later $I(S, Y)$ converges to $H(Y)$. When the mutual information of a set of SNPs reaches the entropy value of the phenotype, it has reached his maximum. This is the limit of the multiple SNP genetic association.

As it corresponds to a real case, it can be observed that the curves have a more staggered appearance in figure 7.3b. In this particular case, the information of the phenotype $H(Y)$ is recovered only with 8 SNPs of the *F7* gene, even if they are selected randomly. That is, with sets of 8 SNPs of the *F7* gene, one is able to recover the information about the phenotype (the FVII levels in blood) and adding more SNPs to this set will not be translated into a relevant gain of information. This indicates that, in this particular case, it is not necessary to look at more than eighth-order interactions when searching combinations of SNPs that provide information about this phenotype.

The characterization of this minimal number of SNPs necessary to recover the information of the phenotype determines the shape of the significance region. In a similar manner than in chapter 6 with the null distribution characterization, it was not possible to identify systematically the minimal number of SNPs that describes a phenotype, since it is strongly influenced by the allelic frequencies of the SNPs as well as on the number of discretization bins of the phenotype. In this case, the curve of Figure 7.3a delimitates the region of significance of $I(S, Y)$, that is the region where the gain of information ΔI produced when adding a new SNP is higher than if a random SNP would have been added. This region is represented by the striped area. When multiple SNPs $S = \{S_1, \ldots S_n\}$ are associated with a phenotype Y, their mutual information $I(S, Y)$ is expected to belong to the barred area. Thus, the main goal of the proposed nonlinear test is to find sets of SNPs that satisfies this condition, computing the mutual information statistical significance.



(a) Schematical representation of an empirical case with totally random SNPs.

(b) A real case: the 49 SNPs of the *F7* gene randomly selected.

Figure 7.3: The evolution of the mutual information for sets of SNPs as the number of SNPs in the set increases.

7.2.2.2.2 Single SNP significance

As for one-locus genetic association, the method proposed for evaluating the increase of information about the phenotype produced by a SNP on a given set of SNPs consisted on generating a null distribution. Given a set of previously selected SNPs S and a candidadte SNP S_i, the mutual information of the set resulting from adding S_i to S ($I_i = I(S+\{S_i\}, Y)$) is compared with a null distribution of mutual information ($\{I_r = I(S + \{S_r\}, Y) : r = 1..N_c\}$) obtained by generating surrogate copies S_r of S_i. The resulting P-value helps to decide if the SNP S_i provides a significant gain of information ΔI_i about the phenotype Y. The surrogate data technique was used for generating the mutual information in order to respect the allelic frequencies of the SNPs, that influence the evolution of the mutual information between multiple SNPs and the phenotype. This method has been called *MISS* (Mutual Information Statistical Significance).

Figure 7.4 illustrates this procedure. It is shown that at each step, a candidate SNP is compared to a particular null distribution in order to

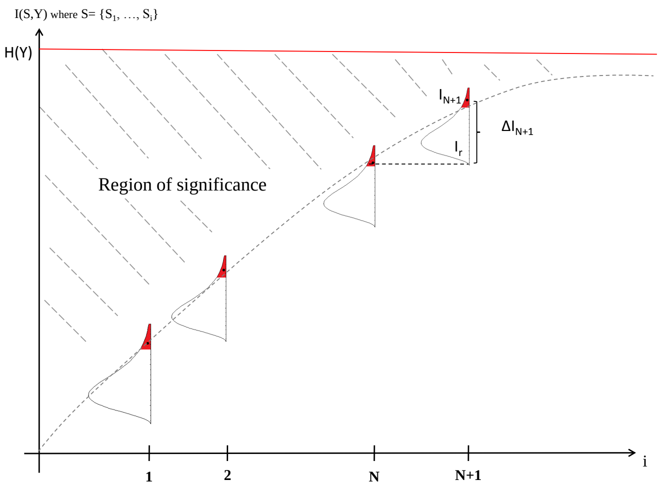


Figure 7.4: Single SNP Mutual Information Statistical Significance.

determine its singular statistical significance. However, the significance of the whole set is expected to be in the significance region.

7.2.2.2.3 SNPs set significance

Once a set of SNPs is selected, another statistical test was applied in order to determine if the selection set of SNPs, jointly, have a significant mutual information against the phenotype. This test consisted in comparing the mutual information of the set against the phenotype $I(S,Y)$ with a null distribution of mutual information $(\{I_r = I(S,Y_r) : r = 1..N_c\})$ generated with N_c surrogate copies of the phenotype. The resulting p-value determines if the selection set, as a set, is significantly related with the phenotype.

7.2.3 The MISS package

The set of functions and algorithms developed for this study were integrated in a package for the R statistical language (R Development Core Team, 2005). This package is called *MISS* and the version 0.2 is already built and available. The *MISS* library contains a documentation that includes examples of the use of the described algorithm and its underlying functions using a SNP dataset. *MISS* allows for parallel and distributed computing through MPI. Parallel implementation has been coded on top of snow R-package, authored by (Tierney et al., 2004). The code is available at `http://www.sisbio.recerca.upc.edu/R/MISS_0.2.tar.gz`.

7.3 Results

The proposed floating feature selection algorithm was applied with the two described criteria, the linear one (*MLR*) and the nonlinear one (*MISS*). Both methodologies were compared to a third approach, the *MECPM* methodology, that uses a greedy search based on a nonlinear criterion of maximum entropy.

The three methodologies were applied at a local scale, in particular to the study of the $F7$ gene. The 93 founders from the GAIT database (section 5) were used for this comparison. The phenotype, corresponding to the FVII levels in blood, was discretized using the kernel-based discretization method described in section 6.1.2. In particular, the factor FVII levels in blood were discretized in 8 categories. Besides that, the *MECPM* algorithm was tested both with the phenotype discretized using kernel functional and with a two-classes quantization. The approach giving the best results was selected, which corresponds to a binary phenotype.

Besides, in order to compare and evaluate the performance of the algorithms in detecting interactions of SNPs in a closed environment, the described methodologies were also applied to a synthetic dataset. This dataset was generated from two polymorphisms of the $F7$ gene ($rs491098$ and $rs36208414$). A phenotype was synthetically generated from the information of these SNPs through an epistatic multiplicative model ([188]). The multiplicative property of the epitstatic model involves that the correlation between the synthetic phenotype and the two selected SNPs, as a set, is nonlinear. The dataset was completed by random SNPs generated by surrogating the remaining SNPs of the $F7$ gene. The surrogate technique allowed to destroy the individual order in the SNPs variables, which guarantees randomness. The knowledge of this dataset was used to ascertain if the applied methodologies were able to detect the correlation between the interaction of the two selected SNPs and the phenotype, without assuming the intrinsic properties of the $F7$ gene.

Table 7.1 shows the results obtained by the three methods for the real dataset corresponding to SNPs in the $F7$ gene and the FVII levels in blood.

It can be observed in table that *MLR* found more sets of SNPs ($N = 151$) than *MECPM* ($N = 62$) and than *MISS* ($N = 48$). Sets obtained with *MLR* contained more SNPs ($n = 6$) than *MISS* ($n = 2$) and *MECPM* which found single SNPs related with the phenotype and did not find higher order interactions of SNPs predicting FVII levels in blood. *MISS* needed less SNPS for recovering the variability of the phenotype in exchange of finding less significance (p-values of 10^{-2}) than *MLR*.

Most of the SNPs obtained in the three cases are reported in the literature as functional variants related with FVII concentrations.

The SNPs' sets obtained using *MLR* were similar to each other since they contained common SNPs. The SNPs that made the sets different ($rs762636$,

Table 7.1: Results obtained by the different methods for the real dataset.

Method	MLR	MISS	MECPM
N	151	48	62
n	6	2	1
P-value	10^{-7}	10^{-3}	10^{-1*}
Relevant sets of SNPs	$\begin{matrix} rs762636^{a,b} \\ rs36208415^{a} \\ rs36208416^{a} \end{matrix} + \left\{ \begin{matrix} rs1755685^{a,b} \\ rs6041 \\ rs36208763^{a} \\ rs36209564^{a} \\ rs36208755 \\ rs3093266 \end{matrix} \right.$	$\begin{matrix} \text{rs493833} \\ \text{rs491098}^{a} \\ \text{rs510335}^{a,b,e,h} \end{matrix} + \left\{ \text{rs9604025}^{a} \right.$	rs561241^{a,b,d,f} rs762636^{a,b} rs493833 rs36208416^{a} rs36209763^{a} rs36208070^{a,b,e,f,g} rs510335^{a,b,e,h} rs564965^{a} rs36208415^{a} rs510317^{a,c} rs36209567^{a}
	$\begin{matrix} rs762636^{a,b} \\ rs36208415^{a} \\ rs36208416^{a} \\ rs510317^{a,c} \end{matrix} + \left\{ \begin{matrix} rs36209564^{a} \\ rs36208755 \\ rs36209569 \\ rs36208763^{a} \\ rs3093266 \end{matrix} \right.$	$\begin{matrix} rs491098^{a} \\ rs493833 \\ rs561241^{a,b,d,f} \\ rs6041 \\ rs36209569 \end{matrix} + \left\{ \text{rs36209564}^{a} \right.$	

N represents the number of sets obtained and n the average number of SNPs in a set. P-values are the order of magnitude of the obtained p-values and relevant sets of SNPs are those that present most statistical significance. Each row on the left side of the columns represent a set of SNPs whereas SNPs in curly brackets are common SNPs appearing in all sets at the left of the +.

* Classification error rates are shown instead of p-values, as *MECPM* does not provide significance levels.

[a] [283], [b] [258], [c] [306], [d] [328], [e] [192], [f] [325], [g] [80], [h] [237]

$rs36208415$, $rs36208416$ and $rs510317$) contain similar information about the phenotype. These SNPs appear in [283] in the same cluster of SNPs with a high probability of posterior effect on the phenotype, and are located in the promoter region or in splice sites as shown in figure 5.2.

The SNPs' sets obtained using *MISS* also contained information common to several of them. SNPs that made the sets different ($rs491098$, $rs510335$, $rs561241$) appear in the same cluster of SNPs in [283]. Moreover, both SNPs $rs493833$ and $rs491098$ belong to the fifth intron of the $F7$ gene 5.2. It is important to remark that the sets obtained with the proposed floating search algorithm contained SNPs that do not give redundant information about the phenotype but that complement each other. SNPs appearing in the same set may not belong to the same haplotype and can belong to different regions of the gene as they do not show significant r^2 in the LD plot in figure 7.1.

Most of the SNPs found with *MECPM* are reported in the literature as functional polymorphisms related with the phenotype. However, as it has been previously mentioned, *MECPM* was designed only for detecting first and second order SNP interactions. This is the reason why the sets of SNPs obtained with this methodology only contained one SNP per set.

The combination of SNPs described in section 7.1, corresponding to the SNPs $rs510335$ and $rs9604025$, was not detected using existing multi-loci techniques (*MLR* and *MECPM*). It can be observed that using the *MISS* methodology, SNPs $rs9604025$ and $rs510335$ were not individually significantly correlated with the phenotype, whereas the combination of these

Table 7.2: Results obtained for the simulated datasets.

	MLR	MISS	MECPM
N	1	4	7
n	1	3	1
SNPs detected	SNP_1^*	$(SNP_1, SNP_2)^*$	SNP_1, SNP_2

* denotes that SNPs were always detected, independently of the size of the dataset.

two SNPs is significantly related with the phenotype (red square). SNP $rs9604025$ appears in [283] as a functional variant related to FVII levels. SNP $rs510335$ is a relevant SNP cited in many works related to the factor VII. The rare T allele is associated with lower plasma concentrations of the FVII protein and fully activated FVII molecules [306]. Moreover, the LD plot presented in figure 7.1 showed that these SNPs did not present a significant correlation. Here, it is shown that the effects of this SNP only became apparent when it was combined with SNP $rs9604025$. This reflects the importance of looking at SNP interactions when designing a genetic association study.

The simulation study was developed to validate the performance of our methodology in detecting true interactions between SNPs and a phenotype defined by an epistatic multiplicative model. For each method, 6 datasets of different size were built, corresponding to matrices of 5, 10, 15, 20, 25 and 50 SNPs and 85 samples. The datasets contained 2 SNPs correlated with a simulated phenotype, whereas the remaining SNPs were randomly generated, being not related with the phenotype.

Table 7.2 shows the results obtained with the simulated datasets. The two real SNPs were labeled as SNP_1 and SNP_2. Results shown in table 7.2 are averages of the results obtained as changing the dimension of the dataset. For this experiment it is not worth to list the obtained sets of SNPs, so only the detection of any of the real and relevant SNPs is annotated.

It is observed that *MLR* detected one of the SNPs, as an individual set, regardless of the size of the dataset. *MECPM* found several SNPs as individual sets, including the two selected SNPs. In contrast, for each dataset size, *MISS* always found sets containing both SNPs, sometimes in combination with an other random SNP. Neither *MLR* nor *MECPM* found the true positive corresponding to the combination of the first 2 SNPs whereas *MISS* was able to detect this interaction. However, this accuracy was obtained by increasing the complexity and sacrificing the computational performance of the algorithm. This was not a critical point for local association studies like this but it may become severer in a Genome-Wide Association Study (GWAS).

In order to evaluate the computational cost required when applying *MISS* with respect to the other methods, table 7.3 shows the CPU time corresponding to each method. All computations were performed on a 12 Intel E7310 processors (4Mb Cache 1.60GHz) with 32Gb random access memory. *MISS* was launched using *snow* on MPI mode over the 12 nodes and the computing time presented corresponds to the total computing time employed by all CPUs involved. In order to make a fair comparison, the three methods were applied with conditions that benefits their performance. The *MECPM* was applied to a binary phenotype, giving faster and better results whereas with *MLR* and *MISS* it was discretized in 8 categories using [312]. In contrast, the parameters of *MISS* were also adjusted in benefit of obtaining a right detection with the minimum computational cost. Thus, the null distribution has been generated with $N_c = 100$ surrogate copies. The computational cost of the real dataset corresponding the 47 SNPs of the *F7* gene is also presented.

Table 7.3: Comparison of the three methods using the synthetic dataset and the real $F7$ founders dataset (*).

size of the dataset	CPU time (in s)		
	MLR	MISS	MECPM
5	0.2	50	44
10	0.3	77.9	103
15	0.5	90.6	475
20	0.7	130.1	914
25	0.8	237.5	1705
50	2.8	335	9500
47*	238	36828.5	9300

It can be observed that the use of *MISS* slowed down the floating search algorithm in comparison with *MLR*. However *MISS* was faster than *MECPM* for the simulated dataset. For the real dataset, *MISS* was computationally more expensive due to the dependence of the parameters of the null distribution generation that were larger for real data.

7.4 Extension to sib-pairs analysis

Paralelly to the study described in section 7.2, a sib-pairs analysis was developed with the 345 pairs of sibs selected from the GAIT database. Generally, family studies are based on comparing the genotypic information of two individuals within the same family. The first approach was to establish a genotypic distance by determining if two individuals share 0, 1 or 2 alleles Identical-by-Descent (IBD) at a given position ([157]). Two alleles are IBD if one is a copy of the other or if both of them are copies of the same ancestral ([321]). In practice, it is not always possible to estimate the number of alleles shared IBD at a given position because the allelic measurements of the ancestors are not always available. Identity-by-State (IBS) methods also estimate the genotypic differences between sib pairs. Two alleles are IBS if there are the same allele, regardless of their ancestral origin. The IBS methodology estimates a probability distribution of sharing 0, 1 or 2 alleles IBS by looking at the allelic frequencies ([29]). For avoiding the computing of these probabilities, the genotypic distance was established directly from the number of alleles shared IBS, as follows. The distance between two identical homozygous genotypes (e.g. A_1A_1 and A_1A_1) was set to *d=0*. The distance between an homozygous and an heterozygous genotype (e.g. A_1A_1 and A_1A_2) was set to *d=1*. The distance between two opposite homozygous genotypes (e.g. A_1A_1 and A_2A_2) was set to *d=2*. For quantitative traits, the number of alleles IBS that two sibs share should present a correlation with the difference of their phenotypes. Thus, the genotypic distance, computed for each sib pair and each SNP of the $F7$ gene, was compared with the phenotypic distance computed as the difference between the FVII levels of each individual. The variable of phenotypic differences was discretized with the methodology described in [312], concretely in 16 categories. The methodology described in section 7.2 was identically applied in a sib-pairs analysis, considering the IBS values between pairs at each SNP as genotypes and the phenotypic differences as the phenotype. Actually, only the *MSSFFS* algorithm was applied with both linear and nonlinear criteria, given that sib-pairs IBS data format is incompatible with the genotypic format required by the *MECPM* algorithm. Table 7.4 shows the results obtained using the *MSSFFS* algorithm with both *MLR* and *MISS* methods.

It can be observed that for sib-pairs data, *MLR* only found 3 sets of SNPS whereas *MISS* found 50 sets, a similar number than using unrelated individuals. As for the population-based study, sets obtained using *MLR* contained more SNPs (n = 4) than sets obtained with *MISS* (n = 3). In contrast, *MISS* obtained higher p-values than *MLR* in the sib-pairs analysis, whereas using only founders, *MLR* presented higher levels of significance than *MISS*.

It can be observed that results obtained with founders were different than results obtained using the sib-pairs. Most of these differences are due to the differences in the datasets, containing different samples and so, different genotypic and phenotypic measures. Contrarily to the genotypic variability, the variance of the phenotypic differences of the sib-pairs was higher (V = 1159.7) than the variance of the phenotypes of the founders (V = 826.3). This variability can only be expressed through the combinations of SNPs. This combinations were more easily found using *MISS* (50 combinations obtained) than using *MLR* (only 3 combinations). These intrinsic differences in the variability of genetic data between individuals may also influence the results. The large variability present in founders genetic data may increase the false positive discovery ([164]). This is observed with the high values of N and n, especially with *MLR*. *MISS* was more conservative as it found a similar number of SNP sets for both datasets ($N \sim 50$).

As for unrelated individuals, SNPs sets obtained with *MLR* were also similar to each other, being differentiated by SNPs $rs762636$, $rs762635$ and jointly by $rs2774033$ and $rs6039$. SNPs $rs762736$ and $rs762635$ are in the same cluster in [283] while SNP $rs6039$ appears in another cluster. Most of the sets obtained using *MISS* were composed by SNPs that appear in different clusters in [283], giving different information about the phenotype. In particular, SNP $rs36209567$, also known as A294V, appeared in several sets and is located in the ninth exon of the $F7$ gene, producing an amino acid change in the resulting protein from an Alanine to a Valine. SNP $rs36208758$ is also located in the third exon but it is a missense mutation that does not produce any amino acid change in the resulting protein as illustrated in figure 5.2.

Table 7.4: Results obtained using *MSSFFS* with both *MLR* and *MISS* for the sib-pairs analysis.

Method	MLR	MISS
N	3	50
n	4	3
P-value	10^{-6}	10^{-18}
Relevant sets of SNPs	$rs762636^{a,b}$ $rs564965^{a}$ $rs9604025^{a}$ $rs36209567^{a}$	$rs1755685^{a}$ $rs762636^{a,b}$ $rs36209567^{a}$
		$rs1755685^{a}$ $rs510317^{a,c}$ $rs36209567^{a}$
	$rs762635^{a,b}$ $rs564965^{a}$ $rs9604025^{a}$ $rs36209567^{a}$	$rs564965^{a}$ $rs36209567^{a}$ $rs36208070^{a,b,e,f,g}$
	$rs2774033$ $rs564965^{a}$ $rs9604025^{a}$ $rs36209567^{a}$ $rs6039^{a,b}$	$rs510317^{a,c}$ $rs36209567^{a}$ $rs36208758$ $rs564965^{a}$

N represents the number of sets obtained and n the average number of SNPs in a set. P-values are the order of magnitude of the obtained p-values and relevant sets of SNPs are those that present most statistical significance. Each set is presented in square brackets.

[a] [283], [b] [258], [c] [306], [d] [328], [e] [192], [f] [325], [g] [80], [h] [237]

7.5 Discussion

In this chapter, a novel methodology, called *MISS* was proposed. It is a multivariate nonlinear method for multi-loci genetic association with the goal of detecting association between combinations of SNPs and a phenotype.

