UNIVERSITY OF LIÈGE

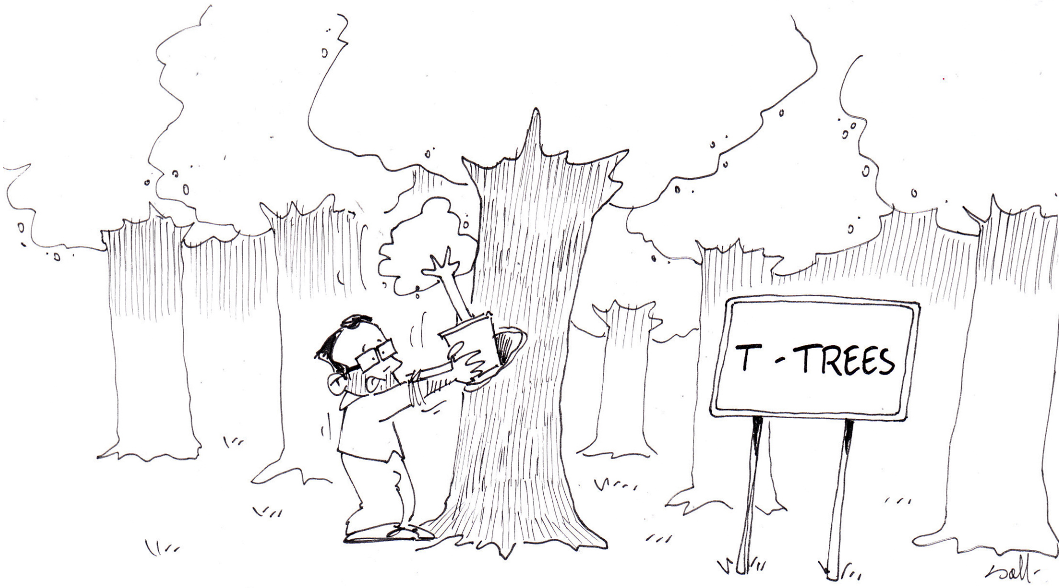
Faculty of Applied Sciences

Department of Electrical Engineering and Computer Science

PhD dissertation

A walk into random forests

Adaptation and application to Genome-Wide Association Studies



Author
Vincent BOTTA

Supervisor
Prof. Louis WEHENKEL

September, 2013

Summary

Understanding underlying mechanisms of common diseases is one of the major goals of current research in medicine. As most of these disorders are linked to genetic factors, identification of the associated variants forms an excellent strategy towards the elucidation of molecular and cellular dysfunctions, and *in fine* could lead to better personalised diagnostics and treatments.

Genome-Wide Association Studies (GWAS) aim to discover variants spread over the genome that could lead, in isolation or in combination, to a particular trait or an unfortunate phenotype such as a disease. The basic idea behind these studies is to statistically analyse the genetic differences between groups of healthy (controls) and diseased (cases) individuals. Advances in genetic marker technology indeed allow for dense genotyping of hundreds of thousands of Single Nucleotide Polymorphisms (SNPs) per individual. This allows to characterise representative samples composed of several hundreds to several thousands of cases and controls, each one characterised by up to a million of genetic markers sampling the genomic variations among these individuals.

The standard approach to genome wide association studies is based on univariate hypothesis tests. In this approach each genetic marker is analysed in isolation from the others, in order to assess its potential association with the studied phenotype, in practice by the computation of so-called *p-values* based on some statistical assumptions about the data-generation mechanism. Because of the very high ratio between the large number of SNPs genotyped and the limited number of individuals, multiple-testing corrections need to be applied when carrying out these analyses, leading to reduced statistical power.

While this standard approach has been at the basis of many novel loci unravelled in the last years for several complex diseases, it has several intrinsic limitations. A first limitation is that this approach does not directly account for correlations among the explanatory variables. A second intrinsic limitation of GWAS is that they can't account for genetic interactions, i.e. causal effects that are only observed when specific combinations of mutations and/or non-mutations are present at the same time. The third limitation of univariate approaches is that they do not directly allow to assess the genetic risk, since many of the identified markers (with similarly small p-values) actually account for the same underlying causal factor: exploiting their information to predict the genetic risk is hence far from straightforward.

Within bioinformatics, machine learning has actually become one of the major potential sources of progress. As a matter of fact, biology has become nowadays one of the main drivers of research in machine learning, and is by itself already a very competitive research field.

Among the subfields of machine learning, supervised learning and its extensions such as semi-supervised learning, stand out as the most mature and at the same time most rapidly evolving area of research. Within this context, the purpose of this thesis was to study the application of random forest types of methods to genome wide association studies, with the twofold goal of (i) inferring predictive models able to asses disease risk and (ii) to identify causal mutations explaining the phenotype. The choice of this family of methods was originally motivated by the fact that these methods are a priori well suited for that kind of analysis due to some of their interesting properties. They are indeed able to deal efficiently with very large amounts of data without relying on strong assumptions about the underlying mechanisms linking genetic and environmental factors to phenotypes, and they can also provide interpretable information, in the form of scorings and/or

rankings of SNPS so as to help in the identification of causal genetic loci.

In the first part of this manuscript, we analyse the state-of-the art in the application field of genome wide association studies and in supervised machine learning, and subsequently describe in details the three tree-based ensemble methods that we have implemented and applied in our research; in Part II, we report our empirical investigations, in three successive steps, namely i.) a preliminary study on simulated datasets yielding controlled conditions with known ground-truth and allowing for a first sanity check of the T-Trees methods, in ideal conditions; ii.) a detailed study on a given real-life dataset concerning Crohn's disease, where we try to understand the main features of the three different algorithms in terms of predictive accuracy and capability of identification of relevant genetic information, and their sensitivity with respect to various kinds of quality control procedures and algorithmic parameters; iii.) a systematic replication study, where we confirm, on 7 different datasets from the *Wellcome Trust Case Control Consortium*, the main outcomes of our study on the Crohn's disease, while using default parameter settings.

Remerciements

Durant ces dures années de labeur, de nombreuses personnes m'ont aidé, accompagné, épaulé, encouragé, soutenu ou tout simplement écouté. Toutes ces contributions, qu'elles soient scientifiques ou non, ont été déterminantes et souvent essentielles, pour le bon déroulement de cette épopée.

Parmi les essentiels, je commencerai par remercier Louis Wehenkel, promoteur de cette thèse, pour ses précieuses interventions, sa disponibilité et sa patience. La confiance et le respect qu'il m'a accordés m'ont permis d'évoluer et de découvrir le monde passionnant de la recherche. Nos nombreuses discussions ont plus que largement contribué à la bonne réalisation de ce travail. Toujours dans les essentiels, je remercie également Pierre Geurts, co-promoteur de ce travail, pour le partage de son expertise, sa disponibilité, ses nombreuses réponses éclairées et éclairantes, son dévouement et sa gentillesse.

Je remercie également les membres du jury de cette thèse qui ont dédié leur précieux temps à sa lecture et à son évaluation.

Ensuite, il y a les satellites, toutes ces personnes qui ont gravité, de près ou de loin, autour de cette recherche. Ceux avec qui j'ai eu l'occasion de partager et discuter, ceux qui ont élargi mes horizons et ouvert de nouvelles voies, ceux qui ont ajouté leur grain de sable et contribué, parfois sans s'en rendre compte, à la bonne conduite de mes recherches. Tom Druet pour ses précieuses remarques et le temps qu'il m'a accordé. Michel Georges pour tous ces échanges, parfois furtifs mais intenses et constructifs. J'ai également une pensée particulière pour l'impulsion qu'aura pu donner Sarah Hansoul à cette recherche. Je remercie aussi Gilles pour son coup de pouce efficace, indispensable et linéaire de la dernière ligne droite. Enfin, dans le désordre et non exhaustivement, je remercie Christophe, Jean-François, Fabien, Raphaël, Benjamin, Yannick, Olivier, ... et tous les collègues.

Après, il y a les persévérants. Ceux qui ont cru en moi sans en démordre une seconde, ceux qui ont été présents pour me maintenir sur la route, ceux qui m'ont entouré de leurs joies. Je ne remercierai sans doute jamais assez mes parents, mes deux merveilleuses soeurs, leurs magnifiques plus si petits bambins, mes grands-parents, Delphine, et Antoine (qui n'y est pas pour rien). Je remercie également Monique pour son authenticité et Jean-Claude pour son humour. Je remercie aussi tous ceux qui ont partagé leur amitié ainsi que de bien agréables instants (et surtout de bonnes bières) tout en refaisant le monde à mes côtés: Gérôme pour sa salsa, Samuel pour son coup de crayon sur la couverture, Jean-Marc pour son accent, Jérome pour ses anti-corps, ... et puis tous les autres.

Je remercie également l'Université de Liège, le GIGA, l'Institut Montefiore et le FNRS qui m'ont accueilli, couvert et chauffé quand il faisait froid.

Enfin, il y a Sylviane. Mon amour, mon étoile, mon refuge, mon pilier. Elle m'a accompagné tout au long de ce parcours. Jour et nuit, elle m'a soutenu et conforté. A mes côtés, elle a participé à cette épreuve et a vécu intensément les humeurs, bonnes et moins bonnes, qui accompagnent cette longue traversée du doctorat. Merci pour ta persévérance, ta force et la constance de ton amour. Merci de m'avoir attendu. Merci à toi, du fond du coeur.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

In this chapter we successively discuss the motivations of our research, present the overall approach used to work, and then outline the structure of the rest of this manuscript and conclude by stating our personal contributions and by providing an annotated list of our publications.

1.1 Motivations

Understanding underlying mechanisms of common diseases, such as cancer, cardiovascular diseases, inflammatory and allergy disorders, is one of the major goals of current research in medicine. As most of these disorders are linked to genetic factors, identification of the associated variants forms an excellent strategy towards the elucidation of molecular and cellular dysfunctions, and *in fine* could lead to better personalized diagnostics and treatments.

Genome-Wide Association Studies (GWAS) aim to discover variants spread over the genome that could lead, in isolation or in combination, to a particular trait or an unfortunate phenotype such as a disease [Man10, TKTJ11, LCVK11, Wel07]. The basic idea behind these studies is to statistically analyze the genetic differences between two populations: a group of healthy individuals (the controls) versus a group of sick ones (the cases). Advances in genetic marker technology indeed allow for dense genotyping of hundred of thousands of Single Nucleotide Polymorphisms (SNPS) per individual. This allows to characterize, at an acceptable cost, representative samples composed of several hundreds to several thousands of cases and controls, each one characterized by up to a million of genetic markers sampling the genomic variations among these individuals.

In addition to the genetic measurements and the binary case/control classification, the individuals may also be characterized by additional information, such as for example, additional phenotypes refining their biological condition, and a multitude of environmental factors that may interact with genetic ones and often significantly impact disease status. Furthermore, meta-datasets may be constructed by merging information from several independent studies about the same or related diseases [B$^+$08, T$^+$12b, S$^+$12].

The very high practical importance of all theses studies, and the rapidly growing amount and complexity of the data generated by all these experiments raise many interesting questions for their analysis, and hence foster intensive research on the development of novel bioinformatics and statistical methodologies to help extracting in a more effective way the relevant information from these datasets.

The standard approach to genome wide association studies is based on univariate hypothesis tests. In this approach each genetic marker is analyzed in isolation from the others, in order to assess its potential association with the studied phenotype, in practice by the computation of so-called p*-values* based on some statistical assumptions about the data-generation mechanism [Bal06, M$^+$08, BCB04]. Because of the very high p/n ratio in GWAS (here p denotes the number of explanatory variables, i.e. the number of SNPS genotyped, while n denotes the sample size, i.e. the number of individuals used as cases and controls), multiple-testing corrections need to be applied when carrying out these analyses, leading to reduced statistical power.

While this standard approach has been at the basis of many novel loci unravelled in the last years for several complex diseases, it has several intrinsic limitations.

A first limitation is that this approach does not directly account for correlations among the explanatory variables, while in the context of GWAS this correlation is often very strong, in particular due to the fact that genetic mutations are transmitted from parents to children through a combination of chromosome replication and cross-over, which leads to a high probability that mutations that are closely located on a DNA strand are inherited in combination, hence leading to strong correlations among closely located genetic markers. Apart from this unavoidable physical correlation, correlations among different markers may also appear as artefacts induced by the experiment design (sample selection, experimental batches, imputations of missing variables) routinely found in the datasets used for GWAS. The net result is that p-value based marker rankings need to be carefully analyzed by hand and subsequently experimentally replicated and validated before they can be confirmed as pin-pointing to genuine causal effects. This situation led to a recurring difficulty in reproducing findings published in the literature and made the scientific community become extremely cautious [INTCI01] and demanding on the soundness of the statistical approaches used in GWAS.

A second intrinsic limitation with the univariate approaches to GWAS is that they can't account for genetic interactions, i.e. causal effects that are only observed when specific combinations of mutations and/or non-mutations are present at the same time. Even though opinions are divided [HGV08, ZHSL12], such potential epistatic effects should however be taken into account in order to increase the power of the statistical analyses. Furthermore, it is not unlikely that many genetic factors are in some way coupled with environmental factors, and taking these couplings systematically into account is as well beyond the capabilities of simple univariate approaches.

The third limitation of univariate approaches is that they do not directly allow to assess the genetic risk, since many of the identified markers (with similarly small p-values) actually account for the same underlying causal factor: exploiting their information to predict the genetic risk is hence far from straightforward, even more so if we want to take into account potential gene-gene or gene-environment interactions.

Since the mid-eighties, the field of machine learning has emerged at the intersection of algorithmics and statistics. The overall goal of the field is to design and theoretically characterize algorithms to extract in a reproducible way relevant information from observational data. The field is driven by a large diversity of applications, such as text mining [FS06], image analysis [MGW09, MGW07], extraction of knowledge from the internet [CMFF10], analyzing data from experimental sciences such as astronomy [B+09], earth monitoring [K+11], high energy physics [P+12b], and – last but not least – biology and medicine [MAW10]. Within bioinformatics, machine learning has actually become one of the major potential sources of progress, as one can contemplate from the growing number of conferences and journals that focus on the application of machine learning to biology. As a matter of fact, biology has become nowadays one of the main drivers of research in machine learning, and is by itself already a very competitive research field.

Among the subfields of machine learning, supervised learning and its extensions such as semi-supervised learning, stand out as the most mature and at the same time most rapidly evolving area of research: the general statistical theory underlying the analysis of supervised learning algorithms has been established at the end of the last century [PRMN04, Vap98a], and in the meantime several powerful paradigms have been developed allowing to leverage supervised learning to a very broad class of problems. Among these supervised learning methods, both kernel-based models [lS02] and tree-based models stand out. In particular, random forest types of methods [Bre01, GEW06], have been shown to provide state-of-the-art results in many applications (e.g., image analysis, bioinformatics, reinforcement learning, etc.), specially in terms of their excellent accuracy vs computational complexity compromise.

Within the above context, the subject of this thesis was defined a few years ago, in the establishment of a collaboration between the research unit in Systems and Modeling of the Department of Electrical

Engineering and Computer Science, on the one hand, and the research unit in Animal Genomics of the Faculty of Veterinarian Medicine, on the other hand. These groups were and still are respectively active in machine learning, and specifically in tree-based supervised learning, and in genome wide association studies, and specifically the study of complex genetic diseases.

The purpose of our work was to study the application of random forest types of methods to genome wide association studies, with the twofold goal of (i) inferring predictive models able to asses disease risk and (ii) to identify causal mutations explaining the phenotype. The choice of this family of methods was originally motivated by the fact that these methods are a priori well suited for that kind of analysis due to some of their interesting properties. They are indeed able to deal efficiently with very large amounts of data without relying on strong assumptions about the underlying mechanisms linking genetic and environmental factors to phenotypes, and they can also provide interpretable information, in the form of scorings and/or rankings of SNPS so as to help in the identification of causal genetic loci.

1.2 Approach to research

Given the limitations of the standard approach discussed in the preceding section, and the acknowledged capability of tree-based methods to handle complex problems in a very flexible way, it is of interest to investigate whether and how these methods could be used in order to improve on univariate approaches in the context of GWAS.

To carry out this investigation, we have worked during our thesis along the following work plan:

- We have started our work by implementing our own software for supervised learning with ensembles of trees; to do so we have been inspired by the experience in the Systems and Modeling team, but we have developed our own software from scratch in order to facilitate later adaptations and benchmarkings.

- At the onset of our thesis, we have participated in some GWAS carried out in the companion group of animal genetics, to become familiar with the nature of the datasets, and most importantly with the quality control and other preprocessings needed prior to statistical analysis.

- We then have applied standard random forest types of algorithms to some synthetic and real-life datasets, in order to gain some first hands-on experience. This also led to some improvements in our software implementation, so as to make it sufficiently efficient (memory, CPU time, adaptation to grid environment) to handle very big datasets.

- In order to cope with the correlation structure implied by linkage disequilibrium, we have designed a novel tree-based ensemble method called *T-Trees* and tested it on synthetic data. The method is based on the segmentation of the vector of SNPS into blocks of fixed size along the genome, and then uses the SNPS inside each block in a homogenous way by first selecting at each tree node a block and then jointly exploiting the SNPS inside this block to create a split. We found that it yields both improved predictive accuracy and a better precision in the detection of causal loci.

- Finally, we have carried out a systematic large-scale empirical investigation based on state-of-the-art and publicly available GWAS datasets about several human diseases. This study allowed us to better understand the features of tree-based ensemble methods in real-life conditions, in particular by identifying the role played by the interaction of rare variants with some pathological behaviors of the score measures used for tree induction. We also assess in this study the effect of quality control on the apparent predictive power of the induced classifiers, by comparing results according to different quality control procedures. The results found in this study should help other machine learning researchers to more effectively analyze their results when using complex black-box procedures such as tree-based ensemble methods, and at the same time help biologists to gain confidence in these results.

1.3 Organization of the manuscript

The main body of the manuscript is divided in two parts.

In part 1, we start by describing the state-of-the-art in genome wide association studies (Chapter 2), and then provide in Chapter 3 the required background in supervised machine learning. Chapter 4 gives a precise description of the algorithms that we have developed and applied in this work.

In part 2, we present our experimental results. We start, in Chapter 5 by analyzing the behavior of standard random-forest types of methods and our proposed method called T-Trees on synthetic datasets, where the ground truth is known, and where we can easily vary experimental conditions (noise level, number of samples etc.) In chapter 6, we then provide detailed results on the real-life *Crohn*'s disease (CD) dataset from the WTCCC [Wel07], and link our results with the scientific literature. Chapter 7 complements our empirical study by investigating the six remaining datasets related to other diseases provided by the WTCCC. Some complementary simulations results are collected in the appendices.

Finally, we conclude by discussing in a retrospective way our findings and by suggesting future directions of research.

1.4 List of publications

In [BGHW08a, BGHW08b], we started to tackle the problem of correlated descriptors in GWAS by considering two different representations of the input data: the raw genotypes described by a few thousand to a few hundred thousand discrete variables each one describing a single nucleotide polymorphism and, on the other hand, haplotype block contents represented by the combination of 10 to 100 adjacent genotypes. The blocks were defined by the HapMap hotspot lists. We adapted the Random Forests to exploit those blocks and compared the results with the use of raw genotypes in terms of predictive power and localization of causal loci. The adaptation consisted in modifying the splitting rule based on estimation of the conditional probability that the observed haplotype is drawn from the population of cases (reps. controls) reaching the current node (assuming class conditional independence of the SNPS in the block). That methodology was applied on simulated datasets with one or two interacting causal mutations. We obtained marginally superior results with our adaptation of the state-of-the-art tree-based method than their direct application to the raw genotype data. That first contribution opened the path we followed in the present thesis.

Also, at the beginning of my PhD, I had the opportunity to develop a graphical interface allowing biologists to annotate images and perform different measurements while extracting subimages used as inputs for tree-based automatic image classification. That work lead to a publication [G$^+$08b] related to the effectiveness of inhaled doxycycline to prevent allergen-induced inflammation in a mouse model of asthma.

An article [BLGW13] presenting the core results of this thesis, namely the *T-Trees* algorithms, and its application to seven real GWAS datasets is under preparation and will be submitted to a journal. In this paper, the capabilities of various tree-based ensemble methods to assess disease risk and to localize causal mutations are evaluated. We are also preparing a short technical note to be submitted to a bioinformatics journal, where we present our findings about the impact of the normalisation of the splitting-criterion used in random forests methods and their bias towards markers with small minor allele frequencies (appendix A.1).

Part I

Background and Methods

Chapter 2

Genome-Wide Association Studies

Contents

Humans are unique but genetically 99% equal. The remaining 1% of genetic differences participate in their rich diversity. Deoxyribonucleic acid (DNA) depicts the essential information needed for building up a human living from the biological point of view. This genetic material can be seen as a linear code underlying the development, the functioning and the reproduction of organisms. Nevertheless, some coding errors may occur which, unfortunately, could cause dysfunction at many levels and eventually lead to diseases. The aim of genome-wide association studies is to locate genetic differences between two sub-populations that are responsible of the differences of one or several phenotypes observed between these sub-populations, and in particular that are related to complex genetic diseases.

This chapter first aims at providing a gentle introduction to the field of genome-wide association studies to non specialists. On the way, we will also present and discuss the current state-of-the-art in terms of statistical analysis techniques commonly used in this context.

2.1 DNA : 3 letters for 3 billions bases

All humans have a sequence of roughly 3 billion DNA bases spread over their 23 pairs of chromosomes. DNA sequences can be viewed as a code containing genetic instructions.

From the informatics viewpoint, DNA is essentially a very long linear string built over an alphabet of four different letters defined by chemical bases (or nucleotides) : A, C, T, G. Almost every cell of an organism contains two copies of this DNA string and this information is transmitted in a quite reliable way from one cell to its daughters, and from one individual to its off-springs (actually exactly one half from both parents). Indeed, the genetic machinery responsible of DNA replication is quite robust, and thanks to coding redundancy and error-correction mechanisms, the genetic information encoded by DNA strands is normally reproduced with great fidelity.

However, during the replication process, errors may occur over time and survive in the offspring (cell lines and/or sub-populations), for example by changing one base to another at some positions in the code, or by yielding multiple copies of some DNA subsequence. If we screen the genetic material of a sample of individuals of a population, at a position where such a mutation phenomenon occurred in the past, we will therefore observe that some individuals (typically a large majority of them) are holding the original genetic material, the so-called wild type, while others (typically a small minority) hold the mutant variant.

Among the various types of mutations that may occur, we focus in this thesis on point-wise mutations characterized by the change of a single letter (a replacement, a deletion, or an insertion) in the DNA string. The resulting genetic variability is called Single Nucleotide Polymorphism (SNP) and its alternate forms observed in the population are called alleles; in most of the cases, an SNP is characterized by only two alleles which translate into three different combinations for diploid organisms.

SNPs are the most abundant source of genetic variation (aside from structural variations) within the human genome, notably because many of these appear in non-coding regions. (this is the redundant code defining the mapping between DNA strings and protein sequences). From recent genetic surveys, it is known that these SNPs occur approximately once every 100 to 300 base-pairs on the average. The International HapMap Project [The03] has studied these variations and has identified about ten million SNPs (where the rarer SNP allele has a frequency of at least 1%) in three sub-populations of humans (Europeans, Africans and East Asians). In the continuity, the 1000 Genomes Project [G$^+$10] sequenced the genomes of more than 1000 people to obtain a more detailed and comprehensive catalogue of human genetic variation. The 1000 Genomes data are available to the scientific community. They can be used, for example, to impute genotypes not directly typed thus avoiding important genotyping costs.

Since genetic material is transmitted in a way such that nearby bases are transmitted together with a high probability, because of the mechanics of DNA replication and re-combination, when two individuals share the same alleles at a SNP locus, it is likely that they also share the same material in the nearby areas of the DNA string. Therefore, even if SNPs only describe a very small part of the DNA of an individual, they are expected to provide a significant amount of information about their genetic differences. Hence, studying the correlations between SNPs and phenotypes may help to identify genetic regions where mutations occurred that are functionally related to the phenotype variability. For the same reason SNPs may potentially be used to predict phenotypes and in particular genetic disease risks.

2.2 SNP : 3 letters (again) for 3 values

Polymorphisms are what make every one of us unique from the genetic point of view. Most of these have no known effect and may be of little or no importance while some of them influence physical appearance, disease risk or drug response. SNPs are involved in the early steps of development. Depending of their nature and location in the genome, they can change the encoded amino acids (non synonymous) or can be silent (synonymous) or just occur in noncoding regions [Sha09]. Thus they can influence more or less one trait, e.g., they may impact the encoding of mRNAs responsible for proteins synthesis (figure 2.1). At the surface, they influence together with environmental factors our general phenotype : the way we look, the hair and eyes color, our weight, size etc. Below the surface, they also may impact how our individual cells will grow, replicate and interact with the others (which may be less obvious to directly observe). Nevertheless, small variations in the DNA sequence can also lead to undesirable effects such as diseases. Mutations can indeed be the starting point of cascades leading to an unexpected and possibly counter-productive trait.

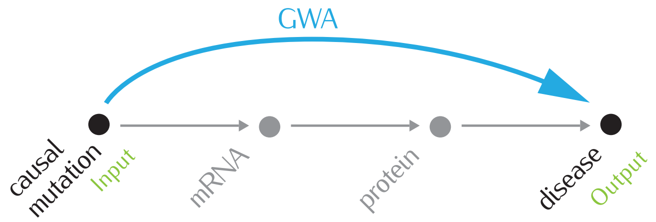


FIGURE 2.1 One way for a mutation to influence phenotype.

Most of the SNPs are biallelic, giving rise, in diploid organisms such as humans, to 3 types of observed genotypes describing how many mutant variants are collected by this individual at a given position. Thus, a genotype can take 3 values : 0, 1 or 2. A value of 0 means that no mutation has been observed, the two alleles are of the wild type. In that case we say that the SNP is homozygous wild. A value of 1 means that one mutation is observed at that position, one wild allele on one of the chromosomes and one mutant allele on the other chromosome, also called the heterozygous. Finally, 2 represents the case where the two alleles are mutants and is called a homozygous mutant genotype.

Figure 2.2 illustrates these ideas: the right-most part shows part of the DNA inherited by a child from its two parents, by depicting a chromosome pair around five nucleotides; the two central parts show the corresponding DNA of the corresponding chromosomes of its two parents; finally the left-most part shows the content of the corresponding wild-type chromosome of the reference population. In green, mutations are highlighted, yielding eventually the genotype of the child and its encoding in orange.

Note that sometimes that representation is not the exact one being used; one can indeed decide to code these values by using any specific allele as a reference (most of the time, the mutant allele is the one that is less frequent in a reference population).

2.3 Genome-wide associations studies

SNPs have the potential to help identify the multiple genes associated with many phenotypes. Of course, SNPs generally do not directly cause an illness but they can help us to identify genomic regions potentially containing mutations causally affecting the biology of the studied phenotype and they could hence help us to evaluate the risk that someone will develop a disease. Identification of causal mutations will provide better diagnostic information that will allow for early diagnosis, prevention and better treatment of human diseases. In the following, we will indifferently use the word "phenotype" to refer to the trait under study which can be a disease or any observable characteristic also called a trait (such as morphology, development, biochemical or physiological properties, behavior...). The phenotype can be qualitative (e.g., disease status) or quantitative (e.g., treatment response status). Analyzing DNA can help to understand what is happening beyond the genetic code, in other words, what are the underlying molecular mechanisms leading to a given phenotype. Genome-wide association studies are designed for that purpose. However, as those genetic variations are transmitted through generations, they also are directly related to ancestry and family relations among individuals.

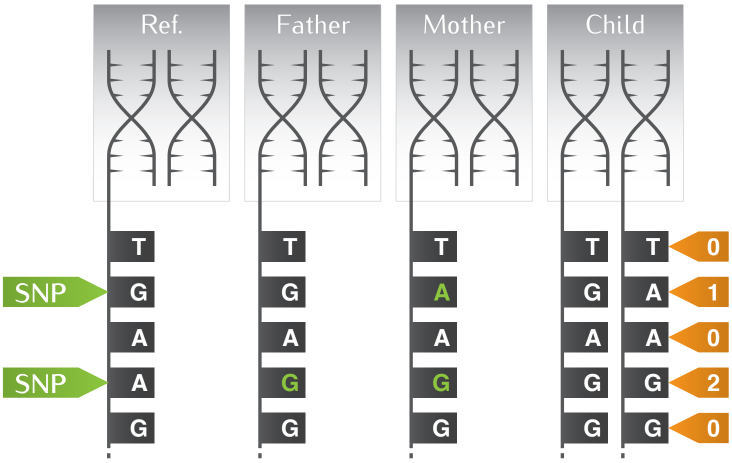


FIGURE 2.2 Mother and father chromosomes DNA differ from reference DNA at one locus (A/G mutation, in green); the mother in addition differs at a second locus (G/A mutation in green). The child's genotype is obtained by combining these variations resulting from the inherited chromosome pair (in orange).

Unlike monogenic Mendelian diseases, a disease is said to be complex when multiple interacting genes and environmental factors are responsible for the phenotype. These complex diseases are not caused in a deterministic way by a single genetic mutation; rather they must have an intricate molecular architecture and may therefore be influenced by a potentially large number of genetic and environmental factors which may act in additive, complementary and/or more complex ways. Most of the time, in that case, it has been observed that each individual genetic variant only makes a very small contribution to the overall heritability of the disease. Since complex diseases are intrinsically related to the perturbation of a complex biological sub-system, this may explain the dispersed association among individual genetic variations and disease phenotype, as well as a rather high sensitivity to environmental factors. In addition to this dispersive effect, it may be the case that part of the heritability of genetic risks towards complex diseases can only be explained by conditional effects, i.e. effects which imply a conjunction of genetic and environmental factors. It might be the case that a multitude of such dispersed and/or conjunctive effects in the end is responsible of the fact that most of these complex diseases are not rare in the population.

When performing a genome-wide association study, the main questions we are trying to answer are the following:

- How many genetic variants are involved ?
- Where are these genetic variations located on the genome ? Do they appear in exons or introns ? Which genes do these changes affect ?
- What is the type and biological consequence of each alteration ? Which allele is protective and which one is causative ?
- What are the functional consequences of these changes ?
- how often such mutations occur (allele frequency, mutation rate) ?
- Are these variants more important than environmental factors ?
- Are there any interactions between the genetic and environmental effects ?
- Can we infer the disease risk, or more generally predict a phenotype, based on genetic factors alone or in combination with environmental ones ?

Typically, to carry out such a study one disposes of a cohort of a few hundred to several thousand individuals, a fraction of them (typically about 50%) having a certain phenotype which are called *cases*, and the rest of them being individuals representative of the genetic variation in the studied population and who do not present the studied phenotype which are called *controls*. This is schematically represented at Figure 2.3.

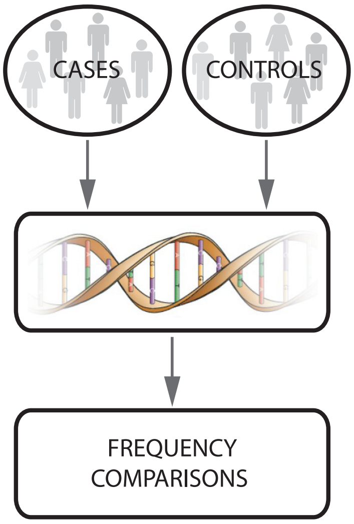


Figure 2.3 Overall principle of a genome-wide association study: step 1 (top) consists in collecting a cohort of cases and controls (experiment design); step 2 (middle) consists in extracting DNA from the individuals and carrying out measurements to characterize their genetic variations (e.g. genotyping at SNP loci); step 3 (bottom) consists in analyzing the resulting dataset so as to identify significant associations among groups of SNPs and phenotype and to determine risk prediction models (statistical inference).

2.3.1 Thanks to linkage disequilibrium

Advances in genotyping technologies allow for genome-wide association studies. In a short time, hundreds of thousands (and even more) SNPs spread over the whole genome can be genotyped for large sets of individuals at low cost. Denser genotyped variations over the whole genome should allow to detect causative DNA regions even if the biologically causative mutation is not directly observed, indeed, those variants are known to be strongly correlated. Thus, the chances of finding an SNP "linked" with the causal one are highly increased (indirect association vs. direct association. See Figure 2.4).

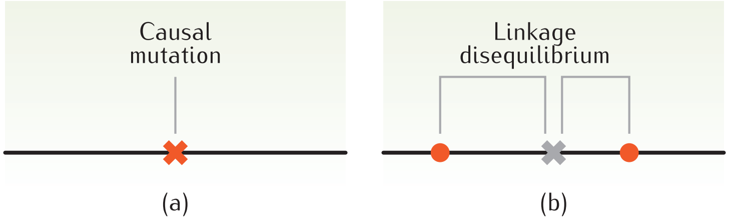


Figure 2.4 We talk about direct association (a) when the causal mutation is directly genotyped. On the other hand, if only variations in LD with the causal mutation are genotyped then we talk about indirect association (b).

That "link" is also called the linkage disequilibrium (LD). It denotes the nonrandom association of alleles at two or more loci. One way of measuring LD between two variants is to compute a simple statistic for a pair of SNPs. Doing so for each pair of SNPs will generate a matrix of values from which LD patterns can be deduced.

The International HapMap Project [The03] investigated LD patterns across the entire human genome. They started by gathering anonymized samples from four different populations: 90 Yoruba (30 parent-offspring trios) from Ibadan (Nigeria); 90 individuals (30 parent-offspring trios) of European ancestry from Utah, 45 unrelated Han Chinese from Beijing and 45 unrelated Japanese from Tokyo. Their findings pointed out that there exist hotspots of recombination in the genome which drive the observed LD patterns. They observed blocks of high LD separated by sharp breakdown of LD corresponding to hotspots.

Those blocks of high LD are also referred to as haplotype blocks. The term haplotype is a contraction of haploid genotype. Haplotypes are the combinations of alleles at different positions along the same chromosome that are transmitted together. It may be much more informative to analyze them simultaneously instead of independently. These haplotypes have a particular structure which provides information on evolution history. According to the International HapMap Project, we now know that chromosomes are structured in many blocks, i.e., haplotype blocks within which there is a limited haplotype diversity (where little to no recombination events occured) separated by small regions of high haplotype diversity. This structure is population dependent.

Technically, the main benefit of those low haplotype diversity regions is that only a few markers need to be genotyped to capture the whole haplotype information. Selecting the minimal number of SNPs that uniquely identify common haplotypes is called haplotype tagging. That property has driven the current GWAS in selecting the right amount of markers along the genome in order to capture a maximum of the variations present in the population under study. However, being focused on common variations, a direct (and maybe not so desirable) consequence is that the possible presence of rarer mutations may be missed and their potential implications underestimated.

To explain the notion of linkage disequilibrium, let us consider two SNPs, the first has alleles A and B, and the second has alleles C and D. In a given population, let us suppose that the marginal allele frequencies are f_A, $f_B = 1 - f_A$, f_C and $f_D = 1 - f_C$, respectively. We define a haplotype as being a particular combination of the alleles of these two SNPs on one chromosome at variant sites and denote the haplotype (joint) frequencies as f_{AC}, f_{BC}, f_{AD} and f_{BD}.

	A	B	Total
C	$f_{AC} = f_A f_C + \mathcal{D}$	$f_{BC} = f_B f_C - \mathcal{D}$	f_C
D	$f_{AD} = f_A f_D - \mathcal{D}$	$f_{BD} = f_B f_D + \mathcal{D}$	f_D
Total	f_A	f_B	

TABLE 2.1 In this table, $\mathcal{D}$ represents the departure from the uncorrelated state in which the joint frequencies is equal to the product of the marginal frequencies. When $\mathcal{D}$ is equal to 0 the two SNPs are said to be in linkage equilibrium (LE).

The situation can be summarized in Table 2.1 where the standard coefficient $\mathcal{D}$ of LD between the two loci is defined by:

$$\mathcal{D}_{AC} = f_{AC} - f_A f_C. \tag{2.1}$$

Equation 2.1 expresses that the expected haplotype frequency in the absence of LD is the product of the marginal frequencies. $\mathcal{D}_{AC}$ represents the departure from the uncorrelated state. Simple algebraic rearrangement shows that:

$$\mathcal{D}_{AC} = -\mathcal{D}_{BC} = -\mathcal{D}_{AD} = \mathcal{D}_{BD}. \tag{2.2}$$

$\mathcal{D}_{AC} = 0$ would suggest independence of the two SNPs but could also simply reflect a low marginal frequency of one of the alleles. Also, the sign of $\mathcal{D}$ is sensitive to the allele code which can be chosen arbitrarily. To circumvent those two drawbacks, two derived LD statistics which both are frequency normalized and always positive are more commonly used:

$$|\mathcal{D}'| \quad = \quad \begin{cases} \frac{-\mathcal{D}_{AC}}{min(f_A f_C, f_B f_D)} & \text{if } \mathcal{D}_{AC} < 0, \\ \frac{\mathcal{D}_{AC}}{min(f_A f_D, f_B f_C)} & \text{if } \mathcal{D}_{AC} > 0, \end{cases} \tag{2.3}$$

$|\mathcal{D}'|$ ranges from 0 to 1, 0 means linkage equilibrium, a value of 1 corresponds to complete LD (the two loci are not separated by recombination, i.e. at most three of the four possible haplotypes are present in the population) but does not necessary indicates that one locus can predict the other with high accuracy.

$$r^2 \quad = \quad \frac{\mathcal{D}_{AC}^2}{f_A f_B f_C f_D}. \tag{2.4}$$

The r^2 is the squared Pearson correlation coefficient, a value of one corresponds to perfect LD for which at most two haplotypes are possibly present. In other words, knowing the allele at one locus allows to predict the allele at the other one.

From these metrics, it is possible to compute a matrix of LD. Figure 2.5 shows an example of observable patterns in a small chromosome 1 region in the Hapmap CEU population. It allows to visually detect blocks of SNPS in high LD clearly separated by LD breakdown.

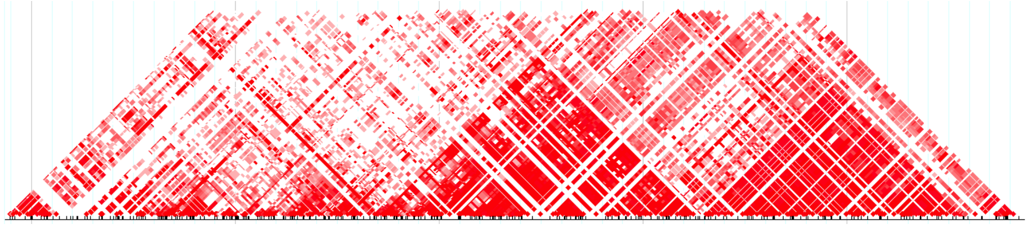


FIGURE 2.5 An example of LD pattern in the Hapmap CEU population, on chromosome 1 (67.68..68.18Mb) region. The different pairwise values of $\mathcal{D}'$ are represented by different intensity of red. We clearly see "triangles" of higher LD depicting the haplotype block structure in that region.

2.4 GWAS : how ?

A genome-wide association study is driven by the following steps (see also Figure 2.3):

1. Choosing and collecting samples: maybe the most difficult part of a GWAS, collecting samples isn't easy at all. Cases may sometimes be rare and for some diseases, getting DNA samples is delicate. An accurate definition of the trait under study is required to minimize the heterogeneity of the underlying causal factors and increase the power of the study. Another major difficulty arises from matching case and control populations in order to avoid (or at least minimize) sample stratification.

2. Genotyping: using recent technologies allows now for genotyping massive amounts of markers at low cost. Genotyping arrays now provide up to one million common SNPS which currently approximately costs 400$ (and that cost will continue to decrease over time, while the number of variants assayed will increase). For that task, two main manufacturers (*Affymetrix* and *Illumina*) respectively provide hybridization-based and enzyme-based genotyping technology. In the end, these genotyping arrays allow to measure allele intensities at several locations in the genome from which genotypes can be

deduced. For further information concerning genotyping arrays, we refer the reader to [K+03] and [S+06].

3. Quality controls: due to the previous steps, it is necessary to check the quality of the resulting data. Those quality control methods (QC) will be further discussed in Section 2.4.1.

4. Statistical analysis: there are two main approaches, the single-locus analysis where each variant is tested in turn and the multi-locus approaches where haplotypes, gene-gene (and possibly higher-order) interactions can be considered. Basically, this step consists in frequency or similarity comparisons between the cases and the controls. These tests are presented in Sections 2.4.2 and 2.4.3.

5. Replication: once some loci have been identified as being statistically associated to the studied phenotype, replication allows to validate or invalidate those results. The replication study is often carried by the use of a different genotyping platform, which may help to remove spurious associations due to technical artefacts.

2.4.1 Quality controls

Once the data have been collected and the genotypes assayed, we want to avoid confounding of signal with something that is not linked to the trait under study. Many quality control (QC) procedures exist to reduce the risk of false-positive and false-negative findings. Basically, those QC procedures can be based on samples or markers. In the following, the main QC procedures used in practice are discussed. For further reading we suggest the recently published tutorial from the genomics group of the *eMERGE* [T+11].

Sample-based QC

- Sample mix-ups and plating errors: it is possible that during the preparation, some samples are mixed up on the plate. It can be messy and sometimes two or more samples are inverted on the array (usually composed of 96 wells). One way of detecting such errors is to check the sex recorded when collecting information about the samples and the one that can be estimated using the X chromosome. E.g., a female with a low heterozygosity rate across the X chromosome markers is probably a good indication of sample mix-up. Also, by mistake, if two samples are placed in the same well it will produce an excess of heterozygosity (on the other hand, low heterozygosity indicates inbreeding).

- Low-quality DNA samples: quality and concentration of the collected DNA may vary from one sample to another, especially when the cases and the controls are not collected and extracted in the same place which can introduce spurious association. Even on the same plate, it is inevitable to observe variations in the quality and concentration, bad quality and/or low concentration samples often lead to failure of signal amplification causing genotypes to be uncallable which results in missing genotypes. In that case, individuals with an insufficient overall call rate should be removed from the study.

- Plate effects: it is also required to check if there are no differences in genotyping frequency between plates. Especially in the situation where cases and controls are typed on different plates, such differences will cause confounding. A common practice is to evenly distribute cases and controls across plates. When it is not possible, a comparison of allele and/or genotype frequencies between one plate against the others will help identify significant differences and allow to discard samples from a study.

- Population stratification: one source of spurious associations is the presence of individuals coming from different ancestral and demographic history. Especially when cases and controls strongly differ regarding these features. It has been clearly observed that many markers carry such an information and demographic information can thus be confounded with disease status. In order to avoid associating that kind of markers to the disease, the first thing that can be done is to remove the outliers from the population under study. Most of the time, using principal component analysis and the addition of known and various ancestry (such as HapMap) genotypes allow for the identification of population structure and permit to confront the collected samples provenance informations to known populations. It has been observed that only two principal components are sufficient to separate clusters of different ancestry individuals. Another approach consists in calculating a genetic distance between pairs of individuals and using the resulting distance matrix to cluster the individuals. Similarly, duplicate and related samples can be identified using the same genetic distance; an abnormally small distance between a pair of individuals would indicate sample duplication and/or relatedness.

Marker-based QC

- Call-rate and Allele Frequency: unfortunately, genotyping platforms are not 100% reliable and could cause some uncertainty for many reasons. When the genotype calling[1] algorithm is not able to determine the genotype at a given position for a given sample, this measurement is normally considered as missing. If a few genotypes are missing, it is not so problematic, since SNPs with missing values can be removed without loosing too much information and if the genotyping is dense enough we could hope that another SNP in LD with the missing one has been correctly genotyped.

  At each locus, for each sample, the genotyping technique measures the intensity of allele A and B. For all the individuals, each genotype can be summarized by a genotype cluster plot (Figure 2.6) which are used for the determination of genotypes. In other words, the analysis of allele intensities is used to determine the genotype at that given locus. Poor quality genotype cluster plots lead to poor confidence in genotype calling, which can lead to a missing genotype or calling errors. Many algorithms are used to this end, we refer the interested reader to [VSKZ09] for a detailed comparison of genotype calling algorithms, to [M$^+$10] which highlights the potential inconsistencies between calling algorithms that can impact downstream analyses and to [L$^+$10] for a proposal of a statistic to evaluate the imputation reliability. Most of these calling algorithms outputs a confidence score for each SNP, when that score is too low, the corresponding genotype is considered as missing.

  Depending on the study, marker with an overall high missing rate should be removed from the study. Otherwise, they can be imputed by predicted values that are based on the observed genotypes at neighboring SNPs. To this end, softwares such as *IMPUTE2* [HDM09] exists. Basically, the imputation process exploits the missing genotype surrounding LD structure and the associated known haplotypes to “guess” what would be the genotype. These algorithms relies on a reference haplotype panel (such as the HapMap samples).

  Another common practice is to remove SNPs with a low minor allele frequency (MAF). Such variables are difficult to study as they require large sample size to gain sufficient statistical power. However, [M$^+$09] suggests that part of the missing heritability could be explained by such rare and recent variations.

- Hardy-Weinberg Equilibrium: the Hardy-Weinberg equilibrium (HWE) principle states that genotype frequencies at any locus are a simple function of allele frequencies (in the absence of migration, mutation, natural selection and assortative mating). In other words, at a given locus, where the alleles A and B are observed, respectively, at frequencies $f_A = p$ and $f_B = q = 1-p$, the following genotype frequencies are expected:

$$f_{AA} = p^2, \tag{2.5}$$

$$f_{BB} = q^2, \tag{2.6}$$

$$f_{AB} = 2pq. \tag{2.7}$$

[1] The genotype calling is the transformation of the allele intensities outputted by the genotyping platform into genotypes.

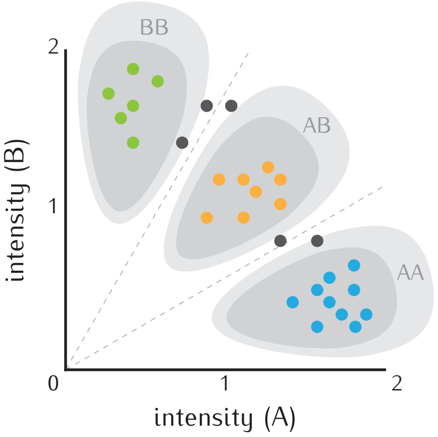


FIGURE 2.6 Genotype cluster plot: the green dots are clearly pointing out BB genotyped samples, the blue ones the other homozygous and the orange ones the heterozygous. The remaining dark gray dots correspond to samples for which the genotype is difficult to determine.

Where f_{AA}, f_{AB} and f_{BB} are the expected genotype frequencies for the homozygous wild, heterozygous and homozygous mutant respectively. It has been observed that these expectations hold for most human populations and in practice, deviations from HWE can indicate inbreeding, population stratification and genotyping problems. A χ^2 test between the expected frequencies and observed ones may detect such deviations. But these deviations can also pinpoint association [NEW98], since deviations from HWE can be due to a deletion polymorphism or a segmental duplication that could be responsible for a phenotype. Thus, one must be careful before discarding loci based on that test.

2.4.2 Single-locus test of association

Suppose that we look at a particular SNP in two sub-groups of a population: cases (individuals affected by the disease under study) and controls (individuals not affected). We denote the number of individuals in the two sub-groups by n_{case} and n_{cont} respectively. At that loci, let us say that we observe the 2 alleles: A, B. Genotype counts can be summarized in a two-way contingency table, as illustrated in Table 2.2. Such a table can be analyzed using an *observed-expected* test statistic which has a χ^2 distribution with two degrees of freedom in order to detect whether there is any relationship, or association, between the genotype and the disease status.

	AA	AB	BB	Total
$Cases$	a	b	c	n_{case}
$Controls$	d	e	f	n_{cont}
Total	$n_{AA} = a + d$	$n_{AB} = b + e$	$n_{BB} = c + f$	n

TABLE 2.2 Full genotype table for a general genetic model : 2×3 table

Based on Table 2.2, the idea is to spot genotype significant differences between cases and controls. The main question is to find whether or not there is an association between the genotype (columns) and the phenotype (rows). In Table 2.2, there is no association when the proportion of each genotype remains the same regardless the disease status. These counts are said to be expected under the null hypothesis that there is no association.

The genotype frequencies, assuming independence from the disease status, can thus be calculated as follows:

$$f_{AA} = \frac{(a+d)}{n}, \tag{2.8}$$

$$f_{AB} = \frac{(b+e)}{n}, \tag{2.9}$$

$$f_{BB} = \frac{(c+f)}{n}, \tag{2.10}$$

and from these frequencies, expected counts are derived, as shown in Table 2.3. It has to be noted that the total counts remain the same as in Table 2.2.

