Universidade de Santiago de Compostela

PhD Dissertation

Nonparametric inference for classification and association with high dimensional genetic data

Author:
Manuel Garcia Magariños

Supervisors (in alphabetical order):
Dr. Ricardo Cao Abad
Dr. Wenceslao Gonzalez Manteiga
Dr. Antonio Salas Ellacuriaga

November, 2009

D. Ricardo Cao Abad, Catedrático de Estadística e Investigación Operativa de la Universidade da Coruña, D. Wenceslao González Manteiga, Catedrático de Estadística e Investigación Operativa de la Universidade de Santiago de Compostela, y D. Antonio Salas Ellacuriaga, Profesor Contratado Doctor de Anatomía Patológica y Ciencias Forenses de la Universidade de Santiago de Compostela.

HACEN CONSTAR:

Que la presente memoria que lleva por título "**Nonparametric inference for classification and association with high dimensional genetic data**", del licenciado en Matemáticas por la Universidade de Santiago de Compostela Manuel García Magariños, ha sido realizada bajo su dirección, considerándola en condiciones para optar al Grado de Doctor y autorizándola para su presentación ante el tribunal correspondiente.

Y para que así conste, firman la presente en Santiago de Compostela, a 2 de noviembre de 2009.

Fdo.: Prof. Dr. Ricardo Cao Fdo.: Prof. Dr. Wenceslao González

Fdo.: Prof. Dr. Antonio Salas

Acknowledgements

Surely, the doctoral thesis is the end of a path, and hopefully it means the beginning of a new (better?) one. When trying to remember all the people who helped me along these last years, I realize that they are a lot and I am obviously going to forget somebody, so I will hold the events earlier apologizing for it. Everybody who really know me also think that I would like to complete this page with a list of people who have not been helpful to me, but maybe this is not the right place to do it. Therefore, it is fair to thank:

To my PhD supervisors. Each one of you has taught me not only scientific knowledge, but also ways to face the different problems you find inside this complex world. To Wenceslao, who introduced me in statistics and trusted me in the first place, so I will be always grateful to him. To Toño, who has guided me in the course of genetics, always finding time to solve my doubts or to help me when I was stuck. To Ricardo, who always has the right idea to solve an intricate problem when everybody else is lost. It is because of their support and their help that I am here now.

To Angel, Maviki, Maria B and Paula. Angel Carracedo opened to me the doors of the Genomic Medicine group and the Legal Medicine Institute. There is no a lot of people so open-minded to allow a mathematician-statistician to enter inside a biological-medical scientific group and to think he is worth the trouble. I am in debt to you. Maviki received me with open arms and gave me her constant support. I will always be thankful for allowing me to settle in your office when I arrived to the laboratory and also for your fruitful advices. Maria B and Paula have been always there to help me, besides being a mirror to look at.

To the people at the Statistics and OR department. They have been always there to help me in everything I could need. In this point, I would like to do special mention to the people who has shared with me my first steps in the world of universitary teaching: Manolo Febrero, Pedro Galeano and Rosa. Special thanks to Rosa, and also to Bea, for her help and her advices, and also for opening the path to those coming behind you. To Cesar, my excuses and my gratitude for all your help in the bureaucratic part.

To my friends and fellows in Medicine. Those who are not with us any more (Nuria, Francesca, I will never forget how much I enjoyed our time in Harry Potter's little room; Gloria, Maria M), those who were here when I arrived (Maria C, Alex, Meli, Vanesa, Eva) and those who arrived after me (Cata, Ana M, Ana F, Angela, Maria D, Alberto, Luis, Montse, Ana P, Yarimar, Laura, Liliana, Carla, Olalla, Miguel, Paula, Jens). I would really like to have time and space to write a dedication to each one of you. Thanks,

Fonde, for being such a good comrade: believe me, you are a person really difficult to find, and also to Raquel, for being such a good, real, friend for me. Unlike my fellows here, I do not spend much time in the hospital; even so, I would like to remind some people that, in some way or other, have been there: Jorge Amigo, Noa, Susi, ...

To my friends and fellows in Mathematics. Eduardo, Abelardo, Chor, Rafa, Fran, Roberto, ... with special mention to Rebeca and Adela. Thanks, Rebeca, for those fruitful discussions and chats in the scholar room.

To Anestis, and the people in Grenoble. Thanks to Anestis for being so welcoming, helpful and for finding always time to work with me. You accepted me in the laboratory with no questions and no requirements. A great part of this work is really yours. I would like to spread part of these acknowledgements to the people who helped me during my stay there: Juana, Bertrand, and specially Azmi.

To my friends and partners along the degree. Although it is only from time to time that we see each other, the true friendship never disappears. Lupe, Eva, Mariña, Julio, Ana, ...; I have a lot to thank to all of you. Antonio (and Patri) taught me what is really to be very good flatmates and a couple of real friends.

To my friends. Terceiro, Andres, Diego, Camoiras, Fran, Victor, ... it has been a lot of years sharing things with all of you, and I wish it will be many more.

To Rocio. Thanks for being the way you are; for your affection and your constant support. Don't ever change.

To my family. Thanks to my parents for being so supportive and for being always there. Thanks for encouraging my education. Thanks also to my grandparents and to the rest of my family.

Success is the ability to go from one failure
to another with no loss of enthusiasm.
Winston Churchill

In the end the Party would announce that two and two
made five, and you would have to believe it. It was
inevitable that they should make that claim sooner or later:
the logic of their position demanded it...
George Orwell, 1984

Contents

Chapter 1

Introduction

1.1 The human genome

Human genome is, technically, the expression used to design the genome of the *Homo Sapiens*, namely, the DNA sequence contained in each human cell, putting together the 23 chromosome pairs in the cell core and the mitochondrial DNA. It carries the basic information for the correct physical development of any human being, coding what is needed to produce the different proteins. The haploid genome (that is, with only one element from each pair) has a total length of around 3.2 billions of base pairs (3200 Mb), containing about 20000–30000 genes.

Since the boom of genetics in the nineties, a lot of hope has been set on this field, as the one thought to have the answers to numerous questions in the world of health and disease [394]. The Human Genome Project started oficially in the United States in 1990 and some of its aims are to study the variability in the human genome, to improve and develop sequencing technologies or, undoubtedly, the most published advance, to complete the human genome sequence, something achieved in 2000-2001 [203, 412]. All this entails too many advances in genomics, as the field in charge of the study of complete genomes. Obviously, scientific complexity of the task has been (and still is!) quite big, and it has caused an exponential growing of some sciences practically unknown before this adventure, like bioinformatics, applied to the study of the informative content of genomes by means of the intensive use of computational and technological resources. Likewise, the different branches of genomics, as functional genomics or comparative genomics, have substantionally evolved during the last decade [302].

One of the most prominent questions to answer about the genome is the one referring to its own variablity. As commented above, the number of base pairs making DNA up is enormous, but only a small percentage of these positions (locus) vary between humans, and it is in this variability where the reason for the different traits in humans can be found. The set of traits in

a human being is known as his phenotype, that is, the environmental expression of his genotype. Comparative genomics carry out the comparative analysis of genomes in different animal species and also estimate the proportion of coincidence between them. For instance, similarity of apes with humans is above 90%; two human beings have more than 99.9% of the DNA in common.

The existence of genetic diseases where only the genotype is responsible of its appearance is widely known since centuries ago. Hemophilia is undoubtedly joint to history as a damn heritage bequeathed to crown princes in royal families like spanish, russian or german. However, it is not this kind of illness the ones clinical genetics is interested in. Genetic basis of diseases caused by one single variant in the genome were unravelled time ago by former generations of geneticists. Moreover, most of them (Wilson's disease, acondroplasia, phenylketonuria, ...) are rare, and therefore their impact in global health is, though important, at least limited. Nowadays clinical genetics focus their interest on trying to discover the genetic basis of common diseases which genetic basis is thought to be complex, not only because they combine genetic with environmental triggers, but also because it is expected to be intricate in the sense of having numerous genes involved either separately or interacting. This means unarguably a challenge for mathematicians, since the pattern generating common disorders seems to be a complex one.

So, the conjunction of large amounts of data with complex patterns and the need to recognize them are the reasons for the landing of statistics in a biological field, as genetics is. But before moving to statistics, let us take some time to study the environment we are going to work with.

1.1.1 Nuclear DNA

DeoxyriboNucleic Acid (DNA) is a macromolecule present in every cell. It carries the genetic information needed for the correct development in every alive organism and some virus, being the main responsible of heritage. Chemically, DNA is a strand of nucleotides. Each nucleotide is compound of a common sugar (deoxyribose) and a common phosphate group, while the nitrogenated bases are specific: adenine (A), cytosine (C), guanine (G) and thymine (T). They are the foundation for the diversity and the evolution not only in humans, but also in the rest of the species. The sequence of bases along the chain is actually the code for the genetic information. Proteins to be produced in each moment of the cell life cycle strongly depend on this code.

The nature of the DNA macromolecule was not known before the discovering of the DNA structure in 1953 [423, 424]. Since the twenties, when some experiments on bacterium cells were carried out [162], scientist knew that DNA was the genetic transmitter, but its physical structure could not

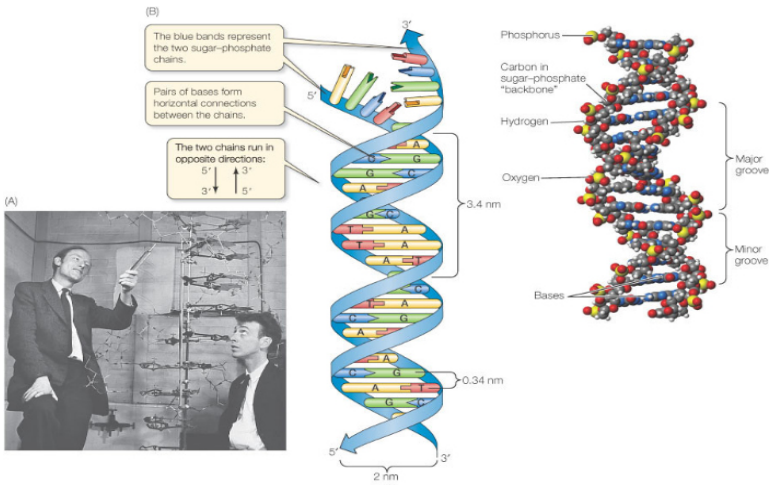

Figure 1.1: DNA double helix structure (B) as proposed by Watson (right) and Crick (left) (A). Image obtained from `www.nature.com`.

be explained. Watson and Crick proposed a revolutionary model for DNA physical structure, as it gave rise to the definition of the gene in chemical terms and meant the starting point to understand gene function and heritage. Even before Watson and Crick's discovery, it was known in science that genetic material should fulfill three main properties:

- It allows to be faithfully copied.
- The content has to be informative.
- Given the fact that evolution is slow along time, genetic material would change rarely.

Watson and Crick suggested a 3D structure compound by two strands or threads that bend making a double helix and joint by the effect of hydrogen chemical bonds. Figure 1.1 shows DNA double helix structure as proposed by them in 1953. This theory explained successfully previous results, and proposed immediately a way for the genetic material to be copied.

At the same time, double helix structure suggests how DNA would establish protein structure, so nowadays we know that the nucleotidic sequence generates the amino acidic sequence in each protein, by means of a genetic code that is degenerate (in the sense that more than one nucleotidic sequence can give rise to the same amino acid) but not ambiguous (a nucleotidic sequence generates always the same amino acid). The human genetic code is often depicted as in Figure 1.2, where each possible sequence of trinucleotides is shown together with the amino acid being generated.

Genes are the physical and functional units of heritage. Heritage of DNA from parents to the offspring consists of a complex process (meiosis)

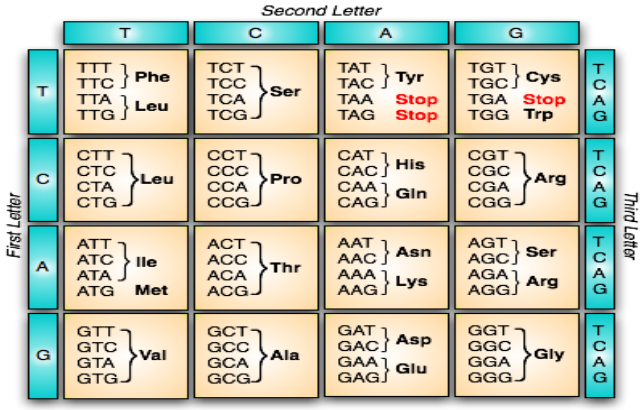

Figure 1.2: Human genetic code. 64 sequences of trinucleotides (codons) give rise to the 20 different amino acids in humans (degenerate code). 3 codons are stop codons, indicating the point where translation finishes. Image obtained from the Encyclopedia of Philosophy of the University of Stanford.

where the two DNA macromolecules belonging to each parent lead to a single final macromolecule in the child. Thus, the two parent chromosomes recombine during meisis producing new "mosaic" chromosomes that are in reality different combinations of the two parent ones. Therefore, the final child DNA can be interpreted as a mix of the two parents strands; one mosaic chromosome per parent. From a mathematical point of view, the most interesting point is about the recombination produced as a result of this mixture: combination of the nucleotidic sequence not only is not random, but certain regions of the genome segregate together, and as a consequence remain together throughout generations. This is going to be a main subject in population genetics, where different populations inherit different variants in each region, but also in clinical genetics, as association studies have to cope with linkage desequilibrium (see below).

As commented before, genetic material establish the basis for evolution. This happens by means of mutations in the nucleotidic sequence (mutability of the DNA sequence), little changes in the nucleotidic bases of the DNA strand, consistent of addings of bases (insertions), elimination of bases (deletions) or replacements of one base to another. Mutations can be the result of failures in any step of the replication process or as a product of the effect of mutational agents; this can occur in the germline or in somatic cells. Although human body is in possession of mechanisms that can repair this biological failures, this is not always the case. As a result, a mutation can be successful in a population and, with the pass of generations, become a new variant (a polymorphism) in the genome, or can be removed, either because the genealogy where it appears dies out or because the mutation

entails some kind of handicap for the individual. As expected, natural selection will favor those mutations giving rise to benefits, whilst those ones producing disadvantages (and especially the lethal ones) will tend slowly to disappear. Somatic mutations are also in the root of the emergence of sporadic cancer; cancer is a disease caused by the appearance of genetic abnormalities in cells, more common as the individual grows older. This leads to think that many types of cancer with a genetic basis can lie in those regions of the genome that deal with the mechanisms of biological repair of the human body, so most of the association studies involving cancer include those regions as possible genetic targets.

Population genetics is the branch of genetics that studies the factors determining the genetic composition of populations and the expected change in this composition along time. Genetic composition of a population is the set of frequencies of the different possibilities (called alleles) in a genotype. These allelic frequencies are the result of different processes that occur inside a population, like kind of mating, migrations, mutations, genetic recombinations or natural selection. Random fluctuations have also their importance, but obviously their influence tends to be null as a population tends to grow or evolve. The Hardy–Weinberg principle [174, 426] states that both allele and genotype frequencies in a population remain constant, that is, they are in equilibrium from generation to generation unless specific disturbing influences are introduced. In the simplest case of a single locus with two alleles: the dominant allele is denoted A and the recessive a and their frequencies are denoted by p and q: $\text{freq}(A) = p$, $\text{freq}(a) = q$ with $p + q = 1$. If the population is in equilibrium, then we will have $\text{freq}(AA) = p^2$, $\text{freq}(aa) = q^2$ and $\text{freq}(Aa) = 2pq$ in the population. Hardy–Weinberg principle has also a vital importance in clinical genetics, as little deviances from this equilibrium can mean association of a locus with a certain disease. For instance, if we detect much less homozygotes in the rare variant than expected, this could mean that these individuals are prone to suffer any disease, or simply that they have any disorder or physical handicap which makes them to be at a clear disadvantage with regard to other individuals.

When any locus take allelic frequencies significantly different than those established by the Hardy–Weinberg principle, this is usually reported as a Hardy–Weinberg disequilibrium status. This disequilibrium can be due to the action of different evolving forces. For instance, natural selection produces changes in allele frequencies in the sense that the best gifted individuals will tend to increase their proportion in a population. Random events can also modify allelic frequencies in a population, due to its finiteness, by means of what is known as genetic drift. Final state of a genetic drift event is that a population can reach homozigosity ($p = 1$) in a locus due to e.g. the random nature of mating. Genetic drift events are more probable in endogamic small populations, producing what is commonly know as kinship [164].

1.1.2 Mitochondrial DNA

Despite nuclear DNA fills most of the informations related with genetics, it is necessary to say that this is not the only genetic material in the human cell: mitochondrial DNA (mtDNA) is a small macromolecule that can be found in mitochondrions, namely, organelas that are located in the citoplasm of the cells. The mtDNA is a small molecule of about 16569 kb, giving rise to only 37 genes coding information about the production of RNA (RiboNucleic Acid) and polypeptides. In Figure 1.3, a graph with the mtDNA macromolecule is shown.

One of the most important subjects about mtDNA is related with the fact that there is no recombination associated with it: mtDNA is directly inherited all through the maternal line, so its only way to change is by means of mutation. Statistical studies conclude that a new mutation appears in mtDNA each 10000 years approximately. Some of these mutations mean only a new step in human evolution, while others are the cause for developing mitochondrial diseases. There are several dozens of mtDNA disorders, all of them rare and uncommon. However, all together are relatively prevalent (about 1 in 3000 individuals).

The fact of maternal heritage in combination with the low mutation rate makes mtDNA a key tool in the study of population genetics and the development of hypothesis about migrations and origin of prehistoric populations. Haplogroup is an useful and pragmatic concept in mtDNA research. It refers to a group of sequences (clades) that are closely phylogenetically related. Haplogroups are defined by series of diagnostic variants. Therefore, any mtDNA usually carries the diagnostic variants that allow to classify

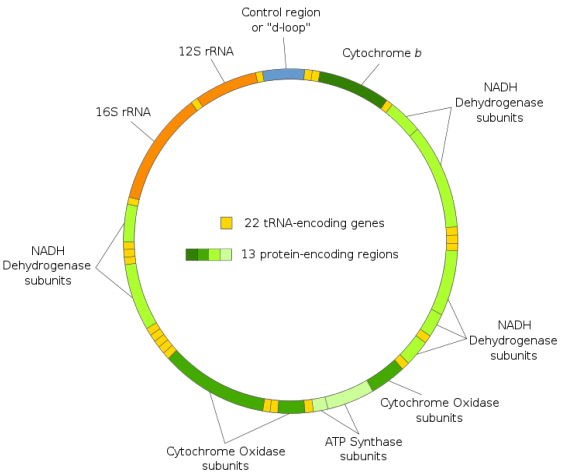

Figure 1.3: mtDNA macromolecule and its main regions. Image obtained from Wikipedia.

it into one particular haplogroup plus a series of "private" variants that are specific of this mtDNA. Variation at the mtDNA molecule is strongly stratified in populations such that different geographical regions or ethinic groups can be identified (from a matrilineal point of view) by particular set of variants/haplogroups. Analysis of the spatial distribution of haplogroups is generally known as phylogeography. These studies allow to reconstruct the demography and the origin of human population in ancient and more recent events.

1.1.3 Chromosomes

Chromosomes are each of the small, tiny bodies with the shape of a cane made by nuclear DNA plus proteins that arise as a result of the cell divisions (mitosis and meiosis). More specifically, chromosomes are the way chromatin is organized during these divisions, so we can say chromatin and chromosomes are morphologically different ways to represent the same entity: the nuclear DNA associated with certain kind of proteins called histones. Chromosomes show characteristic shapes and sizes. Each one has a condensed, constricted region called centromere placed around the middle. An important matter about them is that the number of chromosomes is constant for the individuals belonging to the same species: this number is called diploid number. The word diploid refers to the fact that in most of the alive organisms chromosomes are organized in pairs, being each one a single copy (almost exact) of the other; these are called homologous chromosomes. Therefore, each human being carries two complete genomes or two complete chromosome sets. This diploid number is commonly represented as $2n$ and change among species, being 46 in humans. Arrangement of the chromosomes during the mitosis cycle is known as karyotype. Figure 1.4 shows the shape karyotype takes in humans.

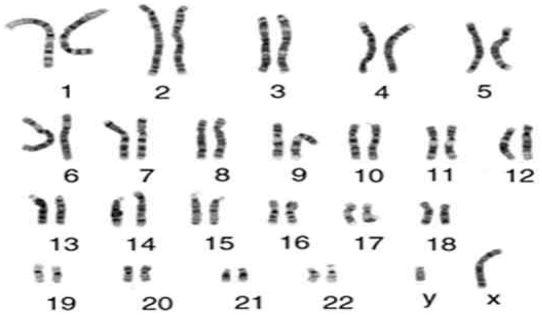

Figure 1.4: Karyotype of a human male. As can be observed, chromosomes differ both in shape and size. Y chromosome is substantially smaller than the X one. Image obtained from the Biotechnology web page of the Australian Government Initiative.

Given that homologous chromosomes are virtually alike, they contain the same genes in the same position, called locus (loci, in the plural). In many organisms, it happens that one of the homologous chromosome pairs is different to the others, stating the gender of an individual. In the human karyotype, there are 22 pairs of chromosomes not linked to gender, called autosomes, and one pair of sex chromosomes, making a total of 46.

As commented previously, sex chromosomes state the gender of an individual. Women have a pair of identical sex chromosomes called X chromosomes; men have a pair of different chromosomes consisting of one X chromosome and one Y chromosome. Name of sex chromosomes is due to their shape. Y chromosome is substantially smaller than X. For instance, SRY gene is contained in it, being responsible of testis development and thus determining sex; likewise, some phenotypic features typical of men are thought to be located in Y chromosome.

The same as happened with mitochondrial DNA finds an equivalent with the Y chromosome: due to obvious reasons, Y chromosome is inherited directly from father to son, and mutation is the only way for changing. As a consequence, Y chromosome is also very useful in relation with developing hypothesis and theories about prehistoric populations, migrations and evolution [405].

Y-chromosomes can be also allocated into haplogroups. The same variants in polymorphic loci are normally shared by those individuals belonging to the same population, or at least having common recent ancestries. There are several diseases associated to sexual chromosomes. Undoubtedly, the most famous one is hemophilia, which is determined in the X chromosome. Hemophilia is a disease characterized by the existence of problems of blood coagulation, leading to hemorrhages that can be eventually fatal. Coagulation is a body function regulated by a certain gene located in the X chromosome. If this gene is damaged, women are probably covered, as they have two copies, but men will suffer from the disease. Hemophilia has been common along generations of royal families in Europe, partially due to inbreeding.

1.1.4 Genes

A gene is a nucleotide sequence inside the DNA molecule containing the necessary information to synthesize a macromolecule carrying out a certain cellular function. Each one of them is understood as the basic unity for genetic information storage and heritage, as this information will be eventually passed on to descendants. Genes are located along chromosomes. The position occupied by them is known as locus.

Inside a gene sequence, two different kinds of regions can be found: codifying regions are known as exons [152], and they are responsible for carrying the necessary information to produce a protein, managing each one the elab-

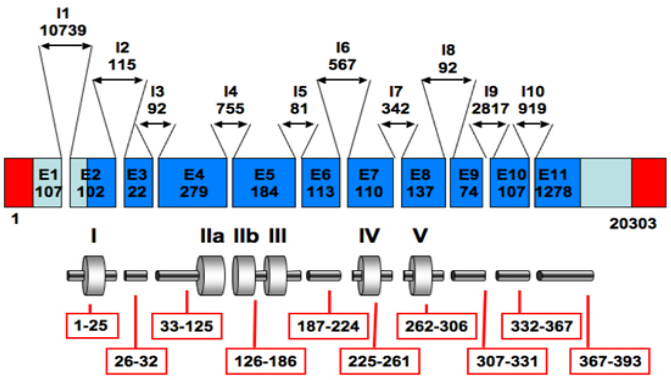

Figure 1.5: Organization of the human *P53* gene. 22000 bp for 11 exons (in blue). Translation begins in exon 2. Sizes of exons and introns are shown in bp. Image obtained from `p53.free.fr`.

oration of a different portion of the protein; exons are separated among them by large DNA regions called introns. Genetic information contained in introns is not used to build proteins, so these are called non-coding sections of the genome. Splicing is the process by means of which they will be removed when moving to mature RNA. This is one of the several steps carried out in the cell to elaborate the final protein. Figure 1.5 shows the sketch of the *P53* gene, dividing in exonic and intronic sequences. A particular case happens when, due to processes related with evolution like mutations, deletions,..., a gene stops being functional but remains inside DNA (as there is no an immediate process to remove it). These ones are called pseudogenes, and use to be similar to other genes in the same organism having specific function.

So genes carry the information needed for elaborating and sorting out the amino acidic chain giving rise to a protein. But, as expected, some mechanism regulating for time and location where each genome region is going to be translated has to be available. As its own police, DNA contributes also with the codifying regions necessary for carrying out this task, providing the means for self-regulation and interaction with information about cell physiological condition. Many steps are needed to produce a protein from the sequence of nucleotides inside a gene. Gene regulation process is shared among most of them, even after translation, by means of applying modifications to proteins. However, it is common thought that the great majority of gene regulation is carried out during transcription, the process by means of which DNA sequence is initially translated into messenger RNA (mRNA).

Regarding gene regulation, transcription changes according to the organism, and becomes more complicated as more complex this is, being therefore quite different in eukaryotics than in prokaryotes. Given the fact of the extremely large number of genes in eukaryotic organisms, it seems to be clear that most of the genes will be inactive in a given moment. This is under-

standable since the fact that many genes are only transcripted in unlikely events, as certain viral infections, or in early phases of the growing of an individual. Bearing in mind these facts, eukaryotic gene regulation has to be able to:

- Ensure that gene expression is generally inactive in the majority of the genome, and only the correct subset of genes are being transcripted.
- Generate thousands of gene expression patterns.

During transcription, gene regulation happens when RNA polymerase merges with the beginning of the DNA sequence, as well as when starts moving along the sequence. Anyway, gene expression always depends on the cell physiological state, in the sense that this is going to control if a gene is going to be read.

Purpose of this essay is to bring the reader closer to some of the many statistical approaches applicable in genetics, besides introducing some advances, hopefully profitable for (a part of) the scientific community. In this sense, one of the aims of statistical geneticists is to help to unravel the genetic basis of disease. Going one step beyond this, a major aim of statistical genetics could be to relate disease with heritage, and, regarding this, it is quite interesting to understand the mechanisms involved in heritage.

Gene heritage patterns were first studied by Gregor Mendel in the 19th century. However, his studies suggested a model where transmission is independent between couples of traits, equivalently, independent transmission among genes (though the term gene was not defined till some decades after). At the beginning of the 20th century, Bateson and Punnet studied the heritage pattern in two genes of the pea, noticing that the results deviated from the proportions expected by Mendel's theories. This suggested a more complex model where not all the genes are transmitted independently, as in many cases after Bateson and Punnet's studies they also proved to be linked. Recombination maps are obtained from several genes inside the same chromosome, studying the frequency of recombination among them and establishing the amount of linkage. Linkage between a pair of genes is usually associated with physical distance inside the chromosome, though there is no a mathematical exact relation. In Figure 1.6, a linkage map for a set of close genetic markers is shown. In mathematical terms, linkage translates as existence of high correlations between genes and deviations from Mendelian proportions. Moreover, recombination also happens inside genes and, as physical distance seems to be the critical point, it is more marked as the size of the gene increases. As it will be mentioned later in the Methodology section, existence of linkage makes association studies even more complicated, as in several times it will be difficult to say if the genetic basis of some disease is located in a certain genetic marker or in other linked with it.

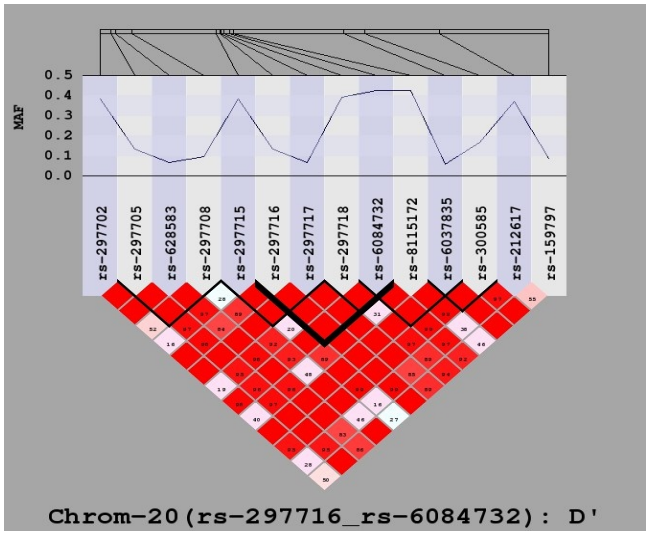

Figure 1.6: Linkage map for a set of genetic markers inside chromosome 20. This kind of graphs try to give a visual idea of how some close markers are usually inherited together. A colour palette indicates the level of linkage (correlation) between markers and sets of close markers. Image obtained using the software genomeSIMLA [107].

As heredity is not an independent process among genes, neither is its way of operating. Genes interact as the different parts of the production line in a factory. In this sense, genetic pathways is the name used to design the groups compound by genes which carry out complementary functions acting like steps of a more complex cell process. There are thousand of them, and interaction is common in and between connected pathways. But interaction also exists inside a gene, as nature and effect of the different alleles is highly changeable.

As mentioned above, the different possibilities or alleles that can be found in the DNA sequence appear as a result of mutations that occur in a population with the pass of generations. The group of known mutations inside a gene is called allelic series. Relation between the different alleles of a gene can take different forms:

- Complete dominance and recessivity. A dominant allele express itself with only one copy, as in heterozygosity, while its alternative allele will be totally recessive, that is, it will be expressed only in presence of two copies.
- Incomplete dominance. Term used for those cases where heterozygotes show an intermediate phenotype between the two homozygotes.
- Codominance. A heterozygote individual express both alleles, as it

happens with blood types in humans.

Gene–gene interaction is going to be, anyway, a recurrent topic along this essay. In the next section, a small introduction about common diseases with a genetic basis will be presented. As this basis is expected to be deeply complex, it is not crazy to think in the possibility of interactions leading to the appearance of such diseases, and the need for statistical methods grows exponentially with the amount of genetic databases.

1.1.5 Variation. Genetic markers

As statistics look for patterns in data with the aim of developing models able to explain reality, it is clear that its eye is going to be focused on DNA regions showing variability. Above we already made a few comments about variability between the human being genome and other species and also between two different unrelated individuals. Obviuosly, this enormous genomic similarity simplifies the task of looking for variants giving rise to different traits, but even so it is necessary to bear in mind that the approximately 0.1% of the DNA sequence where humans show differences represent millions of single positions along DNA. Adding interactions to this sketch does not make things easy.

Variation is, according to Darwin, the principle of evolution. Variation is fundamental for a population to evolve. In this sense, kinship, understood as crossing between closely related individuals, leads to loss of variation that eventually can lead to the appearance of diseases due to endogamy. Variation shows itself mainly within but also between populations. The level of population stratification, that is, how variation is stratified within a population, is an issue of interest in case-control association studies because stratification can be a confounding factor leading to false positives of association (see below). In this context, investigation of the genetic causes of complex diseases is facilitated by studying homogeneous populations (e.g. isolated populations).

Variation can take different forms in the DNA. Structural variations is the name used to include deletions, insertions or copy number variants (CNVs) of large segments in the human genome. These variants involve a large proportion of the genome and as a consequence the scientific community thinks their importance could be comparable with SNPs (see below). Figure 1.7 shows an approximated distribution of CNVs along the human genome. This field of study is at the peak now [221, 310, 413] and, despite being relatively recent, a new project has been created to study this kind of variables over the same individuals used in the HapMap Project [389]. Although global inventory of CNVs is still incomplete [361], they are thought to be important factors which variants contribute to common phenotypes of biomedical importance [273]. In fact, CNVs have already been used in

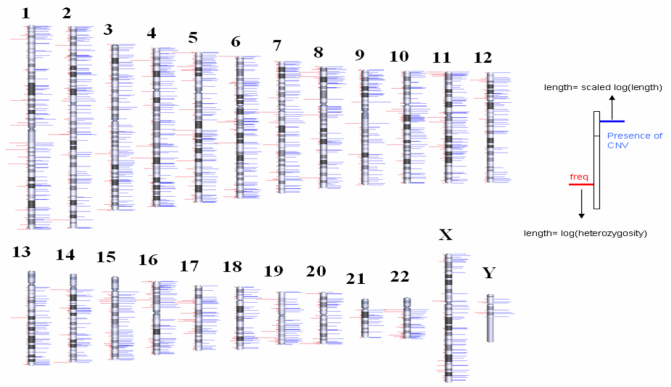

Figure 1.7: Localization and frequency of CNVs in the human genome. Image obtained from the Wellcome Trust Sanger Institute web page.

association studies looking for genetic variants involved in development of common diseases [428].

Along this essay, prediction of the condition of an individual by means of statistical methods will be focused on other kind of variations.

In genetics, the term polymorphism refers to the existence of multiple alleles of a gene in a population. For a locus to be considered as polymorphic in any population, its rare allele has to have an allelic frequency above 1%; in any other case, it will be considered a mutation, as it has not settled enough for ensuring its "survival" in the population. In this sense, it is clear that mutation is the main source of variation. Polymorphism includes different types of genetic markers. Let us take a quick look at them:

- Restriction Fragment Length Polymorphism (RFLP). DNA specific sequences cut by restriction enzymes and variable among individuals.
- Variable Number Tandem Repetition (VNTR). Term used to refer to locations inside the genome where a short nucleotide sequence is organized as a tandem repeat. They often show variations in length between individuals, which makes them extremely useful in the different branches of genetics. There are different classes of VNTRs (microsatellite, minisatellite, ...) but probably the most used are the Short Tandem Repeats (STRs). STRs occur when a pattern of 2–10 nucleotides is repeated, being the repeated sequences adjacent to each other. By examining the number of repetitions of enough STR loci, a unique genetic profile can be created for each individual. Due to this, STR analysis has become the prevalent analysis method for determining genetic profiles in forensic cases.
- Single Nucleotide Polymorphisms (SNP). Variations in the DNA sequence that affects only to one position in the genome (that is, one

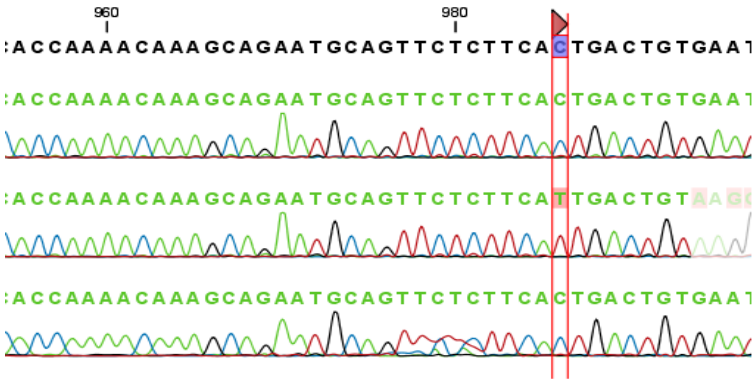

Figure 1.8: Identification of SNPs using sequencing analysis. In this illustration a C–T SNP is seen in position 986 of the sequence. Image obtained from `www.clcbio.com`.

nucleotide). A simple example is shown in Figure 1.8. SNPs form around 90% of all the human genomic variations, so they are very common, appearing frequently along DNA. It is a common thought in the scientific community that SNPs can be in the core of the development of some common diseases which could have a complex pattern. As a result, large resources have been used in this area, giving rise to an extensive area of study. This essay will be partly devoted to the study of statistical methods applied on SNP case–control data.

As mentioned above, analysis of common diseases with a likely genetic basis will be a main issue in the present essay. Looking for the genetic basis of such diseases is one of the greatest challenges today [335]. Cancer, asthma or psychiatric disorders, to name some examples, are likely to have complex genetic basis. This complexity refers not only to the fact of having several locus associated with a particular disease, but also with the environment having an effect (in interaction with genotype). Many other factors have to be taken into account when carrying out a case-control association study, some of them of a biological nature, such as phenocopy, low penetrances, etc., whereas some are more properly related to the study design, such as sample sizes, approaches to correct for multiple test hypothesis, etc. [394]. Due to this complexity, the study of the multifactorial diseases represent a main challenge for statistics, as methods of analysis of genetic databases dealing with this complexity have to be adjusted or developed, at the same time they prove to be computationally feasible.

1.1.6 Human Genome Project

In 1997, a research group from the University of Munich published the sequence of a mtDNA region of 379 bp from a piece of bone of an original

fossil from a Neandertal discovered in 1856 [223]. This study was considered to be an amazing technical achievement, due to all the technical difficulties related to it. Besides, it enabled to explain that Neandertal genealogy became extinct without contributing to the mtDNA of modern humans. This achievement was followed by studies with similar significance and even more technical difficulties that later will lead to the sequence of the first Neandertal mtDNA genome. This period illustrates the enormous technological advances that have allowed the boom of genomics, understood as the science studying nuclear complete genomes. The Human Genome Project (HGP), developed in 1990-2006, is the paradigm of all the advances in the genomic science in these last two decades.

HGP stated in 1998 a series of aims to be fulfilled. Among them we find:

- To complete the human genome sequence in 2003. This initial aim was forced to be sped up because of the competition created with the private initiative of Celera Genomics, so finally the first draft of the genome was published in 2001 in Nature [203] (the map from the HGP) and Science [412] (the map from Celera Genomics).

- To study variation in the human genome. It is expected that deep knowledge of variation can lead to know genes and loci involved in the development of complex diseases.

- To identify all the genes and determine the function of each one of them.

- To encourage development and appearance of sequencing technologies. As a result, sequencing technologies have evolved unbelievably fast, giving rise to high amounts of large datasets that have to be analyzed.

Apart from these, HGP had other aims, like popularization of genomics or achievement of advances in different genomic branches. All together represent a lot of benefits, many of them still to be achieved, and most of them related with advances in medicine and evolution, but also in other branches of the genetics field. Ethical, legal and social issues also took away part of the budget: confidenciality, genetic diagnosis, detection of genetic variants associated with diseases and traits, ... will be topics of constant debate in the immediate future and need therefore to be spreaded in society, so ethical basis supporting biomedical advances in the future can be well-based.

The tremendous advances in sequencing technologies have given rise to the arising of large databases full of genetic data. As a consecuence, bioinformatic tools have also evolved. These bioinformatic tools intend to provide with the ways to deal with data so all the meaningful information contained in it can be correctly obtained. In this sense, bioinformatics is probably the field of science which has better exploited the boom of genomics.

Consequences of the HGP will let to be felt in science during a while. That will be of advantage to mankind, as advances in biomedicine and biotechnology are expected to be huge.

1.2 Gene expression

As told in the previous section, the nucleotide sequences in certain DNA fragments contain the necessary information for elaborating proteins, the main structural and functional elements in the human cell. In fact, these nucleotide sequences fix the amino acids and the order in which they have to be added during protein synthesis. The process by which this information inside DNA is decoded and translated giving rise to proteins is commonly known as gene expression. This process comprises two essential steps: transcription (copy of a DNA sequence in the messenger RNA) and translation (from the messenger RNA to the protein). A general sketch of the process is shown in Figure 1.9.

Study of gene expression processes provides geneticists with the ability to detect the entire complement of genes whose expression pattern is perturbed in an organism with a given phenotype or trait, aiming to discover the genetic basis of complex traits [28]. It consists of measuring the amount of mRNA molecules (transcriptome) that are being produced in one or a population of cells for each gene. Microarray methods are the tools needed to measure the transcriptome. Everything began in the early 1990s, when two groups pioneered microarray technology. Steve Fodor and co-workers at Affymetrix developed commercial microarrays, using photolithography and

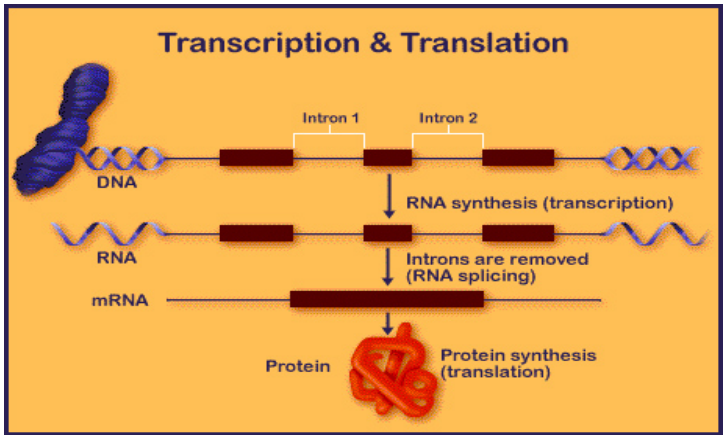

Figure 1.9: Explanatory diagram of the transcription and translation processes to produce a protein. Only exonic regions inside DNA are translated to give rise to the aminoacidic chain. Image obtained from `members.cox.net`.

solid–phase chemical synthesis to build short oligonucleotides in high density on a solid surface [132, 309]. Today, Affymetrix controls around 70 per cent of the market. At the same time as Fodor and co-workers, Patrick Brown and colleagues at Stanford University School of Medicine were developing a microarray that was manufactured by mechanically printing small spots of DNA solutions onto a glass microscope slide; this last technology was widely adopted, especially in academic settings, where its relatively low cost and flexibility were important.

Search for genetics of complex diseases by means of gene expression has its zenith in the use of some gene expression datasets [5, 6, 161] that have become widely used in the statistical literature, with the aim of compare and contrast abilities and disabilities of different statistical methods developed to be used in the field. Nevertheless, many of the scientific articles about gene expression aim to study genetics beyond disease, aiming to discover gene function, complete gene pathways or gain knowledge in what respect to cellular processes [198, 263, 373, 374, 396]. Anyway, statistics have reached a great importance in this field, as the science designed to deal with data and uncertainty.

A typical microarray experiment follows several steps, briefly depicted in Figure 1.10. First, mRNA is isolated from a biological sample of interest. At this point, complementary DNA (cDNA) is usually synthesized, because DNA is more stable and easier to work with than RNA. Samples are then labelled with a fluorescent dye. The labelled DNA is then hybridized to the microarray surface. After hybridization, the microarrays are washed to remove non-specific signal and then scanned to obtain an image at the wavelength of the labels used. These images show the level of fluorescent label hybridizing to each spot on a microarray. The images are then processed with one of a variety of data acquisition software packages that calculate important measurements for each spot on the array, such as total intensity, local background, or pixel–by–pixel intensity. These measurements are what are usually referred to as raw results of gene expression for microarray data. Raw results are used to calculate an indicator of mRNA levels in the original biological sample. Different microarray platforms exist for measuring gene expression, such as Affymetrix, Illumina,...

As a logical consequence of progressive cost reduction and increase of computational capacities, gene expression data has proliferated during the last decade. The function of bioinformatics is now essential to the effective interrogation of gene expression data. This makes expertise in bioinformatics a prerequisite for efectiveness in genetics. Bioinformatics could be understood as the science making up biology and the computational tools needed to extract, organize and analyze all the information inside biological data. This task is highly difficult even if we only consider the gene expression field inside the inmensity of the biological science, as it poses a large number of statistical problems, different microarray platforms, lack of stan-

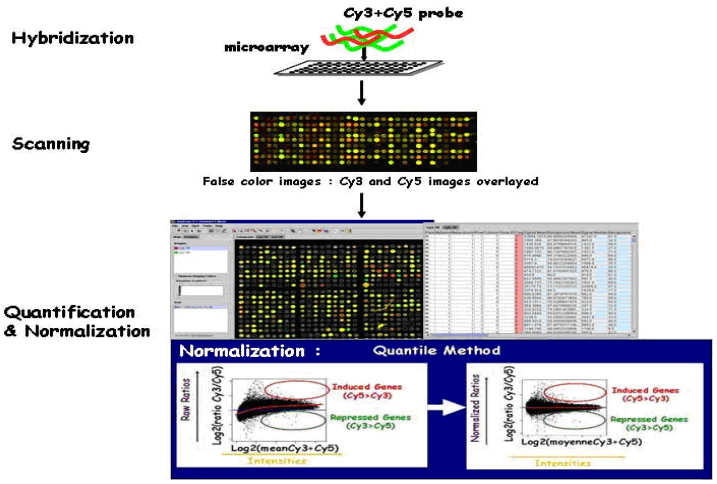

Figure 1.10: Steps carried out to obtain the final product of any microarray experiment, understood as numerical expression intensities for each sample in each gene. Image obtained from the IGBMC Microarray and Sequencing Platform web page.

dardization,... This problem is easily observed just making a quick search on the web, and realizing that the same original dataset can be found in different formats, it is used with different purposes and so on. On the other hand, one of the most important advantages of the gene expression field is the existence of an enormous variety of public databases, as opposed to the field of case–control SNP studies, where data is rarely made public, and only results are available for the scientific community.

This section is organized as follows: next subsection is devoted to explain briefly how gene expression measurements are obtained from spotted images; last subsection lists some of the areas where statistical methods are recursively used. Anyhow, we do not make a deep study, as next section will deal with classification and prediction methods for this kind of genetic data.

1.2.1 Image processing

Even though the study of the technical steps needed to obtain gene expression measurements is not within the scope of the present introduction, it is worth summarizing how image processing in a microarray slide is carried out. As briefly explained previously, cDNA microarrays are prepared by automatically printing thousands of cDNAs in an array format on glass microscope slides, which provide gene–specific hybridization targets. The image analysis task is to extract the average fluorescence intensity from each target site (cDNA region).

Image processing generally involves three stages, graphically explained in Figure 1.11. First, the spots representing the arrayed genes must be identified and distinguished from spurious signals that can arise due to precipitated probe or other hybridization artifacts or contaminants such as dust on the surface of the slide. This task is simplified to a certain extent because the robotic systems used to construct the arrays produce a regular arrangement of the spotted DNA fragments. The second stage is the estimation of background. For microarrays, it is important that the background is calculated locally for each spot, rather than globally for the entire image as uneven background can often arise during the hybridization process. Finally, the background–subtracted hybridization intensities for each spot must be calculated. There are currently two schools of thought regarding the calculation of intensities: the use of the median or the mean intensity for each spot [325]. Some studies [421] add a fourth stage, consisting of determining the quality of each measurement.

Common problems are noise, irregularities of spot shape, size, position, ... Therefore, users need to be able to acquire quality data, to control for imperfections that happen during printing and hybridization. Without a good scheme to produce reliable, high quality data, any complex data mining tools one may use can lead to misleading results [421].

Most of the commercially available microarray scanner manufacturers provide software that handles image processing; moreover, there are several additional image processing packages available. Many of them are listed in [325, 327]. At the same time, different numerical/probabilistic methods are

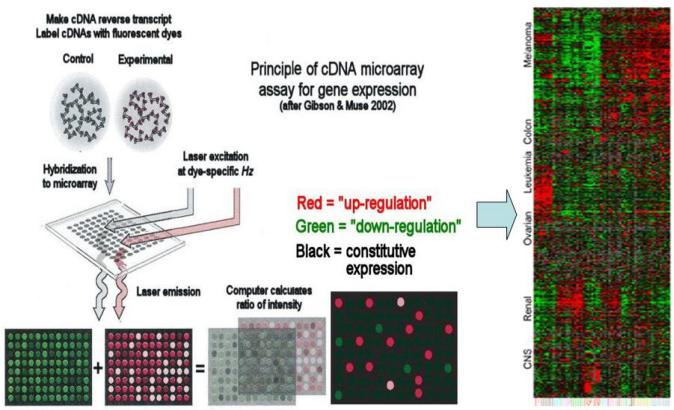

Figure 1.11: General stages to obtain gene expression images. This task is usually handled by specific image processing software packages. Joint of images obtained from the Centre de Recherche Public of Luxembourg (left) and the web page of Crossover Bioinformatics (right).

proposed to correctly measure the intensities for each spot, as for example [70, 421].

1.2.2 Research lines: challenges

Undoubtedly, there are many sections inside the whole gene expression measurement process that could be explained here. However, we will only focus on mentioning two of them, normalization of cDNA and oligonucleotide microarray data and classification of gene expression samples, as both require the use of statistical methods to be solved. Both of them mean active research lines in statistics nowadays.

1.2.2.1 Normalization of cDNA and oligonucleotide microarray data

There are a number of reasons why data should be normalized before statistical processing, including unequal quantities of starting RNA, differences in labelling or detection of efficiencies between the fluorescent dyes used [328]. Some sources of variability are random but most are systematic and due to specific features of the particular microarray technology. Systematic effects resulting from the biological process under study are of interest whereas other systematic sources should be removed [434]. For instance, in two-color cDNA microarrays, where each microarray has been hybridized with RNA from two sources labelled with different fluors, the two color channels obtained are usually referred as red and green (by convention). After image processing (see above), the red and green intensities must be normalized relative to one another so that the red/green ratios are as far as possible an unbiased representation of relative expression. In any other case, analysis and interpretation of gene expression profiles will be exposed to unfavorable and unreal results due to incorrect data processing [382].

The purpose of normalization is to adjust for effects which arise from variation in the microarray technology rather from biological differences between the RNA samples or between the printed probes. Imbalances between the red and green dyes may arise from differences between the labelling efficiencies or scanning properties of the two fluors complicated perhaps by the use of different scanner settings. If the imbalance is more complicated than a simple scaling of one channel relative to the other, as it usually will be, then the dye bias is a function of intensity and normalization will need to be intensity dependent. An example of gene expression normalization extracted from [261] can be seen in Figure 1.12. Differences between arrays may arise from differences in print quality, in ambient conditions when the plates were processed or simply from changes in the scanner settings. Therefore, normalization between as well as within arrays will need to be considered.

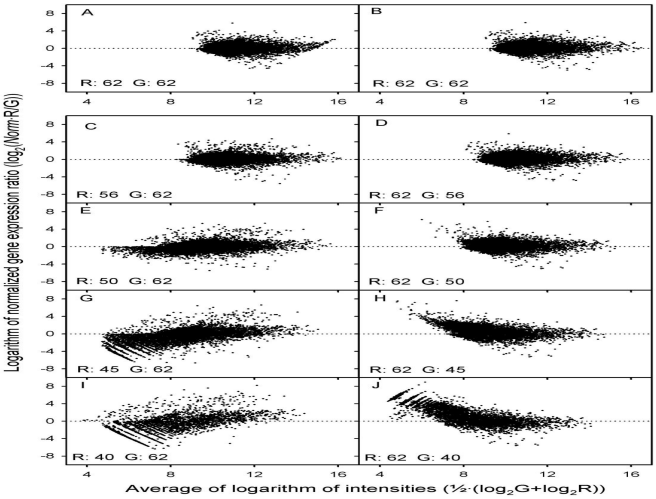

Figure 1.12: Normalized expression ratio *versus* average intensity in the red (R) and green (G) channel plotted on a double logarithmic scale. Different kinds of correction are carried out for the intensities of the red (C, E, G, I) and the green (D, F, H, J) channels. Image obtained from [261].

Scientific literature covering this topic is dense. Many approaches can be found with the aim of normalizing expression levels. Some of them are based on extremely simple assumptions, while there are a number of alternative approaches based on statistical techniques, including linear regression analysis [68], log centering, rank invariant methods [404] and Chen's ratio statistics [70]. Locally weighted linear regression (lowess) [74] analysis has also been proposed [437, 438]; a robust semiparametric normalization technique using local regression is shown in [210]; robust non–linear methods using cubic splines are developed in [434].

The Bioconductor project site (`www.bioconductor.org`) contains software packages to carry out different normalization methods, as those described in [382]. The Bioconductor packages use the free statistical programming environment R. For normalization of cDNA arrays, the relevant packages are *marrayNorm* [102, 103] and *limma*.

Most of the scientific literature addresing microarray normalization concerns cDNA array data, whereas only a few examples can be found for oligonucleotide arrays [242, 243, 357, 358]. Differences between these two kinds of array refer to the different DNA products that can be used to fill the probes in spotted microarrays.

Defining objective criteria for the quality of a DNA microarray assay is highly necessary, as microarray assays have become widespread and subsequent results of analysis applied to this data are highly dependent of the normalization carried out. A quality standard should be approved by the

scientific community. Information about the standards usually carried out can be found in the Normalization Working Group of the Microarray Gene Expression Data (MGED) web page (`www.mged.org`).

1.2.2.2 Classification of gene expression samples

One of the most important research lines in statistics nowadays is focused on developing new statistical tools for classification of gene expression samples. This classification can be made from known classes (supervised learning) or trying to recognize different classes from data patterns (unsupervised learning).

A fast revision of the scientific literature is enough to conclude that (the different types of) cancer dominate most of the articles, opposite to other genetic association studies, where it is common to find any disease with a likely genetic basis. Furthermore, a few databases can be recurrently found along the literature, as the standard data used to compare many of the statistical methods that keep arising with the existing ones. They cover some of the diseases more studied in gene expression, all of them related with cancer: breast cancer [343, 408, 429], leukemia [161], colon cancer [6], prostate cancer [380] and B–cell lymphoma [5], just to mention a few.

Gene expression studies have meant a chance to develop new statistical tools needed to deal with high–dimensional data, specially those cases where number of covariables (genes) p widely exceeds the number of samples n ($p >> n$). These high–dimensional problems also carry around computational matters so, when looking for proper statistical methods, not only classification and prediction abilities, but also computational feasibility and efficiency have to be beared in mind. Another unresolved issue refers to the way multiple test problems are corrected. Next section will be partially devoted to the study of these and other questions.

1.3 Classification and prediction in genetics: challenges

Along the last two sections the different genetic concepts needed to correctly understand the results of this work have been pinpointed. Polymorphic genetic variants received special attention (see Section 1.1), as high–dimensional studies involving them will represent a main part inside this essay.

Microarray gene expression studies have been also explained in great detail (see Section 1.2). Technical details about how to obtain the data (see Subsection 1.2.1) have been given, just as for the shape of the final data, understood as the fact of having much more variables than samples ($p >> n$ problem), often in a proportion of 100 to 1. Moreover, scientific areas where

statistics seem to have great importance have been listed (see 1.2.2). Use of statistical methods within these areas is expanding, giving rise to the development of new tools. This is a fundamental fact to obtain valuable scientific knowledge.

Therefore this section is thought to introduce the readers to forthcoming chapters, where new approaches ready to deal with genetic data will be explained. Empirical studies shown here will move in the sphere of clinical genetics, leaving aside statistical studies in forensic and population genetics.

The rest of this section is organized as follows. Case–control SNP association studies are reviewed in Subsection 1.3.1. Over there, a state–of–the–art of the field will be described from a statistical point of view. Furthermore, we will talk a little about genome–wide association studies (GWAS), that represent present and future of the genetics of disease. In Subsection 1.3.2 a revision on classification in the gene expression field is given in a supervised (training data where the classes are known) and unsupervised (trying to define new classes from data) way of learning. Emphasis will be pointed toward the former one, as some of the techniques developed in this essay deal with this kind of problems.

1.3.1 Case–control SNP association studies

Before the early 1980s, genetic risk factors for a disease or trait could be identified only through direct analysis of candidate genes, usually through association studies. Starting soon after their discovery, blood–group systems as ABO, MN and Rh were tested directly against an array of human diseases, tipically with little replicability [335].

Nowadays, two different approaches can be distinguished with the aim of discovering genetic regions involved in disease development:

Linkage analysis. A method for localizing genes that is based on the co–inheritance of genetic markers and phenotypes in families over several generations.

Association studies. A gene–discovery strategy that compares allele frequencies in cases and controls to assess the contribution of genetic variants to phenotypes in specific populations.

At a fundamental level, association and linkage analysis rely on similar principles and assumptions [40]. Both rely on the co–inheritance of adjacent DNA variants, but over few generations for linkage studies and over many generations for association. Thus, association studies can be regarded as very large linkage studies. Considering this idea, it is theorized that disease gene regions that are identified by linkage will often be large, and can encompass hundreds or even thousands of possible genes across many megabases of DNA. By contrast, association studies draw from historic recombination

so disease–associated regions are (theoretically) extremely small. Authors supporting association studies base his opinions under this idea, as disease regions are thought to be not large, due to the effect of recombination across time.

Linkage analysis leaded to the discovery of many genes for Mendelian diseases and traits years ago. Some examples are shown in [335]. Robustness of linkage analysis applied to Mendelian traits can be seen by its historic low false–positive rate [330]. But so far, all genes first identified by linkage analysis are those with low allele frequency and high displacement, that is, near Mendelian inheritance. Differences between Mendelian inheritance and non Mendelian inheritance are shown in Figure 1.13, and can be measured in terms of displacement, denoted by t, which is the number of standard deviations difference between the mean values of the two homozygotes AA and aa.

By contrast, no genes with moderate displacement have been identified in this way. Moderate displacement is what is expected to be found in common diseases with a genetic basis. Linkage studies have had only limited success in identifying genes for such diseases, like heart disease, asthma, diabetes and psychiatric disorders. Some recognized limitations of existing linkage

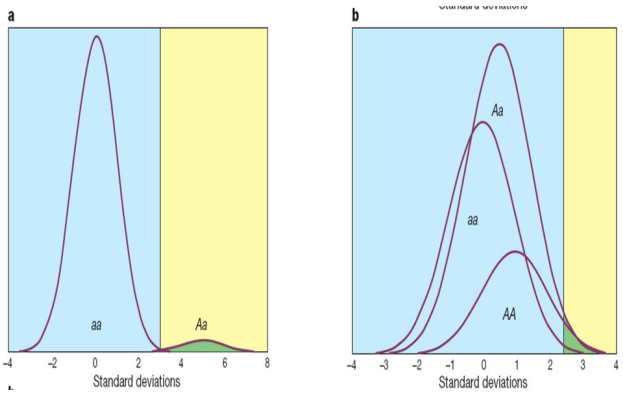

Figure 1.13: Examples of Mendelian and non Mendelian inheritance using a gaussian model. Graph (a) shows a dominant Mendelian locus with low allele frequency $p = 0.00275$ and displacement $t = 5$ standard deviations (sd; signaled in yellow background in the figure). Disease occurs above the threshold of 3 sd. Disease risk for heterozygotes (Aa) is 98% and for homozygotes aa is 0.13%. Prevalence is 0.67%. Graph (b) shows a non Mendelian additive locus with allele frequency $p = 0.4$ and displacement $t = 0.5$ sd for each A allele (total displacement $t = 1$). Disease occurs above the threshold of 2.5 sd. Disease risk for high–risk homozygotes AA is 6.7%, for heterozygotes Aa it is 2.3% and for homozygotes aa it is 0.62%. Prevalence is 2.4%. Image obtained from [335].

strategies in complex disorders are listed in [368].