Similarly than in chapter 6, the proposed method was based on the statistical significance of the mutual information gain produced by a SNP on a set of previously selected SNPs about a phenotype under study. This method was applied as a novel relevance criterion of a new a multi-solution floating feature selection algorithm (*MSSFFS*), proposed in the context of multi-loci genetic association for complex diseases. *MISS* was compared with *MLR*, a standard linear method used for genetic association, also applied as a relevance criterion of the same feature selection algorithm, and with *MECPM*, an algorithm for searching predictive multi-loci interactions with a criterion of maximum entropy. The different methods were tested with SNPs of the *F7* gene, and the FVII levels in blood, with the data from the GAIT sample using both only unrelated individuals and sibpairs in two paralel studies. As the study was a local association analysis focused on the $F7$ gene, functional studies about $F7$ polymorphisms were used to validate the results.

The proposed nonlinear method (*MISS*) improved the results of traditional genetic association methods, detecting new SNP-SNP interactions. Most of the obtained sets of SNPs were in concordance with the functional results found in the literature where the obtained SNPs have been described as functional elements correlated with the phenotype. The results presented in [283] were confirmed by the same group in [258] by functional assays. Moreover, some of these results have been also replicated by association analysis and/or functional assays in [306], [328], [192], [325], [80] and [237]. Moreover, through a particular case it was shown that a specific SNP known to be related to FVII levels in blood, the $rs510335$ SNP, was not detected using one-locus association tests but its effect on the phenotype only became apparent when it was combined with another SNP that is not individually significantly related with the phenotype either. Moreover, this particular combination was only detected using *MISS* but not when applying the other referenced methods.

On one hand, this confirms that multi-loci association improved the results obtained with one-by-one SNP association strategies, showing that combinations of SNPs may contain information about the phenotype that single SNPs are not able to capture. On the other hand, the proposed nonlinear method (*MISS*) was not only able to recover the results the traditional linear regressions but it also improved them, finding correlations between genotype and phenotype not detected with the other tested methodologies.

The originality of this method lies in three specific aspects: (1) the multi-solution characteristic of the feature selection algorithm, (2) the floating strategy of the search algorithm and (3) the generation of the null distribution of the mutual information of multiple SNPs and a phenotype. The multi-solution strategy allows to find several relevant interactions between loci, including those interactions that may be hidden behind the most significant one. However this involves an higher computational cost of the algorithm. The floating search solves the problem of finding redundant SNPs in the same set, removing at each step of the algorithm the uninformative SNPs from the selection set.

In order to evaluate and compare the performance of the poroposed algorithms, the three methodologies were also applied in a controlled environment through a synthetic dataset. This dataset was generated using a multiplicative epistatic model from two real SNPs for the simulation of a phenotype. Remaining SNPs were generated randomly. The nonlinear association between the combination of the two real SNPs and the phenotype was been detected using the *MISS* methodology, proving its capacity for finding true associations with respect to the other methods. However, this accuracy was obtained at the cost of an increased computational task of the algorithm that should be improved for its use in genome-wide association studies or for its application to other diseases or phenotypes. This improvement involves finding an analytical expression for the mutual information null distribution, which is out of the scope of this thesis, and it also invites to apply suboptimal feature selection algorithms, such as genetic algorithms or the Branch and Bound strategy, and compare its computational efficiency with the proposed floating feature selection algorithm, which is emplaced as future work.

Chapter 8

Genetic Association in multiphenotypic schemes

8.1 Introduction

First genome-wide association studies were carried out on two sets of individuals, one healthy control group and one case group affected by a disease. Nowadays, genome-wide association studies for complex diseases are often conducted on collections of patients in which multiple quantitative traits are recorded. These traits, also known as intermediate phenotypes, generally correspond to variables collected as risk factors for this particular syndrome. Some regulatory elements may jointly affect several phenotypes belonging to the same metabolic pathway. When it occurs, it is particularly interesting to study the traits as a whole in order to identify these genetic elements related with the entire underlying pathway. For example, a master regulatory gene is a single gene whose expression is both necessary and sufficient to trigger activation of many other genes in a coordinated fashion, such as transcription factors or other enhancers.

This chapter presents a methodological guide for the simultaneous analysis of several phenotypes that interact together within a given biological pathway. In particular, the method is illustrated with the coagulation cascade, with the aim of identifying regulatory genes related to thrombosis. The methodology is based on the definition of canonical phenotypes, named metaphenotypes, that capture the covariance among the different phenotypes involved in the coagulation pathway, thus explaining the joint regulation or activity of the phenotypes involved in it. The hypothesis of this work is that applying genetic analysis to this new variable will lead to identify regulatory genes that could affect the whole pathway or part of it, by regulating its components.

8.2 Methodology

The proposed methodological framework consists in building new phenotypic variables, called metaphenotypes, that capture the joint activity of sets of phenotypes involved in any metabolic pathway. In order to determine the pleiotropic effects of genetic variants on this set of phenotypes, the metaphenotypes are considered as new phenotypic unities and are subjected to genetic analyses. The methodology was divided in three steps. Firstly, the a data cleaning procedure was performed and the set of phenotypes defining the pathway under study were selected. Afterwards, the metaphenotypes were build. Finally, genetic analyses were performed with these new variables.

8.3 Data pre-processing.

8.3.0.1 Genotypic data cleaning.

The study proposed in this chapter was developed for the analysis of thrombosis disease, using the GAIT project sample, described in chapter 5. The following quality control procedure was performed on the genotypic data. Individuals with a low call rate ($< 0.5\%$), a too high IBS ($> 0.95\%$) and a too high heterozygosity (FDR $< 1\%$) were removed from the sample. In addition, markers with a low call rate ($< 0.95\%$) and a low MAF ($< 0.0064\%$) were also discarded. A total of 96 individuals and 18439 SNPs were removed from the study. A clean dataset containing 364 individuals and 277191 SNPs was obtained for further analyses. This procedure was implemented in R using the GenABEL package [16].

8.3.0.2 Phenotypic data imputation.

The phenotypic dataset was imputed in order to avoid missing data. The imputation was carried out with a bayesian PCA method (bPCA) [288]. This technique applies PCA on incomplete data and uses it to impute missing values. bPCA uses a Bayesian estimation method to calculate the likelihood of imputed values. In particular, the methodology was applied using 3 principal components. This optimal number of principal components for this dataset was determined by cross validation on the captured variance. Two parameters were used for this, the NRMSEP (Normalized Root Mean Square Error in Prediction) and the Q2. The NRMSEP normalizes the root mean square standard deviation (RMSD) between the original data and the imputed data using the variable-wise variance. In contrast, Q2 is the cross-validated correlation parameter and can be interpreted as the ratio of variance that can be predicted independently by the PCA model. Low Q2 values indicate that the PCA model only describes noise and that the model is unrelated to the true data structure.

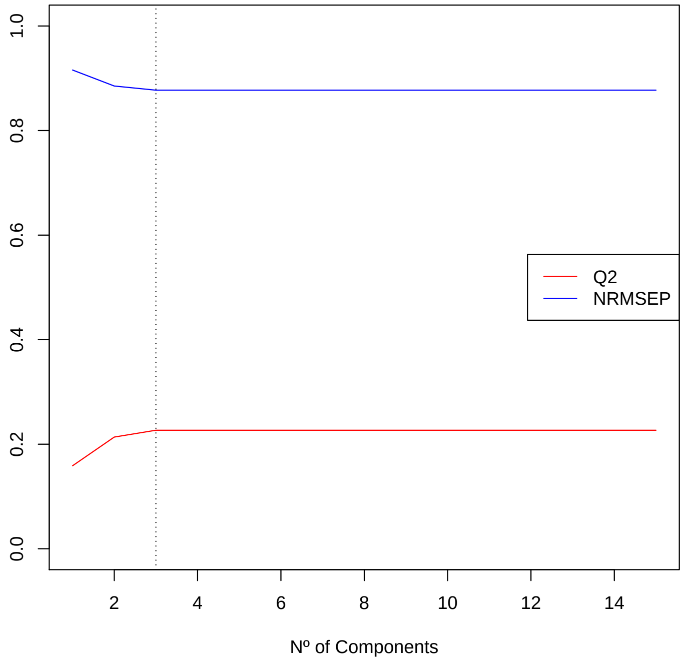


Figure 8.1: NRMSEP and Q2 cross-validation parameters determining the optimal number of components in bpca.

Figure 8.1 shows the NRMSEP and Q2 parameters for different numbers of components. It is observed that the minimum value for the NRMSEP error parameter and the maximum of Q2 parameter are both obtained with 3 components.

8.3.0.3 The phenotypes involved in the coagulation pathway

Among the collection of 80 phenotypes available for the GAIT sample, 32 phenotypes involved in the coagulation pathway were selected in order to study their joint activity within this metabolic process. These phenotypes were selected as they are defined in the literature [167]. Table 8.1 lists the phenotypes included in multiphenotypic models defined in the next section. In order to facilitate the biological interpretation of further results, each phenotype has been associated to the particular pathway of coagulation it belongs to, corresponding to the extrinsic pathway, or tissue factor pathway, the intrinsic pathway, or contact activation pathway and the common pathway of coagulation and the fibrinolysis pathway, as described in Figure 5.1.

Table 8.1: Phenotypes involved in the coagulation pathway.

Pathway	Phenotype
Extrinsic	Factor VII (FVII)
	Tissue Factor (TF) or Factor III (FIII)
	Tissue factor Pathway Inhibitor (TFPI)
Intrinsic	Factor XII (FXII)
	Factor XI (FXI)
	Factor IX (FIX)
	Factor VIII (FVIII)
	von Willebrand Factor (FvW)
	Protein C (PC)
	Protein S Total (PST)
	Protein S Free(PSF)
	Protein S Functional Total(PSFT)
	Protein S Free Ratio (psfR)
	Histidine-rich Glycoprotein (HRG)
	Prekalikrein
	P-selectin (PSEL)
Common	Factor II (prothrombin) (FII)
	Factor V (FV)
	Factor X (FX)
	Factor XIII (FXIII)
	Factor XIII activated (FXIIIa)
	Fibrinogen (FIB)
	Antithrombin (AT)
	Heparin Cofactor II (HC2)
Fibrinolysis	Plasminogen
	Plasminogen Activator Inhibitor (PAI)
	Plasminogen Tissue Activator (TPA)
	Urokinase-type plasminogen activator (u-PA)

8.3.1 Metaphenotype construction

8.3.1.1 An index of joint activity

As they respond in cascade in the coagulation pathway, the 32 phenotypes selected for this study may show a common pattern of activity. Moreover it is known that the genes coding for the different coagulation factors share a joint ancestry, so that there may also exist some regulatory elements jointly affecting them. In order to capture the information shared by the 32 coagulation phenotypes, the concept of ”metaphenotype” is proposed.

A metaphenotype is defined as a synthetic phenotypic variable obtained from a set of real phenotypes through a given mathematical model. This new variable should be able to capture the structure of the original data with the goal of describing them as a whole.

Here, several metaphenotypes were characterized to describe the several phenotypes involved in the coagulation cascade. These variables aim to capture the variability shared by the phenotypes belonging to the pathway. In order to determine the pleiotropic effects of genetic variants on this set of phenotypes, these variables are considered as new phenotypic unities.

Identifying genetic variants related with these metaphenotypes may help to ascertain the genetic bases of the complete metabolic process.

Two mathematical models were used for building the metaphenotypes, Principal Component Analysis (PCA) and Independent Component Analysis (ICA). Both of them transform the original data in a new subspace, generally of lower dimension. In both cases, the components correspond to the new system of coordinates. In PCA, the components are obtained with the criterion of maximizing the proportion of the phenotypic covariance. On the other hand, in ICA, the components are obtained with a criterion of maximizing the independence of their projections, in order to ensure that the different components obtained are mutually independent in a complete statistical sense [302]. PCA has been widely used in statistics for feature extraction and more particularly it has already been used in the context of the genetic analysis of multiphenotypic schemes [197]. In contrast, ICA has been employed in a wide range of potential applications in telecommunications or medical signal treatments but it is still not commonly used in statistics and has not been applied in multiphenotypic problems.

8.3.1.2 Principal Component Analysis

The PCA methodology consists on using an orthogonal transformation to convert a set of observations of possibly correlated variables (here the phenotypes of the coagulation pathway), into a set of values of uncorrelated variables called principal components. Each principal component is a linear combination of the original variables. In PCA, the criterion used for selecting the components is based on maximizing the covariance of the original data captured by the components. The principal components correspond to the basis vectors of the new system of orthogonal axes and their dimensionality is generally the same than the original data, even if the variance of the original data is usually explained by the first few principal components.

The mathematical model underlying the metaphenotype obtention is described as follows. Let X be the matrix of phenotypes, of size $M \times N$ where M is the number of individuals and N the number of phenotypes involved (here 32). Each measurement in this N-dimensional space is defined by a point. The purpose of PCA is to introduce a new set of n orthogonal axes (generally $n \leq N$) in such a way that the projection of the original data on the first principal axis shows the highest variance, the second highest variance component is projected on the second principal axis, and so on, with the remaining variance being shown along the remaining axes. These axes are referred to as principal component axes or simply principal components.

The original phenotypes X may be expressed as a linear combination of the principal components in the new axis system, as defined in equation 8.1.

$$X = T^\top \cdot P^\top + \epsilon \tag{8.1}$$

where T are the linear coefficients, also called scores, and P are the weight of the phenotypes, also called loadings, and where ϵ denotes the residual error.

The principal components are orthogonal to each other so they are uncorrelated and do not contain redundant information. They are obtained by diagonalizing the covariance matrix of the original phenotypes. The eigenvectors are the loadings of the PCA, and they correspond to the direction of the principal components. Each eigenvalue is proportional to the portion of the variance captured by the corresponding component.

8.3.1.3 Independent Component Analysis

While the goal in PCA is to maximize the covariance of the data, the goal of ICA is to minimize the statistical dependence between the basis vectors. As for the PCA model, the original phenotypes X can be expressed as a linear combination of the independent components as in equation 8.2.

$$X = AS + \epsilon \tag{8.2}$$

where A are the linear coefficients, or weights, and S are the independent components and where ϵ denotes the residual error.

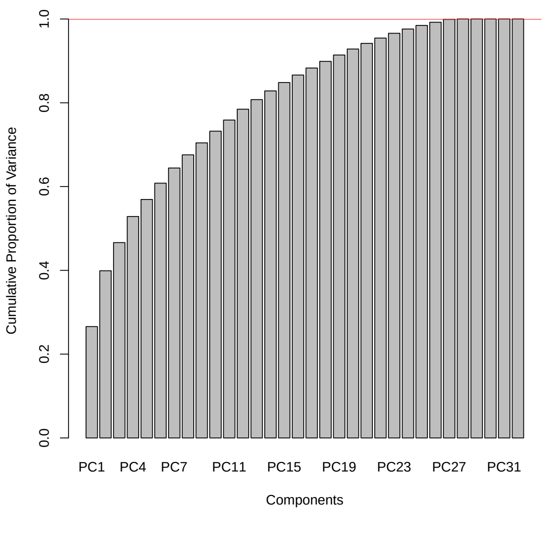


Figure 8.2: Cumulative proportion of variance captured by the principal components. Red lines represent two thresholds corresponding to 85% and 100% used for determining the suitable number of components to be used in ICA.

Unlike PCA, the basis vectors in ICA are neither orthogonal nor ranked in order. Also, there is not a closed form expression to find ω_i that maximize the independence of the components x_i. To do so, S are expressed as in equation 8.3.

$$S = WX \tag{8.3}$$

The independence of the components is guaranteed by finding W that maximizes the nongaussianity of S. Measures of nongaussianity are commonly applied for ICA algorithms such as the kurtosis or the negentropy [131]. Otherwise, other criteria exist such as a maximum likelihood criterion or the minimization of the mutual information between components. Among the several existing ICA algorithms, the fastICA procedure was applied, using a particular approximation of the negentropy measure for maximizing the nongaussianity. FastICA was selected due to its desirable properties when compared with other existing ICA methods [131].

The employed fast ICA algorithm does not include any criterion for determining how many components represent the dynamic structure of the data. As other ICA implementations, it previously applies a PCA to the data in order to ensure that the components are uncorrelated. Generally, the number of components of ICA is determined by the number of components in PCA. Since there is not a standard practice to determine this number of components, two strategies were explored. On one hand, a cross-validation approximation of the selection of the optimal number of components in PCA was considered [141]. This strategy determined that the optimal number of components was 15. On the other hand, a criterion based on the proportion of variance captured by the principal components was established. Figure 8.2 shows the proportion of variance captured by the principal components for the PCA applied to the original coagulation phenotypes. This criterion selected 30 components since it is the value for which the cumulate proportion of variaance achieves the 100% of the total variance. Thus, ICA was applied twice, with 15 and 30 components respectively, as suggested by these two criteria.

8.3.2 Genetic Analyses

The third part of the methodology consists in finding genetic association between genetic variants and the new phenotypic variables defined by the metaphenotypes.

A traditional GWAS design for familiar data was carried out [46]. This methodology is an extension of the traditional test of association based on linear regressions when individuals are correlated. It fits a simple variance-components model to the data which provides a vector of fitted values of the phenotype and an estimate of the variance-covariance matrix for each family [17]. The polygenic mixed model defined in equation 8.4 was applied for each metaphenotype with the age and gender covariables for testing the association as they present a significant correlation with almost all the metaphenotypes.

$$Y_i \sim \mu + \sum_j \beta_j c_{ji} + G_i + \epsilon_i \tag{8.4}$$

where Y_i is the metaphenotype, μ is the overall mean, β_j is the estimate of the j-th covariate, c_{ji} is the j-th covariate, G_i is the random additive polygenic effect (breeding value) which variance is defined as $\Phi\sigma_G$ where Φ is the kinship matrix and σ_G is the additive genetic variance due to polygenes. Finally ϵ_i are the residuals of the model.

GWAS were performed for each metaphenotype with GenABEL (v. 1.7-3) in R.

8.4 Results and discussion

A total of 77 metaphenotypes were obtained with the different methods described in section 8.3.1. 32 of them correspond to the 32 components extracted from PCA. The remaining metaphenotypes correspond to the application of the ICA algorithm with 15 and 30 components respectively. 12 of them presented significant results in association. These results are summarized in table 8.2. Presented significance levels correspond to adjusted p-values [16].

The results obtained for the 12 metaphenotypes presented in table 8.2 have been studied in more detail. The full results of the GWAS are plotted in even-numbered Figures 8.3 to 8.25.

In order to graphically describe the metaphenotypes, the coagulation cascade is plotted using a simple graph, where each coagulation phenotype is represented by a node whose color is determined proportionally to its contribution to the corresponding metaphenotype. This contribution was measured by the loading values of the trait in the corresponding model, representing the weight of the trait within the metaphenotype. Odd-numbered Figures 8.4 to 8.26 show the graph of the coagulation cascade for the metaphenotypes.

For both the 8-th and the 9-th components of the PCA model, SNP $rs9898$ was found to be significantly associated with the metaphenotype, indicating that it may be related with the entire coagulation process. $rs9898$ is located in the *HRG*, the functional gene of the Histidine Rich Glycoprotein, a protein involved in the coagulation pathway, and more specifically in the intrinsic pathway of the coagulation. This SNP is reported to be related with thrombosis [216, 123]. As observed in figures 8.4 and 8.6, in both cases the HRG protein has a high weight in the corresponding metaphenotype. In the case of the 8-th component, HRG has a low negative weight, whereas for the 9-th the component, it has a high positive weight. This result was expectable, since the obtained association signal may be majorly explained by this specific phenotype.

For the 10-th component, *rs*2731672, the only SNP located in the *F12* showed a strong signal of association (p-value of 1.48×10^{-11}). It is observed in Figure 8.8 that the FXII protein has a dominant negative weight in this metaphenotype. As before, this result was predictable.

For the 22-nd component, *rs*1553514, located in the *CNTN5* gene was significant. However this gene has no apparent relation with the coagulation process.