	AA	AB	BB	Total
Cases	$n_{case} f_{AA}$	$n_{case} f_{AB}$	$n_{case} f_{BB}$	n_{case}
Controls	$n_{cont} f_{AA}$	$n_{cont} f_{AB}$	$n_{cont} f_{BB}$	n_{cont}
Total	$a+d$	$b+e$	$c+f$	n

Table 2.3 Expected genotype counts

The idea is now to detect if there is a significant difference between the observed values (Table 2.2) and the expected ones under the independence hypothesis (Table 2.3). This can be achieved using the standard Pearson's χ^2 statistical test for independence of the rows and columns:

$$\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2.11}$$

where O_{ij} is the observed count and E_{ij} is the expected count in the cell in row i and column j. If the null hypothesis of no association is true, then the calculated test statistic approximately follows a χ^2 distribution with $(r-1) \times (c-1)$ degrees of freedom (where r is the number of row variants and c is the number of column variants) i.e. in our case $(r-1) \times (c-1) = 2$. This approximation can be used to obtain a p-value. A small p-value will suggest that there is association between the variables (genotype-phenotype) but the test will not indicate which are the cells (genotypes) in the contingency table that are the most associated. The χ^2 p-value of an observation corresponds to its probability plus all the more extreme ones under the null hypothesis (represented in green at Figure 2.7).

From those univariate tests, a p-value is assigned to each SNP, and thus a ranking of variables may be performed and a Manhattan plot (see Figure 2.8) is usually used to visualize the results. Spotting regions of interest is then a question of correct thresholding over the resulting p-values. Depending of the size of the spotted regions and on the genotyping density, it is often expected to see several variants for which the p-values are way under the significance level α.

One of the major issues at a genome-wide level is the multiple testing problem. Indeed, the larger the number of hypothesis tests, the larger is the probability of getting significant results due to chance. One way to correct for multiple testing is to adjust α (using methods like the *Bonferroni*, the *Šidák* or the *Benjamini-Hochberg* correction) in order to "control" the Type I error rate. Other methods exist such as permutation based adjustments. Resampling can be performed exhaustively, leading to so-called exact tests (if the set of observations is small enough) or approximate test (otherwise), such as *Monte Carlo* simulations. These tests allow for an estimation of how often a random observation can be as extreme as the observed one.

As a recommendation for the χ^2 test, the sample size should be such that no cells in the table have an expected count of less than one and no more then 20% of the cells should have an expected count of less than five. If samples are small, then Fisher's exact test can be used at the price of a higher computational cost. The exact p-value can be calculated by considering all the tables with the same row and column totals

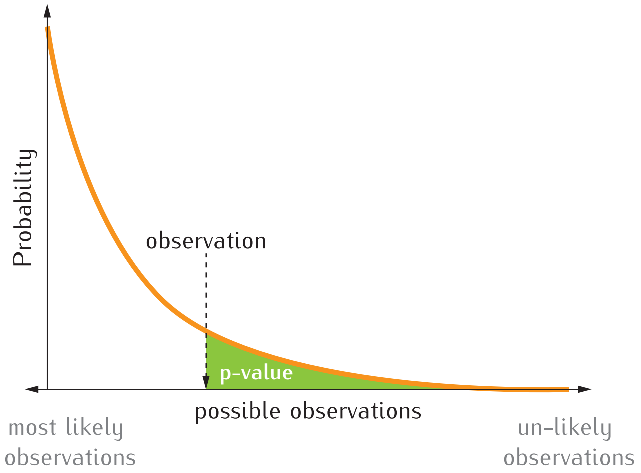


FIGURE 2.7 The curve represents the probability of every observed outcome under the null hypothesis. The p-value is the probability of the observation plus all the more extreme ones, represented by the green area.

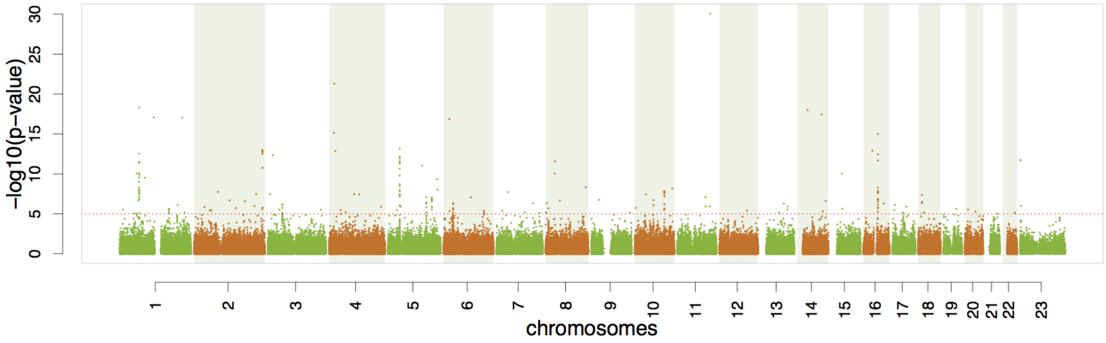


FIGURE 2.8 A *Manhattan* plot is used to identify regions where p-values are under the significance level. Each point represent a SNP (at its chromosomal position, chromosome by chromosome) and its associated χ^2 p-value. The red dotted line represents a significance level of $10^{-5} = 0.00001$, p-values under that threshold may pinpoint causal regions. On the CD_1 Crohn's disease related dataset (see Chapter 6), among others, noticeable regions are found on chromosomes 1, 5 and 16 (which have been confirmed as being involved in the disease).

as the original one but which are as or more extreme in their departure from the null hypothesis. One way of evaluating the probability to find such a table at random under the independence hypothesis is to use Monte-Carlo simulations of random permutations of the dataset.

2.4.3 A bit further

So far, single locus tests of association allowed for the identification of responsible genetic variations and the elucidation of many mechanisms underlying several diseases and traits. Despite many GWAS successes, only a moderate part of the heritability appears to be explained by those successful findings. Thus, posterior to the identification of a genomic region using a single point analysis, which does not directly indicate a true association, neither describe the exact cause and the real impacts of the causative biological mutation, many downstream approaches may be useful to further confirm and contribute to the deep understanding of the role of these candidate broken pieces of DNA. The following is a non exhaustive overview of possibilities (which can be combined):

- One approach would be to identify the model of the genetic disease, little modifications of the Table 2.2 allow to investigate different hypotheses using statistical tests of association such as described in Section 2.4.2. For example, if we suspect an allele at a given position to increase the disease susceptibility, Table 2.4 can be easily derived from Table 2.2 to which a test of association can be applied. Similarly, the recessive model (Table 2.5) can be investigated as well as other models of genetic diseases. This may help to detect which is the mutant allele and how it is associated to the trait of interest.

	AA	AB+BB	Total
Cases	a	$b+c$	n_{case}
Controls	d	$e+f$	n_{cont}
Total	$a+d$	$b+c+e+f$	n

Table 2.4 Dominant model: under that hypothesis, allele B increases the disease susceptibility

	AA+AB	BB	Total
Cases	$a+b$	c	n_{case}
Controls	$d+e$	f	n_{cont}
Total	$a+b+d+e$	$c+f$	n

Table 2.5 Recessive model: two copies of allele B are required to increase the disease susceptibility

- Replication studies allow to validate or invalidate the findings of an association study. In that case, the use of a different genotyping platform may help to discard spurious associations due to batch effects. In order to reduce the cost, a common practice for those replications is to target candidate genes previously identified at a genome-wide level. Denser sequencing around these regions may be useful to confirm the presence of a true association between the suspected loci and the phenotype under study.

- Alleles at different loci on the same chromosome within a gene may create a "super allele" (or haplotype) that has a larger effect than any of its alleles separately. Similarly to the single locus test, haplotype-based analysis compares haplotype frequencies/similarities between cases and controls. This type of test requires the genotype to be phased (phasing allows to identify the provenance of each allele, either it is on the maternal strand or the paternal one), unfortunately, most of the time, the output of genotyping platforms is not phased and it may be computationally intensive to reconstruct the haplotype structure at the genome-wide level.

- Part of the missing pieces [M$^+$09] might be hidden in potential gene-gene interactions. In [Ste12], the author reviews current trends in the field of gene-gene interaction analysis. In particular, the extension of the multifactor dimensionality reduction (MDR) called Model-based MDR (or MB-MDR) [CUV$^+$08] is a notable example for the investigation of gene-gene interactions in which the multi-locus genotypes are merged (in three categories: high risk, low risk and 'no evidence') in order to reduce the dimensionality (hence the name MDR) and increase the power to detect gene-gene interactions.

- Testing the potential interaction of SNPS with environmental factors (such as smoking habits, diet habits or stress exposure), the investigation of copy number variants (CNV) or the link between expression data and the corresponding genotypes at a given candidate gene, etc.

Actually, most of these approaches are, at the time of the writing of this manuscript, practically infeasible at a genome-wide level due to the growing amount of genetic variants typed and the resulting combinatorial "explosion". The majority of these methods tends to be applied after a first stage of filtering, on a (much) smaller amount of variables spotted by single locus analysis, implying that this subset of candidate SNPS harbors a significant marginal effect (which may not be a good assumption for fully epistatic interactions).

Chapter 3

Machines can learn

Contents

During the two last decades, the amount of data being collected in multiple domains (medicine, genetics, social networks, ...) has been massively increasing. As that volume of data increases, the part of it that people understand drops and the task of analyzing huge datasets has now become infeasible by standard approaches. Those facts render obvious the need to develop algorithms and statistical methods capable of extracting knowledge from an important volume of data, in other words, capable of discovering the underlying hidden but potentially useful information contained in large databases.

Fortunately computational power has increased in parallel with the amount of gathered data. In a given context, that computational power can be used to extract some relevant information from a set of observations. In this thesis, we focus on so-called supervised learning. In this case, the extracted information concerns relations/patterns between variables that could explain a studied outcome. Succinctly, the goal is to predict output labels for new objects given their inputs and a dataset of observed input-output pairs.

This chapter provides a general introduction to the field of supervised machine learning. The different stages of a supervised learning algorithm will be described as well as the associated vocabulary, notations and standard procedures used to evaluate the results of applying supervised learning methods to a dataset.

As stated by Moore's law, processor speed, storage capacity and sensor capabilities improve at exponential rates. They double approximatively every two years. Such advances in acquisition and storage technologies now allow for collecting a massively increasing volume of data at low-cost (time and money) investments. In many domains such as web mining, automatic image classification, diagnosis, medicine and marketing, it is now common to gather gigabytes, even terabytes, of data in very short periods of time.

The goal of supervised machine learning is to learn from labelled data. Such data can be viewed as a set of labelled entities of the same type described by a number of features, i.e., each entity is a point in a multidimensional feature space. The purpose of supervised learning is to find the optimal function of features that allows to predict as well as possible the labels.

3.1 Datasets and notations

Typically, in supervised learning, the data are composed by a collection of n objects described by p features (one also uses the terms *input-variables, attributes*, or simply *inputs*) which, basically, correspond to elementary measurements. These features may have some influence or provide some information on the label of each object (one also uses the terms *output-variable, target*, or simply *output*). When the output is a real number, the task of machine learning is also referred to as regression while if the output is discrete (binary or categorical) one talks about classification. In this thesis, the output will be binary, denoting the disease status of a patient, i.e. healthy (controls) or sick (cases), and will be represented by 0 or 1 respectively.

We will denote by X the space of input vectors of dimension p. Similarly, Y will denote the output space. Observed values are written in lowercase; hence the ith observed value in X for object j is written x_i^j. Likewise, the jth output object label is denoted by y^j. Thus, a dataset can be represented by a $n \times (p+1)$ matrix:

$$\begin{bmatrix} x_1^1 & \cdots & x_p^1 & y^1 \\ \vdots & \ddots & \vdots & \vdots \\ x_1^n & \cdots & x_p^n & y^n \end{bmatrix}$$

Along this manuscript, we will denote such a dataset by $\mathcal{DB}$. Generally, such a $\mathcal{DB}$ is a subset of all possible objects. Indeed, in most of the domains, it is impossible to collect all the existing objects from the studied population, even if the latter is in principle finite. In addition, depending of the field under study, gathering data can be a difficult task. For some reasons, features may be incorrect, missing or not easy to access or measure. For these reasons, most of the time, a $\mathcal{DB}$ is submitted to many quality checks and pre-processing steps prior to the application of machine learning techniques. In the sequel, when we refer to a $\mathcal{DB}$, we refer to the data matrix resulting from this pre-processing.

Finally, for genome-wide association studies, the features x_i^j (with $1 \leq i \leq p$ and $1 \leq j \leq n$) represent genetic variants measured all along the DNA of an individual (e.g. SNPs). The output represents the phenotype, indicating whether or not the individuals are suffering from the disease of interest. Typically, for such an analysis, the value of n ranges from a few hundreds to a several thousand while p can reach several hundreds of thousand (and even more). In later chapters, we will discuss and analyze the pre-processing steps typically applied to these latter problems.

3.1.1 Curse of Dimensionality

Nowadays, whether in biotechnology, finance, multi-media or social networking, we observe and face a growing amount of features for a limited number of objects. The curse of dimensionality refers to the fact that it is more and more common to deal with datasets containing n samples described by a very large number p of variables, leading to a really small n/p ratio. We will see that some machine learning methods are more appropriate for dealing with such a scenario than others. This is typically the case for genome-wide association studies where an individual can be described by a huge number of genetic variants.

3.2 The learning step

Once the data have been collected, checked and cleaned they can be used as learning samples for supervised learning. The basic goal of supervised machine learning is to infer from a learning sample composed of a number of objects described by their input variables and their output labels (figure 3.1), a statistical model able to predict the label of new objects (figure 3.2) based on their input values.

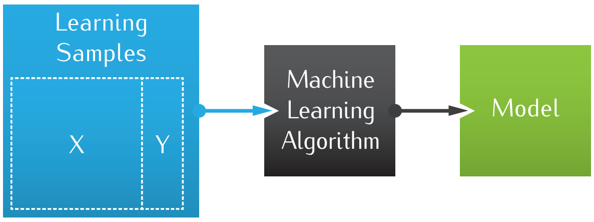


FIGURE 3.1 The learning step is the process of inferring a statistical model from a learning set.

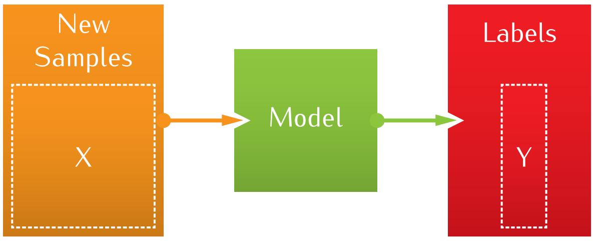


FIGURE 3.2 The common usage of a model is to predict the labels of new samples.

More formally, given a learning set LS of n objects (typically a subset of the available $\mathcal{DB}$):

$$\text{LS} = \{sample^j\}_{j=1}^n = \{((x_1^j, \ldots, x_p^j), y^j)\}_{j=1}^n, \tag{3.1}$$

drawn from some population of objects, the goal of machine learning is to find a function $h : X \to Y$ that on the average predicts as well as possible the value of y for any new object drawn from the population from which the $\mathcal{DB}$ was gathered.

Various additional assumptions have to be made in order to characterize this problem from the mathematical point of view:

- Specification of the *hypothesis-space* of candidate functions $H \subset Y^X$, within which the goal is to find the most accurate predictor. For example, if inputs are encoded as numerical values, one can use either linear or non-linear, parametric or non-parametric hypothesis spaces (see also Section 3.4 and Chapter 4 for some examples).
- Definition of a numerical criterion to measure prediction errors, typically via the choice of a loss function $\ell : Y \times Y \to \mathbf{R}^+$, and by defining the error of a function h as its expected loss over a so-called generative probability distribution $P(X, Y)$, i.e. $E_{P(X,Y)}\{\ell(h(x), y))\}$. For example, when the outputs are binary or categorical, one often uses the so-called 0-1 loss function, defined by $\ell(y, y') = 0$ if $y = y'$ and $\ell(y, y') = 1$ if $y \neq y'$; its expected value is equal to the probability of making a wrong prediction.

- Assumptions about the sample generation mechanism: often it is assumed that the LS is drawn *i.i.d.* (independently and identically distributed) from the same generative probability distribution $P(X, Y)$ used to define the average loss. This assumption is in particular useful in order to characterize the theoretical properties of supervised learning algorithms, such as large sample behavior (asymptotic analyses), and finite sample bias and variance, as a function of the complexity of the hypothesis space H.

One of the main expected qualities of a learning algorithm is its consistency: roughly, consistent behavior means that as the size of the learning sample increases the expected loss decreases, and eventually converges to the lowest possible average loss within the hypothesis space H. Statistical learning theory has established necessary and sufficient conditions for consistent behavior of learning algorithms, and also provides upper bounds on their asymptotic rate of convergence. However, in practice one is interested in the behavior of the used algorithm in finite sample conditions of realistic size. In this context, the theory unfortunately only helps us to understand qualitative behavior and is of little use to make quantitative a priori predictions as concerns the relative performances of different algorithms on a given problem. We will therefore not further elaborate on the theoretical aspects of supervised learning. We refer the interested reader to general text-books [HTF09, Vap98b]. In section 3.3, we discuss techniques for the sound empirical evaluation of models induced by machine learning, which may be used in practice.

Notice that from an algorithmic point of view, different training regimes exist and depend on how the learning samples are used to infer the hypothesized function h. The most common case corresponds to the *batch* mode, when the entire LS is available, all the training objects are used at once to infer h. Another possibility is the *incremental* mode where objects are added (randomly or not) one by one to modify the current hypothesis. A particular case of the incremental mode is the *online* method, as the objects arrive, they are incorporated to the current trained function h. Some methods will use the incremental (and online) mode while some will use the batch mode; some of them can be used in both modes with minor adaptations. In this thesis, we will only consider batch-mode supervised learning algorithms.

3.3 The model evaluation step

The resulting function h can be used as a predictor for new entities. But, prior to that, the accuracy of the corresponding model has to be evaluated. In particular, the following questions are relevant:

- How well the machine has learned ? Does h classify correctly the learning set itself ? Applying h to each training object and comparing the results with the known labels (Figure 3.3) will tell us how well (or how badly) h fits the learning set. This is called the **resubstitution error**.

- How well does h approximate the output variable y over the rest of the population of possible objects ? In other words, how well does the computed model classify objects that are not represented in the learning set. This is also called the **generalization error**.

A common practice is to find a compromise between these two types of errors. Indeed, while one can to some extent expect that the more accurate the model is on the learning sample the more accurate it should also be on the rest of the population, this is not always true. Indeed, if the goal is to minimize the **generalization error**, it is often counterproductive to target a minimal **resubstitution error**. In fact, this scenario is a common issue in machine learning, hereafter referred to as the **overfitting** problem. Indeed, most often a model will not be interesting if it perfectly classifies the training objects while poorly generalizing its classification performances to previously unseen objects.

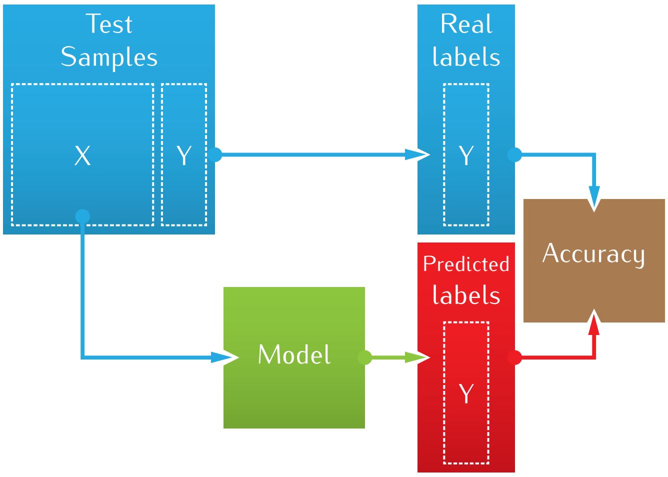


FIGURE 3.3 The evaluation of the generalization error of a model is achieved by a comparison between the predicted labels and the real ones.

3.3.1 Prediction versus Interpretation

Every supervised machine learning algorithm produces models. Those models are potentially useful for:

1. Making **predictions** for previously unobserved objects.

   In the context of GWAS, this **predictive** feature would allow one to assess the genetic risk of disease for a new patient, given his genotype.

2. Making **interpretations** of the underlying relations between the input and output variables that hold in the studied domain. The goal is to improve our knowledge of the studied field. One of the most relevant questions is to see what are the input variables that are the most linked with the studied output variable ? Are there irrelevant variables, what are the relevant ones ? Is there a particular threshold over a particular descriptor that is explaining the output or are there particular combinations among some variables that lead to a particular outcome ? All those questions are possibly of interest and answering them may help towards a better understanding of the situation/domain from which the dataset has been gathered.

   In the context of GWAS, this **interpretability** feature would help biologists to gain insight about the biological mechanisms involved in the disease.

The **interpretability** of a model refers to the possibility to explicitly confront the model with existing knowledge about the problem, to see whether it is coherent or not with this knowledge, and to infer from it new hypotheses that may be validated or invalidated experimentally or from first principles assumed to hold within the considered domain.

Some models have graphical representations which are easily readable. For those, a quick look directly gives a explanation on how the variables are connected to the outcome. Of course when dealing with large numbers of descriptors, graphical representations can become fuzzy and difficult to read. When no such view is available, numbers can still speak by ranking variables according their importances inside the predictive model. We will further elaborate on this aspect in Chapter 4.

3.3.2 Evaluation of the accuracy of binary classification models

In this section we describe several measures of accuracy that are used in practice for supervised learning problems when the target output is a binary variable, as these are relevant for our empirical evaluations in later chapters. In this context, we use the term *classifier* to denote the model inferred by supervised learning, and replace the target variable y by the term $class \in \{0, 1\}$.

Once a classifier $\mathcal{C}$ is built, evaluating its predictive capability can be achieved using an independent test set $\mathcal{T}$ of size n' (this test set is typically the part of the $\mathcal{DB}$ that was not used to create the learning sample):

$$\mathcal{T} = \{sample^j\} = \{(input^j, class^j)\} \qquad j = 1, ..., n' \tag{3.2}$$

The most standard and obvious way to proceed is to compute the **accuracy** which corresponds to the ratio between the number of objects correctly classified over the test set size:

$$Accuracy = \frac{\#\{sample^j : \mathcal{C}(x^j) = class^j, j = 1, ..., n'\}}{n'} \tag{3.3}$$

which is generally expressed as a percentage. That quality measure is the easiest to understand, but it has two drawbacks. Indeed, it relies on:

- the number of objects of each class represented in the test set: in some cases, it could be difficult to get well balanced datasets. One class may for example be strongly over-represented, and this situation would then lead to an average error rate mainly reflecting the rate of correctly classifying objects from this latter class.
- the decision threshold: most of the time, a learned model actually outputs a class-probability $\in [0, 1]$ for each input vector of features. In order to transform this into a class prediction, a threshold has to be chosen. For example, in the case of a binary classification task the common choice is to use a threshold of 0.5, but sometimes this choice might not be the optimal one.

3.3.3 ROC curves

To circumvent those two previous disadvantages, it is possible to characterize the type of each prediction given their real classes. In particular, for binary classification tasks, if we associate the positive class to 1 and the negative class to -1, when the classifier is predicting the good class, that class can be positive for positive samples or negative for negative samples. On the other hand, when the classifier is wrong, the prediction can be negative for positive samples or positive for negative ones. Those two types of errors are also known as the *Type I* (false alarm), *Type II* (miss) errors respectively.

Given these prediction characteristics, it is then possible to derive a contingency table (Figure 3.4), also called a confusion matrix, that will allow to compute metrics such as:

1. the true positive rate, equivalent with *sensitivity* or *recall* : $TPR = \frac{TP}{P}$,
2. the false positive rate : $FPR = \frac{FP}{N}$,
3. the accuracy : $\frac{TP+TN}{P+N}$ or the error rate : $1 - \frac{TP+TN}{P+N}$.

Of course, those metrics are still relying on a chosen decision threshold. For one decision threshold, we can compute the true positive rate and the corresponding false positive rate which gives a point in the *receiver operating characteristic* (ROC) space. Doing this for every possible decision cutoff produces a ROC curve. Figure 3.5 represents the ROC space where:

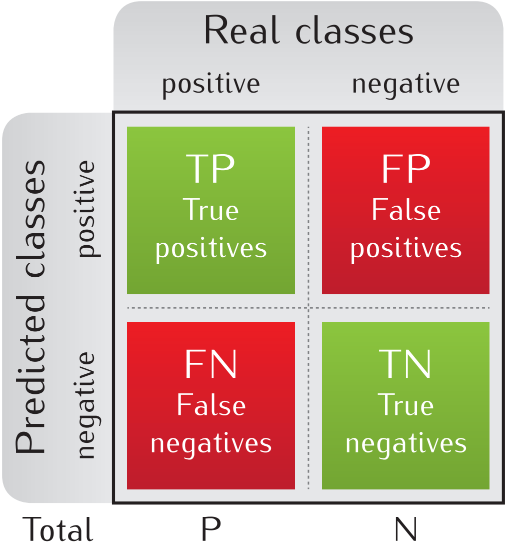


FIGURE 3.4 The confusion matrix: characterization of predictions given the real classes.

- point a at $(0,0)$ represents the cutoff 1, in other words all the predictions are negative,
- point b at $(1,0)$ represents the ideal situation where the true positive rate is maximal and the false positive rate is minimal,
- point c at $(1,1)$ represents the cutoff 0, all the predictions are positive,
- the dashed line represents a ROC curve for random predictions, the area under that curve is equal to 0.5,
- the orange line represents an example of ROC curve,
- the green shaded area represents the area under the ROC curve (AUC)

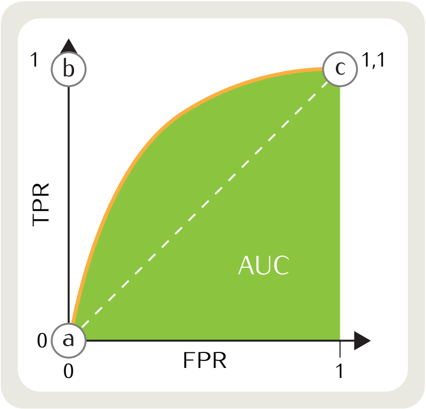


FIGURE 3.5 The important points in the ROC space and an example of ROC curve (in orange) and the corresponding AUC (in green).

The area under the resulting curve is called the AUC and gives a new quality measurement that is independent of the decision threshold and the class distribution. Given the scenario, it is possible to choose the right decision cutoff that would be an acceptable tradeoff between type I and type II errors. For example, in medicine, sometimes it will be preferable to tell a patient he is sick while he is not rather than the opposite. In other word, it will be better to minimize the type II errors. In practice the AUC varies between 0.5, for a model that classifies objects at random, to 1, for a model that is able to classify them perfectly. The

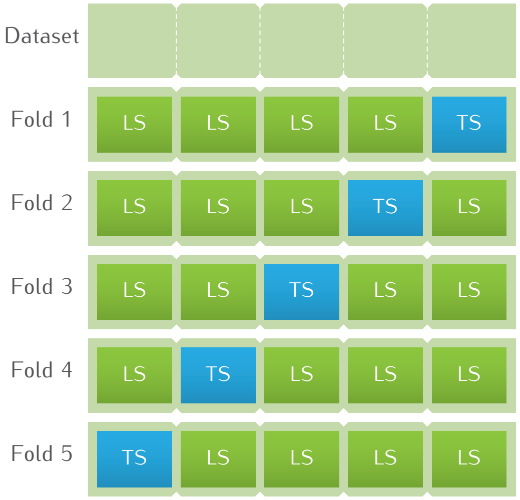


FIGURE 3.6 An example of a 5-fold cross-validation: the initial dataset is divided (randomly) into 5 subsamples. In turn, each fold is used as a test set (TS) to evaluate the learning algorithm applied to the full dataset minus the given TS.

AUC can also be interpreted as the probability that the classifier will assign a higher score to a randomly chosen positive sample than to a randomly chosen negative one. In our empirical investigations on synthetic and real-world GWAS datasets, we will mostly use the AUC as a criterion to evaluate the relative accuracies of models obtained in different conditions.

3.3.4 Cross-validation

In order to assess the predictive accuracy of a model inferred by supervised learning, the ideal situation would be to dispose of a very large test sample, independent from the learning sample used to infer the model, and then to assess its accuracy according to any one (or all) of the preceding criteria (from error rates to AUCs).

Given a dataset of classified objects, we would thus have to split this dataset into two parts, one used for learning and the other used for testing. When the number of available observations is limited, this will lead to even smaller learning and test samples, leading both to suboptimal models and inaccurate evaluations of its accuracy.

To circumvent this dilemma, machine learning researchers commonly use the so-called 'v-fold cross-validation' approach, which works as follows:

- First the overall available sample is divided (randomly) into v subsamples of (approximately) the same size (typical values of v are 5 or 10, depending on the conditions).
- Then, for each one of the v subsamples the following procedure is applied:
  - The supervised learning algorithm is applied to the full dataset minus the given subsample.
  - The accuracy (error rate, AUC, etc.) is evaluated on the subsample.
- The so obtained v values of the accuracy measure are averaged and used as an approximation for the accuracy of the model trained on the whole dataset.

One can show that the larger v (it is upper bounded by the size of the available dataset), the closer the expected value of the 'v-fold cross-validation' approach to the expected error rate of the model built on the whole sample, while small values of v lead in practice to pessimistic estimates. Since the computational complexity of this procedure is directly proportional to v, a value of $v = 10$ typically leads to a good compromise between accuracy and computing times. Figure 3.6 illustrate a 5-fold cross-validation approach.

Notice that since the resulting statistics depend on the random division of the sample into v subsamples, it is important to control this random effect, by using the same folds to assess different models built in different settings (e.g. different algorithms, or different subsets of input features).

3.4 Summary

In this chapter, we introduced the concept of supervised machine learning, the idea of statistical model inference from a dataset and the possible measures of evaluation of such model. It constitutes the basis of our research and we strongly believe that the overall methodology "package" is well suited to the field of genome-wide association studies. Indeed, there is a strong overlap between the two fields in the tasks they try to achieve. The first one is the identification of variables/SNPS of interest and the second one is the ability to exploit those descriptors to estimate the probability of an object/individual to be of a given class/phenotype. Also, the commonly achieved tasks in a GWAS are easily transposable to the field of supervised machine learning.

In the next chapter, we discuss the details of the tree-based supervised-learning methods and motivate why they constitute a good choice for GWAS.

Chapter 4

Tree-based methods for GWAS

Contents

In the present chapter we focus on the sub-class of supervised learning algorithms based on decision trees that we want to study and improve in the context of genome-wide association studies. We start by describing single trees, random forests, and extremely randomized trees which are state-of-the-art methods, and then we explain our own proposal called T-Trees. All along this presentation we try to provide at the same time intuitions about the methods and their algorithmic choices and precise descriptions of the algorithms that we have implemented and applied in GWAS studies.

4.1 Motivations

The nature of genome-wide association studies puts several constraints on candidate supervised learning methods. First of all, the method should help to identify the genomic regions which contain causal mutations which in isolation or in combination are associated with the studied phenotype. In the machine learning language, this implies that the method should be able to identify relevant variables (i.e. SNPS) among a very large number of irrelevant ones, and thus incorporate some feature selection mechanism. Also, given the possible epistatic effects, and the emerging hypothesis [CG10] that rare mutations could affect the phenotype of complex diseases as well as common variants, and given the various sources of correlations (physical linkage, and sampling artefacts), the method should not make any strong a priori assumption about the underlying relationship among genotype and phenotype and rather let the data speak for themselves to reveal this information. Specifically, when considering common complex diseases, we do not know in advance how many causal loci will eventually be revealed by our analysis, and we can't exclude that some mutations that are marginally benign could lead to increased risks when they appear together with others. Similarly we can't exclude that a significant part of the heritability is carried by a relatively large number of variants to rare to be detected by current GWAS univariate approaches.

From the computational point of view, given the very large number of candidate input variables considered in GWAS, and a growing number of cases and controls available, the method should as well be efficient enough, so as to allow its application in realistic scenarios, with millions of variables and thousands of individuals. In addition, in order to avoid cumbersome trial and error iterations, and to ensure reproducibility of results, the machine learning algorithms should be as far as possible 'off-the-shelf', in other words as much as possible free of meta-parameters that need to be tuned in a dataset dependent way.

Moreover, as we will see in our experimental studies, we want to be able to apply the supervised learning method in different settings corresponding to different subsets of input variables, e.g. by using subsets of SNPS based on their allele frequencies, and also with different scenarios corresponding to different ways of pre-processing the datasets. These kind of studies can indeed reveal sampling and preprocessing artefacts, and help to assess the relative importances of different subgroups of variables corresponding to different biological hypotheses about the genetic nature of a disease. The used methods must therefore be flexible and almost fully automatic.

Because of their intrinsic features that fit very well a priori with the above requirements, our research focuses on the application of tree-based classifiers to GWAS, and more specifically on variations around the random forest method. In the subsequent chapters, we will study their behaviour in simulated and real-life conditions. In the present chapter we describe the used algorithms in detail, so as to allow the reader to understand their mechanics, weaknesses and strengths, and so as to ensure reproducibility of our investigations.

The rest of the present chapter is organized as follows. We start, in section 4.2 by describing state-of-the-art methods from the literature, namely standard classification trees, random forests and extremely randomized trees. We customize the description of these methods to our practical context, namely discrete SNP-based input variables and a binary phenotype of type case/control. We then propose an adaptation of these ensemble methods which aims at taking into account the correlation structure of the genotypic information, that we called T-Trees, because it is based on a two-level combination of tree-based methods, the outer layer screening SNP blocks, and the inner layer exploiting the information inside a given block. We proceed by briefly discussing extensions of these algorithms to a broader class of GWAS, where phenotypes may be quantitative or multi-class, and input variables could comprise other than genotypic information, such as environmental factors, although we did not implement these extensions yet. Finally, we provide a discussion of related works from the literature on machine learning applications to GWAS, mostly focusing on the use of tree-based methods.

4.2 State-of-the-art in tree-based supervised learning

In this section we describe the state-of-the-art in ensembles of tree-based methods used in our investigations. While, for reasons that will become clear from our explanations, we do not use single trees in our study, we nevertheless start with a detailed discussion of this method, since it is at the core of these ensemble methods and also of our own proposal described in the subsequent sections.

4.2.1 Single Decision Trees

Basically, a (binary) decision tree is a series of (binary) questions which can be programmatically schematised as nested **`if-then-else`** statements and represented as an acyclic graph directed outwards from a root-node, and having the following properties:

- each node can have zero (a leaf) or two descendants (a test node)
- each node has one parent, except for a single node called the *root* which has no parent.

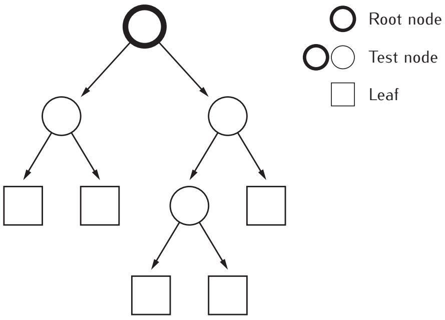


FIGURE 4.1 A directed tree is an acyclic graph directed outwards from a root-node

Figure 4.1 generically illustrates those different types of nodes of a binary directed tree. The directed edges refer to the link from a parent to one of its children. Such a structure is called a full binary tree as each test node has exactly two children. Notice that the total number of nodes of a binary decision tree is equal to the sum of its number of test (or internal) nodes and of its number of leaves (or terminal nodes); since the number of leaves is always equal to the number of test nodes plus one, the total number of nodes is an odd number equal to $2c+1$, where c is the number of test nodes also called tree *complexity*, which is also equal to half the number of edges of the tree. In addition to the number of test nodes, sometimes one also uses another parameter to measure the size of a tree, namely its maximal (respectively average) *depth*, which refers to the maximal (respectively average) length of a path from the root to a leaf of the tree. In our example of Figure 4.1, we have $c = 4$, a maximal depth of 3, and an average depth of $(2+2+3+3+2)/5 = 12/5$. Note that a tree is said to be (almost) *balanced* if its average and maximal depths are (almost) equal. Notice that, while one can also use non-binary decisions trees, in supervised machine learning one generally restricts to binary ones, for reasons that will be explained later on.

Figure 4.2 provides an example of a decision tree for a prediction task. In this case the test nodes are decorated with questions comparing an input feature to a threshold, while the leaves are associated with information about the output target variable (here a simple "yes/no" label). This graphical representation is often easy to understand and interpret by human experts.

A decision tree can be translated into a nested set of 'if-then-else' statements defining an algorithm for carrying out predictions based on the tree. For our example, this is illustrated in Algorithm 1, and we observe that these rules are easy to interpret: reading them allows us to immediately grasp how predictions are linked to input variables. From a practical point of view, the more complex a decision tree is, and the

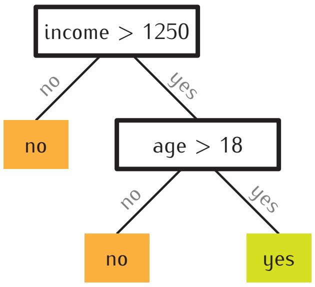


Figure 4.2 An example of a simple binary decision for a hypothetical credit card allocation problem.

more complex the set of 'if-then-else' rules and the predictor h it implements are, and the more difficult it is to interpret this function.

```
if $(income > 1250)$ then
    if $(age > 18)$ then
        yes
    else
        no
else
    no
```

Algorithm 1: Translation of the decision tree of Figure 4.2 in a cascade of `if-then-else` statements.

From a statistical point of view, as we will see, the complexity of a tree is directly related to the number of degrees of freedom of the corresponding model. We can thus anticipate that the more complex a tree, the larger the number of samples necessary to validate its overall significance.

The stakes of supervised learning of (binary) decision trees

Supervised learning aims at automatically inferring a decision tree based on a learning sample. To explain the rationale behind this algorithm, we will discuss a simple supervised learning problem.

Let us consider a bank client listing as the one given at Table 4.1, which we chose very small for the sake of legibility, and let us consider these data as a learning set.

client number	age	monthly income	credit card
001	17	550$	*no*
002	19	900$	*no*
003	21	1500$	*yes*
004	16	1200$	*no*
005	28	1550$	*yes*
006	32	2000$	*yes*
007	25	1150$	*no*
008	34	2900$	*yes*

Table 4.1 Example of data for a credit card marketing problem

The last column of Table 4.1 is the output we want to learn. Our goal is to answer the following question: “Should we spend budget and time to propose a credit card to a client?” The clues that our decision tree should use to answer this question are the two descriptors of the person corresponding to the second and third columns: its age and its monthly income. Indeed, at this stage we can a priori exclude the first column as irrelevant, since we don't normally want to infer a decision strategy based on client number. Note however, that in another context we might be interested in assessing whether or not the client number is correlated to the decisions recorded in the dataset, which could raise interesting questions about the quality of our dataset and/or about our past decisions.

By examining the learning set, we can check that the decision tree shown in Figure 4.2 actually is a consistent explanation of the record of past situations, since its predictions are perfectly fitting the target output over the learning sample. We will say that the tree perfectly fits this dataset, and if we consider that the dataset is also representative of future conditions of usage of the tree, we may be tempted to accept it as a potential decision strategy, and if we consider that the dataset indeed reflects our past experience, we can infer that so does also the decision tree. In this particular case, we might also be attracted by the fact that the interpretation of the rule is conform with our intuition, namely that consumers with small income or that are very young are unlikely to purchase a credit card.

To get further insight, let us depict a geometric interpretation of our problem. To this end, we observe that each object of the table 4.1 may be viewed as a point in a two dimensional space spanned by the variables "age" and "monthly income". These points are plotted on figure 4.3, where the dashed line is the decision boundary that separates the clients asking for a credit card (in green) from those who do not (in orange), as expressed by the decision tree of Figure 4.2.

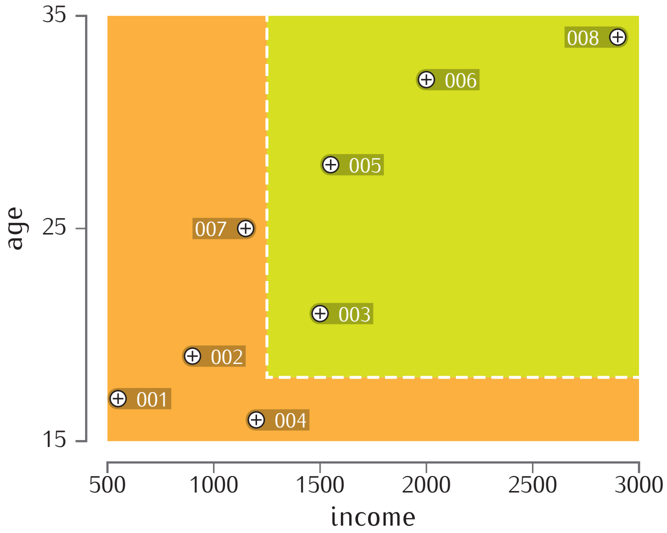


FIGURE 4.3 The decision rule in a two dimensional space

Notice that from this graphic, we may also infer that actually the two subsamples labelled respectively "yes" and "no", could as well be perfectly separated by a simple threshold of say 1250$ on the "income" feature, and almost perfectly by a simple threshold of say 18 on the "age" feature. So, we now end up with three different explanations of our past observations, and the question is which one is the best. Intuitively, it is difficult to compare the first explanation (involving the two features but yielding a perfect fit to the learning sample) and the third one (which involves only one feature but is not perfect anymore on the sample). However, the second rule dominates these two rules, both in terms of data fit and complexity, so that it looks at this stage as being the best explanation among the three alternatives, and maybe the most effective decision rule to use for our future marketing campaigns.

From our discussion, we may infer that good decision trees are those that are both simple and fit well to the dataset, but that it is not so easy to a priori define the correct compromise between these two criteria. We have also seen that for a given level of fit on the training set, it is in general possible to formulate a number of alternative explanations that are possibly of different complexity. For a given level of fit, we would certainly prefer the simplest tree, but it turns out that from a computational point of view solving this problem is so-called *NP-hard* [HR76], and hence out of reach in any practical setting (as soon as the number of features and observations are larger than a few tens). Conversely, for a given complexity, we would prefer the tree that fits best to the learning sample, but solving this problem is again *NP-hard*.

So even if we could a priori guess the right level of fit or of tree complexity for a given problem, finding a good tree in these conditions has to be based on heuristics and possibly suboptimal algorithmic approaches. Furthermore, the optimal level of complexity of a decision tree turns out to be highly problem specific, for example if the features are effectively unrelated to the target output we would like a supervised learning algorithm to tell us this by producing a "trivial" tree composed only of its root, while for a problem where some of the inputs are indeed correlated to the output we would like a tree to tell us this information, but if the correlation structure is intricate we do not expect to be able to achieve this goal unless the dataset is of rather large size.

To summarize our discussion, supervised learning of decision trees leads inevitably to solving a dataset specific compromise between complexity and learning set fit, and can't be solved in a computational efficient way without resorting to heuristic algorithmic approaches. Research in single decision tree induction has started in the early 1960's [MS63, Hun66, HJF69], and was addressed by researchers with very different perspectives (survey analysis in sociology, questionnaire design, artificial intelligence, and last but not least computational statistics [Fri77, Qui83, KBR84]). This work has explored many different alternatives but eventually culminated with the publication of the book on *Classification and Regression Trees* [Bre84] in the mid 1980's. In the next subsection we describe the resulting algorithm. We will see that it is not only interpretable from the viewpoint of the results that it computes but also from the viewpoint of its algorithmic behavior. Later on we will see that this latter kind of interpretability is quite important to help researchers to produce novel algorithms overcoming the basic limitations of this method.

Overall principle of top-down induction of decision trees

The standard strategy for supervised learning of decision trees uses the available dataset in order to build the tree in two steps, i.e. by first growing an overly complex tree and then pruning it to the right level of complexity [Bre84]. Tree growing uses a top-down divide and conquer strategy in order to build a tree of small complexity fitting very well the learning sample. Tree-pruning typically uses an independent test sample, in order to simplify the tree in a bottom-up fashion as much as possible and so as to maximize its accuracy on this independent test-sample.

The goal of tree growing is to divide the attribute space (in our example the plane spanned by 'income' and 'age') into an as small as possible number of regions (in our case rectangles) which contain essentially only samples of a single class. In a nutshell, the method starts with a trivial tree composed only of its root and the complete learning sample: it then tries to split the learning sample by finding a test (or a question) based on one of the input features, in such a way that objects of different classes correspond as much as possible to different outcomes of the test. Once this test has been found, the method splits the learning sample in two subsamples corresponding to the two possible outcomes of the test and proceeds by recursively building the corresponding subtrees based on these subsamples. Notice that before deciding to expand any node the algorithm verifies whether or not the current node should or not become a leaf of the final tree.

Thus, the three key ingredients for growing a decision tree are the following.

1. Definition of a set of candidate splits: based on observations, we need to find a "question" that will partition the learning set in two subgroups. In the standard method, a split is defined by choosing an input feature and a question about its values that will divide the current learning set in two parts. Depending on the nature of the feature (discrete or continuous) different types of splits can be defined. For numerical features with k different values there are $k-1$ possible splits and for categorical variables with i categories there are $2^{i-1}-1$ possible splits.

2. Evaluation of the splits: a score measure has to be used in order to decide which is the best question to ask, what is the best feature to use at the current node and with which threshold. Since the idea is to generate the purest learning subsets in terms of the output labels, the basic principle is to measure the purity improvement made by any candidate split, yielding score measures based on the class purity improvement between the sub-sample reaching the current node and those of its resulting two sub-sub-samples.

3. Deciding under which conditions should a node become a leaf: the stopping rules will ensure that the tree has a finite number of nodes. Different rules exists, some of them are data driven while the others are user defined:

    - **data driven** :
        - a node is pure: if all the objects in a node are of the same class, the fit is locally perfect so that no additional split should be made.
        - identical value for each variable: sometimes, objects may have identical descriptors while being of different classes. In that particular case, it is not possible to find a split that will improve the fit.
    - **user defined**: these are essentially pre-pruning variants to limit the size of a tree and avoid overfitting problems. Several criteria have been proposed in the litterature:
        - limit the maximum depth of a tree
        - limit the number of test nodes
        - limit the minimum number of objects at a node required to split
        - do not split a node when, at least, one of the resulting subsample sizes is below a given threshold
        - stop developing a node if it does not sufficiently improve the fit.

The above framework captures the large majority of tree growing algorithms which have been proposed in the literature. A simplified version of the resulting recursive procedure is depicted in Algorithm 2, where A_{cand} denotes a subset of the input attributes used in the particular application context of the method. The tree induction algorithm is recursive, starting from the root node, the left and right child nodes are created and then expanded by calling the algorithm on their corresponding subsamples. At some point, if the local learning set becomes pure or all the attributes constant, the algorithm stops creating children and the node becomes a leaf. Those leaves are typically labelled with the majority class present in the sub-sample of the learning set reaching the leaf, or if several classes are present in the corresponding subsample by a vector of relative class-frequencies, as suggested in our pseudo-code.

Notice that this simplified procedure continues splitting until no sensible additional split can be found, i.e. produces what we will call in the sequel a fully developed tree. In practice this leads to overly complex trees, typically overfitting the learning sample. Therefore, the tree growing procedure is in practice completed by a so-called tree-pruning method: roughly this consists in generating from the grown tree a sequence of shrinking trees obtained by successively replacing test nodes by leaves, by assessing the accuracy of each one of these trees on an independent test set (or by using a cross-validation technique), and by eventually selecting among the pruned trees one that yields an appropriate compromise between complexity and accuracy. Since in this work we will not use these tree pruning methods, we refer the interested reader to the literature for more information about these methods [Bre84, Min89, WA93].

```
BuildDecisionTree
input  : $\mathcal{LS}$, $A_{cand}$
output: A tree : $T$                    // Its root node

if const(attributes) or const(output) then
    return a leaf labeled by class frequencies in $\mathcal{LS}$
else
    Select all attributes $\in A_{cand}$ : $\{a_1, ..., a_p\}$
    $\{s_1, ..., s_p\}$ : $s_i = $ pickOptimalSplit($\mathcal{LS}, a_i$)
    $s_* = max_{i=1,...,p}$Score($s_i, \mathcal{LS}$)
    Split $\mathcal{LS}$ into $\mathcal{LS}_{left}$ and $\mathcal{LS}_{right}$ according to $s_*$
    $T_{right} \leftarrow$ BuildDecisionTree($\mathcal{LS}_{right}$,$A_{cand}$)
    $T_{left} \leftarrow$ BuildDecisionTree($\mathcal{LS}_{left}$,$A_{cand}$)
    return createNode($s_*$,$T_{left}$,$T_{right}$)
```

Algorithm 2: The recursive algorithm for growing fully developed decision trees from a dataset.