Therefore, although traditional approaches like linkage analysis may identify a few of the genetic susceptibility agents, it seems to be clear that this problem should be rethought from a forward–genetics perspective. Most authors claim that genetic association studies provide greater power and resolution of location than linkage studies [334]. As a consequence, they have become the common approach to dissect the genetic etiology of complex traits. However, association studies suffer from many limitations, commonly referred as factors complicating genetic analysis (see forthcoming section). Despite these known limitations of association studies, their power to detect genetic contributions to complex diseases can be much greater than that of linkage studies. Most of the statistical methodologies developed along the last decade have been thought to attenuate some of these widely perceived limitations.

Two fundamentally different designs are used in genetic association studies: those that use families (family–based) and population designs that use unrelated individuals, called case–control designs. The approach often used is the case–control design, in which a difference in allele frequency is searched between affected individuals and unrelated unaffected controls. Although both designs have their own supporters, here we will try to offer an objective point of view.

Case–control gene association studies are undoubtedly the most common in the scientific literature. Many authors [62, 335] claim they are a more powerful and efficient approach, ensuring robustness when studying a large number of independent SNPs. From an epidemiological perspective, a major limitation of this approach is the potential for confounding leading to artefactual as opposed to causal associations, giving rise to false positives. Conventional case–control gene association studies have a long track record of false–positive results [205, 227].

On the other side, family–based designs have unique advantages over population–based designs, as they are robust against population admixture [394] and stratification, and allow both linkage and association to be tested. Furthermore, the fact that family–based designs contain both within and between–family information has substantial benefits in terms of multiple hypothesis testing, especially in the context of whole–genome association studies. As a drawback of family–based studies, it is often said that they require more genotyping, which increases the popularity of population designs. Anyway, family-based study designs continue to contribute much to the modern era of genome-wide association studies. A complete work showing the role of family studies in modern genetics research, using results from the Framingham Heart Study as examples, can be seen in [83].

The simplest version of the family–based design, the transmission disequilibrium test (TDT) developed in [384], is well known. Family trios are the basis of the TDT. A straightforward explanation is shown in Figure

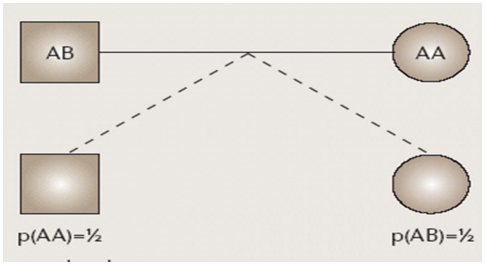

Figure 1.14: For the pedigree in the figure, the mother can only transmit the A allele because she is homozygous for A. Such a parent is not informative about transmissions to affected offspring. However, the father transmits A and B with equal frequency, yielding offspring with AA or AB genotypes with equal frequency. The transmission disequilibrium test (TDT) discards all homozygous parents and just looks at transmissions from a heterozygous parent to an offspring. Assuming the null hypothesis is correct, each transmission of A occurs with a probability of 1/2. Image obtained from [227].

1.14. This test compares the observed number of alleles of type A that are transmitted to the affected offspring with those expected from Mendelian transmissions. An excess of A (or B) alleles among the affected indicates that a disease susceptibility locus for the trait is in linkage and in linkage disequilibrium (LD) with the marker locus [226, 227].

Differently than thinking about them like face–to–face, in many studies, as for instead [227], is believed that both designs, which have different strengths and weakness, should be viewed as complementary and not as competitive in the effort to overcome the challenges of association studies for complex diseases. In terms of statistical power, the differences between the two approaches are generally small, specially when the use of trios in family designs is compared to case–control studies, as can be observed in Figure 1.15, obtained from [227].

1.3.1.1 Background

Although, as commented in the last section, gene–disease associations have been searched since much time ago, this last decade represent the explosion of genetic association studies.

The number of diseases suspicious of having a genetic basis is countless. The scientific literature offers a straightforward opportunity to find a high number of articles trying to detect a positive association. Here we will only show some examples of gene–disease association studies involving some of the most renowned diseases: different cancer types, like breast [248, 303, 369, 371], bladder [147], esophagus [192], lung [449], ...; psychiatric disorders

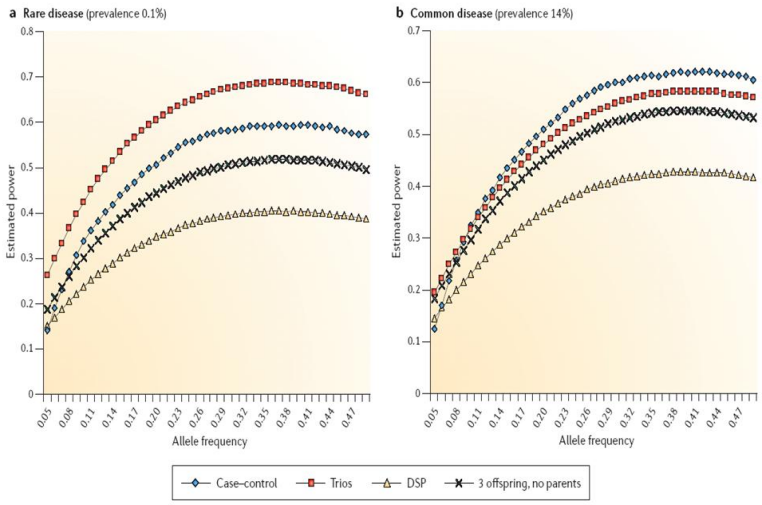

Figure 1.15: The estimated power levels for a case–control study with 200 cases and 200 controls are compared with those for various family–based designs: 200 trios (of an affected offspring plus parents); 200 discordant sibling pairs (DSPs: one affected and one unaffected) without parents; 200 trios of discordant offspring (at least one affected, at least one unaffected) and no parents. Comparisons are carried out for rare diseases (a) and common diseases (b). Image obtained from [227].

like bipolar disorder [191], schizophrenia [279] (these two are many times studied together, for instance in [31, 279]), autism [262], eating disorders [86], attention–deficit hiperactivity disorder (ADHD) [414]; neurological diseases like Parkinson's disease [150] or Alzheimer's disease [239]; diabetes type 1 [321] and type 2 [72]; myocardial infarction [249] and heart–related disorders like blood pressure (BP) [375] and hypertension [272]; ischemic stroke [80]; rheumatoid arthritis [249]; asthma [178]; obesity [324] or even traits like hair color or skin pigmentation [39, 332].

Unfortunately the literature is teeming with reports of associations that either cannot be replicated or for which corroboration by linkage has been impossible to find [62, 146, 388, 427]. Explanations for this lack of reproducibility are well–rehearsed, and tipically include poor study design, incorrect assumptions about genetic architecture and simple overinterpretation of data. Some estimations about a high incidence of false positives in case–control studies exist [205]

The common errors encountered in association studies of complex diseases are: small sample size, lack or improper correction for multiple testing, poorly matched control group, failure to attempt study replication, overinterpretation of results and publication bias and more. While many of those errors are related with mistakes in the study design or financial problems,

publication bias is the result of the imperative need to find a positive association, to publish the result in a journal with high impact index. Violations of Hardy–Weinberg equilibrium are also in the core of many non–replicable associations [402].

Credibility of genetic association studies is studied deeply in [204] by means of calculation of the Bayes factor

$$B = \sqrt{1 + \frac{m}{n_0}} \exp\left[\frac{(-z_m^2)}{2(1 + \frac{n_0}{m})}\right]$$

where m is the effective number of events in the study, z_m is the standardized test statistic for the distribution of the observed effect size θ under the null hypothesis $N(\theta, \sigma^2/m)$ and n_0 takes into account differences between expected values of the effects found under the null (H_0) and the alternative (H_1) hypothesis. As a bayesian approach, the Bayes factor is a poststudy measure of the odds of association being increased or decreased from the results in the study. $B < 1$ means that the study increases the odds that some probed association exists compared with previous thoughts, while $B > 1$ means a decrease in the odds. Figure 1.16 shows the estimated Bayes factors for 50 meta–analyses versus the p–values obtained for the corresponding variables, all of them under 0.05.

Anyway, focus of genetic association studies has not been only on the search for lonely, unique variants with a marginal effect. Interaction or

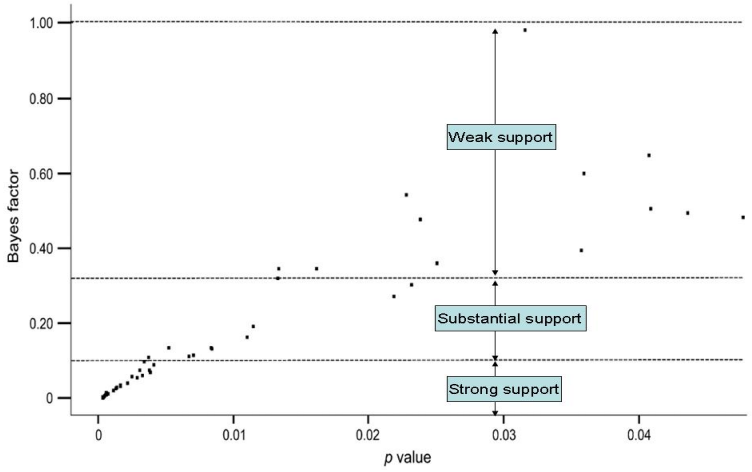

Figure 1.16: Estimated Bayes factors for 50 meta–analyses of genetic associations with formally statistically significant results. The Bayes factor is plotted against the observed p–value. Dashed lines correspond to threshold values (1, 0.32 and 0.1) separating different Bayes factor categories (weak support, substantial support and strong support). Image obtained from [204].

	Hom. A + Heter.	Hom. a	Total
Cases	r_0	r_1	R
Controls	s_0	s_1	S
Total	n_0	n_1	N

Table 1.1: 2×2 contingency table showing the observed frequencies in a case–control association study under a dominant model.

epistasis between different DNA regions is thought to have an effect on disease development, especially regarding complex diseases. This effect is in some cases believed to be even larger than that of marginal variants [287]. A large proportion of studies have been devoted to develop or evaluate statistical or machine learning tools with the aim of discovering gene–gene interactions [80, 81, 185, 239, 266, 276, 303, 318, 369, 375], despite the existing difficulties to settle a common definition for epistasis and interaction [286, 316, 317].

1.3.1.2 Review of statistical methods

This subsection is devoted to explain some of the most used statistical procedures in case–control association studies with SNPs. Although scientific literature is full of *ad hoc* non–prosperous methods, those mentioned here are commonly found in many studies.

Single point analysis Use of statistical techniques in genetic association studies widely varies from one study to another. In any case, single point analysis [3], looking for association between each SNP marker and the disease under study, are almost always carried out.

As biallelic SNPs can take three possible values (homozygous for the rare (A) or the common (a) allele and heterozygous), case–control data can usually be arranged in a 2×2 contingency table for each SNP, simply assigning the heterozygous individuals to one of the homozygous variants. This assignment is made according to either a dominant or a recessive model. An example of 2×2 contingency table is shown in Table 1.1.

The odds ratio (OR) or cross–product ratio is obtained as the quotient of the odds.

$$\mathrm{OR} = \frac{r_0/r_1}{s_0/s_1} = \frac{r_0 s_1}{r_1 s_0}$$

It takes values in $(0, \infty)$. Independence of case–control status and genotype is equivalent to $\mathrm{OR} = 1$. When $1 < \mathrm{OR} < \infty$, allele A is more likely to be associated with case status, and vice versa if $0 < \mathrm{OR} < 1$. Values of OR further away from 1 represent stronger levels of association. Confidence intervals for the OR values are given by

$$\text{OR} \cdot \exp\left[\pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{r_0}+\frac{1}{r_1}+\frac{1}{s_0}+\frac{1}{s_1}}\right]$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\alpha/2)$–quantile of the standard normal distribution.

The chi–square (χ^2) statistic is a goodness–of–fit test first introduced by Karl Pearson [308]. With data from a 2×2 contingency table, it tests the null hypothesis H_0 stating no dependence between case–control status and genotype. In that case, the expected cell frequencies are

For sample counts, Pearson proposed the test statistic

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where the O_i are the observed frequencies (Table 1.1) and the E_i the expected ones (Table 1.2). For large samples X^2 has approximately a χ^2 null distribution with 1 degree of freedom (for 2×2 contingency tables). Significant deviations from this distribution, usually measured by means of p–values, indicate association. Pearson χ^2 statistic is easily extended to $I \times J$ contingency tables.

The Cochran–Armitage trend test [21] and the Fisher's exact test [128, 129] are other goodness–of–fit tests tipically used in genetic association studies (see GWAS section).

Logistic regression (LR) Over the last 30 years, the logistic regression model [403] has become a standard method of analysis in epidemiology [189]. It is suited for studies with a binary response variable.

From a dataset $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})$ are the variable measurements (genotype) in each individual and the vector of binary responses $Y = (y_1, \ldots, y_n) \in \{-1, 1\}^n$ informs about membership to cases (1) or controls (-1), the logistic regression model assumes that

$$P(y_i = 1|\mathbf{x}_i) = \pi(\mathbf{x}_i) = \frac{1}{1 + \exp(-\beta_0 - \mathbf{x}_i'\beta)}$$

where $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ is the vector of coefficients for each covariate and β_0 the independent coefficient. The decision of whether to assign the

	Hom. A + Heter.	Hom. a	Total
Cases	$\frac{n_0 R}{N}$	$\frac{n_1 R}{N}$	R
Controls	$\frac{n_0 S}{N}$	$\frac{n_1 S}{N}$	S
Total	n_0	n_1	N

Table 1.2: 2×2 contingency table showing the expected frequencies in a case–control association study.

i sample to cases or controls is usually accomplished comparing the probability estimate with a threshold (e.g. 0.5). The coefficients β are obtained maximizing the log–likelihood function

$$L(\beta) = \sum_{i=1}^{n} \{y_i \ln [\pi(\mathbf{x}_i)] + (1 - y_i) \ln [1 - \pi(\mathbf{x}_i)]\}$$

Coefficient values are easy to interpret, so the effect of each covariate (SNP) on the outcome can be known. Furthermore, significance of these same coefficients can be easily tested by means of different strategies. Logistic regression allows also for categorical variables (e.g. SNPs) as predictors; this is made translating categories to several binary (dummy) variables.

Use of logistic regression have been common in genetic association studies [75, 283, 304, 420], carrying out both conditional and unconditional models.

Classification and regression trees (CART) Tree methodology [50] is a product of the modern computer era. Classification trees (CART) provide a meaningful tool to discover associations in intricate high–dimensional problems, besides being easy to interpret. An example of a classification tree can be seen in Figure 1.17. Nodes or branches $T_0, \ldots, T_3$ ask a question to the dataset that split it into two new different and more homogeneous (in terms of the response) datasets. The so called terminal nodes $t_1, \ldots, t_5$ end with data division and assign a class to each subset. The splits are selected among a set $S = \{s_i\}$ of questions involving the covariates.

Like other classification procedures, CART usually divides each dataset in a training and a test sample. The entire construction of a tree revolves around three elements:

1. The selection of the splits.
2. The decisions when to declare a node terminal or to continue splitting it.
3. The assignment of each terminal node to a class.

The third element is solved in a straightforward way assigning to the class with more individuals in the training sample. To select the optimum split s_i in a binary class problem is made from an impurity measure

$$\Phi : D \to \mathbb{R}$$

being $D = \{(p_1, p_2) : p_1 + p_2 = 1, p_1 \geq 0, p_2 \geq 0\}$. The impurity measure Φ has to fulfill:

(i) Φ takes a maximum in $\frac{1}{2}, \frac{1}{2}$.

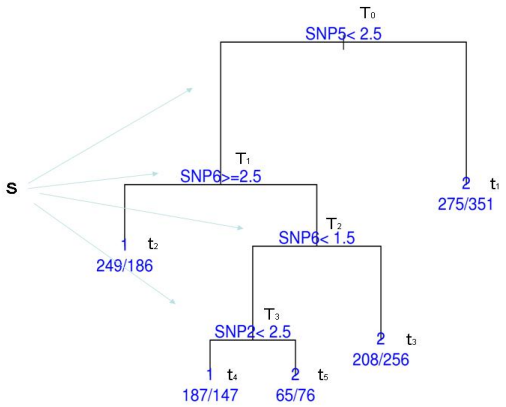

Figure 1.17: Example of a classification tree in a case–control classification problem with SNP data. The possible genotypes for a binary SNP are AA, Aa, and aa; that can be coded as 1, 2, and 3. Data is splitted from the selected SNP values in each branch/node. Terminal nodes classify as a case or a control depending on the rate of training samples showing each category. For instance, the most left leave of the tree indicates that SNP5 with genotypes AA or Aa, and SNP6 with genotype aa, are able to classify the cohort into 249 + 186 cases, where 186 are in reality control individuals; this indicates the classification error. Package *rpart* in R allows for graphical displays of classification trees.

(ii) Φ takes a minimum in $(0, 1)$ and $(1, 0)$.

(iii) Φ is symmetrical.

So the impurity in a node is defined as

$$i(T) = \Phi(\text{proportion of cases}, \text{proportion of controls})$$

and the split selected in a node will be the one giving rise to a higher decrease in node impurity from the previous node to the resulting nodes.

To decide when to declare a node terminal is so simple as choosing a threshold of decrease in node impurity from which no splitting can be carried out. Another option leading to the same end is to allow the tree to grow to their maximum and prune.

An advantage of CART in comparison with similar methods is that solves the missing data problem by means of surrogating splits. The surrogate split of a split is the split giving rise to the most similar data division.

Besides some of the results shown in this essay, CART has been widely used in association studies [69, 147, 321, 371, 436, 445].

Random forests (RFs) Random forests (RFs) [48] are a combination of tree predictors so it could be considered an ensembling procedure like bagging or boosting (see below); however, due to its unquestionable connection with CART, we think it should be better presented separately.

The generalization error for forests converges to a limit as the number of trees in the forest becomes large. It depends on the strength of the individual trees in the forest and the correlation between them. An upper bound for the generalization error is obtained in [48]:

$$\mathrm{Err}_{RF} \leq \rho \cdot \frac{1 - ST^2}{ST^2}$$

where ρ is the mean correlation between trees and ST is an estimate of the strength of the tree classifiers. More theoretical results and asymptotic properties about ensembles of trees can be found in [9, 48].

Trees constructed in RFs show slight differences with those developed in CART procedures. Mainly, they can be summarized in:

1. The best split at each node is selected from among a random subset of m predictor variables.

2. The training set used to grow each tree is a bootstrap resample of the observations. Due to this, some observations are represented multiple times, while others are left–out. The left–out observations are called out–of–bag (OOB) and are used to accurately estimate prediction error.

3. Trees are allowed to grow to their full size and there is no pruning.

A main advantage of trees (e.g. CART) is its interpretability. Nevertheless, this property is lost in random forest. To alleviate this problem, different variable importance measures, like the mean decrease accuracy (MDA) or the Gini index, have been developed. Appropriate explanations about these measures can be found in Chapter 5. Figure 1.18 shows an output of the randomForest package [245] in the R software. MDA and the Gini index are graphically displayed for a problem involving many dimensions; two of the variables (SNPs) show the highest values for both indices.

Due to recurrent construction of trees and use of resampling techniques, RF needs of an efficient computer implementation to reduce computation times, especially in high–dimensional problems. Random forests have been used in studies involving SNP markers [26, 56, 258]

Multifactor–dimensionality reduction (MDR) Multifactor dimensionality reduction (MDR) [337] is a result of the boom of genetics and the search for gene–gene interactions along last decade. MDR aims to

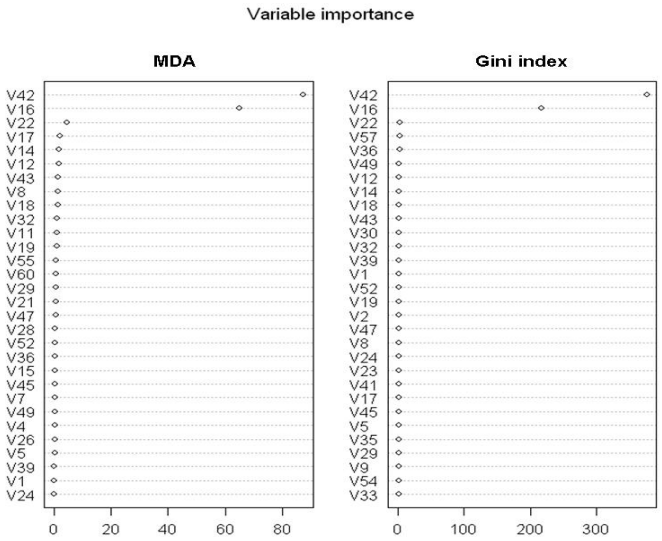

Figure 1.18: Variable importance measures (MDA and Gini index) graphically displayed using R. Variables are sorted from large to small importance.

identify gene–gene (and higher order) interactions, by means of reducing the dimensionality of genetic data.

Figure 1.19 illustrates the steps to be carried out to implement the MDR procedure for case–control study designs. The first step involves partitioning data into t equal parts for cross–validation (t–fold CV). Effect of the number of CV intervals in MDR has been discussed in [293]. In Step two, a set of N genetic markers is selected. These markers and their multifactor classes or cells are represented in Step three, and the ratio of cases to controls is evaluated within each cell. In Step four, each cell is labelled as high–risk or low–risk depending on if the ratio exceeds or not a pre–established threshold (e.g. 1). All possible combinations of N factors are tried and the best one in terms of classification error is selected. Step six is used to estimate the prediction error of the best model. This entire procedure is repeated for each CV division, and the best global set of markers is selected.

MDR has unquestionably prospered in the genetic field: a Java software has been developed to make its use easier [169], MDR properties have been studied [336], even it has been pointed out as a particular case of a classification tree [32]. Many studies can be found in the scientific literature using MDR to search for interactions [72, 75, 289, 297].

Logic regression Logic regression is an adaptive regression methodology that attempts to construct predictors as Boolean combinations of binary covariates [345]. As the name indicates, it is based on the construction of logical orders like $x_1 \vee (x_4 \wedge x_5^c)$ to detect interactions [369].

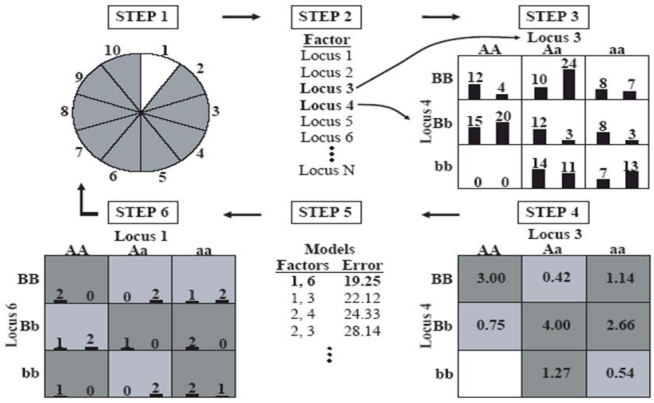

Figure 1.19: MDR steps. Dimension reduction is carried out by means of assigning high or low risk to each of the SNP–SNP combinations in Step 4. Image obtained from [169].

Let $x_1, \ldots, x_k$ be binary predictors and Y the response variable. The aim of logic regression is to fit regression models of the form:

$$g(E[Y]) = \beta_0 + \sum_{j=1}^{t} \beta_j B_j$$

where $B_1, \ldots, B_t$ are Boolean expressions like

$$B_j = (x_i \vee x_r^c) \wedge x_o$$

Note that the number t of Boolean expressions do not necessarily coincides with the number k of binary predictors, or the number p of variables in the model.

The above framework includes different regression models. For example

$$\begin{aligned} g(E[Y]) &= E[Y] && \rightarrow \text{linear regression} \\ g(E[Y]) &= \log\left(\tfrac{E[Y]}{1-E[Y]}\right) && \rightarrow \text{logistic regression} \end{aligned}$$

Classification trees can be seen as a logic regression model [345]. In fact, search of the best model in logic regression is similar in terms of splitting, deleting and pruning to the one carried out in CART, but now, optional changes in logical operators increase the set of possibilities to be chosen for splitting. The range of search algorithms for the best model is wide. Similarities with CART are abundant. Logic regression models are usually displayed like trees. Figure 1.20 shows a logic tree and the sequence of

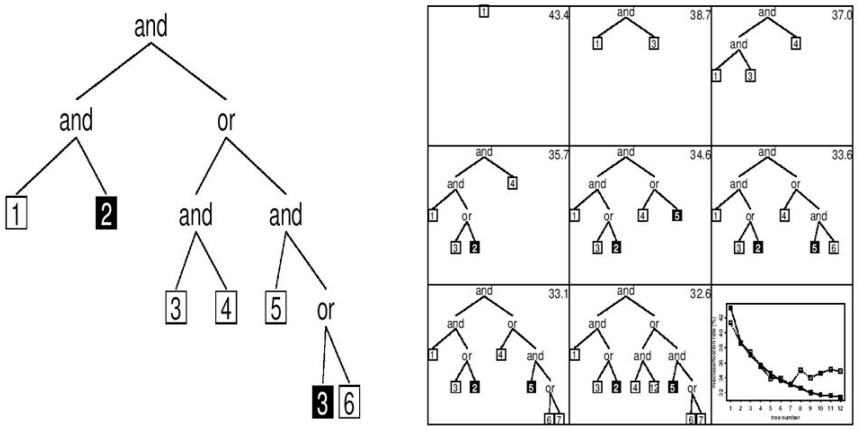

Figure 1.20: Logic regression tree (left), where nodes include logic operators, and sequence of steps to construct a logic tree (right). Image obtained from [345].

construction of a logic model by means of trees, obtained using a greedy search.

Use of logic regression models in genetics have been common [73, 219, 220, 300, 369], although not so wide as with the methods explained previously.

Combinatorial partitioning methods (CPMs) The combinatorial partitioning method (CPM) [298] is another parameter–free approach. As opposed to the other methods studied here, CPM focus only on quantitative phenotype responses. Its aim is to form subsets of SNPs containing different numbers of loci having maximum association with the response.

Let M be a set of m loci; the set of observed m–locus genotypes is denoted as G_M with size g_M. A genotypic partition is defined as a partition that includes one or more of all possible genotypes from the set G_M. A set of genotypic partitions, denoted K with size k, is a collection of two or more disjoint genotypic partitions. The application of the CPM to identify the subset of m loci that divide g_M genotypes into k partitions that are similar within and most dissimilar between partitions for the mean of a quantitative trait can be broken down into three steps. Figure 1.21 (left) shows a diagram of these steps. In step one, the estimation of the genetic variance is measured by variation among the means of the k partitions of the g_M genotypes.

$$
\begin{aligned}
s^2_K &= \sum_{i=1}^{k} \frac{n_i(\bar{Y}_i - \bar{Y})^2}{N} - \frac{k-1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{(Y_{ij} - \bar{Y}_i)^2}{N-k} \\
&= \frac{SS_K}{N} - \frac{k-1}{N} MS_w
\end{aligned}
$$

where $\bar{Y}$ is the sample mean for the response, $\bar{Y}_i$ and n_i are the sample mean and the sample size for partition i, Y_{ij} is the phenotype of individual j in partition i, SS_K is the sum of squared differences among partition means for set K and MS_w is the mean squared estimate of the phenotypic variability among individuals within genotype partitions. This is made for the complete space of sets of genotypic partitions, so it is very demanding, as the number of ways to partition g_M genotypes into a set of k genotypic partitions is given by

$$
S(g_M, k) = \frac{1}{k!} \sum_{i=0}^{k-1} (-1)^i \binom{k}{i} (k-i)^{g_M}
$$

For $m = 2$ and $g_M = 9$, there are 21146 ways to partition G_M into $k = 2, \ldots, 9$ partitions. The computational feasibility of the CPM is a matter of discussion [298] so many applications are restricted to the two–locus case. Once all the genetic variances have been estimated, the sets of genotypic partitions that predict more than a prespecified level of trait variability are retained for further analysis. Figure 1.21 (right) illustrates step one. In the second step each of the retained sets of genotypic partitions are validated by cross–validation methods. Finally, the third step is to select the best sets of genotypic partitions, on the basis of the results of the cross-validation from Step 2, and proceed to draw inferences about the combinations of variable loci and the relationships between the distribution of phenotypic variability and the distribution of the genotypes.

Some applications of CPM can be found in association studies [225, 288], just as discussions about CPM properties and advantages [179, 185].

Support vector machines (SVMs) SVMs are the main subject of Chapter 4 inside this essay. Proper explanation about the method can be found there.

Bayesian approaches Since long time ago, mutual attraction has existed between genetic association studies and Bayesian methods. The terms "Bayesian approaches" include many different methods, having all of them the foundations of Bayes theory behind. Take as example the use of Bayesian methods to detect interactions in [201] or to use known population allele frequencies as priors in [20]. In [85] Bayesian methods are used in

combination wit frequentist ones. An example of use of Bayesian networks can be seen in [265]. Different Bayesian approaches for variable selection in association studies are described in [141].

As showing all the existing approaches is unfeasible and there is no reason to enhance one over the others, here we will focus on giving a few explanations about Bayes decision theory [97], bearing in mind the two–class classification problem.

Bayesian approaches are based on the assumption that the decision problem is posed in probabilistic terms, and that all of the relevant probability values are known. So the *a priori* probabilities $P(y = -1)$ and $P(y = 1)$ reflecting our prior knowledge about the categories are needed. Let $P(\mathbf{x}|y = j)$, $j \in \{-1, 1\}$ be the state–conditional probability density for $\mathbf{x}$, then the Bayes rule is given by

$$P(y = j|\mathbf{x}) = \frac{P(\mathbf{x}|y = j)P(y = j)}{\sum_{k=-1,1} P(\mathbf{x}|y = k)P(y = k)}$$

and it is the basis for the leap from the prior probability to the posterior probability.

If we have an observation (genotype) $\mathbf{x}$ for which $P(y = -1|\mathbf{x})$ is greater than $P(y = 1|\mathbf{x})$, we would be naturally inclined to decide that the true state of nature is $y = -1$, and vice versa. To justify this procedure, let us calculate the probability of error whenever we make a decision. For a general decision rule d ($d(\mathbf{x}) \in \{-1, 1\}$), if we observe a particular genotype $\mathbf{x}$,

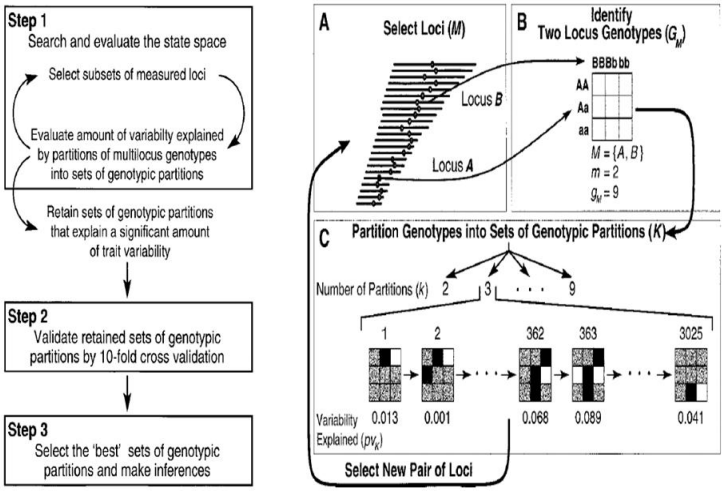

Figure 1.21: Schematic of CPM steps (left) and graphical explanation of the stages to be carried out in step 1 (right). Image obtained from [298].

$$P(\text{error}|\mathbf{x}) = I_{d(\mathbf{x})=1}P(y=-1|\mathbf{x}) + I_{d(\mathbf{x})=-1}P(y=1|\mathbf{x})$$

Thus it is clear that the Bayes decision rule minimizing the probability of error is

$$\begin{aligned} d(\mathbf{x}) &= 0 \quad \text{if } P(y=-1|\mathbf{x}) > P(y=1|\mathbf{x}) \\ d(\mathbf{x}) &= 1 \quad \text{if } P(y=1|\mathbf{x}) > P(y=-1|\mathbf{x}) \end{aligned}$$

Now we take λ_{01} as the loss incurred for deciding $y=-1$ when the true state of nature is $y=1$ and λ_{10} for the contrary case. Then, the conditional risk for each decision is given by

$$\begin{aligned} R(y=0|\mathbf{x}) &= \lambda_{00}P(y=0|\mathbf{x}) + \lambda_{01}P(y=1|\mathbf{x}) \\ R(y=1|\mathbf{x}) &= \lambda_{10}P(y=0|\mathbf{x}) + \lambda_{11}P(y=1|\mathbf{x}) \end{aligned}$$

Common choice for the loss function is the 0–1 loss used in classification. Nevertheless, as a matter of clinic diagnosis, it often does not have the same impact a case that is declared a control than a control that is declared a case. Hence the need of different loss functions.

Ensemble procedures: bagging and boosting Ensemble learning is the process by which multiple models such as classifiers or experts are strategically generated and combined to solve a particular problem. An ensemble–based system is obtained by combining diverse models (henceforth classifiers). Figure 1.22 shows the main idea behind ensemble algorithms.

A brief approach to the two most famous ensemble procedures, boosting and bagging, will be given here, apart from enumerating some other ones.

Boosting Boosting [140] is one of the most powerful learning ideas introduced in the last ten years [177]. It was originally designed for classification problems [46, 139, 140]. Boosting is a procedure that combines the outputs of many weak classifiers to produce a powerful "committee".

Let $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be the sample, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})$ the genotype of the i individual and $y_i = 1$, $y_i = -1$ the way to code now cases and controls, respectively. If $g(\mathbf{x})$ is a classifier producing a prediction in $\{-1, 1\}$, the error rate on the training sample is

$$\bar{\text{err}}_g = \frac{1}{n}\sum_{i=1}^{n} I\left[y_i \neq g(\mathbf{x}_i)\right] \tag{1.1}$$

A weak classifier is one whose error rate is only slightly better than random guessing. The purpose of boosting is to sequentially apply the

weak classification algorithm to repeatedly modified versions of the data, thereby producing a sequence of weak classifiers $g_m(\mathbf{x})$, $m = 1, \ldots, M$. The predictions from all of them are then combined through a weighted majority vote to produce the final prediction:

$$G(\mathbf{x}) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m g_m(\mathbf{x})\right)$$

The values $\alpha_1, \ldots, \alpha_M$ are computed by the corresponding boosting algorithm and weigh the contribution of each classifier. Their effect is to give more influence to the more accurate classifiers. Figure 1.23 (left) gives the schematics of the algorithm AdaBoost.M1 [139], also called Discrete AdaBoost [144]. The data modifications at each boosting step consist of applying weights $\omega_1, \ldots, \omega_n$ to each of the training observations $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$. Initially all the weights are set to $\omega_i = 1/n$, so that the first step simply trains the classifier on the data in the usual manner. Different implementations of the boosting procedure differ in the choice of the ω_i and α_j. AdaBoost takes as weights for classification

$$\alpha_m = \log\left(\frac{1 - \text{err}_{g_m}}{\text{err}_{g_m}}\right)$$

while the data weights at each step are updated in the following way

$$\omega_i^{(l)} = \omega_i^{(l-1)} \exp\left\{\alpha_m \cdot I\left[y_i \neq g_m(\mathbf{x}_i)\right]\right\}$$

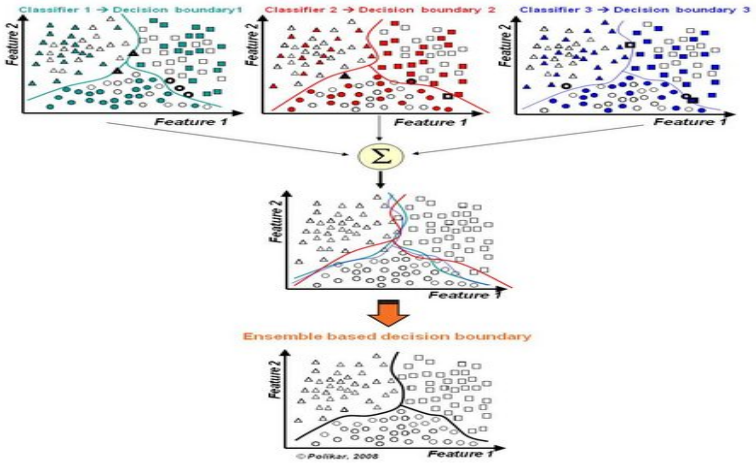

Figure 1.22: The main idea underlying ensemble learning algorithms is to reach a consensus from different classifiers. This is supposed to give competitive results in complex classification problems. Image obtained from Scholarpedia.

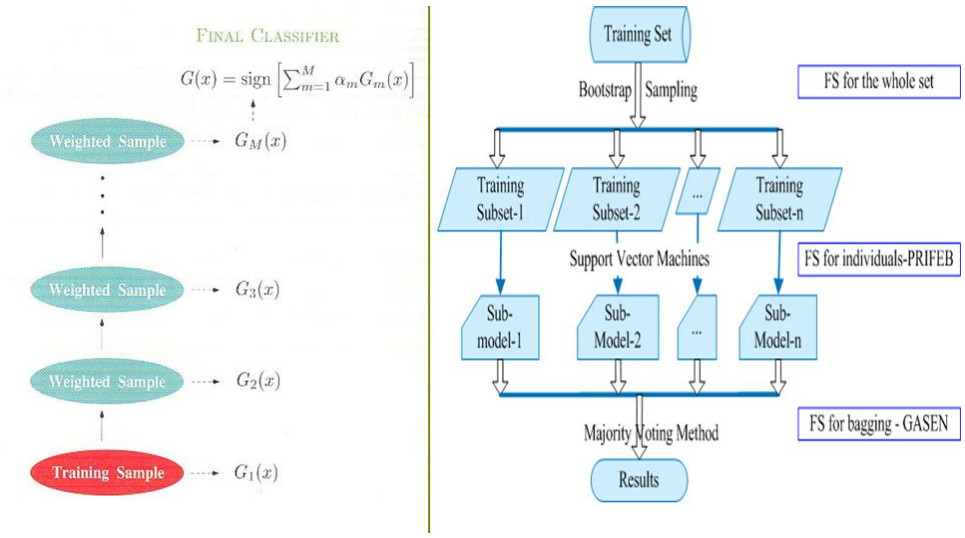

Figure 1.23: Schematics of the AdaBoost algorithm (left) and bagging in a feature selection problem (right) with support vector machine (SVM) classifiers. Images obtained from [177] and [244].

bearing in mind that the error for each classifier is weighted, not as in equation (1.1)

$$\text{err}_{g_m} = \frac{\sum_{i=1}^{n} \omega_i I\left[y_i = g_m(\mathbf{x}_i)\right]}{\sum_{i=1}^{n} \omega_i}$$

Application of boosting procedures is not common in genetic association studies. However, some empirical studies can be found [417], even in pharmacogenetic association studies [347].

Bagging Bagging [47] (from bootstrap aggregation) is a way to use bootstrap resamples [108, 111, 171] to improve classification or prediction.

If from our training data $\mathbf{Z} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ we obtain bootstrap samples $\mathbf{Z}^{*b}$, $b = 1, \ldots, B$ and a classifier $\hat{g}^{*b}(\mathbf{x})$ is applied to each one, the bagging estimate is defined by

$$G^B(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{g}^{*b}(\mathbf{x})$$

Bagging can be used with a wide variety of classifiers, although the most common ones are classification trees [94, 190] and classification trees with only one branch, called "stumps" [89]. Applications of bagging to association studies involving SNP markers are highly uncommon [371].

Other ensemble procedures Bagging and boosting have unquestionably been the most successful ensemble learning algorithms. Nevertheless, there are another ensemble procedures which have not received attention in the genetic field.

Stacking or stacked generalization [49, 235, 433] is a cross–validation method in which an ensemble of classifiers is first trained using bootstrapped samples of the training data.

Mixture of experts [206] generates several experts (classifiers) whose outputs are combined through a (generalized) linear rule. The weights of this combination are determined by a gating network, tipically trained using the expectation–maximization (EM) algorithm [88].

Other statistical methods Apart from the methodologies explained above, there are many others which use has not been wide in association studies. We will mention here some of them.

The k–nearest neighbors procedure has not spread as a classifier in genetic studies [24], although it has been used as an imputation method for missing genotype data [338, 441]. Neural networks [13, 275] have received limited attention [84, 179]. An interesting discussion can be found in [294]. Multivariate adaptive regression splines (MARS) [142] was developed as an adaptive procedure for regression, that can be adapted to handle classification problems [177]. It is well suited for high–dimensional problems in genetics [246]. Principal components analysis (PCA) [14] is a common method in population genetics [66], and it is also used in other branches of genetics to correct for stratification [323]. PCA is carried out in many association studies to account for population substructure in their case–control samples [12, 61, 67]. Apart from all these techniques, scientific literature is also full of *ad hoc* procedures [280, 305, 432] which do not usually have much impact.

1.3.1.3 Factors complicating genetic analysis

Common diseases with a genetic basis are likely to have a complex etiology. Some common errors usually made in association studies [62] were commented above. To solve them is often a simple matter of common sense, while in other cases remain an open research problem [92, 307]. Apart from these ones, which are usually related with the study design, there are numerous complicating factors that can be involved in complex genetic diseases [394].

Basically, these complicating factors can be divided into two categories: heterogeneity and interaction. Heterogeneity factors involve multiple predictor variables and/or multiple outcomes that complicate the analysis by creating a heterogeneous model. Definitions for the distinct types are:

Allelic heterogeneity: two or more alleles of a single locus are indepen-

dently associated with the same disease or trait.

Locus heterogeneity: two or more DNA variants in distinct loci are independently associated with the same disease or trait.

Phenocopy: presence of disease phenotypes with non–genetic basis (random or environmental).

Trait heterogeneity: insufficient specificity defining a disease, causing two or more distinct underlying diseases are considered to be only one.

Phenotypic variability: variation in the degree, severity or age of onset of symptoms exhibited by persons actually having the same disease.

Locus heterogeneity is expected to be common in complex diseases, as can be deduced from many genome–wide association studies [386, 390]. Existence of phenocopy cases support gene–environment studies. Trait heterogeneity appears as a result of the existing ignorance about some diseases yet to be fully studied. In this sense, some cluster analysis carried out with gene expression data aim to discover new variants of known diseases [22, 439]. Finally, phenotypic variability requires careful study and deep knowledge about the disease by the person in charge of collecting new samples for a particular association study. Figure 1.24 shows a small outline with definitions and examples of these complicating factors.

Interaction factors have already been discussed in this essay, especially regarding gene–gene interaction. The other important type of interaction is

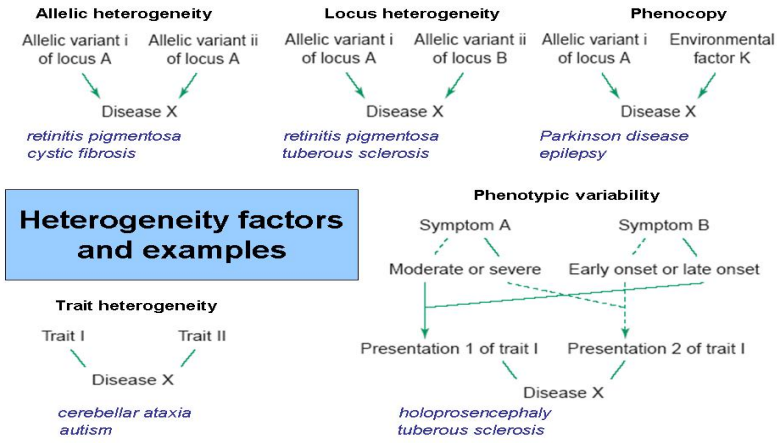

Figure 1.24: Sketch of the heterogeneity factors complicating genetic association analysis. Below them (in blue) some examples of diseases showing these factors can be observed. Image obtained from diagrams shown in [394].

gene–environment interaction, understood as the existence of combined effects between DNA variation and environmental factors. A certain amount of studies have been devoted to seek gene–environment interactions [119, 449], which have been proved to be in the core of some diseases, like depression [65] or bovine spongiform encephalopathy (BSE) [4, 77].

The biggest challenge nowadays in statistical genetics is the development of approaches addressing these factors, although bearing in mind that the source of some of them (e.g. trait heterogeneity) can only be corrected through deeper studies about the etiology of the disease(s) considered.

1.3.1.4 Simulation studies with SNP data

Not all the genetic studies in the literature are fully carried out with real data. Many of them involve the use of simulated data. To generate these simulations, different software packages have been developed in the last decade. Even so, it is difficult to find two of them serving the same purposes, as the range of aims to be achieved is broad.

Most of these programs simulate population genetic data [232], which complexity lies on accurately emulate complex demographic histories dependent on recombination rates, migrations and bottlenecks. There are also packages for pedigree haplotype data [234]. their main dificulty consists of correctly simulate patterns of linkage disequilibrium (LD) present in microsatellite data (expected to be weak) and in dense SNP panels (expected to be strong). Nevertheless, the aim of this section is addressed to simulation of case–control SNP association studies.

Although it has many other applications, the SNaP software [301] seems to fulfill many of the requirements usually demanded to a simulation package for SNP case–control data. It simulates different patterns of LD and haplotypes. Phenotype can be in different ways (continuous or categorical, including case–control). Several output formats are available and most importantly, parameters like amount of noise SNPs, penetrance models, allelic frequencies or sample sizes are defined by the user. Furthermore, a definition (multiplicative, additive or heterogenetic) for the kind of locus interaction to be simulated (if wanted) could be provided.

Probably the most difficult task when simulating genetic data is related to the need for constructing an LD structure in agreement with the one found inside the human genome and common haplotype blocks [418]. This combines with a more recent demand: try to simulate sample and SNP sizes as the ones carried out in GWAS, that is, dense SNP panels comprising tens or hundreds of thousands of variables. Appearance of simulation packages fulfilling those tasks would be invaluable in economical terms, as GWAS real data could be at least approximated without carrying out very expensive investments.

A recent approach is genomeSIMLA [107]; genomeSIMLA is a forward–

time population simulation method that can simulate realistic patterns of LD in both family–based and case–control datasets. As a consequence, it allows simulation of whole–genome association data in reasonable computation times. This software was developed from two previous approaches: genomeSIM [99] and SIMLA [30, 364]. More specific simulations to obtain high–dimensional data have been also carried out; see for instance [281].

A survey of different programs developed to simulate population genetic and genetic epidemiological data can be found in [251].

1.3.1.5 Genome–wide association studies (GWAS)

Genome–wide association studies (GWAS) are studies in which a dense array of genetic markers, capturing a substantial proportion of genome variation, is typed in a usually large set of DNA samples that are informative (case–control) for a trait of interest. The aim is to map susceptibility effects through the detection of associations [274]. Dense arrays are usually made up of tens or hundreds of markers, while sample sizes are about several thousands of individuals. Complexity and economical efforts required to develop such studies obliged to the establishment of consortia composed by different research groups [390].

GWAS are thought as the great hope to finally discover the genetic basis of common diseases. Last years have seen a boom of GWAS involving different diseases: diabetes, different types of cancer, coronary heart disease, traits like height or fat mass, ... The recent study [274] reviews more than 40 GWAS studies recently carried out in different disorders. Another well-known review in the field, although not so recent, was carried out in [181].

Despite having advantages typical of their characteristics (whole–genome under study, large sample sizes), GWAS suffer from many of the same problems as common association studies. Moreover, new ones are added.

For instance, control sample selection problem and population stratification are also present. Strategy of taking a joint sample of controls for different sample cases of different diseases is carried out in [390] and discussed in [131].

Heterogeneity problems, which were commented in a previous section, do not disappear either. In [274] pleiotropy, defined as a phenomenom whereby a single allele could affect several aspects of the phenotype, is pointed out as a possible factor of confusion in GWAS.

Multiple test correction in GWAS should undoubtedly be a matter of further research, given the huge number of SNPs under study. Without a proper correction, false–positive results would proliferate, increased also by biases generated due to genotyping in different laboratories [320]. In many cases, researchers are still using excessively conservative corrections from the past like Bonferroni's [38]. It has not happened the necessary transfer from new techniques developed in the statistical field [34, 109, 112, 113, 340]

	Hom. A	Heter.	Hom. a	Total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

Table 1.3: Contingency table for a case–control association study involving SNP markers.

to empirical association studies published in genetics. There are also plenty of studies developing their own *ad hoc* procedures, which have not usually been previously tested.

Most of the GWAS in the literature do not try complex or recent statistical approaches to detect association, but focus in single point analysis, aiming to detect associations separately for each marker. The most common approach is the Cochran–Armitage trend test [3, 21, 356].

The Cochran–Armitage test is tipically used in ordinal data analysis, to test for association in a $2 \times k$ contingency table (2×3 with SNP markers). It happens that the usual χ^2 test may not be able to detect a trend, but Cochran–Armitage test may be able to do so because a test statistic is chosen to reflect the anticipated trend. Distribution of case–control genotype counts can be represented as in Table 1.3.

The test statistic is given by

$$T = \sum_{i=0}^{2} t_i (r_i S - s_i R)$$

where the t_i are weights, which have to be chosen depending on the type of associations expected to be found. The null hypothesis of no association is expressed as

$$P(\text{Case}|\text{genot.}i) = P(\text{Control}|\text{genot.}i) = \frac{n_i}{N}$$

Assuming this holds, then

$$\begin{aligned} E(T) &= 0 \\ \text{Var}(T) &= \frac{SR}{N}\left(\sum_{i=0}^{2} t_i^2 n_i (N - n_i) - 2 \sum_{i=0}^{1} \sum_{j=i+1}^{2} t_i t_j n_i n_j\right) \end{aligned}$$

and as a large sample approximation

$$\frac{T}{\sqrt{\text{Var}(T)}} \sim N(0, 1)$$

More complex models looking for epistasis or interaction have been also developed and tried [150, 326]. Bayesian approaches [184, 448] are common,

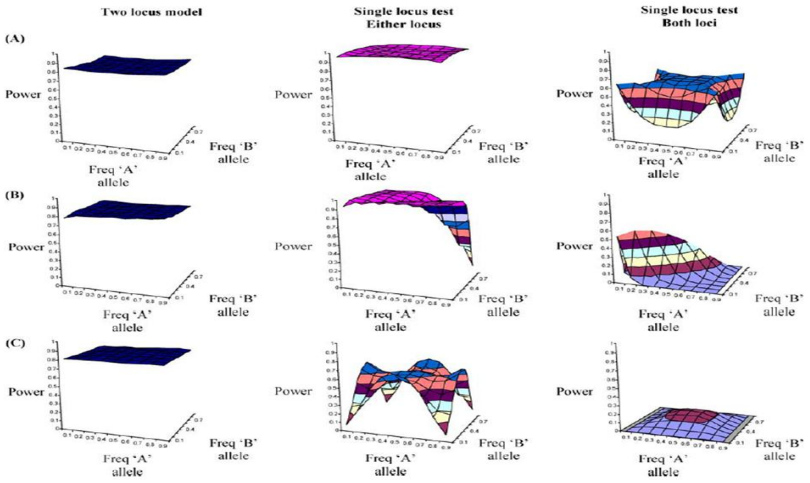

Figure 1.25: Power to detect association for 1500 individuals where both loci are responsible for 5% of the variance. (A) additive model with no epistasis, (B) epistatic model in which an individual requires at least one copy of the increaser allele at both loci to increase the phenotype and (C) exotic model, where an individual has to be heterozygous at both loci to have the trait. Image obtained from [122].

too. Two–stage models to seek epistasis [122, 266] have proved to be efficient with GWAS data. Figure 1.25 shows a comparison between the power of different approaches looking for gene–gene interaction in GWAS.

When results from GWAS are evaluated by an external expert, it is common to ask for replication studies in an independent group of samples, or even in independent populations [386]. The aim of replication is to fully confirm authenticity of the possitive association found. There is great controversy about this point, as carrying out a replication study leads to more economical expenses and new problems relative to the design of the study. Some studies [381] defend that joining of the two independent samples would have a positive impact, increasing significantly the power of the studies.

In any case, it is clear that, due to sample size and power considerations, many variants remain unidentified [202]. Meta–analyses association studies [254, 443] are the most feasible approaches to address these problems, as they enhance power while requiring low costs. Anyway, they are not exempt from problems either.

Choice of the set of SNP markers to be genotyped is also critical. The optimum would be to cover the whole–genome variability [104]. This is a matter of research nowadays, where different points of view are being defended [208]. Once a DNA region is identified as a target, resequencing [184] and fine mapping strategies [255] allow recovery of a more complete inventory of sequence variation within it.

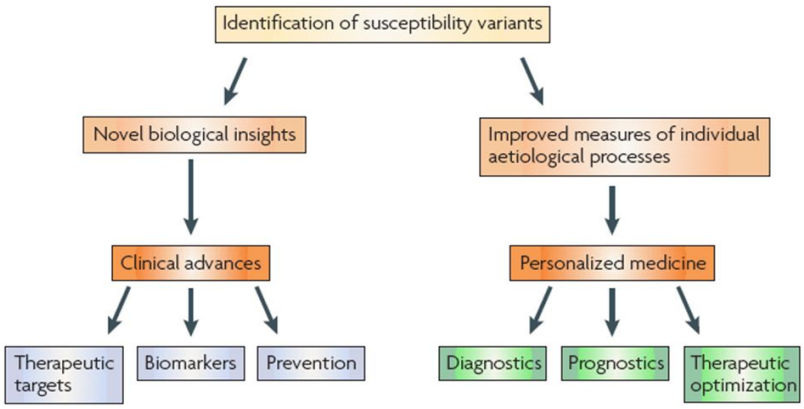

Figure 1.26: Recent successes in the identification of susceptibility variants have opened a debate about how to translate them to clinical practice. There are two principal routes through which such translation might be affected. Image obtained from [274].

In conclusion, GWAS are expected to discover many of the DNA variants giving rise to common diseases. A problem still to be handled is how to translate this knowledge to the clinical field. A general sketch [274] is shown in Figure 1.26. Even so, it is generally accepted that a high number of genetic markers associated with complex diseases will remain undiscovered. A graphical explanation [274] can be observed in Figure 1.27. Despite the

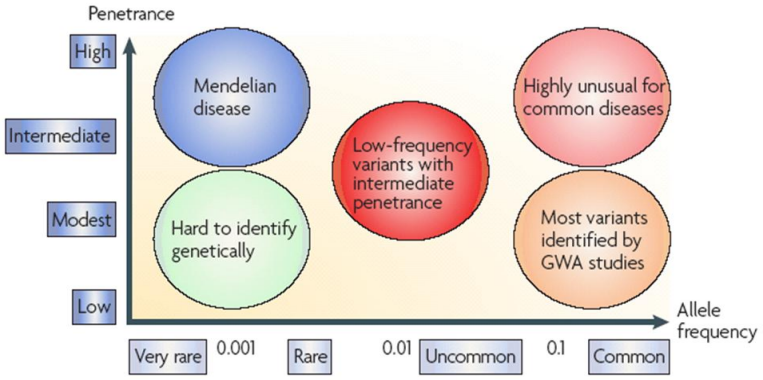

Figure 1.27: Sketch showing the characteristics of the different genetic markers depending on allele frequency and penetrance. Image obtained from [274].

large sample sizes obtained by international consortia, those markers with extremely low allele frequencies would need even larger ones. Estimate of the necessary sample sizes to detect SNP variants showing different degrees of risk (from low to moderate) is carried out in [104]. A straightforward graph can be seen in Figure 1.28.

1.3.1.6 Future of case–control association studies: copy number variants (CNVs)

It is still distant the moment when the field of genetic association studies could be thought to be finished. In this subsection we will focus on what seems to be a new approach to study variation across the genome: copy number variations (CNVs). Although a little advance was given in previous sections, here we will try to bring the reader closer to the current situation.

A copy number variation (CNV) is a segment of DNA in which copy–number differences have been found by comparison of two or more genomes. The segment may range from one kb to several Mb in size [79, 310]; in Figure 1.29 a histogram of the sizes of CNVs in two different databases can be observed. Humans ordinarily have two copies of each autosomal region, one per chromosome. This may vary for particular genetic regions due to deletion or duplication. CNVs may either be inherited or caused by *de novo* mutations. The fact that DNA copy number variation is a widespread and common phenomenon among humans was first shown up [200, 372] in the studies which leaded to the completion of the human genome project. It is estimated that approximately 0.4% of the genomes of unrelated people typically differ with respect to copy number [211]. In humans, CNVs encompass more DNA than single nucleotide polymorphisms (SNPs). Different articles trying to compile the inventory of CNVs along the genome have been

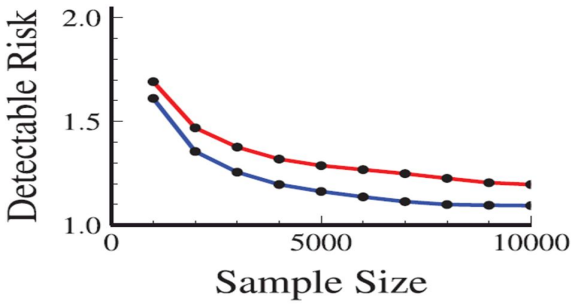

Figure 1.28: Minimum risk detectable with a 80% power and p–value $p = 0.05$ after a Bonferroni correction for several sample sizes. Multiplicative (red) and additive (blue) models are under study. Image obtained from [104].

already published [331]; nevertheless, completeness is still far to be reached.

Due to these reasons, CNVs are expected to have a strong contribution to common phenotypes of medical importance. CNVs associations studies have been already carried out related to different diseases: autism [428], schizophrenia [339], mental retardation [256], prostate cancer [250] and more. An interesting review can be found in [180].

Further research involving CNVs aims to convert intensity traces and SNP–based data into CNV genotypes [273]. New statistical tools, suitable for the specific features of CNV data, need to be developed, together with methods facilitating integration of CNV and SNP information.