For the 23-rd component of PCA, *rs*11057761 obtained a significant p-value (2.14×10^{-07}). This is an intergenic SNP between *SCARB1* and *NCOR2* genes. *SCARB1* corresponds to the Scavenger receptor class B, member 1. It is a a plasma membrane receptor for high density lipoprotein cholesterol (HDL). The encoded protein mediates cholesterol transfer to and from HDL. HDL reportedly functions as a cofactor to the anticoagulant activated protein C (APC) in the degradation of factor V [229]. On the other hand *NCOR2* (nuclear receptor co-repressor 2) is a transcriptional co-regulatory protein that serves as a repressive co-regulatory factor (co-repressor) for multiple transcription factor pathways, such as the HDAC, SIN, HNF4A and JUN transcription factor families. It is noteworthy that HNF4A transcription factor has an important role in the transcription control of coagulation factors [133]. JUN also regulates several genes related to the coagulation. Among them, the gene PLAU (Plasminogen activator) has a direct relation with the coagulation cascade. In addition, JUN regulates three interleukins (*ILB1*, *IL2* and *IL6*), related with the coagulation as markers of inflammation [293, 56] and five genes of the MMP family, related with arteriosclerosis [310, 275].

In addition, looking at protein-protein interactions, *JUN* was found close to the coagulation cascade, concretely at distance 2 from FV. The distance *JUN* to the coagulation pathway was defined as the minimum distance from the candidate gene to all genes in the pathway, where the distance between two genes is the length of the shortest path from one gene to the other in the PPI network. The gene that separates them is *CSNK2A1*, a protein kinase that phosphorilates the FV, inducing FV inhibition from Protein C. AT the same time, *CSNK2A1* also phosphorilates *JUN*. Then, *JUN* and *FV* could be competitors substrates of the CSNK2A1.

In order to illustrate the obtained results, Figure 8.29 shows the coagulation pathway adapted from the KEGG database with the paths relating the candidate genes with the coagulation cascade. Added paths were represented with a dashed line as they are only partially detailed.

For the ICA model built with 30 components, significant associations were found for 5 components. For the 2-nd component, SNP *rs*17255413 obtained a p-value of 4.26×10^{-09}. This SNP is located in the *BOC* gene (cell adhesion molecule-related/down-regulated by oncogenes). It codes for a cell surface receptor of the immunoglobulin/fibronectin type III repeat family involved in myogenic differentiation. Despite the low MAF of the SNP

(0.007), the *BOC* gene shows plausibility to be related with coagulation, since cell adhesion molecules have much to do with the activation of the coagulation process [168, 13]. For the 15-th component, $rs6687825$ and $rs6691481$ located in the *UBR4* gene were found to be associated with the metaphenotype. The *UBR4* (Ubiquitin Protein Ligase E3 Component N-Recognin 4) is a component of the N-end rule pathway with apparently no relation to the coagulation process. However, several miRNA were found in this genomic region, with plausibility to be in LD with the SNP. MiRNA are small non-coding RNA molecule that have transcriptional functions on some specific target genes. In particular, one of these miRNAs, miRNA-4695, targets the *F8* gene expression. It is in concordance with Figure 8.16 where it is observed that FVIII has an important weight in this metaphenotype. Finally, for the 22-nd component, $rs867186$, located in the *PROCR* (Protein C receptor), obtained a p-value of association of 1.12×10^{-08}. This result is in concordance with Figure 8.18, where it is observed that the protein C has an important weight in this metaphenotype. Significant association of SNPs in *PROCR* have been previously reported [14].

Four of the components obtained using an ICA model of 15 components presented significant results in GWAS. For the 3-rd component, 4 SNPs were found in the same region of chromosome 3 that have a strong evidence of genetic association with the metaphenotype (p-value 9×10^{-18} for SNP $rs9898$). As described for the 8-th component of PCA, this SNP belongs to the *HRG* gene and is related to the coagulation and with thrombosis disease. However, in this case the result is more surprising since, as observed in Figure 8.20, the HRG protein has a neutral weight in this metaphenotype.

In addition, for this metaphenotype, two SNPs belonging to the *KNG1* gene (Kininogen-1)showed significant association. *KNG1* has been previously reported as a genetic determinant related to the coagulation since it plays an important role on both FXI and FXII activations [259]. Given that these two proteins have not an important weight in this component, the result is especially interesting suggesting that *KNG1* could have a more global effect on the coagulation cascade.

For the 4-th component, SNP $rs17255413$, located in the *BOC* gene presents a significant score of association. This SNP was also found with the 2-nd component of the ICA applied with 30 components.

For the 10-th component, where the FXII protein has an important weight, the only SNP of the *F12* is recovered in GWAS (p-value 1.05×10^{-14}).

It is observed that two reported results are obtained using both PCA and ICA with 15 components. In order to compare the results, the relationship between both metaphenotypes was graphically represented in Figures 8.27 and 8.28. In both figures, the metaphenotypes obtained using both methods are compared in terms of the weights and the scores of the models. The loadings of the associated protein were plotted in red. Individual scores wee differentiated in color by the genotype they carry at the obtained SNP.

In both cases, the scores of both models show a moderate, yet significant correlation (p-value 2×10^{-16}). However, in the case of the HRG, the loadings are not related, whereas for the *F12*, an evident correlation is observed in Figure 8.28. In the former case, as previously commented, HRG has an important contribution in the metaphenotype extracted with PCA, whereas it does not have a particularly relevant weight in the metaphenotype extracted from ICA. This confirmed the validity of the proposed methodology, showing its capacity to capture the propagating effect of the *HRG* in the coagulation cascade.

Table 8.2: Significant SNPs obtained in GWAS for different metaphenotypes. For each metaphenotype, the SNPs with an adjusted p-value lower than 1×10^{-06} are presented, jointly with their MAF and their chromosomic region and the closest gene to the loci.

Method	Component	SNP	Chromosome	MAF	Gene	p-value
PCA	8	rs9898	3	0.35	HRG*	1.02×10^{-07}
	9	rs9898	3	0.35	HRG*	4.26×10^{-08}
	10	rs2731672	5	0.17	F12*	1.48×10^{-11}
	22	rs1553514	11	0.25	CNTN5	4.21×10^{-07}
	23	rs11057761	12	0.29	SCARB1, NCOR2	2.14×10^{-07}
ICA 30 components	2	rs17255413	3	0.007	BOC	7.6×10^{-09}
		rs2037516	2	0.1	DSU	6.47×10^{-07}
	15	rs6687825	1	0.12	UBR4	5.81×10^{-07}
		rs6691481	1	0.17	UBR4	6.43×10^{-07}
		rs4747989	10	0.07	CAMK1D	5.81×10^{-07}
	22	rs867186	22	0.1	PROCR*	1.12×10^{-08}
		rs3795149	20	0.02	C20orf135	3.28×10^{-07}
		rs6062561	20	0.02	TPD52L2	3.37×10^{-07}
ICA 15 components	3	rs9898	3	0.35	HRG*	9×10^{-18}
		rs3733159	3	0.34	FETUB	6.58×10^{-09}
		rs1621816	3	0.24	KNG1*	4.96×10^{-08}
		rs1403694	3	0.32	KNG1*	6.72×10^{-07}
	4	rs17255413	3	0.007	BOC	2.62×10^{-08}
	5	rs3113727	4	0.24	COL25A1	3.83×10^{-07}
	10	rs2731672	5	0.17	F12*	1.05×10^{-14}

* Previously reported genes related to the coagulation pathway.

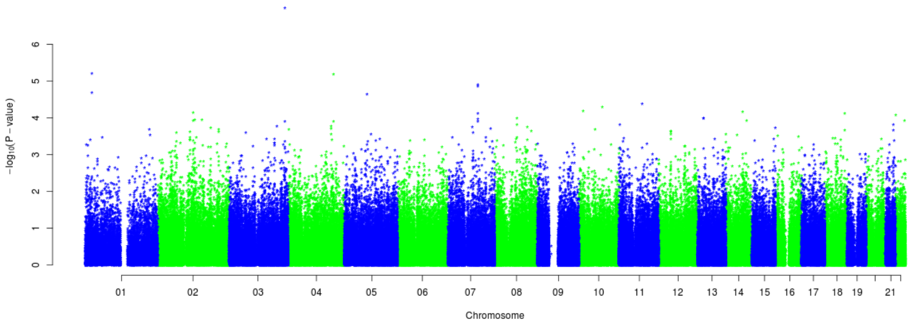


Figure 8.3: GWAS for the 8-th component of the PCA model.

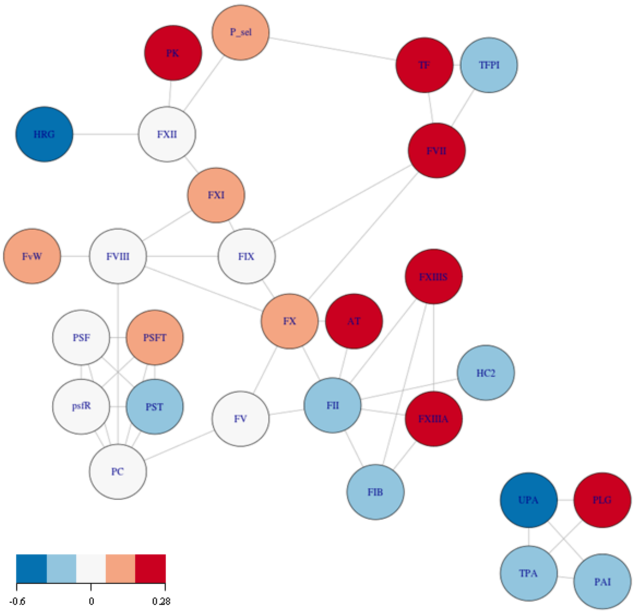


Figure 8.4: Graphical representation of the contribution of each coagulation phenotype on the metaphenotype corresponding to the 8-th component of the PCA model.

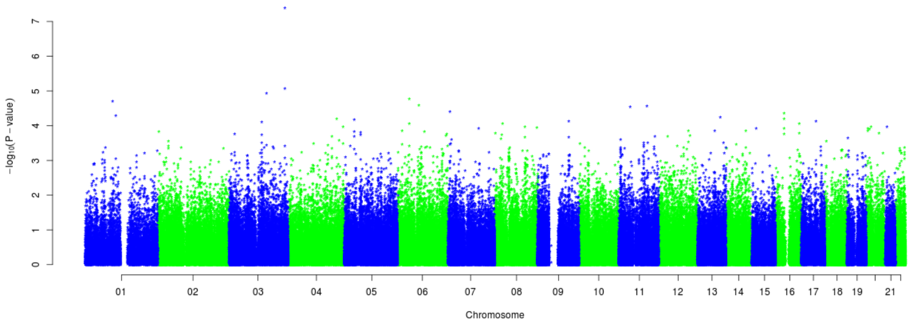


Figure 8.5: GWAS for the 9-th component of the PCA model.

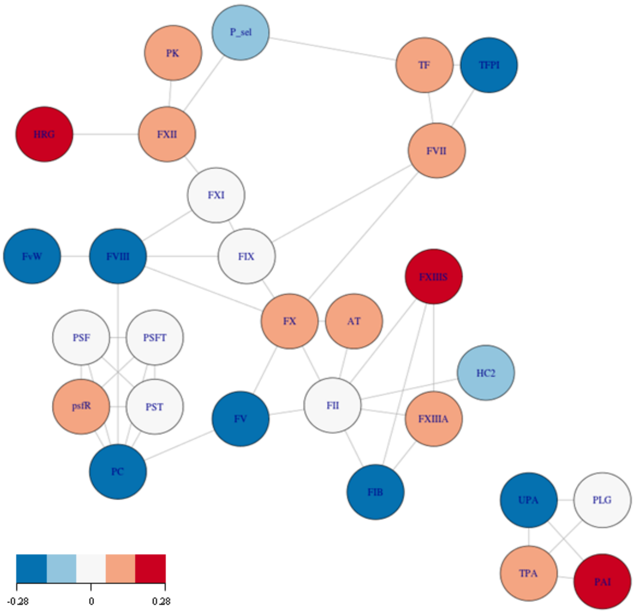


Figure 8.6: Graphical representation of the contribution of each coagulation phenotype on the metaphenotype corresponding to the 9-th component of the PCA model.

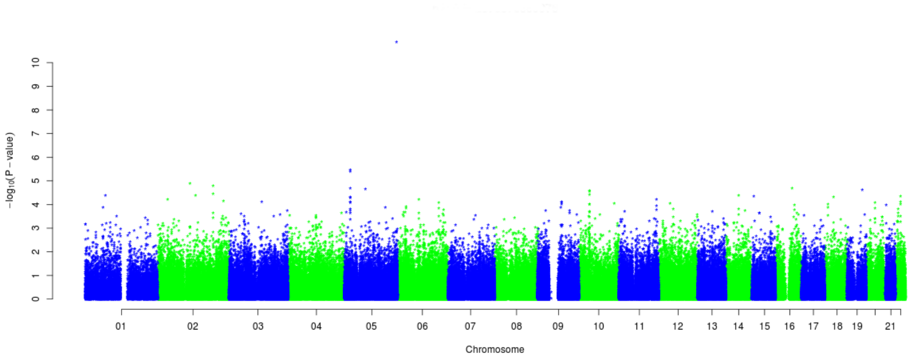


Figure 8.7: GWAS for the 10-th component of the PCA model.

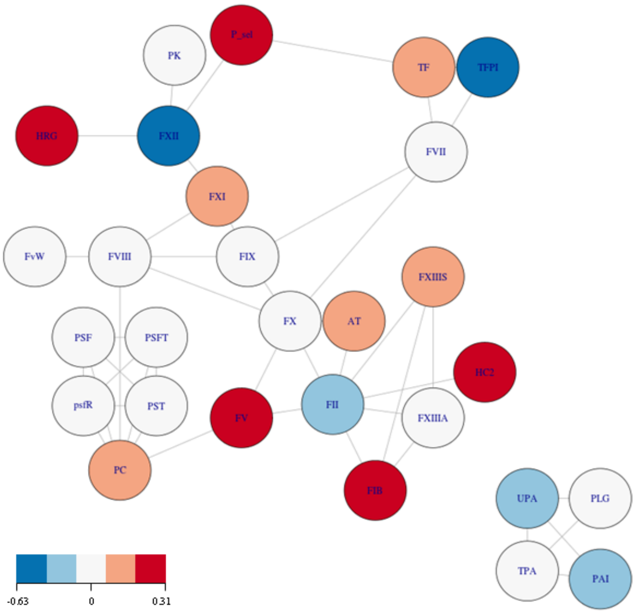


Figure 8.8: Graphical representation of the contribution of each coagulation phenotype on the metaphenotype corresponding to the 10-th component of the PCA model.

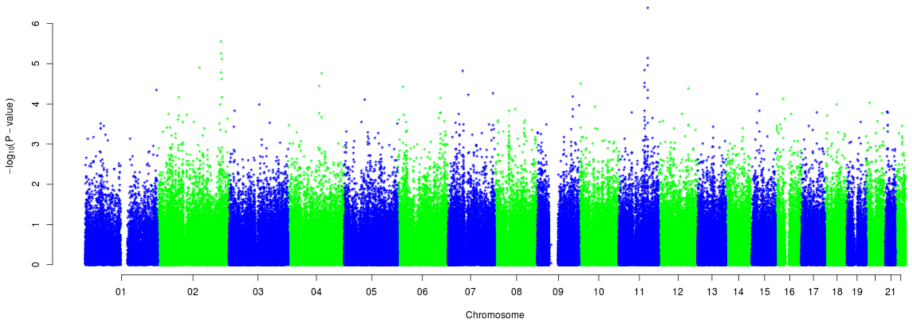


Figure 8.9: GWAS for the 22-nd component of the PCA model.

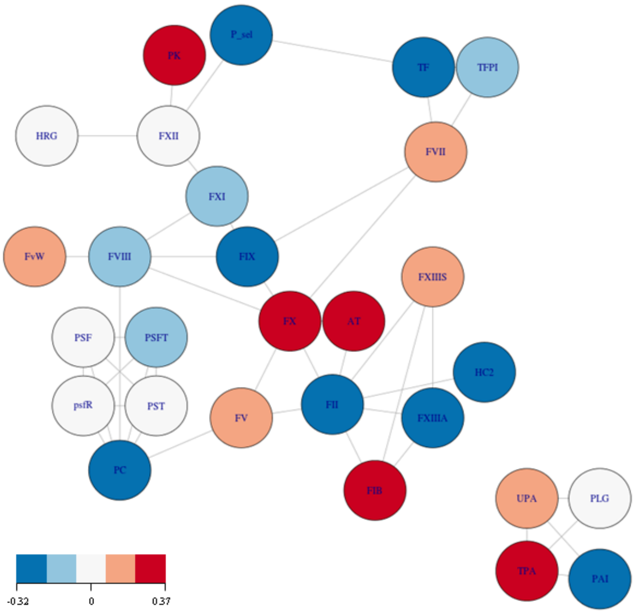


Figure 8.10: Graphical representation of the contribution of each coagulation phenotype on the metaphenotype corresponding to the 22-nd component of the PCA model.

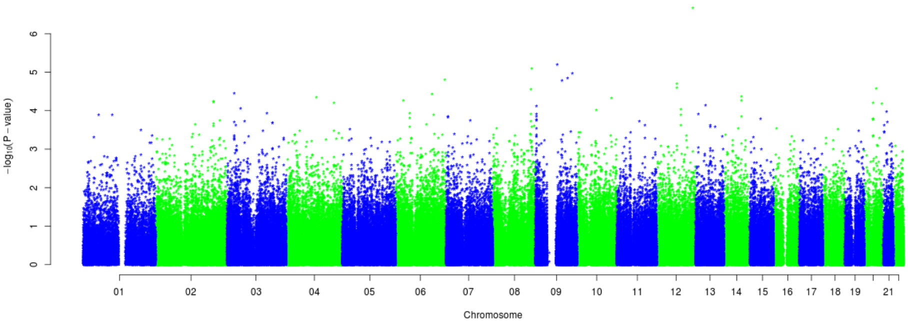


Figure 8.11: GWAS for the 23-rd component of the PCA model.

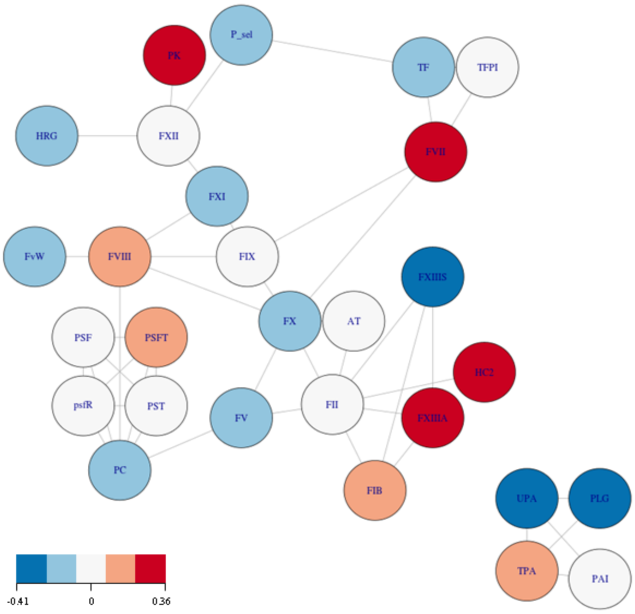


Figure 8.12: Graphical representation of the contribution of each coagulation phenotype on the metaphenotype corresponding to the 23-rd component of the PCA model.

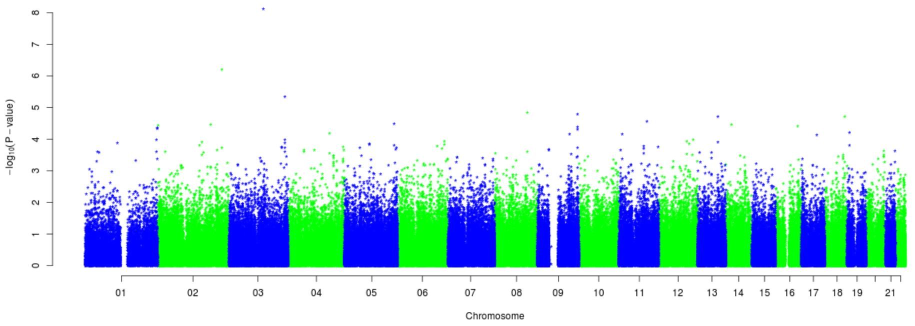


Figure 8.13: GWAS for the 2-nd component of the ICA model built with 30 components.