Implementation details used in this thesis

In order to be precise, in this section we describe the specific choices we have made in our implementation of single trees, used within the methods described in the sequel.

- **Evaluation of the splits** – The score measures we will use in our experiments are based on the well-known logarithmic or *Shannon* entropy. Let t denote a test outcome at a $node$ of a decision tree and c the class which we are trying to predict, t and c are two discrete random variables of respective distribution $(p(t_1), ..., p(t_k))$ and $(p(c_1), ..., p(c_m))$ (in our case, as from a machine learning point of view a GWAS is a binary classification problem and the decision trees are binary trees, $m = k = 2$). Basically, the class entropy allows to measure the impurity at a given $node$:

$$\mathbf{H}_C(node) \triangleq -\sum_{i=1}^{m} p(c_i)\log_2 p(c_i) \tag{4.1}$$

$$\triangleq -p_{case}\log_2 p_{case} - p_{control}\log_2 p_{control} \tag{4.2}$$

  where p_{case} (where $p_{control}$) correspond to the proportion of cases (controls) reaching the current $node$. Similarly the test entropy is defined as follows:

$$\mathbf{H}_T(node) \triangleq -\sum_{j=1}^{k} p(t_j)\log_2 p(t_j) \tag{4.3}$$

$$\triangleq -p_{left}\log_2 p_{left} - p_{right}\log_2 p_{right} \tag{4.4}$$

  where p_{left} (p_{rigth}) denotes the proportion of objects propagated to the left (right) at the current test node. Also we can define the average conditional entropy of the class given the test:

$$\mathbf{H}_{C|T}(node) \triangleq -\sum_{i=1}^{m}\sum_{j=1}^{k} p(c_i, t_j)\log_2 p(c_i|t_j) \tag{4.5}$$

Thus, a score measure can be defined as follow:

$$\begin{aligned}
\mathbf{score}(node) &\triangleq \mathbf{H}_C(node) - p_{left}\mathbf{H}_C(node_{left}) - p_{right}\mathbf{H}_C(node_{right}) && (4.6)\\
&\triangleq \mathbf{H}_C(node) - \mathbf{H}_{C|T}(node) && (4.7)\\
&\triangleq \mathbf{I}_C^T(node) && (4.8)
\end{aligned}$$

and corresponds to the difference of the current node impurity and the weighted impurity of the two resulting child nodes. It reflects the goal of the tree induction which aims to reduce the impurity at each test node. The split that maximizes such score is the one that reduces the more the class entropy from one node to its descendants. It is also called the mutual information I_C^T and it quantifies the reduction of the uncertainty of c given t. As this information quantity is upper bounded by the prior entropy $\mathbf{H}_C(node)$, that measure is sensitive to the number and prior distribution of classes rendering it difficult to interpret. Also in the context of decision tree induction, it has been observed to favor tests at a node with a larger number of outcomes. For these reasons, various normalizations have been introduced and are discussed in details in [Weh96]. Two of these normalizations will be investigated in our work. Equation 4.9 defines the "gain ratio" introduced by Quinlan which aims to reduce the bias towards tests with many successors:

$$\mathbf{Q}_C^T \triangleq \frac{\mathbf{I}_C^T}{\mathbf{H}_T} \tag{4.9}$$

However, low $\mathbf{H}_T$ could lead to an overestimate of the value of a split. In the literature [Tor01], this issue is called the "end-cut" preference of the "gain ratio" criterion as, for ordered attributes, it tends to be maximized at extreme cutoff values.

Second, Equation 4.10 defines a symmetrical (in C and T) score measure:

$$\mathbf{S}_C^T \triangleq \frac{2\mathbf{I}_C^T}{(\mathbf{H}_C + \mathbf{H}_T)} \tag{4.10}$$

In our experiments, we will investigate three variants of score measure ($\mathbf{I}_C^T$,$\mathbf{Q}_C^T$ and $\mathbf{S}_C^T$) and discuss their impact on the results. In practice, as for the χ^2 test of association, scores are computed from a contingency table. For example, if e samples reach a given node, where we test SNP$_{345}$ against the threshold 0.5 (in other words, if we separate the homozygous wild samples (0) from the heterozygous (1) and homozygous mutant ones (2), which corresponds to the dominant model), it is possible to summarise this test in Table 4.2, where a (c) represents the numbers of homozygous wild controls (cases), b (d) represents the number of heterozygous and homozygous mutant controls (cases). The proportions p_{left}, p_{right}, p_{case} and $p_{control}$ are then used in combination with the rest of the table to compute our different score measures.

	SNP$_{345} > 0.5$	SNP$_{345} \leq 0.5$	
Controls	a	b	$p_{control} = \frac{a+b}{e}$
Cases	c	d	$p_{case} = \frac{c+d}{e}$
	$p_{left} = \frac{a+c}{e}$	$p_{right} = \frac{b+d}{e}$	e

TABLE 4.2 Similarly to the χ^2 test of association, the different score measures are based on a contingency table. In this example, e samples are reaching a node at which SNP$_{345}$ is tested against the dominant model.

- **Complexity control** – In some situations, it may be useful to prevent a node from being further splitted. In the following, we mainly use two types of complexity control parameters:

  - N_{min}: this (user defined) number corresponds to the required minimum number of objects (i.e. local sub-sample size) reaching a node for it to continue splitting. For example, setting N_{min} to $n-1$ will produce a one-level decision tree with its root node directly connected to two leaves (this type of tree is also called a *decision stump*). Practically, a simple condition is added to check whether or not the learning set is big enough to create a new split. The typical default value for N_{min} is 2; meaning that the tree is fully developed.

  - N_{node}: this limit corresponds to the maximum number of test nodes allowed in a tree. Similarly, setting N_{node} to 1 will produce a decision stump. In order to obtain a fully developed tree, N_{node} is set to $+\infty$. At this point, from an algorithmic point of view it is important to notice that the nodes are developed in a given order which is implementation dependent. For example, in the Algorithm 2, we see that the right node is always the first one being expanded (and recursively, its right child is also being developed first). Now, imagine that we know our tree will be balanced (maximum and average depth equal 3) and that we use a $N_{node} = 3$ to limit the number of nodes in the induction of such a tree, we end up with a highly unbalanced tree looking like a linked list also called a degenerate tree (see Figure 4.4). That is why we must control the order in which the nodes are developed when we use a limit over the maximum number of nodes allowed in a tree. In that case, in our implementation we choose to randomise the order in which nodes are expanded. In our method proposal, we will see that this configuration will be particularly useful to produce *weak learners*.

- **Labeling the leaves** – In a fully developed tree, (most of the time) the objects reaching one leaf are all of the same class, in that case, we say that the terminal node is pure. One approach would be to label the leaves with their corresponding class, but, as explained before, if we stop earlier the induction of the tree, more than one class may be represented in a leaf. Thus, another approach would be to label the leave with the majority class. The propagation of an object through the tree will lead to labeling it with the majority class. Yet another approach would be to keep in the leaves the proportion of objects of each class that reached that terminal node. That proportion would somehow reflect the confidence of the corresponding prediction. For example, an object reaching a terminal node with 98 cases and only 2 controls is more susceptible to be classified as a case than another object that arrives in a leaf with 32 cases and 25 controls.

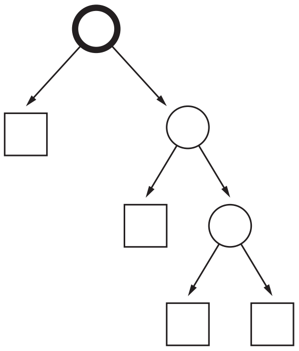


FIGURE 4.4 A degenerate tree is a highly unbalanced tree looking like a linear graph. Each test node has only one test node as child node.

Feature selection ability and input feature importance measures

In tree-based approaches, the feature selection mechanism, i.e. the ability to identify from a large set of candidate attributes the maximal subset of relevant ones, is said to be embedded into the methods. Indeed, in the particular case of single decision trees, at each node, a full scan of the feature space is performed in order to select the optimal split (the one that locally maximizes the reduction of entropy), thus, irrelevant features will naturally be discarded.

Also, single decision trees with a small number of nodes are easily interpretable due to the graphical representation. Now that we know how the tree is inferred, it is easy to understand that the tests that are appearing closer to the root are the ones carrying the most valuable part of the information regarding the output. Indeed, at each node we maximize the score and near the root we do that on bigger parts of the learning set. On the other hand, tests at the bottom of the tree are less informative as they concern a smaller part of the dataset from which the tree was learned.

However, in many applications, the tree complexity increases rapidly rendering its graphical representation difficult to read and understand. Still, numbers can speak and should allow to rank the variables according to their respective "contributions" or importances in a tree. Such numbers should reflect how well that variable helped to reduce the impurity of a node during the learning stage. Intuitively, such a "variable importance" measure should give more credit or weight to a variable that is used near the root while according less importance to the ones used at the bottom of the tree.

It is then possible to use the mutual information for that purpose. For each variable x_i used (maybe more than once) in a decision tree, we compute its importance as follow:

$$\mathbf{V}_{imp}(x_i) = \sum_{n \in Nodes(x_i)} p_n \mathbf{I}_C^T(n) \tag{4.11}$$

where $Nodes(x_i)$ is the set of tree nodes where the variable x_i is used to split, p_n denotes the relative sample size of node n, and $\mathbf{I}_C^T(n)$ is the local reduction of entropy resulting from the selected split at this node.

Doing so, variables appearing in many and "bigger" nodes (i.e. closer to the root) should be more important than the other ones.

Discussion

Single decision tree induction is the core of our research which aims to apply machine learning tree-based methods for GWAS. The main advantages of this method are as follows:

- simplicity: the method is essentially a parameter free "plug-and-play" method. It can be applied with no prior knowledge on the nature of the supervised learning problem, the combination of tree-growing and tree-pruning techniques ensures the asymptotic consistency;
- algorithmic efficiency: the learning stage is linear in the number p of input features and between $O(n \log n)$ and $O(n^2)$ in the sample size, and in practice typically $O(n \log n)$. While the testing step complexity ranges from $O(0)$ to $O(n)$ and typically is $O(\log n)$[1];
- interpretability: the method incorporates mechanisms of feature selection, an easily interpretable graphical representation and variable ranking capabilities which are all natural by-products obtained without significant computational overhead;

Nevertheless, single decision trees as such do not meet the requirements for a genome-wide association study because of their limitations:

[1] The big O notation is used in Computer Science to describe the complexity of an algorithm.

- the number of features used by the model is limited to the size of the learning sample, especially for the GWAS where the ratio n/p is really small and the number of relevant variables is expected to be large, a single decision tree may not be able to exploit all the features that are associated with a disease and may lead to a suboptimal statistical model;
- some score measures (such as $\mathbf{Q}_C^T$ and its "end-cut" preference), used in the context of single decision trees, are suffering from a pathological behavior when dealing with almost constant features or, in the GWAS context, rare variants;
- high variance, leading to low accuracy in generalization and jeopardizing to some extent the interpretability;
- for a given implementation, the induction of a tree is deterministic meaning that, for a given learning set, the corresponding inferred tree will always be the same. In consequence, in the presence of large numbers of strongly correlated variables (i.e. variables carrying the same information), only a few of them will be effectively used, potentially hiding its surrogates in the variable importance ranking.

4.2.2 Random Forests

In this section we explain the rationale behind the random forest method, and then describe the algorithm implementation used in our practical studies carried out in the subsequent chapters, and discuss its three main meta-parameters.

Historical notes and motivation

Single decision trees are subject to several limitations, and in particular a (very) high variance which makes them often suboptimal in practical applications. Driven by this fact, a standard technique for reducing the variance of a machine learning algorithm was proposed in the early nineties by Leo Breiman [Bre96], since then called *Bagging*. The term *bagging* stands for *b*ootstrapping and *agg*regat*ing*: instead of building a single predictor (in our case a single decision tree) the method generates an *ensemble* of predictors by bootstrapping over the learning sample and then aggregates their predictions, in the following way:

1. Generate T randomized versions of the learning sample, by sampling randomly with replacement n objects from the initial learning set. For each one of these T so-called *bootstrap copies* of the learning set, use a supervised learning algorithm (in our case a classification tree growing method).
2. In order to predict the output of a new case, use in turn all the T built models to get as many predictions (in our case each tree provides an estimate of the conditional probability of the output classes, given the input feature values of the new case), and then aggregate these predictions. In practice there are two different ways, one that we call 'soft' voting where the prediction becomes the average class-probability, and one called 'majority' voting where the prediction becomes the relative number of times, among the T predictions, where a given class was of majority probability, i.e. higher than 0.5 if we have only two classes.

Breiman showed that the resulting ensemble model has a smaller variance than the original supervised learning method (called base learner in this context), and that the variance reduction effect is proportional to the number T of ensemble terms. When the base learner consists of growing fully developed trees, the method leads hence to a very strong reduction of variance, and typically at a price of only a very moderate increase in bias [Geu02], so that in the end, the resulting model is typically much more accurate than a single tree built on the original learning sample. Bagging does not improve or fundamentally change the asymptotic properties with respect to those of the used base learner, but it leads in practice to much better small sample behavior in terms of accuracy, essentially at the price of an increased computational budget, since instead of a single run on the full dataset, the supervised base learner is used T times (in practice $T \in [10, 10000]$). The nice feature of this algorithm is that it is fully generic and *any-time*: it can be applied to any base learner, yields monotonic improvement with the number of terms T and can be interrupted at any-time to produce a classifier with a ensemble generated at this stage. Moreover the algorithm may benefit from straightforward parallelization, since each individual model of an ensemble may be learnt independently of the others.

During the same period of the early 1990's other researchers investigated the idea of building ensembles of tree-based models by transforming explicitly the deterministic tree-growing algorithm into a stochastic method, e.g. by locally or globally randomizing the choices of the algorithm, and in particular the subset of input features exploited [Ho98, Die00]. Leo Breiman realized the interest of connecting the theoretical analysis of these methods with bagging of tree-based models, and proposed the *random forest* as a synthesis of the two ideas [Bre01]. Since then this method has been considered as one of the most effective supervised learning methods on the shelf [HTF09]. As we will see, not only it is generically more effective than bagging from the accuracy point of view, but it also has strong advantages in terms of computational complexity, specially in the context of very high dimensional input spaces covered by many correlated features.

Our implementation of the random forest algorithm

In order to randomize the tree induction algorithm, Breiman proposes two levels of randomization:

1. **tree level**: to build each tree of the ensemble, a bootstrap copy of size n is drawn randomly with replacement from the learning set.
2. **node level**: at each node, instead of a search for the optimal split among all the features, only a random subset of K features is investigated ($K \in \{1, \ldots, p\}$ where $p = \#A_{cand}$).

That method introduces two meta-parameters :

1. **T the total number of desired trees in the forest**: the choice of T is essentially driven by time/computation limitation. Indeed, theory and empirical results show that the larger T the better. Of course, given the data, after a certain number of trees in the ensemble, the results are expected to converge. Thus, when it is possible, one recommendation to follow is to build trees until the error rate measured on an independent test set (or via any other unbiased estimation procedure such as cross-validation, or out-of-bag estimates [Bre01]) no longer changes.
2. **K the number of tested variables at each node**: the choice here is dependent on the nature of the problem. If we know that many variables are relevant, a small value of K would be a good choice, on the other hand, when only a few descriptors are informative, a large value of K would be well suited. Still, in most of the cases, it has been observed that $K = \sqrt{p}$ is often a good choice and will produce near-optimal results (in the context of classification trees).

The top-level iteration of the random forest method is stated in Algorithm 3, where we introduce one more parameter: $N_{\min}$. This parameter is used to limit the complexity of the individual trees composing the ensemble model. The sub-routine used within the random forest method to grow individual tress is described in Algorithm 4. It is quite similar to the algorithm 2 except that K random attributes are selected at each node instead of all of them and in that it uses the "pre-pruning" parameter $N_{\min}$ representing the minimum number of samples at a tree node to allow splitting of that node. Fully grown trees correspond to $N_{\min} = 2$; higher values of $N_{\min}$ allow to reduce the complexity of the method (and the resulting trees) and may be either favorable or detrimental in terms of accuracy (in practice , in problems where the features are providing complete information about the output smaller values are better, and vice-versa, if residual uncertainty is present, higher values are better).

```
BuildRFEnsemble
input  : $\mathcal{LS}$,$A_{cand}$,$T$,$K$,$N_{\min}$
output: A tree ensemble $\mathcal{T} = \{t_1, ..., t_T\}$

for $i \leftarrow 1$ to $T$ do
    $\mathcal{LS}_i \leftarrow$ Bootstrap($\mathcal{LS}$)
    $t_i \leftarrow$ BuildRFTree($\mathcal{LS}_i$,$A_{cand}$, $K$, $N_{\min}$)
    Append($\mathcal{T}$,$t_i$)
return $\mathcal{T}$
```

Algorithm 3: Random Forest algorithm: building an ensemble of trees over bootstrapped versions of the learning set. Randomization at the tree level in red; meta-parameters in blue

```
BuildRFTree
input  : $\mathcal{LS}$,$A_{cand}$,$K$,$N_{\min}$
output: A tree : $t$                              // Its root node

if const(attributes) or const(output) or $\#\mathcal{LS} \leq N_{\min}$ then
    return a leaf labeled by class frequencies in $\mathcal{LS}$
else
    Select $K$ random attributes $\in A_{cand}$ : $\{a_1, ..., a_K\}$
    $\{s_1, ..., s_K\}$ : $s_i$ = pickOptimalSplit($\mathcal{LS}, a_i$)
    $s_* = max_{i=1,...,K}$Score($s_i, \mathcal{LS}$)
    Split $\mathcal{LS}$ into $\mathcal{LS}_{left}$ and $\mathcal{LS}_{right}$ according to $s_*$
    $T_{right} \leftarrow$ BuildRFTree($\mathcal{LS}_{right}$,$K$,$A_{cand}$,$N_{\min}$)
    $T_{left} \leftarrow$ BuildRFTree($\mathcal{LS}_{left}$,$K$,$A_{cand}$,$N_{\min}$)
    return createNode($s_*$,$T_{left}$,$T_{right}$)
```

Algorithm 4: Random Forests: sub-routine for building one tree by randomizing the choice of features at each node. Randomization in red and meta-parameters in blue.

Once the ensemble of trees has been learned, predictions are made by combining those of each tree of the forest. As suggested by Figure 4.5, an object is propagated into each tree, leading that object to T different leaves. The probability vectors $\mathbf{v}_i$ associated to these T leaves are averaged as follows:

$$\mathbf{v} \quad = \quad \frac{1}{T}\sum_{i=1}^{T}\mathbf{v}_i \tag{4.12}$$

and the prediction for the propagated object becomes the average class-probability $\mathbf{v}$. This aggregation method is also referred to as the 'soft' class probability aggregation. In the binary classification case, the resulting predictions $\mathbf{v} = (c_0, c_1)$ where c_0 (reps. c_1) corresponds to the probability of being classified as an object of class 0 (resp. class 1) and $c_1 = 1 - c_0$.

Extension of variable importances to ensembles

While the direct interpretation of the ensemble of trees is lost, it is still possible to rank the variables according to the importance in the forest (their occurrence and proximity to the root of a tree). Similarly to the case of single decision trees, variables importances are computed for tree ensembles by using the mutual

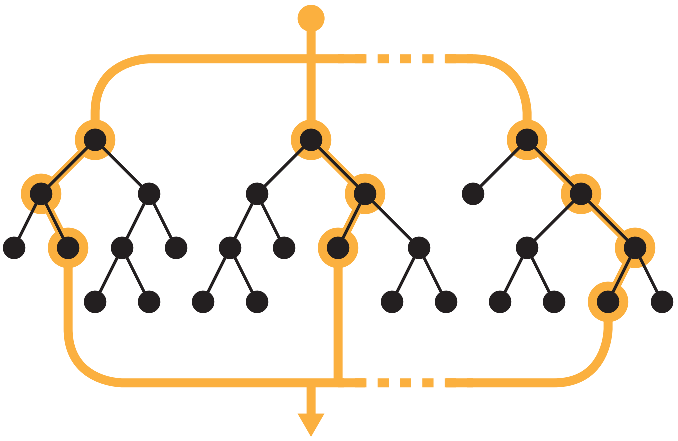


FIGURE 4.5 Ensemble of trees used in prediction mode.

information. In Equation 4.11, instead of looking at nodes from a single tree, the set $Nodes(x_i)$ becomes now the set of all nodes in the forest of T trees where variable x_i is used.

As such, this measure is however dependent on the number of trees T of the ensemble, and on the initial impurity of the dataset and the impurity reduction yielded by the trees, and is thus difficult to interpret. Hence the two possible following normalizations may be used:

$$\mathbf{V}^1_{imp}(x_i) = \frac{\mathbf{V}_{imp}(x_i)}{\sum_{i=1}^{p} \mathbf{V}_{imp}(x_i)}, \tag{4.13}$$

or

$$\mathbf{V}^2_{imp}(x_i) = \frac{\mathbf{V}_{imp}(x_i)}{\max_{i\in\{1,\ldots,p\}} V_{imp}(x_i)}. \tag{4.14}$$

In Equation 4.14, the variable importances are normalized such that the maximum is equal to 1, whereas in equation 4.13 they are normalized so as to sum up to 1.

Discussion of intrinsic features of this algorithm

In the Random Forest algorithm (Algorithm 3), we introduced 3 meta-parameters: T, K and N_{min}. Depending on the nature of the problem, these parameters influence the induced model, the computation time required for the learning stage and the quality of the results. Here is a small discussion about each of these meta-parameters and some recommendations:

- T, the number of trees: the larger the number of trees in such a forest of decision trees and the better will be the variance reduction of the resulting aggregation. It is then recommended to build as many trees as possible in order to obtain better models and smaller generalization errors. Essentially, the choice of this parameter is driven by the available time, computational power and the learning set size. As in our GWAS problem, we face the small n/p ratio, in conjunction with the limited number of test nodes in a single tree, growing a larger forest and randomising the subset of candidate attributes at each node increases the chances of investigating each SNP at least once;

- K, the number of randomly selected variables at each node: the possible value for this meta-parameter ranges from 1 to p (where p denotes the total number of attributes). Of course, as K increases, the required computation time to construct a single node increases too. From an accuracy point of view the optimal value of K is problem dependent, and we will thus investigate its impact on our different datasets.;

- $N_{\min}$, the minimum sample size for splitting a node: also called the *smoothing strength*, it reduces the depth of the trees by limiting their complexity. A larger value of $N_{\min}$ tends to produce results with higher bias and smaller variance. Its optimal value is related to the noise level in the dataset as it prevents "small" nodes from splitting (thereby avoiding the resulting trees to fit the noise). Larger values of $N_{\min}$ lead to earlier stopping of the tree induction process, and hence smaller computation times.

4.2.3 Extremely Randomized Trees

Motivation

The Extra-Tree method (standing for *ext*remely *ra*ndomized *tree*s) was proposed in [GEW06], with the main objective of further randomizing tree building in the context of numerical input features, where the choice of the optimal cut-point is responsible for a large proportion of the variance of the induced tree.

With respect to random forests, the method drops the idea of using bootstrap copies of the learning sample, and instead of trying to find an optimal cut-point for each one of the K randomly chosen features at each node, it selects a cut-point at random.

This idea is rather productive in the context of many problems characterized by a large number of numerical features varying more or less continuously: it leads often to increased accuracy thanks to its smoothing and at the same time significantly reduces computational burdens linked to the determination of optimal cut-points in standard trees and in random forests.

From a statistical point of view, dropping the bootstrapping idea leads to an advantage in terms of bias, whereas the cut-point randomization has often an excellent variance reduction effect. This method has yielded state-of-the-art results in several high-dimensional complex problems [MWG13, LG12, G$^+$05a].

From a functional point of view, the Extra-Tree method produces piece-wise multilinear approximations, rather than the piece-wise constant ones of random forests [GEW06].

Our implementation of the Extra-Tree algorithm

The pseudo-code is given in Algorithms 5 and 6. Essentially the method iterates T times by using the initial learning sample to grow a tree, in the following way:

1. at each node of each tree a random subset of K features is investigated, as in the random forest method,

2. for each of the K features, instead of searching for the optimal split, a random question is picked (in the case of numerical features, this is done by selecting a threshold from a uniform sampling distribution spanning the values of the considered feature in the sub-sample of the current tree node). The K couples (feature, threshold) are evaluated by computing their score and the best one is selected to split the node.

```
BuildETEnsemble
input  : $\mathcal{LS}$, $A_{cand}$,$T$,$K$,$N_{\min}$
output: A tree ensemble $\mathcal{T} = \{t_1, ..., t_T\}$

for $i \leftarrow 1$ to $T$ do
    $t_i \leftarrow$ BuildExtraTree($\mathcal{LS}$,$A_{cand}$,$K$,$N_{\min}$)
    Append($\mathcal{T}$,$t_i$)
return $\mathcal{T}$
```

Algorithm 5: Extra-Trees: the top-level iteration of the algorithm (there is no randomization; meta-parameters are exposed in blue).

```
BuildExtraTree
input  : $\mathcal{LS}$,$A_{cand}$,$K$,$N_{\min}$
output: A tree : $t$                    // Its root node

if const(attributes) or const(output) or $\mathcal{LS} \leq N_{\min}$ then
    return a leaf labeled by class frequencies in $\mathcal{LS}$
else
    Select $K$ random attributes $\in A_{cand}$ : $\{a_1, ..., a_K\}$
    $\{s_1, ..., s_K\}$ : $s_i$ = pickRandomSplit($\mathcal{LS}$, $a_i$)
    $s_* = max_{i=1,...,K}$Score($s_i$, $\mathcal{LS}$)
    Split $\mathcal{LS}$ into $\mathcal{LS}_{left}$ and $\mathcal{LS}_{right}$ according to $s_*$
    $T_{right} \leftarrow$ BuildExtraTree($\mathcal{LS}_{right}$,$K$,$A_{cand}$,$N_{node}$)
    $T_{left} \leftarrow$ BuildExtraTree($\mathcal{LS}_{left}$,$K$,$A_{cand}$,$N_{node}$)
    return createNode($s_*$,$T_{left}$,$T_{right}$)
```

Algorithm 6: Extra-Trees: algorithm for building one extremely randomize tree. Meta-parameters exposed in blue, and randomization in red.

Discussion of interest w.r.t. random forests

In the Extra-Trees, the different inputs needed are the same as for the random forests but $K = 1$ has now a particular meaning. Indeed, setting K to 1 makes the resulting trees totally randomized. Indeed, given the fact that the cut-point is randomized and that there is only one random attribute being checked, the generated splits are totally independent from the output variable information provided in the learning sample. This means that this version (Extra-Trees with $K = 1$ are called Totally Randomized Trees) is producing models that are very close to the k-nearest neighbour method [GEW06].

In the particular case of SNP attributes, there are at most two possible cut-points (0.5 and 1.5). In the random forests, the search for the optimal split at each node is straightforward while in the Extra-Trees the randomization of the cut-point is not so random anymore. This reduces the difference between the choices that are made at each nodes in the two types of methods. Also, the expected speed-up due to the absence of optimal split search in the Extra-Trees is way less significant in such configuration of the feature space.

Finally, variable importances are computed exactly as for the random forest, using the mutual information.

4.3 Trees inside Trees

In this section we describe our method called *T-Trees* (which stands for Trees inside Trees). It is based on the two preceding ones: the random forests and the Extra-Trees. This novel approach aims at addressing some specific features of GWAS.

4.3.1 Motivation

In all the previous algorithms, test-nodes are exploiting only one variable at a time. The basic idea under our extension proposition of decision tree algorithms is to treat more than one variables inside the splitting nodes. One of the main reasons to modify the splits is because of the particular structure of the feature space. In the GWAS context, due to linkage disequilibrium, we expect at a given position a limited haplotype diversity. Thus based on the physically ordered nature of the SNPs, the idea behind the T-Trees algorithm is to partition the feature space into blocks of contiguous and (potentially highly) correlated variables. The splits will be made on a block of SNPs instead of a single one, taking advantage of the local information potentially carried by the region covered by the corresponding block. Figure 4.6 illustrates the structure of a T-Tree.

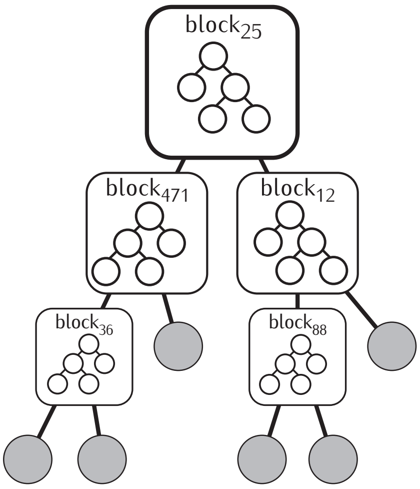


FIGURE 4.6 Overview of a T-Tree. Instead of single variable test-nodes, a T-Tree allow to split on blocks of variables.

This constraint allows to reduce the size of the feature space. By doing so, we expect to increase the chances of:

- capturing the interactions between (blocks of) variables: as the size of the feature space is divided by the size of the blocks, we increase the chance of finding interactions in consecutive splits (or at least along one branch of a tree),
- discovering a particular SNP combination (i.e. a haplotype) that is linked to the disease. Indeed, in the classical tree-based methods, the chances of testing consecutively two SNPs falling into the same haplotype block are relatively small (it will depend on the total number of SNPs, the parameter K and the presence of other informative variables),
- exploiting a group of surrogate variables: if two or more variables are in perfect LD, they share the exact same information about the output variable, due to the randomisation in the previously presented algorithms, each of these variables will be asymptotically equally selected in a forests of decision trees (random forests or Extra-Trees), their respective importances will drop as the number of surrogates increases. The ability to rank a block instead of a single variable will help to identify a group of highly correlated variables.

4.3.2 Algorithm

Our method is based on the two previously presented tree-based methods : the random forest and the Extra-Trees algorithms (respectively abbreviated RF and ET in the following).

As a reminder, those two methods are quite similar, but differences occurs during the learning stage, RF grows T trees by recursively partitioning a bootstrap sample (n samples extracted with replacement from the learning set of size n). At each node, a search for the optimal split among a random subset (of size $k \leq p$) of all the p variables is performed. On the other hand, ET grows T trees by recursively partitioning the initial learning set. At each node, K random splits on a random subset of K variables are picked, the best one is kept.

Basically, the node splitting rule in RF is modified to test a group of variables using a weak learner[2] developed on a small subset of the feature space (e.g. a small number of consecutive SNPS). The predictions of this small learner will produce a vector of probability for each learning samples reaching the corresponding node. That vector is used as a new numerical attribute corresponding to the group over which the split was made. A cut point is then optimally chosen between 0 and 1 as in the standard RF algorithm. In our proposal, we choose to use a single Extra-Tree with a limited number of nodes as a weak learner.

These small trees are used inside the splitting nodes, they are developed on a subset of variables. The Algorithms 9 and 10 uses the *Extra-Trees* ensemble method where the number of tested attributes at each node (K_{int}) is set to the total number of variables contained inside the groups/blocks and T is fixed to 1. Two important modifications are added :

1. the maximal number of nodes is limited by the *internal complexity* parameter: IC. The choice of this value will depend on the nature of the variable groups. $IC = 1$ will be an interesting choice for strongly correlated attributes as they all carry the same information, higher values will be well suited when a combination of several attributes is required to explain the outcome. Note that when the internal complexity is set to 1, the method does not reduce to standard random forests as in T-Trees the K variables will be selected each in a different bloc, thus widening the scan coverage of the feature space while searching for an optimal split.

2. in most of the decision tree induction algorithms a fixed order is used to expand the nodes (left to right or vice versa). As we choose to limit the number of nodes, we want to avoid the resulting tree to degenerate (i.e. to become a branch). For that reason we expand the nodes in a randomised order (to break the depth-first order). For the sake of clarity, that randomisation step has been intentionally omitted in the Algorithm 10.

In the following, as we use trees inside trees, what we call *external nodes* are the group test nodes containing the weak learners. These nodes correspond to RF nodes. *Internal nodes* are the ones being part of the weak learner (the node limited Extra-Tree), testing a single variable (Figure 4.7).

[2] In the machine learning context, a weak learner is a model that performs at least better than random guessing.

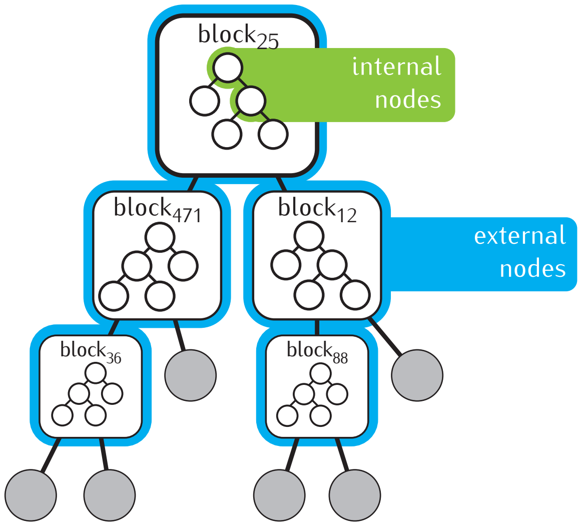


FIGURE 4.7 T-Trees terminology: difference between internal and external nodes.

As input, instead of a pool of candidate attributes, this method needs a *block map* defining how variables are grouped. In the GWAS context, the block map partitions the initial feature space into sets of contiguous SNPS. For this novel method, instead of selecting K random attributes (Algorithm 4), we select K random groups of variables. Algorithms 7 and 8 respectively detail the T-Trees forest induction and one T-Tree induction.

Figure 4.8 shows an example of a T-Tree splitting node. Three external nodes are represented, they partition the learning set using group 8, 27 and 1. In this example, the $IC = 3$. Out of the $Group_1$ 3 SNPS are tested: snp_3^1, snp_6^1 and snp_2^1.

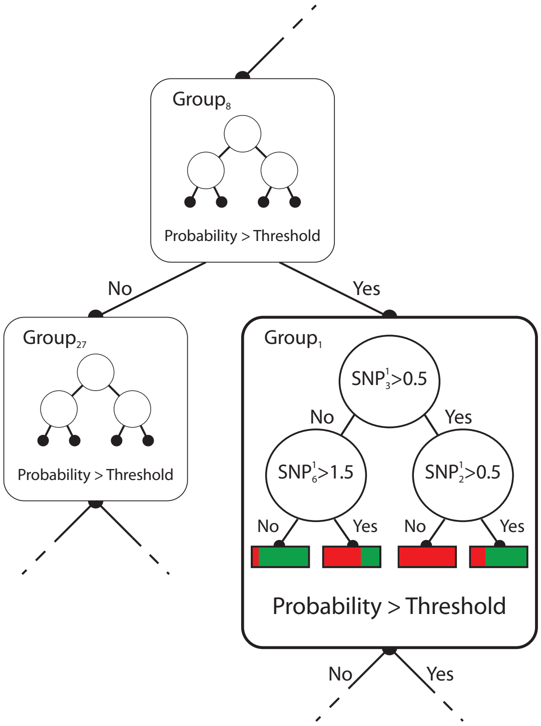


FIGURE 4.8 A closer look into a T-Tree test-node shows how the weak learner is used to split on more than one variable.

At each external node, attribute groups are tested and the prediction obtained with the internal tree corresponding to this node is thresholded to propagate the object to a successor.

```
T-Trees
input : $\mathcal{LS}$,$\mathcal{B}$,$T$,$K$,$K_{int}$,$IC$,$N_{min}$
output: $\mathcal{T}$
----------------------------------------------------------------
for $i \leftarrow 1$ to $T$ do
    $\mathcal{LS}_i \leftarrow$ Bootstrap($\mathcal{LS}$)
    $T_i \leftarrow$ BuildTTree($\mathcal{LS}_i$,$\mathcal{B}$,$K$,$K_{int}$,$IC$,$N_{min}$)
    Append($\mathcal{T}$,$T_i$)
return $\mathcal{T}$
```

Algorithm 7: The T-Trees algorithm is quite similar to the random forest algorithm. It adds two metaparameters: K_{int} and IC; and needs a block map $\mathcal{B}$.

```
BuildTTree
input : $\mathcal{LS}$,$\mathcal{B}$,$K$,$K_{int}$,$IC$,$N_{min}$
output: A TTree $T$                                // Its root node
----------------------------------------------------------------
if const(attributes) or const(output) or $\#\mathcal{LS} \leq N_{min}$ then
    return a leaf labeled by class frequencies in $\mathcal{LS}$
else
    Select $K$ random blocks $\in \mathcal{B}$ : $\{g_1, ..., g_K\}$
    $\{s_1, ..., s_K\} : s_i =$ pickGroupSplit($\mathcal{LS}, g_i, K_{int}, IC$)
    $s_* = max_{i=1,...,K}$Score($s_i, \mathcal{LS}$)
    Split $\mathcal{LS}$ into $\mathcal{LS}_{left}$ and $\mathcal{LS}_{right}$ according to $s_*$
    $T_{right} \leftarrow$ BuildTTree($\mathcal{LS}_{right}$,$\mathcal{B}$,$K$,$K_{int}$,$IC$)
    $T_{left} \leftarrow$ BuildTTree($\mathcal{LS}_{left}$,$\mathcal{B}$,$K$,$K_{int}$,$IC$)
    return createNode($s_*$,$T_{left}$,$T_{right}$)
```

Algorithm 8: The T-Tree building algorithm

```
pickGroupSplit
input  : $\mathcal{LS}, g, K_{int}, IC$
output: $[p < th]$

$T$ = BuildExtraTTree($\mathcal{LS}$,$g$,$K_{int}$,$IC$)
Propagate $\mathcal{LS}$ in $T$, $p$ = vector of resulting probabilities
Search optimal threshold $th$ over $p$
return $[p < th]$
```

Algorithm 9: The `pickGroupSplit` function is based on the Extra-Trees algorithm. A single Extra-Tree is built and its predictions allow to transform a group of attributes into a new numerical value.

```
BuildExtraTTree
input  : $\mathcal{LS}$,$g$,$K_{int}$,$IC$
output: A tree : $t$                         // Its root node

if const(attributes) or const(output) or $\#nodes \leq IC$ then
    return a leaf labeled by class frequencies in $\mathcal{LS}$
else
    Select $K_{int}$ random attributes $\in g$ : $\{a_1, ..., a_{K_{int}}\}$
    $\{s_1, ..., s_{K_{int}}\} : s_i$ = pickRandomSplit($\mathcal{LS}, a_i$)
    $s_* = max_{i=1,...,K_{int}}$Score($s_i, \mathcal{LS}$)
    Split $\mathcal{LS}$ into $\mathcal{LS}_{left}$ and $\mathcal{LS}_{right}$ according to $s_*$
    $T_{right} \leftarrow$ BuildExtraTTree($\mathcal{LS}_{right}$,$g$,$K_{int}$,$IC$)
    $T_{left} \leftarrow$ BuildExtraTTree($\mathcal{LS}_{left}$,$g$,$K_{int}$,$IC$)
    return createNode($s_*$,$T_{left}$,$T_{right}$)
```

Algorithm 10: Inside the external nodes, the weak learner that is used is a single Extra-Tree with an IC-limited number of (internal) test nodes.

4.3.3 Evaluation of variable and group importances

The variable ranking ability of tree-based methods is conserved. The T-Trees uses two types of features: variables (in the internal nodes) and groups of variables (in the external nodes). This leads to two importance measures:

- variable importances: instead of considering all the nodes, we only take into account the internal ones. For each variable, we can sum over all the internal nodes where a variable appears (regardless of its group) the local reduction of entropy weighted by the local sample size.

- group importances: to rank the groups of variables, we consider the external nodes only. As they also produce a split over a learning set, it is also possible for each group to sum the local reduction of entropy at nodes where a group is used (regardless they exploit all variables of a group or not).

Given the features we want to rank, Equation 4.11 is used but the n_j are internal nodes for the variable importances or external nodes for group importances.

4.4 Extension to more quantitative or multiple phenotypes and environmental effects

In this chapter we presented various tree-based methods and focused on binary classification problems. It has to be noted that these methods can easily handle more than two phenotypes, the same score measure can be used on larger contingency tables (instead of a 2×2 table we consider a $2 \times \#outcome$ table).

Also, regression trees are well suited in case of quantitative traits. Instead of measuring the reduction of entropy, variance reduction is used as a score measure.

Finally, the incorporation of environmental factors could be easily achieved as splitting rules could easily handle categorical variables as well.

4.5 Related works

During the last decade, many researchers investigated machine learning approaches in the field of genome-wide association studies. Many of these researches focused their effort on the ability to localise loci associated with a phenotype while a few concentrated their attention on the prediction power of such machine learning techniques. Beside direct/"black-boxed" application of the method, a few proposed small adaptations in order to circumvent some of the method flaws, but none of these investigations proposed an intrinsic modification of the tree-based methodology. Among these, many promising results opened the path and motivated our current research. More precisely, these researches investigated:

- **The general power of machine learning and tree-based approaches on real GWAS datasets**

  In [BDH$^+$03], random forests were applied to the Genetic Analysis Workshop 13 small simulated dataset. Using the IBD (identity by descent) score of sibling pairs as variables in a candidate gene and a genome scan approach, they predicted several quantitative phenotypes (HDL, triglycerides and glucose). The true model being known, they showed the ability of the random forests to detect both susceptibility genes and markers.

  In [H$^+$06], the authors evaluate strengths and weaknesses of several machine learning analysis approaches on large numbers of SNPs (logistic regression, neural network, combinatorial partitioning, multifactor dimensionality reduction and Random Forests). Among the criteria, the following important ones have been considered: the ability to handle large scale datasets, interactions, correlations and heterogeneity. Especially for Random Forests, they highlight the need to deal with correlated SNPs which motivated our T-Trees proposal.

  More recently, in [GHCB10], a random forest based GWAS (with more than 300K SNPS as input variables describing 3000 individuals) on a multiple sclerosis dataset has proven successful, identifying four new candidate MS genes in addition to a few more already reported as associated. They studied the impact of the number of tested attribute at each node (k) and the number of trees on the out-of-bag error rate and variable importances stability. They proposed to remove strong associations in order to find weaker ones (in this case the MHC region on chromosome 6 that is well known to be associated with MS). To address variable correlations, they used an LD-pruning approach using Plink, mostly similar to the one proposed in [MYC$^+$09], removing redundant SNPs. Not surprisingly, they found that, as we expect a small proportion of the variables to be relevant, larger values of k are better.

  In [GRF11], the authors compared two types of Bayesian methods with two types of tree based methods (random forests and boosting). They compared the predictive capability of these methods in simulated and real genetic data analysis. On simulated and purely additive scenarios, they found out that the tree-based methods, and in particular random forests, were slightly superior in terms of precision when dealing with a small number of associated loci. When that number increased, they noticed that the

Bayesian approaches were a bit superior in terms of their Pearson correlation evaluation criteria but still the random forests produced the best AUC results. Similar observations were drawn on the real data sets: random forests produced the best AUC results in every experiments, in particular, they noted that the method perfectly classified the most extreme animals in the different test sets.

Also, in [T+12c], the authors proposed an overview of the random forests in the life sciences and the GWAS context. Discussing the common RF usages and pitfalls, they also briefly tackled some interesting research directions and less known direct by-products of tree-based methods. Among these, they propose to determine similarity between individuals using proximity scores, to compute variable importance locally instead of globally (e.g. by considering only a subpopulation in the variable importance calculation) or to analyse the forest structure in order to detect recurring cascades of interacting SNPs along the tree branches.

- **The ability of tree based methods to detect SNP interactions**

  In [LHSVE04], the authors studied the capabilities of random forest importance measure to correctly rank interacting SNPs. They compared standardised RF importance scores (so called Z_T in the publication) to Fisher p-values on simulated datasets with 100 and 1000 variables under different heterogeneous interaction models. Their finding were the following: Z_T ranking outperforms Fisher test in presence of interacting SNPs, Z_T increases as the number of interacting SNPs does too and if there is no interactions Z_T performs similarly to the univariate Fisher Exact test.

  In [JTWF09], the authors use random forest algorithms and the Gini importance measure as a selection tool to search for epistatic interactions in a framework called *epiForest*. They propose a sliding window sequential forward feature selection procedure to select a subset, significantly smaller than the initial set of variables, of candidate SNPs that minimise the classification error and than test up to three-way interactions using the B-statistic from [ZL07]. They compared their method with BEAM, the stepwise logistic regression and the $\mathcal{X}^2$ test on three simulated disease models and on a genome-wide case-control dataset for age-related macular degeneration (100k SNPs, 96 cases versus 50 controls). They found that Gini variable importances were negatively correlated with the p-values and identified two SNPs that were already reported as being linked to the disease. Nevertheless, due to the small sample size, they found no significant interactions after the Bonferroni correction.

  Finally, in [W+12], the authors studied the ability of random forests to identify SNP interaction in high-dimensional space such as in GWAS. Using different simulation with a fixed number of interacting and non-interacting variables, they noticed that as the total number of variables increased, the probability of detecting interacting SNPs drops more rapidly than for the non-interacting ones. Their experiments were conducted using standard/recommended RF parameters, notably they suggest a limitation on the number of terminal nodes (i.e. limit the depth of the trees) which certainly could limit the decline they observed. This suggests that a dimension reduction would limit the decrease in the probability of detecting interacting loci as we propose in our tree-based adaptation.

- **The behaviour of variable importances in presence of linkage disequilibrium**

  In [BGHW08b], we started to treat haplotype blocks instead of single SNP inside test nodes using a maximum likelihood based estimation of the conditional probability that the observed haplotype block is drawn from the population of cases (resp. controls) assuming class conditional independence of the SNPs in the block. The results obtained on simulated data representing five different disease models provided marginally superior results than the direct application to the SNP representation.

  In order to limit the information dilution among variables in LD, [MYC+09] proposes to prevent two SNPs in linkage disequilibrium to appear in the same tree in a forest so they can not act as surrogate for each other. They change the tree building procedure by growing each tree only with SNPs in linkage equilibrium using a threshold over the pair-wise genotypic correlation (r^2). They modified the permutation variable importances calculation in consequence, for a particular variable v, variable importance is based only on the trees in the forest were v appears. They compared their adaptation of the original random forest method under various synthetic genetic models. Their results suggest that when a risk SNP appears in a tree, SNP in LD might also appear in the same tree which causes the corresponding variable importances to decrease as the number of markers in LD increases. They recommend using the original random forest with their revised importance measure when the genetic model and the number of SNPs in LD are unknown (which is a common situation in a real GWAS). They also applied their methodology on a realistic GWAS dataset and successfully identified a reported risk gene and four new candidate loci missed by the single SNP approach.

In [NM09, NMSZ10, Nic11], the authors investigated the behaviour of the Gini importances in presence of high LD bloc of SNPs. Under the null hypothesis, their experiments show that the higher the redundancy, the lower the variables importances. They noted that pruning the trees diminished that bias. This expected result confirms previous hypotheses about the possible information dilution that could be observed when dealing with redundant variables. Especially in presence of signal, once one of the correlated variables appears in a tree, its surrogates have way less chance to be selected and the bottom of the tree is more inclined to exploit the other non correlated variables. Although their simulations might be considered as extreme cases (almost perfect LD and a low number of variables), they point out some flaws in the Gini variable importances. In addition, they also noticed a bias towards continuous variables (as noticed by [SBZH07, BBLBS12]).

- **Pathological issue and bias related to the Gini variable importances**

  The authors of [SBZH07] suggests that the original variable importances of Brieman in random forest are not reliable when attributes vary in their scale of measurement or their number of categories. They evaluated three different variable importance measures (the "selection frequency" or the number of times a variable appears in a forest, the Gini importance and the permutation accuracy importance) on synthetic data while subsampling with and without replacement at each trees (bootstrap and no bootstrap). In the null case scenario, they demonstrate the preference toward variables with more categories of the frequency and Gini importances. On the other hand, the permutation importance seems less sensitive to the scale of measurement but requires much more computations.