1.3.2 Gene expression: state–of–the–art, challenges and expectations

1.3.2.1 Background

cDNA and oligonucleotide microarray techniques were developed to monitor the expression of many genes in parallel [359, 360]. They have been widely used for tumor diagnosis and classification, prediction of prognoses and treatment, and understanding of molecular mechanisms, biochemical pathways, and gene networks. Proper statistical analysis is vital to the success of array use. What makes microarray data analysis different from traditional statistics is the systematic biases inherent in the variations of experimental conditions and distinguishing features associated with the microarray outputs: high dimensionality (making simultaneous inferences on thousands of genes) and sparsity (only a small fraction of genes are statistically differentially expressed) [126]. Bioconductor (`www.bioconductor.org`)

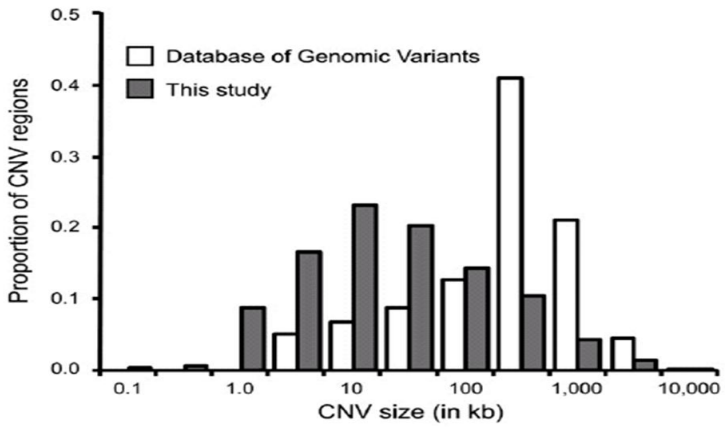

Figure 1.29: Size distribution of CNVs from the Database of Genomic Variants (`http://projects.tcag.ca/variation/`) and the study carried out in [310]. Image obtained from [310].

comprises most of the software packages developed to work with gene expression data. Figure 1.30 summarizes the working scheme of microarray studies and their most important areas of research.

Among these areas of research, our interest here will be focused on classification, which usually needs of variable selection. We will also present a brief approach to cluster problems. Time course and gene regulatory networks [383] also mean a huge work field for statistics [240]. Functional data analysis (FDA) [329] has been commonly used in such studies [173, 295], just as spline approaches [27, 257], hidden Markov models [362] and others.

So the two following subsections will differentiate between supervised (an outcome variable like disease status guides the learning process) and unsupervised (there are only features and no measurements of the outcome) learning.

Classification is included in supervised learning. The aim is usually to predict the outcome, detecting which variables are the most associated with the outcome. Gene expression datasets comprise measurements of thousands of genes for only a few dozens of individuals (the $p >> n$ problem). As a consequence, multicollinearity (strong correlations between genes) and overfitting are the two sticking points to be addressed by statistical methods [118]. Variable selection approaches are carried out to reduce dimensionality [18] and get interpretable models. Filtering [37, 95] and wrapping [145, 168] explicity select the variables (genes) to be employed by discrimination methods. Nevertheless, penalized regression methods [137, 183, 395] are likely among the most used nowadays. Here, variable selection is made implicitly to get sparse models where only a small fraction of the coefficients

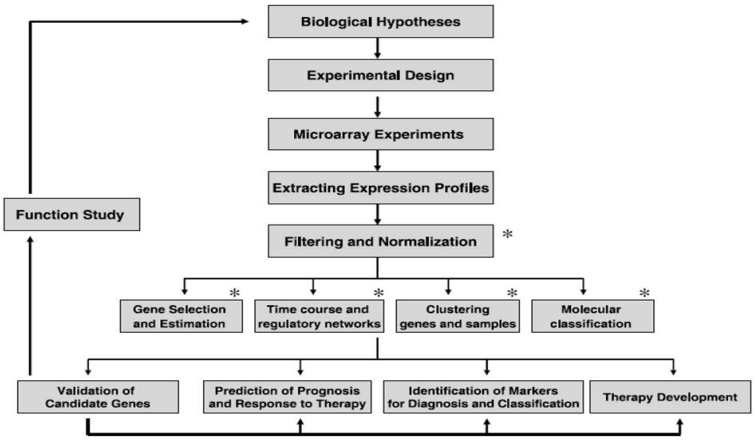

Figure 1.30: Schematic representation of microarray strategy. Steps where statistical analysis is needed are marked with an asterisk. Image obtained from [126].

$(\beta_1, \ldots, \beta_p)$ take nonzero values. Two interesting reviews compiling common statistical methods in gene expression studies can be found in [101, 370]. A penalized regression study with microarray data is properly shown in Chapter 3 of this essay.

Clustering genes is an unsupervised learning procedure, as there is no a response variable acting like a "teacher". In microarray studies clustering aims to reveal groups of genes which act together and whose collective expression follows similar patterns [90].

1.3.2.2 Supervised learning. Statistical classification and prediction

Classification and prediction with gene expression data is usually carried out in data where the response is binary (case–control, two disease subtypes,...) or, in any case, categorical. As opposed to SNP studies, most gene expression studies focus on different types of cancer: melanoma [36], hepatic cancer [236], leukemia [161], breast cancer [408], lymphoma [5], thyroid cancer [163] and others. A low proportion of studies are devoted to non–carcinogenic diseases: for instance, neuropathologies like postmortem Rett syndrome [76] or schizophrenia [284]. A few amount of databases [5, 6, 161] have been made public and their use has become common to compare different methods performed over them. Gene expression data needs sometimes to be preprocessed [101] to avoid redundancies and remove genes which contribution is null.

Traditionally, the aim of classification methods is to reduce misclassification results. However, two–class gene expression studies need to discover the group of genes associated with the response (expected to be small in most cases). Penalized regression methods for classification also emphasize the importance of reducing the bias and variance of the coefficient estimators.

A wide range of methodologies have been tried on expression data. Some of them are listed in [158]. The greatest challenge statistics have to face up with microarrays is the curse of dimensionality, as a result of scanning large regions of the genome. When evaluating a certain statistical technique is not only necessary to check its ability to classify and detect association, but also its efficiency and computational feasibility to manage high–dimensional data. Classification methods like random forests [93] or SVMs [385] have been used, due to their ability to deal with noise. Two–stage methods are also common [376]. A first stage is generally used to select the small amount of covariates that will classify on the second stage. Feature selection (FS) methods are abundant in the field, to discard most of the non associated genes. An interesting review of FS techniques in bioinformatics can be found in [346]. In contrast with two–stage methods, penalized regression approaches carry out variable selection and classification simultaneously, giving rise to sparse models where only a few genes have nonzero coefficients. This makes models interpretable from a biological point of view.

Penalization is usually carried out by means of an objective function like

$$L_P(\beta, \lambda) = L(\beta) + P(\beta, \lambda)$$

where λ is the penalization parameter, $P(\beta, \lambda)$ the penalization term and $L(\beta)$ is the log–likelihood of the sample for some method like logistic regression. Penalized logistic regression [189] is widely used due to the fact of having categorical variables as response.

The use of penalized regression methods with high–dimensional genetic data has increased lately. Different approaches [118, 195, 196, 214, 224] have been tried on gene expression data. These methods differ in their choice of the penalization term:

- Lasso (l_1) [395] penalizes the absolute value of the coefficients β_j

$$P(\beta, \lambda) = \lambda \sum_{j=1}^{p} |\beta_j|$$

- Ridge regression (l_2) [183] penalizes their squared values

$$P(\beta, \lambda) = \lambda \sum_{j=1}^{p} \beta_j^2$$

- Bridge regression (l_q, $0 < q < 1$) [137] applies a similar penalization to lasso, using the l_q norm

$$P(\beta, \lambda) = \lambda \sum_{j=1}^{p} |\beta_j|^q$$

- The elastic net (l_e) [452] is a convex combination of lasso (l_1) and ridge (l_2) penalizations, $l_e = (1 - \alpha)l_1 + \alpha l_2$ with $0 < \alpha < 1$

$$P(\beta, \lambda) = \lambda \left[(1 - \alpha) \sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j=1}^{p} \beta_j^2 \right]$$

All but ridge give rise to sparse models. Our work here will be focused on lasso. Lasso has been largely studied in high–dimensional data contexts [151, 196, 214, 277, 451]. Chapter 3 in this essay is devoted to show a lasso logistic regression approach to work with gene expression data.

1.3.2.3 Unsupervised learning. Cluster analysis

Good clustering algorithms are very much desired for analyzing data with high dimension where the number of variables is considerably larger than the number of observations. DNA microarray analysis is a tipical example that involves such high–dimensional data. Clustering microarray data can be very helpful for certain types of biological studies, such as cancer research. For example, based on the gene expression profiles, interesting cluster distinctions can be found among a set of tissue samples, which may reflect categories of diseases, mutational status or different responses to a certain drug [419]. Biological pathways, understood as groups of genes working together to carry out the same task, are often recognizable by means of cluster analysis [55]. Anyway, in gene expression studies it is not always clear which type of information is wanted to be clusterized: samples [182] or genes [90]. A wide range of objectives are pursued when analyzing gene expression data. A list can be found in [263]:

1. Grouping of genes according to their expression under multiple samples.

2. Classification of a new gene, given the expression of other genes, with known classification.

3. Grouping of samples based on the expression of a number of genes.

4. Classification of a new sample, given the expression of the genes under some experimental conditions.

Among the clustering methods used, there are obviously several *ad hoc* procedures, and also common cluster approaches like hierarchical clustering [175, 422], k–means clustering [136, 253], self–organizing maps (SOM) [216, 217, 218], etc. Ensemble methods like bagging have been also proposed [100] to improve the performance of clustering procedures. A review of cluster analysis carried out with gene expression data can be found in [450]. Figure 1.31 shows the results from clustering samples in a case–control study of the primary Sjogren's syndrome [182]. Red and green indicate higher and lower expressions, respectively.

However, the results from the application of standard clustering methods to genes are limited. For this reason, a number of algorithms that perform simultaneous clustering on the row and column dimensions of the data matrix has been proposed [71, 233, 377]. The goal is to find submatrices, that is, subgroups of genes and subgroups of samples, where the genes exhibit highly correlated activities for every sample. This has been called bicluster analysis [71]. Reviews and comparisons of different biclustering algorithms can be found in [263, 322].

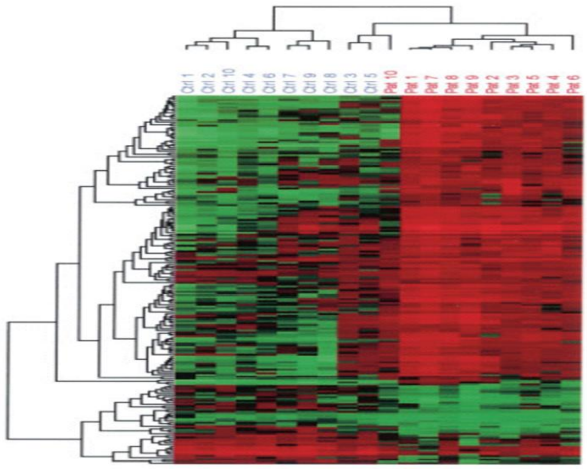

Figure 1.31: Hierarchical clustering of data from the microarray analysis of gene expression in minor salivary glands from patients with primary Sjogren's syndrome and from healthy control subjects. Samples with similar patterns of expression of the genes studied will cluster together, as indicated by the dendrogram. The hierarchical clustering of the 200 genes that were most differentially expressed in patients vs. the healthy controls is illustrated. Image obtained from [182].

1.4 Statistical tools in non–clinical genetics: population genetics and forensic genetics

1.4.1 Sets of markers

1.4.1.1 Short Tandem Repeats (STRs)

As explained previously (see Section 1.1.5), Short Tandem Repeats (STRs) are genetic markers consisting of tandemly repeated sequences, between 2 and 10 bp in length, which exhibit a high degree of length polymorphism due to variation in the number of repeat units. Analysis of STR sequences has become the standard method in forensic identification nowadays.

To select a set of STR markers to be used in the forensic casework relative to a certain population, the most important measures to be taken into account are their combined power of exclusion and their combined power of discrimination between individuals. These measures, to be explained in the next subsection, depend on STR features such as polymorphic nature, mutation rates or independence between the different markers of the set. There is a wide choice of STR loci. Choice of the right ones for human identification cases is critical [51, 406]. Some STRs are more useful than others in forensic analysis, as they produce less amplification artifacts, or simply their polymorphic structure in the population under study is more

suitable or their heterozygosity is higher.

STR markers were first described as effective tools for human identity testing in the early 1990s [105, 106]. Early successes in STR typing were obtained in the U.K. [213] and Canada [138]. After that, the FBI Laboratory's Combined DNA Index System (CODIS) selected a core of 13 STR marked to be used in U.S.A. (`www.fbi.gov/hq/lab/codis/index1.htm`). Some of these STRs are displayed on Figure 1.32, where they are showed in terms of chromosomal location. The same set was subsequently adopted in other populations [58, 154, 207].

Throughout time, new STRs have been discovered as useful in forensic cases [231]. Technical guidelines for validation of STR markers can be found in [260]. Currently, every day new sets of STR markers are proposed to be used with forensic purposes in different populations or subpopulations [166]. A very interesting review about STRs for human identity testing is given in [59].

1.4.1.2 Commercial kits of STRs

The STR Project was an initiative launched in 1996, with the aim of knowing the best STR systems in forensic studies. Since the beginning, important companies like Promega Corporation (Madison, WI) or Applied Biosystems (Foster City, CA) got involved. As a logical consequence, these two companies have been responsible for the development of the great majority of the STR commercial kits used in forensic casework [188, 241, 247, 415]. Table 1.4 summarizes the various STR kits that have become available in the past decade.

Since the turn of the century, new multiplex assays have been developed that amplify all 13 CODIS core loci in a single reaction. Nowadays, there are esentially two available commercial kits standing out over the rest. These

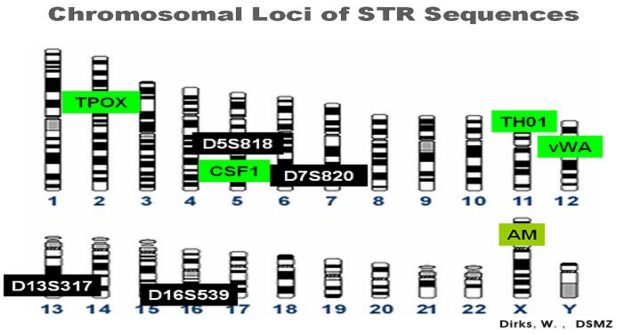

Figure 1.32: STR locus used in the PowerPlex1.2 system and their approximated chromosomal locations. Image obtained from `http://cellbank.nibio.go.jp/str2/str_locus.html`.

are:

- The PowerPlex 16 kit was released by the Promega Corporation in 2000, amplifying the 13 core loci with amelogenin (to determine gender) and two pentanucleotide loci (Penta D and Penta E) [222].

- The 16plex Identifiler kit was released by Applied Biosystems in 2001, amplifying the 13 core loci with amelogenin and two tetranucleotide loci (D2S1338 and D19S433) [78].

Other forensic kits have been developed to be used in special cases (e.g. degraded samples). New STR loci, regarded as very important in forensic studies due to their high polymorphic nature (e.g. SE33 [365, 431]) have been added to the STR kits or directly in forensic applications. Notwithstanding, use of the two commercial kits above have not decreased throughout time.

1.4.1.3 Use of SNPs in forensic cases

SNPs have a number of characteristics that make them ideal markers for human identification. Some of them are listed in [354]:

Kit Name	STR Loci Included	Random Match Probability with Author's Profile*
	Promega Corporation	
PowerPlex 1.1 and 1.2	CSF1PO, TPOX, TH01, VWA, D16S539, D13S317, D7S820, D5S818	7.4×10^{-10}
PowerPlex 2.1 (for Hitachi FMBIO users)	D3S1358, TH01, D21S11, D18S51, VWA, D8S1179, TPOX, FGA, Penta E	3.4×10^{-11}
PowerPlex ES	FGA, TH01, VWA, D3S1358, D8S1179, D18S51, D21S11, SE33, amelogenin	1.3×10^{-10}
PowerPlex 16	CSF1PO, FGA, TPOX, TH01, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, Penta D, Penta E, amelogenin	1.2×10^{-18}
PowerPlex 16 BIO (for Hitachi FMBIO users)	CSF1PO, FGA, TPOX, TH01, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, Penta D, Penta E, amelogenin	1.2×10^{-18}
	Applied Biosystems	
AmpFlSTR Blue	D3S1358, VWA, FGA	1.0×10^{-3}
AmpFlSTR Green I	Amelogenin, TH01, TPOX, CSF1PO	7.8×10^{-4}
AmpFlSTR Cofiler (CO)	D3S1358, D16S539, Amelogenin, TH01, TPOX, CSF1PO, D7S820	2.0×10^{-7}
AmpFlSTR Profiler Plus (Pro)	D3S1358, VWA, FGA, Amelogenin, D8S1179, D21S11, D18S51, D5S818, D13S317, D7S820	2.4×10^{-11}
AmpFlSTR Profiler Plus ID	D3S1358, VWA, FGA, Amelogenin, D8S1179, D21S11, D18S51, D5S818, D13S317, D7S820 (extra unlabeled D8-R primer)	2.4×10^{-11}
AmpFlSTR Profiler	D3S1358, VWA, FGA, Amelogenin, TH01, TPOX, CSF1PO, D5S818, D13S317, D7S820	9.0×10^{-11}
AmpFlSTR SGM Plus (SGM)	D3S1358, VWA, D16S539, D2S1338, Amelogenin, D8S1179, D21S11, D18S51, D19S433, TH01, FGA	4.5×10^{-13}
AmpFlSTR Sefiler (SE)	FGA, TH01, VWA, D3S1358, D8S1179, D16S539, D18S51, D21S11, D2S1338, D19S433, SE33, amelogenin	5.1×10^{-15}
AmpFlSTR Identifiler (ID)	CSF1PO, FGA, TPOX, TH01, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, D2S1338, D19S433, amelogenin	7.2×10^{-19}

Table 1.4: Summary of available commercial STR kits that are commonly used. The random match probabilities measure, in some way, the power of discrimination obtained with each kit. * Allele frequencies used for random match probability calculations (to unrelated individuals) from U.S. Caucasian population data. Image obtained from [59].

1. They have lower mutation rates than the STR and VNTR (variable number tandem repeat) loci typically used for relationship analysis in paternity and immigration testing.

2. SNPs are preferable for anthropological and crime case investigations where the DNA is often degraded, due to technical issues.

3. SNPs can be genotyped with a growing range of high–throughput technologies.

4. As binary polymorphisms, are comparatively easy to validate, because precise allele frequency estimates, required for the accurate interpretation of forensic genotyping data, can be obtained by analysing fewer samples compared to those needed for allele frequencies estimates of STRs and VNTRs.

These characteristics have made SNPs a very nice alternative, or complement, to the standard use of STRs in forensic genetics. Nevertheless, there are also some points making SNPs less attractive. For instance, seeking to match the discriminatory power of the 10–15 multiple allele STRs routinely used in forensic investigations, a set of about 50 polymorphic SNP markers are required [11, 153]. Furthermore, as happened with STRs, SNPs that are polymorphic in one population may be almost or completely monomorphic in another population [311, 379]. Thus, it should be possible to select SNPs that are useful for human identification purposes in the majority of populations, and to supplement these with SNPs showing highly contrasting allele frequency distributions in particular populations.

The SNPforID group (`www.snpforid.org`) is a consortium supported by the EU GROWTH programme with the following objectives:

(i) Selection of at least 50 autosomal SNPs suitable for the identification of persons of unknown population origin and determination of allele frequencies in the major population groups.

(ii) Development of a highly efficient DNA amplification strategy for the simultaneous analysis of up to 50 independent SNPs in a single assay.

(iii) Assessment of automated, high–throughput DNA–typing platforms for reliable and accurate multiplex SNP typing.

(iv) Assessment of the forensic application of the high–throughput SNP–typing methods developed.

A set of 52 unlinked autosomal SNPs (52plex) that are highly polymorphic in European, Asian and African populations is presented in [354]. Discrimination power of this 52plex was tested in different European, Asian

	European	Somali	Asian
Mean match probability	5.0×10^{-21}	1.1×10^{-19}	5.0×10^{-19}
Power of discrimination	>99.99999%	>99.9999%	>99.9999%
Mean exclusion probability	99.98%	99.95%	99.91%
Typ. paternity index (trios)	549000	337000	336000
Typ. paternity index (duos)	4640	3160	2880

Table 1.5: Various discrimination power measures in forensic for the 52plex. Data obtained from [354].

and African populations. Some statistical results about the discrimination power of this 52plex in forensic cases are given in Table 1.5.

Until now, only a few large SNP multiplexes have been reported [96], but larger multiplexes that are constructed based on the same principles as the present 52plex are emerging, e.g. packages with Y chromosome SNPs [53, 353] or autosomal SNPs with contrasting allele frequency distributions in different populations useful for the estimation of the population of origin [315].

1.4.2 Common statistics

Probability theory has been used for a long time in forensic casework. Concepts used are not very complex and do not need a strong mathematical background, so a lot of people can understand them. Bayes probability theory and some ideas about heritage shape the base to understand the majority of problems that can be found in forensics. Although there are different forensic problems, all the probabilistic approaches to solve them are similar [64, 124]. Therefore, here we only show two of the most common ones: criminal cases and paternity tests.

1.4.2.1 Criminalistic cases

Let us place ourselves in a criminal case: a crime has been committed and a human sample (blood, hair,...) has been collected at the crime scene. Later, a suspect is arrested. Genetic profiles from the sample and the suspect are obtained to subsequently observe if they match. In that case, we call G to this common genetic profile.

To calculate a probability measure of the suspect being the real criminal, we have to define the events. We call E to the scientific evidence obtained. This event E consist of two events, namely, S_G, meaning the sample at the crime scene has genetic profile G and E_G, meaning the suspect has genetic profile G. Here, we will focus on the simplest case of having only one suspect and one sample. So, this problem can be identified as a hypothesis testing problem, where the null (H_0) and the alternative (H_1) hypotheses are:

H_0: the suspect is guilty.

H_1: another randomly picked individual different from the suspect is guilty.

Bayes theorem provides a way to calculate the posterior probability $P(H_0|E)$:

$$\begin{aligned} P(H_0|E) &= \frac{P(E|H_0)P(H_0)}{P(E)} \\ &= \frac{P(E_G|H_0)P(S_G|E_G,H_0)P(H_0)}{P(E_G)P(S_G|E_G)} \end{aligned}$$

Despite in the formula $P(E_G|H_0)$ is expressed as a conditional probability, E_G and H_0 are independent events, so $P(E_G|H_0) = P(E_G)$. On the other hand, $P(S_G|E_G, H_0)$ is the probability of the sample having the genetic profile G, assuming it was produced by the suspect, and the suspect has genetic profile G, so it is obvious that $P(S_G|E_G, H_0) = 1$. Value of the prior probability $P(H_0)$ has to be decided from other nongenetic evidences.

Nevertheless, the International Society for Forensic Genetics (ISFG) recommendations on biostatistics [157, 290] suggest that the biological evidence should be based on likelihood ratio (LR) principles. The likelihood ratio is defined as the quotient between the probability of the evidence E assuming H_0 and the probability of E assuming H_1:

$$\text{LR} = \frac{P(E|H_0)}{P(E|H_1)}$$

This can be expressed in terms of the posterior probability $P(H_0|E)$ as follows,

$$\text{LR} = \frac{P(H_0|E)(1-P(H_0))}{(1-P(H_0|E))P(H_0)}$$

The LR is the usual way to express bets and it is also the only result communicated to a judge in a courtroom, so this value is going to be fundamental facing imprisonment or freedom for a certain individual.

1.4.2.2 Paternity and relationship tests

To properly explain the role of statistics in paternity tests we will place ourselves in one of the most common cases: paternity between an individual (F) and a child (S) is tested with no more information than their genotypes.

Scientific evidence is now $E = (S_{GS}, F_{GF})$, where S_{GS} means the son has the genetic profile GS and F_{GF} means the individual F has the genetic profile GF. The null (H_0) and alternative (H_1) hypotheses are now:

H_0: F and S have the relationship father–son.

H_1: F and S have not the relationship father–son.

This problem can become more complex if the alternative hypothesis H_1 is divided in different hypotheses, like, for example, F and S have another kind of relationship and F and S are not relatives at all.

The posterior paternity probability by Bayes theorem is:

$$\begin{aligned} P(H_0|E) &= \frac{P(S_{GS}|F_{GF},H_0)P(F_{GF}|H_0)P(H_0)}{P(S_{GS}|F_{GF})P(F_{GF})} \\ &= \frac{P(S_{GS}|F_{GF},H_0)P(H_0)}{P(S_{GS}|F_{GF})} \end{aligned}$$

The prior probability $P(H_0)$ takes a certain value based on nongenetic evidences for or against paternity. It is also common to take $P(H_0) = 0.5$, indicating statistics are not inclined towards neither of the options (paternity or not). The remaining probabilities are calculated from coincidences and exclusions between the genotypes GS and GF, mutation rates,.... Anyway, as happened with criminal cases, evidence of paternity is based on LR results. Likelihood ratio is referred in paternity testing as the paternity index (PI). Therefore, the PI formula is:

$$\begin{aligned} \text{PI} &= \frac{P(E|H_0)}{P(E|H_1)} = \frac{P(S_{GS},F_{GF}|H_0)}{P(S_{GS},F_{GF}|H_1)} \\ &= \frac{P(S_{GS}|F_{GF},H_0)P(F_{GF}|H_0)}{P(S_{GS}|F_{GF},H_1)P(F_{GF}|H_1)} \\ &= \frac{P(S_{GS}|F_{GF},H_0)}{P(S_{GS}|F_{GF},H_1)} \end{aligned}$$

The meaning of the PI value obtained in a paternity test need always to be explained by an expert. Both Evett [123] and Carracedo and Barros [63] fixed the evidence strength from the PI values in the terms given in Table 1.6.

PI	Evidence strength
1–33	Weak
33–100	Fair
100–330	Good
330–1000	Strong
>1000	Very strong

Table 1.6: Approximated evidence strength for different ranges of PI values in paternity testing. Data obtained from [63, 123].

1.4.3 Intricate problems in forensic and population genetics

Paternity tests are subject to several problems, more or less complex. Seriousness of these problems becomes bigger due to everything around this kind of tests: violation cases, inheritances, feelings, etc. Regarding the scientific ones, a list of some of them includes:

- Thresholds of decision. Once a LR–PI or probability result has been reached, we need to know from which value we can guarantee paternity/guilt. This has been a matter of discussion in forensic genetics during many years [63].
- The choice of a correct value for the prior probability $P(H_0)$ has been always critical, as this value determines in some way the final result obtained.
- The choice of the reference population is not so easy as we could possibly think, because sometimes what is thought to be a homogeneous population is composed of different populations in which allelic frequencies differ substantially. This is called population stratification and, if it is not deal with it properly, it can give rise to erroneous results, as it is shown in [401]. Each homogeneous population should be represented by a suitable database summarizing the allelic frequencies profile inside this population.
- Inbreeding is also a source of problems in paternity tests. Many times, paternity cases become complex because the individual to be tested as father is not available (disappearance, death, ...) and a relative (brother, father, ...) is required. This kind of cases are specifically studied in [314] (see also Chapter 6).
- Highly degraded DNA cases, in which DNA results are difficult to obtain. SNPs are very useful in those cases as they offer greater success than standard STRs with highly degraded DNA [133, 314, 354].

Chapter 2

Motivation and aims

Since its birth some centuries ago, due to the need to understand the mechanics of some gambling games, statistics have evolved and their use has spreaded to other areas and subjects with not so markedly economical purposes, and yet more scientific ones. Finances, medicine, physics, etc. (and obviously gambling) have used statistical tools to obtain knowledge from data.

Genetics is the term used, since Gregor Mendel's experiments more than a century ago, to refer to the science studying heritage patterns and how they are expressed. To comprehend heritage and its mechanisms is expected to offer many of the answers mankind has been looking for along centuries.

So this combination of gambling and peas is what it is today called statistical genetics. This essay is a compendium of statistical approaches and applications to different problems in different branches of genetics. This small chapter aims to offer the reader a general idea about the current situation and the motivation to carry out this work.

Research in genetics is continuously growing and evolving, as the hope to find answers to many diseases (cancer, psychiatric disorders, diabetes, ...) in this field remains intact. Furthermore, the scope of genetics goes beyond clinical application. Forensic genetics are fundamental nowadays regarding successful resolution of criminal cases or parental studies. Population genetics constitute a very powerful tool to investigate human migrations. Many other uses could be listed, but let us stick here to the scope of this essay.

High–throughput technologies and cost reductions are producing a huge mass of data that needs to be analyzed. Statistics have to provide with the appropriate tools to efficiently discover the genetic patterns emerging. Anyway, the economic costs required to carry out competitive studies are sometimes difficult to take. Simulated data can be often a good replacement, specially when evaluating the abilities of different statistical methods in several scenarios.

Data types produced in genetic research are esentially two–fold nowa-

days:

1. Gene expression measurements usually move in continuous ranges. However, categorization has been carried out in a reduced number of studies, with the aim of simplyfying gene expression patterns. Anyway, this is not a common approach and it is not often recommended, due to the loss of information that entails. Therefore, from now on, we will consider gene expression data as continuois data.

2. Nuclear DNA is composed by two nucleotidic chains, containing each the information inherited from parents. In each chain, and for each position (locus), this information is binary (two options), so the combination of the information in each physical position of both chains will still have a very limited number of possibilities, namely, categorical data. This is applicable not only to SNP data but also to the common markers used in forensics.

So it seems clear that genetics are a huge battlefield for statistical tools, not only due to this variety of data types, but also because genetic studies pursue different aims, even in a branch like clinical genetics: discovering of disease association, gene functions, gene pathways, new disease subtypes based on genetic patterns, etc.

The rest of this essay is organized as follows. Chapter 3 contains a penalized regression approach to get sparse, genetically interpretable, models in case–control genetic studies. A support vector machine (SVM) adaptation to SNP data can be found in Chapter 4. In Chapter 5, two different evaluations of statistical methods in SNP case–control association studies, with emphasis on tree–based methods, cover Sections 5.1 and 5.2. Chapters 3 to 5 could be considered a set of several methods for classification/prediction with different types of genetic datasets. Statistical tools for forensic and population genetics are revised in Chapter 6, together with the use of intensive simulation to solve intricate problems. Finally, Chapter 7 is devoted to general conclusions about this work and Chapter 8 lists several lines of further research.

Chapter 3

Penalized regression in gene expression studies: lasso logistic regression to obtain sparsity

3.1 Lasso logistic regression, GSoft and the cyclic coordinate descent algorithm. Application to gene expression data

This chapter consists of a penalized regression study to be applied to continuous gene expression data. This work has been recently finished and it is now in process to be submitted soon to an international journal. The great majority of it was made during a three months stay of the author in the Laboratoire Jean Kuntzmann of the Universite Joseph Fourier in Grenoble (France) with the professor Anestis Antoniadis.

3.1.1 Abstract

Statistical methods generating sparse models are of great value in the gene expression field, where the number of covariates (genes) under study moves about the thousands, while the sample sizes seldom reach a hundred of individuals. For phenotype classification, we propose different lasso logistic regression approaches with specific penalizations for each gene. These methods are based on a generalized soft–threshold (GSoft) estimator. We also show that a recent algorithm for convex optimization, namely the cyclic coordinate descent (CCD) algorithm, provides with a fast way to solve the optimization problem posed in GSoft. Re-

sults are obtained for simulated and real data. The leukemia and colon datasets are commonly used to evaluate new statistical approaches, so they come in useful to establish comparisons with similar methods. Furthermore, biological meaning is extracted from the leukemia results, and compared with previous studies. In summary, the approaches presented here give rise to sparse, interpretable models, competitive with similar methods developed in the field.

3.1.2 Introduction

Advent of high–dimensional data in several fields (genetics, text categorization, combinatorial chemistry,...) is an outstanding challenge for statistics. Gene expression data is the paradigm of high–dimensionality, usually comprising thousands (p) of covariates (genes) for only a few dozens (n) of samples (individuals). Feature selection in regression and classification is then fundamental to get interpretable, understandable models, which might be of use to the field. First approaches to this problem [167, 170, 237, 430] were based on filtering to select a subset of covariates related with the outcome, usually a binary response. Nevertheless, common methods developed nowadays search for variable selection and classification carried out in the same step. Sparse models are needed to account for high–dimensionality (the $p >> n$ problem) and strong correlations between covariates.

Penalized regression methods have received much attention over the past few years, as a proper way to get sparse models in those fields with large datasets. Lasso [395] was originally proposed for linear regression models, and subsequently adapted to the logistic case [344, 378]. Lasso applies a l_1 penalization that, as opposed to ridge regression [183], gives rise to sparse models, ruling out the influence of most of the covariates on the response. Consistency properties of lasso for the linear regression case have been full well studied [215, 259, 278, 444, 446]. An evolution of lasso that allows for specific penalizations in the l_1 penalty (adaptive lasso) is developed in [451]. Lasso has been also adapted to work with categorical variables [16, 25, 277, 442] and multinomial responses [224]. Other penalized regression methods include bridge estimators [137], which replace the l_1 penalization with l_q penalization, being $0 < q < 1$, and the elastic net [452], that penalizes by means of a linear combination of l_1 and l_2 penalties. Consistency studies about bridge and elastic net can be found in [194] and [87], respectively. Application of both approaches to high–dimensional genetic data is carried out in [252]. Optimization of the lasso log–likelihood function is also an important subject of study [238, 363], as a result of the non–differentiability problems of the l_1 penalty around zero.

In this study, we adopt an adaptive lasso logistic regression approach based on the generalized soft–threshold estimator (GSoft) [214]. A theoret-

ical connection between existence of solution in GSoft and convergence of the cyclic coordinate descent (CCD) algorithm [447] is established, allowing the solutions obtained with the latter to take advantage of the asymptotic properties of the former. We try different vectors Γ for the specific penalization of each covariate (gene) and some consistency results [196] are shown for each one. Extensive comparisons with similar approaches are carried out using simulated and real microarray data.

The rest of this chapter is organized as follows: a short introduction about the CCD algorithm, GSoft and some of its asymptotic properties is given in Section 3.1.3, together with the theoretical connection between both and the three different Γ choices for the specific penalizations. Some consistency results for each one are added. Results of simulated and real data are shown in Section 3.1.4. Simulations include approximations of the variance–covariance matrix for the estimated coefficients. Real data includes leukemia [161] and colon [6] datasets. Finally Section 3.1.5 is devoted to conclusions, and the Appendix A contains the proof of Theorem 2.

3.1.3 Methods

Our aim is to learn a binary gene expression classifier $y_i = f(\mathbf{x}_i)$ from a set $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ of independent and identically distributed observations. In each sample i, the vector

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \in \mathbb{R}^p$$

comprises gene expression measurements. The $n \times p$ design matrix is then $X = (\mathbf{x}_j, j \in \{1, \ldots, p\})$ where the $\mathbf{x}_j$'s represent the expression measurements of gene j along the entire set of samples. The vector of binary responses

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

informs about membership (+1) or nonmembership (-1) of the sample to the category. The logistic regression model with vector of regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ assumes that

$$P(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{x}_i'\boldsymbol{\beta})}.$$

Adopting a generalized linear model framework, the associated linear predictor $\boldsymbol{\eta}$ is defined as

$$\boldsymbol{\eta} = X\boldsymbol{\beta} = \begin{pmatrix} \mathbf{x}_1^{'}\boldsymbol{\beta} \\ \vdots \\ \mathbf{x}_n^{'}\boldsymbol{\beta}, \end{pmatrix} \text{ where } X = \begin{pmatrix} \mathbf{x}_1^{'} \\ \vdots \\ \mathbf{x}_n^{'} \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}.$$

The decision of whether to assign the i sample to the category or not is usually accomplished by comparing the probability estimate with a threshold (e.g. 0.5). Consequently, minus the log–likelihood function is

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln \left[1 + \exp(-y_i \mathbf{x}_i^{'}\boldsymbol{\beta})\right] \tag{3.1}$$

The lasso like logistic estimator $\hat{\boldsymbol{\beta}}$ with specific penalizations for each covariate is then given by the minimizer of the function

$$L_1(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \gamma_j |\beta_j| \tag{3.2}$$

where λ is a common nonnegative penalty parameter and the vector $\boldsymbol{\Gamma} = (\gamma_1, \ldots, \gamma_p)$ with nonnegative entries penalizes each coefficient. The standard lasso regularization [395] takes $\gamma_j = 1\ \forall j$. Minimization of these objective functions makes use of their derivatives. We refer to the gradient of $L(\boldsymbol{\beta})$ as the score vector whose components are defined by:

$$s_j(\boldsymbol{\beta}) = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j}$$

The negative Hessian with respect to the linear predictor η is defined as

$$H(\eta) = -\frac{\partial^2 L(\eta)}{\partial \eta \partial \eta^{'}}$$

The basic requirement for the weights γ_j is that their value should be large enough to get $\hat{\beta}_j = 0$ if the true value β_j is zero, and small otherwise. Obtaining of a sparse, interpretable model is of paramount importance in those areas where the number of variables usually outperforms the sample size ($p >> n$ problem). The choice of the $\boldsymbol{\Gamma}$ vector is therefore essential to get an accurate estimator $\hat{\boldsymbol{\beta}}$.

3.1.3.1 Cyclic coordinate descent (CCD) algorithm

The choice of a proper algorithm to solve the minimization of (3.2) is a main issue, as it needs to be capable of dealing with the problem of non–differentiability of the absolute value function around zero. Furthermore, efficiency of the algorithm is fundamental, given the high–dimensionality of the problems at hand.

A number of different algorithms have been developed to obtain the optimum for the objective function. In [160] a "Split–Bregman" method is applied to solve l_1-regularized problems, while in [435] an algorithmic framework for minimizing the sum of a smooth convex function with a non-smooth nonconvex one is proposed. A similar algorithm is used in [212] to obtain the solution for the SCAD estimator in high–dimensions. Two new approaches are developed in [363], together with a comparative study. An efficient algorithm is carried out in [238], using LARS [110] in each iteration. A local linear approximation (LLA) algorithm was recently proposed by [453], while [416] developed a method of least squares approximation (LSA) for lasso estimation, making use of the LARS algorithm.

Finding the estimate of $\boldsymbol{\beta}$ is a convex optimization problem. The cyclic coordinate descent algorithm is based on the CLG algorithm of Zhang and Oles [447]. An exhaustive description of the algorithm is beyond the scope of this paper, and interested readers are referred to the detailed description in [151]. The basis of all cyclic coordinate descent algorithms is to optimize with respect to only one variable at the time while all others are held constant. When this one–dimensional optimization problem has been solved, optimization is performed with respect to the next variable, and so on. When the procedure has gone through all variables it starts all over with the first one again, and the iterations proceed in this manner until some pre–defined convergence criterion is met. The one–dimensional optimization problem is to find β_j^{new}, the value for the j–th parameter that maximizes the penalized log–likelihood assuming that all other β_j's are held constant. In the end, the update equation for β_j becomes

$$\beta_j^{new} = \begin{cases} \beta_j - \Delta_j & \text{if } \Delta v_j < -\Delta_j \\ \beta_j + \Delta v_j & \text{if } -\Delta_j \leq \Delta v_j < \Delta_j \\ \beta_j + \Delta_j & \text{if } \Delta_j < \Delta v_j \end{cases}$$

where the interval $(\beta_j - \Delta_j, \beta_j + \Delta_j)$ is an iteratively adapted trust region for the suggested update Δv_j. The width of this interval is determined based on its previous value and the previous update made to β_j. The suggested update is given by

$$\Delta v_j = -\frac{s_j(\boldsymbol{\beta}) - \lambda\gamma_j \text{sign}(\beta_j)}{Q(\beta_j, \Delta_j)} \tag{3.3}$$

The essential idea in CCD is $Q(\beta_j, \Delta_j)$ to be an upper bound on the second derivative of $L_1(\boldsymbol{\beta})$ in the interval around β_j:

$$\frac{\partial^2 L_1(\boldsymbol{\beta})}{\partial \beta_j^2} = \sum_{i=1}^{n} \frac{x_{ij}^2 \exp(-y_i \mathbf{x}_i' \boldsymbol{\beta})}{\left[1 + \exp(-y_i \mathbf{x}_i' \boldsymbol{\beta})\right]^2}$$

The function $Q(\beta_j, \Delta_j)$ is given by the expression:

$$Q(\beta_j, \Delta_j) = \sum_{i=1}^{n} x_{ij}^2 F(y_i \mathbf{x}_i' \boldsymbol{\beta}, \Delta_j x_{ij})$$

with the function F being defined by

$$F(B, \delta) = \begin{cases} 0.25 & \text{if } |B| \leq |\delta| \\ [2 + \exp(|B| - |\delta|) + \exp(|\delta| - |B|)]^{-1} & \text{otherwise.} \end{cases}$$

A proof of Q being an upper bound in the aforementioned interval is straightforward. Advantages of CCD can be summarized in efficiency of the algorithm, stability and ease of implementation. Efficiency is due to several factors: CCD works following a cycling procedure along the coefficients. From a certain iteration, CCD only visits the active set, reducing considerably its computational demands. Implementation has been carried out by means of the R package *glmnet*. This approach is explained in [143], where it is proved to be faster than its competitors.

3.1.3.2 GSoft

The generalized soft–threshold estimator or GSoft [214] is claimed to be a compromise between approximately linear estimators and variable selection strategies for high dimensional problems. Our interest in GSoft lies in the fact that once a solution $\boldsymbol{\beta}$ exists, a bunch of asymptotic properties can be derived. The next theorem from [214] establishes necessary and sufficient conditions for the existence of such solution.

Theorem 1. *The following set of conditions is necessary and sufficient for the existence of an optimum* $\hat{\boldsymbol{\beta}}$ *of* $L_1(\boldsymbol{\beta})$

(a)

$$\left\{ \begin{array}{ll} |s_j(\beta)| \leq \lambda\gamma_j & \text{if } \beta_j = 0 \\ s_j(\beta) = \lambda\gamma_j & \text{if } \beta_j > 0 \\ s_j(\beta) = -\lambda\gamma_j & \text{if } \beta_j < 0 \end{array} \right\}$$

(b)

$$X_\lambda' H(\eta) X_\lambda \text{ is positive definite,}$$

where X_λ *retains only those columns (covariates)* $\mathbf{x}_j$ *of* X *fulfilling* $|s_j(\boldsymbol{\beta})| = \lambda\gamma_j$*, that is,* $X_\lambda = (\mathbf{x}_j, |s_j(\boldsymbol{\beta})| = \lambda\gamma_j)$*.*

Approximation of the covariance matrix for the estimated coefficients

Approximations to the variance–covariance matrix of $\hat{\boldsymbol{\beta}}$ have to deal with the non–differentiability problem of the penalization term around $|\beta_j| = 0$. This fact is solved by taking a differentiable approximation $a(\beta_j, \delta)$ to the absolute value function, obtained by smoothing it around zero

$$a(\beta_j, \delta) = \begin{cases} |\beta_j| & \text{if } |\beta_j| > \delta \\ \frac{(\beta_j^2+\delta^2)}{2\delta} & \text{if } |\beta_j| \leq \delta \end{cases},$$

with $\delta > 0$ and satisfying $\lim_{\delta \to 0} a(\beta_j, \delta) = |\beta_j|$.

So an approximation can be constructed from the well–known sandwich form developed in [197]

$$V_\delta(\hat{\boldsymbol{\beta}}) = \left\{ H(\hat{\boldsymbol{\beta}}) + \lambda \boldsymbol{\Gamma} G\left(\hat{\boldsymbol{\beta}}, \delta\right) \right\}^{-1} \text{Var}\left\{ s(\hat{\boldsymbol{\beta}}) \right\} \left\{ H(\hat{\boldsymbol{\beta}}) + \lambda \boldsymbol{\Gamma} G\left(\hat{\boldsymbol{\beta}}, \delta\right)) \right\}^{-1}$$

where $H(\hat{\boldsymbol{\beta}})$ is the negative Hessian of L but now as a function of $\hat{\boldsymbol{\beta}}$ and G is the diagonal matrix made up of the second derivatives of the approximations $a(\beta_j, \delta)$:

$$G(\boldsymbol{\beta}, \delta) = \text{diag}\left(\frac{I\{|\beta_1| \leq \delta\}}{\delta}, \ldots, \frac{I\{|\beta_p| \leq \delta\}}{\delta} \right)$$

In these conditions it is clear that, when $\delta \to 0$, the diagonal elements of the matrix $G(\hat{\boldsymbol{\beta}}, \delta)$ corresponding to $\beta_j = 0$ tend to ∞, making the covariance matrix $V_\delta(\hat{\boldsymbol{\beta}})$ become singular in the limit. So regularity conditions of the asymptotic theory are not fulfilled with GSoft when any of the coefficients take the value zero. This is a major concern, since it is just one of the desirable characteristics in a proper variable selection method.

GSoft solves this problem developing an estimator of the covariance matrix that smooths the discontinuity in $G(\hat{\boldsymbol{\beta}}, \delta)$ when $\delta \to 0$ by means of approximating using the expectation of G and a continuous variable (e.g. normal) with mean in $\hat{\boldsymbol{\beta}}$. The estimator is

$$\hat{V}(\hat{\beta}_j) = \left\{ H(\hat{\boldsymbol{\beta}}) + \lambda \boldsymbol{\Gamma} G^*\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}\right) \right\}^{-1} \hat{F}(\hat{\boldsymbol{\beta}}) \left\{ H(\hat{\boldsymbol{\beta}}) + \lambda \boldsymbol{\Gamma} G^*\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}\right) \right\}^{-1} \quad (3.4)$$

where

$$\begin{aligned} &G^*\left(\hat{\boldsymbol{\beta}}, \sigma\right) = \text{diag}\left\{ \tfrac{2}{\sigma_1}\varphi(\hat{\beta}_1/\sigma_1), \ldots, \tfrac{2}{\sigma_p}\varphi(\hat{\beta}_p/\sigma_p) \right\} \\ &\left(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2\right) = \text{diag}\left[H(\hat{\boldsymbol{\beta}})^{-1} \hat{F}(\hat{\boldsymbol{\beta}}) H(\hat{\boldsymbol{\beta}})^{-1} \right] \\ &\varphi \text{ density function of the normal distribution} \end{aligned}$$

Anyhow, the main point to get a well established approach to the real variance–covariance matrix is to use an accurate estimator $\hat{F}$ of the Fisher matrix given by

$$F(\eta) = -E\left\{ \frac{\partial L(\eta)}{\partial \eta \partial \eta'} \right\}$$

Firstly, we made use of the approach carried out in [17]. Nevertheless, after some tests we realized that such a choice really underestimates the true

variance–covariance values. Our solution consists of rescaling this matrix multiplying it by a factor equal to the number p of variables in the model. So

$$\hat{F}(\hat{\boldsymbol{\beta}}) = I(\hat{\boldsymbol{\beta}}) = \frac{p[\partial^2 L(\hat{\boldsymbol{\beta}})/\partial\beta_i\beta_j]}{n} \tag{3.5}$$

Goodness–of–fit for this estimator is discussed in the results section.

3.1.3.3 Connection GSoft – CCD algorithm

The main aim of this article is to establish a theoretical connection between the convergence of the CCD algorithm and the existence of an optimum for the objective function with GSoft. This theoretical connection is established by the next theorem (proof in Appendix A).

Theorem 2. *The following two statements are equivalent:*
(1) The CCD algorithm for the lasso case converges.
(2) An optimum for the objective function under the terms of the theorem in [214] exists.

In this way, positive results of convergence obtained with the CCD algorithm can take advantage of the asymptotic properties of GSoft. Similarly, solutions obtained with GSoft are consistent in the way proved in [277].

Choice of Γ

As we mentioned above, we use a global threshold λ together with a vector of specific thresholds $\boldsymbol{\Gamma} = (\gamma_1, \ldots, \gamma_p)$ corresponding to the coefficients $\beta_1, \ldots, \beta_p$ of each variable in the model. In this study, we will evaluate the performance of three different choices for the Γ vector:

1. $\gamma_j = \sqrt{\text{var}(\mathbf{x}_j)}$. This is one of the choices carried out in [214]. As a consequence, we will refer to it as γ–Klinger. Adjusting the thresholds like this is equivalent to standardization.

2. $\gamma_j = \frac{1}{|\beta_j^{\text{ridge}}|}$. Ridge logistic regression was performed on data with a small global threshold λ_0, obtaining coefficients $\beta_j^{\text{ridge}} \neq 0$, $\forall j = 1, \ldots, p$. This choice is related to penalize according to the importance of the variable in ridge, and it is based on a special case of the adaptive lasso [451]. This choice will be designated as γ–ridge.

3. $\gamma_j = \frac{1}{|\beta_j^{\text{lasso}}|}$. Lasso logistic regression was performed on data with a small global threshold λ_0 and without using specific thresholds γ. Obviously, some coefficients β_j^{lasso} will take zero values. In this case, these variables are excluded from the final model, which is equivalent to take $\gamma_j = \infty$. It will be named as γ–lasso.

Consistency results

Variable selection consistency results in lasso can be found in the recent related literature. Oracle property [125] for the adaptive lasso in linear regression models is proved in [195]. Consistency results shown here are based on the subsequent adaptation of these results to the logistic case, carried out in [196], for the γ–lasso, there called iterated lasso.

The number of covariates p will be taken as a function of sample size, so the notation p_n will be used. For a set of indices $B \subseteq \{1, \ldots, p_n\}$ we consider $X_B = (\mathbf{x}_j, j \in B)$ and $C_B = X_B^{'} X_B / n$. From them we define:

$$\begin{aligned} \underline{c}(m) &= \min_{|B|=m} \min_{\|v\|=1} v^{'} C_B v \\ \overline{c}(m) &= \max_{|B|=m} \max_{\|v\|=1} v^{'} C_B v \end{aligned}$$

The Sparse–Riesz Condition (SRC) [444] is satisfied by the covariance matrix X with rank q and spectrum bounds $0 < c_* < c^* < \infty$ if

$$c_* < \underline{c}(q) < \overline{c}(q) < c^*$$

Let us take the subset of indices with true nonzero coefficients $B_0 = \{j, \beta_j \neq 0\}$. Let $k_n = |B_0|$ and $m_n = p_n - k_n$ be the number of nonzero and zero coefficients, respectively, and $b_{n1} = \min_{j \in B_0} |\beta_j|$, $b_{n2} = \max_{j \in B_0} |\beta_j|$ the minimum and the maximum of the true nonzero coefficients. Let us assume the following conditions:

(i) Bounds for the true coefficients and the covariates:

  (i1) For some constant $0 < b < \infty$, it is fulfilled that $b_{n2} < b$.

  (i2) For some constant $M > 0$, it is fulfilled that $|x_{ij}| < M$ for all $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, p_n\}$.

(ii) The design matrix X satisfies the SRC with bounds $\{c_*, c^*\}$ and rank $q_n = M_1 n^2 / \lambda_0^2$ being M_1 a positive constant.

(iii) When $n \to \infty$, the following convergence is satisfied

$$\frac{\sqrt{\ln\, k_n}}{b_{n1}\sqrt{n}} + \frac{\sqrt{n \ln\, m_n}}{\lambda r_n} + \frac{\lambda \sqrt{k_n}}{n b_{n1}} \to 0$$

  where r_n is the order of consistency at zero [196] of the primary lasso estimator.

Under (i)–(iii) it has been proved that

$$P(\text{sign}(\hat{\boldsymbol{\beta}}) = \text{sign}(\boldsymbol{\beta})) \to 1 \tag{3.6}$$

where the sign function is now taken in a slightly different way than in (3.3): $\text{sign}(\theta_1, \ldots, \theta_p) = (\text{sign}(\theta_1), \ldots, \text{sign}(\theta_p))$ and

$$\text{sign}(t) = \begin{cases} -1 & \text{if } t < 0 \\ 0 & \text{if } t = 0 \\ 1 & \text{if } t > 0 \end{cases}$$

so nonzero coefficients are correctly selected with γ–lasso with probability converging to one. From the same assumptions a result for the asymptotic distribution of the estimated nonzero coefficients of $\hat{\boldsymbol{\beta}}$ wih respect to the true ones $\boldsymbol{\beta}$ can be constructed. The following definitions are needed:

$$\begin{aligned} \boldsymbol{\beta}_{B_0} &= (\beta_j, j \in B_0)' \\ \hat{\boldsymbol{\beta}}_{B_0} &= (\hat{\beta}_j, j \in B_0)' \\ \mathbf{x}_{iB_0} &= (x_{ij}, j \in B_0)' \\ \epsilon &= (\epsilon_1, \ldots, \epsilon_n)' \\ \Sigma_{B_0} &= \frac{1}{n} X'_{B_0} D X_{B_0} \end{aligned}$$

where $\epsilon_i = y_i - (2P(y_i = 1|\mathbf{x}_i) - 1)$ and D is the diagonal matrix composed by the products of the logistic probabilities of case and control in each individual sample. Then, for $s_n^2 = \sigma^2 \alpha' \Sigma_{B_0}^{-1} \alpha$ with α any vector of length k_n fulfilling $\|\alpha\| \leq 1$, the following asymptotic property is satisfied by logistic lasso estimators $\hat{\boldsymbol{\beta}}$ with the γ–lasso choice:

$$\frac{\sqrt{n}}{s_n} \alpha' (\hat{\boldsymbol{\beta}}_{B_0} - \boldsymbol{\beta}_{B_0}) = \frac{\sum_{i=1}^n \epsilon_i \alpha' \Sigma_{B_0}^{-1} \mathbf{x}_{iB_0}}{\sqrt{n} s_n} + o_p(1) \to_D N(0, 1) \tag{3.7}$$

whenever $\frac{\lambda \sqrt{k_n}}{\sqrt{n}} \to 0$.

These two results, (3.6) and (3.7), together mean the γ–lasso choice has the asymptotic oracle property. The proof can be found in [196], which also refers to the proof for the linear case in [195]. A careful study of both proofs is enough to realize that only minor changes in the assumptions have to be applied to transfer the oracle property to the γ–ridge choice of specific penalizations.

When γ–Klinger penalizations are selected, this is equivalent to standardization, as proved in [214]. Therefore, only usual consistency lasso results [196, 277] can be proved in this case, and oracle property does not hold. An upper bound for the number of estimated nonzero coefficients in lasso is given in [196]. There, it is proved that the dimension of the model selected by lasso is directly proportional to n^2 and inversely proportional to the penalization parameter λ.

3.1.4 Results

3.1.4.1 Simulated data

We have simulated two scenarios with binary response according to one of the examples in [199]. In both of them, the response follows:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{x}'\boldsymbol{\beta})}$$

and the complementary probability for $y = -1$. This example has been adapted to two specifical scenarios carried out in [416] (Simulation 1) and [453] (Simulation 2), with the aim of comparing our results with those obtained there. Furthermore, a third bunch of simulations have been developed following [199]. We have also used the scenario in [453] to obtain the results of approximation of variance as explained in the last section.

Simulation 1

Our aim is to compare our results with those obtained with the least squares approximation (LSA) estimator. Comparisons with the results of the Park and Hastie (PH) algorithm of [306] shown in [416] are also established. The model is 9–dimensional with coefficients $\boldsymbol{\beta} = (3, 0, 0, 1.5, 0, 0, 2, 0, 0)'$. The components of $\mathbf{x}_i$ are standard normal and the correlation between each pair of variables $\mathbf{x}_{j_1}$ and $\mathbf{x}_{j_2}$ is fixed to $0.5^{|j_1 - j_2|}$. The sizes of the training samples are $n = 200$ and $n = 400$, and 500 simulation replications have been obtained each time. The BIC criterion is used to obtain the best solution for LSA and PH, while for the choice of λ in our models, we follow a slightly different approach. As choosing the λ giving rise to the smallest error rate (ER) does not necessarily produce a sparse model, we take the largest λ having an error rate smaller than $\min_\lambda ER + 2 * \text{sd}(ER)$. Results are shown in Table 3.1. From now on, lasso logistic regression will be referred with the abbreviation LLR.

The different estimators are compared in terms of model size (MS) and percentage of correct models identified (CM). Unlike [416], here we will not use the relative model error as a comparative measure, since it puts too much weight to the model error without penalty. Besides, in problems involving large amounts of noise, detection of the variables associated with the response is much more important than precise estimation of the true coefficients. Results obtained with our models are slightly better than those in [416], despite improvement of the results of LSA and PH seemed to be highly difficult. Comparisons between the different choices for the Γ vector are favorable to γ–ridge and γ–lasso, as the γ–Klinger seems to be more imprecise than those two regarding detection of the correct model. This imprecision grows when sample size decreases, until reaching the standard of LSA and PH.

Sample size	Estimation Method	MS Mean	(SE)	CM Mean	(SE)
200	LLR γ–Klinger	3.266	(0.025)	0.762	(0.019)
	LLR γ–ridge	2.896	(0.025)	0.812	(0.017)
	LLR γ–lasso	2.96	(0.028)	0.798	(0.018)
	LSA	3.178	(0.026)	0.798	(0.018)
	PH	3.272	(0.033)	0.716	(0.020)
400	LLR γ–Klinger	3.046	(0.011)	0.956	(0.009)
	LLR γ–ridge	2.964	(0.021)	0.860	(0.016)
	LLR γ–lasso	2.982	(0.022)	0.902	(0.013)
	LSA	3.130	(0.018)	0.888	(0.014)
	PH	3.092	(0.023)	0.846	(0.016)

Table 3.1: True model detection results. Comparison between our models and those in [416] is established in the same terms as there.

Simulation 2

Comparisons with the one–step sparse estimates developed in [453] are carried out, along with the SCAD and the other variable selection models used there. The second model is 12–dimensional with vector of coefficients $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0)^{'}$, while $\mathbf{x}$ is obtained as in Simulation 1, but with one important difference: variables with even index are translated to binary according to their sign. Size of the training sample is $n = 200$ and 1000 replicated datasets were obtained. Choice of the optimal λ for our models is carried out in a similar way to Simulation 1, but taking the largest λ having an error rate smaller than $\min_{\lambda} ER + 0.2 * \mathrm{sd}(ER)$. Results are shown in Table 3.2.

Same terms as in [453] are used: columns 'C' and 'IC' measure the average number of nonzero coefficients correctly estimated to be nonzero and the average number of zero coefficients incorrectly estimated to be nonzero, respectively; "Under–fit" and "Over–fit" show the proportion of models excluding any nonzero coefficients and including any zero coefficients through-

Method	C	IC	Proportion of Under-fit	Correct-fit	Over-fit
LLR γ–Klinger	2.84	1.68	0.16	0.14	0.70
LLR γ–ridge	2.77	0.82	0.22	0.40	0.37
LLR γ–lasso	2.71	0.71	0.29	0.40	0.31
one-step SCAD	2.95	0.82	0.051	0.565	0.384
one-step LOG	2.97	0.61	0.029	0.518	0.453
one-step $L_{0.01}$	2.97	0.61	0.028	0.516	0.456
SCAD	2.92	0.51	0.076	0.706	0.218
P-SCAD	2.92	0.5	0.079	0.707	0.214
AIC	2.98	1.56	0.021	0.216	0.763
BIC	2.95	0.22	0.053	0.800	0.147

Table 3.2: True model detection results. Comparison between our models and those in [453] is established in the same terms as there.