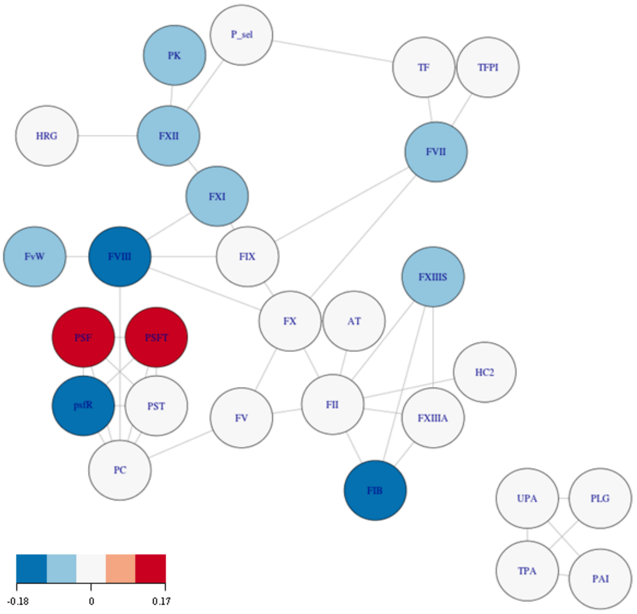


Figure 8.14: Graphical representation of the contribution of each coagulation phenotype on the metaphenotype corresponding to the 2-nd component of the ICA model built with 30 components.

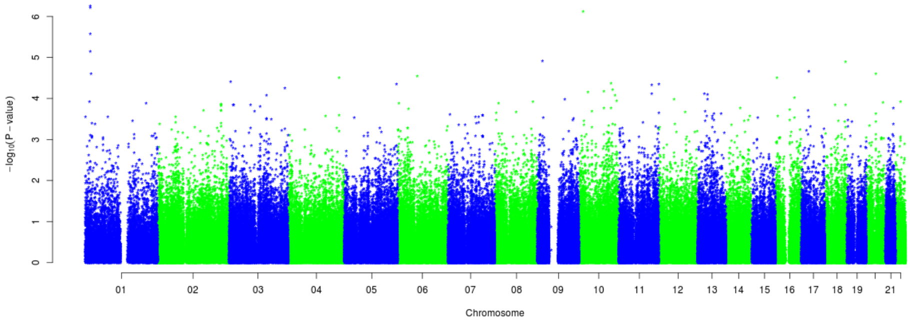


Figure 8.15: GWAS for the 15-th component of the ICA model built with 30 components.

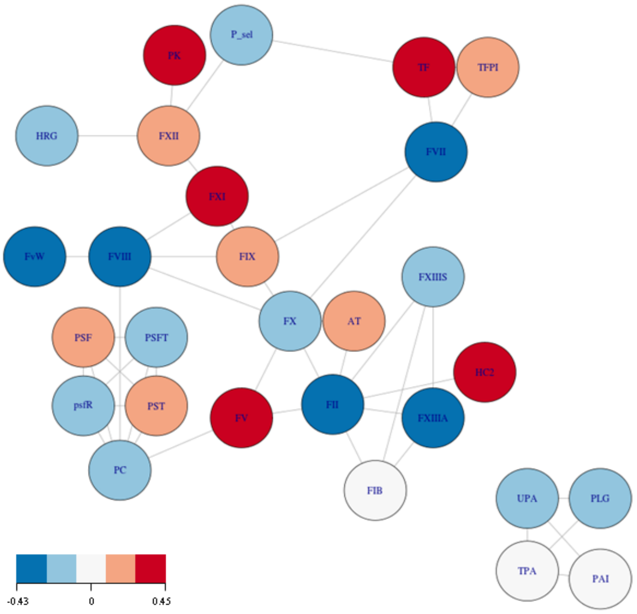


Figure 8.16: Graphical representation of the contribution of each coagulation phenotype on the metaphenotype corresponding to the 15-th component of the ICA model built with 30 components.

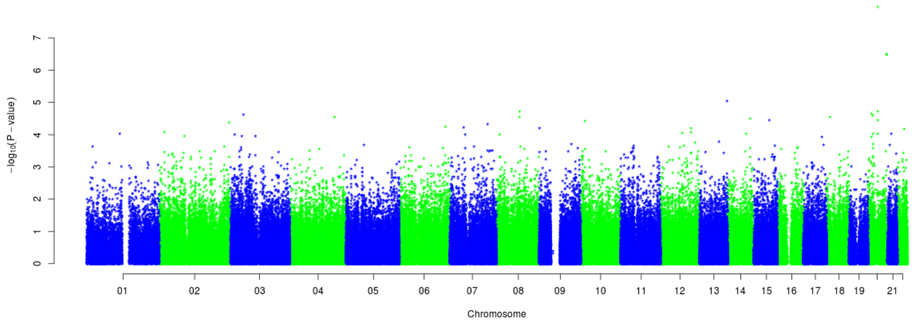


Figure 8.17: GWAS for the 22-nd component of the ICA model built with 30 components.

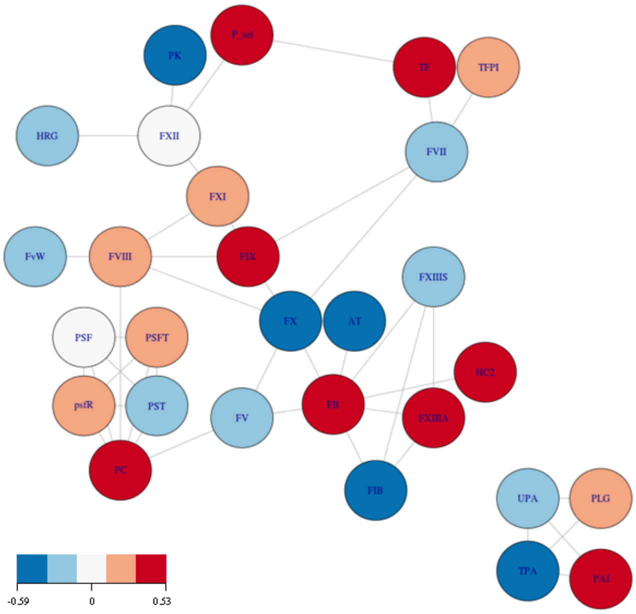


Figure 8.18: Graphical representation of the contribution of each coagulation phenotype on the metaphenotype corresponding to the 22-nd component of the ICA model built with 30 components.

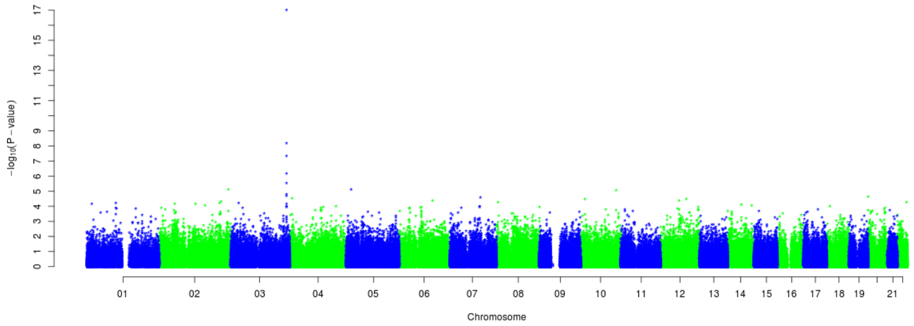


Figure 8.19: GWAS for the 3-rd component of the ICA model built with 15 components.

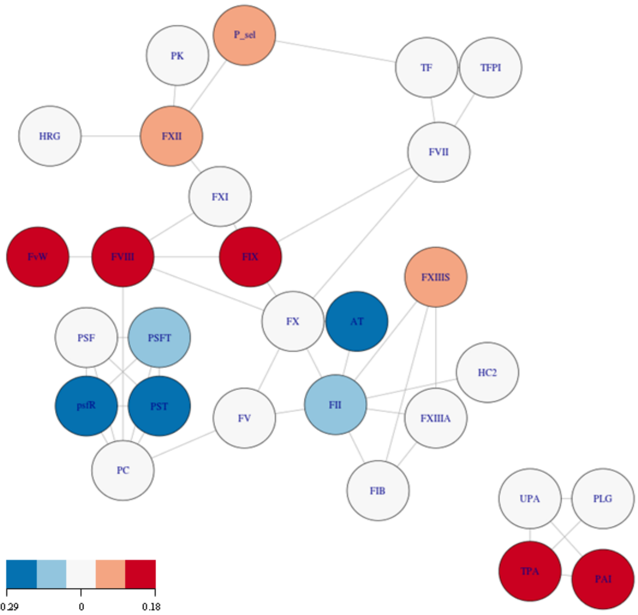


Figure 8.20: Graphical representation of the contribution of each coagulation phenotype on the metaphenotype corresponding to the 3-rd component of the ICA model built with 15 components.

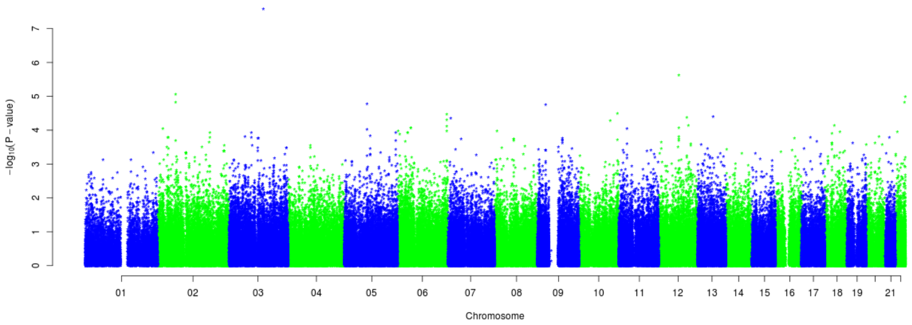


Figure 8.21: GWAS for the 4-th component of the ICA model built with 15 components.

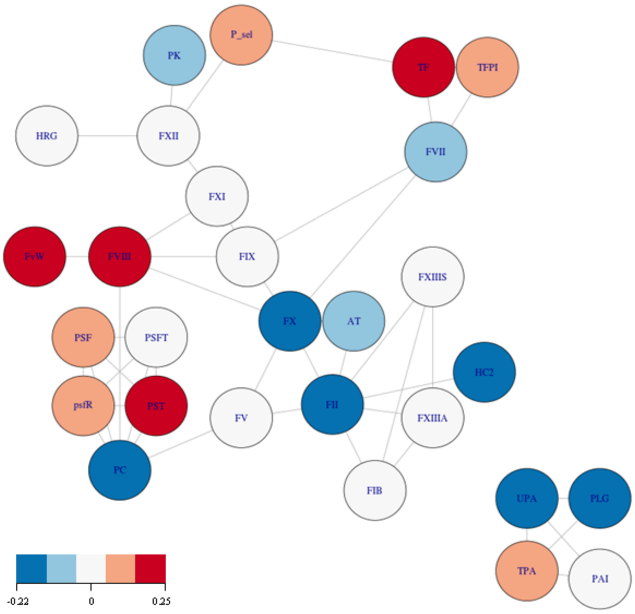


Figure 8.22: Graphical representation of the contribution of each coagulation phenotype on the metaphenotype corresponding to the 4-th component of the ICA model built with 15 components.

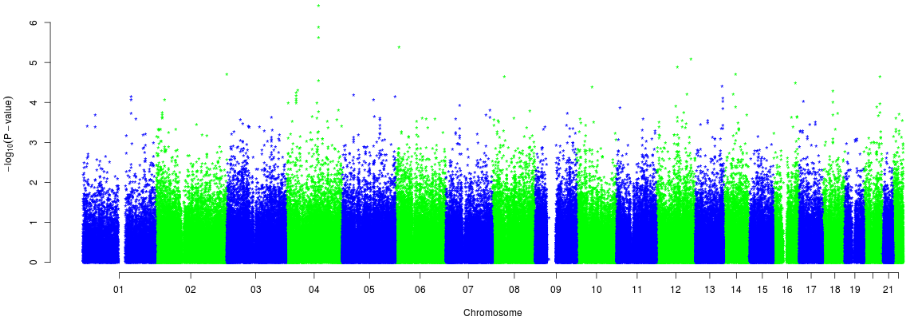


Figure 8.23: GWAS for the 5-th component of the ICA model built with 15 components.

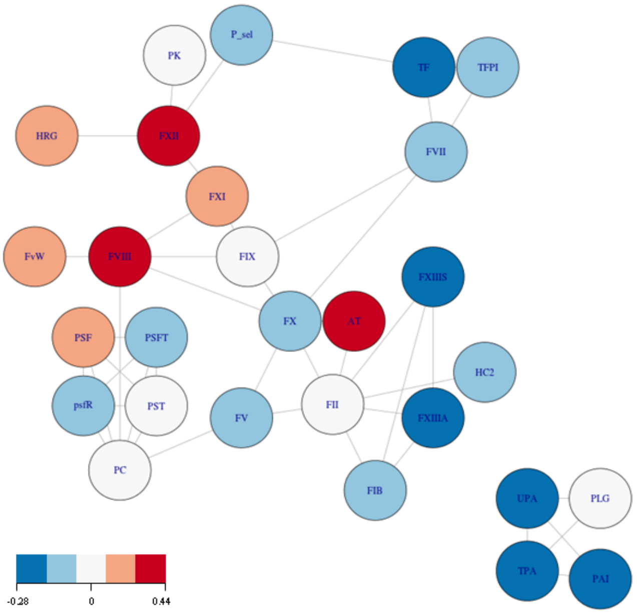


Figure 8.24: Graphical representation of the contribution of each coagulation phenotype on the metaphenotype corresponding to the 5-th component of the ICA model built with 15 components.

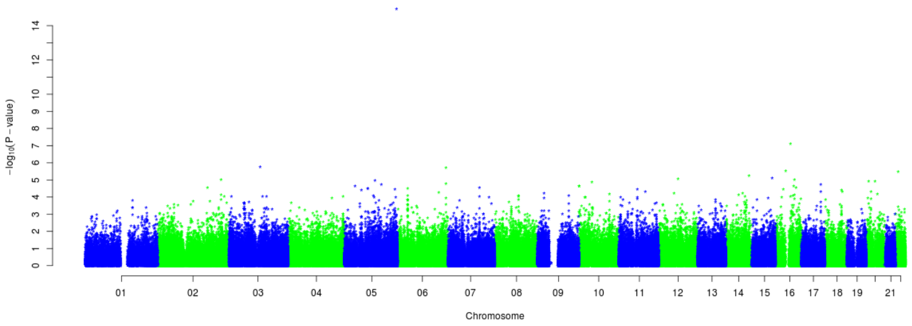


Figure 8.25: GWAS for the 10-th component of the ICA model built with 15 components.

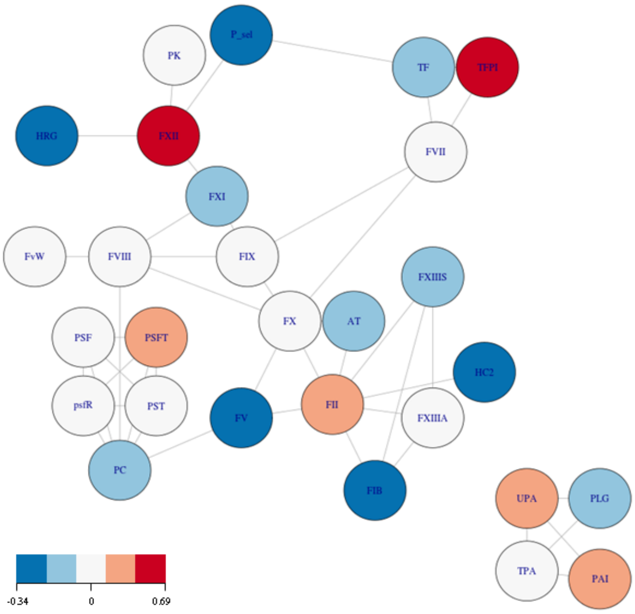


Figure 8.26: Graphical representation of the contribution of each coagulation phenotype on the metaphenotype corresponding to the 10-th component of the ICA model built with 15 components.

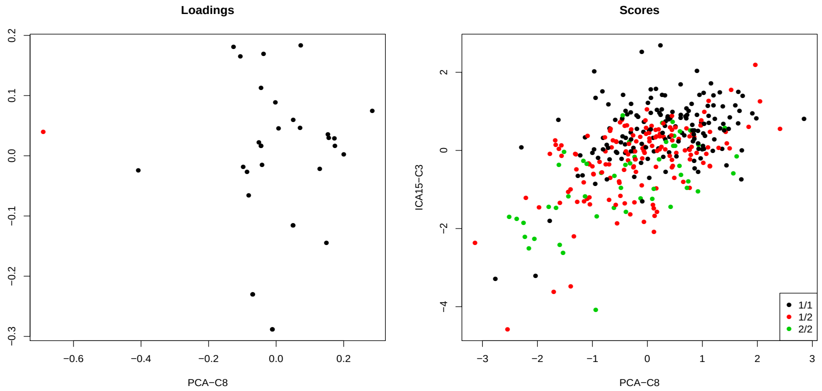


Figure 8.27: Relationship between the metaphenotypes corresponding to the 8-th component of the PCA model and the 3-rd component of the ICA model built with 15 components, related with the SNP $rs9898$ in the *HRG* gene.

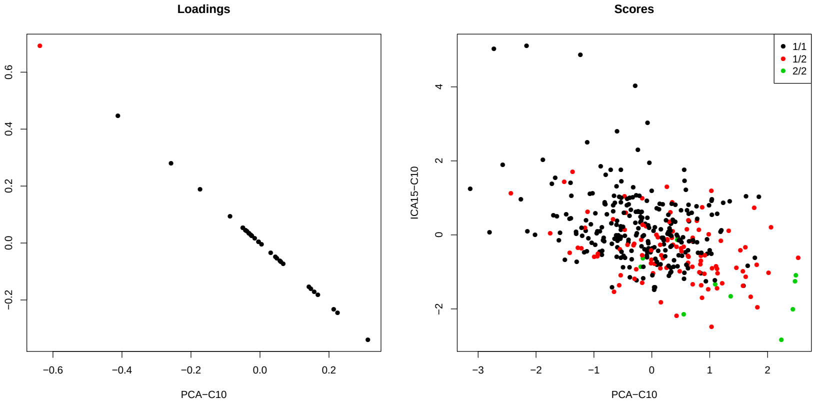


Figure 8.28: Relationship between the metaphenotypes corresponding to the 10-th component of the PCA model and the 10-th component of the ICA model built with 15 components, related to the SNP $rs2731672$ in the *F12* gene.