  In the continuity, using only non-informative SNPs with varying minor allele frequencies, [BBLBS12] found out that Gini importance tends to favour larger minor allele frequency variables (in particular at the bottom of the trees). Meaning that non informative SNP with large MAF might hide the presence of, maybe, interesting variables with smaller MAF. That bias is similar to the one that is expected when dealing with variables of different types. As previously observed in [SBZH07], discrete variables with high number of categories are preferred by the Gini importance measures. Also, in presence of causal markers, if two variables are causative but differ in their respective MAF, the one with the larger minor allele frequency is also the one that is about to affect a bigger part of the learning set. Thus, it might not be surprising to observe a similar "bias". They noted that permutation variable importances were less sensitive to minor allele frequencies.

4.6 Summary

In this chapter, we described 3 types of decision trees algorithms and proposed an extension able to handle more than one descriptor inside the test-nodes.

Single decision trees are really easy to understand and interpret due to their graphical representation: descriptors appearing near the root node are clearly more important than those appearing into the leaves. Beside that simplicity, a single tree is clearly not enough to test all the variables as the number of nodes is limited by the sample size. This is especially problematic for the study of complex diseases where a large number of variables is expected to be causative and the intrinsically small n/p ratio.

Ensemble methods proposes to circumvent that previous problem by growing a forest of trees which is achieved by introducing some randomization during the learning stage and multiplying the number of trees. In that case, the direct interpretation of the ensemble of trees is lost but it is still possible to rank the variables according to their importance in the forest (their occurrence and their proximity to the root of a tree).

Despite all these important improvements, there are still cases where the previous algorithms are not as good as we expect them to be. In the particular case where the descriptors are structured, the chances of using the correct cascade of attributes drastically decreases as the total number of attributes increases. Also, in presence of correlated attributes, classical methods are victims of the information dillution among the variables. That is why we propose to extend the random forests by allowing splitting over a group of descriptors.

With the success of genome-wide studies and technologic advances, it is clear that denser genotypes are becoming usual. As the density increases, the number of descriptors increases as well and, obviously, stronger correlations appear between the variables splitting the genome into logical pieces. In the following chapter, we will show that the T-Trees method outperforms random forests and Extra-Trees on such data.

Beside their particularities, each of these algorithms can be summarised by Figure 4.9. Some of the steps may be optional, less or more important depending on the algorithm. When a forest is required, this flowchart is repeated a certain amount of time (Figure 4.10).

In these methods, as all terms of an ensemble are independent, it is easily feasible to parallelise the construction of a forest using several CPU.

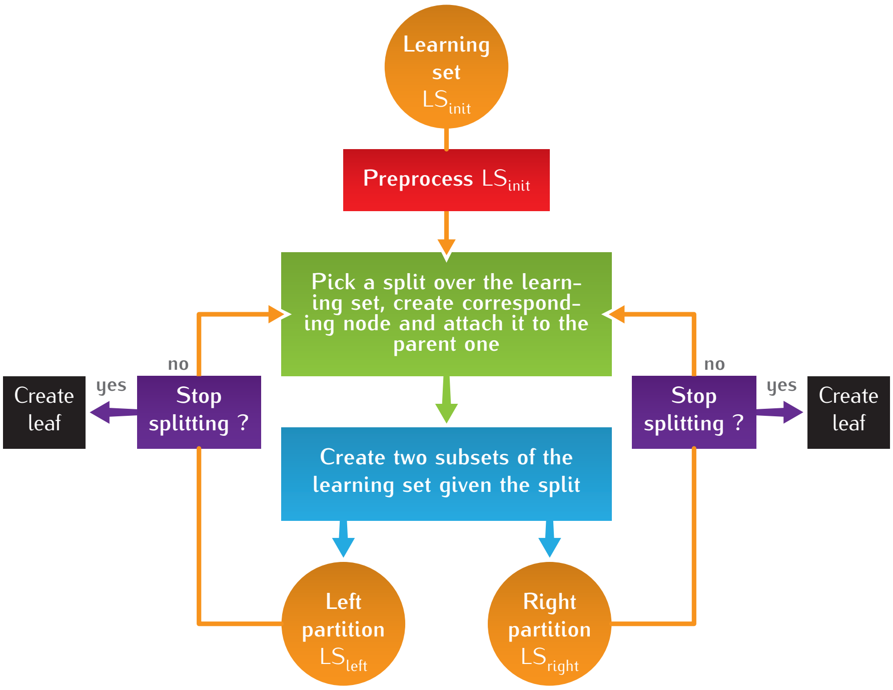

Learning set LS_{init}
Preprocess LS_{init}
Pick a split over the learning set, create corresponding node and attach it to the parent one
no
no
Create leaf
yes
Stop splitting ?
Stop splitting ?
yes
Create leaf
Create two subsets of the learning set given the split
Left partition LS_{left}
Right partition LS_{right}

FIGURE 4.9 Flowchart : growing a decision tree.

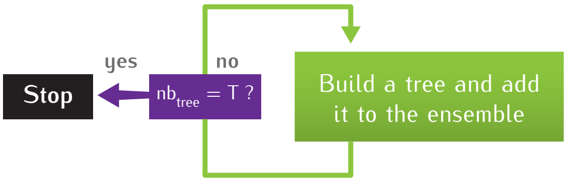


FIGURE 4.10 Flowchart : growing a ensemble of trees.

Part II

Validations

Chapter 5

Comparison of Random Forests and T-Trees on synthetic datasets

Contents

In this chapter, the T-Trees (which stands for trees inside trees) method is evaluated on different synthetically generated datasets aiming at mimicking the structure of GWAS datasets. Its predictive accuracy and its ability to identify relevant input variables are compared with those of the Random Forest method, and its capability to identify causal loci is also compared with that of univariate p-values derived from the exact Fisher test. This study, carried out under controlled conditions, should be considered as a first sanity check and also aims at evaluating the effect of the meta-parameters of the method, so as to evaluate its robustness and to identify appropriate default values of these meta-parameters, before the method is applied in the upcoming chapters to real-life GWAS datasets.

5.1 Synthetic 'GWAS' dataset generation

Our goal is to generate datasets which reproduce the structure of GWAS under controlled conditions, namely with known block structure and a known model of genotype-to-phenotype mapping. To this end, we proceed in two steps:

- first, we generate a sample of input (genotype) vectors reproducing in a way the linkage-disequilibrium structure among genetic markers, based on a segmentation of the set of input variables into a number of blocks of linked variables;
- next, we choose a subset of causal blocks among those genotype blocks together with a genotype to phenotype relation (including a choice of background noise level) and then use this model to associate a binary output (phenotype) value to each input vector, in such as way that the two classes are balanced over our dataset.

5.1.1 Principle of the synthetic 'genotype' generation

We suppose that the input variables are grouped into g consecutive blocks denoted by $b_i, i = 0, \ldots, g-1$, where each block has a given size $s_i \in \mathbb{N}_0$ in terms of the number of 'SNPS' it contains. To each 'SNP' corresponds a ternary input variable. Let us denote by x_{i_j} the jth input variable (SNP) belonging to the ith block. Thus

$$x_{i_j} \in \{0, 1, 2\}, \forall i = 0, \ldots, g-1, \forall j = 1, \ldots s_i,$$

and hence any block b_i may in principle take 3^{s_i} different 'haplotypes' of length s_i.

In order to mimic 'linkage disequilibrium' inside these blocks, we fix once and for all for each one of them a number $m_i \ll 3^{s_i}$ of possible block-wise configurations $\{b_i^1, \ldots, b_i^{m_i}\}$ and then restrict the observed values of the block configurations to these latter. To choose one of the possible m_i configurations for a given block, we select its individual SNP values independently while using uniform probabilities over the three possible values $\{0, 1, 2\}$.

On the other hand, we assume that the haplotypes of two different blocks are statistically independent. Thus, the selection of the genome-wide genotype of an individual is carried out by choosing at random for each block b_i a configuration out of the fixed set $\{b_i^1, \ldots, b_i^{m_i}\}$.

To yield a complete dataset of 'genome-wide' genotypes, we then select independently according to the above principle a number n of genome-wide genotypes.

5.1.2 Principle of the synthetic 'phenotype' generation

Given a set of possible block-wise genotypes and a dataset of n genome-wide genotypes of n individuals, $\{(b_0^{j_0^k}, \ldots, b_{g-1}^{j_{g-1}^k})\}_{k=1}^n$, both built according to the methods described in the previous section, we proceed in the following way to compute the corresponding output class $y^k \in \{0, 1\}$ for each individual.

First, we choose a number $c \in \mathbb{N}$ of distinct 'causal' blocks; let us denote by $c_1, \ldots, c_c$ the corresponding block identifiers ($c = 0$ means that actually no causal locus at all is present). Next, we associate to each one of the m_{c_i} haplotypes of each one of these c blocks a random number chosen uniformly (and independently) in the interval $[0, 1[$. Let us denote by $z(b_{c_i}^j), \forall i = 1, \ldots, c, \forall j = 1, \ldots, m_{c_i}$ the corresponding numbers.

We then compute a numerical value for each individual k according to the following 'additive over blocks' model

$$z^k = \beta\epsilon^k + \sum_{i=1}^{c} \alpha_i z(b_{c_i}^{j_{c_i}^k}), \tag{5.1}$$

where ϵ^k is a random number uniformly distributed in $[0, 1[$, and where the positive parameters, β and $\alpha_i, i = 1, \ldots, c$, define respectively the level of noise, and the strength of the different genetic effects that are associated to the different causal blocks.

Finally, in order to associate a discrete class to each individual of our dataset, and to ensure at the same time that the two classes are balanced in terms of the number of individuals, we compute for each observation k its class y^k by

$$y^k = 1(z^k \geq \theta),$$

where θ is adjusted to the median of the z^k values of the dataset.

5.1.3 Comments

We understand that the above synthetic data model is still far from reproducing the real nature of GWAS datasets. We are also aware of existing simulators published in the literature [SMD11, E$^+$08, LL08, GGS11, P$^+$07]. Nevertheless, for the sake of reproducibility, and to facilitate further research along our line of thinking, we preferred to design our own synthetic model for the study carried out in this chapter.

In the rest of this chapter, we will consider various conditions, corresponding to different settings of our synthetic model, so as to compare the Random Forest and the T-Tree models.

To keep the computational burden under control, we will restrict to $1000 - 1500$ genetic markers (which is, admittedly, quite small with respect to the 500k or 1M markers used in real-life GWAS). On the other hand, we will consider block sizes $s_i \in [10, 50]$, a number of block modalities $m_i \in [10, 500]$, and a number of causal blocks $c \in \{1, 2, 3\}$, with or without background noise. The precise settings of the conditions will be given for each sub-study carried out in the next sections.

5.2 Evaluation protocol

In all our simulations in this chapter, we used a dataset composed of $n = 10,000$ samples in which we select learning samples of small sizes (between 100 and 500 individuals), so as to have large enough test samples for the evaluation of predictive accuracies of the learned models.

Our evaluation protocol for assessing accuracies, for a given learning sample size and given values of the algorithm parameters, operates in the following way:

1. select a learning sample of the considered size at random among the $n = 10,000$ individuals of the dataset, and define the test sample as the remaining individuals;

2. build a model on the learning sample, and assess its accuracy in the form of the value of its AUC, on the corresponding test sample (remaining individuals);

3. repeat ten times steps 1 and 2, and then display the average AUC values obtained over these ten runs.

We note that within a given comparison of RF and T-Trees, we always use the same 10 learning sample vs test sample splits, for all method variants and for all different parameter settings that are studied, so as to minimise the effect of learning and test sampling variance.

Concerning the Random Forest and T-Tree methods, we did not study the effect of parameter T (number of trees in the ensemble) nor that of $N_{\min}$ (degree of pre-pruning). For the accuracy evaluations, these parameters are kept constant over all our simulations, with $T = 100$ (a moderate but sufficient number of trees, given the relatively small numbers of variables considered in our experiments) and $N_{\min} = 2$ (no pruning).

Notice that all computations of input variable *importances* and/or block-wise *importances* that will be displayed in the subsequent sections are computed as averages over the ten learning-samples (i.e. without using any of the test samples). Because of the higher variance of the importance measures derived from tree-based ensembles, we used for these computations a larger number T of trees in each ensemble, namely $T = 1000$.

5.3 Simulation results

In this section, we successively study the sample efficiency of Random Forests and T-Trees in a single locus model, then we evaluate their robustness in noisy conditions, then we study the effect of prior information quality in terms of the given block maps, then the effect of increasing the number of causal blocks, and finally we study the bias of the T-Trees method towards blocks of higher modalities.

5.3.1 Sample efficiency of RF vs T-Trees in the context of a single causal block

We consider in this section a genotype distribution over $g = 100$ blocks, each one composed of $s_i = 10$ variables, and we took for each block $m_i = 80$ possible modalities. Concerning the phenotype, we assume that there is no background noise (i.e. $\beta = 0$) and that there is only a single 'causal' block ($c = 1$, and we chose block number 0 to be the causal one, arbitrarily but without any restriction) so that in principle perfect prediction of the phenotype from only the SNPs located in the single causal block b_0 is possible. The T-Trees method exploits full knowledge of the blocks, while the Random Forest method treats the variables in a fully agnostic way.

We constructed our models based on learning samples composed of (LS_{size}) $100, 250$ and 500 individuals and then evaluated them on the remaining (out of learning sample) objects. The AUC values displayed are average values obtained over ten such random LS/TS splits.

Figure 5.1 depicts the AUCs we obtained. On this figure, rows correspond to the different learning set sizes and columns correspond to the different methods. From left to right, the black lines represent the AUCs for Random Forests with a value of K ranging from 100 to 1000 (notice that $K = 1000$ corresponds here to Tree Bagging), the middle column represents the AUCs obtained with the T-Trees with $K_{int} = 1$ (i.e. Totally Randomised, which are the weakest learners, inside the internal nodes) with a (block-wise) K ranging from 10 to 100 and the right column shows the results with the T-Trees with $K_{int} = s_i$ ($N = s_i = 10$; these are thus very strong models, almost identical to regular decision trees). The results for the T-Trees are coloured as we investigated different values of the internal complexity parameter: from red to blue, the internal complexity IC ranges from 1 to 10. Recall that these numbers are average values obtained over 10 LS vs TS splits of a dataset composed of 10,000 individuals.

Not surprisingly, and regardless of the method, we observe an AUC increase as the LS_{size} increases; the same applies for the value of K. If we compare the Random Forests to the two variants of the T-Trees, we see that as the IC grows the T-Trees performances get better than the RF. In every case, with an IC of 10, T-Trees obtain the best results.

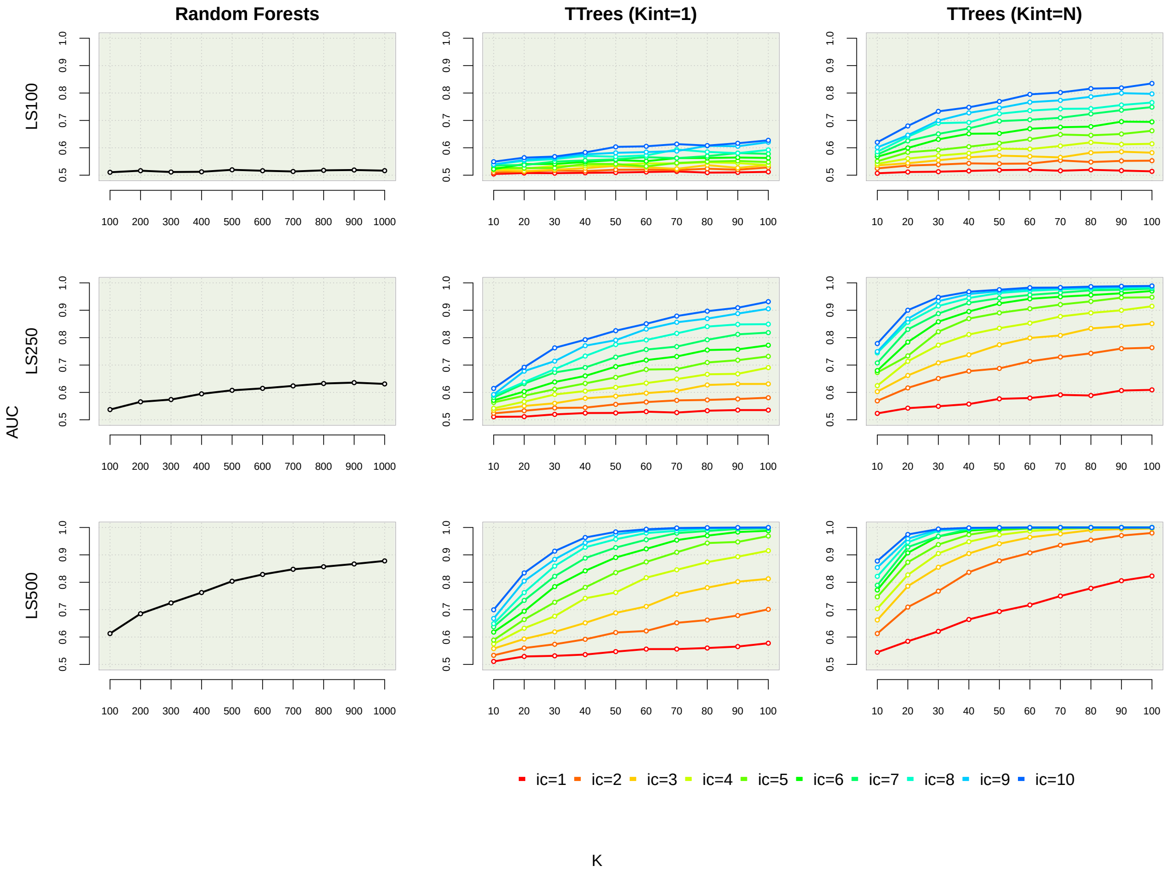


FIGURE 5.1 Influence of learning set size, K and IC on the AUC. The first row represents results with $LS_{size} = 100$, the second one with $LS_{size} = 250$ and the third one with $LS_{size} = 500$. The first column represents the Random Forest results for values of K ranging from 100 to 1000. The two last columns correspond to results obtained with the T-Trees for different values of the internal complexity (from 1 to 10). The middle column corresponds to the T-Trees with K_{int} set to 1 and the right columns with $K_{int} = 10$.

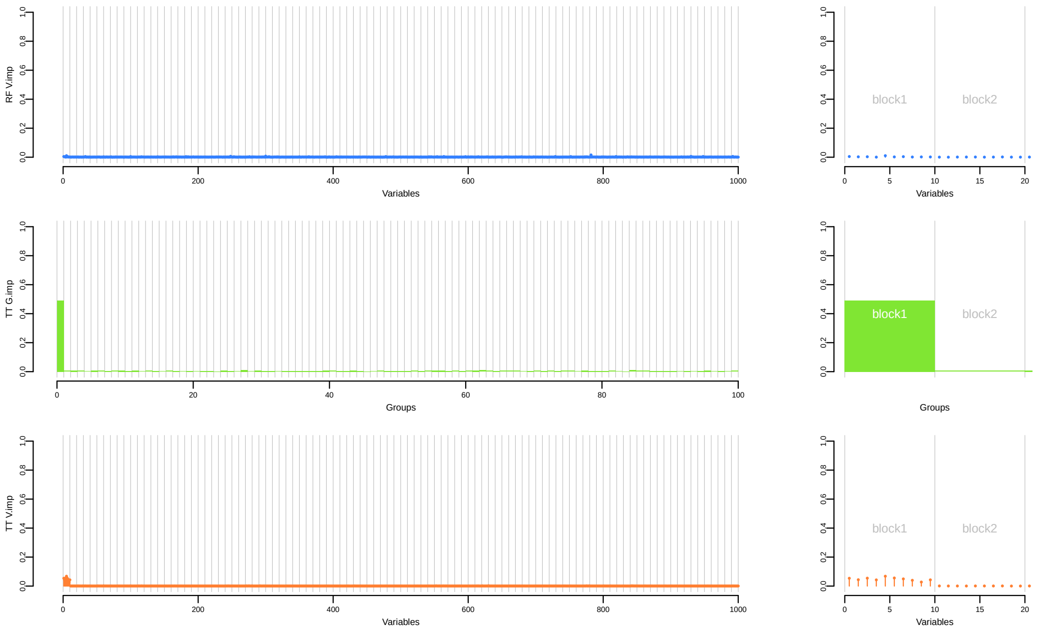


FIGURE 5.2 $LS_{size} = 100$: variable and group importances. In blue, the random forest ($K = 1000$) variable importances. In orange, the T-Trees variable importances and in green, the T-Trees group importances ($K = 100$, $K_{int} = 10$, $IC = 10$). The left part display an overview of all the variables while the right part shows a zoom around the causal block.

We also see that with small learning sample sizes the *T-Trees* very much outperform the Random Forests. While the effective size 250 is still insufficient for Random Forests, *T-Trees* seem already to learn quasi perfect models that correctly classify the remaining samples; *T-Trees* thus clearly show to have better generalisation capabilities in this setting. These results show in a rather spectacular way that taking into account prior knowledge is a very promising avenue for enhancing machine learning methods in the context of high-dimensional problems. Comparing the middle and right part of the figure, we observe that it is preferable to use stronger learners inside the tree-nodes, specially when the sample size is small.

Figures 5.2, 5.3 and 5.4 show variable and group importances (average values over the ten runs; $T = 1000$) obtained for the three previous values of LS_{size}, and with $K = 1000$ for the Random Forests and $K = 100$, $IC = 10$, $K_{int} = 10$ for the *T-Trees*. Notice that the 10 first variables are those corresponding the causal block used to compute the output class. We see that, with $LS_{size} = 100$, the Random Forests are not at all able to identify any relevant variables while the variable and group importances computed from the T-Trees already correctly rank the first block of ten variables.

To further analyse the capabilities of the two methods to identify relevant variables, we display on Figures 5.5, 5.6, 5.7 the distribution of variable rankings over the 10 learning samples, obtained by these two methods and in comparison with a classical univariate p-value based ranking of the individual SNPs. For each method (from left to right : exact Fisher test based p-value, RF based importances, T-Tree based importances), these figures report a box plot giving the mean ranks for the 20 first variables (the 10 first variables are those of the causal block and are thus potentially relevant; on the other hand, the next ten are for sure completely irrelevant in the setting studied in this section) and the distribution of these ranks over the ten runs corresponding to the ten learning samples. We notice that with a small learning set (of size 100 - Figure 5.5), neither the p-values nor the random forest variables importances allow to identify the causal variables while the T-Tree variables importances already correctly rank the 10 first variables. With $LS_{size} = 250$ (Figure 5.6), p-values and random forest variable importances start to correctly rank some of the 10 potentially relevant variables, and with $LS_{size} = 500$ (Figure 5.7) the random forests do correctly rank all the 10 first variables while the p-ranking still only catches a subset of them.

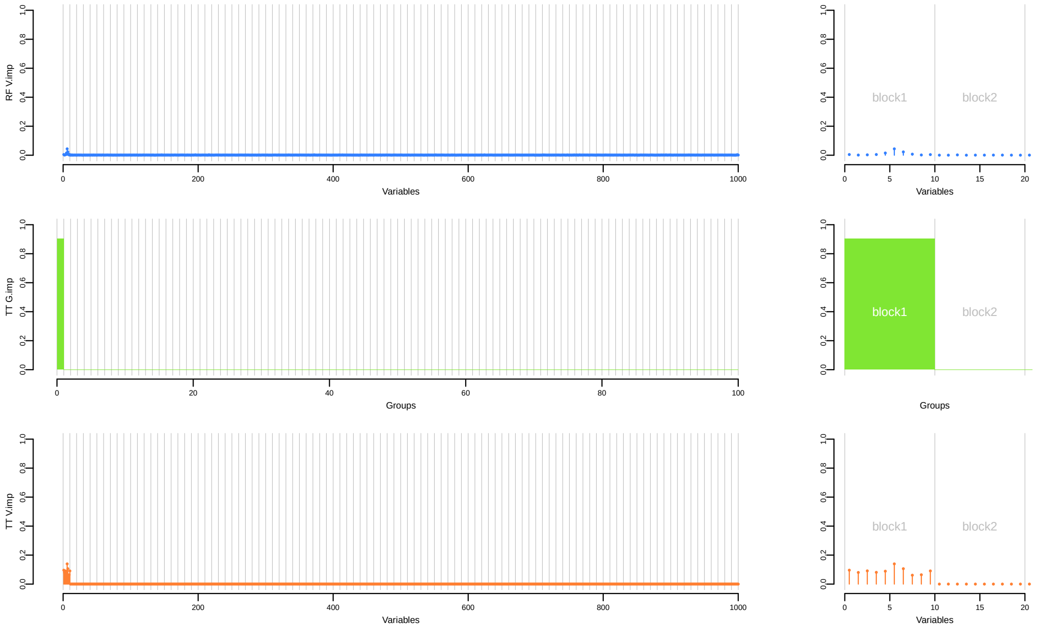


FIGURE 5.3 $LS_{size} = 250$: variable and group importances. In blue, the random forest ($K = 1000$) variable importances. In orange, the T-Trees variable importances and in green, the T-Trees group importances ($K = 100$, $K_{int} = 10$, $IC = 10$). The left part display an overview of all the variables while the right part shows a zoom around the causal block.

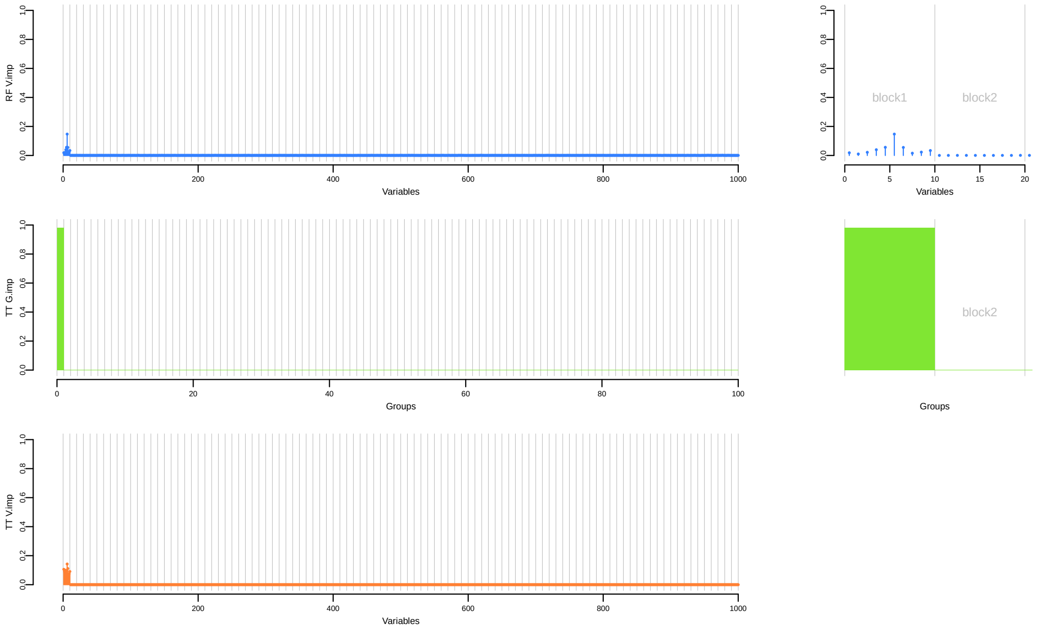


FIGURE 5.4 $LS_{size} = 500$: variable and group importances. In blue, the random forest ($K = 1000$) variable importances. In orange, the T-Trees variable importances and in green, the T-Trees group importances ($K = 100$, $K_{int} = 10$, $IC = 10$). The left part display an overview of all the variables while the right part shows a zoom around the causal block.

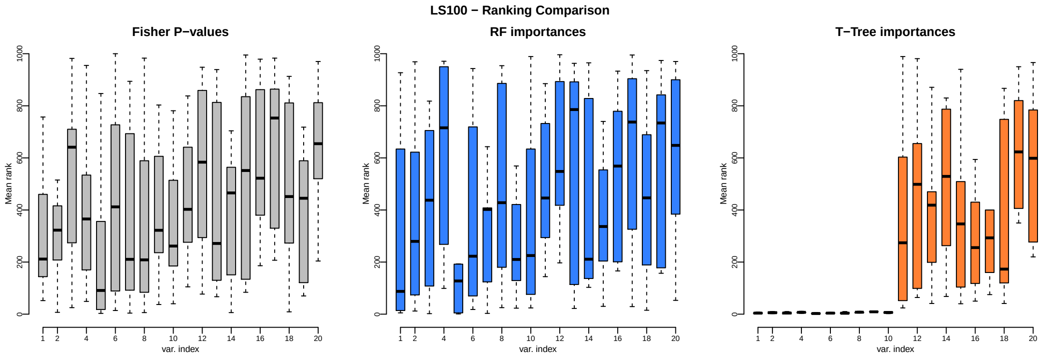


FIGURE 5.5 $LS_{size} = 100$: comparison of three different rankings. The three box plots represents the mean rank for the 20 first variables, the 10 first being the causal ones. In grey, the mean rank according to the Fisher p-value, in blue the mean rank according to the random forest variable importances ($K = 1000$) and in orange, the mean rank according to the T-Trees variables importances ($K = 100$, $K_{int} = N$ and $IC = 10$).

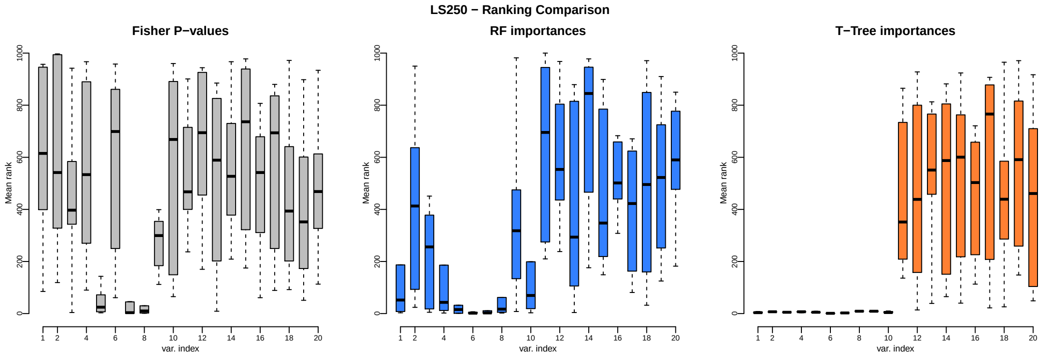


FIGURE 5.6 $LS_{size} = 250$: comparison of three different rankings. The three box plots represents the mean rank for the 20 first variables, the 10 first being the causal ones. In grey, the mean rank according to the Fisher p-value, in blue the mean rank according to the random forest variable importances ($K = 1000$) and in orange, the mean rank according to the T-Trees variables importances ($K = 100$, $K_{int} = N$ and $IC = 10$).

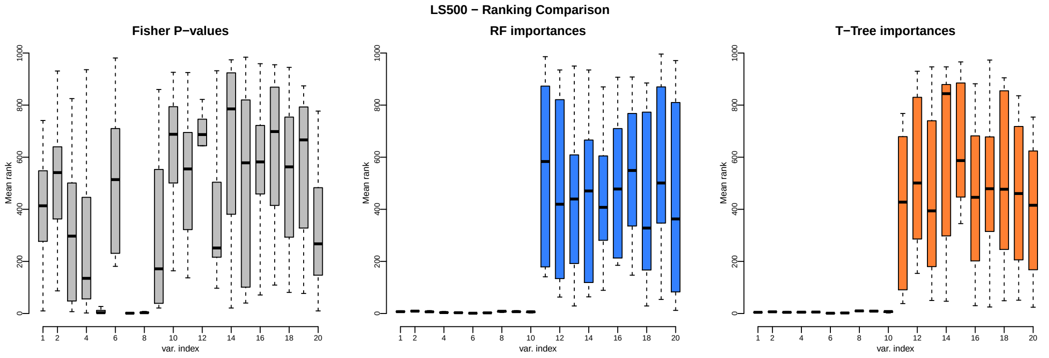


FIGURE 5.7 $LS_{size} = 500$: comparison of three different rankings. The three box plots represents the mean rank for the 20 first variables, the 10 first being the causal ones. In grey, the mean rank according to the Fisher p-value, in blue the mean rank according to the random forest variable importances ($K = 1000$) and in orange, the mean rank according to the T-Trees variables importances ($K = 100$, $K_{int} = N$ and $IC = 10$).

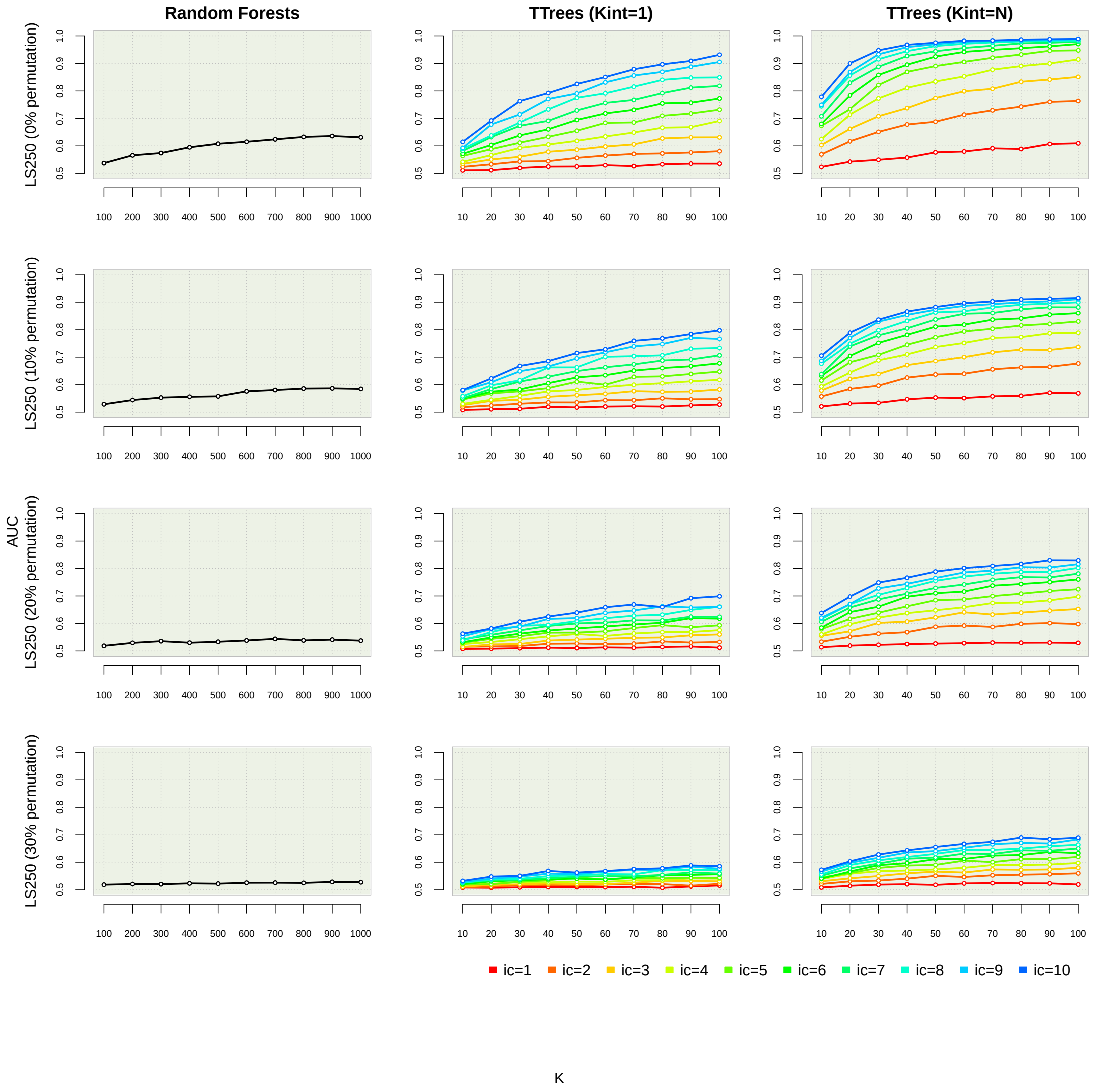


FIGURE 5.8 Robustness against label errors, influence of K and IC on the AUCs. The columns correspond to different methods (from left to right: Random Forests, T-Trees with $K_{int} = 1$, and T-Trees with $K_{int} = 10$). The rows correspond here to different percentages of permuted outputs, from 0 to 30%. For these simulations we used $LS_{size} = 250$.

5.3.2 Robustness against label errors

In order to test the robustness against labelling errors of our analysis of the previous section, we successively permuted 10%, 20% and 30% of the labels of the used dataset and then looked at the effect of constructing models on a learning set of size 250, while using the same protocol as in the previous section.

We report on figure 5.8 the corresponding AUC values. We observe that, as the percentage of permuted labels increases, the AUC decreases for all methods. But the performance of T-Trees AUCs are still much better (specially for $K_{int} = 10$ and $IC = 10$) than those of the RF method, and the best values obtained with this method are actually very close to the maximum possible values that could be expected in these conditions (i.e. respectively 100%, 90%, 80% and 70%, for degrees of permutation of 0%, 10%, 20% and 30%).

5.3.3 Influence of the quality of prior information about the block structure

In this subsection we want to see how much the T-Trees method is sensitive to the quality of the provided block map; indeed, on real datasets, we do not always know how the variables should be structured/grouped.

To assess this, the first row of Figure 5.9 exposes results obtained from a learning set of 250 objects with the actual block map used to generate the data as in the previous simulations (block size = 10), while various 'wrong' maps are used by the T-Trees algorithm (either where each of its used blocs includes 20 adjacent variables, or where each used bloc contains 50 adjacent variables) in the two last rows. On this figure, for the T-Trees, notice that the horizontal axis represents the internal complexity while the different shades of colour correspond to different values of K. For blocks of 20 variables, 50 blocks are available, and K thus ranges from 10 to 50 while IC ranges from 1 to 20. For blocks of 50 variables, K ranges from 10 to 20 while IC ranges from 1 to 50.

Interestingly, AUCs do not decrease so much with blocks of size 20 even with $IC = 20$. If we consider blocks of 50 adjacent variables, good results are still obtained but AUCs start dropping around $IC = 40$ as the internal trees become more complex, which probably leads the resulting trees to overfit the learning set.

In terms of variables ranking, Figure 5.11 displays a comparison of mean ranks according to the three previously used methods (Fisher p-values (in grey), random forests (in blue) and T-Trees (in orange) variable importances). We observe that the 20 variables in the first block are correctly ranked, we also see that the 10 first are the one coming from the real causal blocks. Inspection of the variables importances [results not shown here] (not the ranks) clearly shows that the 10 (real) causal variables are the most important. Figure 5.12 display similar results for a less extreme set of T-Trees parameters: $K = 10$ and $IC = 5$. It seems that when using a "not so bad block map", some of the variables may be wrongly associated as they are, in a way depending on the internal complexity, used jointly with the true signal (probably at the bottom of the small test-nodes located at the bottom of the T-Trees). We also notice that a simple inspection of the variable and group importances and of their ranking, and the investigation of different values of the internal complexity, allow to (visually) guess the true structure of variables (at least the causal ones).

Finally, we wanted to see how the T-Trees deal with a totally random (wrong) bloc map. Figure 5.10 compares the results we obtained with the correct bloc map (non overlapping blocks of 10 contiguous variables) versus the results with a random bloc map (blocks of 10 variables chosen randomly along the 'genome'). As expected, we see that when the variables are not correctly ordered the method is not able anymore to combine them efficiently. The best results of the T-Trees method are then obtained when $IC = 1$, which, as we observed already previously, reduces the performances of the method roughly to that of Random Forests using a same level of randomisation.

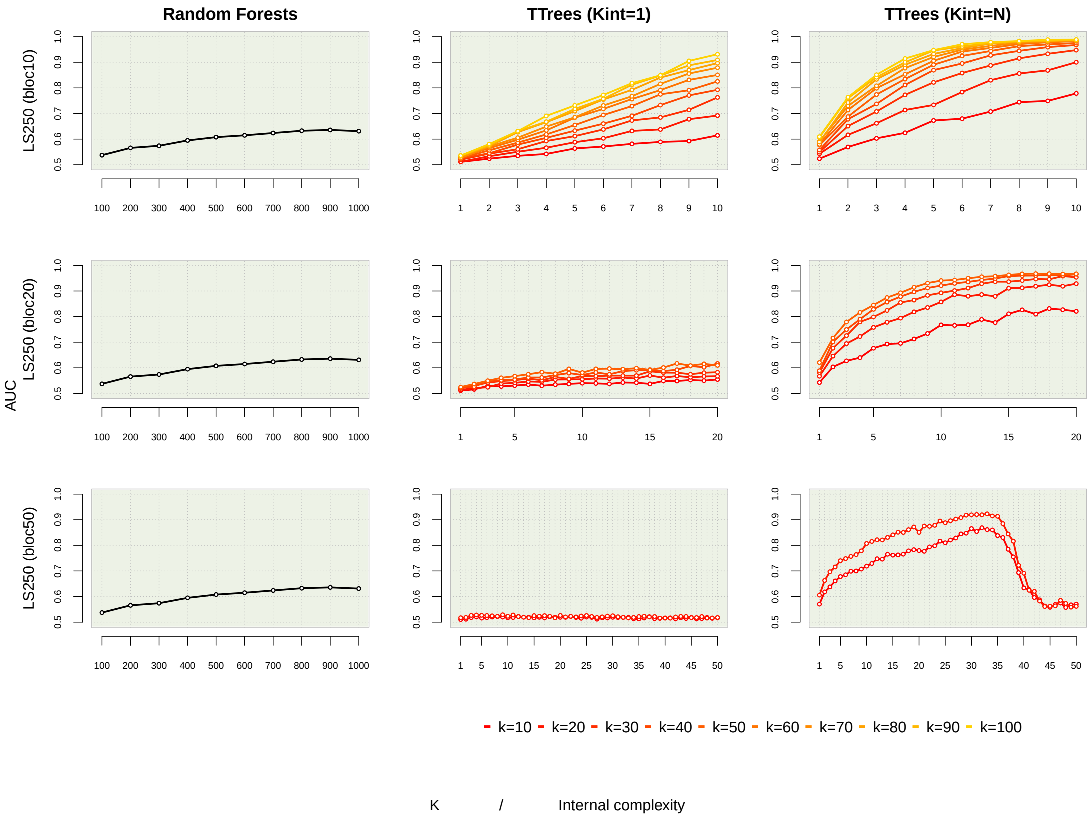


FIGURE 5.9 Influence of bloc size, K and IC on the AUC. The first row represent results with correct block size (10), the second one with blocks of size 20 and the third one with blocks of 50 variables. The first column represent the random forest results for values of K ranging from 100 to 1000. The two last columns correspond to results obtained with the T-Trees for different value of the internal complexity. The middle column correspond to the T-Trees with K_{int} set to 1 and the right columns with $K_{int} = 10$.

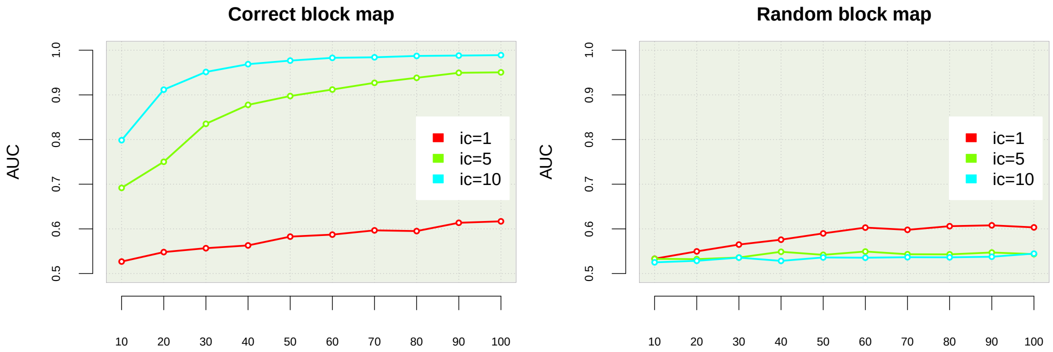


FIGURE 5.10 $LS_{size} = 250$ T-Trees ($K_{int} = N$): on the left, results obtained with the good block map (blocks of 10 contiguous variables). On the right results obtained with a random block map (blocks of 10 randomly chosen variables).

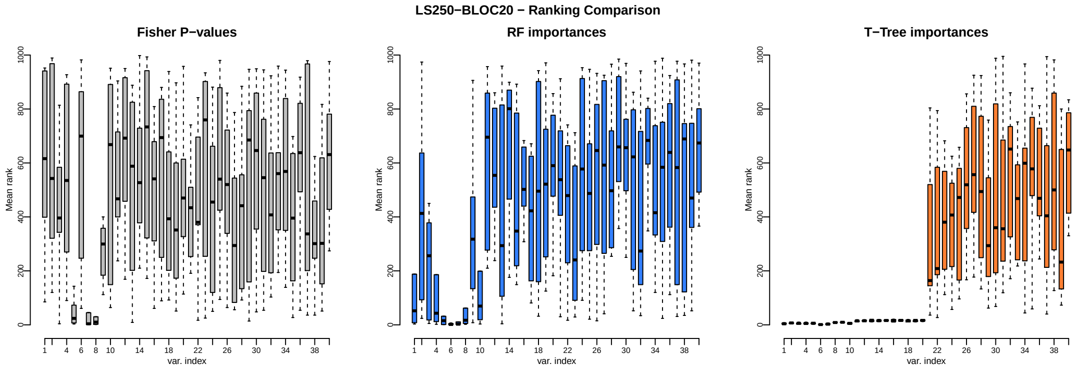


FIGURE 5.11 $LS_{size} = 250$ and block of 20 variables: comparison of three different rankings. The three box plots represents the mean rank for the 40 first variables, the 10 first being the causal ones. In grey, the mean rank according to the Fisher p-value, in blue the mean rank according to the random forest variable importances ($K = 1000$) and in orange, the mean rank according to the T-Trees variables importances ($K = 50$, $K_{int} = N$ and $IC = 20$).

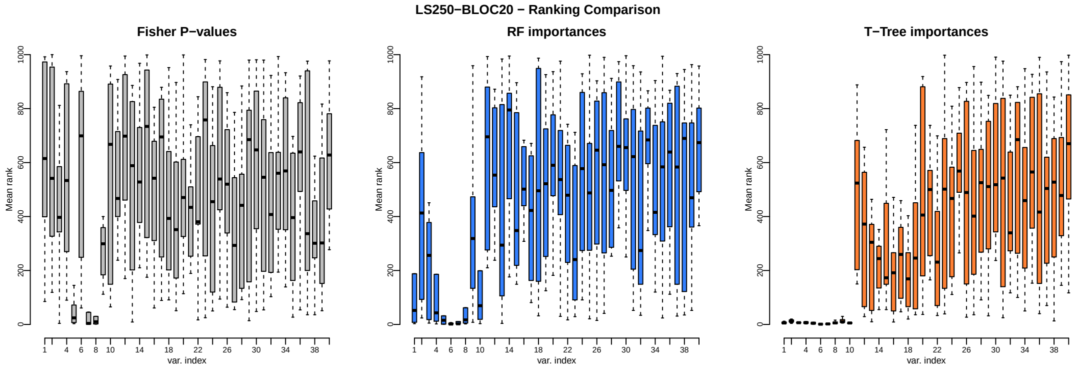


FIGURE 5.12 $LS_{size} = 250$ and blocks of 20 variables: comparison of three different rankings. The three box plots represents the mean rank for the 40 first variables, the 10 first being the causal ones. In grey, the mean rank according to the Fisher p-value, in blue the mean rank according to the random forest variable importances ($K = 1000$) and in orange, the mean rank according to the T-Trees variables importances ($K = 10$, $K_{int} = N$ and $IC = 5$).

5.3.4 More than one causal block

Too further analyse the comparative behaviours of the Random Forests and the T-Trees methods, we investigated a number of 'multi-loci' genotype-to-phenotype models based on Formula 5.1, as indicated in Table 5.1: with and without noise, with 2 and 3 causal blocks.

Dataset	#Variables	α_1	α_2	α_3	β
DB-1	1000	1	1	/	0
DB-2	1000	1.2	1	/	0
DB-3	1000	1.4	1	/	0
DB-4	1000	2	1	/	0
DB-5	1000	1	1	/	0.5
DB-6	1000	2	1	/	0.5
DB-7	1500	1	1	1	0

TABLE 5.1 Summary table of investigated phenotypes. The parameters α_i and β correspond to the weight of each causal block and the noise level, as defined by Formula 5.1.