	$\rho = 0.25$		$\rho = 0.75$	
Method	C	I	C	I
LLR γ–Klinger	5.96	0.034	5.562	0.326
LLR γ–ridge	5.9	0.166	5.912	0.778
LLR γ–lasso	5.9	0.176	5.916	0.76
New	5.922	0	5.534	0.222
LQA	5.728	0	4.97	0.090
BIC	5.86	0	5.796	0.304
AIC	4.93	0	4.86	0.092

Table 3.3: True model detection results. Comparison between our models and those in [199] is established in the same terms as there.

out the 1000 replications, respectively. "Correct–fit" shows the proportion of correct models obtained.

Our methods show a worse behaviour than those in [453]. After some tests (results not shown) we realized that the reason was that they suffer a lot from the presence of binary variables. This is not a major concern, since our aim was to apply these methods to gene expression data, where all the variables move in a continuous way. Therefore, with the intention of testing them in a continuous environment, conditions in [199] were replicated. These conditions are the same as in Simulation 1 but the correlation between variables is now fixed to $\rho = 0.25$ and $\rho = 0.75$. Sample size was also fixed to $n = 200$. Results are shown in Table 3.3.

Optimal λ is chosen as in Simulation 1. "C" and "I" measure the average number of coefficients correctly and incorrectly set to zero, respectively. Comparisons are made with a new proposed algorithm in [199], a local quadratic approximation (LQA) algorithm developed in [125] and best subset variable selection using BIC and AIC scores. Competitive results are obtained with respect to the procedure in [199]. The best variable selection is obtained using BIC. The results obtained with the γ–Klinger are similar to the ones with γ–ridge and γ–lasso.

Approximation of variance

Covariance matrix estimation for the estimated coefficients have been obtained according to the approach previously explained. The same model as in Simulation 2 has been used, without the translation to binary (for simplicity). In Figure 3.1 the behaviour of variance estimation for $\beta_1 = 3$, $\beta_2 = 1.5$ and $\beta_3 = 0$, respectively, is shown in comparison with the true variance, as a function of λ. The estimation, obtained as the median on 1000 replications, fits almost perfectly to the variance except for small deviations when λ approachs zero (maximum likelihood estimator), as the true variance increases enormously.

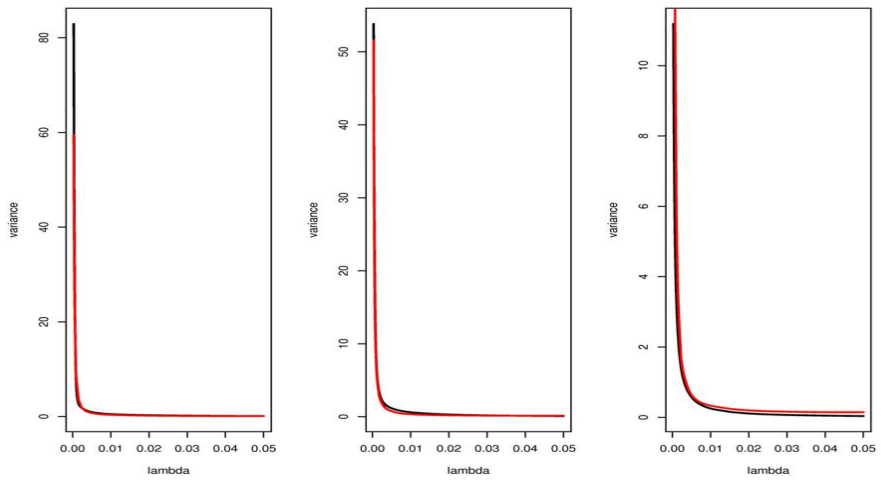

Figure 3.1: Variance estimation (in red) for the estimated values of β_1, β_2 and β_3 in Simulation 2 according to the estimator (3.4) with $\hat{F}$ taken as in (3.5). True variance (in black) was approximated by means of recursive simulation–estimation. Variance is displayed as a function of the penalty parameter λ.

3.1.4.2 Real data

The leukemia dataset [161] has been used on countless occasions through the gene expression literature. It comprises gene expression data for 72 bone marrow and peripheral blood samples (47 cases of acute lymphoblastic leukemia (ALL) and 25 cases of acute myeloid leukemia (AML)) in 7129 genes. Initially [161] the total sample was divided into a training sample (38 bone marrow samples) and a test sample (34 bone marrow and peripheral blood samples).

The colon dataset was analyzed initially by [6]. As leukemia, it is another commonly used dataset in genomic studies. A number of 62 samples (40 tumors and 22 controls) were measured in 2000 human genes. Absolute measurements from Affymetrix high–density oligonucleotide arrays were taken for each sample in each gene in both datasets. Here, we have worked with data in two different ways. On one side, we have carried out preprocessing steps (P) following [101], (i) thresholding of the measurements, (ii) filtering of genes, (iii) base 10 logarithmic transformation. On the other, we have also tried our models over the raw data (RD). With preprocessing, leukemia and colon datasets reduce their dimensionality to 3571 and 1225 genes, respectively.

As a result of combining these two ways to deal with data with the three different choices for γ, we have six different models. Table 3.4 shows the results for the leukemia dataset. To obtain accurate and precise measures

Leukemia	Test error	SD	Genes
RD-γ Klinger	0.062	(0.044)	67 (of 7129)
RD-γ ridge	0.064	(0.039)	11 (of 7129)
RD-γ lasso	0.102	(0.055)	6 (of 7129)
P-γ Klinger	0.079	(0.032)	16 (of 3571)
P-γ Zou	0.067	(0.030)	5 (of 3571)
P-γ Lasso	0.064	(0.028)	5 (of 3571)

Table 3.4: Test error and sparsity results for the leukemia dataset.

for the error and its standard deviation, we split 50 times the set of 72 samples into a training set of 38 samples and a test set of 34 samples. We also record the number of genes with nonzero coefficient for the optimal lambda, in terms of cross–validation (CV) error.

Table 3.5 shows the results for the colon dataset. The 62–sample has been randomly splitted 50 times into a training subsample of 50 observations and a test subsample of 12 observations.

When looking for other error test results obtained with different methods, it is common and correct to think that leukemia and colon datasets have been often used in the scientific literature since its appearance years ago. Nevertheless, it is difficult to find a fair comparison between methods, since each author uses a different way to obtain an error measure. Some of them only focus on a leave–one–out cross–validation rate (too optimistic); others center on the same data subdivision carried out by [161]; finally, the fairest way to know the real performance of each method is to randomly split the total sample N times into two disjoint samples, training and test. Table 3.6 compare our best results with those from methods obtaining their error rate following the latter way.

Comparisons with the following methods have been established. In [43], a CART-based method is developed to discover the emerging patterns inside the set of variables. BagBoosting [89] is a combination of bagging and boosting, two ensemble learning algorithms, applied to stumps, decision trees with only one split and two terminal nodes. Different algorithms are presented in [90]. *Pelora* is a penalized logistic regression method. *Forsela* is similar to *Pelora*, but making a search of single genes instead of groups, *Wilma* [91] shares some characteristics with *Pelora*, but suffers from a few limitations

Colon	Test error	SD	Genes
RD-γ Klinger	0.195	(0.130)	10 (of 2000)
RD-γ ridge	0.147	(0.116)	17 (of 2000)
RD-γ lasso	0.200	(0.128)	9 (of 2000)
P-γ Klinger	0.152	(0.096)	11 (of 1225)
P-γ Zou	0.182	(0.111)	15 (of 1225)
P-γ Lasso	0.215	(0.133)	10 (of 1225)

Table 3.5: Test error and sparsity results for the colon dataset.

Dataset	Method	Test error
Leukemia	Our best	0.062
	CART-Fisher [43] (*)	0.024–0.050
	BagBoosting [89] (**)	0.0408
	Pelora [90] (**)	0.0569
	Wilma [90] (**)	0.0262
	Forsela [90] (**)	0.0415
	PLS [299] (***)	0.033–0.047
	PCA [299] (***)	0.039–0.108
Colon	Our best	0.147
	CART-Fisher [43] (*)	0.128–0.234
	BagBoosting [89] (**)	0.161
	Pelora [90] (**)	0.1571
	Wilma [90] (**)	0.1648
	Forsela [90] (**)	0.1381

Table 3.6: Test error rates obtained using different methods from the scientific literature for the leukemia and colon datasets. (*) In each random split, 10 observations in the test set. (**) In each random split, 2/3 of the data to the training set, 1/3 of the data to the test set. (***) In each random split, 1/2 of the data to the training set, 1/2 of the data to the test set.

[90]. [299] uses dimension reduction through partial least squares (PLS) and principal component analysis (PCA), classifying with discriminant analysis. Our error results are only slightly worse than the others for the leukemia dataset, and among the best for colon. In any case, all the error rates are quite similar. Many of the methods we compare with stand out for grouping genes ([43], [299], *Pelora* and *Wilma* in [90]) in one way or another. Gene preselection is carried out by means of preexisting methods in [43] and [89]. Our logistic lasso methods neither makes use of grouping or gene preselection nor it is necessary to select a lot of different parameters, as in [43], appart from the penalty λ. Moreover, its sparsity (see Tables 3.4 and 3.5) and the interpretability associated with it are merits not fulfilled by these other methods.

Gene expression data is seen as the paradigm of the case $n << p$, as Affymetrix or oligonucleotide arrays map large parts of the human genome while only tens or hundreds of individuals are sampled. This situation makes most of traditional statistical methods inapplicable, so new variable selection approaches had to be developed to deal with this *curse of dimensionality* problem. Lasso selects a group of $p^{'} \leq n$ genes with high importance in the classification of samples, and assign a zero coefficient to the rest. Use of the CCD algorithm to solve the optimization problem is highly desirable, as it provides with the global solution of GSoft in the fastest way.

In a more biological way, we have also studied which genes are more related with the ALL/AML status in leukemia. Observations of the genes with nonzero coefficients for each model have been carried out. As expected,

some recurrences have been found in the six different models. Table 3.7 shows those genes appearing more frequently.

The fact that some genes are discovered in some models and not in others can be explained from the correlations between them. These correlations arise as a result of co–inheritance of nearby genes throughout generations. For instance, gene *M19507* takes a nonzero coefficient with all but two of the models, and gene *M92287* takes nonzero coefficients only in these two models. If we take a careful look to the correlation between them, we detect it as abnormally high. A correlation study between all the genes with nonzero coefficient in any of the models has been carried out. With the aim of knowing the real significance of each correlation value, we have obtained a significance value as the proportion of values, in a set of 10000 random correlations between pairs of genes from the entire dataset, higher than the correlation. This way, significance of the correlation *M19507*–*M92287* is 0.0558; the one between *M84526*–*Y00787* is 0.048, which explains why they are partly complementary. Significances of correlations between gene *Y00787* and the last eight genes in Table 3.7 are also very low, as they are detected specifically in those two models where *Y00787* is not. In a similar way, pairwise correlations in this 8–gene group are often high. Complementarity in the detection by the different models emphasizes one of the biggest problems of lasso selection, also marked in [452]: when there is a group of significan variables with high pairwise correlation lasso selects only one, and does not care which one.

Genes	RD-γ Klinger	RD-γ Zou	RD-γ Lasso	P-γ Klinger	P-γ Zou	P-γ Lasso
M27891	X	X	X	X	X	X
M19507	X	X	X		X	
M84526	X			X	X	X
Y00787		X	X		X	X
M92287				X		X
U05255		X	X			
M17733		X	X			
M63138	X		X			
M96326	X	X				
L07633	X			X		
U82759	X			X		
HG1612	X			X		
M13690	X			X		
M23197	X			X		
X95735	X			X		
Y07604	X			X		
X85116	X			X		

Table 3.7: Genes with nonzero estimated coefficients in the different models for the leukemia dataset. Here we show the seventeen ones which are detected in more than one model.

A bunch of articles can be found in the gene expression literature looking for the genes associated with the ALL/AML status. It is expected that exists some kind of intersection between the sets of genes given by the different studies. First five genes in the relation of Table 3.7 (*M27891*, *M19507*, *M84526*, *Y00787* and *M92287*) are also discovered in [237], being *M27891* the one showing the strongest association with disease, as happens here. Three of the four genes pointed out in [168] (*U82759*, *HG1612* and *X95735*) are also discovered here. On the other hand, coincidences with the list given in [392] are more limited.

3.1.5 Conclusion

We study lasso logistic regression by means of a generalized soft–threshold (GSoft) estimator. An equivalence between existence of a solution in GSoft and convergence of the CCD algorithm to the same solution is given. An approximation of the covariance matrix for the estimated coefficients $\hat{\beta}$ based on the GSoft approach produces very accurate results. The CCD algorithm is fast, stable and efficient, and allows different kinds of implementations. Efficiency of the optimization algorithm is a main issue nowadays, as the datasets used in many fields (text categorization, image processing, ...) have extraordinary high dimensions.

We tried different options for the vector $\mathbf{\Gamma}$ of specific penalizations in GSoft. Some of them are based in the variability shown by each covariate, while others depend on previous application of penalized regression approaches to data. Their consistency properties follow from appropriate developments in the recent literature.

Finally, we applied these methods to simulated and real gene expression data. The same simulations carried out in other studies were used here, in order to provide honest and fair comparisons. Common real gene expression datasets, like leukemia or colon, allow us to know the ability of these methods to detect genes related with the disease or trait under study. The penalized regression approaches performed in this work are expected to give rise to sparse models, where only a very small percentage of covariates (genes) have weight in classification/prediction.

Chapter 4

Machine learning tools: Support Vector Machine (SVM) approach to classify in genetic association studies

4.1 A SVM adaptation to SNP data

This chapter contains a new SVM approach developed to work with SNP data in two–class classification problems. Although unpublished, this SVM tool is two–fold and has been therefore presented in two different conferences: the 24th International Biometric Conference [148] (statistical aspects), and the 2nd Iberian Grid Infraestructure Conference [355] (information technology aspects).

4.1.1 Abstract.

Support vector machines (SVMs) arose in the machine learning field in the nineties as a pattern classification technique. Their main idea is to construct a separating hyperplane between classes in a feature, high–dimensional space, where separability is easier to achieve. To get that, SVMs use a kernel approach. The choice of the kernel is fundamental facing accurate classification. In this chapter, a new kernel approach, adapted to work with categorical SNP data, is developed. Its classification results are given in comparison with similar techniques. The computational burden generated by this SVM approach is very high. Computer parallelization is carried out using two different GRID infraestructures, in order to reduce computation times and allow for feasibility.

4.1.2 Introduction

Support vector machine (SVM) is a pattern classification technique initially proposed by Vapnik and co–workers [42, 82, 409]. The past years have witnessed an increasing interest in this machine learning method, which introduces new principles to solve old problems. SVMs aims at minimizing an upper bound of the generalization error through maximizing the margin between the separating hyperplane and the data. What makes SVMs attractive is the property of condensing information in the training data and providing a sparse representation by using a very small number of points (support vectors) [8, 156].

The performance of SVMs largely depends on the kernel choice and, hence, on the feature space selected. SVMs transform data from the input space to a feature space, usually high–dimensional, where data separation is easier. This is made by means of a similarity measure, called kernel. In the machine learning literature, there are three types of kernels [42, 409] that can be found. The importance of the kernel choice can be observed in many studies [8, 33, 45, 193]. New kernels, some of them modifications of the previous ones, have been proposed. The search for kernels suitable for feature selection in high–dimensional data problems is also a matter of study [193, 249]. Methods for incorporating prior knowledge about a problem at hand in SVMs are also valuable [54, 366].

SVMs have been applied to data from many different fields. Regarding genetic data, it is easy to find studies both in gene expression [54, 319] and SNP data [248, 249, 371]. In this sense, specific approaches have been developed to handle large high–dimensional datasets [44] or to combine a SVM approach with l_1 penalization (see Chapter 3) [387]. There are plenty of software packages developed to run SVMs in many different ways [176, 319].

Most SVM approaches in use nowadays were thought to work with continuous data. Here, we have developed a new kernel approach specifically designed for SNP categorical data. Its classification abilities are studied in comparison with other classification techniques (see Chapter 5). Computer parallelization is used to reduce computation times, taking advantage of the fact that many calculations can be made independently of each other. The type of computer implementation carried out is very important, due to the computational burden generated by the method.

This chapter is organized as follows: Subsection 4.1.3 contains SVM origins and principles, together with a proper explanation about our approach. Classification and computation time results are given in Subsection 4.1.4. Proof of the mathematical properties fulfilled by the kernel is located in the Appendix B of this essay.

4.1.3 Methods

4.1.3.1 Pattern recognition: from perceptron to SVMs

The first machine learning technique, called perceptron, was proposed more than half a century ago [341, 342]. The main aim of perceptron was to solve pattern recognition problems building a rule to accurately separate data from different classes. A training sample $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ is used. Perceptron connects the outputs $y_i \in \{-1, 1\}$ with the inputs $\mathbf{x}_i = (x_i^1, \ldots, x_i^p)$ by means of a functional dependence model

$$y_i = \text{sign}\left[\langle \omega, \mathbf{x}_i \rangle + b\right]$$

Geometrically, the reference space is divided in two regions, one for $y_i = -1$ and the other for $y_i = 1$. Separation between these two regions is defined by the hyperplane

$$\langle \omega, \mathbf{x} \rangle + b = 0$$

The vector ω and the scalar b establish the direction and position, respectively, of this separating hyperplane. Figure 4.1 shows an example of hyperplane (straight line) separating two different classes in the plane.

During the learning process, the perceptron algorithm uses data from the training sample to find the separating hyperplane minimizing the addition of distances from the hyperplane to incorrectly classified points. Incorrectly classified points are those fulfilling:

$$\begin{aligned} y_i = -1 \quad &\text{and} \quad \langle \omega, \mathbf{x}_i \rangle + b > 0 \\ &\text{or} \\ y_i = 1 \quad &\text{and} \quad \langle \omega, \mathbf{x}_i \rangle + b < 0 \end{aligned}$$

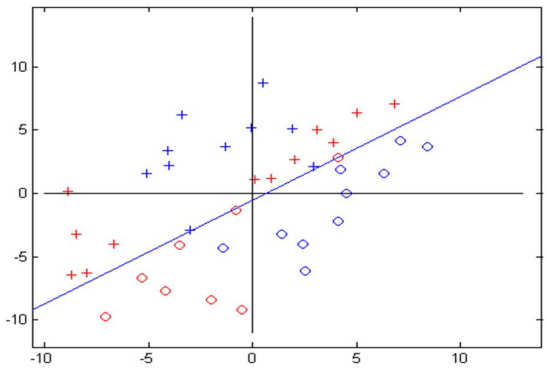

Figure 4.1: Perceptron classification in a two–class problem. Image obtained from `www.celebisoftware.com/Tutorials/neural_networks`.

Therefore, the objective function to be minimized is

$$L(\omega, b) = L^P_{\omega,b} = -\sum_{i \in I} y_i \left[\langle \omega, \mathbf{x}_i \rangle + b\right] \tag{4.1}$$

where the set of indexes I moves along the incorrectly classified points of the training sample. Although the solution of the optimization problem can be easily reached in a finite number of steps, the perceptron suffers from several problems, summarized in [333]:

1. When the data is separable, there are plenty of solutions, depending on the initial values of the optimization algorithm.
2. The finite number of steps to reach the solution could eventually be large, depending on the larger the separation between classes is.
3. When data is not separable, the optimization algorithm will not converge, and it will enter into an infinite loop.

The first problem is easy to solve, adding new restrictions to the separating hyperplane. The optimal separating hyperplanes (OSH) technique [409] looks for separating classes maximizing the distance to the closest point in each class (margin). This way, only one solution expected to give rise to a better classification is obtained. The OSH technique generalizes the perceptron criterion 4.1 formulating the following optimization problem:

$$\begin{gathered} \max_{\|\omega\|=1} C \\ \text{fulfilling} \quad y_i \left[\langle \omega, \mathbf{x}_i \rangle + b\right] \geq C \quad i = 1, \ldots, n \end{gathered}$$

With these restrictions it is ensured that all the observations are at least at a distance C of the separating hyperplane. Taking $\|\omega\| = 1/C$, the problem can be reformulated in terms of ω and b. This is a convex optimization problem (quadratic programming with inequality constraints). The objective function to be maximized now is

$$L^{\mathrm{O}}_{\omega,b} = \frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{n} \alpha_i \left\{y_i \left[\langle \omega, \mathbf{x}_i \rangle + b\right] - 1\right\}$$

It can be proved that the solution ω depends only on the support vectors $\mathbf{x}_i$. The support vectors are those observations lying on the border of the margin (and therefore having $\alpha_i > 0$). Figure 4.2 shows an optimal separating hyperplane with its margins. Support vectors lie on the left and right borders. The fact that none of the training samples fall inside the margin does not imply the same is going to happen with new observations. Simply, intuition indicate us that a large margin in training data would give rise to a good separation with new data.

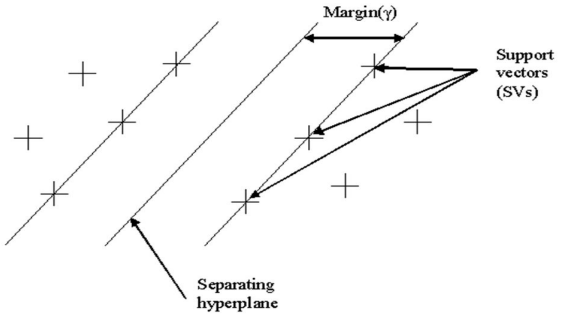

Figure 4.2: Example of optimal separating hyperplane. The support vectors lie on the borders of the margin. A linear combination of them defines the direction ω of the hyperplane. Image obtained from `www.enm.bris.ac.uk/teaching/projects/2004_05/dm1654/svm_classification.html`.

However, the same as with the perceptron, the OSHs do not deal with non–separable data. Data from common problems in real life overlaps in the reference space. A way to deal with overlapping consists of maximizing C, as before, but allowing some samples to fall in the incorrect side of the margin [57, 177]. Working this manner is also expected to overcome one of the main defects of the perceptron and OSHs, called overfitting.

In the non–separable case, a set of slack variables $\xi = (\xi_1, \ldots, \xi_n)$ is defined to allow incorrectness in data. The optimization problem is now posed as

$$\begin{gathered} \max C - \gamma \textstyle\sum_{i=1}^{n} \xi_i \\ \text{fulfilling} \quad y_i \left[\langle \omega, \mathbf{x}_i \rangle + b\right] \geq C(1 - \xi_i) \quad i = 1, \ldots, n \\ \xi_i \geq 0 \end{gathered}$$

where γ sets an upper bound for the amount of incorrectly classified samples; the separable case would correspond with $\gamma = \infty$. The objective function in the present case is

$$L^{\mathrm{N}}_{\omega,b} = \frac{1}{2} \|\omega\|^2 + \gamma \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left\{ y_i \left[\langle \omega, \mathbf{x}_i \rangle + b\right] - (1 - \xi_i) \right\} - \sum_{i=1}^{n} \mu_i \xi_i$$

The so–called dual problem is equivalent and can be obtained simply replacing with the values derived from making the first derivative of the objective function equal to zero:

$$L^{\mathrm{N}_d}_{\alpha} = \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k y_i y_k \langle \mathbf{x}_i, \mathbf{x}_k \rangle \tag{4.2}$$

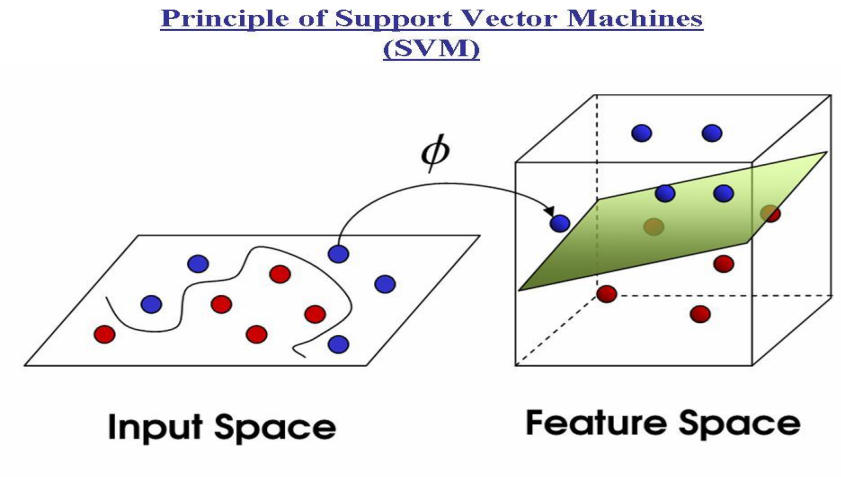

Figure 4.3: The fundamental principle of the SVM procedure is to map from the original reference space to a feature space, usually high–dimensional, where data separation is easier. Image obtained from `imtech.res.in/raghava/rbpred/home.html`.

As in the separable case, the solution will be defined in terms of the support vectors, but it will still be a linear function of the data. SVMs arise as an evolution of the techniques shown here. The main idea of the SVM method is that data separation can be simplified working with higher dimensions (feature space) and coming back then to the original space. This will give rise to a non–linear data separation.

4.1.3.2 Feature spaces, kernel choices and the kernel trick

To work in higher dimensions it is necessary to convert the original reference space X into a feature high-dimensional dot product space F (below we will see the need of the dot product requirement). To this end, we use a map:

$$\begin{aligned} \phi: \quad & X \to F \\ & \mathbf{x} \longmapsto \phi(\mathbf{x}) \end{aligned}$$

Figure 4.3 illustrates the idea of mapping in SVMs to get non–linear transformations back in the original space. SVM methodology is only understood around the concept of kernel. A definite positive kernel K is a similarity measure defined from a map and a dot product in the feature space:

$$\begin{aligned} K: \quad X \times X \quad & \to \mathbb{R} \\ (\mathbf{x}_i, \mathbf{x}_k) \quad & \longmapsto K(\mathbf{x}_i, \mathbf{x}_k) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_k) \rangle \end{aligned}$$

where $i, k \in \{1, \ldots, n\}$. A kernel matrix is the $n \times n$ matrix whose elements are $K_{ik} = K(\mathbf{x}_i, \mathbf{x}_k)$. A common approach in SVM studies is to begin from the kernel instead of from the feature space or the map. It is relatively easy to prove that any definite positive kernel can be represented as a dot product in a given space [367].

The optimization problem to be solved in SVMs is immediately derived from the dual problem in 4.2, but now with the kernel K instead of the general dot product:

$$\begin{gathered} \max L_{\alpha}^{\mathrm{S}_d} = \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k y_i y_k K(\mathbf{x}_i, \mathbf{x}_k) \\ \text{fulfilling } 0 \leq \alpha_i \leq \gamma \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{gathered}$$

A definite positive kernel has to fulfill some mathematical properties. Once we have a definite positive kernel, the kernel trick states that given an algorithm which is formulated in terms of a definite positive kernel K, one can construct an alternative algorithm by replacing K by another positive definite kernel $\widetilde{K}$ [367]. The best known application of the kernel trick is in the case where K is the dot product in the input domain; however, the trick is not limited to that case.

In the scientific literature concerning SVMs, there are three families of kernels that dominate:

- Polynomial kernels

  $$K(\mathbf{x}_i, \mathbf{x}_k) = \langle \mathbf{x}_i, \mathbf{x}_k \rangle^d$$

  with $d \in \mathbb{N}$. Examples of use of this family of kernels can be seen in [2, 54]. There are also SNP association studies making use of them [248, 371, 440].

- Gaussian kernels

  $$K(\mathbf{x}_i, \mathbf{x}_k) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_k\|^2}{2\sigma^2}\right)$$

  with $\sigma > 0$. Some of the kernels used in [2, 54] belong to this family.

- Sigmoid kernels

  $$K(\mathbf{x}_i, \mathbf{x}_k) = \tanh(\tau \langle \mathbf{x}_i, \mathbf{x}_k \rangle + \upsilon)$$

  with $\tau > 0$ and $\upsilon > 0$. In [2] is carried out a SVM feature selection study with kernels from this family, among others.

The importance of the kernel choice has been addressed in many studies [193]. An interesting chapter about design of kernel functions can be found in [367]. Different kernel approaches from the ones explained above are developed, for instance, in [33, 45]. SVM classifiers are improved in [8] by means of modifying kernel functions.

Completely specifying a SVM therefore requires specifying two elements: the kernel function (K) and the magnitude of the penalty for violating the soft margin (γ). The settings of these parameters depend on the specific data at hand [54]. Most of the scientific articles focus on the choice of the kernel. However, it is also possible to find studies trying to assess the influence of the penalty parameter γ on the results [176].

4.1.3.3 Adaptation to SNP categorical data

Samples $\mathbf{x}_i = (x_i^1, \ldots, x_i^p)$ in case–control SNP association studies fulfill $x_i^j \in \{1, 2, 3\}$, depending on if the individual is homozygous in the common allele (1), heterozygous (2) or homozygous in the rare allele (3) for the SNP under study, or a similar coding. The categorical nature of SNP data makes the use of the majority of kernels developed in the literature a complete nonsense. These kernels have been usually developed to work with continuous data.

So there is a need to look for a kernel ready to work with SNP genetic profiles with values in $\{1, 2, 3\}^p \subset \mathbb{R}^p$. It is not easy to define a similarity measure in this set, as such function would understand distance between 1 and 2 is equal to distance between 2 and 3 when there is no reason for such relation to exist in reality. Take as example the sickle cell anemia disease: while homozygote individuals in the common allele (1) and heterozygote individuals (2) are almost equal with respect to disease symptoms, homozygote individuals in the rare allele (3) suffer all the symptoms in the most severe degree.

To look for a kernel is to look for a map of SNP profiles. A basic idea to carry out such a mapping is to consider the two different alleles of each SNP, that is, to map the SNP profiles to binary data. So the genetic profile of an individual could be described by means of $2p$ alleles $\mathbf{z}_i = (z_i^1, z_i^2, \ldots, z_i^{2p-1}, z_i^{2p})$ instead of p SNPs $\mathbf{x}_i = (x_i^1, \ldots, x_i^p)$. Data would be transformed like

$$\begin{aligned}
&\text{if } x_i^j = 1 \text{ then} \quad && z_i^{2j-1} = z_i^{2j} = 0 \\
&\text{if } x_i^j = 3 \text{ then} \quad && z_i^{2j-1} = z_i^{2j} = 1 \\
&\text{if } x_i^j = 2 \text{ then} \quad && z_i^{2j-1} = 1, z_i^{2j} = 0 \quad \text{or } z_i^{2j-1} = 0, z_i^{2j} = 1
\end{aligned}$$

Thus, an ambiguity appears each time we have a heterozygote inside the SNP profile. The best way to deal with this ambiguity is thinking the feature space not as a real space, but as a vector space with all the discrete

random variables in a certain real space. This way, the map $\mathbf{z}$ of a SNP profile $\mathbf{x}$ would be a random variable with $T = 2^{\#\{x^j=2\}}$ values with equal probability. For instance, let $p = 3$ and $\mathbf{x} = (2, 1, 2)$ be. The transformed allele profile would be a random variable in $\mathbb{R}^6$ taking 4 values, each one with probability 1/4:

$$\begin{aligned}\mathbf{z}_{(1)} &= (1, 0, 0, 0, 1, 0)\\ \mathbf{z}_{(2)} &= (1, 0, 0, 0, 0, 1)\\ \mathbf{z}_{(3)} &= (0, 1, 0, 0, 1, 0)\\ \mathbf{z}_{(4)} &= (0, 1, 0, 0, 0, 1)\end{aligned}$$

Intuitively, we are splitting each heterozygote into two allele profiles, and then considering all the allele profiles $\mathbf{z}_{(t)}$, $t = 1, \ldots, T$, for each SNP profile. The weight of each SNP profile is shared equally among each allele profile. When constructing a kernel, we have to bear in mind some aspects:

1. The kernel has to be a similarity measure between profiles. As we work with binary data, the indicator function will be probably necessary...
2. Weight of each SNP profile is shared equally among all the allele profiles. Therefore, probability of each allele profile is $1/T$.
3. As in every discrimination problem, each predictor variable has a certain discriminant power. It will be then necessary to use any kind of measure w_j of the discriminant power of each allele or group of alleles.

Bearing these aspects in mind, a first proposal of kernel for two SNP profiles $\mathbf{x}_i$ and $\mathbf{x}_k$ could be:

$$K(\mathbf{x}_i, \mathbf{x}_k) = \sum_{s=1}^{T_i} \sum_{m=1}^{T_k} \frac{1}{T_i} \frac{1}{T_k} \sum_{l=1}^{2p} w_l I\left\{z^l_{i(s)} = z^l_{k(m)}\right\} \tag{4.3}$$

where I is the indicator function, $T_i = 2^{\#\{x^j_i=2\}}$, $T_k = 2^{\#\{x^j_k=2\}}$ and $z^l_{i(s)}$ is the value of the l allele in the s allele profile of the i individual. As a similarity measure, this kernel estimates the similarity between two SNP profiles $\mathbf{x}_i$, $\mathbf{x}_k$ as the weighted addition along all the possible pairs of allele profiles of all the allele weights w_l where the profile coincides. However, this kernel suffers from two main defects which make it non suitable in case–control association studies.

First, a large part of the research concerning disease–genotype association is focused on the search for interaction effects between SNPs [186, 258, 305, 337, 432]. The kernel in (4.3) only measures similarity between individuals by means of similarities of separate alleles. Interactions between SNP

pairs or trios are not borne in mind. Second, this kernel could be invalid depending on the data dimensions we are working with. A definite positive kernel can be defined as a kernel giving rise to definite positive kernel matrices (hence non–singular) for every sample of profiles. Nevertheless, the kernel in (4.3) has a maximum rank of $p+1$, so if we have $n > p+1$ then the kernel matrix is singular and the optimization problem in SVMs cannot be solved.

Therefore, the need to solve these drawbacks led us to a second proposal of kernel:

$$K(\mathbf{x}_i, \mathbf{x}_k) = \sum_{s=1}^{T_i} \sum_{m=1}^{T_k} \frac{1}{T_i} \frac{1}{T_k} \left(\sum_{l=1}^{2p} \sum_{r=l}^{2p} w_{lr} I \left\{ z_{i(s)}^{lr} = z_{k(m)}^{lr} \right\} \right) \tag{4.4}$$

where $z_{i(s)}^{lr} = (z_{i(s)}^{l}, z_{i(s)}^{r})$ is the pair of alleles (l, r) of the s allele profile. It is obvious that (abuse of notation):

$$\begin{aligned} w_{ll} &= w_l \quad , \quad l = 1, \ldots, 2p \\ z_{i(s)}^{ll} &= z_{i(s)}^{l} \quad , \quad l = 1, \ldots, 2p \end{aligned}$$

This kernel (4.4) can be also expressed as follows:

$$K(\mathbf{x}_i, \mathbf{x}_k) = \sum_{s=1}^{T_i} \sum_{m=1}^{T_k} \frac{1}{T_i} \frac{1}{T_k} \left(\sum_{l=1}^{2p} w_l I \left\{ z_{i(s)}^{l} = z_{k(m)}^{l} \right\} + \sum_{l=1}^{2p} \sum_{r=l+1}^{2p} w_{lr} I \left\{ z_{i(s)}^{lr} = z_{k(m)}^{lr} \right\} \right)$$

where we are adding to the kernel in (4.3) SNP–SNP interaction weights. Thus, the kernel in (4.4) can be considered an extension of kernel (4.3) in the sense that similarities between individuals bear now in mind interactions between pairs of alleles. The proof that expression (4.4) is a definite positive kernel is given in Appendix B.

So this kernel solves, partly, the lack of study of interactions we had with (4.3). The problem of singularity of the kernel matrices generated is also partly solved. The maximal rank of kernel matrices is now $p(p+1)/2$. This is remarkably higher than $p+1$ but in some cases it could still be considered insufficient. Table 4.1 shows the maximal ranks (hence, maximal sample sizes) of kernels (4.3) and (4.4) depending on p.

4.1.3.4 Optimization of the objective function

The SVM optimization problem can be analytically solved only when the training sample size is extremely small, or in the separable case, when the

p	Maximal rank kernel (4.3)	Maximal rank kernel (4.4)
10	11	55
25	26	325
50	51	1275
100	101	5050
200	201	20100
500	501	125250
1000	1001	500500

Table 4.1: Maximal ranks of the kernel matrices depending on p. These maximal ranks determine the maximal sample size which, if exceeded, would give rise to a singular kernel matrix and hence, an unsolvable optimization problem.

support vectors are known in advance. However, in most cases, this optimization problem has to be solved numerically. There are many optimization techniques suitable for this kind of problems. Here we use the primal active set method to solve quadratic programming problems with inequality constraints. A complete explanation about the method can be found in [130].

The primal active set method fits perfectly with the datasets to be used here (see next section); nevertheless, this method works out to be computationally demanding when the dimension p goes beyond. This is the situation in SNP case–control association studies, as every day high–throughput technologies become more and more common. In some cases, optimization techniques discovering an approximate solution instead of the global solution can be acceptable, especially when methods looking for global solutions are computationally unfeasible.

Genetic algorithms (GAs) were first developed in [187]. The GAs are search algorithms based on the mechanics of natural selection and genetics. A very interesting book compiling the basics of GAs is [159]. Figure 4.4 summarizes their working scheme. GAs combine survival of the best data structures with data random crossing among such structures. Each new generation, a new set of data is created mixing pieces of the ones showing best results (in terms of the objective function) in the previous generations, and new data pieces are occasionally added to be tried. Although partly random, GAs are not completely random, as they take advantage of the information acquired in previous steps to speculate about new search points that could most likely improve the objective function values.

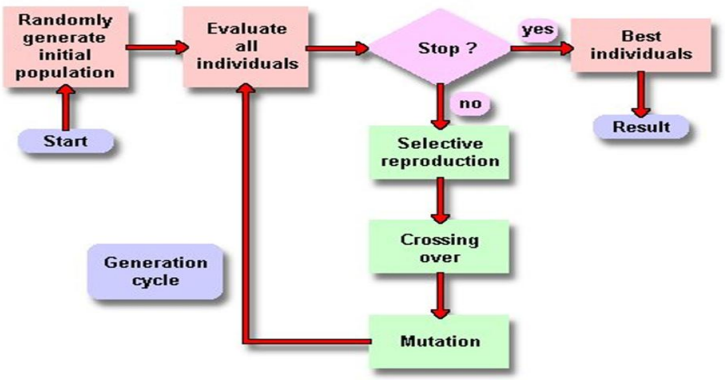

Figure 4.4: Working scheme of GAs. Many of the terms used (mutation, reproduction, population,...) are taken from the genetic field. Image obtained from [159].

4.1.4 Results and discussion

4.1.4.1 SVM classification

To evaluate the classification abilities of this SVM tool, we took some of the datasets simulated in [149] (see next chapter), specifically those including marginal effects. Due to computational feasibility problems, we only kept 20 SNPs (two causal and 18 noise ones). Sample sizes were reduced to 200 individuals for training and 100 for testing the method (half cases and half controls in both).

Misclassification results are shown in Table 4.2 in comparison with the results of classification trees (CART), random forests (RF) and logistic regression (LR) obtained in [149] and the Bayes error rate (also explained there). The six datasets differ in the minimum allelic frequency (MAF) of the causal SNPs (0.2 and 0.4) and the choice of the θ parameter (0.8, 1.4 and 2) in the penetrance model [266]. SVM classification results are worse than tree–based methods ones and only slightly better than LR's. In any case, SVM errors move far from Bayes error rates, even in the most favorable cases to detect association.

Weights w_{lr} are expected to measure in some way the discriminant power of each pair of alleles. We tried different choices, all of them based on differences of allelic frequencies between cases and controls. Results do not differ in a significant way. Further research is needed about this issue. A boosting–based approach to "update" the allele weights could be suitable here.

	SVM	CART	RF	LR	Bayes
MAF $= 0.2$, $\theta = 0.8$	0.49	0.49	0.49	0.49	0.44
MAF $= 0.2$, $\theta = 1.4$	0.47	0.46	0.47	0.49	0.40
MAF $= 0.2$, $\theta = 2$	0.47	0.42	0.45	0.48	0.38
MAF $= 0.4$, $\theta = 0.8$	0.45	0.43	0.44	0.46	0.38
MAF $= 0.4$, $\theta = 1.4$	0.40	0.35	0.37	0.44	0.32
MAF $= 0.4$, $\theta = 2$	0.38	0.32	0.32	0.41	0.29

Table 4.2: Misclassification results of our SVM tool in comparison with the methods carried out in [149] and the Bayes error rate in six different simulated scenarios of minimum allelic frequency (MAF) and penetrance (θ).

4.1.4.2 Computation time

The SVM tool presented here is computationally demanding due to several aspects, like the data mapping carried out or the study of every SNP–SNP interaction. Programming of the method consists of three fundamental stages:

1. Construction of the kernel matrix.

2. Solution of the optimization problem by means of the primal active set method.

3. Testing of the method.

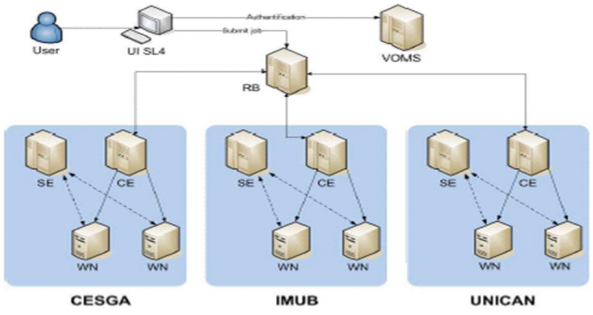

Figure 4.5: Sketch of the GRID infrastructure in the Ingenio Mathematica Project. At each site it has been installed and configured a Computing Element (CE), a Storage Element (SE), a Monitoring System (MON) together with several Worker Nodes (WN). To do this, XEN virtual machines have been used with Dual Core Xeon processors with 2Gb of memory. Image obtained from [355].

There are several factors determining computation times in each one of these stages. Amount of SNPs under study and heterozygosity are the most important, while training and testing sample sizes have also strong influence. To illustrate the computational burden generated by this SVM approach, Table 4.3 shows the computation times needed for constructing the kernel matrix depending on training sample sizes and number of SNPs in a two–year old 512 Mb RAM laptop. A quick look is enough to realize that moving onto high dimensions (more SNPs) is prohibitive.

Computational burden is concentrated on stages 1 and 3 above, especially on the construction of the kernel matrix. This makes the problem manageable, as both stages can be parallelized in a computer cluster (in the kernel matrix each cell is calculated independently of the others; testing each sample is also independent of previous tests with other samples). It reminds a double "bottleneck", as only the middle stage (solution of the optimization problem) needs to be carried out completely at the same computer.

Parallelization allows us to save computation time, sharing computational burden of stage 1 and 3 among several computers connected in a cluster. Here we made use of two different parallelization approaches to obtain SVM classification results:

- A GRID platform with 11 computers Core Duo AMD Opteron 4 Gb RAM located at the Department of Statistics and Operations Research of the University of Santiago de Compostela. Package *Rmpi* in R was

	Number of SNPs		
Training sample	5	10	15
50	7s.	30s.	14m.
75	11s.	1m.	40m.
100	20s.	2m.15s.	–
200	1m.22s.	8m.	–
400	5m.25s.	36m.	–
600	13m.	1h.7m.	–
800	23m.	2h.	–
1000	38m.	–	–
1300	1h.20m.	–	–
1600	1h.35m.	–	–
2000	2h.33m.	–	–

Table 4.3: Approximated computation times of the kernel matrix in a standard laptop as a function of training sample size and number of SNPs under study in hours (h.), minutes (m.) and seconds (s.). Cells marked with a hyphen (–) mean unfeasibility to obtain the matrix. Times obtained with 15 SNPs are highly variable, depending on the heterozygosity of the SNPs; this problem gets worse when more than 15 SNPs are studied.

used to parallelize and optimize the entire process.

- The GRID of the Ingenio Mathematica Project (`www.i-math.org`) in CESGA (Supercomputation Galician Center) has 124 slots. A small portion of them were available to run some tests. Figure 4.5 details the GRID infrastructure, consistent of three nodes at CESGA, UNICAN (University of Cantabria) and IMUB (University of Barcelona). Proper explanation about hardware infrastructure and aims can be found in [355].

Chapter 5

Empirical studies in clinical genetics

5.1 Evaluating the ability of tree-based methods and logistic regression for the detection of SNP–SNP interaction

This section consists of the results included in [149]. Minor changes have been applied here to the original manuscript, either to remove redundant information from previous chapters, or to include some figures which in the article were pushed into supplementary material. Furthermore, the format has been adapted to not alter this essay.

5.1.1 Abstract

Most common human diseases are likely to have complex etiologies. Methods of analysis that allow for the phenomenon of epistasis are of growing interest in the genetic dissection of complex diseases. By allowing for epistatic interactions between potential disease loci, we may succeed in identifying genetic variants that might otherwise have remained undetected. Here we aimed to analyze the ability of logistic regression (LR) and two tree–based supervised learning methods, classification and regression trees (CART) and random forest (RF), to detect epistasis. Multifactor–dimensionality reduction (MDR) was also used for comparison. Our approach involves first the simulation of datasets of autosomal biallelic unphased and unlinked single nucleotide polymorphisms (SNPs), each containing a two–loci interaction (causal SNPs) and 98 "noise" SNPs.We modelled interactions under different scenarios of sample size, missing data, minor allele frequencies (MAF) and several penetrance models:

three involving both (indistinguishable) marginal effects and interaction, and two simulating pure interaction effects. In total, we have simulated 99 different scenarios. Although CART, RF, and LR yield similar results in terms of detection of true association, CART and RF perform better than LR with respect to classification error. MAF, penetrance model, and sample size are greater determining factors than percentage of missing data in the ability of the different techniques to detect true association. In pure interaction models, only RF detects association. In conclusion, tree–based methods and LR are important statistical tools for the detection of unknown interactions among true risk–associated SNPs with marginal effects and in the presence of a significant number of noise SNPs. In pure interaction models, RF performs reasonably well in the presence of large sample sizes and low percentages of missing data. However, when the study design is suboptimal (unfavourable to detect interaction in terms of e.g. sample size and MAF) there is a high chance of detecting false, spurious associations.

5.1.2 Introduction

With the deposition of fresh data on autosomal SNPs in large data repositories such as HapMap [389], population–based studies are becoming very popular among researchers interested in disentangling the genetic causes of complex diseases. Studies of complex diseases such as asthma, schizophrenia, diabetes, etc. generally involve a large number of SNP genotypes.

The determination of which of the SNPs tested modify the risk of disease entails important statistical challenges; even more so when considering the possibility of interaction between SNPs. The definition of epistasis or interaction is highly controversial in the scientific literature [317]. Here, we will use a pragmatic definition of interaction: two or more SNPs contribute jointly to the probability of developing a particular disease. There is an ample spectrum of different statistical approaches for detecting interaction. Logistic regression is probably the most popular one among genetic epidemiologists and geneticists. Over the past few years, the MDR approach [337] has been applied to the analysis of SNPs in many complex diseases [72, 169]. Recently, a number of tree–based techniques, such as CART and RF, have been suggested as suitable for detecting interactions in large–scale association studies [258] or for identifying SNPs predictive of phenotype [56]. The ability of these methodologies to detect association is however a topic of great controversy. Thus, for instance, some authors [147] claim that CART is a suitable technique for detecting interactions, especially in those cases that do not exhibit strong marginal effects [80], while others [394] state that

... CART suffers from the same problem of sequential condi-

> *tioning that can plague many other regression–based methods, which makes it difficult to discover interactions... among predictor variables...*

Since most of the time the detected statistical interaction is considered equivalent to positive real interaction, there is a crucial need for extensive re–evaluation of existing methodologies for detecting SNP interaction in order to minimize the high incidence of false positives in case–control studies [205].

The primary goal of tree–based methods and logistic regression (LR) is to identify SNPs that may increase or decrease susceptibility to disease. This can be achieved by quantifying how much each SNP contributes to the predictive accuracy of these methods by measuring its predictive importance. Finding that a SNP helps distinguishing between cases and controls is an indication that the SNP either contributes to the phenotype or is in linkage disequilibrium with SNPs contributing to the phenotype. While tree–based methods are model–free, the underlying concept of LR is that data follows a given model (logistic function). Tree–based methods are non–parametric statistical approaches for conducting regression and classification analyses by recursive partitioning [177]. The a priori advantage of tree–based methods in comparison with e.g. logistic regression is that they do not require the specification of a model.

Here we examine, by means of simulation of unphased and unlinked SNP genotyping data, the power of different methods for detecting SNP interaction. Several penetrance models, under different conditions of sample size, missing data frequency and minimum allelic frequency (MAF), are considered in order to evaluate which method performs best in different disease scenarios.

5.1.3 Material and methods

5.1.3.1 Simulations

We simulated unphased genotype SNP data using SNaP [301] under several scenarios: different sample sizes (400, 1000 and 2000; 1:1 cases and controls), different percentages of missing data (0, 10 and 20%) equally distributed in cases and controls, and causal SNPs with three different MAFs (0.1, 0.2 and 0.4). In addition, different models of penetrance were also considered. Three of them were built following model 2 of [266]. In this model, the odds of disease have a baseline value, α, unless both loci have at least one disease associated allele. After that, the odds increase multiplicatively within and between genotypes by a parameter, $1 + \theta$. Here we have used $\alpha = 0.05$ and three different values of θ (0.8, 1.4 and 2). The odds were expressed as penetrances, as required by SNaP. Note that these penetrance models involve both marginal SNP effects and SNP–SNP interaction. Besides these, we also simulated two scenarios of pure SNP–SNP interaction, without marginal

PENETRANCES

Pure interaction 1

	BB	Bb	bb
AA	0.3953	0.0015	0.1
Aa	0.0015	0.1328	0.1
aa	0.1	0.1	0.1

Pure interaction 2

	BB	Bb	bb
AA	0.8875	0.0016	0.1
Aa	0.0016	0.1123	0.1
aa	0.1	0.1	0.1

ODDS

Model 2 (Marchini et al.)

	BB	Bb	bb
AA	$\alpha(1+\theta)^4$	$\alpha(1+\theta)^2$	α
Aa	$\alpha(1+\theta)^2$	$\alpha(1+\theta)$	α
aa	α	α	α

Non-pure interaction; $\theta = 0.8$

	BB	Bb	bb
AA	0.3442	0.1394	0.0476
Aa	0.1394	0.0826	0.0476
aa	0.0476	0.0476	0.0476

Non-pure interaction; $\theta = 1.4$

	BB	Bb	bb
AA	0.6239	0.2236	0.0476
Aa	0.2236	0.1071	0.0476
aa	0.0476	0.0476	0.0476

Non-pure interaction; $\theta = 2$

	BB	Bb	bb
AA	0.802	0.3103	0.0476
Aa	0.3103	0.1304	0.0476
aa	0.0476	0.0476	0.0476

Table 5.1: Penetrance matrices for pure and non–pure interaction models. In the top rows we indicate the odd model according to [266]. In order to obtain the penetrance values from the odds in the non–pure interaction models we followed the formula: Penetrance = ODD/(1 + ODD), where $\alpha = 0.05$ and the θ is 0.8, 1.4, and 2 depending on the model. Pure–interaction models are not considered in [266]; therefore, we have employed an *ad hoc* procedure as explained in Subsection 5.1.2. More information concerning these models is shown in Figure 5.2. Image obtained from [149].

SNP effects. Table 5.1 and Figure 5.1 show the different penetrance models in matrix shape and graphically, respectively, as a function of the MAF of the two causal SNPs. Figure 5.2 displays the marginal effects of causal SNPs as a function of their MAFs and genotypes.

All simulated samples consist of unlinked autosomal SNPs: two causal SNPs plus 98 neutral ones (noise). The frequencies of the noise SNPs were randomly taken from a real matrix of SNP frequencies of a panel of autosomal SNPs analyzed in a sample of West European control individuals (data not shown) [410]. The simulated data aims to emulate real case–control association studies where the power to detect real positive association is usually low (due to the presence of large numbers of noise SNPs and/or low penetrance and sample sizes,...).

The combination of the different parameters involved in the simulation consists of 99 different scenarios. Each of them was replicated 100 times in order to account for the randomness of the sampling process. A Perl script (`http://www.perl.org/`) was written in order to reformat the output from the SNaP and to allow the processing of the data according to the different

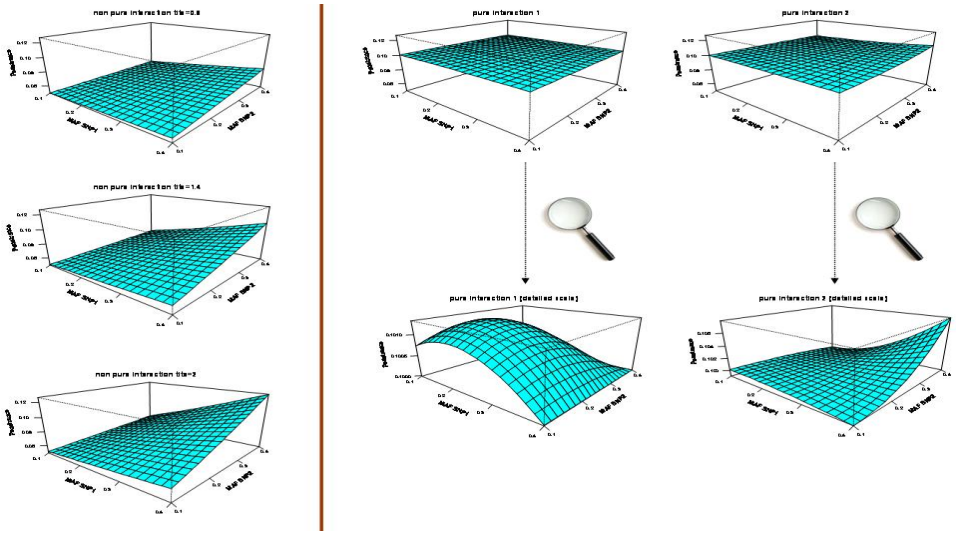

Figure 5.1: 3–D graphics showing the penetrance values as a function of the MAF of the two causal SNPs: non–pure interaction models using $\theta = 0.8$, $\theta = 1.4$ and $\theta = 2$ (left, from top to bottom); and pure interaction models 1 and 2 (right) as defined in Material and methods (see also Figure 5.1). Pure interaction model figures are zoomed to allow an easier interpretation of the effect of MAF on penetrance. Image obtained from [149].

specifications accomplished. This Perl script also generates a random seed for each of the SNaP runs.

5.1.3.2 Statistical methods

CART [1]

The typical approach explained in Chapter 1 to classification trees has been slightly modified here, with the aim of fitting better to the characteristics of genetic data. For instance, as high–dimensionality of data could lead to computational problems, we set to 20 the minimum number of observations required for splitting a node, to avoid the generation of complex and excessive large trees.

Furthermore, trees were pruned following an *ad hoc* procedure: on a first step, we identified the sub–tree with minimum classification error (mCE); on a second step, our choice is the shortest sub–tree whose classification error is below mCE + SD, where SD denotes the standard deviation of the mCE. The aim of this second pruning is to reduce the complexity of the tree with a minimum loss of classification power, leading to an improvement related to

[1] A large part of the explanations about CART, RF, LR and MDR have been removed from [149] to avoid redundancies with Chapter 1. Only issues related to specific performances of these methods are included here.

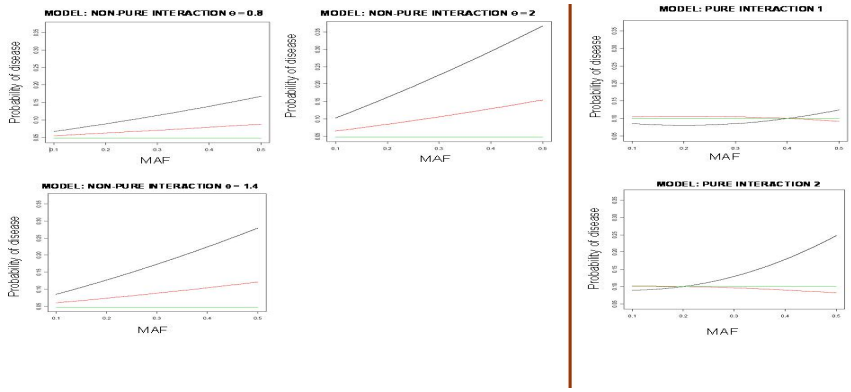

Figure 5.2: Marginal effects of the causal SNPs as a function of their MAF's and genotypes. Each line represents the three possible genotypes: black indicates the homozygote for the rare variant, red the heterozygote, and green the homozygote for the common variant. Note that the disease risk in both the pure–interaction model 1 and 2 for MAF = 0.4 and MAF = 0.2 respectively, is the same for the three possible genotypes. This clearly shows that these SNPs do not show marginal effects for the MAF values employed in our pure–interaction models. These distributions are computed as follows; for the pure-interaction model 1 and using the penetrance values of Figure 5.1: (AA): $0.3953 \cdot p^2 + 0.0015 \cdot 2 \cdot p \cdot (1-p) + 0.1 \cdot (1-p)^2$; (Aa): $0.0015 \cdot p^2 + 0.1328 \cdot 2 \cdot p \cdot (1-p) + 0.1 \cdot (1-p)^2$; (aa): 0.1; being p the allele frequency for the rare allele. The same procedure can be applied to the other models replacing the penetrance values (Figure 5.1) accordingly. Image obtained from [149].

the interpretability of the model. Our choice for pruning has no influence on the association results because we are removing only those branches located far away from the top.

RF We constructed 500 trees in each forest and took $m = 25$ predictor variables in each node. These values yielded the best results in RF but the differences using other parameters were not substantially different (e.g. number of trees ranging from 100 to 1000 at intervals of 100). The recommended number of variables selected for the random set is smaller than $m = 25$. However, for large datasets, a larger number of variables could improve classification errors [48].

For each tree, out–of–bag (OOB) SNP profiles are predicted to a class (i.e., case or control) by running them down the tree. The term OOB profiles refers to those individuals of the training sample which have not been resampled by bootstrap, and therefore have not been used to build the corresponding tree. Each tree gives one vote for each OOB observation, and

the forest prediction for a given observation is the class receiving the most votes (majority vote rule) [245].

Unlike CART, RF requires previous imputation of missing data. For consistency, we also use RF to impute. Specifically, we sequentially consider each SNP as an outcome and the others as covariates, and fill missing data with the values predicted by RF.