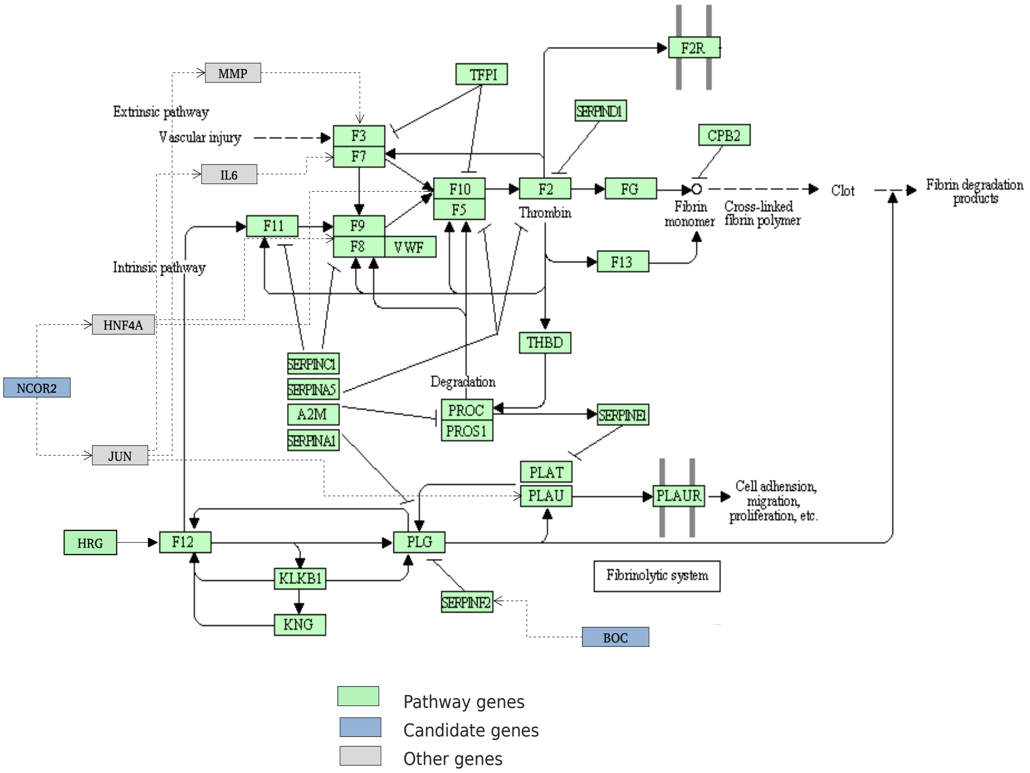


Figure 8.29: Graph relating the candidate genes to the coagulation cascade [Adapted from Kegg [143]].

8.5 Concluding remarks

In this chapter, a novel methodological framework was proposed for the simultaneous analysis of multiple intermediate phenotypes involved in complex diseases. As a practical implementation, the methodology was applied to the GAIT project dataset, with the aim of identifying genetic markers involved in thrombosis.

The methodology consisted in building new canonical phenotypic variables, named metaphenotypes, that capture the joint activity of the intermediate phenotypes involved in the coagulation pathway. Three different mathematical models were built for the definition of the metaphenotypes. First a PCA was carried out obtaining 32 metaphenotypes corresponding to the components of the model. In addition, a novel methodology in this field, ICA, was applied. Two ICA were performed using 30 and 15 components respectively, obtaining a total of 45 new metaphenotypes.

Among the 77 metaphenotypes, 12 presented significant results in GWAS. The obtained results can be classified into two categories, those previously reported and those conforming new plausible candidates to be related to the coagulation process as a whole.

Most of the previously reported results correspond to metaphenotypes in which one particular protein has a significant contribution in comparison with the others. In these cases, as expected, the top hits from the GWAs correspond to SNPs located in the genomic region of the main contributing protein. For example, for two of the metaphenotypes with an important contribution of the FXII, significant scores obtained in GWAS correspond to the only SNP belonging to the *F12* gene.

In contrast, for the metaphenotype corresponding to the 3-rd component of the ICA applied with 15 components, 3 significant SNPs were found in a region of the chromosome 3 containing the *HRG* and *KNG1* genes. Both genes have already been related to thrombosis. However, this particular metaphenotype is not specifically oriented to the related coagulation phenotypes, suggesting an unknown global effect of the gene on the coagulation pathway as a whole.

On the other hand, a SNP from the intergenic region between *NCOR2* and *SCARB1* showed a significant association with the metaphenotype corresponding to the 23-rd component of the PCA. *NCOR2* is a co-regulatory gene that regulates several families of transcription factors. Among them two transcription factors were detected to be specifically related to the coagulation cascade (coded by genes *HNF4A* and *JUN*). Thus, *NCOR2* seems to be a possible candidate to have a global regulatory function on the coagulation pathway has a whole. It has been proposed to biologists for further functional analyses.

The novelty of this work lies in addressing the genetic analysis of thrombosis and of the coagulation phenotypes from a multidimensional perspective, defining new indexes of joint activity of this metabolic process. Using this approach we were able to retrieve already known associations but also to propose new candidates with evidence to have a global regulatory effect on the multiple coagulation phenotypes.

Part III

Conclusions

Chapter 9

Conclusions

This chapter aims to summarize the issues addressed in this dissertation and the contributions that were made towards their solution. This thesis research is focused on the genetic study of complex diseases and in particular on genetic association studies, with the goal of studying statistical associations between genetic polymorphisms and phenotypes or disease states leading to the identification of potential genetic risk factors. The original goal of this thesis was to develop nonlinear methods for carrying out genetic association studies for complex diseases. In section 9.1 we enumerate the original contributions proposed in this direction and in section 9.2 we sketch some future directions to improve this work.

9.1 Original Contributions

- A critical review on genetic association methods was performed. This survey revealed that the methods used for genetic association studies are mainly linear. This motivated the goal of addressing the genetic association problem through nonlinear methods. In this direction, a review on information-theoretic measures and their application to genetic research was also carried out, indicating that mutual information would be an appropriate measure for genetic association.

- An exploratory study was carried out to explore the properties of the sequence variability of the SNPs related to disease. A statistical analysis comparing the SNP sequence variability between SNPs related to disease and neutral SNPs at different genetic regions was carried out. The results showed that for most of the regions, SNPs related to disease tend to be less variable across species than neutral SNPs. This observation was in concordance with previous results showing that functional genetic regions tend to be more conserved across species than nonfunctional regions.

- A nonlinear test for one-locus genetic association was designed and characterized, based on the mutual information as a measure of association. Three different estimations for the mutual information measures were considered. The statistical significance of an association was determined with the design of a statistical test based on a null distribution of mutual information. The proposed association test took into account the genetic stratification of the population. This novel methodology was compared with the standard procedure in genetic association based on linear regressions. The obtained results were neutral, showing that the proposed nonlinear methodology is able to recover the results obtained with standard linear models. The proposed methodology is sensible to false-positive findings in a similar manner than traditional techniques, obtaining performances of the same order of magnitude in terms of classification.

- The methodology employed in the one-locus test was applied in a multi-loci association study. In this case a novel multi-solution floating feature selection algorithm was proposed for the search of multi-loci interactions related with a given phenotype. The mutual information-based association test was defined as the relevance criterion of the algorithm. With a comparison purpose, the feature selection algorithm was applied with the traditional genetic association measure based on linear regressions. Both strategies were compared to MECPM, an existing multi-loci association algorithm. This study was performed with the SNPs of the F7 gene. The results obtained using this methodology were consistent with the information found in in the literature in several functional analysis on the F7 gene. Moreover, the results obtained improved the results from traditional methods found in the literature. An important result was to find combinations of SNPs significantly associated with the phenotype, while individually they did not show a statistically significant association. Although the proposed nonlinear test obtained neutral results at a one-locus scale, it obtained positive and relevant results for detecting multi-loci interactions.

- The set of routines developed for the multi-loci genetic association procedure were integrated in an R package called MISS (Mutual Information Statistical Significance), which is available at `http://www.sisbio.recerca.upc.edu/R/MISS_0.2.tar.gz`.

- A novel methodological framework was proposed for addressing the genetic analysis of complex diseases from a multidimensional point of view. The methodology consisted in building new phenotypic variables, named metaphenotypes, from a set of phenotypes involved in a given biological process or disease. As a practical implementation, the methodology study was applied to the phenotypes of the coagulation pathway for the GAIT project data. 77 metaphenotypes were obtained corresponding to the application of one PCA and two ICA with 15 and 30 components respectively. Among them only 12 obtained significant results in genetic association. Some of the results corresponded to already known associations but new candidates were also proposed as master regulatory genes witha global effect on the coagulation process as a whole.

9.2 Future extensions

This section aims to analyze the principal limitations of the research performed in this thesis and to propose future research directions.

In this work, a nonlinear genetic association test was proposed. The first characterization of this test showed that it is able to compete with other genetic association tests, even if some optimizations could be necessary to improve its performance. Genetic association tests, and in particular the proposed method have some difficulties to detect rare variants. In order to improve the performance of the proposed methodology in these cases, it could be interesting to apply parametric measures of information theory that may be more sensible to low MAFs such as the Renyi divergence measures. Moreover, another limitation of one-locus genetic association studies and in particular of the proposed nonlinear test is the false positive findings. We also propose as a future extension of this work the application of prioritization criteria for complementing the results found with the association test. These criteria may help to determine the genetic variants really related to the phenotype. We propose to incorporate protein-protein interaction networks in order to prioritize the genes that are related with a given biological process. It would also be interesting to expand the exploratory study on the cross-species sequence variability and to establish a prioritization criterion based on the premise that relevant SNPs may be less variable across species than neutral SNPs.

A floating feature selection algorithm was proposed in this thesis for multi-loci genetic association studies. This algorithm was applied with two different criteria, the standard linear regression-based association criterion and a nonlinear criterion based on the mutual information measure, as for the one-locus association test. Both experiments were compared with MECPM, a methodology for multi-loci association based on entropy maximization and using a greedy search procedure. In order to extend this study, it would be interesting to apply both the linear and nonlinear relevance criteria within other feature selection algorithms, that would be suboptimal but faster. As for the one-locus test, the characterization of the analytical expression for the mutual information null distribution would considerably optimize the performance of this methodology.

Regarding the methodology proposed in chapter 8, for the genetic study of multiphenotypic schemes, the main extensions we propose lies in the definition of the metaphenotypes. On one hand, in the present study the metaphenotypes were defined with all the phenotypes involved in the coagulation pathway. As a future extension of this work we propose to define the metaphenotypes looking at reduced sets of original phenotypes of the coagulation or also of other related processes. The second extension concerns the methodology used for building the metaphenotype. In this work, the novel ICA methodology was applied with two number of components selected according to two different criteria. We propose as future work to establish a unique and robust criterion for the selection of the number of components in ICA in order to protocolize this novel methodological framework.

Appendix A

The GAW17 dataset simulation model

The following tables explain the simulation model for the phenotypes of the GAW17 dataset, showing the genes and SNPs influencing each phenotype.

Table A.1: Genes and SNPs with effects on Q1.

Gene	SNP	MAF
ARNT	C1S6533	0.0115
ARNT	C1S6537	0.0007
ARNT	C1S6540	0.0014
ARNT	C1S6542	0.0022
ARNT	C1S6561	0.0007
ELAVL4	C1S3181	0.0007
ELAVL4	C1S3181	0.0007
FLT1	C13S320	0.0014
FLT1	C13S399	0.0007
FLT1	C13S431	0.0172
FLT1	C13S479	0.0007
FLT1	C13S505	0.0007
FLT1	C13S514	0.0007
FLT1	C13S522	0.028
FLT1	C13S523	0.0667
FLT1	C13S524	0.0043
FLT1	C13S547	0.0007
FLT1	C13S567	0.0007
FLT4	C5S5133	0.0014
FLT4	C5S5156	0.0007
HIF1A	C14S1718	0.0007
HIF1A	C14S1729	0.0022
HIF1A	C14S1734	0.0122
HIF1A	C14S1736	0.0007
HIF3A	C19S4799	0.0007
HIF3A	C19S4815	0.0007
HIF3A	C19S4831	0.0007
KDR	C4S1861	0.0022
KDR	C4S1873	0.0007
KDR	C4S1874	0.0007
KDR	C4S1877	0.0007
KDR	C4S1878	0.165
KDR	C4S1879	0.0007
KDR	C4S1884	0.021
KDR	C4S1887	0.0007
KDR	C4S1889	0.0007
KDR	C4S1890	0.0022
VEGFA	C6S2981	0.0022
VEGFC	C4S4935	0.0007

Table A.2: Genes and SNPs with effects on Q2.

Gene	SNP	MAF
BCHE	C3S4834	0.0007
BCHE	C3S4836	0.0007
BCHE	C3S4856	0.0007
BCHE	C3S4859	0.0022
BCHE	C3S4860	0.0007
BCHE	C3S4862	0.0007
BCHE	C3S4867	0.0007
BCHE	C3S4869	0.0007
BCHE	C3S4873	0.0029
BCHE	C3S4874	0.0007
BCHE	C3S4875	0.0007
BCHE	C3S4886	0.0007
BCHE	C3S4880	0.0014
GCKR	C2S354	0.0122
INSIG1	C7S5132	0.0007
INSIG1	C7S5133	0.0007
INSIG1	C7S5144	0.0007
LPL	C8S442	0.0158
LPL	C8S476	0.0007
LPL	C8S530	0.0014
PDGFD	C11S5292	0.0086
PDGFD	C11S5299	0.0007
PDGFD	C11S5301	0.0007
PDGFD	C11S5302	0.0014
PLAT	C8S1741	0.0036
PLAT	C8S1742	0.0007
PLAT	C8S1758	0.0014
PLAT	C8S1770	0.0007
PLAT	C8S1772	0.0014
PLAT	C8S1773	0.0014
PLAT	C8S1799	0.0057
PLAT	C8S1811	0.0014
RABR	C3S635	0.0007
RABR	C3S679	0.005
SIRT1	C10S3048	0.0022
SIRT1	C10S3050	0.0022
SIRT1	C10S3058	0.0007
SIRT1	C10S3092	0.0007
SIRT1	C10S3093	0.0007
SIRT1	C10S3107	0.0007
SIRT1	C10S3108	0.0007
SIRT1	C10S3109	0.0007
SIRT1	C10S3110	0.0022
SREBF1	C17S1007	0.0022
SREBF1	C17S1009	0.0007
SREBF1	C17S1024	0.0043
SREBF1	C17S1030	0.0007
SREBF1	C17S1043	0.0043
SREBF1	C17S1045	0.0036
SREBF1	C17S1046	0.0029
SREBF1	C17S1048	0.0014
SREBF1	C17S1055	0.0014
SREBF1	C17S1056	0.0007
VLDLR	C9S367	0.0007
VLDLR	C9S376	0.0029
VLDLR	C9S377	0.0014
VLDLR	C9S391	0.0007
VLDLR	C9S430	0.0007
VLDLR	C9S443	0.0014
VLDLR	C9S444	0.0014
VLDLR	C9S497	0.0007
VNN1	C6S5378	0.0057
VNN1	C6S5380	0.1707
VNN3	C6S5412	0.0007
VNN3	C6S5426	0.033
VNN3	C6S5439	0.0007
VNN3	C6S5441	0.0983
VNN3	C6S5446	0.0007
VNN3	C6S5448	0.0007
VNN3	C6S5449	0.0104
VWF	C12S181	0.0007
VWF	C12S211	0.0057

Table A.3: Genes and SNPs with effects on disease liability.

Gene	SNP	MAF
AKT3	C1S11396	0.0007
BCL2L11	C2S2286	0.0007
BCL2L11	C2S2288	0.0029
BCL2L11	C2S2307	0.0007
ELAVL4	C1S3181	0.0007
ELAVL4	C1S3182	0.0007
HSP90AA1	C14S3630	0.0007
HSP90AA1	C14S3695	0.0007
HSP90AA1	C14S3704	0.0036
HSP90AA1	C14S3706	0.2583
NRAS	C1S5748	0.0007
PIK3C2B	C1S9164	0.0014
PIK3C2B	C1S9165	0.0007
PIK3C2B	C1S9172	0.0043
PIK3C2B	C1S9173	0.0014
PIK3C2B	C1S9174	0.0007
PIK3C2B	C1S9189	0.0065
PIK3C2B	C1S9200	0.0007
PIK3C2B	C1S9222	0.0007
PIK3C2B	C1S9250	0.0014
PIK3C2B	C1S9266	0.0029
PIK3C2B	C1S9267	0.0022
PIK3C2B	C1S9306	0.0007
PIK3C2B	C1S9320	0.0007
PIK3C2B	C1S9333	0.0007
PIK3C2B	C1S9346	0.0007
PIK3C2B	C1S9373	0.0007
PIK3C2B	C1S9391	0.0007
PIK3C2B	C1S9423	0.0007
PIK3C2B	C1S9432	0.0108
PIK3C2B	C1S9445	0.0007
PIK3C2B	C1S9446	0.0007
PIK3C2B	C1S9449	0.0007
PIK3C2B	C1S9455	0.0029
PIK3C2B	C1S9457	0.0007
PIK3C3	C18S2475	0.0007
PIK3C3	C18S2492	0.0172
PIK3R3	C1S2919	0.0007
PRKCA	C17S4578	0.1664
PRKCA	C17S4581	0.0007
PRKCB1	C16S1894	0.0007
PTK2	C8S4825	0.0007
PTK2	C8S4839	0.0007
PTK2B	C8S886	0.0007
PTK2B	C8S900	0.0014
PTK2B	C8S909	0.0014
RRAS	C19S4929	0.0014
RRAS	C19S4997	0.0014
SHC1	C1S7061	0.0065
SOS2	C14S1381	0.0007
SOS2	C14S1382	0.0036

Appendix B

Publications derived from this thesis

- H. Brunel, A. Perera, A. Buil, M. Sabater-Lleal, J. C. Souto, Jordi Fontcuberta, M. Vallverdú, J. M. Soria and P. Caminal. Obtención de grupos de SNPs mediante un algoritmo de selección de características bajo criterio de información mutua. Aplicación al gen F7. *Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB)*, (2007).

- H. Brunel, A. Perera, A. Buil Demur, M. Sabater-Lleal, J.C. Souto Andrés, J. Fontcuberta Boj, M. Vallverdú, J.M. Soria and P. Caminal. SNP Sets Selection under Mutual Information Criterion, Application to F7/FVII Dataset. *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, (2008).

- H. Brunel, A. Perera, A. Buil Demur, M. Sabater-Lleal, J.C. Souto, J. Fontcuberta, M. Vallverdu, J.M. Soria and P. Caminal. Floating Feature Selection for multiloci association of quantitative traits in sib-pairs analysis. *8th IEEE International Conference on BioInformatics and BioEngineering (BIBE)*, (2008).

- H. Brunel, A. Perera, A. Buil, M. Sabater-Lleal, J.C. Souto, J. Fontcuberta, M. Vallverdú, J.M. Soria, P. Caminal. Algoritmo de selección de características flotante para asociación multiloci en estudios con familias. *Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB)*, (2008).

- H. Brunel, J.J. Gallardo-Chacon, M. Vallverdú, P. Caminal and A. Perera. Conservación de polimorfismos relacionados con enfermedades humanas. *Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB)*, (2009)

- H. Brunel, J.J. Gallardo-Chacón, A. Buil, M. Vallverdú, J.M. Soria, P. Caminal and A. Perera. MISS: a non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis. *Bioinformatics*, 26(15):1811-1818, (2010).

- H.Brunel, R. Massanet, A. Martinez, J.M. Soria and A. Perera. Metafenotipos: una herrramienta para estudios de asociación genética en contextos multifenotípicos.*Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB)*, (2011).

- H. Brunel, J.J. Gallardo-Chacon, M. Vallverdú, P. Caminal and A. Perera. Effect of genetic regions on the correlation between single point mutation variability and morbidity.*Computers in Biology and Medicine*, 43(5):594–599 (2013).

Bibliography

[1] G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*, 30(1):97–101, 2002.

[2] C. Adami. Information theory in molecular biology. *Physics of Life Reviews*, 1(1):3–22, 2004.

[3] E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous, and B. S. Pickard. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 6:55, 2005.

[4] E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous, and B. S. Pickard. Suspects: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, 22:773–774, 2006.

[5] V. Aishwarya, A. Grover, and P. C. Sharma. Eumicrosatdb: a database for microsatellites in the sequenced genomes of eukaryotes. *BMC Genomics*, 8:225, 2007.

[6] L. Almasy and J. Blangero. Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics*, 62(5):1198–1211, 1998.

[7] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[8] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[9] C I Amos. Robust variance-components approach for assessing genetic linkage in pedigrees. *American Journal of Human Genetics*, 54:535–543, 1994.

[10] N. Arshadi, B. C., and R. Kustra. Predictive modeling in case-control single-nucleotide polymorphism studies in the presence of population

stratification: a case study using genetic analysis workshop 16 problem 1 dataset. *BMC Proceedings*, 3(Suppl 7):S60, 2009.

[11] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–29, 2000.

[12] S. Asthana, M. Roytberg, J. Stamatoyannopoulos, and S. Sunyaev. Analysis of sequence conservation at nucleotide resolution. *PLoS Comput Biol*, 3:e254, 2007.