On these different conditions, with a learning set of size 250, the obtained results (see Figure 5.13) highlight the influence of the parameters K and IC on the predictive power. As before, as K and IC increase, the AUC increases as well. We can also conclude that the *T-Trees* method (with the correct block map) clearly outperforms the Random Forests method when IC is appropriately set to 10. As a matter of fact, we see that Random Forests are not able to deal with 3 interacting blocks.

To complete the analysis, Figure 5.14 shows variable and group importances in the context of the 3-loci model of DB-7: we conclude that *T-Trees* are still able to correctly rank the 30 causal variables contained inside the 3 blocks, while the Random Forests are not extracting any valuable signal in these conditions.

5.3.5 Modality, scale of measurement and number of categories

In this section we want to address the problem mentioned in [BBLBS12]: the authors established a variable importance bias in favour of large minor allele frequency SNPs.

Besides the fact that we are working with SNPs, the concept of minor allele frequency is closely related to the variable importance bias introduced by the scale of measurement or the number of categories of each variables (see [SBZH07]). With the *T-Trees*, working with groups of SNPs may introduce such a bias. Indeed, as the linkage disequilibrium markers varies along the chromosome (and also depends on the genotyping density), the numbers of SNP combinations generated from the different blocks may differ. In other words, the numbers m_i of block modalities may introduce the same bias. In order to test that hypothesis we generated blocks of growing modalities, compiled them in a single dataset with a random output. Figure 5.15 shows the resulting variable and group importances for block modalities $10, 20, 30, 40$ and 50 while Figure 5.16 corresponds to modalities $10, 50, 100$ and 500. We observe that as the modality m_i increases the resulting variable and group importances indeed slightly increase as well.

We then introduced one actual causal block (see Figure 5.17) and two actual causal blocks (see Figure 5.18) on those blocks having the smallest number m_i of configurations, so as to test how that small pathological bias influences the importances when there is a true signal. We notice that as the true signal(s) is (are) still quite well detected, so that the above bias does not seem to be a major source of false positives in practice.

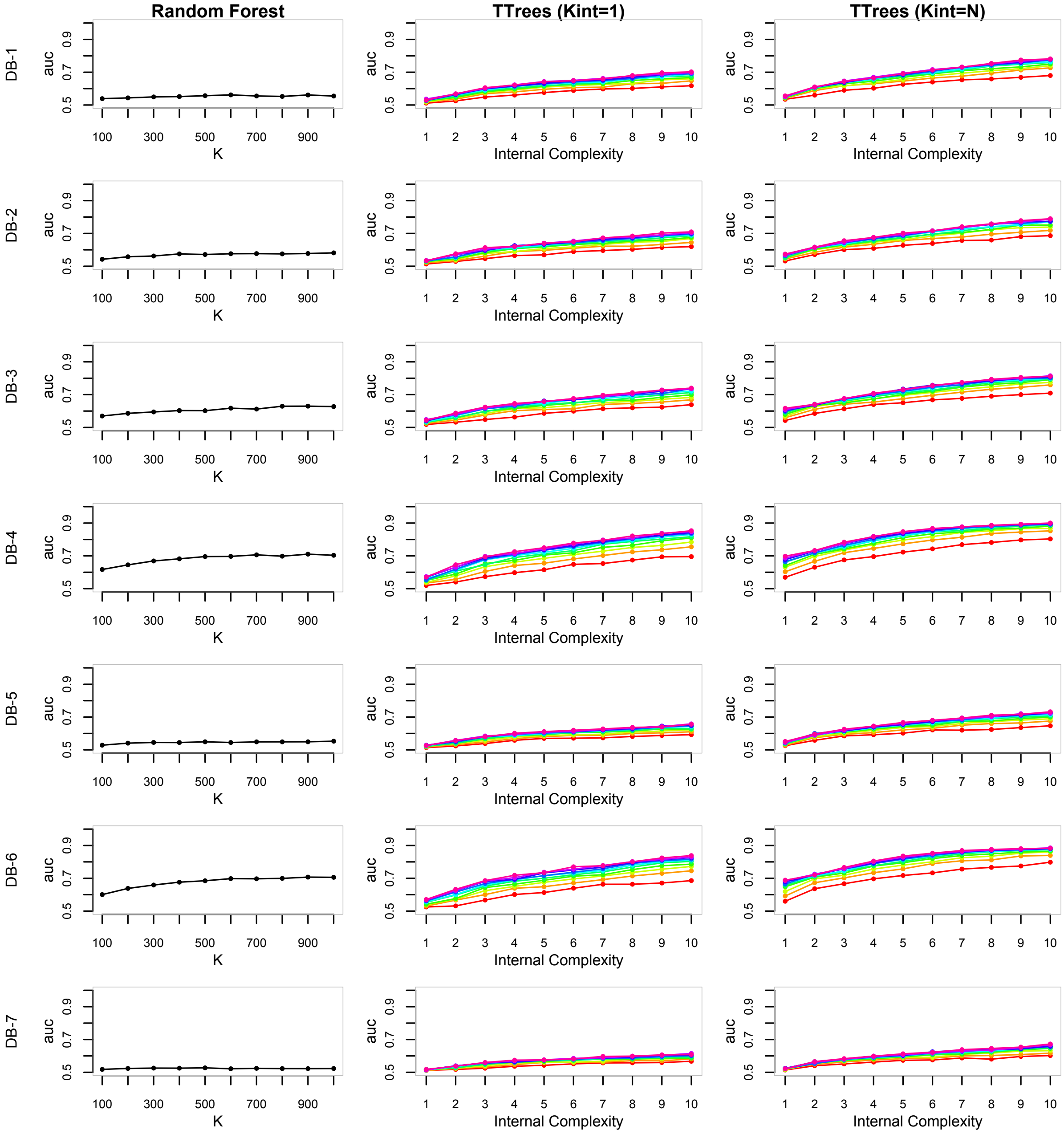


FIGURE 5.13 Influence K and IC on the AUC on different 'phenotype' models. Each row corresponds to one of the models described in Table 5.1. The first column represents the Random Forest results for values of K ranging from 100 to 1000. The two last columns correspond to results obtained with the T-Trees for different value of the internal complexity. The middle column correspond to the T-Trees with K_{int} set to 1 and the right columns with $K_{int} = 10$.

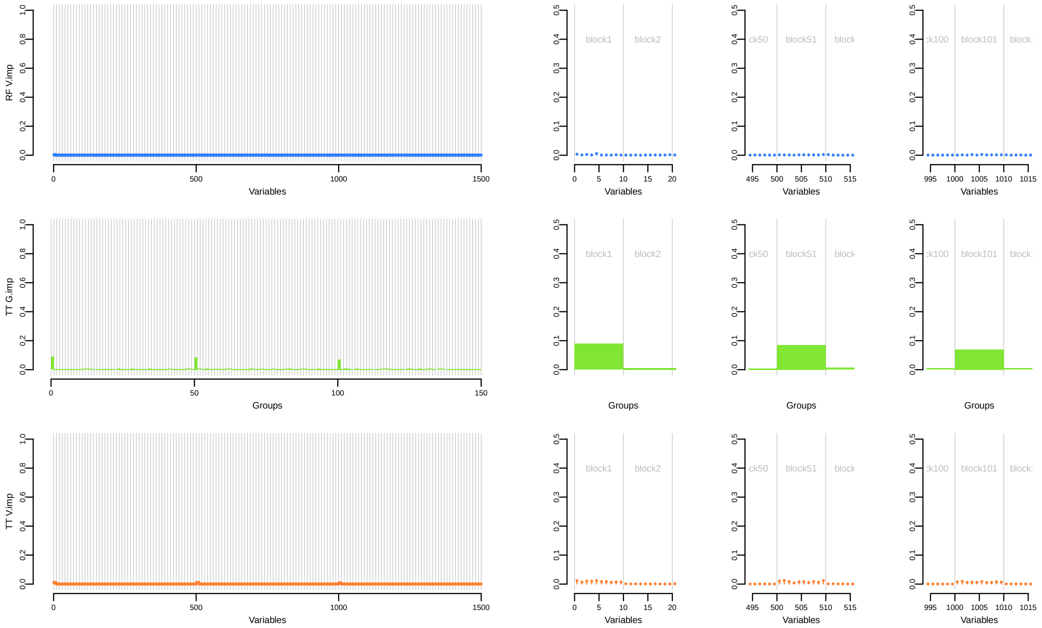


FIGURE 5.14 DB7: variable and group importances. In blue, the random forest ($K = 1500$) variable importances. In orange, the T-Trees ($K = 150$, $K_{int} = 10$ and $IC = 10$) variable importances and in green, the T-Trees group importances. The left part display an overview of all the variables while the right part shows a zoom around the three causal blocks.

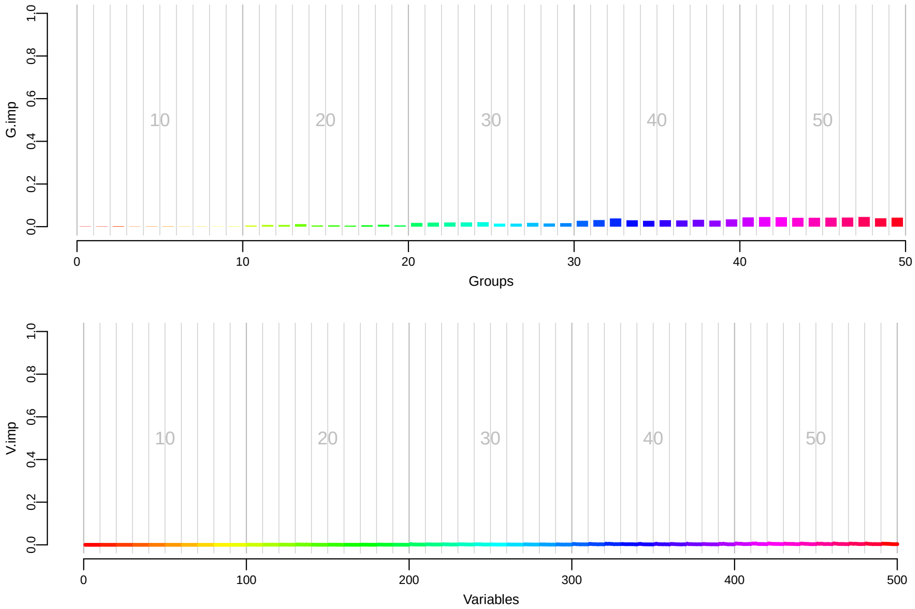


FIGURE 5.15 Results with different block modalities spread over 50 blocks, the 10 first blocks modalities are equal to 10, the 10 following 20 and so on up to 50 for the last ten blocks. The upper plot corresponds to group importances and the bottom plot to variable importances obtained with the T-Trees methods ($K = 25$, $K_{int} = 10$ and $IC = 10$) on a dataset with no causal variable/block.

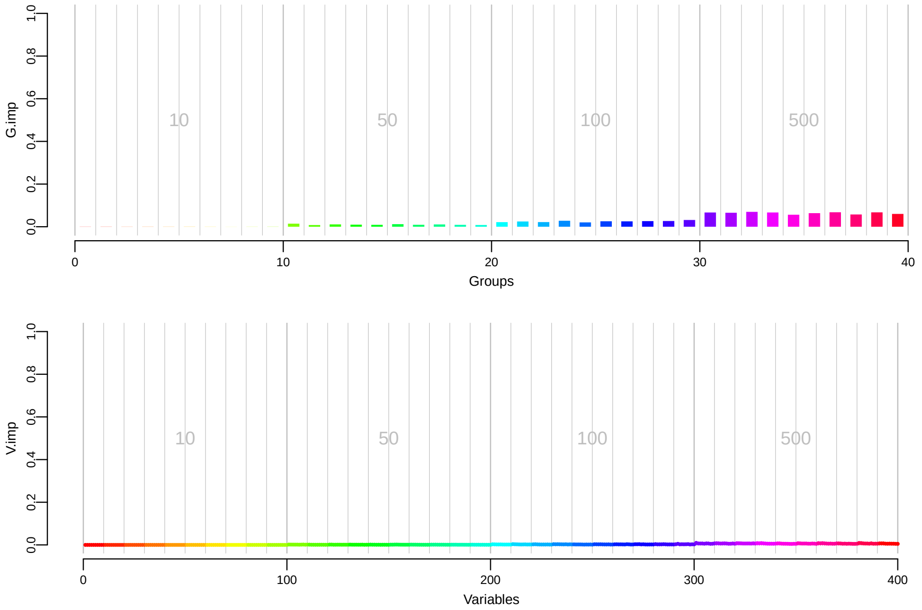


Figure 5.16 Results with different (higher) block modalities spread over 40 blocks, the 10 first blocks modalities are equal to 10, the 10 following 50, followed by 10 blocks of modality equals 100 and to 500 for the last ten blocks. The upper plot corresponds to group importances and the bottom plot to variable importances obtained with the T-Trees methods ($K = 20$, $K_{int} = 10$ and $IC = 10$) on a dataset with no causal variable/block.

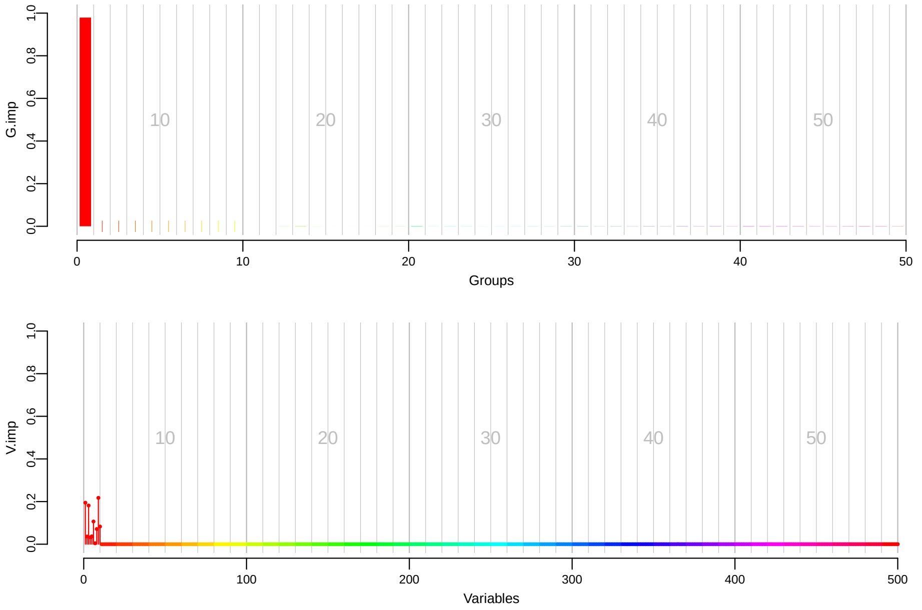


Figure 5.17 Results with different block modalities spread over 50 blocks, the 10 first blocks modalities are equal to 10, the 10 following 20 and so on up to 50 for the last ten blocks. The upper plot corresponds to group importances and the bottom plot to variable importances obtained with the T-Trees methods ($K = 25$, $K_{int} = 10$ and $IC = 10$) on a dataset with one causal block (the first one).

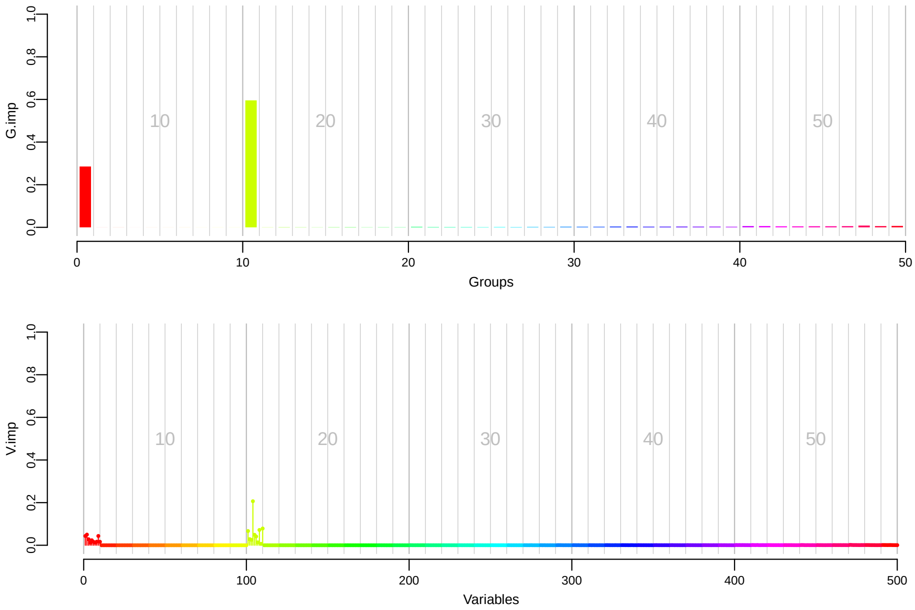


FIGURE 5.18 Results with different block modalities spread over 50 blocks, the 10 first blocks modalities are equal to 10, the 10 following 20 and so on up to 50 for the last ten blocks. The upper plot corresponds to group importances and the bottom plot to variable importances obtained with the T-Trees methods ($K = 25$, $K_{int} = 10$ and $IC = 10$) on a dataset with two causal blocks (block 1 and 11).

5.4 Discussion

In this chapter, we compared the *T-Trees* to the Random Forests on the basis of synthetically generated datasets aiming at mimicking the conditions of real GWAS. Even tough our artificial datasets are certainly a rather naïve approximation of real GWAS datasets, we were able to screen the various questions that one wants to investigate in this context, and we observed an encouraging difference between the T-Trees and the Random Forest methods. That difference is quite striking and may point out a Random Forests weakness to detect the correct combination of variables even with a relatively small number of correlated features.

We also noticed that, specially with a small number of learning samples, the T-Trees produced much better results than the Random Forest method in terms of AUCS. In the case of GWAS where the number of samples is often difficult to collect and genotype, it may be of crucial help to be able to extract relevant information out of a small number of individuals. In terms of the identification of relevant variables, we notice that T-Trees variable importances allowed to correctly rank the 10 causal variables in a very robust fashion. We also found out that the Fisher test of association and the Random Forest derived variable importances seem to require significantly larger sample sizes in order to detect and rank the associated variables, in those conditions that we used to generate our synthetic datasets.

In the case of the previous simulations, we were expecting an $IC = 10$ to be the best choice given the design of our datasets. Still, with most of the values < 10 of IC we obtained better AUCS than with the Random Forest method. In practice, the choice of the value for the internal complexity will depend on how the blocks are structured. When dealing with strongly correlated variables inside the blocks, lower values of the internal complexity should be a good choice. In the extreme case where all the variables of the blocks are perfectly correlated, exploiting one of them would suffice to capture the complete group information, thus an internal complexity of 1 would then be recommended. On the other hand, if we expect combinations of variables of a particular block to be linked to the output, as it was the case in the previous experiences, we recommend using an internal complexity greater than 1.

We noticed that the block map provided to the method does not need to be precise. Indeed, we observed that using “larger than required” block maps did not dramatically change the results while blocks of random variables lead to poor results. This suggests that the main thing that matters when defining a block map is to preserve the structure of the blocks. In real life problems, the assessment of a precise/right block definition is not always possible and may therefore remain uncertain. The investigation of variable and group importances can however help in finding a good value for the block sizes and may thus help to understand how the genotypic variables are structured.

We compared 2 values of the T-Trees meta parameter K_{int} which is used to control the “weakness” of the internal trees. In most cases, we were still able to beat the random forest with the weaker learner inside the splitting nodes. It can reduce the computing times required at the learning step and thus be helpful in the early stage of an experiment. But, we noticed that under some conditions (such as in the bad block map experiment or in the presence of noise), totally randomised trees inside test nodes leads to quite suboptimal performances. Therefore, we will fix the value of K_{int} to the size of the blocks in the two following chapters, when applying the T-Trees to real datasets, in spite of the higher computational cost implied by this decision.

Chapter 6

The case of *Crohn*'s disease

Contents

Crohn's disease is a form of inflammatory bowel disease (IBD). It usually affects the intestines, but may occur anywhere from the mouth to the end of the rectum. It mostly causes abdominal pain, diarrhea, vomiting and may cause weight loss. It is known to be a complex disease, it is a result of environmental, immunological and bacterial factors combined with a genetic predisposition. Current treatments imply lifestyle changes (such as dietary adjustments, proper hydration or smoking cessation), medication and, in the more extreme cases, surgery. Even with a full battery of tests (endoscopy, radiologic tests, blood tests), current diagnoses of *Crohn*'s disease are difficult to establish with 100% certainty. Combining the intrusive aspects of diagnosis, the important required lifestyle changes, the absence of real effective treatments and all the inconveniences related to the disease motivates the efforts of geneticists in identifying the genetic underlying causes of such pathology.

This chapter focuses on a *Crohn*'s disease dataset. The first part investigates the predictive power, the localization of causative genes ability, and the influence of quality control filters applied to the dataset on the different tree-based methods discussed in Chapter 4. The second part is dedicated to the influence of the different meta-parameters of the different methods. This chapter ends with recommendations for the application of tree-based methods in the field of genome-wide association studies.

6.1 Two dataset variants for *Crohn*'s disease

In this section, we will show how tree-based ensemble algorithms can be used in the fields of genome-wide association studies. In order to validate our approach, we had access to a huge data collection coming from the Wellcome Trust Case Control Consortium (WTCCC [Wel07]). It contains 17.000 genotypes, it is composed of 3.000 shared controls and 14.000 cases and it concerns 7 common diseases of major public health (including *Crohn*'s disease, abbreviated here CD). The genotypes are described by 500.000 SNPs (genotyped with the Affymetrix GeneChip 500K Mapping Array Set).

As described in [Wel07], the data quality controls where applied at a study-wide range. The sample-based quality controls excluded:

- 250 samples with $> 3\%$ missing rate across all SNPS,
- 6 samples for excess of heterozygosity ($> 30\%$),
- 3 samples for low heterozygosity ($< 23\%$),
- 16 samples because of discrepancies between WTCCC information and external identifying information,
- 153 individuals clearly not caucasian (compared to the HapMap CEU population),
- 295 duplicated samples ($> 99\%$ identity) and,
- 86 related samples (between 86% and 99% identity).

A total of 809 individuals were removed, leaving 4686 individuals in the WTCCC CD dataset. The marker-based QC were less stringent, the following markers are excluded:

- 26567 SNPS with a missing data rate $> 5\%$ or $> 1\%$ for the markers exhibiting a study-wise minor allele frequency $< 5\%$,
- 4351 markers with a HWE exact p-value $< 5.7 \times 10^{-7}$ in the combined set of controls and,
- 93 markers with a p-value $< 5.7 \times 10^{-7}$ for either a one- or two-degree of freedom test of association between the two control groups.

A total of 469557 SNPS remained in the study. They choose to apply light quality control filters but to visually inspect (the cluster plots of) all the apparently associated SNPS. We will see that this assumption is of crucial importance, particularly when using multivariate approaches such as tree-based methods. Out of the bulk WTCCC download, we used the CHIAMO output to regenerate a dataset using the following rule to determine the missingness of each genotype: are considered as missing the genotypes for which the CHIAMO score is below 0.9. As our methods do not deal with missing values as such, for each SNP we chose to randomly fill the unknown genotypes taking into account the genotypic distribution of the corresponding non missing values of the SNP. Finally, we excluded markers and individuals based on the provided exclusion lists (found in the WTCCC bulk download).

On the other hand, we also got access to a variant of the CD dataset from the *Inflammatory Bowel Disease* (IBD) Consortium. This copy of the WTCCC samples has been strongly filtered. In terms of missingness, a comparison between the WTCCC bulk download and the IBD consortium showed that the missing values were determined following the same procedure (perfect match in position for the missing values in the two genotype matrices). We also randomly filled the missing genotypes following the same approach as for the previous WTCCC dataset.

On that second version of the CD dataset, the following QC filters have been applied:

- missing rate per SNP $< 5\%$ (before sample removal)

- missing rate per individual $< 2\%$
- heterozygosity per individual $+/-0.2$
- missing rate per SNP $< 2\%$
- missing rate difference between case and control $< 2\%$
- HWE p-value $< 10^{-6}$ (controls only)
- HWE p-value $< 10^{-10}$ (cases only)

We will use the two resulting datasets to compare how the QC filters may impact the quality of the inducted trees. In the following, we denote those two variants of the datasets as follow:

- CD_{wtccc}: the WTCCC **lightly** filtered (≈ 470000 SNPs and 4686 individuals (1748 cases and 2938 controls)),
- CD_{ibd}: the IBD **strongly** filtered (≈ 436000 SNPs and 4676 individuals (1739 cases and 2937 controls)).

6.1.1 Predictions

In this section, we will compare the previously discussed methods in terms of their predictive power. All the following results are obtained in a 10-folds cross-validation way and are expressed in 10-folds average AUCs. We used the score measure $\mathbf{S}_C^T$ for all the experiments (see Section C). We study the influence of the different parameters for each of the three methods: Random Forests, Extra-Trees and T-Trees. Our goal here is twofold: first, we want to evaluate the influence of the preprocessing steps applied to the CD dataset and, second, we want to compare the three methods.

Random Forests

Let us start with the Random Forest results on the CD_{wtccc} dataset, shown in Figure 6.1. In this figure, the six upper plots correspond to the different values of $K = 100, 500, 1000, 2500, 5000$ and 10000, and each such plot depicts the influence of the parameters $N_{\min}$ (larger values correspond to stronger pre-pruning) and T, de number of trees in the ensemble. The bottom part shows the evolution of the AUC as K increases: the plain line corresponds to maximum values of the AUC (corresponding to the maximum in each of the six upper plots) while the dotted line corresponds to the minimum values. On the left part is represented the rainbow colour scale used in the upper "heatmaps". From these observations, we see, as expected, that, for K and T, the larger they are, and the better the results in terms AUC. While for the pre-pruning parameter $N_{\min}$ it seems that pruning the trees does not dramatically degrade the quality of the predictions; as a matter of fact, it is even favourable as it slightly increases the AUC while reducing the computation time required. We also notice that pruning is mostly beneficial when we consider a forest with a smaller number of trees. The maximum AUC of 0.919 we observed was obtained for $K = 10000$ with $N_{\min} \approx 110$ while the minimum reached was 0.893 with a $N_{\min} \approx 2000$.

Similarily, Figure 6.2 shows the results we obtained on the CD_{ibd} dataset. The predictive power is significantly smaller on this dataset. The maximum of 0.7 is obtained with $K = 2500$ with $N_{\min} \approx 1050$ while a minimum of 0.673 is reached without pruning the trees. Meanwhile, for $K = 10000$ we obtained AUCs ranging from 0.671 (with no pruning) to 0.697 (with $N_{\min} \approx 870$).

The gap between CD_{ibd} and CD_{wtccc} demonstrates the impact of the stronger QC filters that have been applied to the CD_{ibd} dataset implying the removal of about 34000 variables, as compared to the CD_{wtccc} dataset.

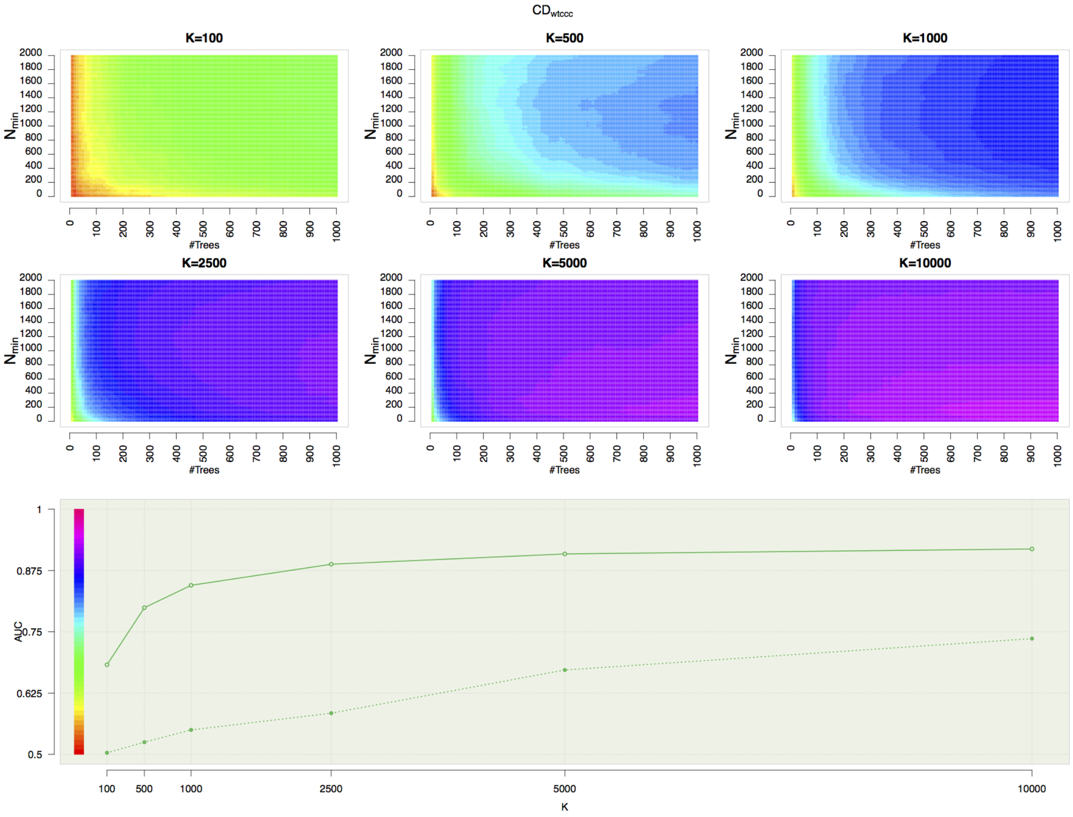


FIGURE 6.1 Random Forest: influence of T, $N_{\min}$ and K on CD_{wtccc}. The six upper panels show the influence of T and $N_{\min}$ for each of the six investigated values of K. The last panel displays the evolution of the AUC as the value of K increases, the plain line corresponds to the maximum and the dotted line to the minimum obtained at each value of K.

The influence of $N_{\min}$ is illustrated at Figure 6.3 for Random Forests with $K = 10000$. On CD_{wtccc}, we see that it is preferable to prune a bit but not too much as it decreases the predictive power. On the other hand, on CD_{ibd}, pruning slightly increases the area under the ROC curve. In these two cases, optimising the pruning parameter $N_{\min}$ allowed for an AUC gain of 0.02. While that gain is negligible in terms of accuracy, it is still of interest as it speeds up the computational time required to construct a forest.

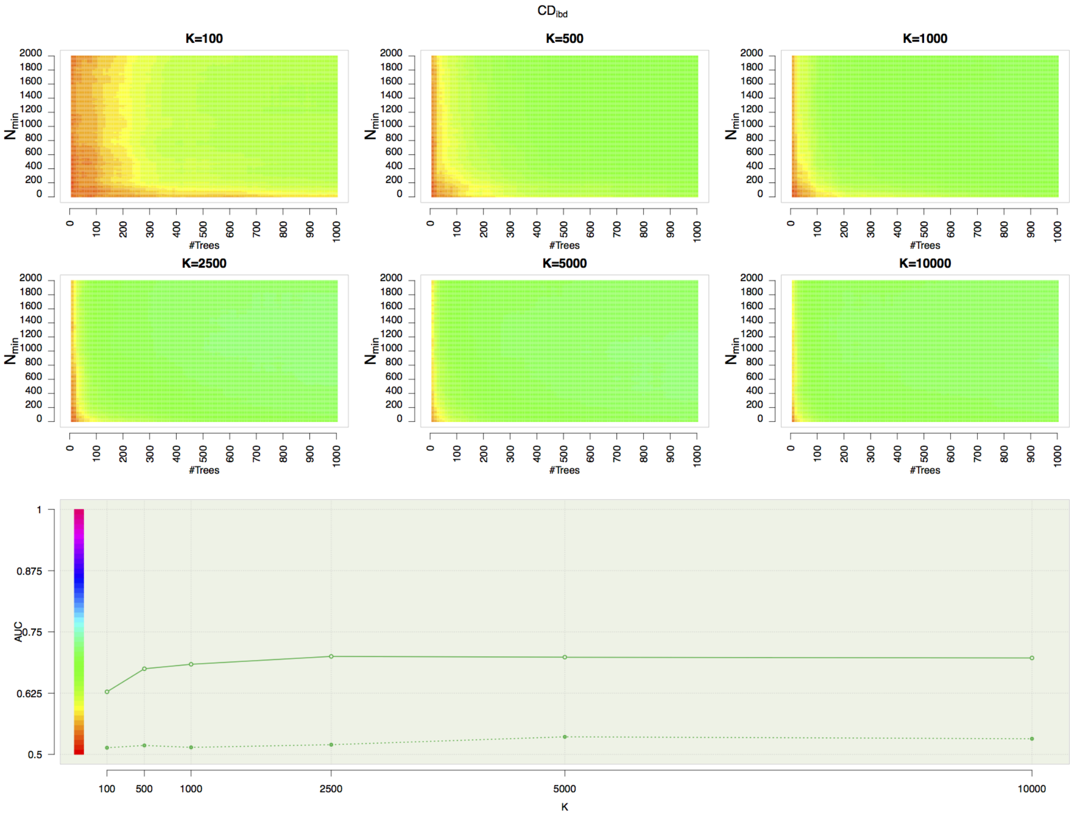


FIGURE 6.2 Random Forest: influence of T, N_{min} and K on CD_{ibd}. The six upper panels show the influence of T and N_{min} for each of the six investigated values of K. The last panel displays the evolution of the AUC as the value of K increases, the plain line corresponds to the maximum and the dotted line to the minimum obtained at each value of K.

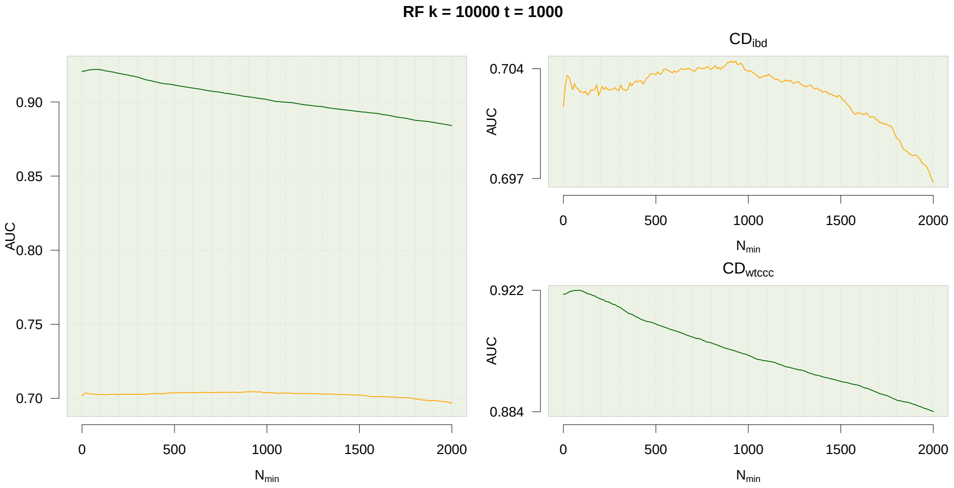


FIGURE 6.3 Dataset comparison: Random Forest $K = 10000$ $T = 1000$. Evolution of the AUC for the optimal value $K = 10000$. CD_{wtccc} and CD_{ibd} results are, respectively, represented in green and orange. The left panel shows a comparison between the two variants while, the two right panels show a detailed view of the corresponding curves.

Extra-Trees

Similarly, Figure 6.4 shows the results we obtained on CD_{wtccc} with the Extra-Trees. Again optimal results are obtained at $K = 10000$, minimum and maximum, respectively are, 0.884 (with $N_{\min} \approx 2000$) and 0.922 (with $N_{\min} \approx 90$).

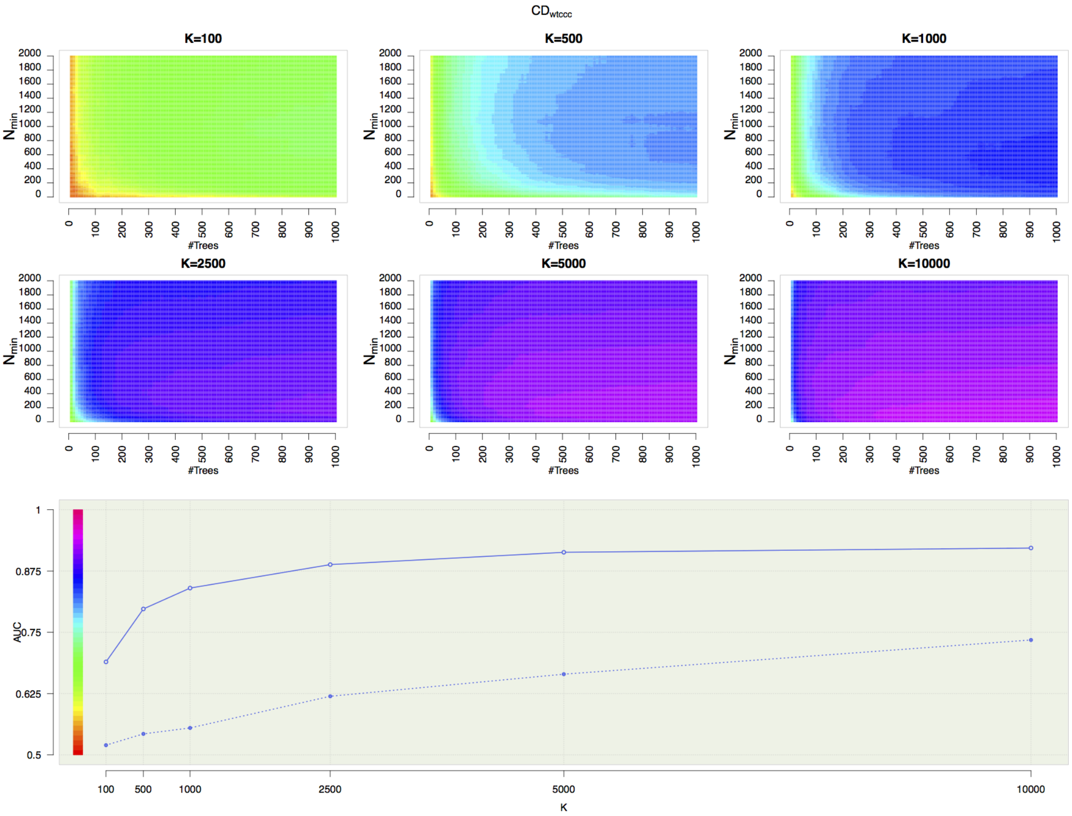


FIGURE 6.4 Extra-Trees: influence of T, $N_{\min}$ and K on CD_{wtccc}. The six upper panels show the influence of T and $N_{\min}$ for each of the six investigated values of K. The last panel displays the evolution of the AUC as the value of K increases, the plain line corresponds to the maximum and the dotted line to the minimum obtained at each value of K.

Finally, Figure 6.5 depicts the results of Extra-Trees for CD_{ibd}: as with Random Forests, optimal AUCs are obtained with $K = 2500$, AUCs ranging from 0.708 ($N_{min} \approx 0$) to 0.723 ($N_{\min} \approx 1160$) while with $K = 10000$, AUCs range from 0.697 ($N_{min} \approx 2000$) to 0.704 ($N_{\min} \approx 1000$).

Finally, Figure 6.6 shows the influence of $N_{\min}$ for $K = 10000$. On CD_{ibd}, the AUC is almost insensitive to $N_{\min}$.

Random forests versus Extra-Trees

Table 6.1 summarises the results. A first comparison can be made regarding the two methods. Extra-Trees perform slightly better than the Random Forests especially on the CD_{ibd} dataset. While the difference is not striking, we noticed that the Extra-Trees produces deeper trees, thus the expected time gain from its algorithmic properties is compromised (in addition to the randomisation that is not so random due to the discrete nature of SNPs: at most there are only two possible splits for each variable), on the other hand the Random Forest produces trees that are a bit smaller. Beside that small difference in term predictive power, the behaviour of the two methods remained consistent across the two dataset variations.

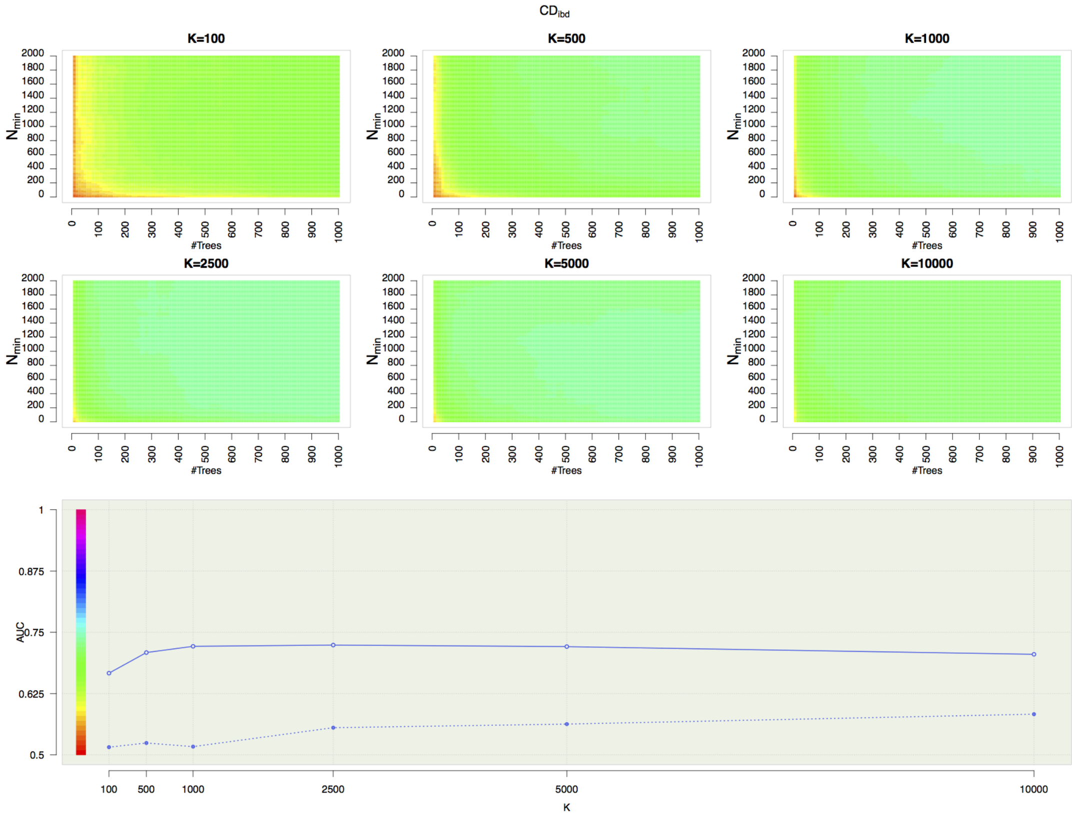


FIGURE 6.5 Extra-Trees: influence of T, $N_{\min}$ and K on CD_{ibd}. The six upper panels show the influence of T and $N_{\min}$ for each of the six investigated values of K. The last panel displays the evolution of the AUC as the value of K increases, the plain line corresponds to the maximum and the dotted line to the minimum obtained at each value of K.

A second comparison between the two datasets pinpoints the influence of the QC filters prior to the learning stage. Regardless of the used method, we notice that the AUCs produced on more lightly filtered version of the dataset (CD_{wtccc}) are way higher than those obtained with the more strongly filtered variant (CD_{ibd}). This shows that QC filters may crucially influence the predictive power obtained with tree-based methods. As we can see, the AUCs drop from 0.919 to 0.7 with the RF and from 0.922 to 0.724 with the ET. Also on CD_{ibd}, we see that AUCs reaches a top around $K = 2500$ while on the other variant it seems that potentially higher values of K (> 10000) could still improve the resulting AUCs.

Besides these observations, from the biological point of view, AUCs reaching a value of 0.92 are probably over-optimistic. Indeed, *Crohn*'s disease being a complex disease, we do not expect our prediction to be so good as we know that environmental factors may play an important role in the disease mechanism in addition to the fact that we do not have at our disposal, for each individual, the complete genetic information.

In Section 6.1.2, we will investigate the variable importances and will determine which of the variables and QC filters are responsible for the important AUC fluctuations we just observed.

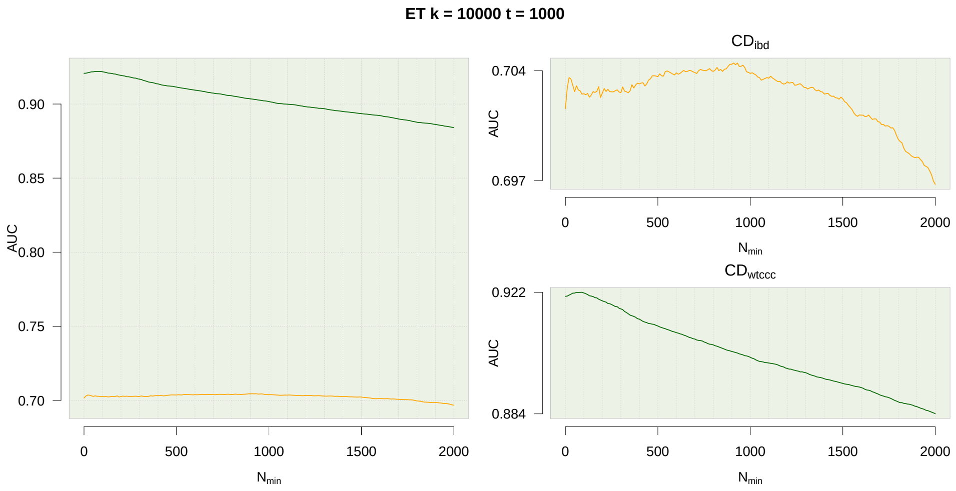


FIGURE 6.6 Dataset comparison: Extra-Trees $K = 10000$ $T = 1000$. Evolution of the AUC for the optimal value $K = 10000$. CD_{wtccc} and CD_{ibd} results are, respectively, represented in green and orange. The left panel shows a comparison between the two variants while, the two right panels show a detailed view of the corresponding curves.

	Random Forest		Extra-Trees	
K	CD_{wtccc}	CD_{ibd}	CD_{wtccc}	CD_{ibd}
100	0.683	0.628	0.690	0.667
500	0.799	0.675	0.798	0.709
1000	0.845	0.684	0.840	0.721
2500	0.888	**0.700**	0.888	**0.724**
5000	0.909	0.698	0.913	0.721
10000	**0.919**	0.697	**0.922**	0.705

TABLE 6.1 Summary table: the maximum AUCs obtained with the RF and the ET for the different values of K and $T = 1000$. Columns maxima are highlighted in bold.

T-Trees

We will now investigate our method: the T-Trees. In the following, we applied our novel approach in the same manner than we did for the two previous methods. In the next results, we decomposed the set of variables into blocks of 10 contiguous SNPs. We will investigate later on the impact of using other block sizes.

Figure 6.7 shows the results we obtained on the CD_{wtccc} dataset using the T-Trees and an internal complexity of 10. The method beats the two previous ones. Previously maximum AUCs are reached with smaller values of K. (note: but, in the context of T-Trees, K now represents the number of blocs being tested at each node which means that potentially $K \times 10$ SNPs are investigated at each tests). We observed an AUC of 0.945 with $K = 1000$ and $N_{\text{min}} = 2000$ while no pruning produced an AUC of 0.941 which is quite better than the best results we obtained with Random Forests and Extra-Trees.