LR LR is extremely prone to overfitting. Here we use a stepwise variable selection algorithm based in the Akaike Information Criterion (AIC). Using the function *stepAIC* in R, we choose the best model in terms of having the lowest AIC (beginning from the model without variables and adding or removing variables one–by–one). This should remove a large proportion of the noisy SNPs and therefore alleviate the effect of overfitting.

As in RF, the method is not able to deal with missing data. For each SNP we impute missing data by drawing randomly from a multinomial distribution whose vector of probabilities consists of the genotype frequencies estimated from complete observations for the same SNP.

Multifactor dimensionality reduction (MDR) MDR was used here as a reference method. Since MDR gives rise to a large computational burden, we have applied this technique in one random run per model (one out of 100). MDR treats missing data as a new category [169].

Selection of the best candidate SNP In CART the best candidate SNP is defined as the one located at the root of the tree. In RF, there are two statistical indices that quantify the relative importance of each SNP in the model: the mean decrease accuracy (MDA) and the Gini index. For each tree, MDA records the prediction accuracy on the OOB portion of the data. The prediction accuracy is also recorded for the same OOB portion of the data after permuting the values of the SNP of interest; the differences between the two accuracies are then averaged over all trees and normalized by their standard error. The Gini index is the total decrease in node impurities from splitting on the variable, averaged over all trees. Only the results obtained using MDA are reported here, since they are highly correlated to those obtained using the Gini index. In LR, two coefficients are assigned to each SNP as follows: (i) each SNP is a categorical variable with three categories (genotypes; e.g. $GG/GA/AA$ are initially coded as 1/2/3), and this involves two degrees of freedom, (ii) next, LR transforms these three categories to a binary code by means of creating two dummy variables that consider the presence of one or two rare alleles (status 1 = 0/0, status 2 = 0/1, status 3 = 1/0); and (iii) the SNP having the dummy variable with the most significant p–value is considered the best candidate SNP. In MDR, the best candidate SNPs are the pair of SNPs that better

classify in terms of disease status. As MDR is carried out using a ten-fold cross–validation, the MDR cross–validation consistency counts the number of times a particular pair of SNPs is detected as the best candidate.

It is also worth measuring the significance of the association of each scenario. Thus, p–values can be computed as the probability of detecting two noisy SNPs as the best candidate SNPs at least the same number of times as observed for the causal SNPs. This can be done assuming a binomial distribution with parameters $n = 100$ (number of runs) and $p = 0.02$ (proportion of causal SNPs).

Bayes error rate Table 5.2 displays error rates from Bayes rule in every simulated scenario. Their values do not depend much on sample sizes, percentages of missing data or amount of noisy SNPs. The Bayes error rate provides the lowest achievable error rate for a given pattern classification problem. As a result, Bayes error rate is commonly used as the gold standard for comparing the classification error obtained using a particular statistical approach.

In a two–category problem like the one presented here (case–control), there are two ways in which a classification error can occur [97]. Either an observation has been classified as a case ($\mathbf{x} \in R_2$) and the true state of its nature is a control ($y = -1$), or it has been classified as a control ($\mathbf{x} \in R_1$) and it is actually a case ($y = 1$). Since these sources of error are mutually exclusive and exhaustive, the probability of error can be calculated as follows:

$$\begin{aligned} P(\text{err}) &= P(\mathbf{x} \in R_2, y = -1) + P(\mathbf{x} \in R_1, y = 1) \\ &= P(\mathbf{x} \in R_2 | y = -1)P(y = -1) + P(\mathbf{x} \in R_1 | y = 1)P(y = 1) \\ &= \int_{R_2} P(\mathbf{x}|y = -1)P(y = -1) + \int_{R_1} P(\mathbf{x}|y = 1)P(y = 1) \end{aligned}$$

The Bayes optimal decision rule minimizes this probability, so gives the

MAF	Non-pure interaction			Pure interaction	
	$\theta = 0.8$	$\theta = 1.4$	$\theta = 2$	1	2
0.1	0.48	0.47	0.46	-	-
0.2	0.45	0.41	0.38	-	0.49
0.4	0.38	0.32	0.29	0.42	-

Table 5.2: Bayes error rates for the different models considered in this study. Image obtained from [149].

lowest error probability (Bayes error rate). In our particular problem, it is easy to obtain the a posteriori probabilities of case and control for each individual based on their penetrance and minimum allele frequency (MAF). When the posterior probability of being a case is larger than the one of being a control, the Bayes optimal decision rule allocates this individual as a case, acting in an analogous way when the posterior probability of being a control is larger. Bayes error rates are a function of penetrances and MAF, regardless of the sample size (an infinite population is assumed), percentages of missing data or amount of neutral (noise) SNPs (such parameters are not considered).

Software Most of the methods above have been run in R software (`http://www.r-project.org/`) using the libraries *rpart* [391] for running CART, *randomForest* [245] for RF and *stats* for logistic regression. RF is more computationally demanding, especially due to the imputation procedure applied. For running MDR we have used the MDR software package [169] version 1.1.0. More information concerning the R code employed can be provided upon request.

5.1.4 Results

5.1.4.1 Association

Table 5.3 reports in table format the percentage of runs in which one of the two causal SNPs is detected as the best candidate (only for non–pure interaction scenarios). Such a result will be referred to here as a positive association. CART, RF and LR show very similar outcomes even in the presence of missing data. The latter is somehow surprising given the different nature of the three methods employed and the fact that each of them used different imputation procedures. Among all the parameters considered in the simulations, MAF is the most determinant one regarding association detection, leading to substantial improvements in detection as its value increases. High penetrances and sample sizes are very important too. Strikingly, percentage of missing data seems to be less influential (especially in CART). Figure 5.3 summarizes the performance of CART, RF and LR.

Table 5.4 shows association results for pure interaction scenarios. All the methods yielded poor results in pure interaction model 2. In fact, none of them was able to detect the existing association between the outcome variable and causal SNPs. It is striking that, for pure interaction model 1, RF actually detects association in those cases where sample sizes are high enough or percentage of missing data is low or zero. In such circumstances, CART and LR fail to detect the association. This is an unforeseen discovery, as RF is a generalization of the CART procedure. Notice that in the best association scenario (sample size 2000 and 0% of missing data) RF detects

		MAF = 0.1									MAF = 0.2									MAF = 0.4								
		θ = 0.8				θ = 1.4				θ = 2				θ = 0.8				θ = 1.4				θ = 2				θ = 0.8			
N = 400	MD = 0	2	1	2	-	5	5	4	-	16	16	17	-	9	12	14	-	50	51	53	-	79	83	88	-	71	79	89	-
	MD = 10	1	2	7	-	13	4	9	-	10	11	10	-	21	13	19	-	41	43	48	-	79	69	80	-	73	78	86	-
	MD = 20	1	2	3	-	6	3	3	-	6	7	6	-	21	10	15	-	37	38	40	5/10	67	56	60	-	66	66	72	4/10
N = 1000	MD = 0	9	6	5	-	18	15	29	-	36	31	41	-	50	51	45	-	92	96	97	-	99	100	100	-	100	100	100	-
	MD = 10	4	4	6	-	18	11	19	-	38	30	30	-	49	37	50	-	91	91	88	-	100	99	100	10/10	96	100	100	10/10
	MD = 20	6	6	7	-	8	7	12	-	25	17	27	-	44	36	41	-	86	76	84	-	97	100	100	10/10	97	97	100	10/10
N = 2000	MD = 0	13	7	13	-	40	39	40	-	73	79	84	-	79	85	89	-	100	100	100	10/10	100	100	100	10/10	100	100	100	10/10
	MD = 10	10	3	12	-	39	34	34	-	67	59	63	-	81	79	78	-	99	100	99	10/10	100	100	100	10/10	100	100	100	10/10
	MD = 20	10	10	15	-	41	19	25	-	58	35	55	-	77	61	75	-	99	99	100	10/10	100	100	100	10/10	100	100	100	10/10

		MAF = 0.4									
		θ = 1.4				θ = 2					
N = 400	MD = 0	98	100	99	10/10	100	100	100	10/10		
	MD = 10	98	100	100	10/10	100	100	100	10/10		
	MD = 20	97	98	97	10/10	100	100	100	10/10		
N = 1000	MD = 0	100	100	100	-	100	100	100	10/10	100	100
	MD = 10	100	100	100	10/10	100	100	100	10/10	100	100
	MD = 20	100	100	100	10/10	100	100	100	10/10	100	100
N = 2000	MD = 0	100	100	100	10/10	100	100	100	10/10	100	100
	MD = 10	100	100	100	10/10	100	100	100	10/10	100	100
	MD = 20	100	100	100	10/10	100	100	100	10/10	100	100

Table 5.3: Detection of the association in models of non–pure interaction. For each value of penetrance (θ), we indicate the percentage of runs a causal SNP was selected as the best candidate SNP in CART, RF and LR, respectively (see Subsection 5.1.2); the fourth column of each tetrad indicates the MDR cross–validation consistency (ten–fold cross–validation). A hyphen indicates that the causal variables were not selected as the best candidate SNP pair [337]. MAF = minimum allele frequency; N = sample size; MD = percentage of missing data. Image obtained from [149].

one of the causal SNPs in 85% of the runs. The ability of RF to detect a positive association decreases with lower sample sizes and/or higher amounts of missing data.

Tables 5.5 and 5.6 show the p–values that indicate the significance of the

		Pure interaction 1				Pure interaction 2			
N = 400	MD = 0	4	10	3	-	3	1	4	-
	MD = 10	1	3	1	-	4	2	3	-
	MD = 20	3	5	3	-	2	3	4	-
N = 1000	MD = 0	2	35	1	10/10	3	0	4	-
	MD = 10	1	13	3	10/10	1	2	3	-
	MD = 20	4	5	3	10/10	3	2	1	-
N = 2000	MD = 0	1	85	2	10/10	5	7	1	-
	MD = 10	2	25	0	10/10	0	2	1	-
	MD = 20	3	12	1	10/10	0	2	3	-

Table 5.4: Detection of the association in pure interaction models. The data is presented as in Figure 5.5 (see its legend for details). Image obtained from [149].

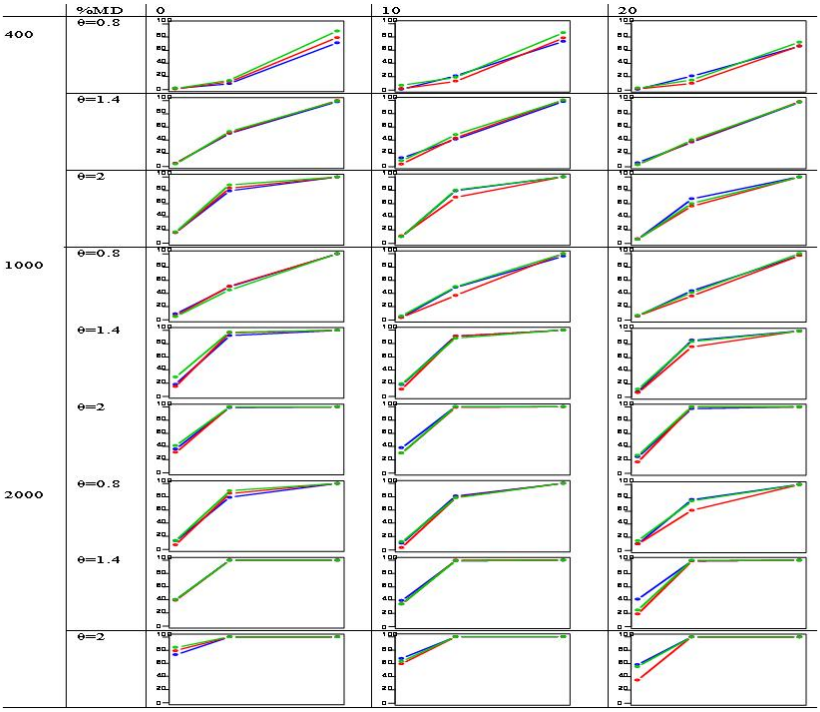

Figure 5.3: Detection of positive association in non–pure interaction models for CART (blue), RF (red) and LR (green). In each scenario, association detection varies as MAF (0.1, 0.2, 0.4) increases. Image obtained from [149].

association for each scenario (see Subsection 5.1.2). These results confirm those summarized in Tables 5.3 and 5.4: all the methods evaluated here fail when MAF and penetrances are not high enough, and with pure interactions. Only RF shows promising results for the most favourable scenarios in terms of association.

5.1.4.2 Performance of MDR versus tree–based methods

MDR performance varies according to the type of interaction between causal SNPs. With non–pure interactions, MDR association results look similar to those obtained with CART, RF and LR, with perhaps a better performance of the latter (Table 5.3).

As occurs with tree–based methods, MDR is unable to detect the existing associations between causal SNPs and the disease in pure interaction model 2. However, with pure interaction model 1, MDR shows a better behavior than RF, since it detects interaction in the presence of missing data, at least up to 20%, and suffers less from shortage of sample size. For sample sizes of 400 cases and controls, MDR is prone to false positives, as is also the case of the tree–based methods and LR (Table 5.4).

In terms of execution requirements, the four methods are very time–consuming. In a standard computer (and only 1GB of memory) it would take from 1 hour (for one of the simplest scenarios; no missing data and 400 cases and controls and using CART) to 45 hours (missing data, sample sizes of 2000 cases and controls, and using RF). CART is the less computationally demanding method, while RF and LR suffer from recursive construction of classification trees and use of a stepwise variable selection AIC criterion, respectively. Finally, computation times for the MDR algorithm are at least three times higher than those of LR.

5.1.4.3 Estimation of classification error

Each run from each simulated scenario was randomly divided into two parts: 80% of the sample is used for training (training sample), and the remaining 20% is employed to (unbiasedly) estimate the classification error (test sample). Imputation of missing data (when needed) is then carried out separately on training and test samples. Classification error of models in each scenario is then averaged over all runs.

CART yields the lowest classification errors (Table 5.7), which are close to Bayes error rates for the most favourable models in terms of MAF (0.4) and no missing data. RF errors are generally only slightly larger than CART

		MAF = 0.1									*MAF* = 0.2									*MAF* = 0.4								
		θ = 0.8			θ = 1.4			θ = 2			θ = 0.8			θ = 1.4			θ = 2			θ = 0.8			θ = 1.4			θ = 2		
N = 400	MD = 0	0.6	0.87	0.6	0.05	0.05	0.14	2E-10	2E-10	2E-11	2E-4	8E-7	1E-8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	MD = 10	0.87	0.6	4E-3	1E-7	0.14	2E-4	3E-5	6E-6	3E-5	9E-16	1E-7	1E-13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	MD = 20	0.87	0.6	0.32	0.02	0.32	0.32	0.02	4E-3	0.02	9E-16	3E-5	2E-9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N = 1000	MD = 0	2E-4	0.02	0.05	2E-12	2E-9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	MD = 10	0.14	0.14	0.02	2E-12	6E-6	1E-13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	MD = 20	0.02	0.02	4E-3	9E-4	4E-3	8E-7	0	2E-11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N = 2000	MD = 0	1E-7	4E-3	1E-7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	MD = 10	3E-5	0.32	8E-7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	MD = 20	3E-5	3E-5	2E-9	0	1E-13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 5.5: Significance of the association in non–pure interaction models. *p*–values were computed as the probability of detecting two noise SNPs as the best candidate SNPs at least the same number of times as observed for the causal SNPs. For each value of penetrance (θ), we indicate the *p*–value obtained with CART, RF and LR, respectively (see Subsection 5.1.2). MAF = minimum allele frequency; N = sample size; MD = percentage of missing data. Image obtained from [149].

		Pure interaction 1			Pure interaction 2		
$N = 400$	MD = 0	0.14	3E-5	0.32	0.32	0.87	0.14
	MD = 10	0.87	0.32	0.87	0.14	0.6	0.32
	MD = 20	0.32	0.05	0.32	0.6	0.32	0.14
$N = 1000$	MD = 0	0.6	0	0.87	0.32	1	0.14
	MD = 10	0.87	1E-7	0.32	0.87	0.6	0.32
	MD = 20	0.14	0.05	0.32	0.32	0.6	0.87
$N = 2000$	MD = 0	0.87	0	0.6	0.05	4E-3	0.87
	MD = 10	0.6	0	1	1	0.6	0.87
	MD = 20	0.32	8E-7	0.87	1	0.6	0.32

Table 5.6: Significance of the association in models of pure interaction. The data is presented as in Figure 5.8 (see its legend for details). Image obtained from [149].

errors, while LR performs worst. These results mirror those obtained for association detection, with the exception of LR. The most plausible reason is the overfitting problem related to LR, aggravated by the fact that the AIC stepwise algorithm employed here is prone to include more neutral (noise) SNPs in the model when either missing data increase or sample sizes decrease. As a result, LR classification errors improve significantly as sample sizes grow. The presence of missing data slightly increases the classification

		MAF = 0.1									MAF = 0.2									MAF = 0.4								
		θ = 0.8 [48.5]			θ = 1.4 [47.3]			θ = 2 [46.3]			θ = 0.8 [44.5]			θ = 1.4 [40.1]			θ = 2 [38.1]			θ = 0.8 [37.7]			θ = 1.4 [32.3]			θ = 2 [28.9]		
$N = 400$	MD = 0	51.1	50.1	49.4	50.4	50.1	49.6	49.7	49.5	49.9	49.3	49.4	49.5	46.6	47.6	48.9	42.3	45	47.7	43	44.3	45.9	35.1	36.7	43.8	31.6	32.2	41.2
	MD = 10	50	48.9	48.7	50.5	49.1	49.9	49	50.1	51	48.4	49	49.4	49	48.6	48.2	43.4	45.5	47.2	44.8	44.3	48.2	36.7	37.6	44	32.1	34.7	40.2
	MD = 20	49.9	48.9	49.2	49.8	49.6	49.6	50	49.5	49.8	49	49.8	49.9	48.2	47.9	47.9	45.4	46.8	48.5	45.8	46.4	47.1	38.2	40.4	44.4	34.5	35.9	41.5
$N = 1000$	MD = 0	49.6	50.3	50.3	50.3	49.7	49.6	48.1	49	48.6	47.7	48.2	49.2	42.6	44.5	45.7	39.4	41.4	44.9	38.4	41.3	42.7	32.4	33.7	37.4	28.8	29.5	34.1
	MD = 10	49.6	50.1	50.1	50.2	49.5	49.8	49.7	48.8	49.3	47.8	48.1	48.5	44.8	46.1	47.1	41.6	42.6	44.7	41	42.3	44.3	34	35.6	38.2	30.7	30.9	35.5
	MD = 20	49.7	49.6	50.3	49.8	49.3	49.2	49.3	49.1	50.1	49	49.6	48.7	45.6	46.6	47.3	42.5	43.2	45.4	41.9	43.5	44.5	35.2	36.6	39.2	32.3	31.9	36.7
$N = 2000$	MD = 0	50.1	50.1	49.9	49.3	49.4	49.7	47.9	48.4	48.9	46.7	47.9	48	42.6	43.4	45.5	38.5	40.7	43.2	37.7	39.7	41.8	32.4	32.9	35.9	28.9	29.1	32.1
	MD = 10	49.7	50.2	49.3	49.8	49.6	49.5	48.4	48.7	48.4	46.8	48.3	48.2	44	44.6	45.6	40.4	41.6	43.1	38.8	40.7	41.9	33.6	34.3	36.4	30.8	30.8	33
	MD = 20	49.6	49.8	49.7	49.2	49.6	49.5	48.6	49.2	49.6	47.9	48	48.4	44.8	44.7	45.9	42.1	42.4	43.8	39.6	41.8	42.9	35.2	34.6	37	32.4	30.8	33

Table 5.7: Classification error in models of non–pure interaction. For each value of penetrance (θ), we indicate the mean classification error using CART, RF and LR, respectively (see subsection 5.1.2 for more details). Codes are as in Figure 5.5. In square brackets (beside θ) we indicate the Bayes error rate for each model. Image obtained from [149].

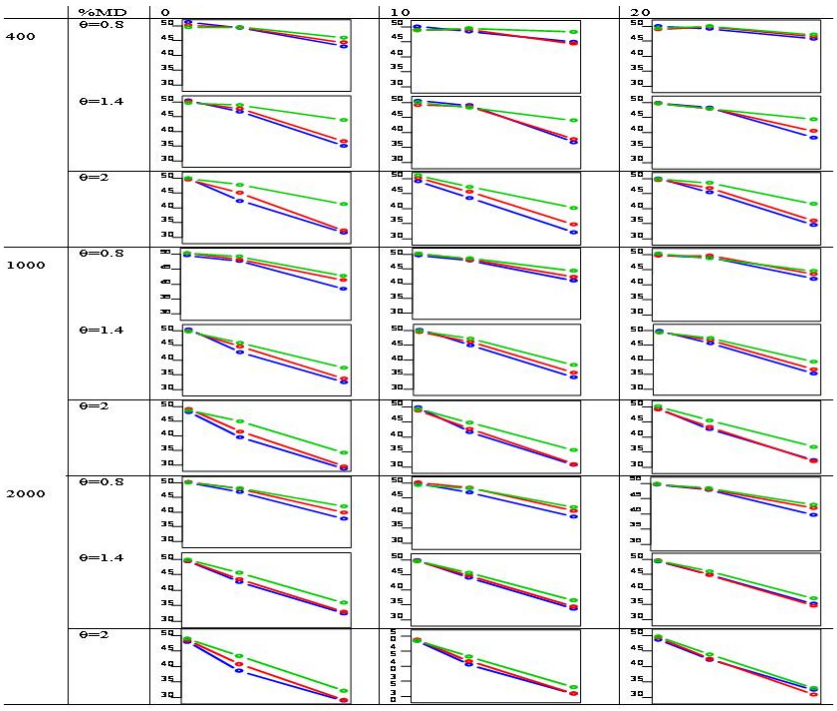

Figure 5.4: Classification error in non–pure interaction models in CART (blue), RF (red) and LR (green). In each scenario, classification error varies as MAF (0.1, 0.2, 0.4) increases. Image obtained from [149].

error. This is especially noticeable for the most favoured models.

None of the methods classify better than tossing a coin when MAF is very low (0.1). MAF is, again, the most determinant of the parameters, followed by penetrance and sample size (Figure 5.4). The effect of MAF is more substantial when combined with high penetrances. Percentage of missing data has little influence on the classification error of CART and RF.

Table 5.8 shows classification error results for pure interaction scenarios. All errors are around 50%. This could be at least expected for pure interaction model 2, bearing in mind that its Bayes error rate (48.6%) is near 50%, while it is unexpectedly discouraging for pure interaction model 1, especially taking into account the reasonable performance of RF in the association results (Table 5.4). A tentative explanation of this result is that RF is able to detect only the effect of one of the two causal SNPs which is insufficient for the detection of two–SNP interaction effects.

5.1.5 Discussion

Human genome analysis and high–throughput techniques are giving rise to a mass of complex biological data. Numbers of candidate SNPs being used

		Pure Interaction 1			Pure Interaction 2		
		[0.42]			[0.49]		
N = 400	MD = 0	48.8	49	49.2	49.6	49	49.8
	MD = 10	49	49.6	49.9	50.5	50	49.8
	MD = 20	49.9	51.1	49.9	50.5	50.4	49.9
N = 1000	MD = 0	50	48.4	49.7	50.2	49.1	50.6
	MD = 10	50.3	50.2	50.1	50.4	49.7	49.1
	MD = 20	50.4	50.1	50	50	51.1	50
N = 2000	MD = 0	50.3	48.9	49.8	49.5	49.6	49.9
	MD = 10	49.7	49.5	50.3	49.8	49.8	49.9
	MD = 20	50.1	49.4	49.9	50.2	49.6	49.6

Table 5.8: Classification error in pure interaction models. The data is presented as in Figure 5.5, thus, each trio of columns shows the results of CART, RF and LR, respectively. The Bayes error rate for each model is included in square brackets. Image obtained from [149].

in association studies are increasing rapidly across a wide range of disease phenotypes. The general discouraging results obtained in case–control association studies of complex diseases [205] are favouring a growing interest in the search for interactions (gene–gene, SNP–SNP, gene–environment,. . .) as a key causal factor in the disease outcome. The analysis of genomic data demands new statistical tools required to deal with one of the major problems in common disease association studies, namely, the curse of dimensionality [394]. Here, we used simulated data for evaluating the ability of three statistical approaches for detecting interaction: CART, RF and LR. Apart from assessing the ability of the different methods for detecting positive association between causal SNPs and the disease we also aimed to estimate the performance of these methods for diagnosis, i.e., to determine their ability to classify disease status as a function of their individual genotype. The latter can be inferred by evaluating the classification error across different simulated scenarios, and comparing them with their corresponding Bayes error rates.

In terms of association, CART, RF and LR yield similar results, though CART seems to be slightly better than the other two, mainly due to its better performance in the presence of missing data. Classification error results mirror those obtained for association detection, except for the fact that differences favoring CART become more pronounced, especially regarding LR. The poor results of LR related to the classification error are more likely due to its sensibility with regard to the curse of dimensionality [179, 394, 432]. The three methods fail to detect weak genotype–disease associations. Their performance improves gradually as the model becomes more favourable (i.e., higher MAF, larger sample sizes, . . .). For the best conditions, classification error measures of CART and RF are close to Bayes error rates, indicating

that the presence of noise SNPs does not interfere significantly with the detection of the interaction and classification [48, 56, 179, 258]. However, the tree–based methods and MDR tested here are probably not useful for large scale genomic projects (e.g. GWAs) if we assume that the candidate SNPs have low penetrances and the sample sizes are not extremely large (as considered here). As seen in the present study, the presence of only 98 noise SNPs could already lead to meaningless conclusions.

In contrast to the conclusions of previous studies [80, 147, 371, 436, 445], our results seem to indicate that CART, as a sequential binary–splitting technique, is not able to discover interactions between predictor variables, unless these predictor variables have an individual effect, independent of the other variables [177, 394]. On the other hand RF has the ability to detect pure SNP–SNP interactions responsible for the disease outcome [179, 258, 371] although at the cost of demanding high sample sizes and low percentages of missing data. The permutation process immanent to the RF procedure allows "amplification" of the SNP–SNP interaction signal [258].

When the SNPs have also some marginal effect, CART and RF perform as well as (or even slightly better than) MDR. However, when there are no marginal effects, MDR is more sensitive to the interaction, even though RF also behaves reasonably well in some circumstances. CART and RF can be very useful in those scenarios that include SNPs with marginal effects, especially when MAFs of the causal SNPs are high enough (above 0.2). These techniques perform well in the presence of a large number of noise SNPs.

We also observed that neutral SNPs are always in Hardy–Weinberg (HW) equilibrium, while causal SNPs show a slight tendency to be in disequilibrium in more runs related to those scenarios with strong association. This could indicate a potential usefulness of the HW test for detecting association, as previously suggested [186]. However, we noticed that this potential is relative if we consider that the maximum percentage of runs in HW disequilibrium for a causal SNP was always below 15%, and this maximum is only sporadically achieved.

Finally, the results indicate that pure interactions are difficult to detect if the scenario is not favourable. Most of the methods considered here have been tested in scenarios where marginal effects are difficult to distinguish from the interaction effects (non–pure interaction models) under different circumstances of MAF, sample size, missing data, ... We are aware that the specific effect of interaction versus marginal effect would require further research.

In addition, we foresee several parameters that could be investigated in future research studies based on simulations (as done in the present chapter). For instance, one could examine in depth different approaches related to imputation of missing data or the effect of linkage disequilibrium in classification error and association detection.

In summary, our results indicate that tree–based methods and LR (with an appropriate variable selection algorithm) can play an important role as statistical tools in large–scale genetic association studies where unknown interactions exist among true risk–associated SNPs with marginal effects and in the presence of a significant number of noise SNPs. In pure interaction models, RF performs reasonably well in the presence of large sample sizes and low percentages of missing data. In our study, its performance is comparable with that of MDR. Empirical simulation studies allow the evaluation of the performance of different statistical tools under controlled conditions. The tree–based methods tested in this chapter could be used as complementary approaches following for instance a two–step strategy: one method (e.g. RF) could be applied for variable selection [56] and other (e.g. CART) for classification [371]. Alternatively, the results obtained could be compared for the three methods when analyzing real data, with the aim of checking to what extent the results coincide. This could indicate something meaningful in terms of the association.

There is a general belief that epistasis does really matter as a risk factor in complex diseases. The lack of proper statistical approaches to deal with the curse of dimensionality is likely one of the causes favoring the unfortunate lack of sensibility and specificity in genomic disease association studies.

5.2 Role of the *ZBTB7* gene on breast cancer development

This section consists of the results included in [350]. Moreover, tree–based results that were obtained for this study and were finally dropped in the final version have been included here, together with two explanatory figures. The format has been adapted to not alter this essay.

5.2.1 Abstract

It has been proposed that the excess of familiar risk associated with breast cancer could be explained by the cumulative effect of multiple weakly predisposing alleles. The transcriptional repressor FBI1, also known as Pokemon, has recently been identified as a critical factor in oncogenesis. This protein is encoded by the *ZBTB7* gene. Here we aimed to determine whether polymorphisms in *ZBTB7* are associated with breast cancer risk in a sample of cases and controls collected in hospitals from North and Central Spanish patients. We genotyped 15 SNPs in *ZBTB7*, including the flanking regions, with an average coverage of 1 SNP/2.4 Kb, in 360 sporadic breast cancer cases and 402 controls. Comparison of allele, genotype and haplotype frequencies between cases and controls did not reveal associations using Pearson's chi–square test and a permutation procedure to correct for multiple test. In this, the first study of the *ZBTB7* gene in relation to, sporadic breast cancer, we found no evidence of association.

5.2.2 Introduction

It has been suggested that breast cancer, together with prostate and colorectal, are the cancers with the highest heritable components. A substantial proportion of familiar breast cancer (∼25%) is explained by mutations in the *BRCA1* and *BRCA2* genes [135, 393]. By contrast, the excess of familiar risk associated with sporadic breast cancer (as well as the unexplained genetic risk in familiar breast cancer) may be better explained by the effect of multiple weakly predisposing alleles [19, 312]. The identification of common alleles conferring modest susceptibility to cancer (as opposed to the known high penetrance *BRCA1/2* genes) is a field of growing interest, especially with the development of new genotyping techniques and SNP database facilities [389].

Hence, there is much interest in the search for gene/variants with low penetrance for breast cancer, which could exist with relatively high prevalence in the general population. Many polymorphisms have been proposed

as candidates for susceptibility to sporadic breast cancer but reported positive associations have rarely been replicated in independent studies [15, 35, 292, 425].

Recently [264], the transcriptional repressor FBI1, namely Pokemon (POK erythroid myeloid ontogenic factor), was identified as a critical factor in oncogenesis. This protein is encoded by the *ZBTB7* gene ("zing finger and BTB domain containing 7"; Gene ID: 51341). Mouse embryonic fibroblasts lacking *ZBTB7* are completely refractory to oncogene–mediated cellular transformation. Conversely, FBI1 over–expression led to overt oncogenic transformation both in vitro and in vivo in transgenic mice. FBI1 can specifically repress the transcription of the tumor suppressor gene *ARF* (600160). In [264], it was found that FBI1 is aberrantly over–expressed in human cancers, and its expression levels predict biologic behaviour and clinical outcome. On the other hand, tissue microarray (TMA) analysis in breast carcinomas has revealed high levels of Pokemon expression in a subset of these tumours. In addition, the genomic region where the *ZBTB7* gene resides (19p13.3) is a hotspot for chromosomal translocations (The Cancer Genome Anatomy Project; `http://cgap.nci.nih.gov/`). *ZBTB7* is therefore a good candidate low penetrance breast cancer susceptibility gene.

Here we aim to study the potential implications of common *ZBTB7* variants in sporadic breast cancer in a sample of cases and controls from Spain. To do this, we selected a set of 19 SNPs covering the whole extension of *ZBTB7* and flanking regions at high density.

5.2.3 Material and methods

5.2.3.1 Study subjects and DNA extraction

Cases were 360 Spanish women with breast cancer and mean age at diagnosis of 59 years (range 25 to 85 years), recruited between 2000 and 2004 (48% of cases were recruited within one year of their diagnosis and 79% within five years). All cases were collected from a consecutive series recruited via three public Spanish hospitals: Hospital La Paz (20%), Fundación Jiménez Díaz (50%) and Hospital Monte Naranco (30%). Our samples contain prevalently invasive cases of breast cancer, 96%; while only 4% of *in situ* breast cancer. Controls were 402 Spanish women free of breast cancer at ages ranging from 24 to 85 years (mean = 53 years) and recruited between 2000 and 2005, via the Menopause Research Centre at the Instituto Palacios (50%), the Colegio de Abogados (31%) and the Centro Nacional de Transfusiones (19%), all in Madrid. While data was not available to calculate response rates, our experience is that response rates are very high for cases (∼90%).

Genomic DNA was isolated from peripheral blood lymphocytes using automatic DNA extraction (Magnapure, Roche) according to the manufacturer's recommended protocols. DNA was quantified using picogreen and

diluted to a final concentration of 50 ng/ul for genotyping. Informed consent was obtained from all participants and the study was approved by the institutional review boards of Hospital Clínico Universitario (Santiago de Compostela, Galicia, Spain) and Hospital La Paz, Madrid.

5.2.3.2 SNP selection

SNPs were selected from different sources: the International HapMap Project (The International HapMap Consortium, 2003; 2004; `www.hapmap.org`), Ensemble (Birney *et al* 2004; `www.ensemble.org`), the Sequenom RealSNP database (`www.realsnp.com/default.asp`), and PupaSNP (Conde *et al* 2004; `www.pupasnp.org`). All 22 SNPs described at the time of selection were included, which yielded an average coverage of 1 SNP/1.7 Kb. These SNPs cover the upstream and downstream flanking regions (10000 bp) and the introns of *ZBTB7*, and include only one coding non–synonymous SNP (Table 5.9).

5.2.3.3 SNP genotyping

Genotyping was performed using the MassARRAY SNP genotyping system (Sequenom Inc., San Diego, CA) located at the Universidad de Santiago de Compostela node of the Spanish National Genotyping Center (Centro Nacional de Genotipado; `http://www.cegen.org`), following the manufacturer's instructions. This typing assay uses the extension of a single primer

ID	SNP	Alleles	MA	MAF (%)	Chr. 19 location in Ensembl v.31	aa change	Position (bp)	Intermarker Distance (bp)	Distance from v01 (bp)
–	rs10405522	C/G	G	2.4	3'-down-stream	–	3990056	–	–
v1	rs11882361	A/G	G	2.9	3'-down-stream	–	3993858	3802	3802
v2	rs2121137	C/T	T	2.6	3'-down-stream	–	3993923	65	3867
v3	rs2121136	C/T	T	2.6	3'-down-stream	–	3994024	101	3968
v4	rs2121135	A/G	A	0.2	3'-down-stream	–	3994038	14	3982
v5	rs10414035	A/G	G	0.3	3'-down-stream	–	3996434	2396	6378
v6	rs10412825	A/G	A	0.1	3'-down-stream	–	3996962	528	6906
v7	rs4807540	C/T	T	4.0	intronic	–	4001399	4437	11343
v8	rs7251080	C/T	T	0.8	coding	341syn	4005208	3809	15152
v9	rs3745985	C/T	T	0.3	coding	A177T	4005702	494	15646
v10	rs10409301	C/T	T	4.7	intronic	–	4007346	1644	17290
v11	rs895331	G/T	G	2.1	intronic	–	4011222	3876	21166
–	rs895330	C/G	C	22.7	intronic	–	4011707	485	21651
v12	rs350842	C/T	C	1.1	intronic	–	4012781	1074	22725
v13	rs350841	A/G	A	0.9	intronic	–	4014067	1286	24011
–	rs350840	C/G	C	1.9	intronic	–	4015418	1351	25362
v14	rs1992710	C/G	C	0.2	intronic	–	4015672	254	25616
–	rs350832	C/T	C	22.9	5'-upstream	–	4020426	4754	30370
v15	rs11880023	C/T	T	21.1	5'-upstream	–	4025697	5271	35641

Abbreviations: MA: minor allele; MAF: control minor allele frequency; aa: aminoacid.

Table 5.9: *ZBTB7* SNPs successfully genotyped. Image obtained from [350].

that binds to the sequence flanking the mutation site. Base–specific primer extension products are created 1–4 bases long depending on the substitution present. The different primer extension products are then differentiated by mass. Multiple sites can be typed simultaneously by multiplexing the extension reaction. Detection uses matrix–assisted laser desorption ionization time–of–flight (MALDI–TOF) mass spectrometry with samples automatically genotyped from each mass spectrum produced. The assays were designed using Spectro DESIGNER software. Case and control samples were genotyped using 384–well plates and automated protocols. The allele–calling of all possible SNPs in each DNA sample was performed automatically using SpectroTYPER–RT software. Positive and negative controls were incorporated in each genotyping plate in order to assess genotyping quality. We estimate a genotyping error rate below 0.001%.

5.2.3.4 Statistical analyses

We tested for differences in allele frequencies between cases and controls using Pearson's chi–squared test (the best model is provided in Table 5.10). We adjusted for age in categories <45, 44–49, 50–54, 55–59, and >60 via logistic regression using Stata v8. Disequilibrium coefficients (D') for adjacent SNPs were calculated using Haploview v3.11 [29]. We used Gold software [1] to graphically summarized patterns of linkage disequilibrium in *ZBTB7* because it is well suited to the analysis of dense genetic maps. Assuming a minimum allele frequency (MAF) of 3% (the average MAF of our SNP set) and a genetic effect of 2, the a priori power to detect association under a dominant model is above 70%.

Haploview v3.32 (`www.broad.mit.edu/mpg/haploview`) was used for estimating the genotyping coverage of the selected SNPs (see below) and haplotype block structure.

The Cocaphased program of the Unphased software package [98] was used to check for single SNP and haplotype associations. We tested all two, three, four, and five–SNP haplotypes for association in a sliding window across the gene. The option "drop rare haplotypes" was used in order to restrict the analysis to the haplotypes with a frequency $> 1\%$. We followed the permutation test procedure implemented in Unphased which provides p–values corrected for the multiple haplotypes tested. The EM algorithm was used to impute missing data.

Evaluation of stratification was carried out based on the genotyping of 28 neutral SNPs, as previously described in a separate study that targeted a different set of low penetrance breast cancer genes in overlapping samples [282].

Two different tree–based methods were used to assess the importance of the genotyped SNPs in determining breast cancer risk: CART and RF. In CART, the tree is built in two steps, growing and pruning. First, the tree is

SNP	Rare Allele	Best Model	OR[1]	95% CI	Un-adjusted *P*-value
rs11882361	G	Dom.[2]	1.30	0.70–2.42	0.4
rs2121137	A	Dom.[2]	1.32	0.69–2.53	0.4
rs2121136	T	Dom.[2]	1.36	0.69–2.68	0.4
rs2121135	T	Dom.[2]	0.54	0.05–5.96	0.6
rs10414035	G	Dom.[2]	2.18	0.20–24.1	0.5
rs10412825	A	Dom.[2]	1.11	0.07–17.7	0.9
rs4807540	T	Dom.[2]	0.95	0.54–1.70	0.9
rs7251080	T	Dom.[2]	1.91	0.51–7.18	0.3
rs3745985	T	Dom.[2]	1.08	0.15–7.71	0.9
rs10409301	T	Rec.	0.27	0.06–1.29	0.1
rs895331	C	Dom.[2]	0.99	0.48–2.06	0.9
rs350842	C	Dom.[2]	0.70	0.24–2.08	0.5
rs350841	A	Dom.	0.78	0.24–2.47	0.7
rs1992710	C	Dom.[2]	0.52	0.05–5.81	0.6
rs11880023	T	Rec.	0.69	0.32–1.48	0.3

[1]Using common-allele homozygotes as reference. The *P*-value refers to a Pearson's Chi-squared test.
[2]Only model that could be fit due to zero counts for rare homozygotes.

Table 5.10: OR and p–value for the best fitting model. Image obtained from [350].

allowed to grow to its maximum. Then, CART executes a process of pruning that consists of selecting the sub–tree with the minimum classification error Err_{CART}. The selected tree is further pruned using one standard deviation (SD) relative to Err_{CART}, that is, we take the shorter tree within the range:

$$\text{Err}_{CART} \pm \text{SD}$$

The samples were divided into a training set (∼80% randomly selected cases and controls) and a test set (the remaining ∼20%). A complexity parameter was obtained as a measure of the improvement in classification error as the tree was grown. The main role of this parameter is to save computing time by pruning off splits that are obviously not worthwhile. With RF, the training set used to grow each tree was a 608–individuals bootstrap sample taken from ∼80% of the observations. The number of trees built for each model was set to 1000 and m, the number of SNP markers randomly chosen to split at each node, was set to 10. We computed the classification error and we also recorded the mean decrease in accuracy (MDA) that allows weighting the relative importance of the different markers in the model built by RF. The *rfImpute* function of the *randomForest* library in R was used to impute missing genotypes.

5.2.4 Results and discussion

Three out of the 22 SNPs selected failed genotyping. Four out of the 19 remaining SNPs (namely, rs10405522, rs895330, rs350840, and rs350832) were successfully genotyped in less than 75% of the samples and were therefore

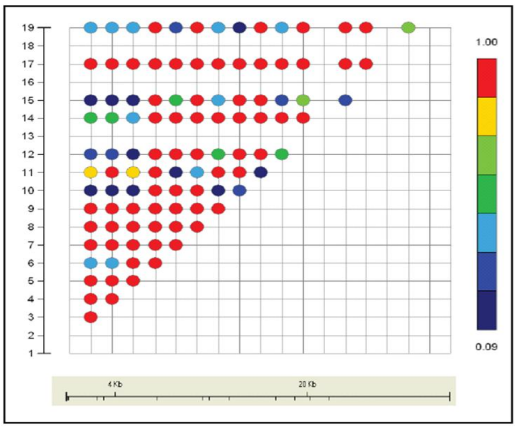

Figure 5.5: D' pairwise linkage disequilibrium values of *ZBTB7* markers in control individuals. Image obtained from [350].

excluded from association analyses. The average call rate for these 15 SNPs was 95% (see also preliminary results in [351, 352, 411]) and none gave evidence of deviation from Hardy–Weinberg equilibrium. Table 5.9 summarizes their location and allele frequencies.

We computed D' values between all 19 markers, and detected moderate levels of linkage disequilibrium (LD) (Figure 5.5). However, under the "four gamete rule" model (see Haploview for more information) we identified haplotype blocks nearly covering the entire extension of the gene (Figure 5.6). This characterization of LD along the *ZTBT7* region could be useful for future association study designs in cancer.

In order to measure the percentage of variability captured by the our selected SNPs, we first collected the HapMap data from the CEPH subset (`http://www.hapmap.org`) and the same chromosome range explored in the present study (chromosome 19: positions 3990056–4025697). Then, we estimated the number of SNPs un–captured in the CEPH–HapMap using our SNP selection under an r^2 threshold of 0.8 and a model of "aggressive tagging". Only one SNP in the HapMap dataset would remain untagged by our selected SNPs, indicating that our set of SNPs covers well the whole gene region under analysis.

No statistically significant differences between cases and controls were observed for individual SNPs based on comparisons of allele frequencies (see Table 5.10 for the best fitting models) whether or not age was adjusted for. Four and three–SNPs haplotypes carrying markers rs350842 and rs350841 had associated p–values below 0.05 but were not significant after correction for multiple testing. Note also that these adjusted p–values overestimates the real value since the software employed (Cocaphased) does not correct for the multiple hypothesis tested running different sliding windows.

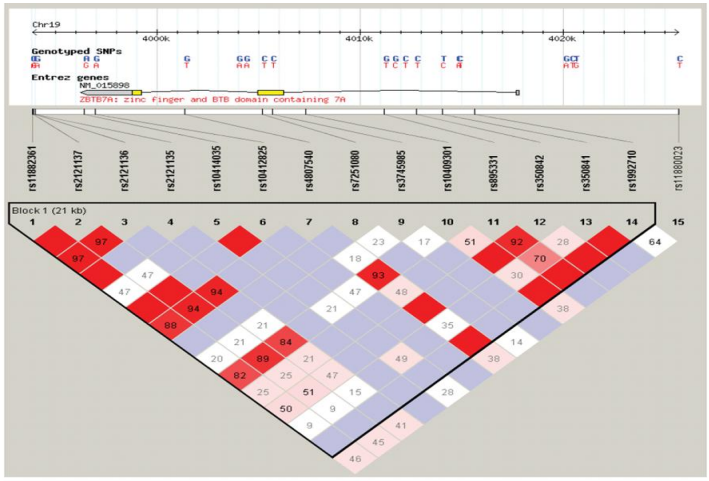

Figure 5.6: Haplotype block structure in our control individuals and HapMap information for the CEPH dataset (top). Image obtained from [350].

We applied two different tree–based methods in order to estimate the relative importance of the genotyped SNPs in our sample of breast cancer patients. CART produced a complex tree of 21 leaves. The tree with 13 leaves (Figure 5.7 left) is simpler and retains only 9 different variables (SNPs) when pruned. Figure 5.7 (right) shows the evolution of the training error as the tree grows. The root node gives a classification error of 0.472 (benchmark). The model constructed by CART gives 0.457, indicating that it performed only marginally better in predicting the disease outcome than tossing a coin. Difference with the benchmark is obviously not statistically significant. Figure 5.8 shows the 15 SNP markers sorted by their MDA values. The most associated SNP in RF, as determined by its higher MDA value in comparison with the rest of SNPs, was rs350841.

To our knowledge, this is the first time that *ZBTB7* has been evaluated as a candidate sporadic breast cancer susceptibility gene. We have not found evidence of an association for *ZBTB7* SNPs nor haplotypes with breast cancer risk. It should be mentioned that most of the *ZBTB7* variants studied are rare in our sample. We are aware that the main drawback in detecting positive associations of rare variants (or haplotypes) is the need for large sample sizes. Therefore, the present result needs further validation in future studies of independent case–control series before a role for *ZBTB7* in breast cancer can be completely ruled out.

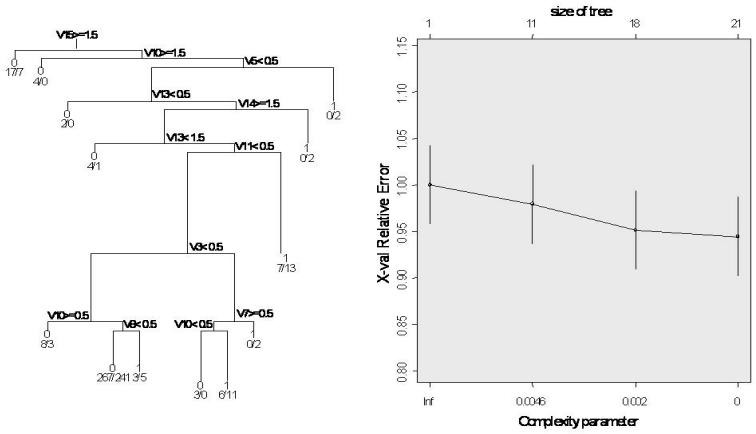

Figure 5.7: Classification tree for the *ZBTB7* case–control data (left) and 2D graphical plot showing the evolution of the training error as the tree grows (right).

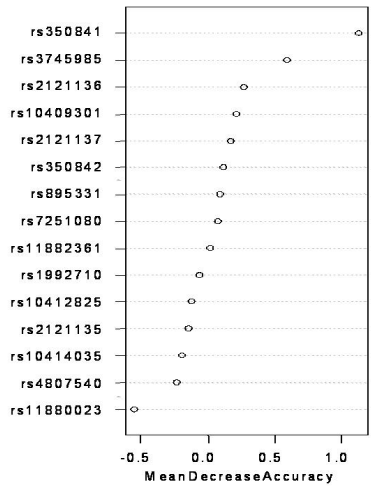

Figure 5.8: R plot displaying the 15 SNPs of the study sorted by their MDA value.

Chapter 6

Statistics in non–clinical genetics: intricate problems in forensic genetics

6.1 Resolving relationship tests that show ambiguous STR results using autosomal SNPs as supplementary markers

This section contains the work published in [314]. We have adapted it to the format and notations used in this essay. Here, statistics help to show how the addition of a set of SNP markers can be very useful to solve complex paternity tests. Although the statistical concepts which appear here were developed long time ago, this work is a model of how to apply common statistical results and intensive simulation to forensic genetic problems. Furthermore, simulation can be really helpful in genetic studies, as obtaining of human genetic data is sometimes a path full of obstacles (for instance, economic costs).

6.1.1 Abstract

When using a standard battery of STRs for relationship testing a small proportion of analyses can give ambiguous results – where the claimed relationship cannot be confirmed by a high enough paternity index or excluded with fully incompatible genotypes. The majority of such cases arise from unknowingly testing a brother of the true father and observing only a small number of exclusions that can each be interpreted as one– or two–step mutations. Although adding extra STRs might resolve a proportion of cases, there are few properly validated extra STRs available, while the commonly added hypervariable SE33 locus is four times

more mutable than average, increasing the risk of ambiguous results. We have found that SNPs in large multiplexes are much more informative for both low initial probabilities or ambiguous exclusions and at the same time provide a more reliable genotyping approach for the highly degraded DNA encountered in many identification cases. Eight relationship cases are outlined where the addition of SNP data resolved analyses that had remained ambiguous even with extended STR typing. In addition we have made simulations to ascertain the frequency of failing to obtain exclusions or conclusive probabilities of paternity with different marker sets when a brother of the true father is tested. Results indicate that SNPs are statistically more efficient than STRs in resolving cases that distinguish first–degree relatives in deficient pedigrees.

6.1.2 Introduction

Most laboratories performing relationship testing will rely on the core forensic sixteen–marker short tandem repeat (STR) sets to obtain an exclusion or strong probability of paternity (i.e. reaching virtual proof). However a small proportion of cases show ambiguous results where the claimed relationship cannot be confirmed by a high enough probability or when an exclusion is suggested by just one or two loci. A large proportion of ambiguous results arise from unknowingly testing a first–degree relative of the true father, usually a brother, so the exclusion rate is markedly reduced and a paternity index using a likelihood ratio against a random man does not apply. Less frequently, ambiguous STR results occur from observing exclusions that may originate from germ–line step mutations (see `www.cstl.nist.gov/biotech/strbase/mutation.htm` and also `www.aabb.org/Documents/Accreditation/Parentage\_Testing\`) [52]. These mutations are characterized by one or two repeat additions or diminutions creating an incompatibility that is impossible to distinguish as a mutation or an exclusion. Ambiguous genotypes are particularly difficult to interpret when a brother of the true father is unknowingly tested, as this reduces the total excluding loci. The main recourse for laboratories finding such results is addition of extra STRs to improve the probability or provide clear, unambiguous exclusions. However outside of the principal commercial kits few additional autosomal STRs are validated and readily applicable. Another source of ambiguity is second order exclusions created when primer binding site substitutions lead to the dropout of an amplifiable allele in both parent and offspring. This phenomenon is observed more frequently in certain STRs (`www.cstl.nist.gov/biotech/strbase/NullAlleles.htm`) and the normal approach is to use complementary marker sets testing identical loci with alternative primer designs [10, 127, 291].

For relationship testing we use extended STR sets comprising 17 markers in two complementary kits: Identifiler® and Powerplex® 16 plus singleplex STRs: D1S1656, D12S391, D18S535 and SE33. Supplementary genotyping has been developed in–house, with three STRs extensively characterized during their initial forensic optimization [228, 229, 230]. This choice of kits plus standalone STRs benefits from using 13 loci common to each marker set with different primer sites to help detecting dropout, plus eight unique STRs providing powerful extra discrimination. In the past three years we have added single nucleotide polymorphisms (SNPs) to STR analysis in an increasing proportion of complex or deficient relationship tests. Although SNPs have a much lower discriminatory power per locus than STRs, we have used a standardized forensic 52plex assay [354] that matches or exceeds the discriminatory power of 15 STRs. Notably SNPs applied to relationship testing offer a much lower overall mutation rate, typically: $\mu = 2.5 \times 10^{-8}$ compared with $\mu = 10^{-3}$ to 10^{-4} in STRs but the 52plex has provided an ideal complementary approach for three additional reasons: (i) the genomic positions of the 52 SNPs are well spaced, both as a set and in relation to common STRs, to facilitate segregation between related individuals; (ii) SNPs, as binary polymorphisms, are more likely than multi–allelic STRs to show informative second order exclusions in deficient cases (i.e. lacking all pedigree members) and; (iii) the 52plex amplified fragments are all less than 120 bp offering greater success than standard STRs with highly degraded DNA [133, 134, 354]. Since a small but consistent proportion of relationship tests we perform involve analysis of human remains, this last characteristic of SNPs provides an important way to avoid a further source of ambiguous results with STRs: uninformative paternity probabilities resulting from incomplete profiles commonly obtained from degraded DNA.

We outline eight cases that failed to give a clear, unequivocal indication of the claimed relationship with STRs alone. Each one showed that adding SNPs improved the paternity index or successfully resolved ambiguous STR exclusions.

6.1.3 Materials and methods

6.1.3.1 Marker sets used

Table 6.1 outlines the 21 STRs used, based on two commercial STR mutliplexes: Identifiler® (Applied Biosystems, Foster City, CA) and Powerplex® 16 (Promega, Madison, WI) providing complementary primer set analysis of 13 loci and two specific to each set plus supplementary singleplex STRs: D1S1656, D12S391, D18S535 and SE33. SNP analysis was based on the well–established SNPforID 52plex assay previously described [354] (supplementary data at: `www.snpforid.org/publications.html`) and shown to be informative for forensic identification [133, 134, 313, 354].

Identifiler®			Powerplex® 16			Supplementary STRs		
No.	STR	μ	No.	STR	μ	No.	STR	μ
1	CSF1PO	0.16		CSF1PO		18	**SE33**	0.64
2	**D2S1338**	0.12	16	**Penta D**	0.14	19	**D1S1656**	0.16*
3	D3S1358	0.12		D3S1358		20	**D12S391**	0.16*
4	D5S818	0.11		D5S818		21	**D18S535**	0.16*
5	D7S820	0.1		D7S820				
6	D8S1179	0.14		D8S1179				
7	D13S317	0.14		D13S317				
8	D16S539	0.11		D16S539				
9	D18S51	0.22		D18S51				
10	**D19S433**	0.11	17	**Penta E**	0.16			
11	D21S11	0.19		D21S11				
12	FGA	0.28		FGA				
13	TH01	0.01		TH01				
14	TPOX	0.01		TPOX				
15	vWA	0.17		vWA				

Table 6.1: STRs sets used and their reported percentage mutation rates (μ). Values of μ with * denote an average rate used in the absence of current estimates. STRs in bold show non–complementary markers analyzed using single primer pairs. Image obtained from [316].

6.1.3.2 Statistical analysis

All STR and SNP genotypes were compared amongst tested individuals using *Familias* pedigree analysis software [115] and locally derived (NW Spain) allele frequencies (SNP data in the SNPforID frequency browser: `spsmart.cesga.es/snpforid.php`). In all cases where a paternity index is given as a percentage probability (P) an a priori value of 0.5 was always used. The *Familias* program specializes in suggesting the most likely relationship given the genotypes of tested individuals by calculating the probability of given sets of possible pedigrees. When second order exclusions and step mutations are observed *Familias* is able to factor in specific mutation rates for the loci to compile a probability of the defined relationships. We added values for μ reported in STRbase and listed in Table 6.1, with range: $\mu = 0.0001$ for TPOX/TH01 to $\mu = 0.0064$ for SE33, with D1S1656, D12S391, D18S535 using an average value of 0.0016 in the absence of current estimates. A universal SNP mutation rate of $\mu = 2.5 \times 10^{-8}$ was used – to date the 52plex SNPs have been validated in trios and extended families without detecting second order incompatibilities for nearly all the SNPs [41, 313].

6.1.3.3 Simulation of testing a first–degree relative of the true father

We developed a computer program in R (`www.r-project.org/`) to assess the probability $P(B)$, that a paternal first–degree relative (simplified here to "brother" but applicable to the father or a son of the true father) has been

tested and is fully compatible with paternity for different combinations of markers, such as 21 STRs or STRs plus 52 SNPs. $P(B)$, for loci: $i = 1, \ldots, l$, can be defined as:

$$P(B) = \prod_{i=1}^{l} P(B_i)$$

for each locus:

$$P(B_i) = P(B_i|C1)P(C1) + P(B_i|C2)P(C2) + P(B_i|C3)P(C3)$$

where $C1$ = two alleles shared by the true father and a brother, so the probability of $C1$: $P(C1) = 0.25$; $C2$ = one allele shared, $P(C2) = 0.5$; $C3$ = no alleles shared, $P(C3) = 0.25$, and $P(B_i|C1)$, $P(B_i|C2)$ and $P(B_i|C3)$ are calculated from the allele frequencies and mutation rates for each locus i. This allowed estimation of the expected proportion of cases where no exclusions are detected in a brother. Additionally we simulated child–father–brother pedigrees to estimate the paternity index considering two exclusive hypotheses: a brother being the true father against a random man being the true father. More details of the algorithms are available on request.

6.1.3.4 Relationship tests examined

The eight cases showing ambiguous STR results can be categorized: (i) a simple disputed paternity trio: 44p06; (ii) paternity analysis of aged, degraded skeletal remains: 70p06 and 20p07; (iii) sibship analysis differentiating half from full sibs: 24p07 and 28p07; (iv) a sib versus paternity counter–claim (individual A claims to be the son of B, B claims to be the half sib of A): 45p06; (v) testing of a sib as proxy for the deceased claimed father: 39p04 and 123p04. With the exception of simple trio 44p06, all families analyzed were deficient, i.e. lacking the mother or the supposed father. Figure 6.1 gives the explanatory pedigrees showing alternative relationships analyzed for 44p06 plus the two most complex cases: 123p04, and 45p06.

6.1.4 Results

Results can be divided into two groups: (i) three cases showing ambiguous STR exclusions resolved by adding SNPs, (ii) five cases with uninformative paternity indices improved by adding SNPs, two due to partial STR profiles obtained from degraded bones that gave near–complete SNP profiles in parallel genotyping.