[13] K.I. Ataga, J. E. Brittain, P. Desai, R. May, S. Jones, J. Delaney, D. Strayhorn, A. Hinderliter, and N.S. Key. Association of coagulation activation with clinical complications in sickle cell disease. *PLoS ONE*, 7(1):e29786, 2012.

[14] G. Athanasiadis, A. Buil, J.C. Souto, M. Borrell, S. López, A. Martinez-Perez, M. Lathrop, J. Fontcuberta, L. Almasy, and J.M. Soria. A genome-wide association study of the protein c anticoagulant pathway. *PLoS ONE*, 2011.

[15] T. Attwood and D. Parry-Smith. *Introduction to Bioinformatics*. Prentice Hall, 1999.

[16] Y. S. Aulchenko, S. Ripke, A. Isaacs, and C. M. van Duijn. GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, 23(10):1294–1296, 2007.

[17] Yurii S. Aulchenko, Dirk-Jan de Koning, and Chris Haley. Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, 177(1):577–585, 2007.

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

[19] W. Bateson. The progress of genetic research. *Report of the Third International Conference on Hybridisation and Plant Breeding*, 1907.

[20] A. D. Baxevanis. *The Importance of Biological Databases in Biological Discovery*. John Wiley & Sons, Inc., 2002.

[21] T. M. Baye, H. He, L. Ding, B. G Kurowski, X. Zhang, and L. J. Martin. Population structure analysis using rare and common functional variants. *BMC Proceedings*, 5(Suppl 9):S8, 2011.

[22] M. A. Beaumont and D. J. Balding. Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, 13:969–980, 2004.

[23] K. G. Becker, K. C. Barnes, T. J. Bright, and A. Wang. The genetic association database. *The Nature Genetics*, 36:431 – 432, 2004.

[24] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[25] J. Bergman. The Functions of Introns: From Junk DNA to Designed DNA. *Perspectives on Science and Christian Faith*, 53(3):170–178, 2001.

[26] F. Besnier and O. Carlborg. A general and efficient method for estimating continuous ibd functions for use in genome scans for qtl. *BMC Bioinformatics*, 8(1):440, 2007.

[27] K. Bhasi, L. Zhang, D. Brazeau, A. Zhang, and M. Ramanathan. Information-theoretic identification of predictive SNPs and supervised visualization of genome-wide association studies. *Nucleic acids research*, 34(14):e101, 2006.

[28] P. Bhatti, D. M. Church, J. L. Rutter, J. P. Struewing, and A. J. Sigurdson. Candidate single nucleotide polymorphism selection using publicly available tools: A guide for epidemiologists. *American Journal of Epidemiology*, 164(8):794–804, 2006.

[29] D. T. Bishop and J. A. Williamson. The power of identity-by-state methods for linkage analysis. *American Journal of Human Genetics*, 46(2):254–265, 1990.

[30] R. Blekhman, O. Man, L. Herrmann, A. R. Boyko, A. Indap, C. Kosiol, C. D. Bustamante, K. M. Teshima1, and M. Przeworsk. Natural selection on genes that underlie human disease susceptibility. *Current Biology*, 18(12):883–889, 2008.

[31] T. Bo and I. Jonassen. New feature subset selection procedures for classification of expression profiles. *Genome biology*, 3(4), 2002.

[32] L. Boltzmann. Weitere studien uber das warmegleichgewicht unter gasmolekulenen. *Sitzungsberichte der Akademie der Wissenschaften*, 6:275–370, 1872.

[33] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[34] D. Brinza, J. He, and A. Zelikovsky. Combinatorial search methods for multi-SNP disease association. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1:5802–5805, 2006.

[35] A. J. Brookes. The essence of SNPs. *Gene*, 234(2):177–186, 1999.

[36] R. T. Brumfield, P. Beerli, D. A. Nickerson, and S. V. Edwards. The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, 18(5):249–256, 2003.

[37] A. V. Buchanan, K. M. Weiss, and S. M. Fullerton. Dissecting complex disease: the quest for the philosopher's stone? *International journal of epidemiology*, 35(3):562–571, 2006.

[38] A. Buil, D. A. Trégouët, J. C. Souto, N. Saut, M. Germain, M. Rotival, L. Tiret, F. Cambien, M. Lathrop, T. Zeller, M. C. Alessi, S. Rodriguez de Cordoba, T. Münzel, P. Wild, J. Fontcuberta, F. Gagnon, J. Emmerich, L. Almasy, S. Blankenberg, J. M. Soria, and P. E. Morange. C4bpb/c4bpa is a new susceptibility locus for venous thrombosis with unknown protein s-independent mechanism: results from genome-wide association and gene expression analyses followed by case-control studies. *Blood*, 115(23):4644–50, 2010.

[39] A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and P. Van Eerdewegh. Identifying snps predictive of phenotype using random forests. *Genetic Epidemiology*, 28(2):171–182, 2005.

[40] D. F. Burke, C. L. Worth, E. M. Priego, T. Cheng, L. J. Smink, Todd J. A, and T. L. Blundell. Genome bioinformatic analysis of nonsynonymous SNPs. *BMC Bioinformatics*, 8:301, 2007.

[41] M. C. Byng, J. C. Whittaker, A. P. Cuthbert, C. G. Mathew, and C. M. Lewis. Snp subset selection for genetic association studies. *Annals of Human Genetics*, 67:543–556, 2003.

[42] T. Calinski, Z. Kaczmarek, P. Krajewski, C. Frova, and M. Sari-Gorla. A multivariate approach to the problem of qtl localization. *Heredity*, 84(3):303–310, 2000.

[43] R. M. Cantor, K. Lange, and J. S. Sinsheimer. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. *The American Journal of Human Genetics*, 86(1):6–22, 2010.

[44] E. Capriotti, R. Calabrese, and R. Casadio. Predicting the insurgence of human genetic diseases associated to single point protein mutations

with support vector machines and evolutionary information. *Bioinformatics*, 22:2729–2734, 2006.

[45] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics*, 74(1):106–120, 2004.

[46] E. Caspari. Pleiotropic gene action. *Evolution*, 6:1:18, 1952.

[47] R. Cavill, A. Kamburov, J. K. Ellis, T. J. Athersuch, M. S. C. Blagrove, R. Herwig, T. M. D Ebbels, and H. C. Keun. Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells. *PLoS Comput Biol*, 7(3):e1001113, 2011.

[48] P. Chanda, A. Zhang, D. Brazeau, L. Sucheston, J. L Freudenheim, C. Ambrosone, and M. Ramanathan. Information-theoretic metrics for visualizing gene-environment interactions. *American Journal of Human Genetics*, 81(5):939–963, 2007.

[49] A. Chao and T. J. Shen. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, 10(4):429–443, 2003.

[50] K. Chen and N. Rajewsky. Deep conservation of microrna-target relationships and 3'utr motifs in vertebrates, flies, and nematodes. *Cold Spring Harbor symposia on quantitative biology*, 71:149–156, 2007.

[51] W. M. Chen and G. R. Abecasis. Family-based association tests for genomewide association scans. *American journal of human genetics*, 81(5):913–26, 2007.

[52] C. Y. Cheng, K. E. Lee, P. Duggal, E. L. Moore, A. F. Wilson, R. Klein, J. E. Bailey-Wilson, and B. E. K. Klein. Genome-wide linkage analysis of multiple metabolic factors: evidence of genetic heterogeneity. *Obesity Silver Spring Md*, 18(1):146–152, 2010.

[53] J. M. Cheverud. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, 87(1):52–58, 2001.

[54] R. Clausius. Ueber die wärmeleitung gasförmiger körper. *Annalen der Physik*, 125(1):353–400, 1865.

[55] W. G. Cochran. Some methods for strengthening the common chi-squared tests. *Biometrics*, 10(4):417–451, 1954.

[56] Harvey Jay Cohen, Tamara Harris, and Carl F Pieper.

[57] P. C. Conilione and D. Wang. A comparative study on feature selection for e. coli promoter recognition a comparative study on feature selection for e. coli promoter recognition. *International Journal of Information Technology*, 11:54–66, 2005.

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

[59] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[60] G. M. Cooper, E. A. Stone, G. Asimenos, E. D. Green, S. Batzoglou, and A. Sidow. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, 15(7):901–913, 2005.

[61] H. J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, 2002.

[62] H. J. Cordell. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, 2009.

[63] A. Cornish-Bowden. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Research*, 13(9):3021–3030, 1985.

[64] F. H. C. Crick. On Protein Synthesis. *The Symposia of the Society for Experimental Biology*, 12:138–163, 1958.

[65] J. F. Crow. Hardy, weinberg and language impediments. *Genetics*, 152(3):821–825, 1999.

[66] J. Y. Dai, I. Ruczinski, M. LeBlanc, and C. Kooperberg. Comparison of haplotype-based and tree-based snp imputation in association studies. *Genetic Epidemiology*, 30(8):690–702, 2006.

[67] M. Dash and H. Liu. Hybrid search of feature subsets. *Lecture Notes in Computer Science*, 1531:238–249, 1998.

[68] Z. Dawy, B. Goebel, J. Hagenauer, C. Andreoli, T. Meitinger, and J. C. Mueller. Gene mapping and marker clustering using shannon's mutual information. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 3(1):47–56, 2006.

[69] Z. Dawy, J. Hagenauer, P. Hanus, and J. C. Mueller. Mutual information based distance measures for classification and content recognition with applications to genetics. 2005.

[70] H. de Vries. *Intracellulare pangenesis.* G. Fischer, 1889.

[71] H. de Vries. *Die Mutationstheorie. Versuche und Beobachtungen über die Entstehung der Arten im Pflanzenreich*, volume II. Veit, Leipzig, 1901.

[72] R. Dersimonian and N. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, 1986.

[73] P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. Mining the gene expression matrix: inferring gene relationships from large scale gene expression data. *Proceedings of the second international workshop on Information processing in cell and tissues*, 1998.

[74] P. B. Dobrokhotov, C. Goutte, A. L. Veuthey, and E. Gaussier. Combining NLP and probabilistic categorisation for document and term selection for Swiss-Prot medical annotation. *Bioinformatics*, 19 Suppl 1:91–94, 2003.

[75] F Dudbridge. Pedigree disequilibrium tests for multilocus haplotypes. *Genetic Epidemiology*, 25:115–121, 2003.

[76] F. Dudbridge and A. Gusnanto. Estimation of significance thresholds for genomewide association scans. *Genetic epidemiology*, 32(3):227–234, 2008.

[77] F. Fabris. Shannon information theory and molecular biology. *Journal of Interdisciplinary Mathematics*, 12:41–87, 2009.

[78] F. Fabris, A. Sgarro, and A. Tossi. Splitting the blosum score into numbers of biological significance. *EURASIP J. Bioinformatics Syst. Biol.*, Special Issue, 2007.

[79] I. Feenstra, J. Fang, D. A. Koolen, A. Siezen, C. Evans, R. M. Winter, M. M. Lees, M. Riegel, B. B. A. de Vries, C. M. A. Van Ravenswaaij, and A. Schinzel. European cytogeneticists association register of unbalanced chromosome aberrations (ecaruca); an online database for rare chromosome abnormalities. *Eur J Med Genet*, 49(4):279–91, 2006.

[80] D. Feng, G. H. Tofler, M. G. Larson, C. J. O'Donnell, I. Lipinska, C. Schmitz, P. A. Sutherland, M. T. Johnstone, J. E. Muller, R. B. D'Agostino, D. Levy, and K. Lindpaintner. Factor VII gene polymorphism, factor VII levels, and prevalent cardiovascular disease: the framingham heart study. *Arteriosclerosis, thrombosis, and vascular biology*, 20(2):593–600, 2000.

[81] L. Feuk, A. R. Carson, and S. W. Scherer. Structural variation in the human genome. *Nature Review Genetics*, 7:85–97, 2006.

[82] W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, and F. Molemans et al. et al. Complete nucleotide-sequence of bacteriophage MS2-RNA - primary and secondary structure of replicase gene. *Nature*, 260(5551):500–507, 1976.

[83] H. V. Firth, S. M. Richards, A. P. Bevan, S. C., M. Corpas, D. Rajan, S. Van Vooren, Y. Moreau, R. M. Pettett, and N. P. Carter. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American journal of human genetics*, 84(4):524–533, 2009.

[84] R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

[85] G. Forman. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3:1289–1305, 2003.

[86] D. Francois, V. Wertz, and M. Verleyson. The permutation test for feature selection by mutual information. *European Symposium on Artificial Neural Networks*, 2006.

[87] I. R. Franklin and R. C. Lewontin. Is the gene the unit of selection? *Genetics*, 65(4):707–734, 1970.

[88] K. A. Frazer, L. Elnitski, D. M. Church, I. Dubchak, and R. C. Hardison. Cross-species sequence comparisons a review of methods and available resources. *Genome Research*, 13(1):1–12, 2009.

[89] K. S. Fu. *Sequential methods in pattern recognition and machine learning (Mathematics in science and engineering)*. Academic Press, 1st edition, 1968.

[90] J. Garnier, D. J. Osguthorpe, and B. Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology*, 120(1):97–120, 1978.

[91] L. L. Gatlin. The information content of dna. *Journal of Theoretical Biology*, 10(2):281–300, 1966.

[92] B. Goebel. An approximation to the distribution of finite sample size mutual information estimates. *Proceedings of the IEEE International Conference on Communications*, 2:1102–1106, 2005.

[93] K. I. Goh, M. E. Cusick, D. Valle, and et al. The human disease network. *Proc. Natl. Acad. Sci.*, 104:8685–90, 2007.

[94] Golden Helix. `http://www.goldenhelix.com`, 2011. [Online; accessed July-2012].

[95] D. B. Goldstein. Common genetic variation and human traits. *The New England Journal of Medicine*, 360(17):1696–8, 2009.

[96] J. R. González, L. Armengol, X. Solé, E. Guinó, J. M. Mercader, X. Estivill, and V. Moreno. Snpassoc: an r package to perform whole genome association studies. *Bioinformatics*, 23(5):654–655, 2007.

[97] C. M. T. Greenwood, J. Rangrej, and L. Sun. Optimal selection of markers for validation or replication from genome-wide association studies. *Genetic Epidemiology*, 31(5):396–407, 2007.

[98] I. Grosse, H. Herzel, S. V. Buldyrev, and H. E. Stanley. Species independence of mutual information in coding and noncoding DNA. *Physical Review E*, 61:5624–5629, 2000.

[99] NIH/CEPH Collaborative Mapping Group. A comprehensive genetic linkage map of the human genome. *Science*, 258(5079):148–162, 1992.

[100] C. M. Gruner. Mutual information calculation using empirical classification. *Neurocomputing*, 44-46:1083–1088, 2002.

[101] C. A. Hackett, R. C. Meyer, and W. T. B. Thomas. Multi-trait QTL mapping in barley using multivariate regression. *Genetical Research*, 77:95–106, 2001.

[102] J. Hagenauer, Z. Dawy, B. Göbel, P. Hanus, and J. Mueller. Genomic analysis using methods from information theory. *IEEE Information Theory Workshop*, 2004.

[103] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17:107–145, 2001.

[104] B. V. Halldorsson, V. Bafna, R. Lippert, R. Schwartz, F. M. De La Vega, A. G. Clark, and S. Istrail. Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. *Genome research*, 14(8):1633–1640, 2004.

[105] B. V. Halldorsson, S. Istrail, and F. M. De La Vega. Optimal selection of SNP markers for disease association studies. *Human heredity*, 58(3-4):190–202, 2004.

[106] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A McKusick. The Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 30(1):52–55, 2002.

[107] R. C. Hardison. Comparative genomics. *PLoS Biology*, 1(2):e58, 2003.

[108] R. V. L. Hartley. Transmission of information. *Bell Syst. Tech. Journal*, 7:535–563, 1928.

[109] J. K. Haseman and R. C. Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, 2(1):3–19, 1972.

[110] S. Hattirat, C. Ngamphiw, A. Assawamakin, J. Chan, and S. Tongsima. Catalog of genetic variations (snps and cnvs) and analysis tools for thai genetic studies. *Computational Systems-Biology and Bioinformatics*, 115:130–140, 2010.

[111] J. Hausser, K. Strimmer, and X. Shen. Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning*, 10:1469–1484, 2009.

[112] H. He, X. Zhang, L. Ding, T. M. Baye, B. G. Kurowski, and L. J. Martin. Effect of population stratification analysis on false-positive rates for common and rare variants. *BMC Proceedings*, 5(Suppl 9):S116, 2011.

[113] J. He. Informative SNP selection methods based on SNP prediction. *NanoBioscience, IEEE Transactions on*, 6(1):60–67, 2007.

[114] J. He and A. Zelikovsky. MLR-tagging: informative SNP selection for unphased genotypes based on multiple linear regression. *Bioinformatics*, 22(20):2558–2561, 2006.

[115] C. M. Hearne, S. Ghosh, and J. A. Todd. Microsatellites for linkage analysis of genetic traits. *Trends in Genetics*, 8(8):288–294, 1992.

[116] S. C. Heath. Markov chain monte carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics*, 61(3):748–760, 1997.

[117] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[118] M. Hirakawa, T. Tanaka, Y. Hashimoto, M. Kuroda, T. Takagi, and Y. Nakamura. JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Research*, 30(1):158–62, 2002.

[119] H. Hiratani, D. W. Bowden, S. Ikegami, S. Shirasawa, A. Shimizu, Y. Iwatani, and T. Akamizu. Multiple SNPs in intron 7 of thyrotropin receptor are associated with Graves´ disease. *J Clin Endocrinol Metab*, 90(5):2898–903, 2005.

[120] J. N. Hirschhorn. Genomewide Association Studies — Illuminating Biologic Pathways. *N Engl J Med*, 360(17):1699–1701, 2009.

[121] A. E. Hirsh and H. B. Fraser. Protein dispensability and rate of evolution. *Nature*, 411:1046–1049, 2001.

[122] B. Hoogendoorn, S. L. Coleman, C. A. Guy, S. K. Smith, M. C. O'Donovan, and P. R. Buckland. Functional analysis of polymorphisms in the promoter regions of genes on 22q11. *Human mutation*, 24(1):35–42, 2004.

[123] L.M. Houlihan, G. Davies, A. Tenesa, S.E. Harris, M. Luciano, A.J. Gow, K.A. McGhee, D.C. Liewald, D.J. Porteous, J.M. Starr, G.D. Lowe, P.M. Visscher, and I.J. Deary. *American Journal of Human Genetics*, 2010.

[124] D. Hoyer, B. Pompe, K. H. Chon, H. Hardraht, C. Wicher, and U. Zwiener. Mutual information function assesses autonomic information flow of heart rate dynamics at different time scales. *Biomedical Engineering, IEEE Transactions on*, 52(4):584–592, 2005.

[125] G. Huang and P. Jeavons. A geometrical model for the SNP motif identification problem. *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering*, 2007.

[126] H. Huang, E. E. Winter, H. Wang, K. G. Weinstock, H. Xing, L. Goodstadt, P. D. Stenson, D. N. Cooper, D. Smith, M. Albà, C. P. Ponting, and K. Fechtel. Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. *Genome Biology*, 5(7):R47, 2001.

[127] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp. The Ensembl genome database project. *Nucleic Acids Research*, 30(1):38–40, 2002.

[128] R. Hubley, E. Zitzler, and J. Roach. Evolutionary algorithms for the selection of single nucleotide polymorphisms. *BMC Bioinformatics*, 4(1), 2003.

[129] J. E. Hutz, A. T. Kraja, H. L. McLeod, and M. A. Province. Candid: a flexible method for prioritizing candidate genes for complex human traits. *Genetic Epidemiology*, 32:779–790, 2008.

[130] S. Hwang, Y. S. Lee, S. C. Kim, and D. Lee. SNP@Ethnos: a database of ethnically variant single-nucleotide polymorphisms. *Nucleic Acids Research*, 35:711–715, 2007.

[131] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13:411–430, 2000.

[132] A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Q., S. W. Scherer, and C. Lee. Detection of large-scale variation in the human genome. *Nature Genetics*, 36(9):949–951, 2004.