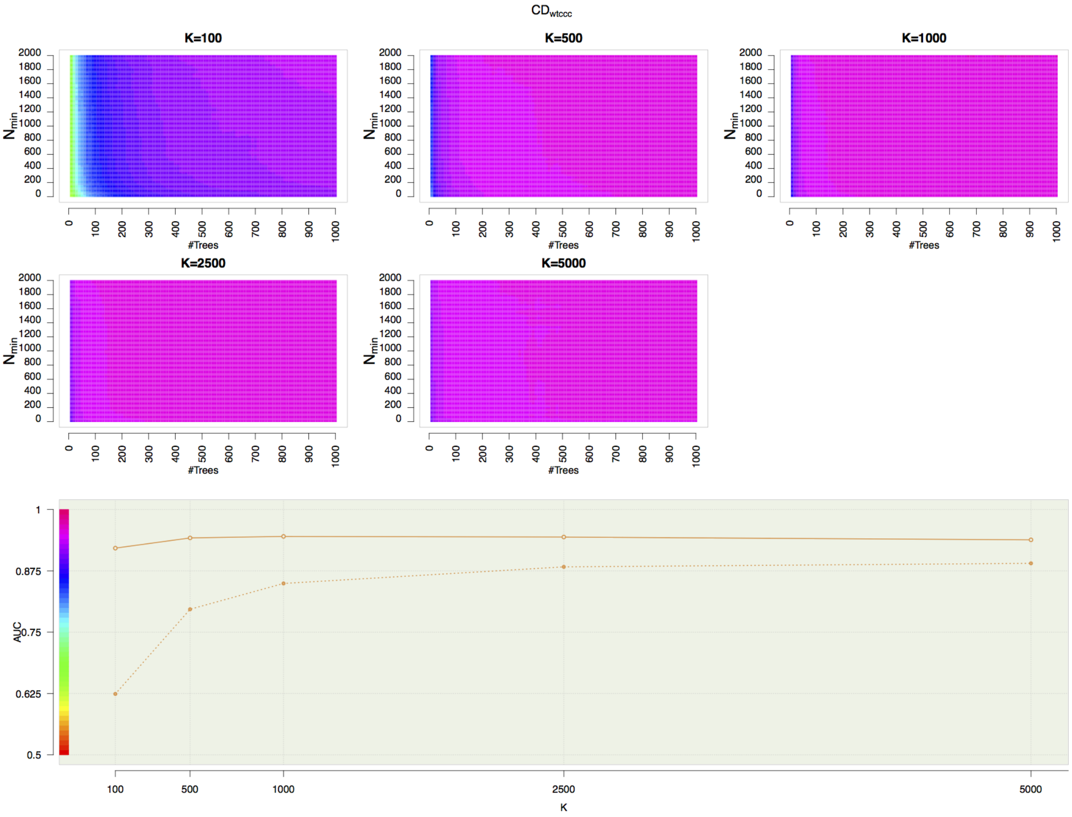


FIGURE 6.7 On CD_{wtccc}: T-Trees prediction performance with 10-SNPs blocks and $IC = 10$. The five upper panels show the influence of T and N_{min} for each of the five investigated values of K. The last panel displays the evolution of the AUC as the value of K increases, the plain line corresponds to the maximum and the dotted line to the minimum obtained at each value of K.

Finally, Figure 6.8 displays the results we obtained on CD_{ibd}. Again, $K = 1000$ is the "winner", a value of 0.748 is reached with $N_{\text{min}} = 2000$ while without pruning we observed a value of 0.744.

We notice that pruning does not drastically modify the results, in fact, it allowed for a small increase in the AUCs we just observed. Therefore, it seems advised to prune the trees that are built in a forest of T-Trees since, on the one hand, it does not deteriorate the accuracy results at all, and, on the other hand, allows to reduce the computing time required. Figure 6.9 summarises these conclusions.

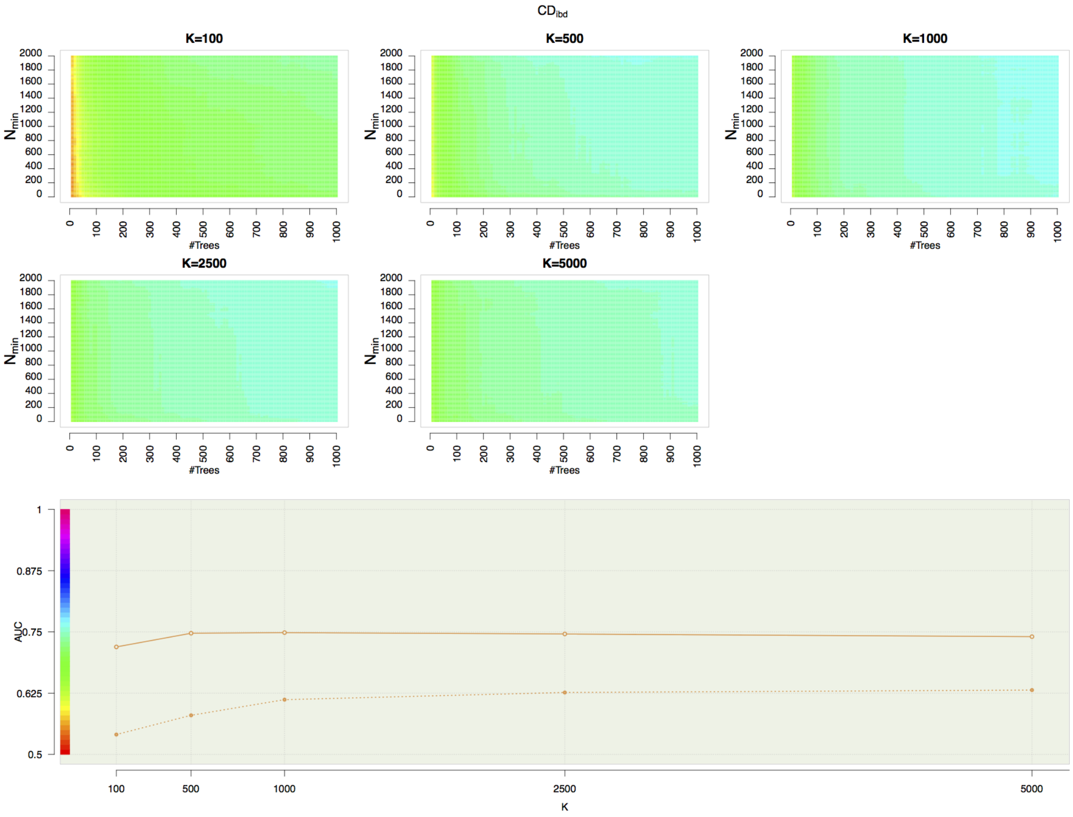


FIGURE 6.8 On CD_{ibd}: T-Trees prediction performance with 10-SNPS blocks and $IC = 10$. The five upper panels show the influence of T and $N_{\min}$ for each of the five investigated values of K. The last panel displays the evolution of the AUC as the value of K increases, the plain line corresponds to the maximum and the dotted line to the minimum obtained at each value of K.

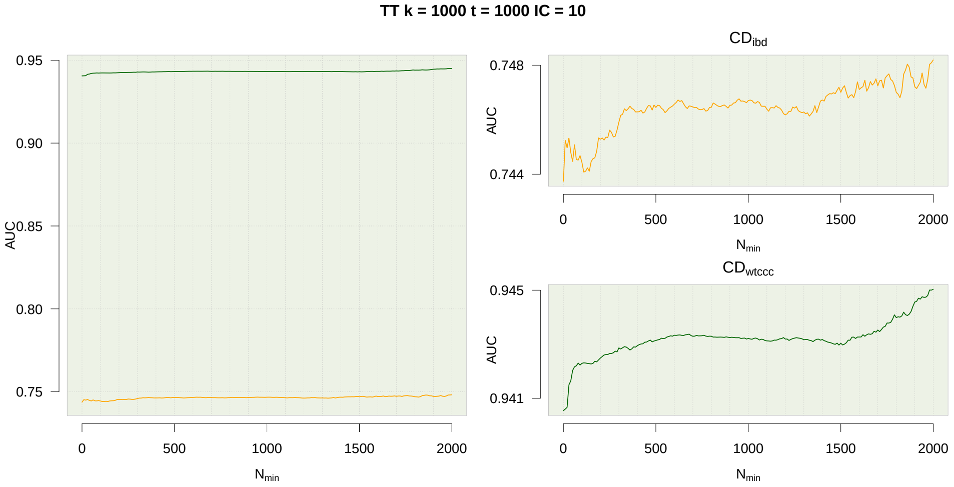


FIGURE 6.9 Dataset comparison: T-Trees $K = 1000$ $T = 1000$ and $IC = 10$. Evolution of the AUC for the optimal value $K = 1000$. CD_{wtccc} and CD_{ibd} results are, respectively, represented in green and orange. The left panel shows a comparison between the two variants while, the two right panels show a detailed view of the corresponding curves.

T-Trees specific metaparameters influence

In this section, we will focus on the influence the bloc map that is provided to the T-Trees algorithm and the influence of the internal complexity parameter.

Figure 6.10 and Table 6.2 show the influence of the IC parameter for a value of $K = 1000$ with 1000 fully developed trees on the two variations of the CD dataset. Green and orange respectively represent results for CD_{wtccc} and CD_{ibd}. Even with an internal complexity of 1, we obtained better results than the random forests with $K = 1000$ (in dashed lines). While for greater values of the IC, we get better AUCs than with Random Forests with $K = 10000$ (in dotted lines).

We just observed that an internal complexity of 1 leads to similar results than the maximal ones we reached with standard Random Forests. That setting allows to span the search of an optimal attribute at each node. Instead of looking at all the variables, the search for an optimal split is done in every 10 SNPS blocks.

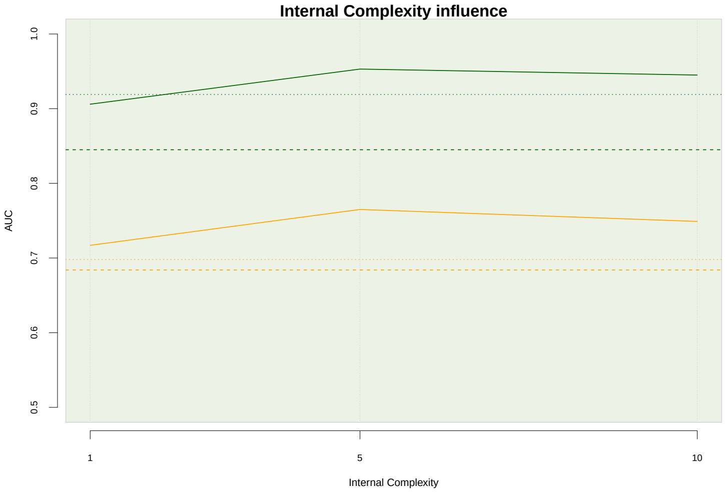


FIGURE 6.10 IC comparison for $K = 1000$ on the two datasets. In green CD_{wtccc} and in orange, CD_{ibd}. The dashed lines represents the maximal AUC obtained with $K = 1000$ and the dotted lines the maximal AUC obtained with $K = 10000$ for the Random Forests.

IC	CD_{wtccc}	CD_{ibd}
1	0.906	0.717
5	**0.953**	**0.765**
10	0.945	0.749

TABLE 6.2 AUCs obtained for different internal complexities on the three CD datasets. Each variant maximum are highlighted in bold.

Table 6.3 summarises the results we obtained while investigating different bloc maps and different values of the internal complexity parameter. In addition of using 10 consecutive SNPS blocks, we also tested 20 and 50 SNPS blocks of contiguous markers. Results remained consistent across the two datasets. When the internal complexity parameter produced better results for one of the dataset variants it also allows for a gain in predictive power on the other dataset variant. These results also suggest that the T-Trees are quite robust against the block composition/choice. For each dataset, regardless of the block map and the IC value, we see that AUCs do not fluctuate that much. The table even suggests that, no matter which block map is used, the only parameter that influences the predictive power is the internal complexity parameter (e.g. results

Bloc size	IC	CD_{wtccc}	CD_{ibd}
10	1	0.906	0.717
	5	**0.953**	**0.765**
	10	0.945	0.749
20	5	**0.955**	**0.755**
	10	0.945	0.740
	20	0.931	0.706
50	10	**0.937**	**0.744**
	25	0.913	0.700

Table 6.3 Block size and internal complexity influence. Resulting AUCs obtained with the T-Trees for different block sizes and internal complexities. Maxima for each block size are highlighted in bold.

with $IC = 10$ for bloc size 10 and 20 are almost identical). Nevertheless, we do observe a slight decrease as the size of the block and the IC increases. We notice that these configurations with both larger block size and large IC force the T-Trees to exploit a larger number of variables for each internal split, and we believe that by doing so they tend to overfit the training data.

T-Trees versus standard tree based methods

Table 6.4 summarises all the previous results we obtained with the three methods on the two datasets. Cross dataset results are consistent among the three methods: CD_{wtccc} produces higher AUCs than CD_{ibd}.

	CD_{wtccc}			CD_{ibd}		
K	RF	ET	TT	RF	ET	TT
100	0.683	0.690	**0.921**	0.628	0.667	**0.719**
500	0.799	0.798	**0.942**	0.675	0.709	**0.747**
1000	0.845	0.840	**0.945**	0.684	0.721	**0.749**
2500	0.888	0.888	**0.944**	0.700	0.724	**0.746**
5000	0.909	0.913	**0.938**	0.698	0.721	**0.740**
10000	0.919	0.922		0.697	0.705	

Table 6.4 Summary table: the maximum AUCs obtained with the RF, the ET and the TT for the different values of K.

Using the T-Trees methodology encouragingly improves all the results, regardless of the dataset variant. Even with the lowest value of K, we observed AUCs that are sometimes slightly better, or at least close to the best results we obtained with the two standard methods. Beside the fact that the AUCs increased, a notable improvement is the downward shift of the optimal value of the K parameters. While with the standard approaches, AUC maxima where observed for $K = 10000$, a $K = 1000$ seems sufficient to reach a maximum value with the trees inside trees approach. In addition to the fact that strongly pruned ($N_{min} = 2000$) T-Trees were observed as optimal, they thus produces trees that are compact in comparison to the two other variants.

These promising AUC improvements suggest that using more than one SNP at each node improves the power of our tree based "prediction machine". Table 6.5, empirically ensures that our approach is effectively taking advantage of the structured nature of the variables i.e. the LD pattern. Indeed, we tested our approach with blocks of randomly positioned SNPs rather than contiguous SNP blocks. We see that with random blocks (i.e. when we break the surrounding LD structure), the AUCs substantially drop, meaning that the gain we observed is not the sole consequence of the dimension reduction introduced by our methodology. This result supports the effectiveness of our approach and confirms the initial intuition that led us to propose such a modification in the standard tree based approaches in the GWAS context.

	CD_{wtccc}		CD_{ibd}	
K	contig.	rand.	contig.	rand.
100	**0.903**	0.753	**0.690**	0.600
500	**0.936**	0.835	**0.728**	0.625
1000	**0.941**	0.853	**0.744**	0.627

TABLE 6.5 AUCs obtained with unpruned T-Trees with $IC = 10$, **contig**uous blocks of 10 SNPs versus **rand**om blocks of 10 SNPs. Using randomized bloc map drastically deteriorates the results.

6.1.2 Identification of suceptibility loci

In this section we investigate the loci identification ability of the tree-based methods in the field of GWAS. Based on comparisons between p-values, variable and group importances we will show that most of the confirmed associated loci identified by standard χ^2/Fisher tests of associations are also more or less selected by the ensemble of tree methods. We will also show that different qc filters may have an important impact on the tree induction process and that the multivariate aspect of our methods renders it more sensitive to these filters. Indeed, as the nodes are expanded conditionally to their parents, weaker effects may be hidden by stronger ones that might not be genuine association but genotyping artefacts.

Besides the important influence on the apparent predictive power of the different tree based methods that such artefacts can introduce, we will see that variable importances still robustly and consistently identify many regions as being important. Especially, we will see that many hits (defined by SNPs appearing in the top ranking according to the tree based variable importances) are consistently concentrated in some regions of the genome across the two variants of the CD dataset.

A first glance at the nine reported regions

Let's start with the 9 regions reported as being strongly associated with the disease on the Crohn dataset in [Wel07]. Using these 9 loci (listed in Table 6.6), we can compare how well the methods are able to identify confirmed susceptibility regions on the 2 dataset variants. With the Random Forests we choose to fix the settings to $K = 10000$, $T = 1000$ and $N_{\min} = 250$ as it corresponds, for the 2 datasets, to the near optimal set of parameters (see Figure 6.3). While for the T-Trees, we choose $K = 1000$, $T = 1000$, $IC = 5$ and $N_{\min} = 2000$.

Chromosome	Start (Mb)	End (Mb)
1p31	67.30	67.48
2q37	233.92	234.00
3p21	49.30	49.87
5p13	40.32	40.66
5q33	150.15	150.31
10q21	64.06	64.31
10q24	101.26	101.32
16q12	49.02	49.40
18p11	12.76	12.91

TABLE 6.6 The nine WTCCC confirmed regions. Positions are expressed in NCBI build-35 coordinates.

Figures 6.11 and 6.12 allow for a rapid comparison between the random forest and the T-Trees applied to the CD_{wtccc} and CD_{ibd} dataset. Each histogram corresponds to the top 100 tree based variables importances. Left column corresponds to random forest variable importances while the right corresponds to T-Trees variable importances.

Firstly, on Figure 6.11, in the first row, variables found in one of the nine confirmed regions are highlighted in red, variables reported as moderately associated in [Wel07] Supplementary Information are highlighted in orange, and similarly, blue highlights variables appearing in two more regions that have been detected by both tree based methods (details about these two regions will follow later in this section). With the RF we see that many "confirmed" variables appear but most of them are ranked below the 50 first. On the other hand, with the T-Trees, we detected more variables coming from two additional regions (in blue). Next, in the second row, purple corresponds to rare variants (SNPS with a MAF < 0.05), some of the rare variants appear in our two top rankings (there are a few more with the RF (39) than with the T-Trees (25)), among these some are also deviating from HWE and/or are associated according to their p-values. In the third row, we highlighted in orange the variables with a HWE p-value $< 10^{-6}$ in the controls only or $< 10^{-10}$ in the cases only. These two thresholds corresponds to the one found in the CD_{ibd} QC filters. We see that almost all of the ten first variables according to the random forests variables importances are deviating from HWE (at least for one of the phenotype). Finally, in the last row, variables highlighted in green are those with a Fisher p-value below 10^{-6}. Many of the markers deviating from HWE are also associated according to the Fisher test of association.

Similarly, Figure 6.12 demonstrates that the removal of suspicious SNPS from the set of candidate variables allowed for a better detection of the nine reported regions (first row, first column). On the other hand, we observed that the T-Trees strongly detects two more regions (in blue, first row, second column). Interestingly, for the random forest, markers from these two regions are also selected but considered as less informative while in the T-Trees, these two regions are mainly the only two that are represented in the 20 first according to the variable importances (see Table 6.9). There seems to be a more important concentration of rare variants in the top 20 with the RF and way less with the T-Trees. We noticed that there were also less of these variables on CD_{wtccc} with the T-Trees. As the T-Trees exploit blocks of SNPS instead of single ones, it is less sensitive to their respective allele frequencies. As more of the confirmed regions are captured by the tree based method in the strongly filtered version CD_{ibd} of the dataset, we will focus on this latter dataset in the rest of this section.

Figure 6.13 shows the results we obtained on the CD_{ibd} data where the grey bars correspond to the exact Fisher test based p-values, the blue bars and dots represent the 10 fold aggregated random forests variable importances (which corresponds to variable importances derived from a forest of 10000 trees), the orange bars and the green boxes respectively corresponds to variable importances and group importances derived from the T-Trees ($T = 1000$, $K = 1000$, $IC = 5$ and $N_{min} = 2000$). Each green box contains 10 markers. This Figure shows how the p-values and the variable importances are correlated. In these nine regions, most of the variables detected by a low p-value are also, more or less, identified by the variable importances. In each region, the most associated SNPS are also the most important according to the RF variables importances and the group importance seems to be the highest where the markers with the locally lowest p-values are. However, the vertical axis scales on these graphs are a bit misleading (each vertical axis maximum is equal to the maximum in the corresponding window) (we refer the reader to Figure A.1 in the Appendix for a normalised vertical axis version).

Table 6.8 details the 20 first SNPS according to the random forest variable importances on CD_{ibd}. In that top, the first variable appears to be `rs11209026` (found in region 1p31) which is also the most important one according to the Fisher p-value (see Table A.3). We also notice the presence of several SNPS located in the nine regions. Such as `rs2076756` and `rs2066843` on chromosome 16q12. `rs10210302`, `rs6431654`, `rs6752107`, `rs3828309` and `rs3792106` on chromosome 2q37. `rs7515029` also located in region 1p31. Mainly three of the nine regions are represented in these 20 first variables.

On the other hand, similarly Table 6.9 lists the 20 first SNP according to the T-Trees variable importances. We no longer see that many SNPS as being reported except two, found in region 1p31 (`rs11209026` and

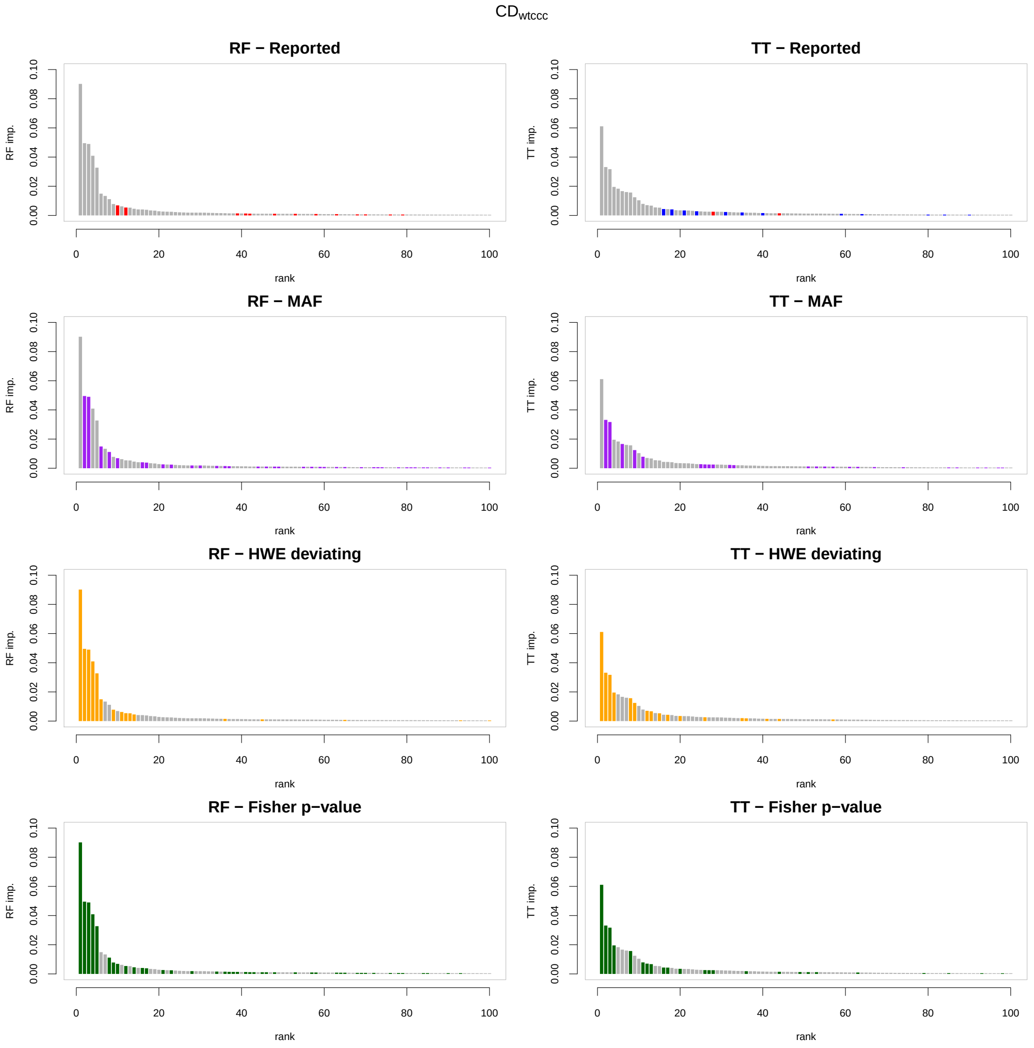


FIGURE 6.11 The first 100 variables according to the tree based importance rankings for CD_{wtccc}. The horizontal axis corresponds to the ranks and the vertical axis to the variable importances. In the first column variables are ordered according to random forests variable importances and in the second column they are ordered according to the T-Trees variable importances. In the first row, red highlights the nine reported regions and blue highlights two more regions mostly detected by tree based methods. In the second row, purple corresponds to rare variants (MAF < 0.05). In the third row, orange highlights SNPs deviating from HWE and in the last row, green represents markers with a low Fisher p-value ($< 10^{-6}$).

rs2201841). However, we notice several hits on chromosome 2p12 (near position 81.5Mb) and on chromosome 7q31 (near position 125.1Mb). In that ranking, the first marker **rs11887827** is also found at position 20 in the random forest ranking. In these two regions, only two markers have a low p-value: rs11887827 (2p12) and rs2107062 (on 7q31) but these do not correspond to the markers with the overall lowest p-values. It seems that region 2p12 is useful for both tree based methods while region 7q31 is only considered as important by the T-Trees method. Those two regions are represented in Figure 6.14. In these two regions, two groups of markers are considered as the most important according to the T-Trees group importances.

Figure 6.15 focuses on the 2p12 region. In that figure, T-Trees variable and group importances are reported in the first row. In the second row, the univariate (Fisher) p-values (axis on the right side) and the haplotype p-values (axis on the left side; derived from the case/control omnibus test with $H - 1$ degree of

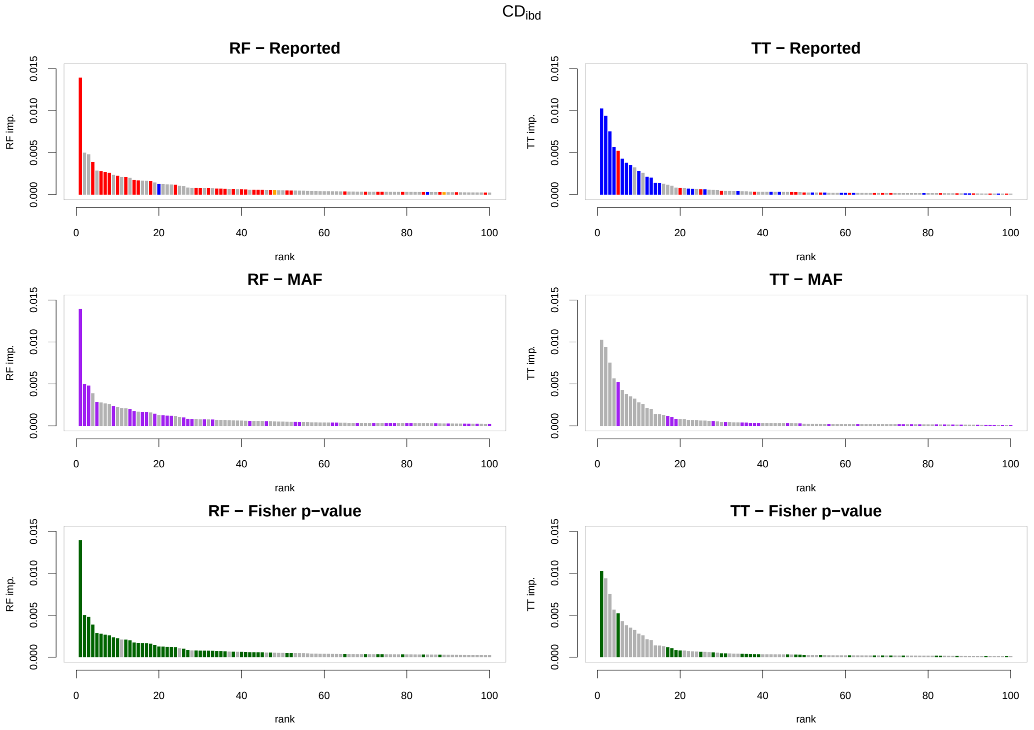


FIGURE 6.12 The first 100 variables according to the tree based importance rankings for CD_{ibd}. The horizontal axis corresponds to the ranks and the vertical axis to the variable importances. In the first column variables are ordered according to random forests variable importances and in the second column they are ordered according to the T-Trees variable importances. In the first row, red highlights the nine reported regions and blue highlights two more regions mostly detected by tree based methods. In the second row, purple corresponds to rare variants (MAF < 0.05). In the third row, green represents markers with a low Fisher p-value ($< 10^{-6}$).

freedom where H corresponds to the number of common[1] haplotypes). The third row represents the number of common haplotypes found in each 10 SNP blocks and the bottom part of the figure shows the LD pattern (r^2) in that region. We see that the block with the highest importance is also strongly associated with the disease and while the univariate p-value (the lowest p-value in that block is $1 \cdot 10^{-7.6}$) fails to strongly identify this association the haplotype p-value is extremely low ($1 \cdot 10^{-67}$). The LD pattern suggests that there are two haplotype blocks, and the 10 SNP block we identified with the T-Trees falls in a strongly correlated subregion in the second haplotype block. Similarly (Figure 6.16), for the 7q31 region (details not reported), the same analysis shows that the corresponding block has a haplotype p-value of $1 \cdot 10^{-43}$ while the flanking blocks are not associated at all. Finally, Table 6.7 reports the 6 common haplotypes found in the 2p12 and 7q31 regions.

2p12	7q31
GGCATGTGGG	GGTGTTAGTC
AGCACGTGGG	ATGCCTGACT
AGCACGTAGG	ATGCCTGACC
GATGTAAGGC	GTGCCTGACC
AATGTAAGGC	GGGGTTAGTC
GATGTAAGTC	AGGCCCGACT

TABLE 6.7 The 6 common haplotype found in the 2p12 and the 7q31 regions.

[1] Here, a haplotype is said to be common if its frequency > 0.01 in the population under study

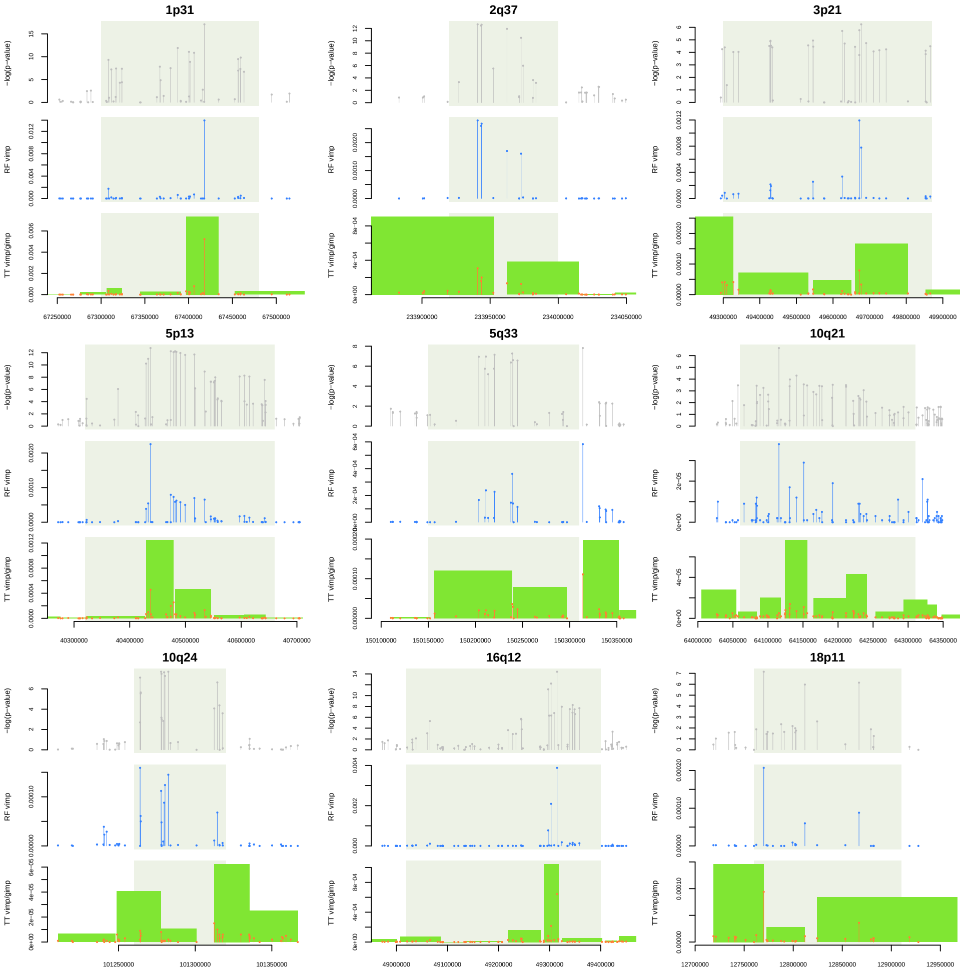


FIGURE 6.13 The nine reported regions for Crohn's disease. In grey, the exact Fisher test based p-values, in blue random forest variable importances ($K = 10000$, $T = 1000$ and $N_{\mathrm{min}} = 250$), in orange the T-Trees variable importances while the green boxes denote the T-Trees group importances ($T = 1000$, $K = 1000$, $IC = 5$ and $N_{\mathrm{min}} = 2000$). The light grey shaded boxes delimit the nine regions as reported.

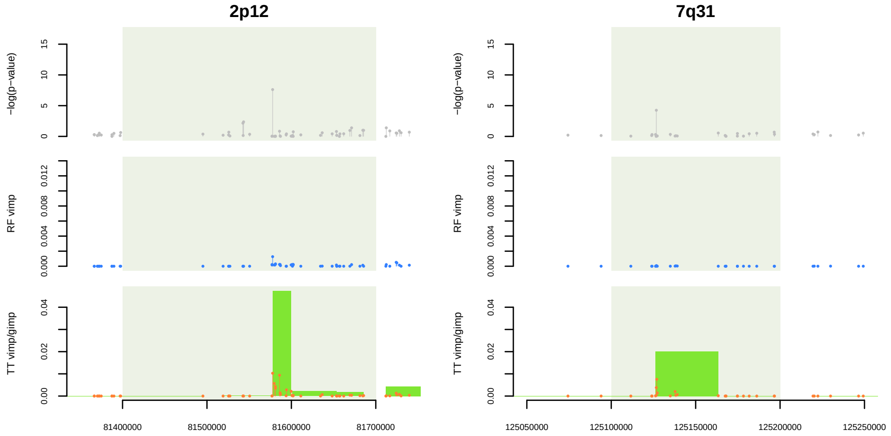


FIGURE 6.14 Two more important regions for Crohn's disease according to the tree based methods. In grey, the exact Fisher test based p-values, in blue random forest variable importances ($K = 10000$, $T = 1000$ and $N_{\mathrm{min}} = 250$), in orange the T-Trees variable importances while the green boxes denote the T-Trees group importances ($T = 1000$, $K = 1000$, $IC = 5$ and $N_{\mathrm{min}} = 2000$).

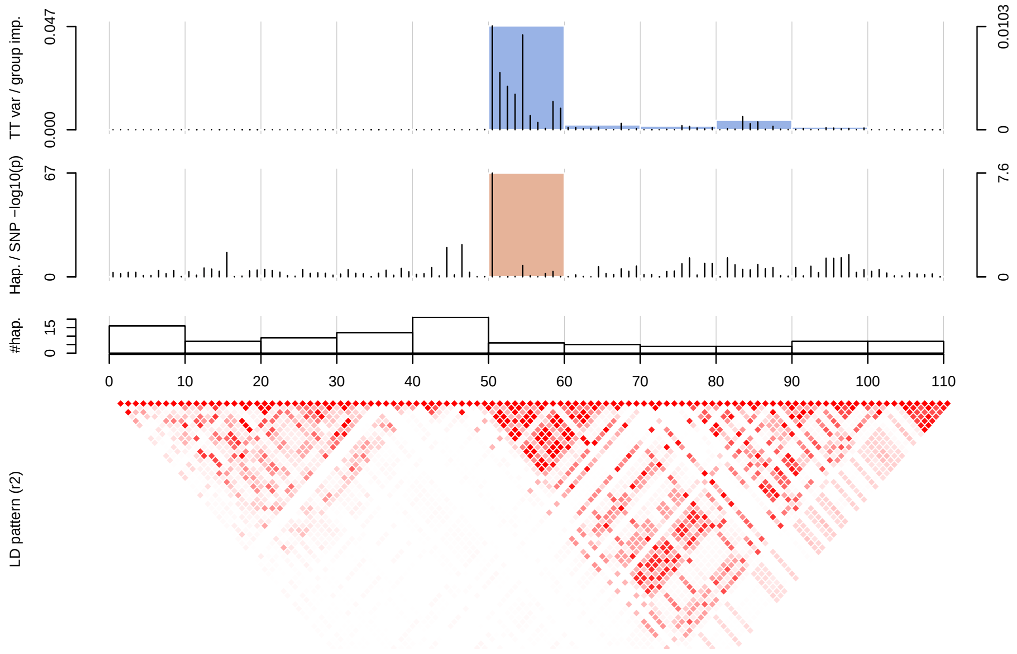


FIGURE 6.15 The region 2p12 on CD_{ibd} analysed according to the T-Tree block map. In the first row, T-Trees variable and group importances. In the second row, SNP and haplotype p-values. In the third row, the number of haplotypes. The bottom of the figure represents the corresponding LD pattern (r^2) in that region.

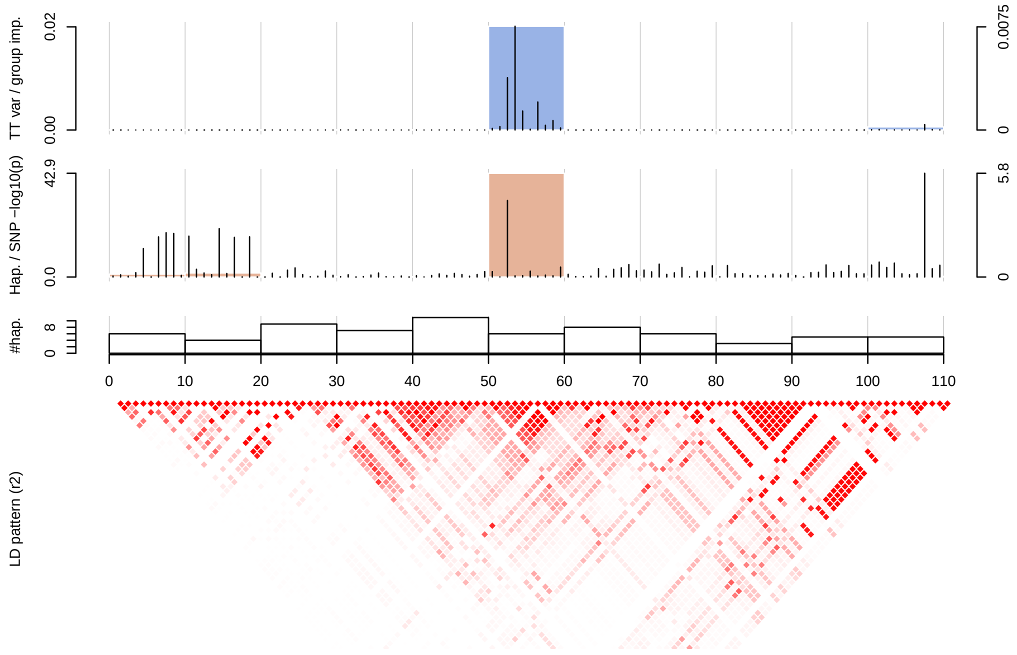


FIGURE 6.16 The region 7q31 on CD_{ibd} analysed according to the T-Tree block map. In the first row, T-Trees variable and group importances. In the second row, SNP and haplotype p-values. In the third row, the number of haplotypes. The bottom of the figure represents the corresponding LD pattern (r^2) in that region.

#	Chr	Pos.	SNP	RF imp.	Fisher p-value	χ^2 p-value	MAF	f_{miss}	pmiss	0	1	2	HWE
1	1	67417979	**rs11209026**	$1.40 \cdot 10^{-2}$	$8.24 \cdot 10^{-18}$	$2.09 \cdot 10^{-16}$	0.0451	0.006 63	$4.45 \cdot 10^{-3}$	0	419	4226	$1.03 \cdot 10^{-4}$
								0.009 19		0	342	2568	$3.52 \cdot 10^{-5}$
								0.002 30		0	77	1658	1
2	14	84425325	rs10144260	$5.02 \cdot 10^{-3}$	$1.18 \cdot 10^{-9}$	$4.40 \cdot 10^{-8}$	0.007 64	0.005 99	$9.48 \cdot 10^{-10}$	0	71	4577	1
								0.000 681		0	67	2868	1
								0.0150		0	4	1709	1
3	4	86127973	rs1872321	$4.81 \cdot 10^{-3}$	$6.88 \cdot 10^{-9}$	$1.48 \cdot 10^{-8}$	0.002 04	0.005 56	$3.15 \cdot 10^{-1}$	0	19	4631	1
								0.006 47		0	0	2918	1
								0.004 03		0	19	1713	1
4	16	49314382	**rs2076756**	$3.88 \cdot 10^{-3}$	$3.95 \cdot 10^{-15}$	$2.25 \cdot 10^{-15}$	0.270	0.007 91	$1.64 \cdot 10^{-2}$	373	1758	2508	$9.11 \cdot 10^{-3}$
								0.005 45		174	1065	1682	$7.62 \cdot 10^{-1}$
								0.0121		199	693	826	$4.50 \cdot 10^{-3}$
5	8	30332834	rs7842024	$2.88 \cdot 10^{-3}$	$4.09 \cdot 10^{-8}$	$6.42 \cdot 10^{-8}$	0.001 83	0.006 63	$5.81 \cdot 10^{-11}$	0	17	4628	1
								0.000 681		0	0	2935	1
								0.0167		0	17	1693	1
6	2	233940839	**rs10210302**	$2.79 \cdot 10^{-3}$	$2.22 \cdot 10^{-13}$	$2.31 \cdot 10^{-13}$	0.452	0.000 428	$5.33 \cdot 10^{-1}$	936	2349	1389	$3.30 \cdot 10^{-1}$
								0.000 681		646	1529	760	$1.98 \cdot 10^{-2}$
								0		290	820	629	$4.25 \cdot 10^{-1}$
7	2	233943769	**rs6431654**	$2.68 \cdot 10^{-3}$	$2.55 \cdot 10^{-13}$	$2.75 \cdot 10^{-13}$	0.451	0.001 07	$6.57 \cdot 10^{-1}$	933	2345	1393	$3.60 \cdot 10^{-1}$
								0.001 36		645	1524	764	$2.91 \cdot 10^{-2}$
								0.000 575		288	821	629	$4.85 \cdot 10^{-1}$
8	2	233943448	**rs6752107**	$2.60 \cdot 10^{-3}$	$3.61 \cdot 10^{-13}$	$3.89 \cdot 10^{-13}$	0.452	0.000 428	$5.33 \cdot 10^{-1}$	937	2348	1389	$3.45 \cdot 10^{-1}$
								0.000 681		646	1528	761	$2.19 \cdot 10^{-2}$
								0		291	820	628	$3.97 \cdot 10^{-1}$
9	4	178272461	rs1595154	$2.36 \cdot 10^{-3}$	$1.08 \cdot 10^{-7}$	$8.80 \cdot 10^{-8}$	0.002 48	0.006 63	$1.25 \cdot 10^{-4}$	0	23	4622	1
								0.009 87		0	2	2906	1
								0.001 15		0	21	1716	1
10	5	40437266	**rs17234657**	$2.26 \cdot 10^{-3}$	$1.72 \cdot 10^{-13}$	$8.09 \cdot 10^{-14}$	0.146	0.001 71	1	113	1132	3423	$9.90 \cdot 10^{-2}$
								0.001 70		51	628	2253	$3.51 \cdot 10^{-1}$
								0.001 73		62	504	1170	$4.18 \cdot 10^{-1}$
11	5	33150395	rs6894272	$2.10 \cdot 10^{-3}$	$4.49 \cdot 10^{-3}$	$3.95 \cdot 10^{-3}$	0.0776	0.001 28	$9.07 \cdot 10^{-2}$	22	681	3967	$2.60 \cdot 10^{-1}$
								0.002 04		2	415	2514	$3.47 \cdot 10^{-5}$
								0		20	266	1453	$7.05 \cdot 10^{-2}$
12	16	49302700	**rs2066843**	$2.09 \cdot 10^{-3}$	$5.96 \cdot 10^{-13}$	$4.26 \cdot 10^{-13}$	0.285	0.003 42	$1.26 \cdot 10^{-1}$	422	1813	2425	$1.99 \cdot 10^{-3}$
								0.002 38		206	1106	1618	$3.61 \cdot 10^{-1}$
								0.005 18		216	707	807	$1.92 \cdot 10^{-3}$
13	5	167826491	rs888775	$2.01 \cdot 10^{-3}$	$2.98 \cdot 10^{-7}$	$3.73 \cdot 10^{-7}$	0.001 62	0.007 91	$6.23 \cdot 10^{-11}$	0	15	4624	1
								0.001 36		0	0	2933	1
								0.0190		0	15	1691	1
14	1	67308393	**rs7515029**	$1.74 \cdot 10^{-3}$	$5.01 \cdot 10^{-10}$	$1.47 \cdot 10^{-9}$	0.0359	0.002 35	$5.51 \cdot 10^{-1}$	10	315	4340	$9.10 \cdot 10^{-2}$
								0.002 04		9	245	2677	$1.90 \cdot 10^{-1}$
								0.002 88		1	70	1663	$5.29 \cdot 10^{-1}$
15	2	233962410	**rs3828309**	$1.70 \cdot 10^{-3}$	$1.19 \cdot 10^{-12}$	$1.12 \cdot 10^{-12}$	0.452	0.002 99	$5.90 \cdot 10^{-1}$	935	2344	1383	$3.15 \cdot 10^{-1}$
								0.003 41		645	1521	761	$2.89 \cdot 10^{-2}$
								0.002 30		290	823	622	$5.18 \cdot 10^{-1}$
16	8	116374529	rs16887291	$1.68 \cdot 10^{-3}$	$1.15 \cdot 10^{-6}$	$3.75 \cdot 10^{-7}$	0.006 78	0.005 77	$7.72 \cdot 10^{-5}$	1	61	4587	$1.91 \cdot 10^{-1}$
								0.008 85		1	18	2892	$3.22 \cdot 10^{-2}$
								0.000 575		0	43	1695	1
17	3	6848446	rs17046143	$1.67 \cdot 10^{-3}$	$5.75 \cdot 10^{-7}$	$4.96 \cdot 10^{-7}$	0.001 93	0.004 71	$6.65 \cdot 10^{-1}$	0	18	4636	1
								0.005 11		0	1	2921	1
								0.004 03		0	17	1715	1
18	2	233972740	**rs3792106**	$1.60 \cdot 10^{-3}$	$3.32 \cdot 10^{-11}$	$3.66 \cdot 10^{-11}$	0.397	0.001 28	$4.22 \cdot 10^{-1}$	725	2256	1689	$5.41 \cdot 10^{-1}$
								0.001 70		494	1490	948	$2.81 \cdot 10^{-2}$
								0.000 575		231	766	741	$1.42 \cdot 10^{-1}$
19	7	125371971	rs4431537	$1.46 \cdot 10^{-3}$	$1.74 \cdot 10^{-6}$	$9.66 \cdot 10^{-7}$	0.003 74	0.000 214	1	0	35	4640	1
								0.000 341		0	8	2928	1
								0		0	27	1712	1
20	2	81577812	rs11887827	$1.27 \cdot 10^{-3}$	$2.42 \cdot 10^{-8}$	$2.73 \cdot 10^{-8}$	0.311	0.002 14	$5.15 \cdot 10^{-1}$	499	1903	2264	$1.04 \cdot 10^{-3}$
								0.001 70		322	1299	1311	1
								0.002 88		177	604	953	$1.54 \cdot 10^{-7}$

TABLE 6.8 The 20 first markers according to the random forest variable importances (denoted RF imp. in the gray shaded column) on the CD_{ibd} datasets and the corresponding statistics. Green shaded cells refer to statistics related to controls only while red shaded cells refer to cases only statistics. Bold typeface SNP identifiers denote markers found in one of the nine reported regions.