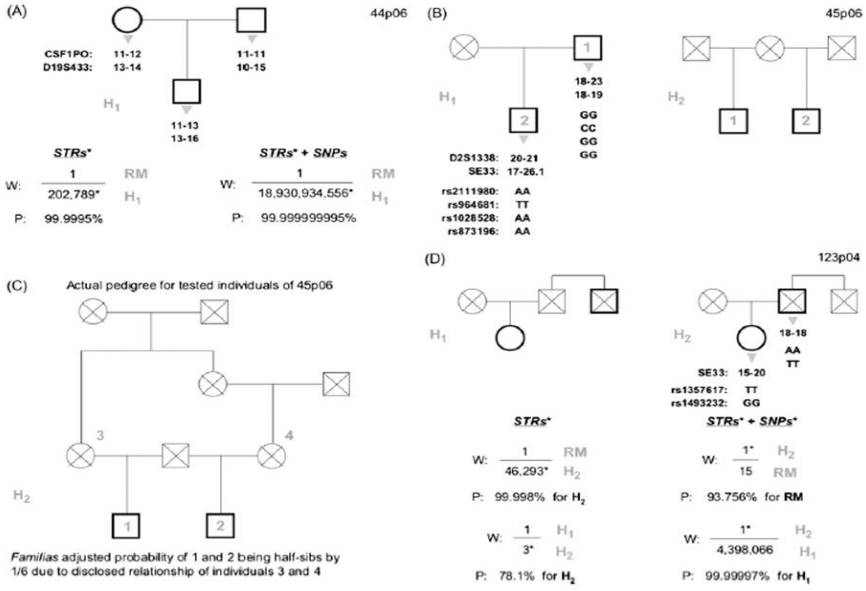

Figure 6.1: Three cases showing ambiguous exclusions with likelihood ratios (W) and probabilities (P) for the most likely relationships (hypotheses: H_1, H_2 or RM = random man) using 21 STRs on the left and with the addition of 52 SNPs on the right. Bold pedigree components denote the tested individuals. Likelihoods marked with * were calculated adding mutation rates for excluding loci. Panel C shows the actual pedigree for 45p06 (previously summarized in panel B) disclosed by the family during testing. Image obtained from [316].

6.1.4.1 Cases with ambiguous exclusions

Cases 44p06 and 45p06 (Figure 6.1 A and B respectively) showed an interesting contrast in their final interpretations although both gave two 1– or 2–step genotype differences after typing 21 STRs. In the simple trio 44p06 these comprised a maternal one–step or paternal two–step incompatibility in CSF1PO plus a maternal two–step or paternal one–step incompatibility in D19S433. A reasonable interpretation at this stage would be that two independent mutations are highly unlikely so the tested man is excluded although he may be closely related to the true father. A high paternity index when factoring in the mutation rates also suggested that a brother of the true father could have been tested, but this case remained ambiguous because the incompatibilities were each one– or two–step differences. The addition of SNPs resolved the case since the final paternity index from STRs and SNPs combined with mutation rates, reached 99.999999995% with a predicted probability of failing to exclude a brother of 0.00017 (final row, Table 6.2).

The sib versus paternity counter–claim case 45p06 had a deficient pedigree: compromising the ability to unambiguously exclude the tested man, while the alternative possibility that the tested men were half–sibs also reduced the excluding power. Additionally the excluding STRs showed one– or two–step differences possible from either paternal allele in both D2S1338 and SE33. Adding SNPs provided four independent second order exclusions emphasizing the enhanced ability to resolve deficiency cases provided by binary markers. In fact this case proved to be more challenging than originally supposed, as the true pedigree disclosed by the family showed that one man was the offspring of the others aunt (Figure 6.1 C), with *Familias* allowing a straightforward adjustment to the probability estimates.

Case 123p04, outlined in Figure 6.1 D, was a fully deficient pedigree (both parents deceased) testing the brother of the deceased man. The tested man claimed paternity of the sole offspring (a daughter, precluding mitochondrial and Y–chromosome analysis). STR analysis gave a single two–step incompatibility in SE33. Factoring in the mutation rate of SE33 gave a probability of paternity against a random man of 99.9978%, but more significantly paternity for the tested man was three times more likely than for the deceased. As SE33 has a mutation rate four times higher than average but the probabilities were not considered strongly indicative of paternity this case remained ambiguous. Addition of SNPs provided two further exclusions of the tested man and, more importantly for resolving the case, when conservative mutation rates of $\mu = 0.00001$ were included for each SNP *Familias* gave a 99.99997% probability in favour of paternity for the

Marker set (number of loci)	Probability of no detected exclusions in a brother (standard deviation in brackets)	Proportion of PI values higher than 1 (%)
Identifiler® (15)	0.02657 (0.011)	6.9
MiniFiler® (8)	0.12044 (0.035)	11.9
Powerplex® 16 (15)	0.02503 (0.01)	6.4
Profiler Plus® (9)	0.09648 (0.029)	10.7
Identifiler® + Powerplex®16 (17)	0.01395 (0.006)	5.0
17 core STRs + 4 supplementary	0.00277 (0.001)	2.4
SNP*for*ID ID-SNPs (52)	0.05165 (0.018)	6.1
21 STRs + 52 SNPs	0.00017 (0.0001)	0.5

Table 6.2: Predicted probabilities of a brother of the true father being compatible with paternity (no exclusions detected) for different marker sets and their combinations. The right column lists the proportion of uninformative PI values that simulations suggest can be expected from each marker set (i.e. a PI value higher than 1, when a brother is more likely than a random man to be the father). Table obtained from [316].

deceased man against the brother.

6.1.4.2 Cases with uninformative probabilities for the claimed relationship

Table 6.3 outlines the five cases where SNP analysis provided a significant improvement in the probability of the claimed relationship. The severely degraded skeletal remains tested in cases 20p07 and 70p06 involved respectively: a 35–year–old femur where 9 of 17 STRs were successfully typed and a 10–year–old doubly degraded femur [296] where all 17 STRs failed. SNP profiles detecting 51/52 loci were obtained in both cases [134]. Case 39p04 was identical in structure to 123p04 described above and in Figure 6.1 D, but here addition of SNPs provided a strong indication that the tested brother was the true father by increasing the paternity index 35–fold to 99.994% against the deceased man.

6.1.4.3 Probability of failing to exclude first–degree relatives of the true father

We calculated the probability of a brother of the true father showing no exclusions against the tested child. Here an exclusion denotes a mendelian incompatibility given the hypothesis of the tested man's brother being the true father. Probabilities are shown in Table 6.2 with the corresponding standard deviations for common STR sets, the 52 ID–SNP set and their

Case	Test	Relationship hypothesis			STRs	STRs + SNPs
			H_1	H_2	$W = H_1/H_2$ %P	$W = H_1/H_2$
28p07	Sib analysis		Half-sib	Full-sib	W: 1/3 P: 75%	1/1,193 99.91%
24p07	Sib analysis				1/897 99.89%	1/12,140,628,977 99.999999%
		RM	H_1		$W = RM/H_1$	$W = RM/H_1$
20p07	Identification of remains (paternity)	Random man is father	Tested man is father		1/139* 98.28%	1/58,823‡ 99.9983%
20p06	Identification of remains (paternity)				No profile	1/14,286‡ 99.993%
		RM	H_1	H_2	$W = RM/H_2$	$W = RM/H_2$
39p04	Brother of deceased man tested as proxy	Random man is father	Deceased man is father	Brother of deceased is father	1/11,156,811 99.99999%	1/353, 340,169 99.99999999%
					$W = H_1/H_2$	$W = H_1/H_2$
					1/496 99.799%	1/17,178 99.994%

Table 6.3: Five cases testing three different sets of alternative pedigrees. RM (random man), H_1 and H_2 relationship hypotheses were assessed with likelihood ratios (W) and percentage probabilities (P) for 21 STRs alone and STRs plus 52 SNPs. Values marked with a suffix denote partial profiles. Image obtained from [316].

combinations. Profiler Plus® and MiniFiler® are included as we now regularly use these in combination with SNPs to analyze degraded DNA when Identifiler® and Powerplex® 16 give incomplete profiles. The values reveal that both the core STR sets have a comparable failed exclusion rate of ∼2%, while SNPs alone show a rate of ∼5%: indicating that in about 1 in 50 cases using STRs a brother is completely compatible with paternity since no exclusions are detected. The slightly lower power of SNPs compared to STRs can be partly explained because with binary markers heterozygotes (in either brother or true father) are uninformative for both inclusions and exclusions in deficient families. This loss of discrimination power in SNPs is compensated by using a much higher total number of loci compared to STRs. The addition of six STRs to either core set lowers the failed exclusion rate to 1 in 360 but notably the rate is reduced more than 16–fold to 1 in 5880 when 21 STRs and 52 SNPs are combined.

Figure 6.2 plots the paternity indices obtained from the simulation of father–brother–child pedigrees. Computation of the paternity index for the alternative hypotheses: paternity of a brother against paternity of a random man provides a more realistic simulation of how an actual paternity case is normally approached when no exclusions are detected. Values for this paternity index higher than one indicate that no exclusions have been detected in the brother so he is more likely than a random man to be the father, a typical ambiguous result. Table 6.2 lists the proportion of paternity indices higher than one for each marker set. Figure 6.2 plots the complete range of PI values obtained for each marker set in 6577 simulations, ranked left to right, from most to least informative, so lower plot lines indicate a higher proportion of informative PI values obtained. Although SNPs give a “ladder–shape” plot because only opposite homozygotes between brother and child are informative, the overall proportion of highly informative PI values is seen to be equivalent to the plot for a full set of 21 STRs. Table 6.2 shows the proportion of PI values higher than one obtained for Identifiler with 6.9% and Powerplex 16 with 6.4% are both slightly higher than 52 SNPs with 6.1%. Therefore results indicate that SNPs are more efficient than STRs for resolving cases that attempt to distinguish first–degree relatives in deficient pedigrees. Overall Table 6.2 and Figure 6.2 clearly indicate that combining STRs and SNPs provides the most secure interpretative framework for relationship testing of close relatives, reducing to 0.5% the total proportion of ambiguous paternity indices.

6.1.5 Discussion

Each of the eight relationship tests reported gave some ambiguity in the STR results that was successfully resolved by including SNP analysis. The SNP profiles were generated from a straightforward multiplex assay optimized and validated for forensic identification, where a very low frequency of in-

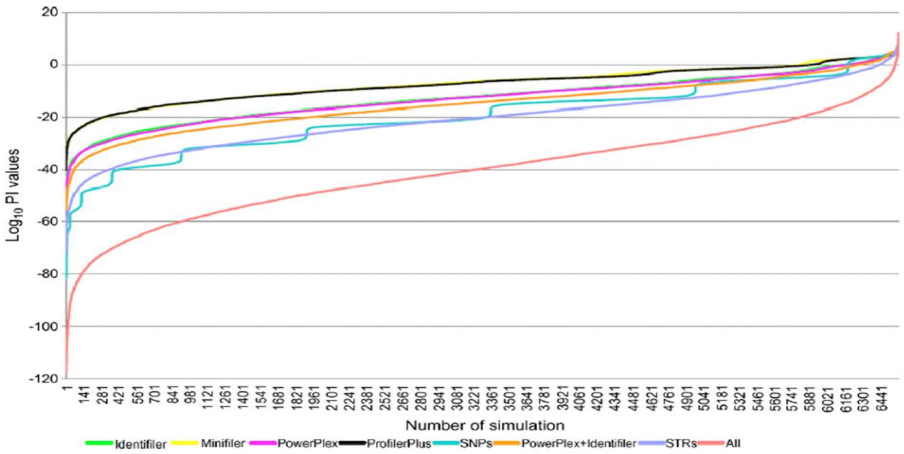

Figure 6.2: Range of logarithms base 10 of PI values from 6577 simulations for different marker sets given the two hypotheses: a brother of the true father is compatible with paternity versus a random man. The "ladder–shape" of the SNP PI values is due to the fact that only opposite homozygotes between brother and child are informative. The plot labelled *STRs* denotes all 21 unique STR loci in combination, the plot labelled *All* denotes 21 STRs plus 52 SNPs. Image obtained from [316].

compatibilities in normal trios has already been established [41, 313]. These cases clearly illustrate that the addition of 52 SNPs removes the element of doubt involved in the interpretation of challenging relationship tests using extended STR typing. We found the combination of adding a large battery of SNPs and using *Familias* to obtain reliable probabilities for each possible relationship created a more secure framework for interpreting results.

The application of SNPs in relationship testing has not been widespread to date because nearly all paternity cases are adequately resolved with existing well validated STR sets. However a characteristic of SNPs often listed in their favour for relationship testing is a comparatively low mutation rate, suggesting SNPs markedly reduce the risk of ambiguous exclusions arising from mutation. The distinction should be made here between exclusions created by allelic instability and those created by allele dropout from primer binding site mutations. SNPs have a much lower rate of allele mutation than STRs, reflected in the rates detailed above. In contrast, SNP analysis of ∼50 loci (assuming use of one extension plus two PCR primers and 20 bp average lengths) will be prone to ∼5 times more allele dropouts from binding site mutations than 15 STRs. However since the average nucleotide substitution rate is extremely low [296] this has a minor effect on the rate of incompatibilities compared to the meiotic instability of STRs. Additionally, the effect of genotyping 50 or more binary markers makes it most likely that

a primer site mutation creates a single second order exclusion contrasting with the overall pattern of results.

There is persuasive evidence in the cases described and previous studies (Figure 5 of [313]) that SNPs can add the extra discrimination power needed to resolve relationship tests that routinely compare closely related individuals. It is likely that this characteristic of SNPs is largely due to the relatively high number of segregations occurring between first–degree relatives with an extensive marker set showing the widest possible genomic distribution. Furthermore the low SNP mutation rate makes the interpretation of any exclusions found amongst closely related individuals much more secure. Applications that can therefore benefit from SNP analysis include disaster victim identification, immigration testing, complex pedigree reconstruction and the analysis of deficient families that forms a large proportion of tests identifying missing persons. The fact that SNPs additionally offer greater success when typing highly degraded DNA indicates that combining SNPs, rather than extra STRs, with the current core markers offers the best way to improve the interpretation of challenging relationship tests in the future.

6.2 Population stratification in Argentina: influence in paternity testing

The work presented here was published in [401]. Statistics and intensive simulation aim here to prove the need of appropriate databases and allelic frequency data to get accurate results in forensic identification cases. Notations and formulas have been adequately adapted to this essay.

6.2.1 Abstract

A simulation–based analysis was carried out to investigate the potential effects of population substructure in paternity testing in Argentina. The study was performed by evaluating paternity indexes (PI) calculated from different simulated pedigree scenarios and using 15 autosomal short tandem repeats (STRs) from eight Argentinean databases. The results show important statistically significant differences between PI values depending on the dataset employed. These differences are more dramatic when considering Native American versus urban populations. This study also indicates that the use of *Fst* to correct for the effect of population stratification on PI might be inappropriate because it cannot account for the particularities of single paternity cases.

6.2.2 Introduction

Historically, non–exclusion in paternity testing was statistically evaluated by means of probability of paternity according to the Essen–Moller formula [120, 121]. Later, the use of the ratio between the probability of the hypothesis of paternity (X) and non–paternity (Y), with the form X/Y, was proposed [165] and this ratio, called the paternity index (PI), was considered to be sufficiently appropriate to report a result [407]. Recently, the Paternity Testing Commission of the International Society for Forensic Genetics (ISFG; `www.isfg.org`) has issued a series of recommendations on biostatistics [157, 290] suggesting that the biological evidence should be based on likelihood ratio principles.

Calculation of PI requires knowing the allele frequency distributions in the reference population. Caution must be taken when population substructure exists, so that appropriate corrections on PI values can be applied [124]. The use of *Fst* to measure (and correct for) the effect of substructure in reference populations is commonly used in forensic genetics [124]. *Fst* measures population differentiation based on allele frequencies. However, in routine casework, one case is generally evaluated at a time and global patterns of variability in the population do not necessarily represent the idiosyncrasies of particular cases and genetic profiles, in the same way as

for haploid data [116, 117, 348]. Therefore, the use of *Fst* to account for population stratification does not always correctly adjust the PI values in every single case.

It is well documented that in Argentina differences exist between allele frequency distributions in populations, for common genetic markers used in forensics, that can have important consequences in routine forensic casework [398, 400]. This view, however, is controversial since other authors claim that population differences within the country are irrelevant in this context [267]. Recently, we used a simulation–based approach to show that these differences actually have implications in the computation of likelihood ratios in forensic casework [400]. The goal of the present study was to determine the impact of the population substructure on paternity testing, using a different simulation–based approach that compares the results obtained when using different datasets for the computation of PI values in several pedigree scenarios. Some analytical expressions can be obtained in order to address these problems [114] in a general population context. These other approaches aim generally to investigate the expected average effect of using different levels of population stratification and mutation rates in hypothesized situations (e.g. artificially created populations). The study by Karlsson *et al* [209] described a very interesting approach related to the evaluation of the risk of erroneous conclusions on DNA testing for immigration cases. The aim of the present study was, however, to exactly measure the real impact of using different datasets from Argentina on particular PI values by simulating paternity cases that could be real in this country, and given the fact that it is a particular PI value that is generally communicated to the courtroom. Therefore, the most theoretical general approaches, although necessary in science, do not help by definition to evaluate singular forensic cases where particular individuals are being judged. On the other hand, the present approach has the advantage that cases where parents come from different populations can easily be handled by sampling from different databases.

6.2.3 Materials and methods

6.2.3.1 Population samples and genotyping data

The study was based on 1906 genotypes belonging to individuals of six urban populations from Buenos Aires ($N = 879$), Neuquen ($N = 355$), Tucumán ($N = 75$), San Luis ($N = 61$), Santa Cruz ($N = 132$), and La Pampa ($N = 232$) and two Native American populations from Colla ($N = 43$) and Toba ($N = 129$) in Argentina.

The genotype data consisted of a set of 15 autosomal STRs from the Powerplex® 16 System kit (Promega, Madison, WI, USA): D3S1358, FGA, D21S11, D18S51, HUMvWA, D5S818, D13S317, D7S820, D16S539, CSF1PO, PENTA D, PENTA E, D8S1179, HUMTPOX, and HUMTH01. No devia-

tions from Hardy–Weinberg equilibrium were detected in any of these population samples.

6.2.3.2 Data simulation

Data simulation involved the following steps:

1. *Generation of artificial profiles.* For each of the 1906 real profiles in the database, a set of new profiles was created by a computer–assisted procedure. First, allele frequencies were obtained for all the original datasets. Second, compatible profiles for both parents of each individual were built as follows: each of the two alleles was randomly assigned to each parent then the other allele of each parent was randomly taken from a vector of allele population frequencies of each STR locus. Parents' sets were tested for Hardy–Weinberg equilibrium and no departures were observed.

2. *Definition of pedigrees to calculate the PI.* With the individuals generated as described in Step 1, we constructed two different types of pedigrees: alleged father–mother–child (trio) and alleged father–child (duo).

3. *Frequency databases.* A total of 50 different allelic frequency matrices were built from each population sample constructed by selecting at random 80% of the individuals of the original datasets. This bootstrap–based approach aim to control for the variability involved in the estimation of allele frequencies due, for instance, to differences in samples size.

4. *PI calculation.* PI values were calculated by contrasting two mutually exclusive hypotheses in trios and duos: (1) the alleged father is the true father of the child and (2) the father is an unrelated individual.

PIs for all the pedigrees in one population were calculated with the corresponding reference database, and also using the databases from the seven other populations. Since 50 different frequency matrices were available for each population, each pedigree yielded 50 PIs for each population database. For each population database, the mean PI value was also calculated for every single pedigree.

6.2.3.3 Statistical analyses

As explained, for each individual ($N = 1906$) a set of 50 PI values were obtained using each of the eight datasets. Three goodness–of–fit tests were employed in order to examine if each set of 50 PI values fits to a normal, namely Kolmogorov–Smirnov, Shapiro–Wilks and Pearson's χ^2 (see e.g. [285]). The

normality assumption was rejected in most of the situations, even for the most conservative test, namely Kolmogorov–Smirnov. Therefore, all the PI values were converted into logarithms and the normality was checked again using the same goodness–of–fit tests. The normality assumption (required to properly carry out the statistical tests below) could then be accepted for the logarithm of the PI values ($\log_{\mathrm{PI}}$).

Next, for each individual an ANOVA analysis was carried out for the eight sets of 50 $\log_{\mathrm{PI}}$. ANOVA allowed testing significant differences among the logPI values obtained when using the different datasets. Due to the fact that the null hypothesis of equality among sets was always rejected, we next used four different statistical tests (namely Tukey, LSD Fisher, Duncan Ranks, and Newman; see e.g. [285]), in order to explore statistical differences between all pairwise comparisons involving the 1906 profiles.

The decision to use several tests for testing normality and several post hoc tests was based on two facts: (a) the need for testing inconsistencies when using different statistical approaches that could reveal, for instance, some technical or conceptual problem in the design of the simulations and (b) select the test providing the most conservative results. Bonferroni's adjustment was used in order to account for multiple test corrections and setting the nominal significant value α to 0.01.

Additionally, for each profile we computed an ad hoc index, the weighted mean difference (WMD) between pairs of populations that quantifies the magnitude of the differences between pairs of PI values. This index is defined here as follows: for each pair of populations i, j,

$$\mathrm{WMD} = \frac{\bar{\mathrm{PI}}_i - \bar{\mathrm{PI}}_j}{\max(\bar{\mathrm{PI}}_i, \bar{\mathrm{PI}}_j)}$$

where $\bar{\mathrm{PI}}$ indicates the mean value for the set of 50 PIs obtained of each individual in each dataset.

6.2.3.4 Double checking the results

All the simulations and statistical analysis were carried out using Visual Basic programming in Microsoft Excel and the freely available statistical package R (`http://www.r-project.org/`). A random subset of the pedigrees was selected from the original pedigree simulations, and the accuracy of the results was double checked by using the shareware software *Familias* v.1.81, `www.math.chalmers.se/~mostad/familias/` [115].

6.2.3.5 Rationale

The aim of the statistical analysis was to evaluate the impact on PIs using a single national database for every forensic case in the country compared to using a regional database. In fact it is common for example that a laboratory

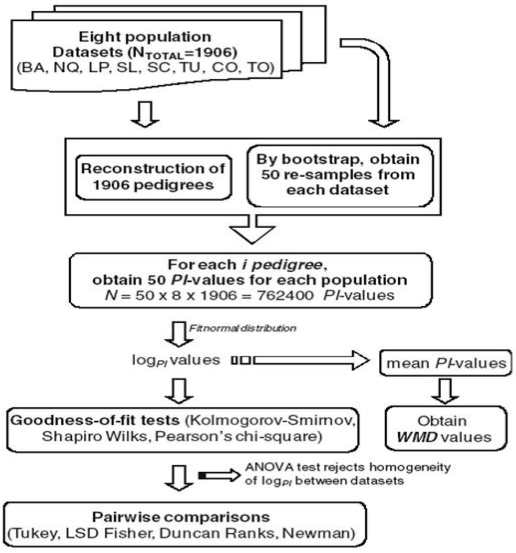

Figure 6.3: Scheme showing the main steps considered in the simulation and statistical analysis carried out in the present study. Image obtained from [401].

in Buenos Aires receives paternity cases from all over the country. If all the populations in Argentina were homogeneous no significant differences would be observed on PIs. On the contrary, if population substructure exists, we would expect to find important differences depending on the database employed. The latter would involve the need to develop local frequency tables representing the main regions from the country instead of using a global one.

One could envisage another simpler potential solution to the problem, i.e. to build a global database of the country and use it as reference population for any paternity test carried out in the territory. However, as demonstrated below, the differences in PI values when using different datasets can be dramatic, and so, the use of a single database would just aggravate the problem; e.g. if one has a case from Buenos Aires, it will be more appropriate to use the Buenos Aires database than a global one. Similar problems were addressed from a theoretical point of view by Ayres [23].

The whole simulation algorithm employed in the present study is summarized in the scheme of Figure 6.3.

	BA	NQ	LP	SL	SC	TU	CO	TO
TRIOS								
BA	–	69.1/55	52.1/32.9	76.9/67.3	74.1/62.9	76.8/67.2	89.3/84.7	92.5/89.2
NQ	11.4	–	64.7/48.4	75.3/62.7	59.5/43.2	73.7/62.6	88.4/82.9	91.8/86.8
LP	3.4	7.5	–	74.9/63.7	65.9/52.3	74.8/63.2	88.0/82.6	92.8/89.0
SL	28.8	23.1	21.9	–	71.9/60.7	68.6/53.7	82.8/74.3	94.3/91.8
SC	20.7	5.6	11.8	17.1	–	67.4/52.1	85.7/80.4	92.0/88.1
TU	27.9	20.6	21.8	11.8	11.1	–	82.6/73.8	92.3/88.0
CO	56.7	56.0	54.8	42.2	50.6	42.7	–	95.6/94.0
TO	71.5	68.2	71.4	76.0	68.2	71.0	82.2	–
DUOS								
BA		67.9/54.6	51.7/34.3	76.8/66.3	74.2/63.1	76.9/67.5	86.7/81.2	93.3/90.2
NQ	11.5		65.2/48.2	73.2/61.2	60.6/43.3	73.3/61.5	86.7/79.1	93.1/88.8
LP	3.0	7.5		74.6/62.3	66.6/53.8	74.9/63.5	86.7/80.2	93.3/90.5
SL	28.6	21.7	20.4		71.5/59.4	67.5/53.1	82.2/72.8	94.8/92.3
SC	22.8	5.5	12.9	16.5		68.0/52.9	85.0/78.3	93.6/90.2
TU	29.7	20.4	23.2	12.7	10.5		80.2/72.9	94.2/91.2
CO	56.3	52.0	52.8	41.1	47.0	40.9		96.4/95.0
TO	74.3	72.7	75.5	78.6	73.6	75.1	84.3	

Table 6.4: Significant difference between populations. The upper diagonals values indicate the percentages of individuals that show significant differences in pairwise comparisons under the test of Tukey for trios (upper) and duos (bottom); the first term is for a $\alpha = 0.01$, while the second term is for the Bonferroni's correction assuming 1906 comparisons. The below diagonals show the percentages of WMD values above 0.8. Population codes: BA Buenos Aires, NQ Neuquen, LP La Pampa, SL San Luis, SC Santa Cruz, TU Tucumán, CO Collas, TO Tobas. Image obtained from [401].

6.2.4 Results and discussion

6.2.4.1 PI values vary significantly depending on the reference population

Several statistical tests were used to measure the percentage of pedigrees from which the PI values statistically differ when using different population datasets. For instance, the Tukey test (Table 6.4) indicates that most of the times the $\log_{\text{PI}}$ values differ significantly among populations (e.g. using a nominal value of $\alpha = 0.01$ coupled with a Bonferroni's correction assuming 1906 comparisons). As expected, the largest percentages of statistically significant PI differences almost always involved comparisons between the two Native American populations versus the other datasets. The largest differences occurred between these two Native American populations.

The other statistical tests employed yielded less conservative results than Tukey (data not shown), since the Tukey test internally controls for global error type I (given the 28 comparisons carried out each time). The different statistical tests are, however, consistent in showing the percentages of PI values statistically significant as showed by a Mantel test. For instance, in trios, $r^2 > 0.997$ and $p < 0.001$ (Pearson's correlation, 10000 permutation tests) for all the comparisons (Tukey versus LSD of Fisher, Tukey versus Duncan Ranks, Tukey versus Newman).

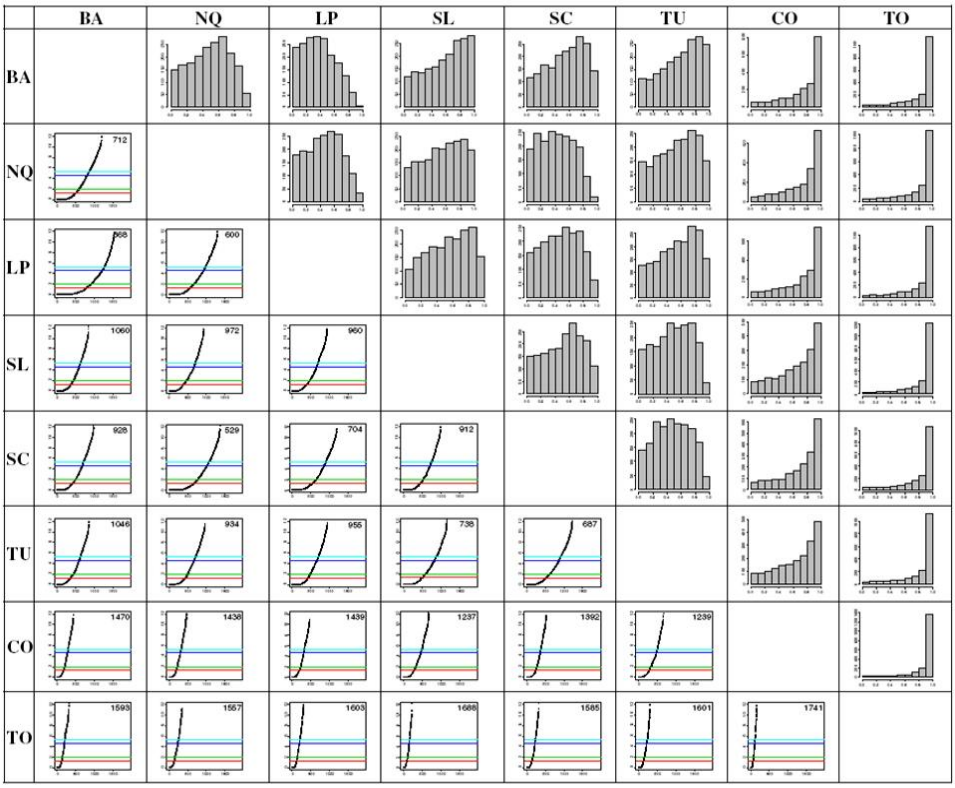

Figure 6.4: WMD values for the 1906 profiles in the database in trios. Above the diagonal are the pairwise distributions of WMD values for the 1906 profiles in the database in trios. Each histogram represents therefore the impact on WMD for a given pair of frequency datasets over the 1906. Below the diagonal are the distributions of -$\log_{10}$(p–values) for Tukey's test; the horizontal lines represent from bottom to top the -$\log_{10}$ values for $\alpha = 0.05$, $\alpha = 0.01$, and the respective values assuming Bonferroni corrections. The numbers in the top–right corner of these distributions pictures indicate the number of tests that fall out of the distribution and that in general correspond to values close to zero. Image obtained from [401].

The distribution of -$\log_{10}$(p–values) obtained using Tukey's test are shown in Figure 6.4 (below the diagonal), for trios and duos. The most outstanding feature of these figures is that the slopes of the distributions are more pronounced in those comparisons involving more distant populations (see also [397]). For instance, those involving Native Americans. It is also remarkable the large number of -$\log_{10}$(p–values) that falls below the most conservative Bonferroni's correction.

6.2.4.2 Measuring inter–population differences in PI values

The main aim of the present analytical approach is to evaluate the magnitude of the differences in PI values and to what extent statistical significances between populations have an impact in substantial PI differences that could be relevant for decisions in court.

WMD values were computed for each individual profile. These values measure the magnitude of the difference between every single pair of mean PI values among populations. For instance, a WMD value of 0.7 indicates that the difference between the two mean values considered is 70% of the absolute value of the largest mean. Therefore, high WMD values indicate large differences between populations and vice versa.

Figures 6.4 and 6.5 (above the diagonal) for trios and duos respectively, show the distributions of WMD values between pairs of datasets. Note again that the two Native American populations show the most skewed distributions towards high WMD values. In particular Toba is more distinct than Colla with respect to the other populations. In general, the histograms of Figures 6.4 and 6.5 indicate large differences between PI values independently of the population dataset used. Table 6.4 (data below the diagonals for trios and duos) indicates the percentage of WMD values above 0.8. Note that these values correspond with the two last bars of the histograms presented in Figures 6.4 and 6.5 (data above the diagonal).

6.2.4.3 Reviewing previous finding concerning population substructure in Argentina

The importance of population substructure in Argentina has been minimized in previous studies [268, 269, 270, 271]. More recently, Marino et al. [267] measured the impact of population substructure in Argentina, analyzing 15 autosomal STRs in ten population samples from the country, and concluded that no substructure could be detected supporting that a single database of the whole country could be suitable for the correct interpretation of paternity testing and forensic casework results. Nevertheless, they found a clear statistical differentiation between the Salta population sample and the rest of the population samples analyzed, which contradict their final conclusion about the possibility of using a unique database for the whole country. Moreover, our previous findings [398] revealed the existence of population substructure in Argentina at autosomal STR level. In addition, population stratification is also supported when looking at the population patterns of Y–STR [270, 399] and mitochondrial DNA data, as can be inferred from the few studies carried out in populations from this country [7, 60, 155, 349].

In the present study, we have employed exactly the same autosomal marker set used by Marino *et al* [267] but our results and conclusions differ substantially. The main reason is that the statistical approaches employed in

these studies are conceptually different. While Marino *et al* [267] employed *Fst* genetic distances to detect and quantify genetic stratification, our approach aimed to measure the effect of population substructure directly on PI values. We demonstrated here that *Fst* corrections might not account for the singularities of the full universe of genetic profiles in a population. Thus, for instance, considering trios, ∼22% of the PI values of the Toba's profiles differs more than three orders of magnitude if we use the database of Buenos Aires and some PI value can differ more than five magnitude orders. To cite one of the many outstanding examples of our results, we have observed a Toba profile with a PI value of 273 using the Toba dataset but 15788114 using Buenos Aires as the reference population in a case of alleged father–son.

It is worth stressing that in forensic routine work the results of the genetic test are directly communicated to the judge by way of a PI value, and these values are therefore those that are finally considered no matter what the values of *Fst* are in the populations. In other words, the use of *Fst*

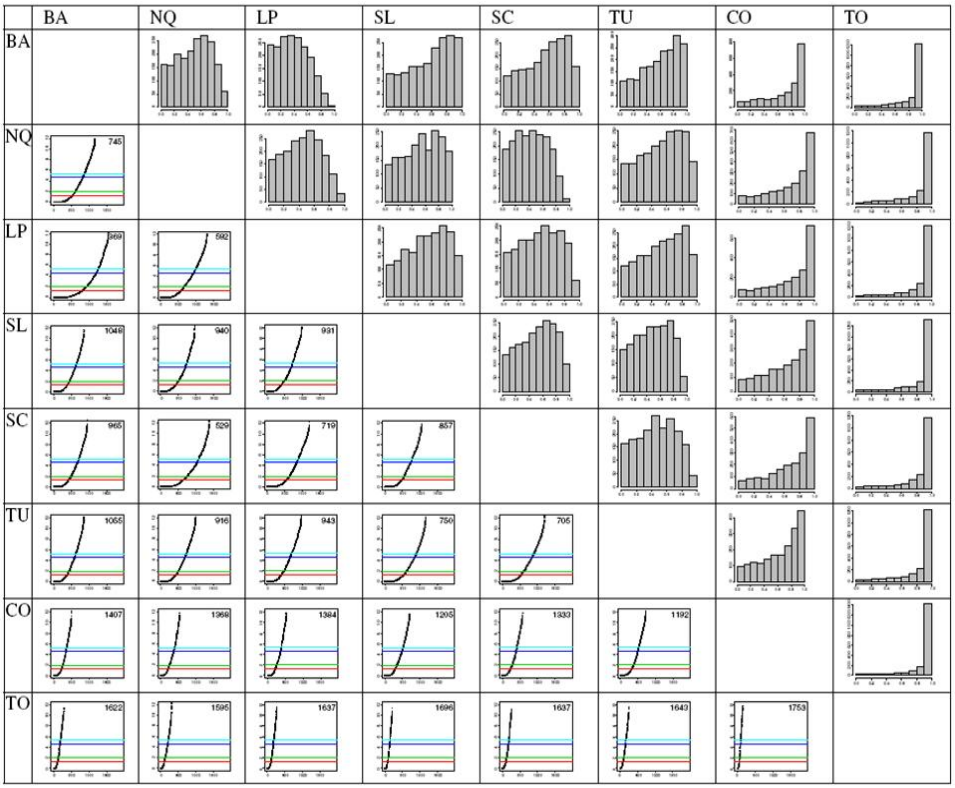

Figure 6.5: WMD values for the 1906 profiles in the database in duos. The table shows the same data as in Figure 6.4 but for duos father–son. See legend of Figure 6.4 for more details. Image obtained from [401].

to correct for population stratification might not be appropriate in court.

6.2.5 Conclusions

The results of the present study clearly support the existence of population stratification in Argentina to a level that can be relevant in forensic routine work. On the other hand, the Argentinean populations show low *Fst* values, indicating that the use of this index to measure and correct for population substructure might be inappropriate in forensics.

We have simulated pedigree scenarios where a set of 15 different STRs are fully genotyped in all the individuals. These simulations emulate the most favorable scenario. However, in real paternity cases DNA profiles can be deficient (missing data) when using highly degraded DNA (e.g. exhumed remains). Moreover, the discrimination power of the 15–plex can be limited in pedigrees where e.g. a paternity relationship has to be inferred indirectly by genotyping family members related to the alleged father. In these cases, the consequences of using inappropriate databases can be even more dramatic because PI values are generally lower.

Using a single database for routine paternity testing in Argentina might not be justified and could lead to serious bias when estimating PI values. The approach used in the present study would be also appropriate to investigate the real effect of population stratification in the paternity testing routine work exercised in other countries.

Chapter 7

Conclusions

Different conclusions and final remarks have been already pointed out for each of the different chapters in the present essay. Therefore, this section just aims to summarize some global ideas about the essentials and the state–of–the–art of present statistical approaches available in the genomic field. The conclusions are heterogeneous as it corresponds to the variety of statistical problems studied here.

The curse of dimensionality problem is present in almost all of the genomic fields of research, being (perhaps) particularly severe in the gene expression field. Current expression arrays generally deal with the analysis of few thousand genes; however, technical improvements will soon allow to include the full set of nuclear genes (around 25000–30000), which would complicate even more the statistical analysis. Furthermore, it can be tentatively said that most likely, sample sizes will remain the same. This also connects with the multiple test correction problem: as there are too many covariates (genes, gene markers, ...), many statistical tests will be needed, in some way or another, so the probability of type I errors will increase notably (see next chapter).

The huge amount of data generated in the new genomic era obviously demands fast, efficient algorithms, and powerful computers; this connects statistics and genetics with several other knowledge areas, such as bioinformatics, neural computation, etc. Further advances in computation are strongly needed to allow the implementation of new, computationally demanding statistical methods. The kernel approach developed in Chapter 4 serves as an example.

As statistics are continuously evolving and providing with new methodologies, and new problems arise each day in the genetics field, there is a growing need to improve the interplay between these two fields of research, statistics and genomics. One of the main drawbacks arises from the fact that many statistical developments are published exclusively in specialized statistical journals and therefore, these developments usually pass unnoticed

to geneticists. It is also important to mention that availability of data is different between the fields of gene expresion and SNP case–control association studies: expression data is generally available to the scientific community for further statistical analysis while SNP genomic data is not (although it seems this tendency is starting to change in the GWAS era [390]).

Another important problem in genetic association studies is the so–called publication bias, as usually journals with high impact factor tend to select for publication only studies reporting positive associations. This has contributed to the appearance of many spurious associations, that are not replicable in other studies using independent samples. Sometimes, a well–designed study that do not find association with a particular disease or trait can bring more light to science than many biased studies showing (false) positive associations that subsequently do not replicate.

Although the statistical basis of the tools used in forensic genetics is simpler than in genomics, and the conceptual framework was developed long time ago, there are many problems that still remain unsolved. Many of these problems connect to the field of population and molecular genetics, namely, population stratification, mutation rates, ... Although they rarely affect most of the routine forensic casework, these problems also demands the attention of statisticians. This essay contains examples of how to combine intensive simulation, statistical tools and forensic knowledge to solve intricate, unresolved problems in forensics.

Chapter 8

Further research

The number of further research lines arising from this work is countless, due to the great variety of areas we work with and the heterogeneity of the studies. Here we will mention some of them, being aware that a lot will remain unsaid. Groups of forthcoming research lines will be suggested for each of the chapters inside this essay: Chapter 3 (penalized regression in studies involving high–dimensional data), Chapter 4 (support vector machines in classification problems), Chapter 5 (statistics in clinical genetics) and Chapter 6 (statistics in forensic and population genetics).

8.1 Penalized regression in studies involving high–dimensional data

- The imperious need of sparse models in those fields where data is high–dimensional (gene expression, text categorization, information retrieval, combinatorial chemistry, ...) makes l_q penalization a very attractive option to be massively carried out. The elastic net also seems to be promising, especially when strong correlations are present.

- The power of penalization in combination with the power of resampling techniques has been already proposed [172], and its consistency has been proved. A long road is in front of us. A great variety of resampling methods and penalization approaches exist, so many researchers will have part along this path.

- Different options for the vector of specific penalizations in Chapter 3 were investigated in this essay, and many other could be possible.

- It is very important to go into the biological interpretation of the gene expression results obtained in depth. This is partly done, on a small–scale, inside this essay, with leukemia dataset results.

8.2 Support vector machines (SVMs) in classification problems

- There is a need to get faster and more efficient kernels, reducing the computational burden, while at the same time allowing for the existence of complex gene interactions associated with disease.

- Recent studies have tried to combine the abilities of SVMs to classify with the power of penalization approaches. More research work is needed.

8.3 Statistics in clinical genetics

- New statistical methods need to be developed, and existing ones need to be adapted, to face the new challenges arising as a consequence of the GWAS age: multifactorial diseases, low penetrances, ...

- A main issue in GWAS studies is the high number of hypotheses carried out, giving rise to the so–called multiple test correction problem. Most of the GWAS inside the scientific literature use Bonferroni–type corrections (which are too conservative) or *ad hoc* procedures which have not been previously tested. Careful studies are needed to determine which are the best corrections to be carried out with high–dimensional genetic datasets. There are some consortia or authors of GWAS studies that have undertaken to make data available (upon request). This provides statisticians with the necessary tools, together with high–dimensional data simulation packages, to test the abilities of the different corrections.

- Careful attention to positive associations detected is required by all the researchers involved in the field. Publication bias is a worrying source of problems. As commented along this essay, many positive associations have not been subsequently replicated in independent studies. Critical studies, bibliographic revisions and meta–analyses are of great utility to detect those studies which results are not reliable or, at least, dubious.

8.4 Statistics in forensic and population genetics

- Although the statistical basis to be used in criminal and paternity cases were established long time ago, the great variety of problems and the importance of each forensic case itself (usually involving inheritances, feelings, sentences and imprisonments, ...), make that the

tools needed to solve this kind of problems have to be continuously reinvented. The study in Section 6.1 serves as an example.

- Both studies in Sections 6.1 and 6.2 show the importance that simulations acquire nowadays. Real data is sometimes unattainable or economical costs required to obtain samples cannot be afforded.

- Further work on correction for population stratification problems is needed. Use of *Fst* or θ corrections has been largely considered as the definitive solution for stratification. Preliminary studies show that *Fst* corrections could not be appropriate in populations where indigenous and immigrant subpopulations are present without admixture (e.g. Argentina).

- Many forensic genetic studies focus nowadays on discovering the genetic regions associated with physical traits like eye colour, hair colour or pigmentation. Phenotypes are highly variable with regards to each of these traits, which means a new challenge for statisticians. There is a substantial need to find the best procedures to deal with this kind of data.

- Ancestrality studies look for the effect of human migrations in the distribution of genetic frequencies. In this sense, the search for those genetic markers (STRs or SNPs) which allelic frequencies have the largest differences among world populations is very interesting, as it can provide with the tools to distinguish ethnically different individuals in a genetic way. Statisticians may look for the best measures allowing to unravel these genetic markers, as they can mean a powerful tool in forensic genetic casework.

Appendix A

Proof of the equivalence GSoft – CCD algorithm

The log–likelihood functions in logistic regression and in lasso logistic regression with specific penalizations are given in (3.1) and (3.2), respectively. The first partial derivatives or score functions are:

$$s_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{-y_i x_{ij}}{1 + \exp(y_i \mathbf{x}_i' \boldsymbol{\beta})}$$

The definition of the Δv_j for the lasso case in [151], applied on a penalized regression problem with specific penalizations for each variable, is given in (3.3). For ease of notation, we will use here S instead of $\text{sign}(\beta_j)$. We will base the entire proof in the steps and the notations used in Figures 4 and 5 in [151]. Many of the terms used there will be repeated here. To clarify the notation, we will use β_j for the true value of the coefficients and $\beta_j^{(I)}$ for the value of the jth coefficient in the iteration I of the CCD algorithm.

We will begin proving the equivalence for the case $\beta_j = 0$, and then we will move to the more general case of $\beta_j > 0$ (analogous proof for $\beta_j < 0$).

Case $\beta_j = 0$

(1) $\Rightarrow$ (2)

We assume that the CCD algorithm, as explained in [151], converges. Therefore, from a certain iteration I we have $\beta_j^{(I)} = 0$ and $\Delta v_j^{(I)} = 0$. The CCD algorithm tries then to improve the objective function value searching in the positive and the negative direction, so:

$$\left\{ \begin{array}{l} S = 1 \text{ and } \Delta v_j^{(I+1)} \leq 0 \;\Leftrightarrow\; s_j(\boldsymbol{\beta}) - \lambda\gamma_j \leq 0 \\ S = -1 \text{ and } \Delta v_j^{(I+1)} \geq 0 \;\Leftrightarrow\; s_j(\boldsymbol{\beta}) + \lambda\gamma_j \geq 0 \end{array} \right\} \Leftrightarrow$$

$$\left\{ \begin{array}{l} s_j(\boldsymbol{\beta}) \leq \lambda\gamma_j \\ -s_j(\boldsymbol{\beta}) \leq \lambda\gamma_j \end{array} \right\} \Leftrightarrow \tag{A.1}$$

$$\Leftrightarrow \ |s_j(\boldsymbol{\beta})| \leq \lambda\gamma_j$$

(2) ⇒ (1)

We assume now that the necessary and sufficient conditions for convergence in the GSoft theorem are fulfilled. That implies, for β_j

$$|s_j(\beta)| \leq \lambda\gamma_j$$

We need to bear in mind also that the initial value for β_j in the CCD algorithm is $\beta_j^{(0)} = 0$. In this situation and from the definitions of the CCD algorithm for the lasso case, we have that

- if we try $S = 1$ (positive direction) then $\Delta v_j^{(0)} \leq 0$ and positive direction failed.
- if we try $S = -1$ (negative direction) then $\Delta v_j^{(0)} \geq 0$ and negative direction failed.

Therefore, following the steps of the CCD algorithm for the lasso case, this means we take $\Delta v_j^{(0)} = 0$, as both directions failed, and then

$$\Delta\beta_j = \min\big(\max(0, -\Delta_j), \Delta_j\big) = \min\big(0, \Delta_j\big) = 0$$

and the CCD algorithm converges.

Case $\beta_j > 0$ (the proof is analogous for $\beta_j < 0$)

(1) ⇒ (2)

Let us suppose that $s_j(\boldsymbol{\beta}) \neq \lambda\gamma_j$ and we will try to show that this gives rise to a contradiction. As the true β_j is positive and the CCD algorithm converges, from any iteration I we will have $\beta_j^J > 0$ for all iteration $J > I$, so $S = 1$ and $\Delta v_j^J \neq 0$ following the definition in (3.3). This way, for any positive constant k,

$$\Delta\beta_j^{(J)} = \min\big(\max(\Delta v_j^{(J)}, -\Delta_j^{(J)}), \Delta_j^{(J)}\big) \neq 0 \quad \Rightarrow$$

$$\Rightarrow \qquad \Delta_j^{(J+1)} = \max\Big(2|\Delta\beta_j^{(J)}|, \tfrac{\Delta_j^{(J)}}{2}\Big) > k > 0$$

and this happens for every iteration $J > I$, which enters in contradiction with the convergence of the CCD algorithm to β_j.

(2) ⇒ (1)

We assume now that necessary and sufficient conditions for convergence in the GSoft theorem are fulfilled; let us suppose that the CCD algorithm converges to a different "solution" $\bar{\beta} \neq \boldsymbol{\beta}$ with $\bar{\beta}_j \neq \beta_j$.

In such case, as the conditions in (a) in the GSoft theorem determine an unique solution, it has to be $s_j(\bar{\beta}) \neq \lambda\gamma_j$; then $\Delta v_j^{(J)} \neq 0$, for all $J > I$ with $I \in \mathbb{N}$ and therefore $\Delta\bar{\beta}_j$ does not converge to 0, which means the CCD algorithm does not converge either, and we have reached a contradiction.

We have not mentioned or used anywhere in the proof the condition about the positive definite nature of the matrix $X_\lambda' H(\hat{\eta}) X_\lambda$. So we have to prove this condition is also fulfilled when the CCD algorithm converges. We will prove this by *reductio ad absurdum*.

Let us assume that $X_\lambda' H(\hat{\eta}) X_\lambda$ is not definite positive. As X_λ is a complete matrix, this implies that $H(\hat{\eta})$ is not definite positive, and therefore

$$\left.\begin{array}{l} -H(\hat{\eta}) \text{ (Hessian) is not definite negative} \\ \frac{\partial L_1(\hat{\beta})}{\partial \beta_j} = 0 \text{ for all } j \in \{1, \ldots, p\} \end{array}\right\}$$

and therefore the estimated linear predictor $\hat{\eta}$ cannot be a maximum of the objective function in [214], which means $\hat{\beta}$ is not a minimum of the objective function in [151] and the CCD algorithm does not converge (contradiction).

Appendix B

Mathematical properties and definite positiveness of the SVM kernel

We need to prove that the kernel (4.4) used in our SVM studies can be thought of as a dot product in the so–called feature space F [367], where a dot product is a symmetric bilinear form that is strictly positive definite in the vector space F.

To do this, we will proceed in three consecutive steps: first, we will define the mapping ϕ and the feature space F. After that, we will observe that F is a vector space over $\mathbb{R}$ with addition and scalar multiplication, to finally show that the kernel (4.4) can be seen as a symmetric bilinear form strictly positive definite in F.

B.1 Mapping ϕ and feature space F

We take the feature space:

$$F = \left\{ \mathbf{\Gamma} = (\mathbf{\Gamma}^{(1)}, \ldots, \mathbf{\Gamma}^{(U)}) \colon \mathbf{\Gamma}^{(u)} \in \mathbb{R}^{8p^2},\, u = 1, \ldots, U,\, U = 2^v \right\}$$

so F can be understood as a space which elements are vectors of vectors. v is a fixed value that can be taken as the maximum of heterozygotes $\max \# \left\{ x^j = 2 \right\}$ that can be found in p SNP markers $(x^1, \ldots, x^p)$ (p is an upper bound for v).

The mapping ϕ is given by:

$$\begin{aligned} \phi : \quad & \{1,2,3\}^p \to F \\ & \mathbf{x} \longmapsto \phi(\mathbf{x}) = \mathbf{\Gamma} \end{aligned}$$

For the sake of convenience, we will call $\phi(\mathbf{x}) = \mathbf{\Gamma} = (\mathbf{\Gamma}^{(1)}, \ldots, \mathbf{\Gamma}^{(U)})$ where $\Gamma^{(u)} = (1/T)\mathbf{A}^{(u)}$, $u = 1, \ldots, U$, $T = 2^{\#\{x^j=2\}}$ and

$$\begin{aligned}\mathbf{A}^{(u)} = \Big(&\sqrt{w_{1,1}}z^1_{(u)}, \sqrt{w_{1,1}}(1-z^1_{(u)}), \sqrt{w_{1,2}}z^1_{(u)}z^2_{(u)}, \sqrt{w_{1,2}}z^1_{(u)}(1-z^2_{(u)}),\\ &\sqrt{w_{1,2}}(1-z^1_{(u)})z^2_{(u)}, \sqrt{w_{1,2}}(1-z^1_{(u)})(1-z^2_{(u)}), \ldots, \sqrt{w_{1,2p}}z^1_{(u)}z^{2p}_{(u)},\\ &\sqrt{w_{1,2p}}z^1_{(u)}(1-z^{2p}_{(u)}), \sqrt{w_{1,2p}}(1-z^1_{(u)})z^{2p}_{(u)}, \sqrt{w_{1,2p}}(1-z^1_{(u)})(1-z^{2p}_{(u)}),\\ &\ldots, \sqrt{w_{2,2}}z^2_{(u)}, \sqrt{w_{2,2}}(1-z^2_{(u)}), \sqrt{w_{2,2}}z^2_{(u)}z^3_{(u)}, \sqrt{w_{2,2}}z^2_{(u)}(1-z^3_{(u)})\\ &\sqrt{w_{2,2}}(1-z^2_{(u)})z^3_{(u)}, \sqrt{w_{2,2}}(1-z^2_{(u)})(1-z^3_{(u)}), \ldots, \sqrt{w_{2,2p}}z^2_{(u)}z^{2p}_{(u)},\\ &\sqrt{w_{2,2p}}z^2_{(u)}(1-z^{2p}_{(u)}), \sqrt{w_{2,2p}}(1-z^2_{(u)})z^{2p}_{(u)}, \sqrt{w_{2,2p}}(1-z^2_{(u)})(1-z^{2p}_{(u)}),\\ &\ldots, \sqrt{w_{2p,2p}}z^{2p}_{(u)}, \sqrt{w_{2p,2p}}(1-z^{2p}_{(u)})\Big) \in \mathbb{R}^{8p^2}\end{aligned}$$

where the $w_{i,j}$ are the weights assigned to the pair i, j of alleles and the binary values $z^j_{(u)}$ are given by:

$$\begin{aligned}&\text{if } x^j = 1 \text{ then} \quad z^{2j-1}_{(u)} = z^{2j}_{(u)} = 0\\ &\text{if } x^j = 3 \text{ then} \quad z^{2j-1}_{(u)} = z^{2j}_{(u)} = 1\\ &\text{if } x^j = 2 \text{ then} \quad z^{2j-1}_{(u)} = 1, z^{2j}_{(u)} = 0 \quad \text{or } z^{2j-1}_{(u)} = 0, z^{2j}_{(u)} = 1\end{aligned}$$

The number U of vectors in each element $\mathbf{\Gamma}$ in F is fixed as the maximum of heterozygotes in a genotype. It could happen (in fact, it will be a common situation) that this maximum will not be reached for some genotype $\mathbf{x}$. In that case, the components $\mathbf{\Gamma}^{(u)}$ with index from $2^{\#\{x^j=2\}} + 1$ to $U = 2^v$ are taken as $\mathbf{0} \in \mathbb{R}^{8p^2}$.

F is a vector space as a consequence of its definition and the properties of Euclidean vector spaces. Addition $(+)$ and scalar multiplication $(\cdot)$ are defined by:

$(+)$ Addition. Let $\mathbf{\Gamma} = (\mathbf{\Gamma}^{(1)}, \ldots, \mathbf{\Gamma}^{(U)})$ and $\mathbf{\Omega} = (\mathbf{\Omega}^{(1)}, \ldots, \mathbf{\Omega}^{(U)})$ be, $\mathbf{\Gamma}, \mathbf{\Omega} \in F$. Then

$$\mathbf{\Gamma} + \mathbf{\Omega} = (\mathbf{\Gamma}^{(1)} + \mathbf{\Omega}^{(1)}, \ldots, \mathbf{\Gamma}^{(U)} + \mathbf{\Omega}^{(U)})$$

$(\cdot)$ Scalar multiplication. Let $\lambda \in \mathbb{R}$ and $\mathbf{\Gamma} = (\mathbf{\Gamma}^{(1)}, \ldots, \mathbf{\Gamma}^{(U)}) \in F$ be. Then

$$\lambda\mathbf{\Gamma} = (\lambda\mathbf{\Gamma}^{(1)}, \ldots, \lambda\mathbf{\Gamma}^{(U)})$$

B.2 The SVM kernel can be expressed as a dot product in the feature space F

The last step of the proof entails expression of the kernel (4.4) as a dot product in the feature space F. Let us define the following application:

$$\begin{aligned} Q: \quad F \times F &\to \quad \mathbb{R} \\ (\mathbf{\Gamma}, \mathbf{\Omega}) &\longmapsto \quad Q(\mathbf{\Gamma}, \mathbf{\Omega}) = \sum_{u=1}^{U} \sum_{o=1}^{U} \left\langle \mathbf{\Gamma}^{(u)}, \mathbf{\Omega}^{(o)} \right\rangle \end{aligned}$$

where $\langle , \rangle$ is the common scalar product in $\mathbb{R}^{8p^2}$. Therefore, we have that the kernel in (4.4) can be expressed in terms of Q:

$$K(\mathbf{x}_i, \mathbf{x}_k) = Q(\phi(\mathbf{x}_i), \phi(\mathbf{x}_k))$$

So it only remains to be proved that Q is a dot product, that is, a symmetric bilinear form strictly positive definite, for (4.4) to be a valid kernel. The requirements demanded are:

- **Bilinearity**. We will prove bilinearity in the first component. The proof is analogous for the second one. Let $\mathbf{\Gamma}, \mathbf{\Omega}$ and $\mathbf{\Psi}$ be elements in the feature space F, and $\lambda_1, \lambda_2 \in \mathbb{R}$. Then:

$$\begin{aligned} Q((\lambda_1 \mathbf{\Gamma} + \lambda_2 \mathbf{\Omega}), \mathbf{\Psi}) &= \sum_{u=1}^{U} \sum_{o=1}^{U} \left\langle (\lambda_1 \mathbf{\Gamma} + \lambda_2 \mathbf{\Omega})^{(u)}, \mathbf{\Psi}^{(o)} \right\rangle \\ &= \sum_{o=1}^{U} \Big(\sum_{u=1}^{U} \left\langle (\lambda_1 \mathbf{\Gamma} + \lambda_2 \mathbf{\Omega})^{(u)}, \mathbf{\Psi}^{(o)} \right\rangle \Big) \\ &= \sum_{o=1}^{U} \Big(\sum_{u=1}^{U} \left[\lambda_1 \left\langle \mathbf{\Gamma}^{(u)}, \mathbf{\Psi}^{(o)} \right\rangle + \lambda_2 \left\langle \mathbf{\Omega}^{(u)}, \mathbf{\Psi}^{(o)} \right\rangle \right] \Big) \\ &= \lambda_1 \sum_{u=1}^{U} \sum_{o=1}^{U} \left\langle \mathbf{\Gamma}^{(u)}, \mathbf{\Psi}^{(o)} \right\rangle + \lambda_2 \sum_{u=1}^{U} \sum_{o=1}^{U} \left\langle \mathbf{\Omega}^{(u)}, \mathbf{\Psi}^{(o)} \right\rangle \\ &= \lambda_1 Q(\mathbf{\Gamma}, \mathbf{\Psi}) + \lambda_2 Q(\mathbf{\Omega}, \mathbf{\Psi}) \end{aligned}$$

- **Simmetry**. Let $\mathbf{\Gamma}, \mathbf{\Omega} \in F$ be, then:

$$Q(\mathbf{\Gamma}, \mathbf{\Omega}) = \sum_{u=1}^{U} \sum_{o=1}^{U} \left\langle \mathbf{\Gamma}^{(u)}, \mathbf{\Omega}^{(o)} \right\rangle = \sum_{o=1}^{U} \sum_{u=1}^{U} \left\langle \mathbf{\Omega}^{(o)}, \mathbf{\Gamma}^{(u)} \right\rangle = Q(\mathbf{\Omega}, \mathbf{\Gamma})$$

- **Strictly positive definiteness**. Positive definite nature is relatively easy to prove. Let $\mathbf{\Gamma} = (\mathbf{\Gamma}^{(1)}, \ldots, \mathbf{\Gamma}^{(U)})$ be any element of F. Then,

making some calculations and using the properties of the scalar product $\langle , \rangle$ in $\mathbb{R}^{8p^2}$ it can be seen that:

$$\begin{aligned} Q(\mathbf{\Gamma}, \mathbf{\Gamma}) &= \sum_{u=1}^{U} \sum_{o=1}^{U} \left\langle \mathbf{\Gamma}^{(u)}, \mathbf{\Gamma}^{(o)} \right\rangle \\ &= \left\langle \mathbf{\Gamma}^{(1)} + \ldots + \mathbf{\Gamma}^{(U)}, \mathbf{\Gamma}^{(1)} + \ldots + \mathbf{\Gamma}^{(U)} \right\rangle \quad \geq 0 \end{aligned}$$

From the above inequality it is immediate that the "strict" nature is not achieved, as we will have $Q(\mathbf{\Gamma}, \mathbf{\Gamma}) = 0$ for those $\mathbf{\Gamma} \neq \mathbf{0}$ fulfilling $\mathbf{\Gamma}^{(1)} + \ldots + \mathbf{\Gamma}^{(U)} = \mathbf{0}$ in $\mathbb{R}^{8p^2}$. Anyway, this will not affect the results obtained with the kernel (4.4), as the sets of (genetic) data involved in our studies give rise to elements $\mathbf{\Gamma}$ in F with all the values non–negative in each component $\mathbf{\Gamma}^{(u)}$.