[133] Yusuke Inoue, L.L. Peters, SunHee Yim, Junko Inoue, and FrankJ. Gonzalez. Role of hepatocyte nuclear factor 4a in control of blood coagulation factor gene expression. *Journal of Molecular Medicine*, 84(4):334–344, 2006.

[134] I. Inza, B. Sierra, R. Blanco, and P. Larrañaga. Gene selection by sequential search wrapper approaches in microarray cancer class prediction. *J. Intell. Fuzzy Syst.*, 12:25–33, 2002.

[135] J. P. A. Ioannidis, P. Boffetta, J. Little, T. R. O'Brien, A. G. Uitterlinden, P. Vineis, D. J. Balding, A. Chokkalingam, S. M. Dolan, W. D. Flanders, J. P. Higgins, M. I. McCarthy, D. H. McDermott, G. P. Page, T. R. Rebbeck, D. Seminara, and M. J. Khoury. Assessment of cumulative evidence on genetic associations: interim guidelines. *International Journal of Genetic Epidemiology*, 37:120–132, 2007.

[136] A. Jain and D. Zongker. Feature selection: evaluation, application, and small sample performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(2):153–158, 1997.

[137] E. T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review Online Archive*, 106(4):620–630, 1957.

[138] C. Jiang and Z. B. Zeng. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, 140(3):1111–1127, 1995.

[139] T. Jirapech-Umpai and S. Aitken. Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. *BMC bioinformatics*, 6:148, 2005.

[140] Andrew D. Johnson. Single-nucleotide polymorphism bioinformatics: a comprehensive review of resources. *Circ. Cardiovasc. Genet.*, 2(5):530–536, 2009.

[141] Julie Josse and François Husson. Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics and Data Analysis*, 56(6):1869–1879, 2012.

[142] L. Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.

[143] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[144] J. N. Kapur. *Measures of Information and Their Applications*. John Wiley & Sons, 1994.

[145] J. Kasturi, R. Acharya, and M. Ramanathan. An information theoretic approach for analyzing temporal patterns of gene expression. *Bioinformatics*, 19(4):449–458, 2003.

[146] S. Keles, M. van der Laan, and M. B. Eisen. Identification of regulatory elements using a feature selection method. *Bioinformatics*, 18(9):1167–1175, 2002.

[147] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.

[148] P. M. Kim, H. Y. Lam, A. E. Urban, J. O. Korbel, J. Affourtit, F. Grubert, X. Chen, S. Weissman, M. Snyder, and M. B. Gerstein. Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome research*, 18(12):1865–1874, 2008.

[149] K. Kira and L. A. Rendell. A practical approach to feature selection. *Proceedings of the ninth international workshop on Machine learning*, 1992.

[150] L. Klei, D. Luca, B. Devlin, and K. Roeder. Pleiotropy and Principal Components of Heritability Combine to Increase Power for Association Analysis. *Genetic Epidemiology*, 32:9–19, 2008.

[151] R. J. Klein, C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C.Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, and J. Hoh. Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, 2005.

[152] G. T. Klus, A. Song, A. Schick, M. Wahde, and Z. Szallasi. Mutual information analysis as a tool to assess the role of aneuploidy in the generation of cancer-associated differential gene expression patterns. *Pacific Symposium on Biocomputing*, (6):42–51, 2001.

[153] S. A. Knott and C. S. Haley. Multitrait least squares for quantitative trait loci detection. *Genetics*, 156(2):899–911, 2000.

[154] A. B. Korol, Y. I. Ronin, and V. M. Kirzhner. Interval Mapping of Quantitative Trait Loci Employing Correlated Trait Complexes. *Genetics*, 140(3):1137–1147, 1995.

[155] A. B. Korol, Y. I. Ronin, E. Nevo, and P. M. Hayes. Multi-interval mapping of correlated trait complexes. *Heredity*, 80(3):273–284, 1998.

[156] A. Krishnamachari, V. Mandal, and Karmeshu. Study of dna binding sites using the Rényi parametric entropy measure. *Journal ofTheoretical Biology*, 227(3):429–436, 2004.

[157] L. Kruglyak and E. S. Lander. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics*, 57(2):439–454, 1995.

[158] L. Kruglyak and D. A. Nickerson. Variation is the spice of life. *Nature genetics*, 27(3):234–236, 2001.

[159] M. Kudo and J. Sklansky. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 33(1):25–41, 2000.

[160] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86, 1951.

[161] S. Kumar, J. T. Dudley, A. Filipski, and L. Liu. Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. *Trends in genetics*, 27(9):377–386, 2011.

[162] N. M. Laird and C. Lange. Family-based designs in the age of large-scale gene-association studies. *Nature reviews. Genetics*, 7(5):385–394, 2006.

[163] C. Lange, K. van Steen, T. A., H. L., D. L. DeMeo, B. Raby, A. Murphy, E. K. Silverman, A. MacGregor, S. T. Weiss, and N. M. Laird. A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. *Statistical Applications in Genetics and Molecular Biology*, 3(1):17, 2004.

[164] R. W. Lawrence, D. M. Evans, and L. R. Cardon. Prospects and pitfalls in genome association studies. *Philosophical Transactions of the Royal Society B*, 360:1589–1595, 2005.

[165] J. H. Lee and G. H. Gonzalez. Towards integrative gene prioritization in alzheimerś disease. *Pacific Symposium on Biocomputing*, 2011.

[166] J. W. Lee, J. Bok Lee, M. Park, and S. H. Song. An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, 48(4):869–885, 2005.

[167] Jerry B Lefkowitz. Coagulation pathway and physiology. *An Algorithmic Approach to Hemostasis Testing*, pages 3–12, 2008.

[168] Marcel Levi, Tom van der Poll, and Harry R. Büller. Bidirectional relation between inflammation and coagulation. *Circulation*, 109(22):2698–2704, 2004.

[169] S. Levy, G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, B. P. Walenz, N. Axelrod, J. Huang, E. F. Kirkness, G. Denisov, Y. Lin, J. R. MacDonald, A. Wing, C. W. Pang, M. Shago, T. B. Stockwell, A. Tsiamouri, V. Bafna, V. Bansal, S. A. Kravitz, D. A. Busam, K. Y. Beeson, T. C. McIntosh, K. A. Remington, J. F. Abril, J. Gill, J. Borman, Y. H. Rogers, M. E. Frazier, S. W. Scherer, R. L. Strausberg, and J. C. Venter. The diploid genome sequence of an individual human. *PLoS biology*, 5(10):e254, 2007.

[170] C. M. Lewis. Genetic association studies: Design, analysis and interpretation. *Briefings in Bioinformatics*, 3(2):146–153, 2002.

[171] R. C. Lewontin. The interaction of selection and linkage. i. general considerations; heterotic models. *Genetics*, 49(1):49–67, 1964.

[172] R. C. Lewontin and K. Kojima. The evolutionary dynamics of complex polymorphisms. *Evolution*, 14(4):458–472, 1960.

[173] M. Li, M. Boehnke, and G. R. Abecasis. Joint modeling of linkage and association: Identifying snps responsible for a linkage signal. *American Journal of Human Genetics*, 76(6):934–949, 2005.

[174] M. Li, M. P. Reilly, D. J. Rader, and L. S. Wang. Correcting population stratification in genetic association studies using a phylogenetic approach. *Bioinformatics*, 26(6):798–806, 2010.

[175] Q. Li and K. Yu. Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genetic Epidemiology*, 32(3):215–226, 2008.

[176] Y. Li and J. Patra. Integration of multiple data sources to prioritize candidate genes using discounted rating system. *BMC Bioinformatics*, 11(Suppl 1), 2010.

[177] D. J. Liu and S. M. Leal. Estimating genetic effects and quantifying missing heritability explained by identified rare-variant associations. *The American Journal of Human Genetics*, 91(4):585 – 596, 2012.

[178] H. Liu and R. Setiono. Incremental feature selection. *Applied Intelligence*, 9:217–230, 1998.

[179] N. LIU, H. Zhao, A. Parki, N. A. Limbdi, and D. B. Allison. Controlling population structure in human genetic association studies with samples of unrelated individuals. *Stat. Interface*, 4(3):317–326, 2011.

[180] I. Lobo. Pleiotropy: One gene can affect multiple traits. *Nature Education*, 1(1), 2008.

[181] D. M. Loewenstern, H. M. Berman, and H. Hirsh. Automated classification of dna structure from sequence information. Technical report, Rutgers University, Dept. of Computer Science, 1997.

[182] N Long, D Gianola, G J M Rosa, K A Weigel, and S Avendaño. Marker-assisted assessment of genotype by environment interaction: a case study of single nucleotide polymorphism-mortality association in broilers in two hygiene environments. *Journal of animal science*, 86(12):3358–66, 2008.

[183] C. G. Loots, R. M. Locksley, C. M. Blankespoor, Z. E. Wang, W. Miller, E. M. Rubin, and K. A. Frazer. Identification os a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, 288:136–140, 2000.

[184] N. Lopez-Bigas and C. A. Ouzounis. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Research*, 32(10):3108–3114, 2004.

[185] S. C. Lovell, X. Li, N. R. Weerasinghe, and K. E. Hentges. Correlation of microsynteny conservation and disease gene distribution in mammalian genomes. *BMC Genomics*, 10:521, 2009.

[186] X. Lu, C. Zhai, V. Gopalakrishnan, and B. G. Buchanan. Automatic annotation of protein motif function with Gene Ontology terms. *BMC Bioinformatics*, 5(1), 2004.

[187] K. L. Lunetta. Genetic association studies. *Circulation*, 118(1):96–101, 2008.

[188] M. Lynch and B. Walsh. *Genetic Analysis of quantitative Traits.* Sinauer Associates, 1 edition, 1998.

[189] J. Mackay and G. A. Mensah. The atlas of heart disease and stroke. World Heart Organization, 2004.

[190] P. Maison, C. D. Byrne, C. N. Hales, N. E. Day, and N. J. Wareham. Do different dimensions of the metabolic syndrome change together

over time? evidence supporting obesity as the central feature. *Diabetes Care*, 24(10):1758–1763, 2001.

[191] B. Mangin, P. Thoquet, and N. Grimsley. Pleiotropic qtl analysis. *Biometrics*, 54:88–99, 1998.

[192] G. Marchetti, P. Patracchini, M. Papacchini, M. Ferrati, and F. Bernardi. A polymorphism in the 5´ region of coagulation factor VII gene (f7) caused by an inserted decanucleotide. *Human Genetics*, 90(5):575–576, 1993.

[193] E. H. Margulies, M. Blanchette, Comparative Sequencing Program, D. Haussler, and E. D. Green. Identification and characterization of multi-species conserved sequences. *Genome Research*, 13(12):2507–2518, 2003.

[194] Eden R. Martin, Stephanie A. Monks, Liling L. Warren, and Norman L. Kaplan. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am.J.Hum.Genet*, 67:146–154, 2000.

[195] R. Massanet, P. Caminal, and A. Perera. Use of gene ontology semantic information in protein interaction data visualization. *Proceedings of the 8th IEEE International Conference on Bioinformatics and Bioengineering*, 2008.

[196] R. Massanet, J. J. Gallardo-Chacón, P. Caminal, and A. Perera. Search of phenotype related candidate genes using gene ontology-based semantic similarity and protein interaction information: Application to brugada syndrome. *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2009.

[197] R. A Mathias, Y. Kim, H. Sung, L. R Yanek, V. J. Mantese, J. E. Hererra-Galeano, I. Ruczinski, Alexander F. Wilson, N. Faraday, L. C. Becker, and et al. A combined genome-wide linkage and association approach to find susceptibility loci for platelet function phenotypes in european american and african american families with coronary artery disease. *BMC medical genomics*, 3:22, 2010.

[198] J. Maynou, J. J. Gallardo-Chacón, M. Vallverdú, P. Caminal, and A. Perera. Computational detection of transcription factor binding sites through differential Rényi entropy. *IEEE Transactions on Information Theory*, 56(2):734–741, 2010.

[199] M. I. Mccarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics*, 9(5):356–369, 2008.

[200] J. L. McCauley, S. J. Kenealy, E. H. Margulies, N. Schnetz-Boutaud, S. G. Gregory, S. L. Hauser, J. R. Oksenberg, M. A. Pericak-Vance, J. L. Haines, and D. P. Mortlock. SNPs in multi-species conserved sequences (MCS) as useful markers in association studies: a practical approach. *BMC Genomics*, 8:266, 2007.

[201] J. McClellan and M. C. King. Genetic Heterogeneity in Human Disease. *Cell*, 141(2):210–217, 2010.

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

[203] H. Mei, M. L. Cuccaro, and E. R. Martin. Multifactor dimensionality reduction-phenomics: a novel method to capture genetic heterogeneity with use of phenotypic variables. *Am J Hum Genet*, 81(6):1251–1261, 2007.

[204] G. Mendel. Versuche über pflanzenhybriden. *Verhandlungen des Naturforschenden Vereines in Brünn*, 4:3–47, 1866.

[205] D. M. Milewicz and C. E. Seidman. Genetics of cardiovascular disease. *Circulation*, 102(20):IV103–11, 2000.

[206] D. J. Miller, Y. Zhang, G. Yu, Y. Liu, L. Chen, C. D. Langefeld, D. Herrington, and Y. Wang. An algorithm for learning maximum entropy probability models of disease risk that efficiently searches and sparingly encodes multilocus genomic interactions. *Bioinformatics*, 25(19):2478–2485, 2009.

[207] G. A. Miller. Note on the bias of information estimates. *Information Theory in Psychology: Problems and Methods*, 1955.

[208] P. I. Missirlis, C. L. Mead, S. L. Butland, B. F. Ouellette, R. S. Devon, B. R. Leavitt, and R. A. Holt. Satellog: a database for the identification and prioritization of satellite repeats in disease association studies. *BMC bioinformatics*, 6, 2005.

[209] T. M. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2):203–226, 1982.

[210] L. C. Molina, L. Belanche, and A. Nebot. Feature selection algorithms: a survey and experimental evaluation. *Proceedings of the IEEE International Conference on Data Mining*, 2002.

[211] R. Moonesinghe, A. Yesupriya, M. H. Chang, N. F. Dowling, M. J. Khoury, and A. J. Scott. A hardy-weinberg equilibrium test for analyzing population genetic surveys with complex sample designs. *American Journal of Epidemiology*, 171(8):932–941, 2010.

[212] S. Mooney. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Briefings in bioinformatics*, 6(1):44–56, 2005.

[213] J. Moore and B. White. Tuning relieff for genome-wide genetic analysis. *Lecture notes in computer science*, 4447:166–175, 2007.

[214] J. H. Moore, F. W. Asselbergs, and S. M. Williams. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–455, 2010.

[215] J. H. Moore and S. M. Williams. Epistasis and its implications for personal genetics. *American journal of human genetics*, 85(3):309–320, 2009.

[216] Pierre-Emmanuel Morange, Tiphaine Oudot-Mellakh, William Cohen, Marine Germain, Noémie Saut, Guillemette Antoni, Marie-Christine Alessi, Marion Bertrand, Anne-Marie Dupuy, Luc Letenneur, Mark Lathrop, Lorna M. Lopez, Jean-Charles Lambert, Joseph Emmerich, Philippe Amouyel, and David-Alexandre Trégouët. Kng1 ile581thr and susceptibility to venous thrombosis. *Blood*, 117(13):3692–3694, 2011.

[217] S. Mottagui-Tabar, M. A. Faghili, Y. Mizuno, P. G. Engstrom, B. Lenhard, W. W. Wasserman, and C. Wahlestedt. Identification of functional SNPs in the 5-prime flanking sequences of human genes. *BMC Genomics*, 6:18, 2005.

[218] P. M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *Computers, IEEE Transactions on*, C-26(9):917–922, 1977.

[219] National Center for Biotechnology Information. `http://eutils.ncbi.nlm.nih.gov/`, 2009. [Online; accessed November-2009].

[220] National Center for Biotechnology Information. `http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp`, 2009. [Online; accessed November-2009].

[221] NCBI Genome Browser. `http://www.ncbi.nlm.nih.gov/sites/genome`, 2011. [Online; accessed July-2011].

[222] I. Nemenman, F. Shafee, and W. Bialek. Entropy and inference, revisited. *Advances in Neural Information Processing Systems*, 14, 2002.

[223] K. Neveling, R. W. J. Collin, C. Gilissen, R. A. C. van Huet, L. Visser, M. P. Kwint, S. J. Gijsen, M. N. Zonneveld, N. Wieskamp, J. de Ligt, A. M. Siemiatkowska, L. H. Hoefsloot, M. F. Buckley, U. Kellner, K. E. Branham, A. I. den Hollander, A. Hoischen, C. Hoyng, B. J. Klevering, L. I. van den Born, J. A. Veltman, F. P. M. Cremers, and H. Scheffer. Next-generation genetic testing for retinitis pigmentosa. *Hum Mutat*, 33(6):963–72, 2012.

[224] M. Ng and L. Chan. Informative gene discovery for cancer classification from microarray expression data. *Proceedings of 2005 IEEE Workshop on the Machine Learning for Signal Processing*, 2005.

[225] M. K. Ng, M. J. Li, S. I. Ao, P. C. Sham, Y. M. Cheung, and J. Z. Huang. Clustering of SNP data with application to genomics. *Proceedings of the Sixth IEEE International Conference on Data Mining Workshops*, 2006.

[226] D. R. Nyholt. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics*, 74(4):765–769, 2004.

[227] C. H. Ooi and P. Tan. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*, 19(1):37–44, 2003.

[228] M. V. Osier, K. H. Cheung, J. R. Kidd, A. J. Pakstis, P. L. Miller, and K. K. Kidd. ALFRED: an allele frequency database for diverse populations and DNA polymorphisms - an update. *Nucleic Acids Research*, 29:317–319, 2001.

[229] C. Oslakovic, E. Norstrøm, and B. Dahlbäck. Reevaluation of the role of hdl in the anticoagulant activated protein c system in humans. *The Journal of Clinical Investigation*, 120(5):1396–1399, 2010.

[230] Jurg Ott, Yoichiro Kamatani, and Mark Lathrop. Family-based designs for genome-wide association studies. *Nature reviews. Genetics*, 12(7):465–474, 2011.

[231] Q. Peng, J. Zhao, and F. Xue. PCA-based bootstrap confidence interval tests for gene-disease association involving multiple SNPs. *BMC genetics*, 11(1):6+, 2010.

[232] L. S. Penrose. The general purpose sib-pair linkage method. *Annals of Eugenics*, 18(2):120–124, 1953.

[233] A. Perera, M. Vallverdu, F. Claria, J. M. Soria, and P. Caminal. Dna binding site characterization by means of r&eacute;nyi entropy measures on nucleotide transitions. *IEEE Trans Nanobioscience*, 7(2):133–41, 2008.

[234] T. M. Phuong, Z. Lin, and R. B. Altman. Choosing SNPs using feature selection. *Proceedings of the Computational Systems Bioinformatics Conference*, 2005.

[235] A. R. Pico, I. V. Smirnov, J. S. Chang, R. F. Yeh, J. L. Wiemels, J. K. Wiencke, T. T., B. R. Conklin, and M. Wrensch. SNPLogic: an interactive single nucleotide polymorphism selection, annotation, and prioritization system. *Nucleic Acids Research*, 37(Suppl 1):D803–809, 2009.

[236] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever. Mutual-information-based registration of medical images: a survey. *Medical Imaging, IEEE Transactions on*, 22(8):986–1004, 2003.

[237] E. S. Pollak, H. L. Hung, W. Godin, G. C. Overton, and K. A. High. Functional characterization of the human factor VII 5´-flanking region. *Journal of Biological chemistry*, 271:1738–1747, 1996.

[238] R. Pong-Wong, A. W. George, J. A. Woolliams, and C. S. Haley. A simple and rapid method for calculating identity-by-descent matrices using multiple markers. *Genetics Selection Evolution*, 33(5):453–471, 2001.

[239] M. D. Prasad, M. Muthulakshmi, K. P. Arunkumar, M. Madhu, V. B. Sreenu, V. Pavithra, B. Bose, H. A. Nagarajaram, K. Mita, T. Shimada, and J. Nagaraju. Silksatdb: a microsatellite database of the silkworm, bombyx mori. *Nucleic Acids Research*, 33(Suppl 1):D403–D406, 2005.