#	Chr	Pos.	SNP	TT imp.	Fisher p-value	χ^2 p-value	MAF	f_{miss}	pmiss	0	1	2	HWE
1	2	81577812	rs11887827	$1.03 \cdot 10^{-2}$	$2.42 \cdot 10^{-8}$	$2.73 \cdot 10^{-8}$	0.311	0.00214	$5.15 \cdot 10^{-1}$	499	1903	2264	$1.04 \cdot 10^{-3}$
								0.00170		322	1299	1311	1
								0.00288		177	604	953	$1.54 \cdot 10^{-7}$
2	2	81585945	rs7593114	$9.39 \cdot 10^{-3}$	$1.44 \cdot 10^{-1}$	$1.41 \cdot 10^{-1}$	0.326	0.00214	$4.60 \cdot 10^{-2}$	498	2046	2122	$8.94 \cdot 10^{-1}$
								0.00102		323	1299	1312	$9.67 \cdot 10^{-1}$
								0.00403		175	747	810	$9.12 \cdot 10^{-1}$
3	7	125126957	rs6947579	$7.55 \cdot 10^{-3}$	$8.54 \cdot 10^{-1}$	$8.45 \cdot 10^{-1}$	0.317	0.000855	$3.04 \cdot 10^{-1}$	485	1994	2193	$3.11 \cdot 10^{-1}$
								0.00136		301	1263	1369	$7.02 \cdot 10^{-1}$
								0		184	731	824	$2.45 \cdot 10^{-1}$
4	2	81579712	rs17020244	$5.66 \cdot 10^{-3}$	$9.76 \cdot 10^{-1}$	$9.69 \cdot 10^{-1}$	0.149	0.000642	1	111	1169	3393	$3.86 \cdot 10^{-1}$
								0.000681		65	743	2127	1
								0.000575		46	426	1266	$1.57 \cdot 10^{-1}$
5	1	67417979	**rs11209026**	$5.23 \cdot 10^{-3}$	$8.24 \cdot 10^{-18}$	$2.09 \cdot 10^{-16}$	0.0451	0.00663	$4.45 \cdot 10^{-3}$	0	419	4226	$1.03 \cdot 10^{-4}$
								0.00919		0	342	2568	$3.52 \cdot 10^{-5}$
								0.00230		0	77	1658	1
6	2	81581046	rs12623313	$4.30 \cdot 10^{-3}$	1	$9.87 \cdot 10^{-1}$	0.149	0.00107	$7.09 \cdot 10^{-3}$	110	1168	3393	$4.18 \cdot 10^{-1}$
								0		65	743	2129	1
								0.00288		45	425	1264	$2.16 \cdot 10^{-1}$
7	7	125126698	rs2107062	$3.81 \cdot 10^{-3}$	$5.60 \cdot 10^{-5}$	$5.43 \cdot 10^{-5}$	0.299	0.00770	$1.86 \cdot 10^{-6}$	483	1811	2346	$2.68 \cdot 10^{-6}$
								0.0119		300	1223	1379	$2.45 \cdot 10^{-1}$
								0.000575		183	588	967	$9.34 \cdot 10^{-10}$
8	2	81581185	rs10520335	$3.52 \cdot 10^{-3}$	$9.04 \cdot 10^{-1}$	$9.06 \cdot 10^{-1}$	0.149	0.00278	1	109	1167	3387	$4.87 \cdot 10^{-1}$
								0.00272		63	742	2124	$9.42 \cdot 10^{-1}$
								0.00288		46	425	1263	$1.56 \cdot 10^{-1}$
9	23	21732484	rs5904497	$3.26 \cdot 10^{-3}$	$4.41 \cdot 10^{-2}$	$4.29 \cdot 10^{-2}$	0.273	0.00749	$3.55 \cdot 10^{-2}$	213	896	1410	$5.81 \cdot 10^{-5}$
								0.00953		122	562	780	$1.50 \cdot 10^{-1}$
								0.00403		91	334	630	$5.97 \cdot 10^{-6}$
10	2	81594169	rs17020301	$2.80 \cdot 10^{-3}$	$3.89 \cdot 10^{-1}$	$3.79 \cdot 10^{-1}$	0.0812	0.00128	$6.77 \cdot 10^{-1}$	30	698	3942	1
								0.00102		16	433	2485	$6.14 \cdot 10^{-1}$
								0.00173		14	265	1457	$6.40 \cdot 10^{-1}$
11	23	21732111	rs4824171	$2.60 \cdot 10^{-3}$	$3.48 \cdot 10^{-1}$	$3.42 \cdot 10^{-1}$	0.284	0.00428	$3.55 \cdot 10^{-1}$	209	995	1335	$2.17 \cdot 10^{-1}$
								0.00511		120	569	791	$2.14 \cdot 10^{-1}$
								0.00288		89	426	544	$6.52 \cdot 10^{-1}$
12	2	81599721	rs12613517	$2.14 \cdot 10^{-3}$	$9.04 \cdot 10^{-1}$	$9.05 \cdot 10^{-1}$	0.149	0.00128	1	110	1170	3390	$4.53 \cdot 10^{-1}$
								0.00136		64	743	2126	1
								0.00115		46	427	1264	$1.86 \cdot 10^{-1}$
13	7	125137814	rs1419584	$2.04 \cdot 10^{-3}$	$8.72 \cdot 10^{-1}$	$8.53 \cdot 10^{-1}$	0.317	0.00214	$7.53 \cdot 10^{-1}$	487	1988	2191	$2.51 \cdot 10^{-1}$
								0.00238		302	1260	1368	$6.40 \cdot 10^{-1}$
								0.00173		185	728	823	$2.02 \cdot 10^{-1}$
14	2	81586635	rs9646997	$1.40 \cdot 10^{-3}$	$8.33 \cdot 10^{-1}$	$8.34 \cdot 10^{-1}$	0.149	0.00150	$4.36 \cdot 10^{-1}$	111	1168	3390	$3.86 \cdot 10^{-1}$
								0.00102		63	744	2127	$8.84 \cdot 10^{-1}$
								0.00230		48	424	1263	$8.99 \cdot 10^{-2}$
15	7	125127253	rs6967968	$1.38 \cdot 10^{-3}$	$8.36 \cdot 10^{-1}$	$8.19 \cdot 10^{-1}$	0.317	0.000855	1	486	1993	2193	$2.95 \cdot 10^{-1}$
								0.00102		302	1263	1369	$6.71 \cdot 10^{-1}$
								0.000575		184	730	824	$2.44 \cdot 10^{-1}$
16	2	81724164	rs12464902	$1.31 \cdot 10^{-3}$	$2.76 \cdot 10^{-1}$	$2.69 \cdot 10^{-1}$	0.0644	0.00128	$4.22 \cdot 10^{-1}$	13	575	4082	$1.44 \cdot 10^{-1}$
								0.00170		10	370	2552	$4.57 \cdot 10^{-1}$
								0.000575		3	205	1530	$2.04 \cdot 10^{-1}$
17	4	86127973	rs1872321	$1.19 \cdot 10^{-3}$	$6.88 \cdot 10^{-9}$	$1.48 \cdot 10^{-8}$	0.00204	0.00556	$3.15 \cdot 10^{-1}$	0	19	4631	1
								0.00647		0	0	2918	1
								0.00403		0	19	1713	1
18	14	84425325	rs10144260	$1.07 \cdot 10^{-3}$	$1.18 \cdot 10^{-9}$	$4.40 \cdot 10^{-8}$	0.00764	0.00599	$9.48 \cdot 10^{-10}$	0	71	4577	1
								0.000681		0	67	2868	1
								0.0150		0	4	1709	1
19	8	30332834	rs7842024	$8.47 \cdot 10^{-4}$	$4.09 \cdot 10^{-8}$	$6.42 \cdot 10^{-8}$	0.00183	0.00663	$5.81 \cdot 10^{-11}$	0	17	4628	1
								0.000681		0	0	2935	1
								0.0167		0	17	1693	1
20	1	67406223	**rs2201841**	$7.94 \cdot 10^{-4}$	$1.41 \cdot 10^{-11}$	$1.15 \cdot 10^{-11}$	0.345	0	1	564	2094	2018	$5.60 \cdot 10^{-1}$
								0		311	1251	1375	$2.89 \cdot 10^{-1}$
								0		253	843	643	$4.19 \cdot 10^{-1}$

TABLE 6.9 The 20 first markers according to the T-Trees variable importances (denoted TT imp. in the gray shaded column) on the CD_{ibd} datasets and the corresponding statistics. Green shaded cells refer to statistics related to controls only while red shaded cells refer to cases only statistics. Bold typeface SNP identifiers denote markers found in one of the nine reported regions.

Focusing on the suspected regions

To further estimate how these nine reported regions and the two additional ones detected by our approach influence the tree induction process, we trained Random Forests and T-Trees while using only the variables from those regions as candidate SNPs or blocks. SNP are found in the different investigated regions. The 9 confirmed loci are defined as in [Wel07] (Table 6.6), while the region denoted 2p12 ranges from 81.4Mb to 81.7Mb and the one denoted 7q31 ranges from 125.1Mb to 125.2Mb. Tables 6.11 and 6.10 summarise the different AUCs we obtained under these conditions for the two datasets. In these tables, the second columns give the number of variables found in each considered subset of variables. As these numbers change, the appropriate values of K change accordingly. Values of K are represented in the light green cells while the dark green cells report the IC values in the T-Trees experiments.

On CD_{ibd} (Table 6.10), we observe that using only the nine reported regions allowed to reach a maximum AUC of 0.655 with Random Forests (with the lowest investigated value of K). The addition of region 2p12 allowed to further increase the AUC to 0.699 with the T-Trees this time (also with the lowest investigated values of IC and K). The adjunction of 7q31 markers also allowed for a further slight AUC gain, a maximum of 0.719 was obtained with the T-Trees. While the use of only 2p12 and 7q31 variables lead to an AUC of 0.628 As the difference between the subsets of variables that were used in both analyses as candidate is small, similar observations are now drawn for the weakly filtered variant of the dataset CD_{wtccc} (the three first row in Table 6.11).

While the gap between the AUC obtained by exploiting all the variables and only considering 9 to 11 regions is important for the light filtered variant of the dataset, that gap becomes much smaller for the strongly filtered version. Obviously, a significant part of the information exploited by the tree methods is located in markers that have been filtered out by the stringent QC filters applied to the CD_{ibd} dataset. The two additional rows at Table 6.11, denoted by HWE and NOT IBD confirm that hypothesis. Using only markers deviating[2] from HWE allowed to reach an AUC of 0.851 while the use of markers being excluded from the CD_{ibd} variant leads to an AUC of 0.878.

Candidates	#	RF			TT			
		100	200	300	10		30	
					5	10	5	10
9loci	322	0.655	0.652	0.651	0.653	0.652	0.648	0.642
9loci,2p12	363	0.677	0.677	0.677	0.699	0.692	0.692	0.685
9loci,2p12,7q31	386	0.682	0.681	0.681	0.719	0.719	0.718	0.714
		20	40	60	2		6	
					5	10	5	10
2p12,7q31	64	0.628	0.628	0.626	0.627	0.625	0.619	0.616
all	436517	0.7			0.749			

Table 6.10 CD_{ibd} AUCs obtained while only using different subset of the available SNPs. "9loci" corresponds to only using markers from the nine reported regions, "9loci,2p12" where we added the 2p12 regions and "9loci,2p12,7q31" where the 2p12 and 7q31 regions are used in addition to the nine reported regions. The last row corresponds to the results we obtained when using all the available variables. Light and dark green shaded cells resp. corresponds to values of K and IC parameters.

[2] Markers with a HWE p-value $< 10^{-5}$ in controls only or a HWE p-value $< 10^{-7}$ in cases only.

Candidates	#	RF			TT			
		100	200	300	10		30	
					5	10	5	10
9loci	340	0.654	0.652	0.652	0.652	0.652	0.647	0.644
9loci,2p12	382	0.674	0.674	0.674	0.700	0.698	0.696	0.693
9loci,2p12,7q31	406	0.678	0.680	0.680	0.725	0.723	0.718	0.713
		20	40	60	2		6	
					5	10	5	10
2p12,7q31	66	0.635	0.635	0.631	0.625	0.622	0.621	0.618
		1000	2000	3000				
HWE	3178	0.851	0.847	0.845				
		1000	2000	3000				
NOT IBD	33123	0.870	0.870	0.878				
all	469612	0.919			0.945			

TABLE 6.11 CD_{wtccc} AUCs obtained while only using different subset of the available SNPs. "9loci" corresponds to only using markers from the nine reported regions, "9loci,2p12" where we added the 2p12 regions and "9loci,2p12,7q31" where the 2p12 and 7q31 regions are used in addition to the nine reported regions. The HWE row corresponds to the case where we only considered the variables deviating from HWE and NOT IBD corresponds to the case where we only used SNPs that are excluded from the CD_{ibd} variant. The last row corresponds to the results we obtained when using all the available variables. Light and dark green shaded cells resp. corresponds to values of K and IC parameters.

A little further down in the ranking

So far, we briefly looked at the 100 first variables and detailed the twenty first ones according to the variable importances of the tree-based methods. Let us now go a little further down in these rankings and see if more reported regions are detected. In addition to the nine reported regions, we will now also include the loci reported in [Wel07] Supplementary Information and the 140 loci reported more recently in [J$^+$13].

Table 6.12 lists the different regions found in the 200 first variables detected by the tree-based methods on the CD_{ibd} dataset. The upper part corresponds to Random Forests variable importances and the bottom part to T-Trees variable importances.

In that Table, we grouped markers and reported regions as follows: given the physical order of the 200 first SNPs found in the variable rankings, we start a region with the first marker and iteratively add the following one if it is at most 20 SNPs away from the previous one. Regions are thus separated by at least 20 SNPs. For the sake of readability, we report only regions containing at least two markers and at the end of each subtable, in gray shaded text, markers that appeared isolated but reported as associated in [Wel07] Supplementary Information. Only details for the most important markers is reported for each group. When a marker is located in a gene[3], the corresponding gene name is reported in parenthesis next to the marker id. The variable rank is also reported in parenthesis next to the importance value in the last column. Using the following conventions, we highlighted the marker IDs:

[3] According to PheGenI: Phenotype-Genotype Integrator (NCBI).

- in red, when they were reported as strongly associated in [Wel07] Supplementary Information (p-value $< 10^{-5}$),
- in orange, when reported as moderately associated in [Wel07] Supplementary Information ($10^{-5} <$ p-value $< 10^{-4}$ and within 200kb of at least one other SNP with a p-value $< 10^{-3}$),
- in cyan, they were found in regions reported with a strong signal when used in the expanded reference group analysis[4],
- in blue, when they correspond to the two regions we identified previously with the T-Trees,
- underlined, when they correspond to markers reported in [J$^+$13],
- with a $^{(*)}$, when they correspond to one of the nine regions previously discussed.

For instance, the first row in Table 6.12 reports a region located on chromosome 1, spanning from 67.31Mb to 67.46Mb. That region is characterised by 10 markers, the most important one in that region is **rs11209026** with an importance of $1.40 \cdot 10^{-2}$ and is ranked at the first position in the Random Forests variables ranking. That SNP was discussed in the [Wel07], it is one of the nine regions we investigated until now, it is also reported in [Wel07] Supplementary Information as strongly associated and is found in the 140 loci reported in [J$^+$13], hence the red, the underline and the asterisk in parenthesis next to the marker name.

As we can see, with Random Forests, 5 of the nine reported regions are selected and well represented in the 100 first variables. There are ten SNPs located in the interleukin 23 receptor regions on chromosome 1, five on chromosome 2 in the **ATG16L1** gene, six around **rs11718165**, twelve on chromosome 5 around **rs17234657** and four in the **NOD2** region on chromosome 16. We also notice the presence of **rs2542151** but it appeared isolated and ranked below the 100 first variables. In addition, **rs931058** on chromosome 5 has been reported in the list of 140 loci. We note that SNP was not reported in [Wel07]. A few other SNPs reported only in [J$^+$13] appeared isolated in these 200 first variables: **rs11260562** at position 108 in the ranking, **rs909813** at position 172, **rs17101358** (104), **rs10923915** (111), **rs11190083** (92) and **rs1751852** at position 171 (not reported in Table 6.12). Interestingly, the most represented region in this ranking corresponds to one of the two regions we mentioned previously as detected only by the tree-based methods. Seventeen SNPs from that region were found in the 200 first variables.

Similarly, with the T-Trees, the six same regions out of the nine are identified. Especially, **rs11209026**, **rs10210302**, **rs17234657** and **rs2076756** were found in the 100 first. We notice the presence of the two "blue" regions. This time, on chromosome 2p12, 35 variables were found in the 200 first and 9 for the 7q31 region. The most important markers in these two regions are positioned first and third in the T-Tree variable ranking. Additionnaly, **rs16884693** was represented by 2 markers and a few more of the 140 loci not reported by the WTCCC appeared isolated and at lower ranks (**rs11260562** (138), **rs17101358** (149), **rs931058** (120), **rs10772590** (169) and **rs2352937** (95)).

[4] The expanded reference group analysis consists in including the cases from other (non related) diseases as controls in the analysis of each disease.

Random Forests

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	67.31	67.46	10	rs11209026$^{(*)}$ (*IL23R*)	0.045	1	$3.52 \cdot 10^{-5}$	$8.24 \cdot 10^{-18}$	$1.40 \cdot 10^{-2}$ (1)
2	45.58	45.58	2	rs3755076	0.087	$2.65 \cdot 10^{-3}$	$6.02 \cdot 10^{-3}$	$5.18 \cdot 10^{-1}$	$5.30 \cdot 10^{-4}$ (48)
2	81.58	81.76	17	rs11887827	0.311	$1.54 \cdot 10^{-7}$	1	$2.42 \cdot 10^{-8}$	$1.27 \cdot 10^{-3}$ (20)
2	233.94	233.97	5	rs10210302$^{(*)}$ (*ATG16L1*)	0.452	$4.25 \cdot 10^{-1}$	$1.98 \cdot 10^{-2}$	$2.22 \cdot 10^{-13}$	$2.79 \cdot 10^{-3}$ (6)
3	49.43	49.68	6	rs11718165$^{(*)}$ (*BSN*)	0.295	$1.33 \cdot 10^{-2}$	$1.60 \cdot 10^{-2}$	$1.70 \cdot 10^{-6}$	$1.19 \cdot 10^{-3}$ (24)
4	114.61	114.62	2	rs17045935 (*ANK2*)	0.095	$2.38 \cdot 10^{-1}$	$1.07 \cdot 10^{-4}$	$5.28 \cdot 10^{-2}$	$6.45 \cdot 10^{-4}$ (39)
5	24.77	24.77	3	rs16893874	0.008	$2.61 \cdot 10^{-1}$	1	$3.18 \cdot 10^{-5}$	$3.32 \cdot 10^{-4}$ (80)
5	40.43	40.61	12	rs17234657$^{(*)}$	0.146	$4.18 \cdot 10^{-1}$	$3.51 \cdot 10^{-1}$	$1.72 \cdot 10^{-13}$	$2.26 \cdot 10^{-3}$ (10)
5	121.75	121.76	2	rs17149128 (*SNCAIP*)	0.122	$1.10 \cdot 10^{-2}$	$1.03 \cdot 10^{-2}$	$4.10 \cdot 10^{-1}$	$1.97 \cdot 10^{-4}$ (166)
5	150.21	150.31	4	rs931058	0.071	$5.64 \cdot 10^{-1}$	1	$1.53 \cdot 10^{-8}$	$5.83 \cdot 10^{-4}$ (44)
6	36.54	36.64	2	rs600382	0.001	1	1	$2.38 \cdot 10^{-5}$	$2.67 \cdot 10^{-4}$ (95)
8	129.88	129.96	4	rs10216909	0.003	1	1	$7.76 \cdot 10^{-5}$	$3.04 \cdot 10^{-4}$ (87)
10	65.96	65.96	2	rs16919914	0.080	$8.80 \cdot 10^{-2}$	$4.00 \cdot 10^{-4}$	$2.22 \cdot 10^{-1}$	$5.20 \cdot 10^{-4}$ (49)
11	130.84	130.84	2	rs1533339 (*NTM*)	0.005	1	1	$2.78 \cdot 10^{-4}$	$2.15 \cdot 10^{-4}$ (145)
16	49.30	49.32	4	rs2076756$^{(*)}$ (*NOD2*)	0.270	$4.50 \cdot 10^{-3}$	$7.62 \cdot 10^{-1}$	$3.95 \cdot 10^{-15}$	$3.88 \cdot 10^{-3}$ (4)
23	89.59	89.64	2	rs6522332	0.160	$3.10 \cdot 10^{-1}$	$5.50 \cdot 10^{-1}$	$3.23 \cdot 10^{-1}$	$2.08 \cdot 10^{-4}$ (155)
7	135.31	135.31	1	rs834771	0.151	$1.01 \cdot 10^{-1}$	$3.37 \cdot 10^{-2}$	$1.25 \cdot 10^{-3}$	$1.91 \cdot 10^{-4}$ (177)
8	77.90	77.90	1	rs10957818	0.024	$7.13 \cdot 10^{-1}$	$6.26 \cdot 10^{-1}$	$2.62 \cdot 10^{-5}$	$2.13 \cdot 10^{-4}$ (151)
14	77.10	77.10	1	rs4903604	0.227	$5.78 \cdot 10^{-3}$	$2.41 \cdot 10^{-2}$	$2.48 \cdot 10^{-3}$	$2.89 \cdot 10^{-4}$ (89)
18	12.77	12.77	1	rs2542151$^{(*)}$	0.180	$3.08 \cdot 10^{-1}$	$9.46 \cdot 10^{-1}$	$7.21 \cdot 10^{-8}$	$2.07 \cdot 10^{-4}$ (156)

T-Trees

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	3.25	3.26	2	rs12409315	0.077	$1.75 \cdot 10^{-1}$	$4.37 \cdot 10^{-3}$	$2.54 \cdot 10^{-3}$	$4.36 \cdot 10^{-4}$ (32)
1	67.31	67.46	10	rs11209026$^{(*)}$ (*IL23R*)	0.045	1	$3.52 \cdot 10^{-5}$	$8.24 \cdot 10^{-18}$	$5.23 \cdot 10^{-3}$ (5)
1	77.61	77.62	2	rs11162341	0.132	$4.01 \cdot 10^{-1}$	$3.78 \cdot 10^{-1}$	$8.99 \cdot 10^{-1}$	$2.28 \cdot 10^{-4}$ (57)
1	236.50	236.50	5	rs6677092 (*RPS7P5*)	0.373	$3.01 \cdot 10^{-6}$	1	$1.77 \cdot 10^{-4}$	$4.15 \cdot 10^{-4}$ (33)
2	81.58	81.85	35	rs11887827	0.311	$1.54 \cdot 10^{-7}$	1	$2.42 \cdot 10^{-8}$	$1.03 \cdot 10^{-2}$ (1)
2	143.22	143.28	2	SNP_A-2293058	0.003	1	1	$1.79 \cdot 10^{-5}$	$1.81 \cdot 10^{-4}$ (78)
2	233.94	233.97	5	rs10210302$^{(*)}$ (*ATG16L1*)	0.452	$4.25 \cdot 10^{-1}$	$1.98 \cdot 10^{-2}$	$2.22 \cdot 10^{-13}$	$3.07 \cdot 10^{-4}$ (48)
3	7.49	7.50	2	rs17047422	0.001	1	1	$3.45 \cdot 10^{-4}$	$1.91 \cdot 10^{-4}$ (73)
3	120.41	120.42	2	rs6774 (*B4GALT4*)	0.108	$2.50 \cdot 10^{-1}$	$7.10 \cdot 10^{-4}$	$1.39 \cdot 10^{-2}$	$3.41 \cdot 10^{-4}$ (43)
3	187.31	187.35	2	rs4686733	0.053	$6.50 \cdot 10^{-5}$	$1.12 \cdot 10^{-2}$	$3.65 \cdot 10^{-1}$	$1.39 \cdot 10^{-4}$ (93)
4	86.13	86.18	2	rs1872321	0.002	1	1	$6.88 \cdot 10^{-9}$	$1.19 \cdot 10^{-3}$ (17)
4	114.61	114.62	2	rs17045935 (*ANK2*)	0.095	$2.38 \cdot 10^{-1}$	$1.07 \cdot 10^{-4}$	$5.28 \cdot 10^{-2}$	$2.57 \cdot 10^{-4}$ (53)
4	178.27	178.28	3	rs1595154	0.002	1	1	$1.08 \cdot 10^{-7}$	$5.70 \cdot 10^{-4}$ (28)
5	40.43	40.53	10	rs17234657$^{(*)}$	0.146	$4.18 \cdot 10^{-1}$	$3.51 \cdot 10^{-1}$	$1.72 \cdot 10^{-13}$	$4.55 \cdot 10^{-4}$ (30)
6	21.33	21.35	2	rs16884693	0.004	1	1	$1.21 \cdot 10^{-3}$	$9.36 \cdot 10^{-5}$ (145)
6	129.84	129.84	3	rs2784899	0.260	$8.06 \cdot 10^{-1}$	$5.25 \cdot 10^{-1}$	$6.48 \cdot 10^{-2}$	$1.26 \cdot 10^{-4}$ (106)
7	35.37	35.37	2	rs10270692	0.066	1	$6.68 \cdot 10^{-1}$	$9.31 \cdot 10^{-2}$	$1.99 \cdot 10^{-4}$ (68)
7	125.13	125.16	9	rs6947579	0.317	$2.45 \cdot 10^{-1}$	$7.02 \cdot 10^{-1}$	$8.54 \cdot 10^{-1}$	$7.55 \cdot 10^{-3}$ (3)
8	129.90	129.92	2	rs10216909	0.003	1	1	$7.76 \cdot 10^{-5}$	$1.03 \cdot 10^{-4}$ (131)
10	38.31	38.38	2	rs11011417	0.001	1	1	$1.85 \cdot 10^{-5}$	$1.31 \cdot 10^{-4}$ (100)
11	14.16	14.16	2	rs9804490	0.459	$1.50 \cdot 10^{-10}$	$1.44 \cdot 10^{-6}$	$2.41 \cdot 10^{-5}$	$1.16 \cdot 10^{-4}$ (117)
12	42.78	42.80	2	rs11613902 (*TMEM117*)	0.099	$3.98 \cdot 10^{-7}$	$9.76 \cdot 10^{-2}$	$9.43 \cdot 10^{-1}$	$3.46 \cdot 10^{-4}$ (41)
14	84.39	84.43	4	rs10144260	0.008	1	1	$1.18 \cdot 10^{-9}$	$1.07 \cdot 10^{-3}$ (18)
14	104.47	104.53	2	rs2819467 (*C14orf79*)	0.011	$3.21 \cdot 10^{-1}$	1	$1.51 \cdot 10^{-3}$	$1.23 \cdot 10^{-4}$ (110)
16	49.30	49.31	3	rs2076756$^{(*)}$ (*NOD2*)	0.270	$4.50 \cdot 10^{-3}$	$7.62 \cdot 10^{-1}$	$3.95 \cdot 10^{-15}$	$6.43 \cdot 10^{-4}$ (25)
23	21.69	21.74	8	rs5904497 (*SMS*)	0.273	$5.97 \cdot 10^{-6}$	$1.50 \cdot 10^{-1}$	$4.41 \cdot 10^{-2}$	$3.26 \cdot 10^{-3}$ (9)
23	70.94	70.94	2	rs6624585 (*NHSL2*)	0.068	$7.76 \cdot 10^{-1}$	1	$2.69 \cdot 10^{-2}$	$2.24 \cdot 10^{-4}$ (58)
3	49.67	49.67	1	rs11718165$^{(*)}$ (*BSN*)	0.295	$1.33 \cdot 10^{-2}$	$1.60 \cdot 10^{-2}$	$1.70 \cdot 10^{-6}$	$7.93 \cdot 10^{-5}$ (159)
5	57.95	57.95	1	rs2279980	0.188	$8.16 \cdot 10^{-2}$	$7.99 \cdot 10^{-1}$	$6.19 \cdot 10^{-5}$	$7.03 \cdot 10^{-5}$ (182)
8	77.90	77.90	1	rs10957818	0.024	$7.13 \cdot 10^{-1}$	$6.26 \cdot 10^{-1}$	$2.62 \cdot 10^{-5}$	$1.06 \cdot 10^{-4}$ (126)
18	12.77	12.77	1	rs2542151$^{(*)}$	0.180	$3.08 \cdot 10^{-1}$	$9.46 \cdot 10^{-1}$	$7.21 \cdot 10^{-8}$	$9.35 \cdot 10^{-5}$ (146)

TABLE 6.12 CD_{ibd}: lists of regions identified by the Random Forests and the T-Trees methods.

Rare and common variants

Another important observation based on Table 6.8 (and also in the two other datasets top rankings at Table A.1 in the Appendix) is that most of the SNPs that are not found in one of the nine regions are of low minor allele frequency. Firstly, it contradicts the finding of [BBLBS12]. Secondly, it suggests that an important part of the information exploited by the random forests is located in rare variants spread all over the genome. Table 6.13 display the AUCs we obtained by using (1) only rare variants (MAF ≤ 0.05) and (2) using only common ones (MAF > 0.05) as candidate attributes with the random forests ($T = 1000$, $N_{min} = 250$). In that Table, the gray numbers correspond to AUCs obtained using all the variables. We see that using only rare/common variants still allows to reach high AUCs (and may comfort the idea that part of the missing heritability might be hidden in rare variants). With rare variants only we also notice that different values of K does not significantly change the AUC value while on the other hand, with common variants only, an increase of K leads to an increase of the AUC. Obviously, when a test node uses a low MAF SNP as a splitting attribute, it certainly produces a child with a small proportion of the current learning set reaching that node (maybe a pure subsample in which case the node becomes a leaf) transforming the structure of the tree in a long branch along which the order of variable appearance is not so important anymore (and thus explains why the AUC is so stable with regards to the K parameter).

MAF	K	$\#_{wtccc}$	CD_{wtccc}	$\#_{ibd}$	CD_{ibd}
(1) ≤ 0.05	1000	102593	0.800 (0.845)	85110	0.668 (0.684)
	2500		0.810 (0.888)		0.668 (0.700)
	5000		0.811 (0.909)		0.671 (0.698)
(2) > 0.05	1000	367019	0.764 (0.845)	351407	0.646 (0.684)
	2500		0.825 (0.888)		0.661 (0.700)
	5000		0.851 (0.909)		0.660 (0.698)
		469612		436517	

TABLE 6.13 AUCs obtained while filtering the list of candidate attributes based on the MAF (minor allele frequency) with the random forest ($T = 1000, N_{min} = 250$). The two columns $\#_{wtccc}$ and $\#_{ibd}$ denote the numbers of SNPs passing the corresponding MAF filters.

Excluding the X chromosome

Because of its specific nature, the X chromosome is often not included in GWAS or is studied separately from the 22 autosomal chromosomes. In our previous experiments, all markers from the chromosome X were considered. Table 6.14 reports the results we obtained when removing the entire chromosome from the pool of candidate variables in our tree-based methods. We found no significant differences between taking and not taking the X chromosome into account for our two Crohn's disease datasets. As the proportion of male and female is well balanced between cases and controls in these datasets, the risk of spurious signal of association related to the individual sex is low.

	CD_{wtccc}		CD_{ibd}	
chr. X	RF	TT	RF	TT
included	0.919	0.945	0.697	0.749
excluded	0.910	0.945	0.696	0.749

TABLE 6.14 Comparison of AUCs obtained while including/excluding chromosome X from the candidate attributes. Parameter settings: RF: $T = 1000$, $K = 10000$, $N_{min} = 250$ and TT: $T = 1000$, $K = 1000$, $IC = 5$, $N_{min} = 2000$.

6.2 A few experiments with linear models

In the present section we carry out a few experiments in order to compare the results obtained with our tree-based methods with several standard approaches to build linear models. While we mostly focus on the assessment of predictive accuracy, we also make a preliminary analysis of the SNPS rankings obtained with such linear models in comparison with those obtained from tree-based variable importances.

6.2.1 Sum of log odds ratio

The odds ratio (OR) corresponds to the ratio between the proportion of cases having a specific allele and the proportion of controls having the same allele. That ratio will be greater than one when the frequency of the allele used as reference is higher in the cases. It denotes how the presence of a specific allele at a given locus increases or decreases the genetic risk. We investigated a simple model summing together the natural logarithm of the allelic OR. Each individual is then assigned with an average score per non-missing SNP defined as:

$$\frac{1}{n}\sum_{i=1}^{n} g_i \times \log(OR_i). \tag{6.1}$$

where n corresponds to the number of non-missing genotypes and g_i denotes the genotype (0,1 or 2) of SNP$_i$ of the considered individual. This model (also called NAIVE-BAYES in the machine learning literature) assumes a class-conditional independence between the markers. We evaluated two variants of this OR based model: in the first experiment, we used the odds ratio derived directly from the datasets as defined in eqn. (6.1), while in the second experiment we used instead the "odds ratio" derived from a logistic regression applied separately to each SNP, denoted OR_{logit} in the following (which corresponds to the β_1 coefficient in the corresponding linear model [HV03]). Both experiments were performed with *plink* [P+07] and the corresponding cumulated scores are used to determine AUCS, according to our 10-fold cross-validation scheme and by using the same folds as before.

	CD_{wtccc}	CD_{ibd}
OR	0.661	0.648
OR_{logit}	0.739	0.729

TABLE 6.15 AUCS obtained with the "log odds ratio" methods on the two CD datasets.

Table 6.15 report the AUCS we obtained. We see that higher AUCS are reached on CD_{wtccc} and that the second experiment (OR_{logit}) produced much better results.

6.2.2 Globally trained linear models

The goal is to learn a linear scoring function in the form $f(x) = w^T x + b$. A common choice to find the model parameters w and b is by minimizing the regularized training error given by:

$$E(w,b) = \sum_{i=1}^{n} L(y_i, f(x_i)) + \alpha R(w) \tag{6.2}$$

where L is a loss function that measures model fit and R is a regularisation term that penalises model complexity, i.e. to prevent overfitting ($\alpha > 0$ is a non-negative hyperparameter). We used the hinge loss function (i.e. $L(y_i, f(x_i)) = max(0, 1 - y_i \cdot f(x_i))$) which is well suited for classification problems. We tested two types of regularisation, namely the $L1$ and the $L2$ norms which are defined as:

$$L1 : R(w) = \sum_{i=1}^{n} |w_i| \tag{6.3}$$

$$L2 : R(w) = \frac{1}{2} \sum_{i=1}^{n} w_i^2 \tag{6.4}$$

In our first experiment, we fitted our linear model with stochastic gradient descent (SGD) [Zha04]. We used an open source machine learning library written in *Python*: *Scikit-learn* [PVG+11]. Table 6.16 reports the averaged AUCs we obtained with our protocol (i.e. using the same 10-fold cross validation while trying to fit a model based on all the available SNPs). For both regularisations, the α values were chosen so as to maximise the reported AUCs. For the $L1$-regularized model we obtained the best result with $\alpha = 1 \cdot 10^{-6}$ and for the $L2$-regularized model the best AUC was reached with $\alpha = 1000$. We notice that the latter regularisation provides better results, while both variants provide slightly better results on the CD_{wtccc} dataset than on the CD_{ibd} one.

	CD_{wtccc}	CD_{ibd}
L1	0.623	0.613
L2	0.643	0.635

TABLE 6.16 AUCs obtained with linear models learnt by stochastic gradient descent, with two types of regularisation and on the two CD datasets.

Also, we investigated the $L2$-regularised logistic regression also available in *Scikit-learn* which is based on LIBLINEAR [FCH+08] and solves the following unconstrained optimisation problem:

$$\min_{w} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(1 + e^{-y_i w^T x_i}) \tag{6.5}$$

Again, under the same conditions and with a grid search for the optimal parameter, we reached an AUC of 0.648 (with $C = 1 \cdot 10^{-5}$) on CD_{wtccc} and 0.638 (with $C = 1 \cdot 10^{-6}$) on CD_{ibd}.

Finally, Table 6.17 summarises all the AUCs we obtained until now. The two first rows in that table corresponds to the best tree-based AUCs. For both datasets, in terms of predictive power the best results are achieved with the T-Trees. While the difference is notable on CD_{wtccc} the gap between the best linear model and the best tree-based one is much smaller on the CD_{ibd} dataset. These preliminary results suggest that:

- As expected and highlighted with the sum of log odds ratio, a notable part of the genetic risk is attributable to a linear combination of the individual SNPs odd ratios. In particular, the fact that variables are considered independent of each other renders this approach quite fast and robust against overfitting. Indeed, in the overall score, small noise effects due to a large number of irrelevant markers, are on average cancelled.
- On the other hand, when we tried to globally fit a linear model on all the variables at once, it appeared as counterproductive. The results we obtained with SGD and the logistic regression are indeed one step behind the sum of log odds ratio method. This may be caused by the very high dimensionality and hence the less effective bias-variance tradeoff of the global weight fitting schemes.
- Although the AUCs difference is less impressive on CD_{ibd} dataset, we saw that the T-Trees were nevertheless able to outperform all the linear methods as well as the random forests.

	CD_{wtccc}	CD_{ibd}
RF	0.919	0.700
TT	**0.955**	**0.765**
OR	0.661	0.648
OR_{logit}	0.739	0.729
SGD-L1	0.623	0.613
SGD-L2	0.643	0.635
Logit	0.648	0.638

Table 6.17 Comparison of AUCs between the tree-based and the linear models.

6.2.3 Preliminary analysis of the SNP rankings

In addition to the assessment of the AUCs of the linear models, we also looked at their ranking of the different SNPs. Given the univariate way of scoring the SNPs in the OR variants, their rankings should essentially be similar to those of the univariate p-values. On the other hand, with the globally trained versions we might expect different rankings. We therefore ranked the SNPs based on $|w_i|$ obtained with the SGD-$L2$ variant, while focusing on the top 100 variables in these rankings. On CD_{ibd} we note, in the overlap, the presence of several SNPs in the 1p31, 2q37, 5p13 and 16q12 WTCCC reported regions and also one SNP in the 2p12 locus (one of the additional regions identified by the T-Trees). Similarly, on CD_{wtccc}, we notice the presence of several markers also found in confirmed regions and, additionally, a few markers with a strong deviation from HWE. These similarities with our previous experiments are comforting, but they also highlight that there is probably more than only linear effects in these datasets, since the tree-based approaches reach higher AUCs than the linear models, particularly on the CD_{wtccc} dataset.

6.3 Discussion

With all the previous results in mind, different conclusions can be drawn. In this chapter, we applied three methods to two variations of the same datasets: a lightly and a strongly QC filtered dataset. We evaluated those methods in terms of their predictions and their ability to identify markers that are associated to *Crohn*'s disease.

6.3.1 Importance of the preprocessings

Independently from the method, we observed consistent behaviour in terms of prediction accuracy on the two variants of the dataset:

- **QC filters**: we saw that the way these kind of data are cleaned is of crucial importance and beside the fact that the score we used is strongly correlated and similar to the standard χ^2/*Fisher* test of association that are of common use in GWAS, it can drastically modify the predictive power and the corresponding variable ranking. At each tree in our different types of forests, nodes are expanded conditionally to the previous ones. When a strongly (but maybe abnormally) associated marker appears in the dataset, it will be considered as important by our methods and appear nearer to the root of trees in an ensemble, prevailing over other weaker (but potentially truly) associated markers.

  Especially, on the weakly filtered dataset, we noticed that the most important variables according to the variable importances in our tree based methods are often those that strongly deviate from the Hardy-Weinberg equilibrium in one of the two groups. The tree based methods seem to be extremely sensitive to this filter, as it reflects an important difference in genotype distributions between the two phenotypes. This issue is strongly related to the quality of the data and is a common problem in the field of machine learning, unlike univariate analysis where suspicious variables can easily be discarded on by one. The stronger filters applied to the CD_{ibd} dataset demonstrate that the removal of such "strangely" deviating markers allowed for a better identification of the nine reported regions. Still, while sometimes such markers appeared isolated (which suggests a potential genotyping error), we noticed, e.g. in a region on chromosome 4, on the CD_{wtccc} that even SNPS not deviating as much as the strongest one were picked as well. In the T-Trees method, one can easily think that the marker deviating from HWE carries with him the other SNPS found in the same block, but that is not true in the random forest context where variables are treated independently from the others.

  Besides all these considerations, some of the QC filters are questionable. How can we be sure about the reasons and the exclusion thresholds used to discard a SNP? As suggested in [ZVSW10], what if, given the way affected individuals are sampled, one or more SNPS just deviate from one of the filters because of the disease. A meticulous inspection of each of the variables seems to be inevitable to guarantee the inference of a reliable tree-based model. Unfortunately, that step, in particular the ones that are SNP based, corresponds to what could be done posterior to a GWAS but becomes a crucial prerequisite to the application of tree-based methods. We noticed that the QC filters essentially affect the predictive power of the methods. Still, even in presence of suspicious markers in the most important ones according to the tree based methods, we saw that many of the reported SNPS were also selected in a consistent manner across datasets and methods in the variable rankings. The presence of such aberrant variables pushed the reported ones a bit downwards in the rankings. To circumvent this problem, one approach would be to repeat the construction of an ensemble of trees while iteratively removing markers that are considered as important while looking suspicious like common genotyping errors do.

  We also noticed that, while the random forest variables importances allowed to detect the markers with the lowest p-values, the T-Trees variable and group importances allowed to spot additional regions that were not associated nor robustly detected by the random forests.

6.3.2 Methods

We are now confident about the power of tree based methods applied to the field of genome-wide association studies. The consistency between the three methods applied to the two dataset variations highlighted the reliability of the ensemble of trees methods behaviour. Most importantly, we notice as expected that the most important variables in the different models are always the ones that are the most statistically associated with the outcome.

ET seemed to beat the RF, it also produces deeper trees while reaching slightly better AUCs. Results between the two methods are almost similar and that's expected given the fact that there are at most two possible splits for each SNP. The only real left difference between those two algorithms being the bootstrap sampling step. As those two methods are quite similar, we decided to focus on the one that is more widely known and the more commonly used in the field of GWAS: Random Forests.

Finally, we evaluated the power of our novel method, taking into account the particular structure of the variables. In every case, T-Trees produced higher AUCs than the two other standard approaches. We confirmed that taking LD into account is an effective way to improve the quality of the model while increasing the robustness against noise of the resulting variable rankings. This method allows to reduce the dimensionality of the problem by dividing it by the size of a bloc. We explored the different parameters and noticed that even with the fastest parameter configurations, the T-Trees reached AUCs comparable to the best results of RF and the ET.

Based on the variable importances, we noticed that most of the time, variables that were confirmed by the WTCCC where also quite well selected by our tree based methods. Besides that promising similarity, interestingly, we also found that some regions were sometimes considered as important by the machine learning approach while being ignored by the univariate statistical approach. Regions such as the one found (on the two datasets variants and with the two tree based methods) on the chromosome 2 (around position 81.5Mb) may draw geneticists attention as it has not been reported in the literature as being associated with the Crohn's disease, at least not yet.

Chapter 7

Six other complex diseases

Contents

In the previous chapter, we focused our effort on and detailed the case of *Crohn*'s disease. Using our latest findings, in this chapter we will investigate the six other diseases related WTCCC datasets. With a systematic comparison between different QC filters and two tree-based methods, we will first look at predictive power and then inspect derived variable importances. This empirical study aims at showing that our previous findings generalise to other datasets exhibiting different "architectures".

7.1 The six other diseases

Following the same principles, we applied our methodology to the six additional WTCCC datasets. For each disease, we generated two dataset variations based on different QC filters. Table 7.1 summarises the 7 diseases (including the Crohn's disease) and the corresponding name conventions we will adopt along the current chapter. The WTCCC filtered versions are subscripted with $_{wtccc}$ while the "ibd"-like filtered are subscripted with $_{qc}$.

For the qc versions, starting from the initial set of 500568 SNPs available on the *Affymetrix* chip, we applied separately the following filters on each disease dataset (in the following order):

1. missing rate per SNP $< 5\%$ (before sample removal)
2. missing rate per individual $< 2\%$
3. heterozygosity per individual $+/-0.2$
4. missing rate per SNP $< 2\%$
5. missing rate difference between case and control $< 2\%$
6. HWE p-value $< 10^{-6}$ (controls only)
7. HWE p-value $< 10^{-10}$ (cases only).

Unfortunately, some of these filters (at least when applied to the CD_{ibd} dataset) were fine tuned under visual inspection of various plots (e.g. heterozygosity rate vs. missing rate), allowing the manual removal of outliers and were not exactly reproducible. For that reason, we obtained another variant of the Crohn's disease dataset called CD_{qc} (as reported in Table 7.1) which is slightly different from the CD_{ibd} variant. Figure 7.1 represents these three resulting datasets (CD_{wtccc}, CD_{ibd} and CD_{qc}).

Disease	Light QC	Strong QC
Bipolar disorder	BD_{wtccc}	BD_{qc}
Coronary artery disease	CAD_{wtccc}	CAD_{qc}
Crohn's disease	CD_{wtccc}	CD_{qc} ($\neq CD_{ibd}$)
Hypertension	HT_{wtccc}	HT_{qc}
Rheumatoid arthritis	RA_{wtccc}	RA_{qc}
Type 1 diabetes	$T1D_{wtccc}$	$T1D_{qc}$
Type 2 diabetes	$T2D_{wtccc}$	$T2D_{qc}$

TABLE 7.1 The 7 diseases from the WTCCC dataset and the corresponding naming conventions adopted along this chapter.

Table 7.2 summarises the QC filter impact for each disease. Columns "#ind." and "#SNP" refers to the initial number of individuals and SNPs respectively. The five next columns denote the remaining number of individuals or SNPs depending on the corresponding filter that was applied. As no filter is related to the minor allele frequency, the final numbers of SNPs in this table includes $3.65 \pm 0.23\%$ of monomorphic markers. Finally, Table 7.3 list the overlap size between the $wtccc$ datasets, the list of excluded SNPs in WTCCC (comprising 30956 SNPs) and the corresponding qc versions.

Dataset	#ind.	#SNP	qc_1 (SNP)	qc_2 (ind.)	qc_3 (ind.)	qc_4 (SNP)	$qc_{5,6,7}$ (SNP)
BD_{qc}	5002	500568	483331	4949	4943	460219	454409
CAD_{qc}	4992	500568	484645	4928	4928	463113	457400
CD_{qc}	5009	500568	484798	4934	4922	463144	457609
HT_{qc}	5005	500568	484293	4947	4937	462089	456448
RA_{qc}	5003	500568	484306	4926	4917	462432	456642
$T1D_{qc}$	5004	500568	484357	4975	4966	462342	456767
$T2D_{qc}$	5003	500568	483915	4942	4930	461284	455671

TABLE 7.2 Quality control filters influence on the number of SNPs and individuals for each of the 6 other WTCCC datasets.

Dataset	$wtccc$	qc	$wtccc \cap qc$	$qc \cap excluded$
BD	469612	454409	449754	4655
CAD	469612	457400	452153	5247
CD	469612	457609	452482	5127
HT	469612	456448	451548	4900
RA	469612	456642	451471	5171
$T1D$	469612	456767	451721	5046
$T2D$	469612	455671	450784	4887

TABLE 7.3 First column: markers included in the $wtccc$ versions, second column: markers included in our qc versions, third column: markers found in both versions and last column: markers found in qc versions that were excluded from the $wtccc$.

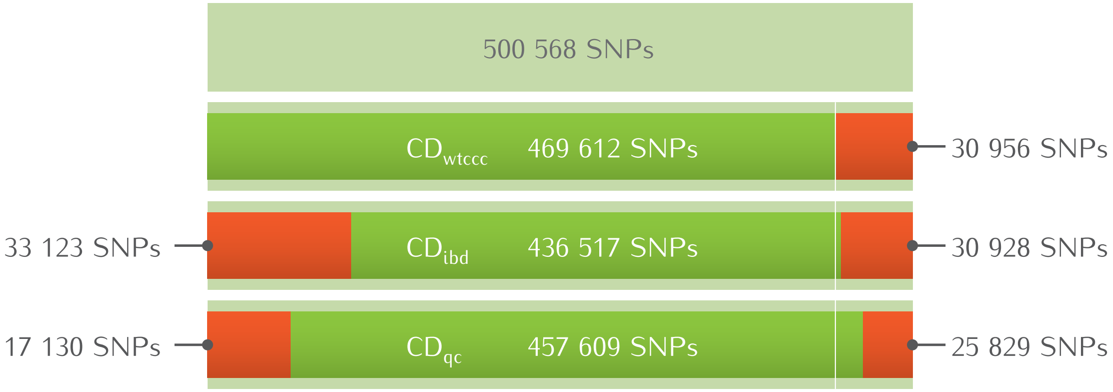


FIGURE 7.1 Proportion of markers being excluded for each of the 3 Crohn's disease dataset versions. In light green, the full set of available SNPs. In green, the markers included in the study and in orange, the excluded markers.