Resumo en galego - Summary in Galician language

Dende o seu nacemento fai varios séculos, indubitablemente debido a necesidade de comprender a lóxica pola cal se rexían moitos xogos de azar, o uso da estatística extendeuse a outras áreas con obxectivos non tan marcadamente economicistas, e si mais científicos. Os campos das finanzas, da medicina, da física, das ciencias da computación, etc. (e tamén dos xogos de azar) fixeron uso neste tempo de diferentes ferramentas estatísticas para obteren coñecemento dos datos.

A xenética é un destes campos. Considérase que o seu estudo comeza a partir dos traballos do monxe checo Gregor Mendel, que realizou no século XIX unha serie de experimentos coa fin de coñecer o patrón de herencia de diferentes especies vexetáis. Nos últimos tempos, os estudos xenéticos sufriron unha explosión, debida ós avances na tecnoloxía e ó incremento das investigacións dirixidas a determinar a influencia do ADN na clínica, no eido das investigacións forenses, na determinación das migracións das poboacións humanas, etc. Esta explosión trouxo consigo importantísimos incrementos nas cantidades de cartos adicadas a investigación xenética (a lo menos en aquelas nacións que otorgan á investigación a importancia que realmente ten), o cal se traduxo, entre outras cousas, na aparición de cantidades masivas de datos que necesitan ser convintemente analizados. É neste punto no que a estatística surxe co obxectivo de dotar de significado a estes datos, proveendo a xenética das metodoloxías e das ferramentas necesarias de cara a obtención de resultados que sexan de proveito tanto para o descubrimento de variantes xenéticas relacionadas con enfermidades, como para relacionar cada xen coa súa función correspondente, dar con aqueles marcadores de relevancia no que se refire ós análisis forenses ou clasificar novos subtipos de enfermidades segundo os patróns de expresión xénica observados.

Esta memoria pretende por un lado introducir ó lector no contexto da estatística xenética e toda a súa complexidade e, por outro, provelo cunha serie de ferramentas que consideramos de utilidade neste ámbito. A contin-

uación pasaremos a resumir capítulo por capítulo o seu contido:

Na introducción (**Capítulo 1**) faise unha ampla descripción de todos aqueles elementos que serán tratados nos traballos contidos na memoria. A **Sección 1.1** repasa, nun certo orden lóxico, todos aqueles elementos involucrados nos mecanismos da herencia. Comezando a partir do ADN (nuclear e mitocondrial), e a súa estrutura en forma de cromosomas, pásase a continuación a explicar o significado das unidades físicas e funcionais da herencia, os xens, para logo enumerar os diferentes tipos de marcadores xenéticos, de máximo interés tanto en xenética clínica, como en xenética forense, de poboacións, evolutiva ou filoxenética. Dado que un par de individuos tomados de forma aleatoria apenas difiren nun 0.1% da súa secuencia nucleotídica, o estudo daqueles marcadores nos que se apreza variabilidade dentro da especie humana (polimorfismos) resulta ser vital de cara a coñecer as bases xenéticas que determinan as diferencias interhumanas.

Diferentes tipos de márcadores xenéticos foron usados no eido da xenética clínica ó longo do tempo. Os mais en boga hoxe en día son os SNPs (do inglés, *Single Nucleotide Polymorphism*), consistentes nun cambio de base na secuencia nucleotídica. Os CNVs (do inglés, *Copy Number Variable*) supoñen tamén unha fonte de información que crece de día en día. Xa dentro do marco da xenética forense, os STRs (do inglés, *Short Tandem Repeats*) úsanse nas probas de identificación forense, moitas veces xunto cos SNPs, o cal supón un avance relativamente recente.

Dentro da **Sección 1.1** tamén se fai un repaso do Human Genome Project (HGP), o proxecto en gran parte responsable da explosión mediática da xenética. O obxectivo principal do HGP foi dende o principio a obtención do xenoma completo dun ser humano, con todas as implicacións e riscos que conleva. Este proxecto foi desenrolado de forma independente por un consorcio público, sendo o governo dos EEUU o maior inversor, e unha iniciativa privada comandada por Celera Genomics. Obviamente, o conxunto de obxectivos a alcanzar por parte do HGP vai mais aló da simple secuenciación do xenoma, xa que isto supón a consecución dunha grande cantidade de obxectivos a unha menor escala.

Na **Sección 1.2** trátase a obtención de datos cuantitativos de expresión xénica, cos cales se traballará ó longo da primeira parte deste traballo. En concreto, explícase como son obtidos os datos a partiren de diferentes paquetes de software e técnicas de procesado de imaxes procedentes de estudos de microarray. A expresión xénica mide específicamente o nivel de expresión (sobreexpresión, infraexpresión, ...) dun xen ou rexión xénica; nestes estudos dito nivel recibe un valor numérico que se move dentro dun rango continuo. Unha vez explicados os pasos da obtención das medidas de expresión, faise un pequeno repaso dalgunhas das líneas de investigación mais en boga neste eido, e nas que a estatística ten un papel mais relevante. Como veremos mais adiante, a principal característica dos datos de expresión xénica

é a súa alta dimensionalidade, xa que estes estudos involucran un número inusualmente alto de variables (e que en ocasións pode comprender a totalidade de xenes dentro do xenoma, arredor de 25000). Este problema coñécese en estatística como a maldición da dimensionalidade (do inglés, *curse of dimensionality*).

Polo tanto, compre comentar que esta memoria vai presentar unha batería de métodos, técnicas e ferramentas estatísticas con aplicación sobre diferentes tipos de variables que poden aparecer no eido da xenética. É por iso que a **Sección 1.3** aparece dividida en dúas diferentes subseccións:

A **Subsección 1.3.1** fai un amplo resumo do estado da arte no que se refire os estudos de asociación en xenética clínica con SNPs, centrándose en aqueles que buscan diferencias entre unha mostra de enfermos ou casos, e outra mostra formada íntegramente por individuos control. Xeralmente, diferentes metodoloxías estatísticas son usadas na busca de diferencias significativas nas frecuencias alélicas presentes nos casos e nos controis. En determinadas circunstancias, estas diferencias significativas son indicativas de asociación (que non necesariamente causalidade) entre unha variante xenética dada e a enfermidade baixo estudio. Diferentes ferramentas son comúnmente usadas en dita busca: árbores de clasificación (CART), random forests (RF), regresión loxística (LR), etc. Os factores que poden complicar os estudos de asociación xenética son moitos e diversos: fenocopia, heteroxeneidade, ...e deben ser tidos en conta á hora de diseñar o estudo. Nunha primeira fase, os estudios de asociación centráronse nunhas determinadas rexións do ADN sospeitosas ou candidatas de conter as variantes responsables das enfermidades. Hoxe en día, os GWAS (do inglés, *Genome Wide Association Study*) conteñen información de centos de miles de SNPs (o cal engloba a maior parte da variabilidade xenética humana) para miles de casos e miles de controis agrupados e obtidos botando man de consorcios internacionais. Por suposto, as inversións económicas necesarias para levar a cabo os devanditos estudios son moi grandes, e imposibles de abordar por parte da maioría dos grupos de investigación. Neste senso, a simulación de xenotipos de individuos adquire unha grande importancia, xa que permite a competitividade incluso en situacións de inferioridade económica. A lista de enfermidades estudadas é longa. Descartadas as enfermidades mendelianas (hemofilia, acondroplasia, ...), cunha base xenética extremadamente simple e descuberta tempo ha, a investigación céntrase agora en intentar descubrir a base xenética das enfermidades comúns (cáncer, enfermidades psiquiátricas, diabetes, asma, ...), obviamente cunha natureza complexa. Desta complexidade surxe a necesidade de técnicas estatísticas capaces de detectar patróns en conxuntos de datos de alta dimensión.

Pola súa parte, a **Subsección 1.3.2** informa acerca da situación actual no que se refire os estudos con datos de expresión. Dado que os obxectivos (busca de asociacións xen–enfermidade, función xénica, rutas metabólicas, ...) difiren en gran medida entre os diferentes tipos de estudos, a lista de

técnicas tamén é ampla e diversa no que se refire a finalidade. As técnicas de aprendizaxe supervisado (e.g. regresión loxística) están pensadas para a clasificación/predicción do estatus caso ou control en novos individuos, polo que son entrenadas sobre unha mostra de entrenamento (*training sample*) e posteriormente probadas nunha mostra de proba (*test sample*). Por outro lado, as técnicas de aprendizaxe non supervisado (e.g. análisis cluster) non teñen en conta en ningún momento o estatus (caso ou control), agrupando ós individuos segundo as súas similaridades xenéticas. A súa finalidade non é en ningún caso a clasificación, e si detectar novos subtipos da enfermidade, descubrir a función dun xen a partir da súa pertenza a un determinado cluster de expresión, etc.

Para finalizar coa introducción, a **Sección 1.4** resume as características mais importantes a ter en conta no que se refire ó uso de ferramentas estatísticas en xenética forense e de poboacións. Os STRs e os SNPs son os marcadores xenéticos xeralmente utilizados neste tipo de estudios. A partir de tales marcadores, perfís xenéticos son construidos e comparados para os diferentes individuos, usando kits comerciais de STRs ou grupos ("plexes") de SNPs altamente polimórficos. O teorema de Bayes e o teorema das probabilidades totais son as ferramentas estatísticas xeralmente utilizadas nos casos forenses mais comúns e simples. Nembargantes, problemas tales coma a estratificación poboacional, a endogamia, ... xeran casos de difícil resolución que requiren de técnicas mais complexas, simulacións intensivas, etc. Todas as implicacións que acarrean os casos de rutina forense (herencias, sentencias xudiciais, encarceramentos, etc.) provocan que a resolución de cada caso sexa vital, a lo menos no que respecta ás partes implicadas, e que a probabilidade de erro se teña que ver reducida ó máximo.

É, polo tanto, esta unha memoria enfocada e orientada a estatística xenética, na que se exporán diversos traballos con diversos tipos de datos (expresión xénica, SNPs, STRs) procedentes de fontes tanto reais coma simuladas. A xustificación e os obxectivos que persigue o traballo resúmense no **Capítulo 2**. Cada un dos seguintes capítulos contén un traballo ou grupo de traballos que se enmarcan dentro dunha mesma liña. O fío conductor é, evidentemente, o uso de ferramentas estatísticas na xenética. Pasamos a continuación a relatar con algo mais de detalle cada un deles:

Capítulo 3. Os microarrays de expresión xénica xeralmente estudan miles de xens para tan só unhas poucas docenas de mostras. O obxectivo é explicar a variable resposta (de carácter categórico) a partir do patrón de expresión, utilizando un modelo que inclúa únicamente unhas poucas variables, polo que os métodos estatísticos dando lugar a modelos "*sparse*" (modelos de regresión nos que únicamente un número reducido de variables ten coeficiente non nulo) son grandemente valorados. Os métodos de regresión penalizada, tales como o lasso, a bridge regression ou a elastic net dan lugar

a modelos "*sparse*" minimizando unha función obxectivo da forma:

$$L_P(\beta, \lambda) = L(\beta) + P(\beta, \lambda)$$

onde β refírese ó vector de coeficientes e λ é o termo de penalización.

Neste estudo nos propoñemos diferentes métodos de penalización lasso usando modelos de regresión loxística. As penalizacións son específicas para cada xen, e poden basarse na súa variabilidade ou en aplicacións previas de métodos de penalización. Todas estas metodoloxías teñen a súa partida nun recente estimador *GSoft* (do inglés, *generalizad soft-threshold*). Un novo algoritmo, chamado algoritmo CCD (do inglés *cyclic coordinate descent*), é utilizado para resolver o problema de optimización numérica que surxe ó minimizar a función obxectivo. O tal algoritmo é capaz de resolver dito problema, sendo ademáis rápido e eficiente, o cal resulta ser unha vantaxe incomparable, tendo en conta a dimensionalidade dos datos coa que estamos traballando. O CCD minimiza a función obxectivo iterando en cada un dos coeficientes namentres os tempos de computación non se resinten.

Obtivéronse resultados tanto para datos reais coma simulados. Dous conxuntos de datos de leucemia e cancro de colon con resposta binaria, moi comúnmente usados na literatura para probar novas técnicas, foron usadas co obxectivo de probar os nosos métodos, e os resultados obtidos foron comparados con outros publicados na literatura científica. Ademáis, extraéronse conclusións de valor no que se refire ós resultados de leucemia, que tamén foron comparados con estudos previos. En resumo, con estes métodos de penalización é posible obter modelos biolóxicamente interpretables, e competitivos cos resultados previos obtidos neste eido.

Capítulo 4. Os support vector machines (SVMs) apareceron no campo da machine learning nos anos noventa como unha técnica de clasificación de patróns, que rápidamente adquiriu moita relevancia. A idea principal dos SVMs é construir un hiperplano separador entre clases (as diferentes categorías da variable resposta) nun espacio transformado de alta dimensión, no que a separabilidade é fácilmente obtible. Con este obxectivo, os SVMs usan unha aproximación tipo kernel, que nun mesmo cálculo obtén unha medida de similaridade entre individuos, traballando no espacio transformado. A elección do kernel é fundamental cara a obter unha clasificación precisa. Neste capítulo construimos un novo método kernel, preparado para traballar con datos categóricos como os dos SNPs. Esto foi necesario dado que aproximacións previas ó problema do kernel pensaban só en problemas con datos continuos. Este novo kernel toma a forma:

$$K(\mathbf{x}_i, \mathbf{x}_k) = \sum_{s=1}^{T_i} \sum_{m=1}^{T_k} \frac{1}{T_i} \frac{1}{T_k} \left(\sum_{l=1}^{2p} \sum_{r=l}^{2p} w_{lr} I \left\{ z_{i(s)}^{lr} = z_{k(m)}^{lr} \right\} \right)$$

onde se calcula a similaridade entre os xenotipos dos individuos $\mathbf{x}_i$ e $\mathbf{x}_k$ no espacio transformado.

Os resultados de clasificación obtidos compáranse cos de técnicas similares. A demanda computacional desta aproximación SVM é moi alta. Traballando en computación paralela utilizando dúas diferentes infraestruturas GRID (CESGA e Departamento de Estatística e IO en Santiago) somos capaces de reducir os tempos de computación e permitir a factibilidade computacional.

Capítulo 5. Este capítulo componse, a diferencia dos anteriores, de dous traballos nos que se evalúan as capacidades de técnicas estadísticas en datos simulados (**Sección 5.1**) e reais (**Sección 5.2**).

Na **Sección 5.1** pártese da idea de que a maioría das enfermedades comúns probablemente posúen unha etioloxía complexa. Os métodos estatísticos que se centran na busca de epístasis entre diferentes marcadores ou rexións xénicas son de crecente interés, dado que espérase con elas identificar zonas que de outro modo serían indetectables. Nesta sección analizamos a capacidade da regresión loxística (LR) a dúas técnicas de aprendizaxe supervisado tipo árbore: árbores de clasificación (CART) e random forests (RF), á hora de detectar epístasis. O método MDR (do inglés, *multifactor dimensionality reduction*) foi tamén usado con fins comparativos. Partindo da simulación de SNPs autosómicos onde dous dos SNPs son causais interactuando entre si e co estatus da enfermidade, modelamos dita interacción en diferentes escenarios de tamaño de mostra, frecuencia alélica mínima (MAF), porcentaxe de datos perdidos e diversos modelos de penetrancia, algúns deles simulando interaccións puras (sen rastro de presencia de efectos marxinais). Todo esto da lugar a 99 escenarios de simulación diferentes.

Inda que CART, RF e LR ofrecen resultados similares no que atingue a detección da asociación, CART e RF funcionan millor no que respecta ó erro de clasificación. O MAF, a penetrancia e o tamaño de mostra semellan ser factores moito mais determinantes que a porcentaxe de datos perdidos. Nos escenarios de interacción pura tan so os RF son capaces de detectar a asociación dun modo similar ó MDR. En conclusión, os métodos tipo árbore e a LR son ferramentas estadísticas de importancia no que se refire a detección das interaccións en situacións de exceso de SNPs de ruido. Nos modelos de interacción pura, só RF e MDR dan resultados mínimamente aceptables. Nembargantes, cando o diseño do estudio non é óptimo existe unha alta probabilidade de detectar asociacións espúreas.

Na **Sección 5.2** pártese da proposta, común na literatura científica, de que o a probabilidade de sufrir cancro de mama podería ser explicada polo efecto acumulativo dunha grande cantidade de alelos cun efecto moi débil. O represor transcripcional FBI1, tamén coñecido como *Pokemon*, foi recentemente identificado coma un factor crítico na oncoxénese. Esta proteína é codificada polo xen *ZBTB7*. Este estudo ten como obxectivo determinar si

os polimorfismos dentro do xen *ZBTB7* están asociados co risco de sufrir cancro de mama. Utilízase unha mostra de casos e controis recolectada en hospitais no norte e centro de España. Quince SNPs foron xenotipados, cunha cobertura promedio dun SNP por cada 2.4 kilobases, en 360 casos de cancro de mama esporádico e 402 controis. A comparativa de frecuencias haplotípicas, xenotípicas e alélicas non revela asociacións significativas. Un procedimento basado na permutación é usado para correxir por test múltiple. Neste primeiro estudio involucrando o xen *ZBTB7* co cancer de mama esporádico non se aprecia evidencia algunha de asociación.

Capítulo 6. Do mesmo modo que o capítulo anterior, este está composto por dous traballos nos que o fío conductor é o uso de ferramentas estatísticas e de simulación en problemas complexos de xenética forense. No primeiro (**Sección 6.1**) estúdase como a adición dun "plex" de SNPs pode ser de gran axuda na resolución de probas de paternidade en situacións complexas, namentres que no segundo (Sección 6.2) demóstrase a existencia dun claro problema de estratificación poboacional na Arxentina, o cal podería ocasionar problemas nas probas de parentesco que alí se realicen. Vexámolo cun pouco mais de detalle:

A **Sección 6.1** evalúa o salto de mellora que se obtén usando un "plex" de SNPs como complemento en problemas de identificación forense. En moitas ocasións, cando se usa unha batería estandar de STRs en casos forenses de parentesco, unha pequena proporción dos casos poden dar lugar a resultados lixeiramente ambiguos. Moitos de estes casos aparecen en estudos de paternidade onde o presunto pai non está dispoñible i é necesario recurrir a un irmán deste. Inda que a adición dunha certa cantidade de STRs podería axudar a resolver estes casos, non son moitos os STRs dispoñibles. Neste estudo móstrase que grandes multiplexes de SNPs son moi informativos naqueles casos onde se obteñen probabilidades de paternidade pouco informativas ou exclusións (da paternidade) ambiguas. Ó mesmo tempo, os SNPs ofrecen resultados dunha maior fiabilidade, o cal é moi importante en casos con mostras de ADN degradado.

Neste estudo móstranse oito casos reais de parentesco nos que a adición de datos de SNPs resolveu o problema de resultado ambiguo previamente obtido usando tan só STRs. Además, realizáronse simulacións que permitiron determinar a frecuencia dos fracasos á hora de obter exclusións ou probabilidades de paternidade concluintes con diferentes conxuntos de marcadores naqueles casos nos que un irmán do verdadeiro pai é utilizado na proba. Os resultados indican que os SNPs son estadísticamente mais eficientes que os STRs de cara a resolver este tipo de casos.

Na **Sección 6.2**, un estudo de simulación foi levado a cabo co obxectivo de investigar os efectos potenciais da subestructuración poboacional en probas de paternidade en Arxentina. O estudio foi realizado mediante a avaliación dos índices de paternidade (PI) calculados en diferentes escenar-

ios simulados de pedigrís familiares, usando 15 STRs autosómicos en oito bases de datos de subpoboacións da Arxentina.

Os resultados mostran importantes diferencias estatísticamente significativas entre os valores de PI obtidos, segundo as frecuencias alélicas utilizadas (que varían en cada unha das subpoboacións presentes na Arxentina). Estas diferencias son mais dramáticas segundo se consideren poboacións nativo–americanas ou poboacións urbanas. O estudo tamén mostra que o uso do indicador *Fst* á hora de correxir no PI o efecto da estratificación poboacional podería ser inapropiado, dado que non ten en conta as particularidades de cada un dos casos de paternidade que chegan ós xulgados.

Inda que cada un dos capítulos e seccións que presentan un traballo de investigación conteñen as conclusións propias correspondentes a cada un deles, o **Capítulo 7** conten un conxunto de conclusións mais "xerais" e, en certo modo, aplicables ó contexto común da estatística xenética.

Para finalizar, o **Capítulo 8** e último fai referencia ás liñas de investigación futuras relacionadas con cada un dos traballos que se expoñen na memoria. Compre decir que estas liñas son só unha pequena mostra das que realmente existen nun campo tan en auxe coma o que se trata neste traballo.

Bibliography

[1] Abecasis G.R. and Cookson W.O. *GOLD–graphical overview of linkage disequilibrium.* Bioinformatics, **16**, (2000), 182–183.

[2] Acir N., Ozdamar O. and Guzelis C. *Automatic classification of auditory brainstem responses using SVM–based feature selection algorithm for threshold selection.* Engineering Applications of Artificial Intelligence, **19**, (2006), 209–218.

[3] Agresti A. *Categorical data analysis.* John Wiley and Sons, San Francisco, (2002).

[4] Aguzzi A. and Weissmann C. *A suspicious signature.* Nature, **383**, (1996), 666–667.

[5] Alizadeh A.A., Eisen M.B., Davis R.E., Ma C., Lossos I.S., Rosenwald A., Boldrick J.C., Sabet H., Tran T., Yu X., Powell J.I., Yang L., Marti G.E., Moore T., Hudson Jr J., Lu L., Lewis D.B., Tibshirani R., Sherlock G., Chan W.C., Greiner T.C., Weisenburger D.D., Armitage J.O., Warnke R., Levy R., Wilson W., Grever M.R., Byrd J.C., Botstein D., Brown P.O. and Staudt L.M. *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature, **403**, (2000), 503–511.

[6] Alon U., Barkai N., Notterman D., Gish K., Ybarra S., Mack D. and Levine A.J. *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.* Proceedings of the National Academy of Sciences USA, **96**, (1999), 6745–6750.

[7] Alvarez–Iglesias V., Jaime J.C., Carracedo A. and Salas A. *Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups.* Forensic Science International: Genetics, **1**, (2007), 44–55.

[8] Amari S. and Wu S. *Improving support vector machine classifiers by modifying kernel functions.* Neural Networks, **12**, (1999), 783–789.

[9] Amit Y. and Geman D. *Shape quantization and recognition with randomized trees.* Neural Computation, **9**, (1997), 1545–1588.

[10] Amorim A., Alves C., Pereira L. and Gusmao L. *Genotyping inconsistencies and null alleles using AmpFlSTR®, Identifiler® and Powerplex® 16 kits.* Progress in Forensic Genetics, **10**, (2004), 176–178.

[11] Amorim A. and Pereira L. *Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs.* Forensic Science International, **150**, (2005), 17–21.

[12] Amundadottir L. et al *Genome–wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer.* Nature Genetics, **41**, (2009), 986–990.

[13] Anderson J. and Rosenfeld E. (editors) *Neurocomputing: Foundations of Research.* MIT Press, Cambridge, Massachussets, (1988).

[14] Anderson T.W. *Asymptotic theory for principal components analysis.* The Annals of Mathematical Statistics, **34**, (1963), 122–148.

[15] Angele S., Romestaing P., Moullan N., Vuillaume M., Chapot B., Friesen M., Jongmans W., Cox D.G., Pisani P., Gerard J.P. and Hall J. *ATM haplotypes and cellular response to DNA damage: association with breast cancer risk and clinical radiosensitivity.* Cancer Research, **63**, (2003), 8717–8725.

[16] Antoniadis A. and Fan J. *Regularization of wavelet approximations.* Journal of the American Statistical Association, **96**, (2001), 939–967.

[17] Antoniadis A., Gijbels I. and Nikolova M. *Penalized likelihood regression for generalized linear models with nonquadratic penalties.* Annals of the Institute of Statistical Mathematics (in press), (2009).

[18] Antoniadis A., Lambert–Lacroix S. and Leblanc F. *Effective dimension reduction methods for tumor classification using gene expression data.* Bioinformatics, **19**, (2003), 563–570.

[19] Antoniou A.C. and Easton D.F. *Models of genetic susceptibility to breast cancer.* Oncogene, **25**, (2006), 5898–5905.

[20] Antonyuk A. and Holmes C. *On testing for genetic association in case–control studies when population allele frequencies are known.* Genetic Epidemiology, **33**, (2009), 371–378.

[21] Armitage P. *Tests for linear trends in proportions and frequencies.* Biometrics, **11**, (1955), 375–386.

[22] Armstrong S.A., Staunton J.E., Silverman L.B., Pieters R., den Boer M.L., Minden M.D., Sallan S.E., Lander E.S., Golub T.R. and Korsmeyer S.J. *MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia.* Nature Genetics, **30**, (2001), 41–47.

[23] Ayres K.L. *Relatedness testing in subdivided populations.* Forensic Science International, **114**, (2000), 107–115.

[24] Baer A., Lie–Injo L.E., Welch Q.B., Lewis A.N. *Genetic factors and malaria in the Temuan.* American Journal of Human Genetics, **28**, (1976), 179–188.

[25] Bakin S. *Adaptive regression and model selection in data mining problems.* PhD Thesis, Australian National University, Canberra, (1999).

[26] Bao L. and Cui Y. *Prediction of the phenotypic effects of non–synonymous single nucleotide polymorphisms using structural and evolutionary information.* Bioinformatics, **21**, (2005), 2185–2190.

[27] Bar–Joseph Z., Gerber G.K., Gifford D.K., Jaakkola T.S. and Simon I. *Continuous representation of time–series gene expression data.* Journal of Computational Biology, **10**, (2003), 341–356.

[28] Barnes M.R. (editor) *Bioinformatics for geneticists.* John Wiley and Sons, San Francisco, (2007).

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

[30] Bass M.P., Martin E.R. and Hauser E.R. *Pedigree generation for analysis of genetic linkage and association.* Pacific Symposium in Biocomputing, (2004), 93–103.

[31] Bass N.J., Datta S.R., McQuillin A., Puri V., Choudhury K., Thirumalai S., Lawrence J., Quested D., Pimm J., Curtis D. and Gurling H.M.D. *Evidence for the association of the DAOA (G72) gene with schizophrenia and bipolar disorder but not for the association of the DAO gene with schizophrenia.* Behavioral and Brain Functions, **5**:28, (2009).

[32] Bastone L., Reilly M., Rader D.J. and Foulkes A.S. *MDR and PRP: a comparison of methods for high–order genotype–phenotype associations.* Human Heredity, **58**, (2004), 82–92.

[33] Baudat G. and Anouar F. *Generalized discriminant analysis using a kernel approach.* Neural Computation, **12**, (2000), 2385–2404.

[34] Benjamini Y. and Hochberg Y. *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society Series B, **57**, (1995), 289–300.

[35] Berstein J.L., Concannon P., Langholz B., Thompson W.D., Bernstein L., Stovall M., Thomas D.C. and the WECARE Study. *Multi–center*

screening of mutations in the ATM gene among women with breast cancer. Radiation Research, **163**, (2005), 698–699.

[36] Bittner M., Meltzer P., Chen Y., Jiang Y., Seftor E., Hendrix M., Radmacher M., Simon R., Yakhini Z., Ben–Dor A., Sampask N., Dougherty E., Wang E., Marincola F., Gooden C., Lueders J., Glatfelter A., Pollock P., Carpten J., Gillanders E., Leja D., Dietrich K., Beaudry C., Berens M., Alberts D., Sondak V., Hayward N. and Trent J. *Molecular classification of cutaneous malignant melanoma by gene expression profiling.* Nature, **406**, (2000), 536–540.

[37] Bo T. and Jonassen I. *New feature subset selection procedures for classification of expression profiles.* Genome Biology, **3**, (2002), 0017.

[38] Bonferroni C.E. *Il calcolo delle assicurazioni su gruppi di teste.* Studi in Onore del Professore Salvatore Ortu Carboni, (1935), 13–60.

[39] Bonilla C., Boxill L.A., Mc Donald S.A., Williams T., Sylvester N., Parra E.J., Dios S., Norton H.L., Shriver M.D. and Kittles R.A. *The 8818G allele of the agouti signaling protein (ASIP) gene is ancestral and is associated with darker skin color in African Americans.* Human Genetics, **116**, (2005), 402–406.

[40] Borecki I.B. and Suarez B.K. *Linkage and association: basic concepts.* Advances in Genetics, **42**, (2001), 45–66.

[41] Borsting C., Sanchez J.J., Birk A.H., Bruun C., Hallenberg C., Hansen A.J., Hansen H.E., Simonsen B.T. and Morling N. *Comparison of paternity indices based on typing 15 STRs, 7 VNTRs and 52 SNPs in 50 Danish mother–child–father trios.* Progress in Forensic Genetics, **11**, (2006), 436–438.

[42] Boser B., Guyon I. and Vapnik V. *A training algorithm for optimal margin classifiers.* Fifth Annual Workshop on Computational Learning Theory, New York, (1992).

[43] Boulesteix A.L., Tutz G. and Strimmer K. *A CART–based approach to discover emerging patterns in microarray data.* Bioinformatics, **19**, (2003), 2465–2472.

[44] Bradley P.S. and Mangasarian O.L. *Massive data discrimination via linear support vector machines.* Optimization Methods and Software, **13**, (2000), 1–10.

[45] Brailovsky V.L., Barzilay O. and Shahave R. *On global, local, mixed and neighborhood kernels for support vector machines.* Pattern Recognition Letters, **20**, (1999), 1183–1190.

[46] Breiman L. *Arcing classifiers.* Annals of Statistics, **26**, (1998), 801–849.

[47] Breiman L. *Bagging predictors.* Machine Learning, **26**, (1996), 123–140.

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

[49] Breiman L. *Stacked regressions.* Machine Learning, **24**, (1996), 51–64.

[50] Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. *Classification and regression trees.* Wadsworth International Group, Belmont, California, (1984).

[51] Brinkmann B. *The STR approach.* In Advances in Forensic Haemogenetics 6, Springer, Berlin, (1996).

[52] Brinkmann B., Klintshar M., Neuhuber F., Huhne J. and Rolf B. *Mutation rates in human microsatellites: influence of the structure and length of the tandem repeat.* American Journal of Human Genetics, **62**, (1998), 1408–1415.

[53] Brion M., Sanchez J.J., Balogh K., Thacker C., Blanco–Verea A., Borsting C., Stradmann–Bellinghausen B., Bogus M., Syndercombe–Court D., Schneider P.M., Carracedo A. and Morling N. *Introduction of an single nucleotide polymorphism–based major Y–chromosome haplogroup typing kit suitable for predicting the geographical origin of male lineages.* Electrophoresis, **26**, (2005), 4411–4420.

[54] Brown M.P.S., Grundy W.N., Lin D., Cristianini N., Sugnet C.W., Furey T.S., Ares Jr M. and Haussler D. *Knowledge–based analysis of microarray gene expression data by using support vector machines.* Proceedings of the National Academy of Sciences USA, **97**, (2000), 262–267.

[55] Bumgarner R.E. and Yeung K.Y. *Methods for the inference of biological pathways and networks.* Methods in Molecular Biology, **541**, (2009), 225–245.

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

[57] Burges C.J.C. *A tutorial on support vector machines for pattern recognition.* Data Mining and Knowledge Discovery, **2**, (1998), 121–167.

[58] Butler J.M. *Forensic DNA typing: biology, technology, and genetics of STR markers.* Elsevier, New York, (2005).

[59] Butler J.M. *Genetics and genomics of core short tandem repeat loci used in human identity testing.* Journal of forensic Science, **51**, (2006), 253–265.

[60] Cabana G.S., Merriwether D.A., Hunley K. and Demarchi D.A. *Is the genetic structure of Gran Chaco populations unique? Interregional perspectives on native South American mitochondrial DNA variation.* American Journal of Physical Anthropology, **131**, (2006), 108–119.

[61] Caporaso N., Gu F., Chatterjee N., Sheng–Chih J., Yu K., Yeager M., Chen C., Jacobs K., Wheeler W., Landi M.T., Ziegler R.G., Hunter D.J., Chanock S., Hankinson S., Kraft P. and Bergen A.W. *Genome–wide and candidate gene association study of cigarette smoking behaviors.* Plos One, **4**:e4653, (2009).

[62] Cardon L.R. and Bell J.I. *Association study designs for complex diseases.* Nature Reviews in Genetics, **2**, (2001), 91–99.

[63] Carracedo A. and Barros F. (editors) *Problemas bioestadisticos en genetica forense.* Universidade de Santiago de Compostela, Santiago de Compostela, (1996).

[64] Carralero Yepes J. *Matematicas aplicadas a la genetica forense.* Ministerio del Interior, Madrid, España, (2007).

[65] Caspi A., Sugden K., Moffitt T.E., Taylor A., Craig I.W., Harrington H., McClay J., Mill J., Martin J., Braithwaite A. and Poulton R. *Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene.* Science, **301**, (2003), 386–389.

[66] Cavalli–Sforza L.L. *The DNA revolution in population genetics.* Trends in Genetics, **14**, (2000), 60–65.

[67] Cerhan J.R., Liu–Mares W., Fredericksen Z.S., Novak A.J., Cunningham J.M., Kay N.E., Dogan A., Liebow M., Wang A.H., Call T.G., Habermann T.M., Ansell S.M. and Slager S.L. *Genetic variation in tumor necrosis factor and the nuclear factor–kappaB canonical pathway and risk of non–Hodgkin's lymphoma.* Cancer Epidemiology, Biomarkers and Prevention, **17**, (2008), 3161–3169.

[68] Chatterjee S. and Price B. *Regression analysis by example.* John Wiley and Sons, San Francisco, (1991).

[69] Chen J., Yu K., Hsing A. and Therneau T.M. *A partially linear tree–based regression model for assessing complex joint gene–gene and gene–environment effects.* Genetic Epidemiology, **31**, (2007), 238–251.

[70] Chen Y., Dougherty E.R. and Bittner M.L. *Ratio–based decisions and the quantitative analysis of cDNA microarray images.* Journal of Biomedical Optics, **2**, (1997), 364–374.

[71] Cheng Y. and Church G.M. *Biclustering of expression data.* Proceedings of the Eight International Conference in Intelligent Systems for Molecular Biology, (2000), 93–103.

[72] Cho Y.M., Ritchie M.D., Moore J.H., Park J.Y., Lee K.U., Shin H.D., Lee H.K. and Park K.S. *Multifactor–dimensionality reduction shows a two–locus interaction associated with type 2 diabetes mellitus.* Diabetologia, **47**, (2004), 549–554.

[73] Clark T.G., De Iorio M. and Griffiths R.C. *Bayesian logistic regression using a perfect phylogeny.* Biostatistics, **8**, (2007), 32–52.

[74] Cleveland W.S. *Robust locally weighted regression and smoothing scatterplots.* Journal of the American Statistical Association, **74**, (1979), 829–836.

[75] Coffey C.S., Hebert P.R., Ritchie M.D., Krumholz H.M., Gaziano J.M., Ridker P.M., Brown N.J., Vaughan D.E. and Moore J.H. *An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene–gene interactions on risk of myocardial infarction: the importance of model validation.* BMC Bioinformatics, **5**:49, (2009).

[76] Colantuoni C., Jeon O.H., Hyder K., Chenchik A., Khimani A.H., Narayanan V., Hoffman E.P., Kaufmann W.E., Naidu S. and Pevsner J. *Gene expression profiling in postmortem Rett syndrome brain: differential gene expression and patient classification.* Neurobiology of Disease, **8**, (2001), 847–865.

[77] Collinge J., Palmer M.S. and Dryden A.J. *Genetic predisposition to iatrogenic Creutzfeldt-Jakob disease.* Lancet, **337**, (1991), 1441–1442.

[78] Collins P.J., Hennessy L.K., Leibelt C.S., Roby R.K., Reeder D.J. and Foxall P.A. *Developmental validation of a single–tube amplification of the 13 CODIS STR loci, D2S1338, D19S433, and amelogenin: the AmpFlSTR Identifiler PCR Amplification Kit.* Journal of Forensic Science, **49**, (2004), 1265–1277.

[79] Cook E.H. and Scherer S.W. *Copy–number variations associated with neuropsychiatric conditions.* Nature, **455**, (2008), 919–923.

[80] Cook N.R., Zee R.Y.L. and Ridker P.M. *Tree and spline based association analysis of gene–gene interaction models for ischemic stroke.* Statistics in Medicine, **23**, (2004), 1439–1453.

[81] Cordell H.J., Barratt B.J. and Clayton D.G. *Case/Pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene–gene and gene–environment*

interactions, and parent–of–origin effects. Genetic Epidemiology, **26**, (2004), 167–185.

[82] Cortes C. and Vapnik V. *Support vector networks.* Machine Learning, **20**, (1995), 273–297.

[83] Cupples L.A. *Family study designs in the age of genome-wide association studies: experience from the Framingham heart study.* Current Opinion in Lipidology, **19**, (2008), 144–150.

[84] Curtis D. *Comparison of artificial neural network analysis with other multimarker methods for detecting genetic association.* BMC Genetics, **8**:49, (2007).

[85] Cussenot O., Azzouzi A.R., Bantsimba–Malanda G., Gaffory C., Mangin P., Cormier L., Fournier G., Valeri A., Jouffe L., Roupret M., Fromont G., Sibony M., Comperat E. and Cancel–Tassin G. *Effect of genetic variability within 8q24 on aggressiveness patterns at diagnosis and familial status of prostate cancer.* Clinical Cancer Research, **14**, (2008), 5635–5639.

[86] Davis C., Levitan R.D., Kaplan A.S., Carter J., Reid C., Curtis C., Patte K., Hwang R. and Kennedy J.L. *Reward sensitivity and the D2 dopamine receptor gene: a case-control study of binge eating disorder.* Progress in Neuro-Psychopharmacology and Biological Psychiatry, **32**, (2008), 620–628.

[87] De Mol C., De Vito E. and Rosasco L. *Elastic–net regularization in learning theory.* Journal of Complexity, **25**, (2009), 201–230.

[88] Dempster A., Laird N. and Rubin D. *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society Series B, **39**, (1977), 1–38.

[89] Dettling M. *BagBoosting for tumor classification with gene expression data.* Bioinformatics, **20**, (2004), 3583–3593.

[90] Dettling M. and Buhlmann P. *Finding predictive gene groups from microarray data.* Journal of Multivariate Analysis, **90**, (2004), 106–131.

[91] Dettling M. and Buhlmann P. *Supervised clustering of genes.* Genome Biology, **3**, (2002), 0069.1–0069.15.

[92] Devlin B. and Roeder K. *Genomic control for association studies.* Biometrics, **55**, (1999), 997–1004.

[93] Diaz–Uriarte R. and Alvarez de Andres S. *Gene selection and classification of microarray data using random forest.* BMC Bioinformatics, **7**:3, (2006).

[94] Dietterich T.G. *An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization.* Machine Learning, **40**, (2000), 139–157.

[95] Ding C. and Peng H. *Minimum redundancy feature selection from microarray gene expression data.* Journal of Bioinformatics and Computational Biology, **3**, (2005), 185–205.

[96] Dixon L.A., Murray C.M., Archer E.J., Dobbins A.E., Koumi P. and Gill P. *Validation of a 21–locus autosomal SNP multiplex for forensic identification purposes.* Forensic Science International, **154**, (2005), 62–77.

[97] Duda R.O. and Hart P.E. *Pattern classification and scene analysis.* John Wiley and Sons, San Francisco, (1973).

[98] Dudbridge F. *Pedigree disequilibrium tests for multilocus haplotypes.* Genetic Epidemiology, **25**, (2003), 115–121.

[99] Dudek S., Motsinger A.A., Velez D., Williams S.M. and Ritchie M.D. *Data simulation software for whole–genome association and other studies in human genetics.* Pacific Symposium in Biocomputing, (2006), 499–510.

[100] Dudoit S. and Fridlyand J. *Bagging to improve the accuracy of a clustering procedure.* Bioinformatics, **19**, (2003), 1090–1099.

[101] Dudoit S., Fridlyand J. and Speed T.P. *Comparison of discrimination methods for the classification of tumors using gene expression data.* Journal of the American Statistical Association, **97**, (2002), 77–87.

[102] Dudoit S. and Yang Y.H. *Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data.* In The Analysis of Gene Expression Data: Methods and Software, Springer, (2002), 73–101.

[103] Dudoit S., Yang Y.H. and Bolstad B. *Using R for the analysis of DNA microarray data.*, R News, **2**, (2002), 24–32.

[104] Eberle M.A., Ng P.C., Kuhn K., Zhou L., Peiffer D.A., Galver L., Viaud–Martinez K.A., Taylor Lawley C., Gunderson K.L., Shen R. and Murray S.S. *Power to detect risk alleles using genome–wide tag SNP panels.* Plos Genetics, **3**, (2007), 1827–1837.

[105] Edwards A., Civitello A., Hammond H.A. and Caskey C.T. *DNA typing and genetic mapping with trimeric and tetrameric tandem repeats.* American Journal of Human Genetics, **49**, (1991), 746–756.

[106] Edwards A., Hammond H.A., Jin L., Caskey C.T. and Chakraborty R. *Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups.* Genomics, **12**, (1992), 241–253.

[107] Edwards T.L., Bush W.S., Turner S.D., Dudek S.M., Torstenson E.S., Schmidt M., Martin E. and Ritchie M.D. *Generating linkage disequilibrium patterns in data simulations using genomeSIMLA.* In Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, Springer, Berlin/Heidelberg, (2008).

[108] Efron B. *Bootstrap methods: another look at the jackknife.* Annals of Statistics, **7**, (1979), 1–26.

[109] Efron B. *Large–scale simultaneous hypothesis testing: the choice of a null hypothesis.* Journal of the American Statistical Association, **99**, (2004), 96–104.

[110] Efron B., Hastie T., Johnstone I. and Tibshirani R. *Least angle regression.* Annals of Statistics, **32**, (2004), 407–499.

[111] Efron B. and Tibshirani R. *An Introduction to the Bootstrap.* Chapman and Hall, London, (1993).

[112] Efron B. and Tibshirani R. *On testing the significance of sets of genes.* Annals of Applied Statistics, **1**, (2007), 107–129.

[113] Efron B., Tibshirani R., Storey J.D. and Tusher V. *Empirical Bayes analysis of a microarray experiment.* Journal of the American Statistical Association, **96**, (2001), 1151–1160.

[114] Egeland T. and Mostad P.F. *Statistical genetics and genetical statistics: a forensic perspective.* Scandinavian Journal of Statistics, **29**, (2002), 297–307.

[115] Egeland T., Mostad P., Mevag B. and Stenersen M. *Beyond traditional paternity and identification cases. Selecting the most probable pedigree.* Forensic Science International, **110**, (2000), 47–59.

[116] Egeland T. and Salas A. *Estimating haplotype frequency and coverage of databases.* PLoS ONE, **3**:e3988, (2008).

[117] Egeland T. and Salas A. *Statistical evaluation of haploid genetic evidence.* The Open Forensic Science Journal, **1**, (2008), 4–11.

[118] Eilers P.H.C., Boer J.M., van Ommen G.J. and van Houwelingen H.C. *Classification of microarray data with penalized logistic regression.* Proceedings of the SPIE, **4266**, (2001), 187.

[119] Eley T.C., Sugden K., Corsico A., Gregory A.M., Sham P., McGuffin P., Plomin R. and Craig I.W. *Gene–environment interaction analysis of serotonin system markers with adolescent depression.* Molecular Psychiatry, **9**, (2004), 908–915.

[120] Essen–Moller E. *Die beweiskraft der ahnlichkeit im vaterschaftsnachweis– theoretische grundlagen.* Mitteilungen der Anthropologischen Gesellschaft (Wien) **68**, (1938), 9–53.

[121] Essen–Moller E. and Quensel C. *Zur theorie des vaterschaftsnachweises aufgrund von ahnlichkeitsbefunden.* International Journal of Legal Medicine, **31**, (1939), 70–96.

[122] Evans D.M., Marchini J., Morris A.P. and Cardon L.R. *Two–stage two–locus models in genome–wide association.* Plos Genetics, **2**, (2006), 1424–1432.

[123] Evett I. *Bayesian inference and forensic science: problems and perspectives.* The Statistician, **36**, (1987), 99–105.

[124] Evett I. and Weir B. *Interpreting DNA evidence. Statistical genetics for forensic scientists.* Sinaur Associates Inc., Sunderland, Massachusets, (1998).

[125] Fan J. and Li R. *Variable selection via nonconcave penalized likelihood and its oracle properties.* Journal of the American Statistical Association, **96**, (2001), 1348–1360.

[126] Fan J. and Ren Y. *Statistical analysis of DNA microarray data in cancer research.* Clinical Cancer Research, **12**, (2006), 4469–4473.

[127] Fimmers R., Henke L., Henke J. and Baur M.P. *How to deal with mutations in DNA–testing.* Advances in Forensic Haemogenetics, **4**, (1992), 285–287.

[128] Fisher R.A. *On the interpretation of χ^2 from contingency tables, and the calculation of P.* Journal of the Royal Statistical Society, **85**, (1922), 87–94.

[129] Fisher R.A. *Statistical methods for research workers.* Oliver and Boyd, Edinburgh, (1954).

[130] Fletcher R. *Practical methods of optimization.* John Wiley and Sons, Chichester, (1987).

[131] Flintoft L. *Onwards and upwards for genome–wide association studies.* Nature Reviews in Genetics, **8**, (2007), 494–495.

[132] Fodor S.P., Rava R.P., Huang X.C., Pease A.C., Holmes C.P. and Adams C.L. *Multiplexed biochemical assays with biological chips.* Nature, **364**, (1993),555–556.

[133] Fondevila M., Phillips C., Naveran N., Cerezo M., Rodriguez A., Calvo R., Fernandez L.M., Carracedo A. and Lareu M.V. *Challenging DNA: assessment of a range of genotyping approaches for highly degraded forensic samples.* Forensic Science International: Genetics Supplement Series, **1**, (2008), 26–28

[134] Fondevila M., Phillips C., Naveran N., Fernandez L., Cerezo M., Salas A., Carracedo A. and Lareu M.V. *Identification of skeletal remains using short–amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur.* Forensic Science International: Genetics, **2**, (2008), 212–218.

[135] Ford D., Easton D.F., Stratton M., Narod S., Goldgar D., Devilee P., Bishop D.T., Weber B., Lenoir G., Chang–Claude J., Sobol H., Teare M.D., Struewing J., Arason A., Scherneck S., Peto J., Rebbeck T.R., Tonin P., Neuhausen S., Barkardottir R., Eyfjord J., Lynch H., Ponder B.A.J., Gayther S.A., Birch J.M., Lindblom A., Stoppa–Lyonnet D., Bignon Y., Borg A., Hamann U., Haites N., Scott R.J., Maugard C.M., Vasen H., Seitz S., Cannon–Albright L.A., Schofield A., Zelada–Hedman M. and the Breast Cancer Linkage Consortium. *Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families.* American Journal of Human Genetics, **62**, (1998), 676–689.

[136] Forgy E. *Cluster analysis of multivariate data: efficiency vs. interpretability of classifications.* Biometrics, **21**, (1965), 768–769.

[137] Frank I.E. and Friedman J.H. *A statistical view of some chemometrics tools.* Technometrics, **35**, (1993), 109–135.

[138] Fregeau C.J. and Fourney R.M. *DNA typing with fluorescently tagged short tandem repeats: a sensitive and accurate approach to human identification.* BioTechniques, **15**, (1993), 100–119.

[139] Freund Y. and Schapire R. *A decision–theoretic generalization of online learning and an application to boosting.* Journal of Computer and System Sciences, **55**, (1997), 119–139.

[140] Freund Y. and Schapire R. *Experiments with a new boosting algorithm.* Machine Learning: Proceedings of the Thirteenth International Conference, (1996), 148–156.

[141] Fridley B.L. *Bayesian variable and model selection methods for genetic association studies.* Genetic Epidemiology, **33**, (2009), 27–37.

[142] Friedman J. *Multivariate adaptive regression splines.* Annals of Statistics, **19**, (1991), 1–141.

[143] Friedman J., Hastie T. and Tibshirani R. *Regularization paths for generalized linear models via coordinate descent.* Technical Report, Department of Statistics, Stanford University, (2008).

[144] Friedman J., Hastie T. and Tibshirani R. *Additive logistic regression: a statistical view of boosting.* Annals of Statistics, **28**, (2000), 337–407.

[145] Furlanello C., Serafini M., Merler S. and Jurman G. *Entropy–based gene ranking without selection bias for the predictive classification of microarray data.* BMC Bioinformatics, **4**:54, (2003).

[146] Gambaro G., Anglani F. and D'Angelo A. *Association studies of genetic polymorphisms and complex disease.* Lancet, **355**, (2000), 308–311.

[147] Garcia–Closas M., Malats N., Real F.X., Welch R., Kogevinas M., Chatterjee N., Pfeiffer R., Silverman D., Dosemeci M., Tardon A., Serra C., Carrato A., Garcia–Closas R., Castaño–Vinyals G., Chanock S., Yeager M. and Rothman N. *Genetic variation in the nucleotide excision repair pathway and bladder cancer risk.* Cancer Epidemiology Biomarkers and Prevention, **15**, (2006), 536–542.

[148] Garcia–Magariños M., Cao R. and Gonzalez–Manteiga W. *Support vector machine adaptation to biallelic SNP data.* The 24th International Biometric Conference, Dublin, Ireland, (2008).

[149] Garcia–Magariños M., López–de–Ullibarri I. , Cao R. and Salas A. *Evaluating the ability of tree–based methods and logistic regression for the detection of SNP–SNP interaction.* Annals of Human Genetics, **73**, (2009), 360–369.

[150] Gayan J., Gonzalez-Perez A., Bermudo F., Saez M.E., Royo J.L., Quintas A., Galan J.J., Moron F.J., Ramirez–Lorca R., Real L.M. and Ruiz A. *A method for detecting epistasis in genome–wide studies using case–control multi–locus association analysis.* BMC Genomics, **9**:360, (2008).

[151] Genkin A., Lewis D.D. and Madigan D. *Sparse logistic regression for text categorization.* DIMACS Working Group on Monitoring Message Streams, Project Report, (2005).

[152] Gilbert W. *Why genes in pieces?* Nature, **271**, (1978), 501.

[153] Gill P. *An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes.* International Journal of Legal Medicine, **114**, (2001), 204–210.

[154] Gill P. *Role of short tandem repeat DNA in forensic casework in the UK– past, present, and future perspectives.* BioTechniques, **32**, (2002), 366–372.

[155] Ginther C., Corach D., Penacino G.A., Rey J.A., Carnese F.R., Hutz M.H., Anderson A., Just J., Salzano F.M. and King M.C. *Genetic variation among the Mapuche Indians from the Patagonian region of Argentina: mitochondrial DNA sequence variation and allele frequencies of several nuclear genes.* In DNA Fingerprints: State of the Science, Birkhauser, Basel, (1993).

[156] Girosi F. *An equivalence between sparse approximation and support vector machines.* Neural Computation, **20**, (1998), 1455–1480.

[157] Gjertson D.W., Brenner C.H., Baur M.P., Carracedo A., Guidet F., Luque J.A., Lessig R., Mayr W.R., Pascali V.L., Prinz M., Schneider P.M. and Morling N. *ISFG: recommendations on biostatistics in paternity testing.* Forensic Science International: Genetics, **1**, (2007), 223–231.

[158] Goeman J.J. and Buhlmann P. *Analyzing gene expression data in terms of gene sets: methodological issues.* Bioinformatics, **23**, (2007), 980–987.

[159] Goldberg D.E. *Genetic algorithms in search, optimization, and machine learning.* Addison–Wesley, Massachussets, (1989).

[160] Goldstein T. and Osher S. *The Split Bregman method for L1 regularized problems.* UCLA CAAM Report 08–29, (2008).

[161] Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeck M., Mesirov P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A. Bloomfield C.D. and Lander E.S. *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, **286**, (1999), 531–537.

[162] Griffith F. *The significance of pneumococcal types.* Journal of Hygiene, **27**, (1928), 113–159.

[163] Griffith O.L., Melck A., Jones S.J.M. and Wiseman S.M. *Meta–analysis and meta–review of thyroid cancer gene expression profiling studies identifies important diagnostic biomarkers.* Journal of Clinical Oncology, **24**, (2006), 5043–5051.

[164] Griffiths A.J.F., Wessler S.R., Lewontin R.C. and Carroll S.B. *Introduction to genetic analysis.* W.H. Freeman and Company, New York, (2008).

[165] Gurtler H. *Principles of blood group statistical evaluation of paternity cases at the University Institute of Forensic Medicine Copenhagen.* Acta Medicinae Legalis et Socialis (Liege), **9**, (1956), 83–94.

[166] Gusmao L., Sanchez–Diz P., Alves C., Lareu M.V., Carracedo A. and Amorim A. *Genetic diversity of nine STRs in two northwest Iberian populations: Galicia and northern Portugal.* International Journal of Legal Medicine, **114**, (2000), 109–113.

[167] Guyon I. and Elisseeff A. *An introduction to variable and feature selection.* Journal of Machine Learning Research, **3**, (2003), 1157–1182.

[168] Guyon I., Weston J., Barnhill S. and Vapnik V. *Gene selection for cancer classification using support vector machines.* Machine Learning, **46**, (2002), 389–422.

[169] Hahn L.W., Ritchie M.D. and Moore J.H. *Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions.* Bioinformatics, **19**, (2003), 376–382.

[170] Hall M. *Correlation–based feature selection for machine learning.* PhD Thesis, Department of Computer Science, Waikato University, New Zealand, (1999).

[171] Hall P. *The Bootstrap and Edgeworth Expansion.* Springer–Verlag, New York, (1992).

[172] Hall P., Lee E.R. and Park B.U. *Bootstrap–based penalty choice for the lasso, achieving oracle performance.* Statistica Sinica, **19**, (2009), 449–472.

[173] Hall P., Poskitt D.S. and Presnell B. *A functional data–analytic approach to signal discrimination.* Technometrics, **43**, (2001), 1–9.

[174] Hardy H.G. *Mendelian proportions in a mixed population.* Science, **28**, (1908), 49–50.

[175] Hartigan J.A. *Clustering algorithms.* John Wiley and Sons, New York, (1975).

[176] Hastie T., Rosset S., Tibshirani R. and Zhu J. *The entire regularization path for the support vector machine.* The Journal of Machine Learning Research, **5**, (2004), 1391–1415.

[177] Hastie T., Tibshirani R. and Friedman J. *The elements of statistical learning.* Springer, New York, (2001).

[178] He J.Q., Hallstrand T.S., Knight D., Chan–Yeung M., Sandford A., Tripp B., Zamar D., Bosse Y., Kozyrskyj A.L., James A., Laprise C. and Daley D. *A thymic stromal lymphopoietin gene variant is associated with asthma and airway hyperresponsiveness.* Journal of Allergy and Clinical Immunology, **124**, (2009), 222–229.

[179] Heidema A.G., Boer J.M.A., Nagelkerke N., Mariman E.C.M., van der A D.L. and Feskens E.J.M. *The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases.* BMC Genetics, **7**:23, (2006).

[180] Henrichsen C.N., Chaignat E. and Reymond A. *Copy number variants, diseases and gene expression.* Human Molecular Genetics, **18**, (2009), R1–R8.

[181] Hirschhorn J.N. and Daly M.J. *Genome–wide association studies for common diseases and complex traits.* Nature Reviews in Genetics, **6**, (2005), 95–108.

[182] Hjelmervik T.O.R., Petersen K., Jonassen I., Jonsson R. and Bolstad A.I. *Gene expression profiling of minor salivary glands clearly distinguishes primary Sjogren's syndrome patients from healthy control subjects.* Arthritis and Rheumatism, **52**, (2005), 1534–1544.

[183] Hoerl A.E. and Kennard R. *Ridge regression: biased estimation for nonorthogonal problems.* Technometrics, **12**, (1970), 55–67.

[184] Hoggart C.J., Whittaker J.C., De Iorio M. and Balding D.J. *Simultaneous analysis of all SNPs in genome–wide and re–sequencing association studies.* Plos Genetics, **4**, (2008), 1–8.

[185] Hoh J. and Ott J. *Mathematical multi–locus approaches to localizing complex human trait genes.* Nature Reviews in Genetics, **4**, (2003), 701–709.