[240] A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson. New approaches to population stratification in genome-wide association studies. *Nature reviews. Genetics*, 11(7):459–463, 2010.

[241] J. C. Principe, D. Xu, Q. Zhao, and J. W. Fisher III. Learning from examples with information theoretic criteria. *Journal of VLSI Systems, Kluwer*, 26:61–77, 1999.

[242] I. Priness, O. Maimon, and I. Ben-Gal. Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics*, 8(111), 2007.

[243] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959, 2000.

[244] S. Purcell, B. Neale, K. Toddbrown, L. Thomas, M. Ferreira, D. Bender, J. Maller, P. Sklar, P. Debakker, and M. Daly. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[245] P. Qiu, A. J. Gentles, and S. K. Plevritis. Fast calculation of pairwise mutual information for gene regulatory network reconstruction. *Comput. Methods Prog. Biomed.*, 94:177–180, 2009.

[246] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, 2005.

[247] D. Rabinowitz and N. Lairdb. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Human Heredity*, 50:211–223, 2000.

[248] V. Ramensky, P. Bork, and S. Sunyaev. Human non-synonymous SNPs: server and survey. *Nucleic acids research*, 30(17):3894–3900, 2002.

[249] T. R. Rebbeck, M. Spitz, and X. Wu. Assessing the function of genetic variants in candidate gene association studies. *Nature Review Genetics*, 5:589–597, 2004.

[250] D. E. Reich, S. F. Schaffner, M. J. Daly, G. McVean, J. C. Mullikin, J. M. Higgins, D. J. Richter, E. S. Lander, and D. Altshuler. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature genetics*, 32(1):135–142, 2002.

[251] A. Renyi. *A Diary on Information Theory (Probability & Mathematical Statistics).* John Wiley & Sons, 1987.

[252] A. Rich. Discovery of the hybrid helix and the first dna-rna hybridization. *The Journal of iological chemistry*, 281(12):7693–7696, 2006.

[253] F. Rodier, J. Campisi, and D. Bhaumik. Two faces of p53: aging and tumor suppression. *Nucleic acids research*, 35(22):7475–7484, 2007.

[254] N. A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, and M. W. Feldman. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet*, 1(6):e70, 2005.

[255] M. Rossbach and M. Garcia Martinez de Lecea. Translational genomics in personalized medicine - challenges en route to clinical practice. *The HUGO Journal*, 6(1):2, 2012.

[256] G. M. Rubin, M. D. Yandell, J. R. Wortman, G. L. Gabor Miklos, C. R. Nelson, I. K. Hariharan, M. E. Fortini, P. W. Li, R. Apweiler, W. Fleischmann, J. M. Cherry, S. Henikoff, M. P. Skupski, S. Misra, M. Ashburner, E. Birney, M. S. Boguski, T. Brody, P. Brokstein, S. E. Celniker, S. A. Chervitz, D. Coates, A. Cravchik, A. Gabrielian, R. F. Galle, W. M. Gelbart, R. A. George, L. S. Goldstein, F. Gong, P. Guan, N. L. Harris, B. A. Hay, R. A. Hoskins, J. Li, Z. Li, R. O. Hynes, S. J. Jones, P. M. Kuehl, B. Lemaitre, J. T. Littleton, D. K. Morrison, C. Mungall, P. H. Farrell, O. K. Pickeral, C. Shue, L. B. Vosshall, J. Zhang, Q. Zhao, X. H. Zheng, and S. Lewis. Comparative genomics of the eukaryotes. *Science*, 287(5461):2204–15, 2000.

[257] M. Ruiz-Marin, M. Matilla-Garcia, J. Cordoba, J. Susillo-Gonzalez, A. Romo-Astorga, A. Gonzalez-Perez, A. Ruiz, and J. Gayan. An entropy test for single-locus genetic association analysis. *BMC Genetics*, 11(1):19:33, 2010.

[258] M. Sabater-Lleal, M. Chillón, T. E. Howard, E. Gil, L. Almasy, J. Blangero, J. Fontcuberta, and J. M. Soria. Functional analysis of the genetic variability in the f7 gene promoter. *Atherosclerosis*, 195(2):262–268, 2007.

[259] Maria Sabater-Lleal, Angel Martinez-Perez, Alfonso Buil, Lasse Folkersen, Juan Carlos Souto, Maria Bruzelius, Montserrat Borrell, Jacob Odeberg, Angela Silveira, Per Eriksson, Laura Almasy, Anders Hamsten, and José Manuel Soria. A genome-wide association study identifies kng1 as a genetic determinant of plasma factor xi level and activated partial thromboplastin time. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 32(8):2008–2016, 2012.

[260] C. Sabatti, S. K. Service, A. L. Hartikainen, A. Pouta, S. Ripatti, J. Brodsky, C. G. Jones, N. A. Zaitlen, T. Varilo, M. Kaakinen, U. Sovio, A. Ruokonen, J. Laitinen, E. Jakkula, L. Coin, C. Hoggart, A. Collins, H. Turunen, S. Gabriel, P. Elliot, M. I. McCarthy, M. J. Daly, M. R. Järvelin, N. B. Freimer, and L. Peltonen. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*, 41(1):35–46, 2009.

[261] S. F. Saccone, R. Bolze, P. Thomas, J. Quan, G. Mehta, E. Deelman, J. A. Tischfield, and J. P. Rice. SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. *Nucleic acids research*, 38:W201–W209, 2010.

[262] Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[263] M. Sarkis, B. Goebel, Z. Dawy, J. Hagenauer, P. Hanus, and J. C. Mueller. Gene mapping of complex diseases - a comparison of methods from statistics informnation theory, and signal processing. *Signal Processing Magazine, IEEE*, 24(1):83–90, 2007.

[264] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, 18(20):6097–6100, 1990.

[265] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeuch. The information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188:415–431, 1986.

[266] N. J. Schork. Genetics of complex disease: approaches, problems, and solutions. *American journal of respiratory and critical care medicine*, 156(4 Pt 2):S103–9, 1997.

[267] T. Schurmann and P. Grassberger. Entropy estimation of symbol sequences. *Chaos*, 6(3):414–427, 1996.

[268] David W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

[269] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y. H. Lee, J. Hicks, S. J. Spence, A. T. Lee, K. Puura, T. Lehtimäki, D. Ledbetter, P. K. Gregersen, J. Bregman, J. S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M. C. King, D. Skuse, D. H. Geschwind, T. C. Gilliam, K. Ye, and M. Wigler. Strong association of de novo copy number mutations with autism. *Science*, 316(5823):445–449, 2007.

[270] S. C. Shah and A. Kusiak. Data mining and genetic algorithm based gene/SNP selection. *Artificial Intelligence in Medicine*, 31(3):183–196, 2004.

[271] C. E. Shannon. *An algebra for theoretical genetics.* PhD thesis, Massachusetts Institute of Technology. Dept of Mathematics, 1940.

[272] C. E. Shannon. A mathematical theory of communication. *The Bell Systems Technical Journal*, 27:379–423, 1948.

[273] S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):683–690, 1991.

[274] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, 2001.

[275] Gerasimos Siasos, Dimitris Tousoulis, Stamatios Kioufis, Evangelos Oikonomou, Zoi Siasou, Maria Limperi, Athanasios G Papavassiliou, and Christodoulos Stefanadis.

[276] W. Siedlecki and J. Sklansky. On automatic feature selection. In *Handbook of pattern recognition & computer vision*. World Scientific Publishing Co., Inc., 1993.

[277] K. A. Skelding, G. S. Gerhard, R. D. Simari, and D. R. Holmes Jr. The effect of hapmap on cardiovascular research and clinical practice. *Nature clinical practice. Cardiovascular medicine*, 4(3):136–142, 2007.

[278] E. S. Snitkin, , and D. Segrè. Epistatic interaction maps relative to multiple metabolic phenotypes. *PLoS Genetics*, 7(2):e1001294, 2011.

[279] H. C. So and P. C. Sham. Multiple testing and power calculations in genetic association studies. *Cold Spring Harbor Protocols*, 2011(1), 2011.

[280] X. Solé, E. Guinó, J. Valls, R. Iniesta, and V. Moreno. Snpstats: a web tool for the analysis of association studies. *Bioinformatics*, 22(15):1928–1929, 2006.

[281] P. Somol, P. Pudil, J. Novovicová, and P. Paclík. Adaptive floating search methods in feature selection. *Pattern Recognition Letters*, 20(11-13):1157–1163, 1999.

[282] J. M. Soria, L. Almasy, J. C. Souto, D. Bacq, A. Buil, A. Faure, E. Martínez-Marchán, J. Mateo, M. Borrell, W. Stone, M. Lathrop, J. Fontcuberta, and J. Blangero. A quantitative-trait locus in the human factor XII gene influences both plasma factor xii levels and susceptibility to thrombotic disease. *Am J Hum Genet*, 70(3):567–74, 2002.

[283] J. M. Soria, L. Almasy, J. C. Souto, M. Sabater-Lleal, J. Fontcuberta, and J. Blangero. The f7 gene and clotting factor VII levels: dissection of a human quantitative trait locus. *Human biology; an international record of research*, 77(5):561–575, 2005.

[284] J. C. Souto. Search for new thrombosis-related genes through intermediate phenotypes. genetic and household effects. *Pathophysiology of haemostasis and thrombosis*, 32(5-6):338–340, 2002.

[285] J. C. Souto, L. Almasy, M. Borrell, F. Blanco-Vaca, J. Mateo, J. M. Soria, I. Coll, R. Felices, W. Stone, J. Fontcuberta, and J. Blangero. Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Genetic Analysis of Idiopathic Thrombophilia. *American Journal of Human Genetics*, 67(6):1452–1459, 2000.

[286] R. S. Spielman, R. E. McGinnis, and W. J. Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American journal of human genetics*, 52(3):506–516, 1993.

[287] V. B. Sreenu, V. Alevoor, J. Nagaraju, and H. A. Nagarajaram. Micdb: database of prokaryotic microsatellites. *Nucleic Acids Research*, 31(1):106–108, 2003.

[288] W. Stacklies, H. Redestig, M. Scholz, D. Walther, and J. Selbig. pcaMethods a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23(9):1164–1167, 2007.

[289] S. D. Stearns. On selecting features for pattern classifiers. *Proceedings of the 3rd International Conference on Pattern Recognition*, 1976.

[290] V. De Stefano, G. Leone, and K. Paciaroni. Epidemiology of factor V leiden: clinical implications. *Seminars in Thrombosis and Hemostasis*, 24(4):367–79, 1998.

[291] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig. The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, 18(Suppl 2):S231–S240, 2002.

[292] G. D. Stormo and G. W. Hartzell. Identifying protein-binding sites from unaligned DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America*, 86(4):1183–1187, 1989.

[293] J M Stouthard, M Levi, C E Hack, C H Veenhof, H A Romijn, H P Sauerwein, and T van der Poll.

[294] B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavare, P. Deloukas, M. E. Hurles, and E. T. Dermitzakis. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853, 2007.

[295] H. A. Sturges. The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66, 1926.

[296] S. C. Su. Single nucleotide polymorphism data analysis - state-of-the-art review on this emerging field from a signal processing viewpoint. *Signal Processing Magazine, IEEE*, 24(1):75–82, 2007.

[297] L. Sucheston, P. Chanda, A. Zhang, D. Tritchler, and M. Ramanathan. Comparison of information-theoretic to statistical methods for gene-gene interactions in the presence of genetic heterogeneity. *BMC Genomics*, 11:487, 2010.

[298] P. Sudbery. *Human molecular genetics.* Harlow, England, 2nd edition edition, 2002.

[299] S. Szymczak, A. Nuzzo, C. Fuchsberger, D. F. Schwarz, A. Ziegler, R. Bellazzi, and B. W. Igl. Genetic association studies for gene expressions: permutation-based mutual information in a comparison with standard anova and as a novel approach for feature selection. *BMC proceedings*, 1:S9, 2007.

[300] M. E. Tabangin, J. G. Woo, and L. J. Martin. The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC proceedings*, 3(Suppl 7):S41, 2009.

[301] The Human Genome Variation Database. `https://gwas.biosciencedbc.jp/`, 2012. [Online; accessed July-2012].

[302] D. A. Tibaduiza, L. E. Mujica, M. Anaya, J. Rodellar, and A. Guemes. Principal Component Analysis vs. Independent Component Analysis for Damage Detection. *6 European Workshop on Structural Health Monitoring*, 2013.

[303] A. L. Tyler, F. W. Asselbergs, S. M. Williams, and J. H. Moore. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *BioEssays news and reviews in molecular cellular and developmental biology*, 31(2):220–227, 2009.

[304] W. S. J. Valdar. Scoring residue conservation. *Proteins: Structure, Function, and Bioinformatics*, 48(2):227–241, 2002.

[305] J. F. Valencia, M. Vallverdu, R. Schroeder, A. Voss, R. Vazquez, A. Bayes de Luna, and P. Caminal. Complexity of the short-term heart-rate variability. *Engineering in Medicine and Biology Magazine, IEEE*, 28(6):72–78, 2009.

[306] F. M. van't Hooft, A. Silveira, P. Tornvall, A. Iliadou, E. Ehrenborg, P. Eriksson, and A. Hamstem. Two common functional polymorphisms in the promoter region of the coagulation factor VII gene determining plasma factor VII activity and mass concentration. *Blood*, 10:3432–3441, 1999.

[307] M. Via, C. Gignoux, and E. González G. Burchard. The 1000 Genomes Project: new opportunities for research and social challenges. *Genome medicine*, 2(1):3+, 2010.

[308] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five Years of GWAS Discovery. *Am J Hum Genet*, 90(1):7–24, 2012.

[309] M. R. Šikonja and I. Kononenko. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Mach. Learn.*, 53(1-2):23–69, 2003.

[310] D. Wagsater, C. Zhu, J. Björkegren, J. Skogsberg, and P. Eriksson. Mmp-2 and mmp-9 are prominent matrix metalloproteinases during atherosclerosis development in the ldlr(-/-)apob(100/100) mouse. *Int J Mol Med*, 28(2):247–53, 2011.

[311] X. Wan, C. Yang, Q. Yang, H. Xue, N. Tang, and W. Yu. Megasnphunter: a learning approach to detect disease predisposition snps and high level interactions in genome wide association study. *BMC Bioinformatics*, 10(1):13, 2009.

[312] M. P. Wand. Data-based choice of histogram bin width. *The American Statistician*, 51:59–64, 1996.

[313] D. Wang, Y. Sun, P. Stang, J. A. Berlin, M. A. Wilcox, and Q. Li. Comparison of methods for correcting population stratification in a genome-wide association study of rheumatoid arthritis: principal-component analysis versus multidimensional scaling. *BMC Proceedings*, 3(Suppl 7):S109, 2009.

[314] J. Wang, L. Dai, and M. Li. Go semantic similarity-based false positive reduction of protein-protein interactions. *Proceedings of the 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*, 2009.

[315] Jinliang Wang. An estimator for pairwise relatedness using molecular markers. *Genetics*, 160(3):1203–1215, 2002.

[316] K. Wang and R. Samudrala. Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinformatics*, 7:385, 2006.

[317] W. Wang, Y. Guo, Y. Zou, and T. Wu. An improved algorithm for tag SNP selection based on pair-wise linkage disequilibrium. *Proceedings of the 2nd International Conference on Bioinformatics and Biomedical Engineering*, 2008.

[318] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, 2009.

[319] Zuoheng Wang. Direct assessment of multiple testing correction in case-control association studies with related individuals. *Genet Epidemiol*, 35(1):70–9, 2011.

[320] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[321] D. E. Weeks and K. Lange. A multilocus extension of the affected-pedigree-member method of linkage analysis. *American Journal of Human Genetics*, 50(4):859–868, 1992.

[322] The Wellcome, Trust Case, and Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[323] J. I. Weller, G. R. Wiggans, P. M. Van Randen, and M. Ron. Application of a canonical transformation to detection of quantitative trait loci with the aid of genetic markers in a multi-trait experiment. *Journal of Theoretical and Applied Genetics*, 92:998–1002, 1996.

[324] J. I. Weller, G. R. Wiggans, P. M. VanRaden, and M. Ron. Application of a canonical transformation to detection of quantitative trait loci with the aid of genetic markers in a multi-trait experiment. *Theoretical and Applied Genetics*, 92:998–1002, 1996.

[325] K. Wulff and F. H. Hermann. Twenty two novel mutations of the factor VII gene in factor VII deficiency. *Human Mutation*, 15:489–496, 2000.

[326] X. Xu, L. Tian, and L. J. Wei. Combining dependent tests for linkage or association across multiple phenotypic traits. *Biostatistics*, 4(2):223–229, 2003.

[327] Xuesen, J. Li, and X. M. Wu. Mutual information for testing gene-environment interaction. *PLoS ONE*, 4(2):e4578, 2009.

[328] Q. Yang, S. Kathiresan, J. P. Lin, G. H. Tofler, and C. J. O'Donnell. Genome-wide association and linkage analyses of hemostatic factors and hematological phenotypes in the framingham heart study. *BMC Medical Genetics*, 8, 2007.

[329] Q. Yang, H. Wu, C. Y. Guo, and C. S. Fox. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genetic Epidemiology*, 34(5):444–454, 2010.

[330] Hubert P. Yockey. An application of information theory to the central dogma and the sequence hypothesis. *Journal of Theoretical Biology*, 46(2):369–406, 1974.

[331] J. Yu, G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208, 2005.

[332] Z. G. Yu, Z. Mao, L. Q. Zhou, and V. V. Anh. A mutual information based sequence distance for vertebrate phylogeny using complete mitochondrial genomes. *Proceedings of the Third International Conference on Natural Computation*, 2:253–257, 2007.

[333] Z. G. Yu, X. W. Zhan, G. S. Han, R. W. Wang, V. Anh, and H. Chu. Proper distance metrics for phylogenetic analysis using complete genomes without sequence alignment. *International Journal of Molecular Sciences*, 11(3):1141–1154, 2010.

[334] X. Yuan, J. Zhang, and Y. Wang. Mutual information and linkage disequilibrium based snp association study by grouping case-control. *Genes &amp; Genomics*, 33:65–73, 2011.

[335] P. Yue, E. Melamud, and J. Moult. SNPs3D: Candidate gene and SNP selection for association studies. *BMC Bioinformatics*, 7:166, 2006.

[336] Q. Yue, V. Apprey, and G. E. Bonney. Which strategy is better for linkage analysis: single-nucleotide polymorphisms or microsatellites? evaluation by identity-by-state – identity-by-descent transformation affected sib-pair method on gaw14 data. *BMC Genetics*, 6(Suppl 1):S16, 2005.

[337] K. Zhang, Z. S. Qin, J. S. Liu, T. Chen, M. S. Waterman, and F. Sun. Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. *Genome research*, 14(5):908–916, 2004.

[338] L. Zhang, J. Liu, and H. W. Deng. A multilocus linkage disequilibrium measure based on mutual information theory and its applications. *Genetica*, 137(3):355–364, 2009.

[339] Shuanglin Zhang, Kui Zhang, Jinming Li, Fengzhu Sun, and Hongyu Zhao. Test of association for quantitative traits in general pedigrees: The quantitative pedigree disequilibrium test. *Genetic Epidemiology*, 21(Suppl 1):S370–5, 2001.

[340] Y. Zhang and J. S. Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, 39(9):1167–1173, 2007.

[341] Y. Zhao, W. T. Clark, M. Mort, D. N. Cooper, P. Radivojac, and S. D. Mooney. Prediction of functional regulatory snps in monogenic and complex disease. *Hum Mutat*, 32(10):1183–90, 2011.

[342] Y. Zhu, M. R. Spitz, C. I. Amos, J. Lin, M. B. Schabath, and X. Wu. An evolutionary perspective on single-nucleotide polymorphism screening in molecular cancer epidemiology. *Cancer Research*, 64:2251–2257, 2004.

[343] E. Zintzaras. The generalized odds ratio as a measure of genetic risk effect in the analysis and meta-analysis of association studies. *Statistical Applications in Genetics and Molecular Biology*, 9(1):21, 2010.

[344] J. Ziv and A. Lempel. Compression of Individual Sequences via Variable-Rate Coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.