[186] Hoh J., Wille A. and Ott J. *Trimming, weighting, and grouping SNPs in human case–control association studies.* Genome Research, **11**, (2001), 2115–2119.

[187] Holland J.H. *Adaptation in natural and artificial systems.* The University of Michigan Press, Ann Arbor, (1975).

[188] Holt C.L., Buoncristiani M., Wallin J.M., Nguyen T., Lazaruk K.D. and Walsh P.S. *TWGDAM validation of AmpFlSTR PCR amplification kits for forensic DNA casework.* Journal of Forensic Science, **47**, (2002), 66–96.

[189] Hosmer D.W. and Lemeshow S. *Applied logistic regression.* John Wiley and Sons, New York, (1989).

[190] Hothorn T. and Lausen B. *Building classifiers by bagging trees.* Computational Statistics and Data Analysis, **49**, (2005), 1068–1078.

[191] Houlihan L.M., Christoforou A., Arbuckle M.I., Torrance H.S., Anderson S.M., Muir W.J., Porteous D.J., Blackwood D.H. and Evans K.L. *A case–control association study and family–based expression analysis of the bipolar disorder candidate gene PI4K2B.* Journal of Psychiatric Research (in press), (2009).

[192] Hu N., Wang C., Hu Y., Yang H.H., Giffen C., Tang Z.Z., Han X.Y., Goldstein A.M., Emmert–Buck M.R., Buetow K.H., Taylor P.R. and Lee M.P. *Genome–wide association study in esophageal cancer using genechip mapping 10K array.* Cancer Research, **65**, (2005), 2542–2546.

[193] Huang C.L. and Wang C.J. *A GA–based feature selection and parameters optimization for support vector machines.* Expert Systems with Applications, **31**, (2006), 231–240.

[194] Huang J., Horowitz J.L. and Ma S. *Asymptotic properties of bridge estimators in sparse high–dimensional regression models.* Technical Report, Department of Statistics and Actuarial Science, The University of Iowa, (2006).

[195] Huang J., Ma S. and Zhang C.H. *Adaptive lasso for sparse high–dimensional regression models.* Statistica Sinica, **18**, (2008), 1603–1618.

[196] Huang J., Ma S. and Zhang C.H. *The iterated lasso for high–dimensional logistic regression.* Technical report No. 392, The University of Iowa, (2008).

[197] Huber P.J. *The behavior of maximum likelihood estimates under nonstandard conditions.* Proceedings of the Fifth Berkeley Simposium on Mathematical Statistics and Probability, (1967).

[198] Hughes T.R., Marton M.J., Jones A.R., Roberts C.J., Stoughton R., Armour C.D., Bennett H.A., Coffey E., Dai H., He Y.D., Kidd M.J., King A.M., Meyer M.R., Slade D., Lum P.Y., Stepaniants S.B., Shoemaker D.D., Gachotte D., Chakraburtty K., Simon J., Bard M. and Friend S.H. *Functional discovery via a compendium of expression profiles.* Cell, **102**, (2000), 109–126.

[199] Hunter D.R. and Li R. *Variable selection using MM algorithms.* Annals of Statistics, **33**, (2005), 1617–1642.

[200] Iafrate A.J., Feuk L., Rivera M.N., Listewnik M.L., Donahoe P.K., Qi Y., Scherer S.W. and Lee C. *Detection of large–scale variation in the human genome.* Nature Genetics, **36**, (2004), 949–951.

[201] Ickstadt K., Schafer M., Fritsch A., Schwender H., Abel J., Bolt H.M., Bruning T., Ko Y.D., Vetter H. and Harth V. *Statistical methods for detecting genetic interactions: a head and neck squamous–cell cancer study.*

Journal of Toxicology and Environmental Health Part A, **71**, (2008), 803–815.

[202] Iles M.M. *What can genome–wide association studies tell us about the genetics of common disease?* Plos Genetics, **4**, (2008), e33.

[203] International Human Genome Sequencing Consortium. *Initial sequencing and analysis of the human genome.* Nature, **409**, (2001), 860–921.

[204] Ioannidis J.P. *Effect of formal statistical significance on the credibility of observational associations.* American Journal of Epidemiology, **168**, (2008), 374–383.

[205] Ioannidis J.P., Ntzani E.E., Trikalinos T.A. and Contopoulos–Ioannidis D.G. *Replication validity of genetic association studies.* Nature Genetics, **29**, (2001), 306–309.

[206] Jacobs R.A., Jordan M.I., Nowlan S.J. and Hinton G.E. *Adaptive mixtures of local experts.* Neural Computation, **3**, (1991), 79–87.

[207] Jobling M.A. and Gill P. *Encoded evidence: DNA in forensic analysis.* Nature Reviews in Genetics, **5**, (2004), 739–751.

[208] Jorgenson E. and Witte J.S. *A gene–centric approach to genome–wide association studies.* Nature Reviews in Genetics, **7**, (2006), 885–891.

[209] Karlsson A.O., Holmlund G., Egeland T. and Mostad P. *DNA testing for immigration cases: the risk of erroneous conclusions.* Forensic Science International, **172**, (2007), 144–149.

[210] Kepler T.B., Crosby L. and Morgan K.T. *Normalization and analysis of DNA microarray data by self-consistency and local regression.* Genome Biology, **3**, (2002), 0037.1–0037.12.

[211] Kidd J.M. et al *Mapping and sequencing of structural variation from eight human genomes.* Nature, **453**, (2008), 56–64.

[212] Kim Y., Choi H. and Oh H.S. *Smoothly clipped absolute deviation on high dimensions.* Journal of the American Statistical Association, **103**, (2008), 1665–1673.

[213] Kimpton C.P., Gill P., Walton A., Urquhart A., Millican E.S. and Adams M. *Automated DNA profiling employing multiplex amplification of short tandem repeat loci.* PCR Methods and Application, **3**, (1993), 13–22.

[214] Klinger A. *Inference in high dimensional generalized linear models based on soft thresholding.* Journal of the Royal Statistical Society Series B, **63**, (2002), 377–392.

[215] Knight K. and Fu W.J. *Asymptotics for lasso–type estimators.* Annals of Statistics, **28**, (2000), 1356–1378.

[216] Kohonen T. *Self–organization and associative memory.* Springer–Verlag, Berlin, (1989).

[217] Kohonen T. *The self–organizing map.* Proceedings of the IEEE, **78**, (1990), 1464–1479.

[218] Kohonen T., Kaski S., Lagus K., Salojarvi J., Paatero A. and Saarela A. *Self organization of a massive document collection.* IEEE Transactions on Neural Networks, **11**, (2000), 574–585.

[219] Kooperberg C. and Ruczinski I. *Identifying interacting SNPs using Monte Carlo logic regression.* Genetic Epidemiology, **28**, (2005), 157–170.

[220] Kooperberg C., Ruczinski I., LeBlanc M.L. and Hsu L. *Sequence analysis using logic regression.* Genetic Epidemiology, **21**, (2001), S626–S631.

[221] Korbel J.O., Kim P.M., Chen X., Urban A.E., Weissman S., Snyder M. and Gerstein M.B. *The current excitement about copy-number variation: how it relates to gene duplications and protein families.* Current Opinion in Structural Biology, **18**, (2008), 366–374.

[222] Krenke B.E., Tereba A., Anderson S.J., Buel E., Culhane S., Finis C.J., Tomsey C.S., Zachetti J.M., Masibay A., Rabbach D.R., Amiott E.A. and Sprecher C.J. *Validation of a 16–locus fluorescent multiplex system.* Journal of Forensic Science, **47**, (2002), 773–785.

[223] Krings M., Stone A., Schmitz R.W., Krainitzki H., Stoneking M. and Pääbo S. *Neandertal DNA sequences and the origin of modern humans.* Cell, **90**, (1997), 19–30.

[224] Krishnapuram B., Carin L., Figueiredo M.A.T. and Hartemink A.J. *Sparse multinomial logistic regression: fast algorithms and generalization bounds.* IEEE Transactions on Pattern Analysis and Machine Intelligence, **27**, (2005), 957–968.

[225] Kristensen V.N., Tsalenko A., Geisler J., Faldaas A., Grenaker G.I., Lingjaerde O.C., Fjeldstad S., Yakhini Z., Lonning P.E. and Borresen–Dale A.L. *Multilocus analysis of SNP and metabolic data within a given pathway.* BMC Genomics, **7**:5, (2006).

[226] Laird N.M. and Lange C. *Family–based methods for linkage and association analysis.* Advances in Genetics, **60**, (2008), 219–252.

[227] Laird N.M. and Lange C. *Family–based designs in the age of large–scale gene–association studies.* Nature Reviews in Genetics, **7**, (2006), 385–394.

[228] Lareu M.V., Barral S., Salas A. and Carracedo A. *Sequence variation of a variable short tandem repeat at the D18S535 locus.* International Journal of Legal Medicine, **111**, (1998), 337–339.

[229] Lareu M.V., Barral S., Salas A., Pestoni C. and Carracedo A. *Sequence variation of a hypervariable short tandem repeat at the D1S1656 locus.* International Journal of Legal Medicine, **111**, (1998), 244–247.

[230] Lareu M.V., Pestoni C., Barros F., Salas A. and Carracedo A. *Sequence variation of a hypervariable short tandem repeat at the D12S391 locus.* Gene, **182**, (1996), 151–153.

[231] Lareu M.V., Pestoni C., Schurenkamp M., Rand S., Brinkmann B. and Carracedo A. *A highly variable STR at the D12S391 locus.* International Journal of Legal Medicine, **109**, (1996), 134–138.

[232] Laval G. and Excoffier L. *SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history.* Bioinformatics, **20**, (2004), 2485–2487.

[233] Lazzeroni L. and Owen A. *Plaid models for gene expression data.* Technical Report, Stanford University, (2000).

[234] Leal S.M., Yan K. and Muller-Myhsok B. *SimPed: a simulation program to generate haplotype and genotype data for pedigree structures.* Human Heredity, **60**, (2005), 119–122.

[235] Leblanc M. and Tibshirani R. *Combining estimates in regression and classification.* Journal of the American Statistical Association, **91**, (1996), 1641–1650.

[236] Lee J.S., Chu, I.S. Heo J., Calvisi D.F., Sun Z., Roskams T., Durnez A., Demetris A.J. and Thorgeirsson S.S. *Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling.* Hepatology, **40**, (2004), 667–676.

[237] Lee K.E., Sha N., Dougherty E.R., Vannucci M. and Mallick B.K. *Gene selection: a bayesian variable selection approach.* Bioinformatics, **19**, (2003), 90–97.

[238] Lee S.I., Lee H., Abbeel P. and Ng A.Y. *Efficient L1 regularized logistic regression.* Proceedings of the Twenty–first International Conference on Machine Learning (AAAI-06), (2006).

[239] Lee S.Y., Chung Y., Elston R.C., Kim Y. and Park T. *Log–linear model–based multifactor dimensionality reduction method to detect gene–gene interactions.* Bioinformatics, **23**, (2007), 2589–2595.

[240] Leng X. and Muller H.G. *Classification using functional data analysis for temporal gene expression data.* Bioinformatics, **22**, (2006), 68–76.

[241] Levedakou E.N., Freeman D.A., Budzynski M.J., Early B.E., Damaso R.C., Pollard A.M., Townley A.J., Gombos J.L., Lewis J.L., Kist F.G., Hockensmith M.E., Terwilliger M.L., Amiott E., McElfresh K.C., Schumm J.W., Ulery S.R., Konotop F., Sessa T.L., Sailus J.S., Crouse C.A., Tomsey C.S., Ban J.D. and Nelson M.S. *Characterization and validation studies of powerPlex 2.1, a nine–locus short tandem repeat (STR) multiplex system and penta D monoplex.* Journal of Forensic Science, **47**, (2002), 757–772.

[242] Li C. and Wong W.H. *Model–based analysis of oligonucleotide arrays: expression index computation and outlier detection.*, Proceedings of the National Academy of Sciences of the USA (PNAS), **98**, (2001), 31–36.

[243] Li C. and Wong W.H. *Model–based analysis of oligonucleotide arrays: model validation, design issues and standard error application.*, Genome Biology, **2**, (2001), 0032.1–0032.11.

[244] Li G.Z. and Liu T.Y. *Feature selection for bagging of support vector machines.* Lecture Notes in Artificial Intelligence, **4099**, (2006), 271–277.

[245] Liaw A. and Wiener M. *Classification and regression by randomForest.* R News, **2**, (2002), 18–22.

[246] Lin H.Y., Wang W., Liu Y.H., Soong S.J., York T.P., Myers L., Hu J.J. *Comparison of multivariate adaptive regression splines and logistic regression in detecting SNP–SNP interactions and their application in prostate cancer.* Journal of Human Genetics, **53**, (2008), 802–811.

[247] Lins A.M., Micka K.A., Sprecher C.J., Taylor J.A., Bacher J.W., Rabbach D., Schumm J.W., Bever R.A. and Creacy S.D. *Development and population study of an eight–locus short tandem repeat (STR) multiplex system.* Journal of Forensic Science, **43**, (1998), 1168–1180.

[248] Listgarten J., Damaraju S., Poulin B., Cook L., Dufour J., Driga A., Mackey J., Wishart D., Greiner R. and Zanke B. *Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms.* Clinical Cancer Research, **10**, (2004), 2725–2737.

[249] Liu Q., Yang J., Chen Z., Yang M.Q., Sung A.H. and Huang X. *Supervised learning–based tagSNP selection for genome–wide disease classifications.* BMC Genomics, **9**:S6, (2008).

[250] Liu W., Sun J., Li G., Zhu Y., Zhang S., Kim S.T., Sun J., Wiklund F., Wiley K., Isaacs S.D., Stattin P., Xu J., Duggan D., Carpten J.D., Isaacs W.B., Gronberg H., Zheng S.L. and Chang B.L. *Association of a germ–line copy number variation at 2p24.3 and risk for aggressive prostate cancer.* Cancer Research, **69**, (2009), 2176–2179.

[251] Liu Y., Athanasiadis G. and Weale M.E. *A survey of genetic simulation software for population and epidemiological studies.* Human Genomics, **3**, (2008), 79–86.

[252] Liu Z., Jiang F., Tian G., Wang S., Sato F., Meltzer S.J. and Tan M. *Sparse logistic regression with L_p penalty for biomarker identification.* Statistical Applications in Genetics and Molecular Biology, **6**, (2007), article 6.

[253] Lloyd S. *Least squares quantization in PCM.* Technical Report, Bell Laboratories, (1957).

[254] Loos R.J.F. et al *Association studies involving over 90,000 people demonstrate that common variants near to MC4R influence fat mass, weight and risk of obesity.* Nature Genetics, **40**, (2008), 768–775.

[255] Lowe C.E., Cooper J.D., Brusko T., Walker N.M., Smyth D.J., Bailey R., Bourget K., Plagnol V., Field S., Atkinson M., Clayton D.G., Wicker L.S. and Todd J.A. *Large–scale genetic fine mapping and genotype–phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes.* Nature Genetics, **39**, (2007), 1074–1082.

[256] Lu X., Shaw C.A., Patel A., Li J., Cooper M.L., Wells W.R., Sullivan C.M., Sahoo T., Yatsenko S.A., Bacino C.A., Stankiewicz P., Ou Z., Chinault A.C., Beaudet A.L., Lupski J.R., Cheung S.W. and Ward P.A. *Clinical implementation of chromosomal microarray analysis: summary of 2513 postnatal cases.* PLoS ONE, **2**, (2007), e327.

[257] Luan Y.H. and Li H.Z. *Clustering of temporal gene expression data using a mixed–effects model with B–splines.* Bioinformatics, **19**, (2003), 474–482.

[258] Lunetta K.L., Hayward L.B., Segal J. and Van Eerdewegh P. *Screening large–scale association study data: exploiting interactions using random forests.* BMC Genetics, **5**:32, (2004).

[259] Lv J. and Fan Y. *A unified approach to model selection and sparse recovery using regularized least squares.* Annals of Statistics, **37**, (2009), 3498–3528.

[260] Lygo J.E., Johnson P.E., Holdaway D.J., Woodroffe S., Whitaker J.P., Clayton T.M., Kimpton C.P. and Gill P. *The validation of short tandem repeat (STR) loci for use in forensic casework.* International Journal of Legal Medicine, **107**, (1994), 77–89.

[261] Lyng H., Badiee A., Svendsrud D.H., Hovig E., Myklebost O. and Stokke T. *Profound influence of microarray scanner characteristics on gene expression ratios: analysis and procedure for correction.*, BMC Genomics, **5**:10, (2004).

[262] Ma D.Q., Rabionet R., Konidari I., Jaworski J., Cukier H.N., Wright H.H., Abramson R.K., Gilbert J.R., Cuccaro M.L., Pericak–Vance M.A. and Martin E.R. *Association and gene-gene interaction of SLC6A4 and ITGB3 in autism.* American Journal of Medical Genetics Part B: Neuropsychiatric Genetics (in press), (2009).

[263] Madeira S.C. and Oliveira A.L. *Biclustering algorithms for biological data analysis: a survey.* IEEE Transactions on Computational Biology and Bioinformatics, **1**, (2004), 24–45.

[264] Maeda T., Hobbs R.M., Merghoub T., Guernah I., Zelent A., Cordon–Cardo C., Teruya–Feldstein J. and Pandolfi P.P. *Role of the proto–oncogene Pokemon in cellular transformation and ARF repression.* Nature, **433**, (2005), 278–285.

[265] Malovini A., Nuzzo A., Ferrazzi F., Puca A.A. and Bellazzi R. *Phenotype forecasting with SNPs data through gene–based Bayesian networks.* BMC Bioinformatics, **10**:S7, (2009).

[266] Marchini J., Donnelly P. and Cardon L.R. *Genome–wide strategies for detecting multiple loci that influence complex diseases.* Nature Genetics, **37**, (2005), 413–417.

[267] Marino M., Sala A., Bobillo C. and Corach D. *Inferring genetic substructure in the population of Argentina using fifteen microsatellite loci.* Forensic Science International Genetics, **1**, (2008), 350–352.

[268] Marino M., Sala A. and Corach D. *Genetic analysis of the populations from Northern and Mesopotamian provinces of Argentina by means of 15 autosomal STRs.* Forensic Science International, **160**, (2006), 224–230.

[269] Marino M., Sala A. and Corach D. *Genetic attributes of 15 autosomal STRs in the population of two patagonian provinces of Argentina.* Forensic Science International, **160**, (2006), 84–88.

[270] Marino M., Sala A. and Corach D. *Genetic attributes of the YHRD minimal haplotype in 10 provinces of Argentine.* Forensic Science International: Genetics, **1**, (2007), 129–133.

[271] Marino M., Sala A. and Corach D. *Population genetic analysis of 15 autosomal STRs loci in the central region of Argentina.* Forensic Science International, **161**, (2006), 72–77.

[272] Martinez F., Mansego M.L., Escudero J.C., Redon J. and Chaves F.J. *Association of a mineralocorticoid receptor gene polymorphism with hypertension in a spanish population.* American Journal of Hypertension, **22**, (2009), 649–655.

[273] McCarroll S.A. and Altshuler D.M. *Copy-number variation and association studies of human disease.* Nature Genetics, **39**, (2007), S37–S42.

[274] McCarthy M.I., Abecasis G.R., Cardon L.R., Goldstein D.B., Little J., Ioannidis J.P.A. and Hirschhorn J.N. *Genome–wide association studies for complex traits: consensus, uncertainty and challenges.* Nature Reviews in Genetics, **9**, (2008), 356–369.

[275] McCulloch W. and Pitts W. *A logical calculus of the ideas immenent in nervous activity.* Bulletin of Mathematical BioPhysics, **5**, (1943), 115–133.

[276] McKinney B.A., Reif D.M., White B.C., Crowe Jr. J.E. and Moore J.H. *Evaporative cooling feature selection for genotypic data involving interactions.* Bioinformatics, **23**, (2007), 2113–2120.

[277] Meier L., van de Geer S. and Buhlmann P. *The group lasso for logistic regression.* Journal of the Royal Statistical Society Series B, **70**, (2008), 53–71.

[278] Meinshausen N. and Buhlmann P. *High dimensional graphs and variable selection with the lasso.* Annals of Statistics, **34**, (2006), 1436–1462.

[279] Miller C.L., Murakami P., Ruczinski I., Ross R.G., Sinkus M., Sullivan B. and Leonard S. *Two complex genotypes relevant to the kynurenine pathway and melanotropin function show association with schizophrenia and bipolar disorder.* Schizophrenia Research, **113**, (2009), 259–267.

[280] Miller D.J., Zhang Y., Yu G., Liu Y., Chen L., Langefeld C.D., Herrington D. and Wang Y. *An algorithm for learning maximum entropy probability models of disease risk that efficiently searches and sparingly encodes multilocus genomic interactions.* Bioinformatics, **25**, (2009), 2478–2485.

[281] Miller M.B., Lind G.R., Li N. and Jang S.Y. *Genetic Analysis Workshop 15: simulation of a complex genetic model for rheumatoid arthritis in nuclear families including a dense SNP map with linkage disequilibrium between marker loci and trait loci.* BMC Proceedings, **1**:S4, (2007).

[282] Milne R., Ribas G., Gonzalez–Neira A., Fagerholm R., Salas A., Gonzalez E., Dopazo J., Nevanlinna H., Robledo M. and Benitez J. *ERCC4 associated with breast cancer risk: a two–stage case–control study using high throughput genotyping.* Cancer Research, **66**, (2006), 9420–9427.

[283] Milne R.L. et al *Risk of estrogen receptor–positive and –negative breast cancer and single–nucleotide polymorphism 2q35–rs13387042.* Journal of the National Cancer Institute, **101**, (2009), 1012–1018.

[284] Mirnics K., Middleton F.A., Lewis D.A. and Levitt P. *Analysis of complex brain disorders with gene expression microarrays: schizophrenia as a disease of the synapse.* Trends in Neurosciences, **24**, (2001), 479–486.

[285] Montgomery D.C. *Design and analysis of experiments.* John Wiley and Sons, New York, (2001).

[286] Moore J.H. *A global view of epistasis.* Nature Genetics, **37**, (2005), 13–14.

[287] Moore J.H. *The ubiquitous nature of epistasis in determining susceptibility to common human diseases.* Human Heredity, **56**, (2003), 73–82.

[288] Moore J.H., Lamb J.M., Brown N.J. and Vaughan D.E. *A comparison of combinatorial partitioning and linear regression for the detection of epistatic effects of the ACE I/D and PAI–1 4G/5G polymorphisms on plasma PAI–1 levels.* Clinical Genetics, **62**, (2002), 74–79.

[289] Moore J.H. and Williams S.W. *New strategies for identifying gene–gene interactions in hypertension.* Annals of Medicine, **34**, (2002), 88–95.

[290] Morling N., Allen R.W., Carracedo A., Geada H., Guidet F., Hallenberg C., Martin W., Mayr W.R., Olaisen B., Pascali V.L. and Schneider P.M. *Paternity Testing Commission of the International Society of Forensic Genetics: recommendations on genetic investigations in paternity cases.* Forensic Science International, **129**, (2002), 148–157.

[291] Mosibay A., Mozer T.J. and Sprecher C. *Promega corporation reveals primer sequences in its testing kits.* Journal of Forensic Science, **45**, (2000), 1360–1362.

[292] Mosquera–Miguel A., Alvarez–Iglesias V., Vega A., Milne R., Cabrera de Leon A., Benitez J., Carracedo A and Salas A. *Is mitochondrial DNA variation associated with sporadic breast cancer risk?* Cancer Research, **68**, (2007), 623–625.

[293] Motsinger A.A. and Ritchie M.D. *The effect of reduction in cross–validation intervals on the performance of multifactor dimensionality reduction.* Genetic Epidemiology, **30**, (2006), 546–555.

[294] Motsinger–Reif A.A. and Ritchie M.D. *Neural networks for genetic epidemiology: past, present, and future.* BioData Mining, **1**:3, (2008).

[295] Muller H.G. *Functional modelling and classification of longitudinal data.* Scandinavian Journal of Statistics, **32**, (2005), 223–240.

[296] Nachman M.W. and Crowell S.L. *Estimate of the mutation rate per nucleotide in humans.* Genetics, **156**, (2000), 297–304.

[297] Namkung J., Kim K., Yi S., Chung W., Kwon M.S. and Park T. *New evaluation measures for multifactor dimensionality reduction classifiers in gene–gene interaction analysis.* Bioinformatics, **25**, (2009), 338–345.

[298] Nelson M.R., Kardia S.L.R., Ferrell R.E. and Sing C.F. *A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation.* Genome Research, **11**, (2001), 458–470.

[299] Nguyen D.V. and Rocke D.M. *Tumor classification by partial least squares using microarray gene expression data.* Bioinformatics, **18**, (2002), 39–50.

[300] Nicodemus K.K., Wang W. and Shugart Y.Y. *Stability of variable importance scores and rankings using statistical learning tools on single–nucleotide polymorphisms and risk factors involved in gene x gene and gene x environment interactions.* BMC Proceedings, **1**:S58, (2007).

[301] Nothnagel M. *Simulation of LD block–structured SNP haplotype data and its use for the analysis of case–control data by supervised learning methods.* American Journal of Human Genetics, **71**:A2363, (2002).

[302] Novo Villaverde F.J. *Genética humana : conceptos, mecanismos y aplicaciones de la genética en el campo de la biomedicina: texto multimedia.* Pearson Prentice Hall, New Jersey, (2006).

[303] Onay V.U., Briollais L., Knight J.A., Shi E., Wang Y., Wells S., Li H., Rajendram I., Andrulis I.L. and Ozcelik H. *SNP–SNP interactions in breast cancer susceptibility.* BMC Cancer, **6**:114, (2006).

[304] Orozco G., Hinks A., Eyre S., Ke X., Gibbons L.J., Bowes J., Flynn E., Martin P., Wellcome Trust Case Control Consortium, YEAR consortium, Wilson A.G., Bax D.E., Morgan A.W., Emery P., Steer S., Hocking L., Reid D.M., Wordsworth P., Harrison P., Thomson W., Barton A. and Worthington J. *Combined effects of three independent SNPs greatly increase the risk estimate for RA at 6q23.* Human Molecular Genetics, **18**, (2009), 2693–2699.

[305] Ott J. and Hoh J. *Set association analysis of SNP case–control and microarray data.* Journal of Computational Biology, **10**, (2003), 569–574.

[306] Park M.Y. and Hastie T. *An L_1 regularization–path algorithm for generalized linear models.* Manuscript, Department of Statistics, Stanford University, (2006).

[307] Patterson N., Price A.L. and Reich D. *Population structure and eigenanalysis.* Plos Genetics, **2**, (2006), 2074–2093.

[308] Pearson K. *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to hove arisen from random sampling.* Philosophical Magazine, **50**, (1900), 157–175.

[309] Pease A.C., Solas D., Sullivan E.J., Cronin M.T., Holmes C.P. and Fodor S.P. *Light–generated oligonucleotide arrays for rapid DNA sequence analysis.*, Proceedings of the National Academy of Sciences of the USA, **91**, (1994), 5022–5026.

[310] Perry G.H., Ben-Dor A., Tsalenko A., Sampas N., Rodriguez-Revenga L., Tran C.W., Scheffer A., Steinfeld I., Tsang P., Yamada N.A., Park H.S., Kim J.I., Seo J.S., Yakhini Z., Laderman S., Bruhn L. and Lee C. *The fine-scale and complex architecture of human copy-number variation.* The American Journal of Human Genetics, **82**, (2008), 685–695.

[311] Pfaff C.L., Barnholtz–Sloan J., Wagner J.K. and Long J.C. *Information on ancestry from genetic markers.* Genetic Epidemiology, **26**, (2004), 305–315.

[312] Pharoah P.D., Tyrer J., Dunning A.M., Easton D.F. and Ponder B.A. *Association between common variation in 120 candidate genes and breast cancer risk.* PLoS Genetics, **3**, (2007), e42.

[313] Phillips C., Fang R., Ballard D., Fondevila M., Harrison C., Hyland F., Musgrave–Brown E., Proff C., Ramos–Luis E., Sobrino B., Furtado M., Syndercombe–Court D., Carracedo A., Schneider P.M. and The SNPforID Consortium. *Evaluation of the Genplex SNP typing system and a 49plex forensic marker panel.* Forensic Science International: Genetics, **1**, (2007), 180–185.

[314] Phillips C., Fondevila M., Garcia–Magariños M., Rodriguez A., Salas A., Carracedo A. and Lareu M.V. *Resolving relationship tests that show ambiguous STR results using autosomal SNPs as supplementary markers.* Forensic Science International: Genetics, **2**, (2008), 198–204.

[315] Phillips C., Salas A., Sanchez J.J., Fondevila M., Gomez–Tato A., Alvarez–Dios J., Calaza M., Casares de Cal M., Ballard D., Lareu M.V. and Carracedo A. *Inferring ancestral origin using a single multiplex assay of ancestry–informative marker SNPs.* Forensic Science International: Genetics, **1**, (2007), 273–280.

[316] Phillips P.C. *Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems.* Nature Reviews in Genetics, **9**, (2008), 855–867.

[317] Phillips P.C. *The language of gene interaction.* Genetics, **149**, (1998), 1167–1171.

[318] Pickrell J., Clerget–Darpoux F. and Bourgain C. *Power of genome–wide association studies in the presence of interacting loci.* Genetic Epidemiology, **31**, (2007), 748–762.

[319] Pirooznia M. and Deng Y. *SVM classifier – a comprehensive java interface for support vector machine classification of microarray data.* BMC Bioinformatics, **7**:S25, (2006).

[320] Plagnol V., Cooper J.D., Todd J.A. and Clayton D.G. *A method to address differential bias in genotyping in large–scale association studies.* Plos Genetics, **3**, (2007), 759–767.

[321] Pociot F., Karlsen A.E., Pedersen C.B., Aalund M. and Nerup J. *Novel analytical methods applied to type 1 diabetes genome–scan data.* American Journal of Human Genetics, **74**, (2004), 647–660.

[322] Prelic A., Bleuler S., Zimmermann P., Wille A., Buhlmann P., Gruissem W., Hennig L., Thiele L. and Zitzler E. *A systematic comparison and evaluation of biclustering methods for gene expression data.* Bioinformatics, **22**, (2006), 1122–1129.

[323] Price A.L., Patterson N.J., Plenge R.M., Weinblatt M.E., Shadick N.A. and Reich D. *Principal components analysis corrects for stratification in genome–wide association studies.* Nature Genetics, **38**, (2006), 904–909.

[324] Price R.A., Li W.D. and Zhao H. *FTO gene SNPs associated with extreme obesity in cases, controls and extremely discordant sister pairs.* BMC Medical Genetics, **9**:4, (2008).

[325] Priti H., Rong Q., Kristie A., Cheryl G., Sonia D., Renee G., Julie G., Erik S., Norman L., and John Q. *A concise guide to cDNA microarray analysis.* Biotechniques, **29**, (2000), 548–562.

[326] Purcell S., Neale B., Todd–Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., de Bakker P.I., Daly M.J. and Sham P.C. *PLINK: a tool set for whole–genome association and population–based linkage analysis.* American Journal of Human Genetics, **81**, (2007), 559–575.

[327] Quackenbush J. *Computational analysis of microarray data.* Nature Reviews in Genetics, **2**, (2001), 418–427.

[328] Quackenbush J. *Microarray data normalization and transformation.* Nature Genetics, **32**, (2002), 496–501.

[329] Ramsay J.O. and Silverman B.W. *Functional data analysis.* Springer, New York, (2005).

[330] Rao D.C., Keats B.J.B., Morton N.E., Yee S. and Lew R. *Variability of human linkage data.* American Journal of Human Genetics, **30**, (1978), 516–529.

[331] Redon R. et al *Global variation in copy number in the human genome.* Nature, **444**, (2006), 444–454.

[332] Rees J.L. *The genetics of sun sensitivity in humans.* American Journal of Human Genetics, **75**, (2004), 739–751.

[333] Ripley B.D. *Pattern recognition and neural networks.* Cambridge University Press, Cambridge, (1996).

[334] Risch N. and Merikangas K. *The future of genetic studies of complex human diseases.* Science, **273**, (1996), 1516–1517.

[335] Risch N.J. *Searching for genetic determinants in the new millennium.* Nature, **405**, (2000), 847–856.

[336] Ritchie M.D., Hahn L.W. and Moore J.H. *Power of multifactor dimensionality reduction for detecting gene–gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity.* Genetic Epidemiology, **24**, (2003), 150–157.

[337] Ritchie M.D., Hahn L.W., Roodi N., Bailey L.R., Dupont W.D., Parl F.F. and Moore J.H. *Multifactor–dimensionality reduction reveals high–order interactions among estrogen–metabolism genes in sporadic breast cancer.* American Journal of Human Genetics, **69**, (2001), 138–147.

[338] Roberts A., McMillan L., Wang W., Parker J., Rusyn I. and Threadgill D. *Inferring missing genotypes in large SNP panels using fast nearest–neighbor searches over sliding windows.* Bioinformatics, **23**, (2007), i401–i407.

[339] Rodríguez-Santiago B., Brunet A., Sobrino B., Serra-Juhé C., Flores R., Armengol L., Vilella E., Gabau E., Guitart M., Guillamat R., Martorell L., Valero J., Gutiérrez-Zotes A., Labad A., Carracedo A., Estivill X. and Pérez-Jurado L.A. *Association of common copy number variants at the glutathione S-transferase genes and rare novel genomic changes with schizophrenia.* Molecular Psychiatry (in press), (2009).

[340] Romano J.P. and Wolf M. *Stepwise multiple testing as formalized data snooping.* Econometrica, **73**, (2005), 1237–1282.

[341] Rosenblatt F. *Principles of neuroDynamics: perceptrons and the theory of brain mechanisms.* Spartan, Washington DC, (1962).

[342] Rosenblatt F. *The perceptron: a probabilistic model for information storage and organization in the brain.* Psychological Review, **65**, (1958), 386–408.

[343] Ross D.T., Scherf U., Eisen M.B., Perou C.M., Spellman P., Iyer V., Jeffrey S.S., Van de Rijn M., Waltham M., Pergamenschikov A., Lee J.C.F., Lashkari D., Shalon D., Myers T.G., Weinstein J.N., Botstein D. and Brown P.O. *Systematic variation in gene expression patterns in human cancer cell lines.* Nature Genetics, **24**, (2000), 227–234.

[344] Roth V. *The generalized lasso.* IEEE Transactions on Neural Networks, **15**, (2004), 16–28.

[345] Ruczinski I., Kooperberg C. and LeBlanc M. *Logic regression.* Journal of Computational and Graphical Statistics, **12**, (2003), 475–511.

[346] Saeys Y., Inza I. and Larrañaga P. *A review of feature selection techniques in bioinformatics.* Bioinformatics, **23**, (2007), 2507–2517.

[347] Saigo H., Uno T. and Tsuda K. *Mining complex genotypic features for predicting HIV–1 drug resistance.* Bioinformatics, **23**, (2007), 2455–2462.

[348] Salas A., Bandelt H.J., Macaulay V. and Richards M.B. *Phylogeographic investigations: the role of trees in forensic genetics.* Forensic Science International, **168**, (2007), 1–13.

[349] Salas A., Jaime J.C., Alvarez–Iglesias V. and Carracedo A. *Gender bias in the multi–ethnic genetic composition of Central Argentina.* Journal of Human Genetics, **53**, (2008), 662–674.

[350] Salas A., Vega A., Milne R.L., Garcia–Magariños M., Ruibal A., Benitez J. and Carracedo A. *The "Pokemon" (ZBTB7) gene: no evidence of association with sporadic breast cancer.* Clinical Medicine: Oncology, **2**, (2008), 357–362.

[351] Salas A., Vega A., Phillips C., Torres M., Quintela I. and Carracedo A. *ZBTB7 HapMap in a worldwide population study.* Breast Cancer Research, **7**, (2005), S26.

[352] Salas A., Vega A., Torres M., Quintela I., Phillips C., Rodriguez–Lopez R., Rivas G., Benitez J. and Carracedo A. *High–density screening of the ZBTB7 gene in breast cancer patients.* Breast Cancer Research, **7**, (2005), S25.

[353] Sanchez J.J., Borsting C., Hallenberg C., Buchard A., Hernandez A. and Morling N. *Multiplex PCR and minisequencing of SNPs–a model with 35 Y–chromosome SNPs.* Forensic Science International, **137**, (2003), 74–84.

[354] Sanchez J.J., Phillips C., Borsting C., Balogh K., Bogus M., Fondevila M., Harrison C.D., Musgrave–Brown E., Salas A., Syndercombe–Court D., Schneider P.M., Carracedo A. and Morling N. *A multiplex assay with 52 single nucleotide polymorphisms for human identification.* Electrophoresis, **27**, (2006), 1713–1724.

[355] Sanchez M.T., Cao R., Fernandez J., Garcia–Magariños M., Garcia–Torre F., Gesto J.M., Gomez A., Gonzalez–Manteiga W. and Gutierrez J.M. *imath.cesga.es, the VO for the european mathematicians.* Proceedings of the 2nd Iberian Grid Infraestructure Conference, Oporto, (2008).

[356] Sasieni P. *From genotypes to genes: doubling the sample size.* Biometrics, **53**, (1997), 1253–1261.

[357] Schadt E.E., Li C., Ellis B. and Wong W.H. *Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data.* Journal of Cellular Biochemistry, **37**, (2001), Suppl 120–125.

[358] Schadt E.E., Li C., Su C. and Wong W.H. *Analyzing high-density oligonucleotide gene expression array data.* Journal of Cellular Biochemistry, **80**, (2000), 192–202.

[359] Schena M., Shalon D., Davis R.W. and Brown P.O. *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, **270**, (1995), 467–470.

[360] Schena M., Shalon D., Heller R., Chai A., Brown P.O. and Davis R.W. *Parallel human genome analysis: microarray–based expression monitoring of 1000 genes.* Proceedings of the National Academy of Sciences USA, **93**, (1996), 10614–10619.

[361] Scherer S.W., Lee C., Birney E., Altshuler D.M., Eichler E.E., Carter N.P., Hurles M.E. and Feuk L. *Challenges and standards in integrating surveys of structural variation.* Nature Genetics, **39**, (2007), S7–S15.

[362] Schliep A., Schonhuth A. and Steinhoff C. *Using hidden Markov models to analyze gene expression time course data.* Bioinformatics, **19**, (2003), I264–I272.

[363] Schmidt M., Fung G. and Rosales R. *Fast optimization methods for L_1 regularization: a comparative study and two new approaches.* European Conference on Machine Learning (ECML), (2007).

[364] Schmidt M., Hauser E.R., Martin E.R. and Schmidt S. *Extension of the SIMLA package for generating pedigrees with complex inheritance patterns: environmental covariates, gene–gene and gene–environment interaction.* Statistical Applications in Genetics and Molecular Biology, **4**, (2004), Article15.

[365] Schneider H.R., Rand S., Schmitter H. and Weichhold G. *ACTBP2–nomenclature recommendations of GEDNAP.* International Journal of Legal Medicine, **111**, (1998), 97–100.

[366] Scholkopf B., Simard P., Smola A. and Vapnik V. *Prior knowledge in support vector kernels.* In Advances in Neural Information Processing Systems, The MIT Press, Cambridge, (1998).

[367] Scholkopf B. and Smola A.J. *Learning with kernels: support vector machines, regularization, optimization and beyond.* The MIT Press, Cambridge, (2002).

[368] Schork N.J., Cardon L.R. and Xu X. *The future of genetic epidemiology.* Trends in Genetics, **14**, (1998), 266–272.

[369] Schwender H. and Ickstadt K. *Identification of SNP interactions using logic regression.* Biostatistics, **9**, (2008), 187–198.

[370] Schwender H., Ickstadt K. and Rahnenfuhrer J. *Classification with high–dimensional genetic data: assigning patients and genetic features to known classes.* Biometrical Journal, **50**, (2008), 911–926.

[371] Schwender H., Zucknick M., Ickstadt K., Bolt H.M. and The GENICA network. *A pilot study on the application of statistical classification procedures to molecular epidemiological data.* Toxicology Letters, **151**, (2004), 291–299.

[372] Sebat J., Lakshmi B., Troge J., Alexander J., Young J., Lundin P., Maner S., Massa H., Walker M., Chi M., Navin N., Lucito R., Healy J., Hicks J., Ye K., Reiner A., Gilliam T.C., Trask B., Patterson N., Zetterberg A. and Wigler M. *Large–scale copy number polymorphism in the human genome.* Science, **305**, (2004), 525–528.

[373] Segal E., Battle A. and Koller D. *Decomposing gene expression into cellular processes.* Proceedings of the Pacific Symposium in Biocomputing, **8**, (2003), 89–100.

[374] Segal E., Taskar B., Gasch A., Friedman N. and Koller D. *Rich probabilistic models for gene expression* Bioinformatics, **17**, (2001), S243–S252.

[375] Sha Q., Zhu X., Zuo Y., Cooper R. and Zhang S. *A combinatorial searching method for detecting a set of interacting loci associated with complex traits.* Annals of Human Genetics, **70**, (2006), 677–692.

[376] Shen R., Ghosh D. and Chinnaiyan A.M. *Prognostic meta–signature of breast cancer developed by two–stage mixture modeling of microarray data.* BMC Genomics, **5**:94, (2004).

[377] Sheng Q., Moreau Y. and De Moor B. *Biclustering microarray data by Gibbs sampling.* Bioinformatics, **19**, (2003), ii196–ii205.

[378] Shevade S. and Keerthi S. *A simple and efficient algorithm for gene selection using sparse logistic regression.* Bioinformatics, **19**, (2003), 2246–2253.

[379] Shriver M.D., Kennedy G.C., Parra E.J., Lawson H.A., Sonpar V., Huang J., Akey J. and Jones K.W. *The genomic distribution of population substructure in four populations using 8525 autosomal SNPs.* Human Genomics, **1**, (2004), 274–286.

[380] Singh D., Febbo P., Ross K., Jackson D., Manola J., Ladd C., Tamayo P., Renshaw A., D'Amico A., Richie J., Lander E., Loda M., Kantoff P., Golub T. and Sellers W. *Gene expression correlates of clinical prostate cancer behavior.* Cancer Cell, **1**, (2002) 203–209.

[381] Skol A.D., Scott L.J., Abecasis G.R. and Boehnke M. *Joint analysis is more efficient than replication–based analysis for two–stage genome–wide association studies.* Nature Genetics, **38**, (2006), 209–213.

[382] Smyth G.K. and Speed T. *Normalization of cDNA microarray data.* Methods, **31**, (2003), 265–273.

[383] Spellman P.T., Sherlock G., Zhang M.Q., Iyer V.R., Anders K., Eisen M.B., Brown P.O., Botstein D. and Futcher B. *Comprehensive identification of cell cycle–regulated gene of the yeast Saccharomyces cerevisiae by microarray hybridization.* Molecular Biology of the Cell, **9**, (1998), 3273–3297.

[384] Spielman R.S., McGinnis R.E. and Ewens W.J. *Transmission test for linkage disequilibrium: the insulin gene region and insulin–dependent diabetes mellitus (IDDM).* American Journal of Human Genetics, **52**, (1993), 506–516.

[385] Statnikov A., Aliferis C.F., Tsamardinos I., Hardin D. and Levy S. *A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis.* Bioinformatics, **21**, (2005), 631–643.

[386] Sulem P., Gudbjartsson D.F., Stacey S.N., Helgason A., Rafnar T., Magnusson K.P., Manolescu A., Karason A., Palsson A., Thorleifsson G., Jakobsdottir M., Steinberg S., Palsson S., Jonasson F., Sigurgeirsson B., Thorisdottir K., Ragnarsson R., Benediktsdottir K.R., Aben K.K., Kiemeney L.A., Olafsson J.H., Gulcher J., Kong A., Thorsteinsdottir U. and Stefansson K. *Genetic determinants of hair, eye and skin pigmentation in Europeans.* Nature Genetics, **39**, (2007), 1443–1452.

[387] Tarigan B. and van de Geer S. *Adaptivity of support vector machines with l_1 penalty.* Bernoulli, **12**, (2004), 1045–1076.

[388] Terwilliger J.D. and Weiss K.M. *Linkage disequilibrium mapping of complex disease: fantasy or reality?* Current Opinion in Biotechnology, **9**, (1998), 578–594.

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

[390] The Wellcome Trust Case Control Consortium. *Genome–wide association study of 14000 cases of seven common diseases and 3000 shared controls.* Nature, **447**, (2007), 661–678.

[391] Therneau T.M. and Atkinson E.J. *An introduction to recursive partitioning using the RPART routines.* Technical Report, Mayo Foundation, (1997).

[392] Thomas J.G., Olson J.M. and Tapscott S.J. *An efficient and robust statistical modeling approach to discover diferentially expressed genes using genomic expression profiles.* Genome Research, **11**, (2001), 1227–1236.

[393] Thompson D. and Easton D. *The genetic epidemiology of breast cancer genes.* Journal of Mammary Gland Biology and Neoplasia, **9**, (2004), 221–236.

[394] Thornton-Wells T.A., Moore J.H. and Haines J.L. *Genetics, statistics and human disease: analytical retooling for complexity.*, Trends in Genetics, **20**, (2004), 640–647.

[395] Tibshirani R. *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society Series B, **58**, (1996), 267–288.

[396] Tibshirani R., Hastie T., Eisen M., Ross D., Botstein D. and Brown P. *Clustering methods for the analysis of DNA microarray data* Technical

report, Dept. of Health Research and Policy, Dept. of Genetics, and Dept. of Biochemestry, Stanford University, (1999).

[397] Toscanini U., Berardi G., Amorim A., Carracedo A., Salas A., Gusmao L. and Raimondi E. *Forensic considerations on STR databases in Argentina.* International Congress Series 1288, Elsevier, Amsterdam, (2006), 337–339.

[398] Toscanini U., Gusmao L., Berardi G., Amorim A., Carracedo A., Salas A. and Raimondi E. *Testing for genetic structure in different urban Argentinian populations.* Forensic Science International, **165**, (2007), 35–40.

[399] Toscanini U., Gusmao L., Berardi G., Amorim A., Carracedo A., Salas A. and Raimondi E. *Y chromosome microsatellite genetic variation in two Native American populations from Argentina: population stratification and mutation data.* Forensic Science International: Genetics, **2**, (2008), 274–280.

[400] Toscanini U., Salas A., Carracedo A., Berardi G., Amorim A., Gusmao L. and Raimondi E. *A simulation–based approach to evaluate population stratification in Argentina.* Forensic Science International: Genetics Suplement Series, **1**, (2008), 662–663.

[401] Toscanini U., Salas A., Garcia–Magariños M., Gusmao L. and Raimondi E. *Population stratification in Argentina strongly influences likelihood ratio estimates in paternity testing as revealed by a simulation–based approach.* International Journal of Legal Medicine (in press), (2009).

[402] Trikalinos T.A., Salanti G., Khoury M.J. and Ioannidis J.P.A. *Impact of violations and deviations in Hardy-Weinberg equilibrium on postulated gene-disease associations.* American Journal of Epidemiology, **163**, (2006), 300–309.

[403] Truett J., Cornfield J. and Kannel W. *A multivariate analysis of the risk of coronary heart disease in Framingham.* Journal of Chronic Diseases, **20**, (1967), 511–524.

[404] Tseng G.C., Oh M.K., Rohlin L., Liao J.C. and Wong W.H. *Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects.*, Nucleic Acids Research, **29**, (2001), 2549–2557.

[405] Underhill P.A., Shen P., Lin A.A., Jin L., Passarino G., Yang W.H., Kauffman E., Bonné-Tamir B., Bertranpetit J., Francalacci P., Ibrahim M., Jenkins T., Kidd J.R., Mehdi S.Q., Seielstad M.T., Wells R.S., Piazza A., Davis R.W., Feldman M.W., Cavalli-Sforza L.L. and Oefner P.J. *Y chromosome sequence variation and the history of human populations.* Nature, **26**, (2000), 358–361.

[406] Urquhart A., Kimpton C.P., Downes T.J. and Gill P. *Variation in short tandem repeat sequences.* International Journal of Legal Medicine, **107**, (1994), 13–20.

[407] Valentin J. *Positive evidence of paternity calculated according to Essen–Moller: the Bayesian approach.* In Inclusion probabilities in parentage testing, Arlington, Virginia, (1983).

[408] van 't Veer L.J., Dai H., van de Vijver M.J., He Y.D., Hart A.A., Mao M., Peterse H.L., van der Kooy K., Marton M.J., Witteveen A.T., Schreiber G.J., Kerkhoven R.M., Roberts C., Linsley P.S., Bernards R. and Friend S.H. *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, **415**, (2002), 530–536.

[409] Vapnik V.N. *The nature of statistical learning theory.* Springer–Verlag, New York, (1996).

[410] Vega A., Salas A., Milne R.L., Carracedo B., Ribas G., Ruibal A., de Leon A.C., Gonzalez–Hernandez A., Benitez J. and Carracedo A. *Evaluating new candidate SNPs as low penetrance risk factors in sporadic breast cancer: a two–stage Spanish case–control study.* Gynecologic Oncology, **112**, (2009), 210–214.

[411] Vega A., Salas A., Phillips C., Sobrino B., Carracedo B., Ruiz–Ponte C., Rodriguez–Lopez R., Rivas G., Benitez J. and Carracedo A. *Large–scale single nucleotide polymorphism analysis of candidates for low–penetrance breast cancer genes.* Breast Cancer Research, **7**:S22, (2005).

[412] Venter J.C. et al *The sequence of the human genome.* Science, **291**, (2001), 1304–1351.

[413] Vrijenhoek T., Buizer–Voskamp J.E., van der Stelt I., Strengman E., Genetic Risk and Outcome in Psychosis (GROUP) Consortium, Sabatti C., van Kessel A.G., Brunner H.G., Ophoff R.A. and Veltman J.A. *Recurrent CNVs disrupt three candidate genes in schizophrenia patients.* American Journal of Human Genetics, **83**, (2008), 504–510.

[414] Waldman I.D. and Gizer I.R. *The genetics of attention deficit hyperactivity disorder.* Clinical Psychology Review, **26**, (2006), 396–432.

[415] Wallin J.M., Holt C.L., Lazaruk K.D., Nguyen T.H. and Walsh P.S. *Constructing universal multiplex PCR systems for comparative genotyping.* Journal of Forensic Science, **47**, (2002), 52–65.

[416] Wang H. and Leng C. *Unified lasso estimation via least squares approximation.* Journal of the American Statistical Association, **102**, (2007), 1039–1048.

[417] Wang L.Y., Comaniciu D. and Fasulo D. *Exploiting interactions among polymorphisms contributing to complex disease traits with boosted generative modeling.* Journal of Computational Biology, **13**, (2006), 1673–1684.

[418] Wang N., Akey J.M., Zhang K., Chakraborty R. and Jin L. *Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation.* American Journal of Human Genetics, **71**, (2002), 1227–1234.

[419] Wang S. and Zhu J. *Variable selection for model–based high–dimensional clustering and its application to microarray data.* Biometrics, **64**, (2008), 440–448.

[420] Wang S.S., Purdue M.P., Cerhan J.R., Zheng T., Menashe I., Armstrong B.K., Lan Q., Hartge P., Kricker A., Zhang Y., Morton L.M., Vajdic C.M., Holford T.R., Severson R.K., Grulich A., Leaderer B.P., Davis S., Cozen W., Yeager M., Chanock S.J., Chatterjee N. and Rothman N. *Common gene variants in the tumor necrosis factor (TNF) and TNF receptor superfamilies and NF–kB transcription factors and non–Hodgkin lymphoma risk.* Plos One, **4**, (2009), e5360.

[421] Wang X., Ghosh S. and Guo S.W. *Quantitative quality control in microarray image processing and data acquisition.* Nucleic Acids Research, **29**, (2001), E75–5.

[422] Ward J.H. *Hierarchical grouping to optimize an objective function.* Journal of the American Statistical Association, **58**, (1963), 236–244.

[423] Watson J.D. and Crick F.H. *Genetic implications of the structure of DNA.* Nature, **171**, (1953), 964–967.

[424] Watson J.D. and Crick F.H. *Molecular structure of nucleic acids: A structure for DNA.* Nature, **171**, (1953), 737–738.

[425] Wedren S., Lovmar L., Humphreys K., Magnusson C., Melhus H., Syvanen A.C., Kindmark A., Landegren U., Fermer M.L., Stiger F., Persson I., Baron J. and Weiderpass E. *Oestrogen receptor alpha gene haplotype and postmenopausal breast cancer risk: a case control study.* Breast Cancer Research, **6**, (2004), 437–449.

[426] Weinberg W. *Über den nachweis der vererbung beim menschen.* Jahreshefte des Vereins für Vaterlandische Naturkunde in Württemberg, **64**, (1908), 368–382.

[427] Weiss K.M. and Terwilliger J.D. *How many diseases does it take to map a gene with SNPs?* Nature Genetics, **26**, (2000), 151–157.

[428] Weiss L.A., Shen Y., Korn J.M., Arking D.E., Miller D.T., Fossdal R., Saemundsen E., Stefansson H., Ferreira M.A., Green T., Platt O.S., Ruderfer D.M., Walsh C.A., Altshuler D.M., Chakravarti A., Tanzi R.E., Stefansson K., Santangelo S.L., Gusella J.F., Sklar P., Wu B.L., Daly M.J. and the Autism Consortium. *Association between microdeletion and microduplication at 16p11.2 and autism.* New England Journal of Medicine, **358**, (2008), 737–739.

[429] West M., Blanchette C., Dressman H., Huang E., Ishida S., Spang R., Zuzan H., Olson J., Marks J. and Nevins J. *Predicting the clinical status of human breast cancer by using gene expression profiles.* Proceedings of the National Academy of Science (PNAS), **98**, (2001), 11462–11467.

[430] Weston J., Elisseeff A., Scholkopf B. and Tipping M. *Use of the zero–norm with linear models and kernel methods.* Journal of Machine Learning Research, **3**, (2003), 1439–1461.

[431] Wiegand P., Budowle B., Rand S. and Brinkmann B. *Forensic validation of the STR systems SE33 and TC11.* International Journal of Legal Medicine, **105**, (1993), 315–320.

[432] Wille A., Hoh J. and Ott J. *Sum statistics for the joint detection of multiple disease loci in case–control association studies with SNP markers.* Genetic Epidemiology, **25**, (2003), 350–359.

[433] Wolpert D. *Stacked generalization.* Neural Networks, **5**, (1992), 241–259.

[434] Workman C., Jensen L.J., Jarmer H., Berka R., Gautier L., Nielsen H.B., Saxild H.H., Nielsen C., Brunak S. and Knudsen S. *A new non-linear normalization method for reducing variability in DNA microarray experiments.* Genome Biology, **3**, (2002), 0048.1–0048.16.

[435] Wright S.J., Nowak R.D. and Figueiredo M.A.T. *Sparse reconstruction by separable approximation.* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2008).

[436] Wu X., Gu J., Grossman H.B., Amos C.I., Etzel C., Huang M., Zhang Q., Millikan R.E., Lerner S., Dinney C.P. and Spitz M.R. *Bladder cancer predisposition: a multigenic approach to DNA–repair and cell–cycle–control genes.* The American Journal of Human Genetics, **78**, (2006), 464–479.

[437] Yang I.V., Chen E., Hasseman J.P., Liang W., Frank B.C., Wang S., Sharov V., Saeed A.I., White J., Li J., Lee N.H., Yeatman T.J. and Quackenbush J. *Within the fold: assessing differential expression measures and reproducibility in microarray assays.* Genome Biology, **3**, (2002), 0062.1–0062.12.

[438] Yang Y.H., Dudoit S., Luu P., Lin D.M., Peng V., Ngai J. and Speed T.P. *Normalization for cDNA microarray data: a robust composite method addresing single and multiple slide systematic variation.* Nucleic Acids Research, **30**, (2002), e15.

[439] Yeoh E.J., Ross M.E., Shurtleff S.A., Williams W.K., Patel D., Mahfouz R., Behm F.G., Raimondi S.C., Relling M.V., Patel A., Cheng C., Campana D., Wilkins D., Zhou X., Li J., Liu H., Pui C.H., Evans W.E., Naeve C., Wong L. and Downing J.R. *Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling.* Cancer Cell, **1**, (2002), 133–143.

[440] Yu R. and Shete S. *Analysis of alcoholism data using support vector machines.* BMC Genetics, **6**:S136, (2005).

[441] Yu Z. and Schaid D.J. *Methods to impute missing genotypes for population data.* Human Genetics, **122**, (2007), 495–504.

[442] Yuan M. and Lin Y. *Model selection and estimation in regression with grouped variables.* Journal of the Royal Statistical Society Series B, **68**, (2006), 49–67.

[443] Zeggini E., Scott L.J., Saxena R., Voight B. and DIAGRAM Consortium. *Meta–analysis of genome–wide association data and large–scale replication identifies several additional susceptibility loci for type 2 diabetes.* Nature Genetics, **30**, (2008), 638–645.

[444] Zhang C.H. and Huang J. *The sparsity and bias of the lasso selection in high–dimensional linear regression.* Annals of Statistics, **36**, (2008), 1567–1594.

[445] Zhang H. and Bonney G. *Use of classification trees for association studies.* Genetic Epidemiology, **19**, (2000), 323–332.

[446] Zhang T. *Some sharp performance bounds for least squares regression with L1 regularization.* Annals of Statistics, **37**, (2009), 2109–2144.

[447] Zhang T. and Oles F. *Text categorization based on regularized linear classifiers.* Information Retrieval, **4**, (2001), 5–31.

[448] Zhang Y. and Liu J.S. *Bayesian inference of epistatic interactions in case–control studies.* Nature Genetics, **39**, (2007), 1167–1173.

[449] Zhou W., Liu G., Miller D.P., Thurston S.W., Xu L.L., Wain J.C., Lynch T.J., Su L. and Christiani D.C. *Gene–environment interaction for the ERCC2 polymorphisms and cumulative cigarette smoking exposure in lung cancer.* Cancer Research, **62**, (2002), 1377–1381.

[450] Zhu Y., Wang Z., Miller D.J., Clarke R., Xuan J., Hoffman E.P. and Wang Y. *A ground truth based comparative study on clustering of gene expression data.* Frontiers in Bioscience, **13**, (2008), 3839–3849.

[451] Zou H. *The adaptive lasso and its oracle properties.* Journal of the American Statistical Association, **101**, (2006), 267–288.

[452] Zou H. and Hastie T. *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society Series B, **67**, (2005), 301–320.

[453] Zou H. and Li R. *One–step sparse estimates in nonconcave penalized likelihood models.* Annals of Statistics, **36**, (2008), 1509–1533.