7.2 Comparison of the tree-based methods predictive power

On these 14 (2×7) datasets, we applied the random forests and the T-Trees with the near optimal parameters we identified and used in the previous chapter. Table 7.4 summarises the AUCs we obtained for each of these experiments.

	qc		$wtccc$		$qc + wtccc$	
	RF	TT	RF	TT	RF	TT
BD	0.743	0.813	0.918	0.959	0.683	0.756
CAD	0.756	0.814	0.998	0.999	0.675	0.771
CD	0.776	0.801	0.919	0.945	0.735	0.762
HT	0.807	0.866	0.938	0.969	0.692	0.799
RA	0.806	0.830	0.993	0.996	0.747	0.763
$T1D$	0.860	0.870	0.900	0.940	0.852	0.860
$T2D$	0.758	0.834	0.959	0.979	0.705	0.788

Table 7.4 AUC comparisons: RF and TT results on two variants of the 7 WTCCC datasets. The $_{qc}$ columns corresponds to the "idb"-like filtered variant and the $_{wtccc}$ to the lightly filtered variant. (parameter settings: RF: $T = 1000$, $K = 10000$, $N_{min} = 250$ and TT: $T = 1000$, $K = 1000$, $IC = 5$, $N_{min} = 2000$)

A first observation can be drawn when comparing our two tree-based methods. T-Trees were still able to outperform the random forests in term of predictive accuracy (even when there is not much room left for improvement). We observed the same tendency on the two CD datasets in the previous Chapter. No matter how the datasets were preprocessed, taking into account the structure of the descriptors allowed for a notable AUC increase in a very consistent way.

Similarly, a comparison between $wtccc$ and qc filtered dataset versions confirmed the impact of the filters. The "lighter" filters of the WTCCC allowed the two type of decision tree forests to reach unexpectedly high AUCs. Especially for CAD and RA, Random Forests and T-Trees were able to almost perfectly predict individual disease statuses. On these two datasets, we noticed that the removal of certain types of variable decreased the predictive power. In the most extreme case, with RF on CAD, AUC dropped from 0.998 to 0.756. In the next section, we will investigate tree-based derived variables importances to see what filters are responsible for this quite important changes.

As a third and last observation, regarding the CD datasets, we obtained a different AUC on CD_{qc} in comparison with CD_{ibd}. We were not able to reproduce the exact same QC filtering and that lead to an AUC increase with a difference of 0.076 with the RF and 0.052 with the T-Trees.

7.3 Variable importances analyses

Out of the 14 models, we computed RF and TT variables importances. As Figures 6.11 and 6.12, the following coloured histograms (Figures 7.2-7.15) allow us to quickly see what type of variables are considered as important in the resulting predictive models. As a reminder, in these variable importance histograms, we use the following colour scheme:

- based on [Wel07] Supplementary Information, in the first row of each figures, we coloured in red the markers reported with a p-value $< 10^{-5}$. In orange those exhibiting a p-value between 10^{-5} and 10^{-4} and within 200kb of at least one other SNP with a p-value below 10^{-3}. In cyan, markers found in regions reported with a strong signal when used in the expanded reference group analysis[1],

[1] The expanded reference group analysis consists in including the cases from other (non related) diseases as controls in the analysis of each disease.

- in purple in the second rows, the rare variants (MAF < 0.05),
- in the $wtccc$ versions, we added a third row where orange corresponds to markers deviating from HWE (either in controls with a p-value $< 10^{-6}$ or cases with a p-value $< 10^{-10}$),
- in green in the last rows, the SNPs with a Fisher exact p-value $< 10^{-6}$.

Also, Tables 7.5-7.19 list the different regions found in the 200 first variables according to the tree-based variable importances (in each Table, the upper part corresponds to Random Forests variables importances and bottom subtable to T-Trees variable importances). Given the physical order of the 200 first SNPs, we start a region with the first marker and iteratively add the following one if it is at most 20 SNPs away from the previous one. Regions are thus separated by at least 20 SNPs. For the sake of readability, we report only regions containing at least two markers and at the end of each subtable, in gray shaded text, markers that appeared isolated but reported as associated in [Wel07] Supplementary Information. Only details for the most important markers are reported for each group. When a marker is located in a gene[2], its name is reported in parenthesis next to the marker ID. The variable rank is also reported in parenthesis next to the importance value in the last column.

In the qc variants, we noticed the presence of SNPs being excluded from the $wtccc$ variants. The corresponding marker identifiers are superscripted by a parenthesised number: (1) corresponds to the first WTCCC exclusion criteria (i.e. missing rate $> 5\%$ or $> 1\%$ for the markers exhibiting a study-wise MAF $< 5\%$) and (3) to the third one (i.e. p-value $< 5.7 \times 10^{-7}$ for either a one- or two-degree of freedom test of association between the two control groups). Finally, based on [Wel07] supplementary information, we highlighted the markers id using the colour scheme as in Figures 7.2-7.15.

Additionally, in the Appendix, Figures B.1-B.7 position these first 100 variables on their respective chromosome. Blue points and triangles correspond to random forest variable importances. Orange points and squares correspond to T-Trees variable importances. Those figures enable us to:

- compare and find regions that are consistently identified in both $wtccc$ and qc filtered variants,
- localise the most important variables, these deviate from the center horizontal line,
- visualise agreement between the two tree based variable rankings (which also corresponds to the gray shaded rows in the previously mentionned Tables),
- spot regions with several concentrated hits. Such loci, where points overlap, are indicated by a higher (triangle and/or square) opacity.

In the seven following subsections, we study and compare for each disease the differences between quality control filters and method behaviours. Since our biological knowledge is far from complete, we will try to correlate when available, in a non exhaustive manner, our findings with other research results and applications related to the same datasets and/or studying the same phenotypes. Finally, we will discuss the overall results.

[2] According to PheGenI: Phenotype-Genotype Integrator (NCBI).

Bipolar disorder

A first look at Figures 7.2 and 7.3 shows that, in the four experiments, only a few reported regions are considered as important. The only common SNP that is selected in the 100 first variables is rs975687 located on chromosome X. Although its p-value is not that low in comparison with other variables present in the datasets and no particular deviation from HWE, both tree-based methods catch that signal in the two dataset variants.

We also notice that on BD_{qc}, rare variants are much more selected with our QC filters. On the other hand, on the WTCCC dataset, there are fewer rare variants considered as important but many important markers strongly deviate from HWE. Table 7.6 reflects these observations. We can see many markers showing both strong deviation from HWE and strong signal of association according to the variable importances of the two tree-based methods. These markers were removed from the BD_{qc} dataset. And among these, we found that rs13126272 on chromosome 4 was also reported in [BSTB10].

In the first 100 variables of both tree-based methods on BD_{qc}, rs420259 is located in one of the WTCCC reported regions although it has a lower rank with the T-Trees. As it appears isolated in the two rankings, it is reported in light gray at the end of the two subtables of Table 7.5. On the same dataset variants, RF spotted another reported region characterised by two variables on chromosome 2.

We notice that several SNPs excluded from the WTCCC are found in the 200 first variables on BD_{qc}. Contrary to what is suggested by the two subtables of Table 7.5, there are more of them in the Random Forests ranking but they just appear isolated, while due to the way variables are treated in the T-Trees, they were less isolated with this latter method.

Interestingly, we found that rs12355606 (located on chromosome 10 and quite well ranked in the four experiments) falls in the CACNB2 gene. That gene has been reported in [A+13] as significantly associated with bipolar disorder. It is related to the CACNA1C gene which has been reported in many other association studies [F+08, SdS+12]. Note that this region was not reported in [Wel07] and is detected by both tree-based approaches on the two BD dataset versions. rs7821190 is located in NRG1 gene which is linked to bipolar disorder and other related diseases [G+08a, T+07, G+05b].

Similarly, rs1553460 is reported in [P+12a] has being a "hub SNP". That marker appears with both Random Forests and T-Trees in groups of ±10 variables on BD_{wtccc}. On BD_{qc}, Random Forest lost that signal, but T-Trees identifies, at a much lower rank, rs1503865 that is located near rs1553460 (even though rs1553460 itself has been removed from this version of the dataset because of its strong deviation from HWE).

Finally, we can see that both methods are in agreement: when several SNPs in the same region are selected in the 200 first variables, they are found by the Random Forests and the T-Trees approaches. On BD_{qc}, the widest region selected correspond to rs852996, which is a rare variant excluded from the *wtccc*. It has one of the lowest p-values on this dataset. The first one being rs10212068 and is found at position 1 by both tree-based approaches and was excluded from the *wtccc* because of its ability to dissociate the two subgroups of controls. We will see that this one re-appears in most of the other diseases considered in the following subsections.

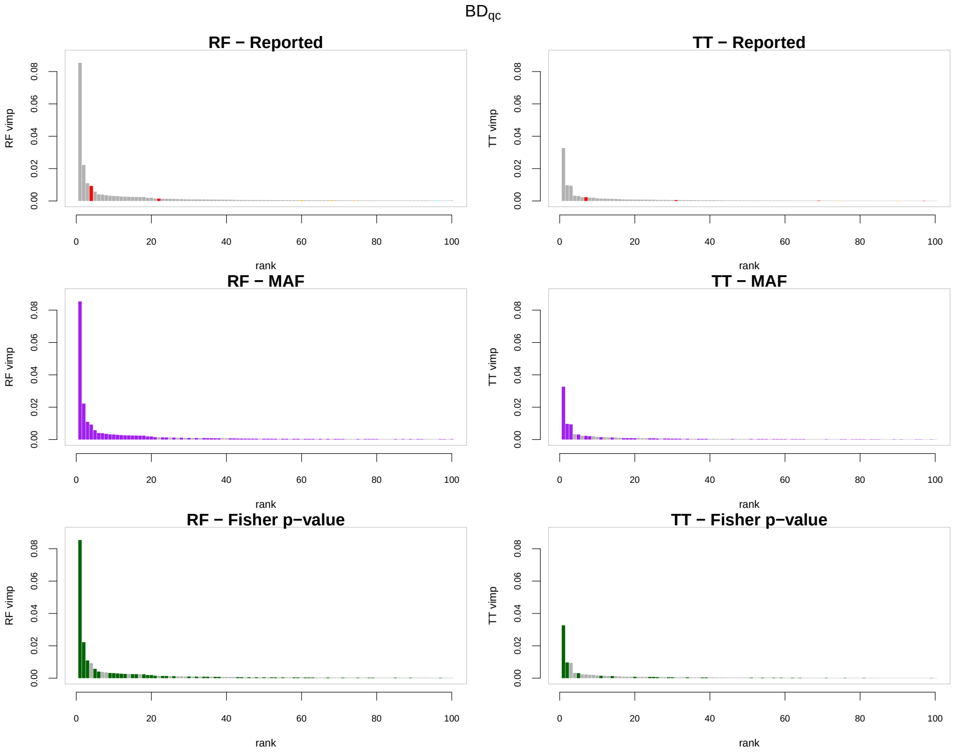


Figure 7.2 The first 100 variables according to the tree based importance rankings for BD_{qc}. The horizontal axis corresponds to the ranks and the vertical axis to the variable importances. In the first column variables are ordered according to random forests variable importances and in the second column they are ordered according to the T-Trees variable importances. In the first row, red highlights the reported strongly associated regions. In the second row, purple corresponds to rare variants (MAF < 0.05). In the third row, green represents markers with a low Fisher p-value ($< 10^{-6}$).

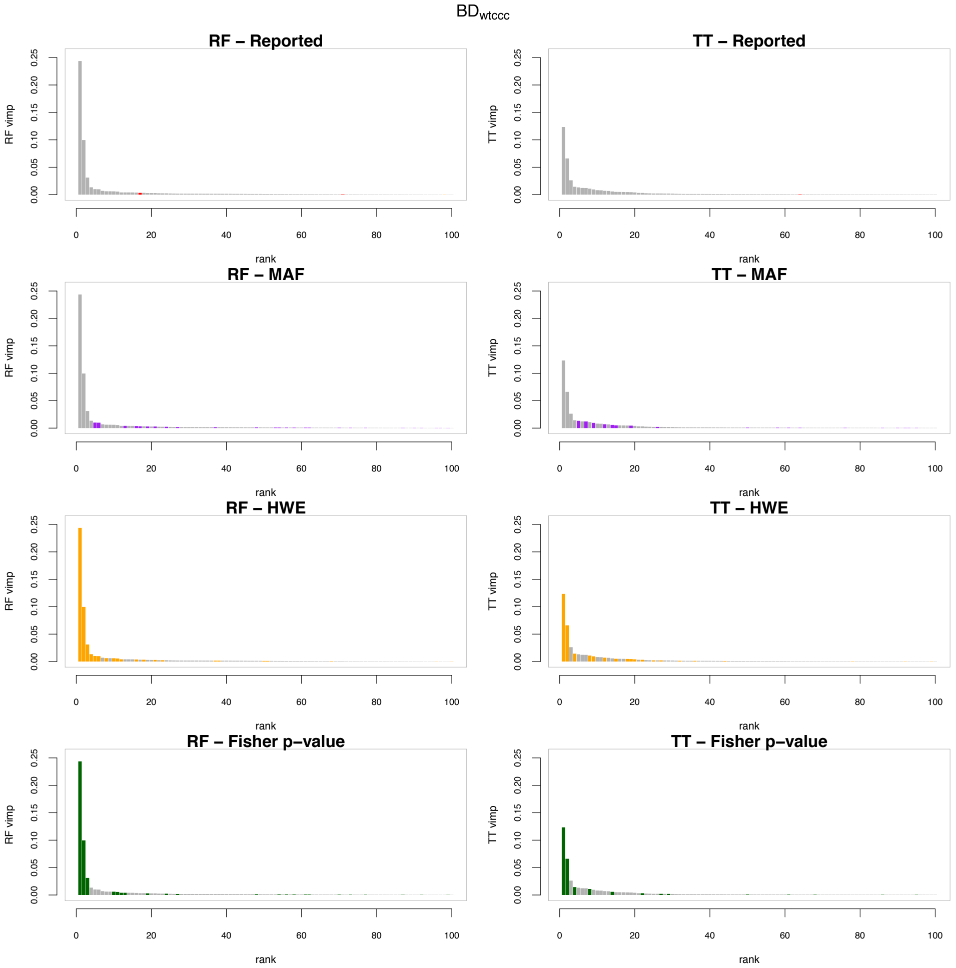


FIGURE 7.3 The first 100 variables according to the tree based importance rankings for BD_{wtccc}. The horizontal axis corresponds to the ranks and the vertical axis to the variable importances. In the first column variables are ordered according to random forests variable importances and in the second column they are ordered according to the T-Trees variable importances. In the first row, red highlights the strongly associated reported regions. In the second row, purple corresponds to rare variants (MAF < 0.05). In the third row, orange highlights SNPs deviating from HWE and in the last row, green represents markers with a low Fisher p-value ($< 10^{-6}$).

Random Forests

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	226.95	226.99	5	rs852996$^{(1)}$	0.009	$2.55 \cdot 10^{-4}$	1	$1.05 \cdot 10^{-17}$	$2.22 \cdot 10^{-2}$ (2)
1	234.34	234.34	3	rs6680230	0.181	$7.57 \cdot 10^{-4}$	$2.25 \cdot 10^{-1}$	$3.48 \cdot 10^{-2}$	$1.78 \cdot 10^{-4}$ (151)
2	241.24	241.24	2	rs2953145 (*RNPEPL1*)	0.212	$6.78 \cdot 10^{-3}$	$9.58 \cdot 10^{-1}$	$2.36 \cdot 10^{-5}$	$2.59 \cdot 10^{-4}$ (102)
6	89.34	89.38	3	rs2610769	0.150	$1.10 \cdot 10^{-1}$	$2.16 \cdot 10^{-3}$	$2.11 \cdot 10^{-2}$	$2.13 \cdot 10^{-4}$ (122)
8	32.03	32.03	2	rs16878694 (*NRG1*)	0.001	1	1	$5.96 \cdot 10^{-6}$	$4.18 \cdot 10^{-4}$ (67)
9	21.85	22.00	2	rs10123713(*MTAP*)$^{(1)}$	0.005	1	1	$4.69 \cdot 10^{-11}$	$1.92 \cdot 10^{-3}$ (20)
10	18.53	18.55	3	rs12355606 (*CACNB2*)	0.006	1	1	$1.02 \cdot 10^{-12}$	$1.10 \cdot 10^{-2}$ (3)
10	77.12	77.15	2	rs7082404	0.004	1	1	$6.77 \cdot 10^{-7}$	$1.23 \cdot 10^{-3}$ (26)
11	10.16	10.19	3	rs1822295 (*SBF2*)	0.057	$4.83 \cdot 10^{-2}$	$2.86 \cdot 10^{-3}$	$7.42 \cdot 10^{-2}$	$2.41 \cdot 10^{-4}$ (109)
14	53.68	53.70	2	rs743276	0.105	$1.64 \cdot 10^{-1}$	$1.16 \cdot 10^{-3}$	$1.67 \cdot 10^{-1}$	$2.08 \cdot 10^{-4}$ (127)
23	32.36	32.37	5	rs3928369 (*DMD*)	0.241	$6.54 \cdot 10^{-3}$	$1.84 \cdot 10^{-1}$	$2.29 \cdot 10^{-3}$	$1.13 \cdot 10^{-3}$ (27)
23	35.97	36.03	4	rs17273161	0.103	$7.33 \cdot 10^{-1}$	$6.78 \cdot 10^{-1}$	$6.06 \cdot 10^{-4}$	$4.31 \cdot 10^{-4}$ (64)
23	110.31	110.32	2	rs975687 (*CAPN6*)	0.030	1	1	$2.43 \cdot 10^{-4}$	$9.27 \cdot 10^{-3}$ (4)
23	134.83	134.83	2	rs12689820 (*SLC9A6*)	0.031	$6.25 \cdot 10^{-1}$	$2.59 \cdot 10^{-1}$	$1.53 \cdot 10^{-2}$	$1.12 \cdot 10^{-3}$ (28)
23	144.00	144.00	2	rs5966463	0.182	$1.28 \cdot 10^{-3}$	$5.57 \cdot 10^{-1}$	$4.12 \cdot 10^{-2}$	$5.43 \cdot 10^{-4}$ (49)
3	7.63	7.63	1	rs1485171	0.158	$3.06 \cdot 10^{-3}$	$7.17 \cdot 10^{-3}$	$1.21 \cdot 10^{-1}$	$3.47 \cdot 10^{-4}$ (74)
3	184.35	184.35	1	rs514636	0.089	$1.55 \cdot 10^{-4}$	$3.70 \cdot 10^{-1}$	$5.54 \cdot 10^{-6}$	$1.44 \cdot 10^{-4}$ (195)
7	22.76	22.76	1	rs2286492	0.097	$9.60 \cdot 10^{-4}$	$5.29 \cdot 10^{-3}$	$4.25 \cdot 10^{-1}$	$4.56 \cdot 10^{-4}$ (60)
8	34.36	34.36	1	rs2609653	0.061	$7.48 \cdot 10^{-1}$	$3.50 \cdot 10^{-1}$	$1.82 \cdot 10^{-6}$	$1.94 \cdot 10^{-4}$ (136)
8	58.48	58.48	1	rs2875734	0.052	$2.86 \cdot 10^{-7}$	$6.80 \cdot 10^{-1}$	$1.40 \cdot 10^{-3}$	$4.09 \cdot 10^{-4}$ (68)
12	73.67	73.67	1	rs1526805	0.052	$1.29 \cdot 10^{-3}$	$6.54 \cdot 10^{-2}$	$4.05 \cdot 10^{-2}$	$2.73 \cdot 10^{-4}$ (96)
14	57.19	57.19	1	rs10134944	0.097	$3.18 \cdot 10^{-1}$	$2.41 \cdot 10^{-1}$	$2.22 \cdot 10^{-6}$	$2.10 \cdot 10^{-4}$ (126)
14	75.15	75.15	1	rs3784005	0.019	1	$6.85 \cdot 10^{-1}$	$3.37 \cdot 10^{-5}$	$1.60 \cdot 10^{-4}$ (173)
16	23.54	23.54	1	rs420259	0.269	$1.36 \cdot 10^{-4}$	$1.11 \cdot 10^{-2}$	$3.78 \cdot 10^{-4}$	$1.47 \cdot 10^{-3}$ (22)

T-Trees

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	18.83	18.84	2	rs2789336$^{(1)}$	0.007	1	1	$3.50 \cdot 10^{-8}$	$8.59 \cdot 10^{-4}$ (20)
1	36.30	36.31	2	rs562929(*MAP7D1*)$^{(3)}$	0.003	1	1	$2.02 \cdot 10^{-6}$	$2.65 \cdot 10^{-4}$ (64)
1	226.94	226.99	10	rs852996$^{(1)}$	0.009	$2.55 \cdot 10^{-4}$	1	$1.05 \cdot 10^{-17}$	$9.71 \cdot 10^{-3}$ (2)
2	5.37	5.38	4	rs1453783$^{(3)}$	0.371	$7.71 \cdot 10^{-1}$	$1.62 \cdot 10^{-2}$	$7.98 \cdot 10^{-1}$	$1.67 \cdot 10^{-3}$ (10)
2	10.91	10.91	2	rs902133$^{(1)}$	0.004	1	1	$3.72 \cdot 10^{-10}$	$1.29 \cdot 10^{-3}$ (14)
2	53.60	53.61	3	rs903229	0.495	$6.51 \cdot 10^{-1}$	$8.54 \cdot 10^{-1}$	$8.69 \cdot 10^{-1}$	$1.82 \cdot 10^{-4}$ (94)
2	111.30	111.30	2	rs12474907 (*ACOXL*)	0.048	$8.07 \cdot 10^{-1}$	$2.24 \cdot 10^{-1}$	$8.48 \cdot 10^{-1}$	$2.29 \cdot 10^{-4}$ (75)
2	133.02	133.02	2	rs10188442 (*GPR39*)	0.204	$9.45 \cdot 10^{-1}$	1	$6.83 \cdot 10^{-1}$	$4.36 \cdot 10^{-4}$ (41)
3	66.60	66.60	2	rs3845903 (*LRIG1*)	0.006	1	1	$2.91 \cdot 10^{-8}$	$2.78 \cdot 10^{-4}$ (62)
3	96.66	96.66	2	rs5000487	0.427	$6.75 \cdot 10^{-1}$	$9.40 \cdot 10^{-1}$	$2.97 \cdot 10^{-1}$	$1.24 \cdot 10^{-4}$ (126)
4	17.91	17.92	3	rs1503865	0.270	$2.65 \cdot 10^{-1}$	$8.87 \cdot 10^{-1}$	$1.80 \cdot 10^{-2}$	$1.09 \cdot 10^{-4}$ (150)
4	23.66	23.68	5	rs582804	0.156	$4.41 \cdot 10^{-1}$	1	$9.32 \cdot 10^{-1}$	$4.94 \cdot 10^{-4}$ (35)
5	36.46	36.50	2	rs2455278$^{(1)}$	0.010	1	1	$1.37 \cdot 10^{-8}$	$1.25 \cdot 10^{-4}$ (125)
5	87.35	87.35	2	rs4916819	0.042	$6.88 \cdot 10^{-2}$	$1.77 \cdot 10^{-1}$	$2.79 \cdot 10^{-1}$	$1.29 \cdot 10^{-4}$ (123)
6	32.87	32.88	3	rs2157082$^{(3)}$	0.445	$8.20 \cdot 10^{-1}$	$3.95 \cdot 10^{-4}$	$1.30 \cdot 10^{-2}$	$2.17 \cdot 10^{-4}$ (82)
6	96.00	96.08	3	rs6928585	0.156	$1.99 \cdot 10^{-1}$	$8.35 \cdot 10^{-1}$	1	$4.41 \cdot 10^{-4}$ (40)
8	58.48	58.49	3	rs2875734	0.052	$2.86 \cdot 10^{-7}$	$6.80 \cdot 10^{-1}$	$1.40 \cdot 10^{-3}$	$2.34 \cdot 10^{-4}$ (74)
8	120.42	120.42	5	rs2469997	0.190	$4.30 \cdot 10^{-1}$	$9.04 \cdot 10^{-1}$	$2.17 \cdot 10^{-1}$	$1.23 \cdot 10^{-3}$ (15)
10	18.53	18.55	6	rs12355606 (*CACNB2*)	0.006	1	1	$1.02 \cdot 10^{-12}$	$3.12 \cdot 10^{-3}$ (5)
10	59.22	59.25	5	rs1505923	0.136	$3.75 \cdot 10^{-6}$	$6.90 \cdot 10^{-1}$	$2.54 \cdot 10^{-1}$	$8.36 \cdot 10^{-4}$ (22)
10	77.08	77.12	3	rs11001473	0.007	1	1	$2.21 \cdot 10^{-1}$	$6.42 \cdot 10^{-4}$ (28)
10	123.23	123.24	2	rs1613776(*FGFR2*)$^{(1)}$	0.036	$7.18 \cdot 10^{-2}$	1	$8.38 \cdot 10^{-2}$	$4.79 \cdot 10^{-4}$ (37)
10	131.97	132.00	2	rs7918047	0.056	$6.42 \cdot 10^{-1}$	$4.04 \cdot 10^{-1}$	$1.28 \cdot 10^{-1}$	$3.65 \cdot 10^{-4}$ (45)
11	37.36	37.44	2	rs11034154	0.020	$4.86 \cdot 10^{-1}$	1	$4.18 \cdot 10^{-1}$	$3.23 \cdot 10^{-4}$ (53)
12	127.05	127.06	5	rs6489228	0.481	$7.52 \cdot 10^{-1}$	$9.12 \cdot 10^{-1}$	$5.37 \cdot 10^{-1}$	$3.25 \cdot 10^{-3}$ (4)
13	78.36	78.36	2	rs1218285	0.042	$7.64 \cdot 10^{-2}$	$5.97 \cdot 10^{-3}$	$7.59 \cdot 10^{-1}$	$2.98 \cdot 10^{-4}$ (58)
14	88.45	88.45	2	rs2401778	0.287	$7.01 \cdot 10^{-1}$	$9.28 \cdot 10^{-1}$	$1.92 \cdot 10^{-1}$	$1.18 \cdot 10^{-4}$ (132)
15	98.87	98.88	2	rs1393940(*CERS3*)$^{(1)}$	0.034	$2.57 \cdot 10^{-1}$	$1.48 \cdot 10^{-2}$	$5.24 \cdot 10^{-7}$	$2.28 \cdot 10^{-4}$ (76)
17	45.48	45.49	2	rs17774763(*LOC284080*)$^{(1)}$	0.028	$4.00 \cdot 10^{-1}$	$7.31 \cdot 10^{-1}$	$8.02 \cdot 10^{-1}$	$5.30 \cdot 10^{-4}$ (32)
17	70.40	70.50	3	rs1873598 (*CDR2L*)	0.002	1	1	$6.10 \cdot 10^{-8}$	$7.67 \cdot 10^{-4}$ (25)
19	22.70	22.71	4	rs12980129	0.028	1	1	$8.27 \cdot 10^{-5}$	$9.09 \cdot 10^{-4}$ (17)
20	0.89	0.89	2	rs2207323 (*RSPO4*)	0.130	$5.35 \cdot 10^{-5}$	$1.30 \cdot 10^{-3}$	$4.25 \cdot 10^{-1}$	$1.56 \cdot 10^{-4}$ (106)
21	39.35	39.37	7	rs999789	0.058	$3.01 \cdot 10^{-2}$	$3.50 \cdot 10^{-1}$	$7.96 \cdot 10^{-4}$	$1.46 \cdot 10^{-3}$ (12)
22	35.92	35.97	8	rs10212068$^{(3)}$	0.028	1	$2.64 \cdot 10^{-3}$	$3.39 \cdot 10^{-58}$	$3.27 \cdot 10^{-2}$ (1)
23	20.34	20.36	2	rs1350838	0.099	$2.02 \cdot 10^{-1}$	$1.66 \cdot 10^{-1}$	$2.16 \cdot 10^{-1}$	$1.16 \cdot 10^{-4}$ (138)
23	32.36	32.39	6	rs3928369 (*DMD*)	0.241	$6.54 \cdot 10^{-3}$	$1.84 \cdot 10^{-1}$	$2.29 \cdot 10^{-3}$	$2.04 \cdot 10^{-4}$ (86)
23	110.26	110.32	6	rs975687 (*CAPN6*)	0.030	1	1	$2.43 \cdot 10^{-4}$	$2.30 \cdot 10^{-3}$ (7)
23	134.83	134.86	3	rs12689820 (*SLC9A6*)	0.031	$6.25 \cdot 10^{-1}$	$2.59 \cdot 10^{-1}$	$1.53 \cdot 10^{-2}$	$2.09 \cdot 10^{-4}$ (84)
1	54.96	54.96	1	rs10888879	0.034	$1.07 \cdot 10^{-1}$	$7.26 \cdot 10^{-1}$	$5.60 \cdot 10^{-6}$	$8.24 \cdot 10^{-5}$ (191)
6	18.29	18.29	1	rs365237	0.077	1	$7.05 \cdot 10^{-5}$	$1.03 \cdot 10^{-4}$	$1.91 \cdot 10^{-4}$ (90)
16	23.54	23.54	1	rs420259	0.269	$1.36 \cdot 10^{-4}$	$1.11 \cdot 10^{-2}$	$3.78 \cdot 10^{-4}$	$1.72 \cdot 10^{-4}$ (97)
22	35.66	35.66	1	rs16997510	0.021	$8.31 \cdot 10^{-2}$	1	$1.85 \cdot 10^{-5}$	$9.88 \cdot 10^{-5}$ (163)

TABLE 7.5 BD_{qc}: lists of regions identified by the Random Forests and the T-Trees methods.

Random Forests

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
2	25.31	25.36	3	rs2164411	0.154	$2.04 \cdot 10^{-10}$	$1.95 \cdot 10^{-1}$	$1.21 \cdot 10^{-3}$	$3.32 \cdot 10^{-3}$ (18)
4	17.73	17.90	9	rs1553460	0.313	$8.85 \cdot 10^{-94}$	$2.88 \cdot 10^{-5}$	$1.28 \cdot 10^{-28}$	$3.11 \cdot 10^{-2}$ (3)
4	186.09	186.11	3	rs13126272 (*ACSL1*)	0.341	$2.30 \cdot 10^{-43}$	$3.02 \cdot 10^{-2}$	$1.27 \cdot 10^{-6}$	$6.02 \cdot 10^{-3}$ (10)
8	19.35	19.51	19	rs17480050 (*CSGALNACT1*)	0.119	0	1	0	$2.44 \cdot 10^{-1}$ (1)
8	32.03	32.03	2	rs7821190 (*NRG1*)	0.004	1	1	$8.11 \cdot 10^{-7}$	$2.74 \cdot 10^{-4}$ (101)
8	82.41	82.41	2	rs1909936	0.018	$2.40 \cdot 10^{-40}$	$6.29 \cdot 10^{-1}$	$1.54 \cdot 10^{-1}$	$2.95 \cdot 10^{-3}$ (21)
10	18.53	18.55	3	rs12355606 (*CACNB2*)	0.006	1	1	$8.35 \cdot 10^{-12}$	$4.06 \cdot 10^{-3}$ (13)
11	113.31	113.31	5	rs17116117 (*HTR3B*)	0.041	$1.20 \cdot 10^{-1}$	$1.16 \cdot 10^{-1}$	$8.54 \cdot 10^{-10}$	$8.77 \cdot 10^{-4}$ (61)
15	40.73	41.32	28	rs12050604 (*UBR1*)	0.274	$1.41 \cdot 10^{-19}$	$9.65 \cdot 10^{-1}$	$4.19 \cdot 10^{-11}$	$4.07 \cdot 10^{-3}$ (12)
16	79.62	79.70	15	rs1048194 (*CENPN*)	0.095	$2.22 \cdot 10^{-322}$	$9.95 \cdot 10^{-6}$	$3.66 \cdot 10^{-39}$	$9.96 \cdot 10^{-2}$ (2)
17	17.27	17.27	2	SNP_A-1948953	0.282	$9.93 \cdot 10^{-24}$	$5.18 \cdot 10^{-3}$	$3.97 \cdot 10^{-5}$	$6.20 \cdot 10^{-3}$ (8)
18	38.16	38.22	3	rs1442650 (*LOC284260*)	0.010	1	$5.24 \cdot 10^{-1}$	$2.32 \cdot 10^{-13}$	$3.02 \cdot 10^{-3}$ (19)
23	32.36	32.37	5	rs3928369 (*DMD*)	0.241	$1.78 \cdot 10^{-2}$	$2.58 \cdot 10^{-1}$	$2.55 \cdot 10^{-3}$	$3.01 \cdot 10^{-4}$ (95)
2	176.72	176.72	1	rs12465451	0.136	$9.82 \cdot 10^{-4}$	$6.21 \cdot 10^{-1}$	$5.02 \cdot 10^{-3}$	$1.18 \cdot 10^{-4}$ (175)
2	241.24	241.24	1	rs2953145 (*RNPEPL1*)	0.211	$6.28 \cdot 10^{-3}$	1	$1.29 \cdot 10^{-5}$	$1.14 \cdot 10^{-4}$ (181)
3	42.38	42.38	1	rs33457	0.021	1	1	$1.68 \cdot 10^{-5}$	$1.12 \cdot 10^{-4}$ (182)
7	22.76	22.76	1	rs2286492	0.097	$1.19 \cdot 10^{-3}$	$1.88 \cdot 10^{-3}$	$5.25 \cdot 10^{-1}$	$2.58 \cdot 10^{-4}$ (107)
8	58.48	58.48	1	rs2875734	0.052	$7.14 \cdot 10^{-7}$	1	$9.84 \cdot 10^{-4}$	$2.87 \cdot 10^{-4}$ (98)
12	73.67	73.67	1	rs1526805	0.051	$6.61 \cdot 10^{-4}$	$6.36 \cdot 10^{-2}$	$3.58 \cdot 10^{-2}$	$2.24 \cdot 10^{-4}$ (118)
16	23.54	23.54	1	rs420259	0.269	$8.62 \cdot 10^{-5}$	$1.57 \cdot 10^{-2}$	$2.25 \cdot 10^{-4}$	$6.57 \cdot 10^{-4}$ (71)
23	110.32	110.32	1	rs975687 (*CAPN6*)	0.031	1	1	$2.67 \cdot 10^{-4}$	$3.33 \cdot 10^{-3}$ (17)

T-Trees

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	80.57	80.57	6	rs1896250	0.400	$9.76 \cdot 10^{-9}$	$9.37 \cdot 10^{-1}$	$3.36 \cdot 10^{-6}$	$1.95 \cdot 10^{-3}$ (29)
1	85.15	85.16	2	rs668860	0.472	$6.74 \cdot 10^{-3}$	$9.41 \cdot 10^{-1}$	$8.48 \cdot 10^{-1}$	$3.21 \cdot 10^{-4}$ (123)
1	118.80	118.80	2	rs12134541	0.480	1	$9.12 \cdot 10^{-1}$	$4.38 \cdot 10^{-1}$	$2.92 \cdot 10^{-4}$ (133)
1	219.15	219.16	2	rs1909194	0.154	$6.53 \cdot 10^{-1}$	$4.12 \cdot 10^{-2}$	$6.64 \cdot 10^{-1}$	$8.57 \cdot 10^{-4}$ (56)
2	25.31	25.36	3	rs2164411	0.154	$2.04 \cdot 10^{-10}$	$1.95 \cdot 10^{-1}$	$1.21 \cdot 10^{-3}$	$1.46 \cdot 10^{-3}$ (32)
2	132.99	132.99	2	rs2315380	0.394	$8.10 \cdot 10^{-1}$	$3.49 \cdot 10^{-1}$	$1.90 \cdot 10^{-1}$	$3.15 \cdot 10^{-4}$ (126)
3	121.76	121.83	4	rs804974 (*HGD*)	0.211	$6.13 \cdot 10^{-2}$	$8.68 \cdot 10^{-1}$	$8.17 \cdot 10^{-1}$	$2.39 \cdot 10^{-3}$ (24)
3	173.37	173.40	3	rs7653441 (*FNDC3B*)	0.018	$5.91 \cdot 10^{-86}$	1	$2.26 \cdot 10^{-2}$	$9.51 \cdot 10^{-3}$ (9)
4	17.85	17.90	10	rs1553460	0.313	$8.85 \cdot 10^{-94}$	$2.88 \cdot 10^{-5}$	$1.28 \cdot 10^{-28}$	$1.44 \cdot 10^{-2}$ (4)
4	23.67	23.67	2	rs615604	0.157	$4.33 \cdot 10^{-1}$	$8.35 \cdot 10^{-1}$	$9.31 \cdot 10^{-1}$	$2.30 \cdot 10^{-4}$ (147)
4	149.61	149.66	5	rs17484678 (*NR3C2*)	0.137	$5.68 \cdot 10^{-2}$	$4.07 \cdot 10^{-1}$	$1.59 \cdot 10^{-2}$	$8.18 \cdot 10^{-3}$ (10)
4	186.09	186.11	5	rs13126272 (*ACSL1*)	0.341	$2.30 \cdot 10^{-43}$	$3.02 \cdot 10^{-2}$	$1.27 \cdot 10^{-6}$	$3.21 \cdot 10^{-3}$ (22)
5	59.67	59.67	2	rs7733705 (*PDE4D*)	0.173	$1.77 \cdot 10^{-1}$	$2.33 \cdot 10^{-1}$	$6.40 \cdot 10^{-2}$	$1.28 \cdot 10^{-3}$ (39)
6	18.18	18.29	4	rs4072775	0.355	$9.18 \cdot 10^{-7}$	$2.71 \cdot 10^{-6}$	$8.44 \cdot 10^{-1}$	$1.08 \cdot 10^{-3}$ (44)
6	91.73	91.75	3	rs6903505	0.409	$3.41 \cdot 10^{-1}$	$7.59 \cdot 10^{-1}$	$2.25 \cdot 10^{-1}$	$9.00 \cdot 10^{-4}$ (52)
6	96.00	96.03	3	rs1319912	0.103	$2.89 \cdot 10^{-1}$	$3.13 \cdot 10^{-1}$	$5.96 \cdot 10^{-1}$	$3.21 \cdot 10^{-4}$ (121)
6	107.09	107.09	4	rs9320174 (*AIM1*)	0.336	$2.94 \cdot 10^{-1}$	$1.45 \cdot 10^{-1}$	$1.08 \cdot 10^{-4}$	$2.57 \cdot 10^{-4}$ (143)
6	148.51	148.51	3	rs9377114	0.444	$3.11 \cdot 10^{-2}$	$4.55 \cdot 10^{-1}$	$8.17 \cdot 10^{-1}$	$5.85 \cdot 10^{-4}$ (72)
6	150.77	150.77	2	rs9322256	0.377	$5.51 \cdot 10^{-1}$	$6.95 \cdot 10^{-1}$	$3.52 \cdot 10^{-1}$	$5.69 \cdot 10^{-4}$ (77)
6	153.40	153.42	3	rs2236014 (*MTRF1L*)	0.259	$5.12 \cdot 10^{-8}$	$1.76 \cdot 10^{-1}$	$7.17 \cdot 10^{-1}$	$5.91 \cdot 10^{-4}$ (71)
7	23.37	23.37	2	rs7781714	0.473	$4.31 \cdot 10^{-1}$	$5.54 \cdot 10^{-1}$	$6.30 \cdot 10^{-1}$	$2.39 \cdot 10^{-4}$ (146)
8	19.35	19.48	16	rs17480050 (*CSGALNACT1*)	0.119	0	1	0	$1.23 \cdot 10^{-1}$ (1)
8	55.10	55.23	6	rs11984645	0.145	$1.07 \cdot 10^{-7}$	1	$4.58 \cdot 10^{-2}$	$9.12 \cdot 10^{-4}$ (51)
8	82.41	82.41	2	rs1909935	0.021	1	$6.29 \cdot 10^{-1}$	$1.70 \cdot 10^{-1}$	$1.33 \cdot 10^{-2}$ (5)
8	120.42	120.42	3	rs2469997	0.189	$6.04 \cdot 10^{-1}$	$8.55 \cdot 10^{-1}$	$3.09 \cdot 10^{-1}$	$4.67 \cdot 10^{-4}$ (88)
10	18.53	18.54	2	rs12355606 (*CACNB2*)	0.006	1	1	$8.35 \cdot 10^{-12}$	$9.24 \cdot 10^{-4}$ (50)
10	43.64	43.64	2	rs7086449	0.126	$8.95 \cdot 10^{-6}$	1	$1.16 \cdot 10^{-4}$	$3.02 \cdot 10^{-4}$ (129)
10	59.23	59.24	2	rs1395043	0.084	$7.47 \cdot 10^{-1}$	$7.71 \cdot 10^{-2}$	$1.36 \cdot 10^{-1}$	$3.97 \cdot 10^{-4}$ (102)
10	77.12	77.15	2	rs7082404	0.004	1	1	$1.52 \cdot 10^{-6}$	$4.78 \cdot 10^{-4}$ (86)
11	113.31	113.31	3	rs1176741 (*HTR3B*)	0.029	$6.27 \cdot 10^{-1}$	$1.19 \cdot 10^{-1}$	$4.58 \cdot 10^{-2}$	$1.22 \cdot 10^{-2}$ (7)
12	37.33	37.38	2	rs826886 (*CPNE8*)	0.452	$1.75 \cdot 10^{-1}$	$8.23 \cdot 10^{-1}$	$6.74 \cdot 10^{-1}$	$1.60 \cdot 10^{-4}$ (177)
12	103.37	103.40	4	rs11112069 (*CHST11*)	0.199	$3.64 \cdot 10^{-22}$	$3.51 \cdot 10^{-1}$	$1.14 \cdot 10^{-7}$	$2.11 \cdot 10^{-3}$ (27)
12	127.06	127.06	3	rs6489228	0.479	$6.43 \cdot 10^{-1}$	$8.83 \cdot 10^{-1}$	$3.46 \cdot 10^{-1}$	$1.57 \cdot 10^{-3}$ (31)
14	43.99	44.06	4	rs435340	0.096	$2.56 \cdot 10^{-6}$	$5.54 \cdot 10^{-2}$	$5.40 \cdot 10^{-3}$	$9.66 \cdot 10^{-4}$ (47)
15	40.88	41.32	20	rs8027733 (*UBR1*)	0.118	$9.09 \cdot 10^{-1}$	$3.39 \cdot 10^{-1}$	$3.47 \cdot 10^{-1}$	$1.23 \cdot 10^{-2}$ (6)
16	79.58	79.70	12	rs1048194 (*CENPN*)	0.095	$2.22 \cdot 10^{-322}$	$9.95 \cdot 10^{-6}$	$3.66 \cdot 10^{-39}$	$6.60 \cdot 10^{-2}$ (2)
17	17.27	17.27	2	SNP_A-1948953	0.282	$9.93 \cdot 10^{-24}$	$5.18 \cdot 10^{-3}$	$3.97 \cdot 10^{-5}$	$1.33 \cdot 10^{-3}$ (36)
18	38.22	38.22	2	rs1442650 (*LOC284260*)	0.010	1	$5.24 \cdot 10^{-1}$	$2.32 \cdot 10^{-13}$	$7.52 \cdot 10^{-4}$ (61)
19	22.69	22.71	3	rs12980129	0.028	1	1	$5.93 \cdot 10^{-5}$	$4.63 \cdot 10^{-4}$ (90)
21	37.37	37.41	3	rs4816560 (*TTC3*)	0.468	$8.52 \cdot 10^{-1}$	$3.74 \cdot 10^{-1}$	$7.05 \cdot 10^{-1}$	$4.09 \cdot 10^{-4}$ (100)
21	39.35	39.37	4	rs999789	0.058	$4.18 \cdot 10^{-2}$	$3.49 \cdot 10^{-1}$	$8.70 \cdot 10^{-4}$	$5.85 \cdot 10^{-4}$ (73)
23	2.58	2.58	2	rs1419930	0.044	$1.81 \cdot 10^{-2}$	$5.71 \cdot 10^{-19}$	$3.33 \cdot 10^{-3}$	$4.54 \cdot 10^{-3}$ (19)
23	110.31	110.32	2	rs975687 (*CAPN6*)	0.031	1	1	$2.67 \cdot 10^{-4}$	$6.66 \cdot 10^{-4}$ (64)

TABLE 7.6 BD_{wtccc}: lists of regions identified by the Random Forests and the T-Trees methods.

Coronary artery disease

At Figures 7.4 and 7.5, we notice fewer rare variants in the 100 first variables in comparison with the bipolar disorder results. On CAD_{wtccc}, we also obtained markers deviating from HWE in the top ranked variables. It seems that, strong deviation from HWE are prefered over rare variants by the tree-based methods.

In Tables 7.7 and 7.8, we also see that none of the reported loci are selected except two different with the T-Trees, one in each dataset versions. On CAD_{qc}, `rs6475606` located in the strongest region reported in [Wel07] Supplementary Information. Although it is the reported as the strongest association for that disease, we found that a rare variant `rs3122348` is exhibiting a (really) stronger associated p-value and appears at position 1 in both ranking on CAD_{qc} (and corresponds to one of the excluded markers in the $wtccc$ version). Many markers are considered as important in that region which corresponds to `LOC645954` pseudogene. Also again, we notice the presence of `rs10212068` in the T-Trees rankings. It is also present in the Random Forests ranking but appeared alone. Additionally, on CAD_{wtccc}, we note that `rs3785579` (on chromosome 17 in `CACNG1` gene) is reported in [P+12a] as a "hub" SNP too.

We also notice the presence of many regions on chromosome X in the four experiments. Much more with the RF than the TT. We found no reported regions located on chromosome X except in [P+00] which report region Xq23-26 (which potentially ranges from 108.5 to 137.7Mb). It is also important to note that the proportion of men in the CAD dataset is higher (about 1500 males and 400 females) (which reflects the higher disease prevalence in males).

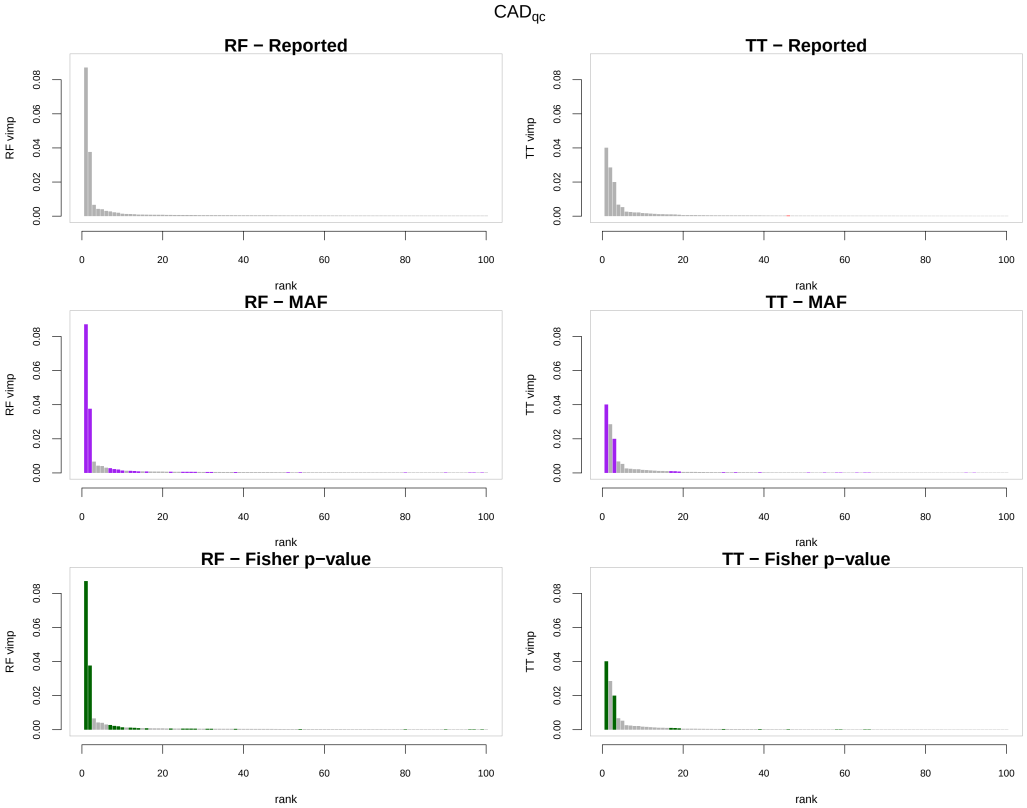


FIGURE 7.4 The first 100 variables according to the tree based importance rankings for CAD_{qc}. The horizontal axis corresponds to the ranks and the vertical axis to the variable importances. In the first column variables are ordered according to random forests variable importances and in the second column they are ordered according to the T-Trees variable importances. In the first row, red highlights the reported strongly associated regions. In the second row, purple corresponds to rare variants (MAF < 0.05). In the third row, green represents markers with a low Fisher p-value ($< 10^{-6}$).

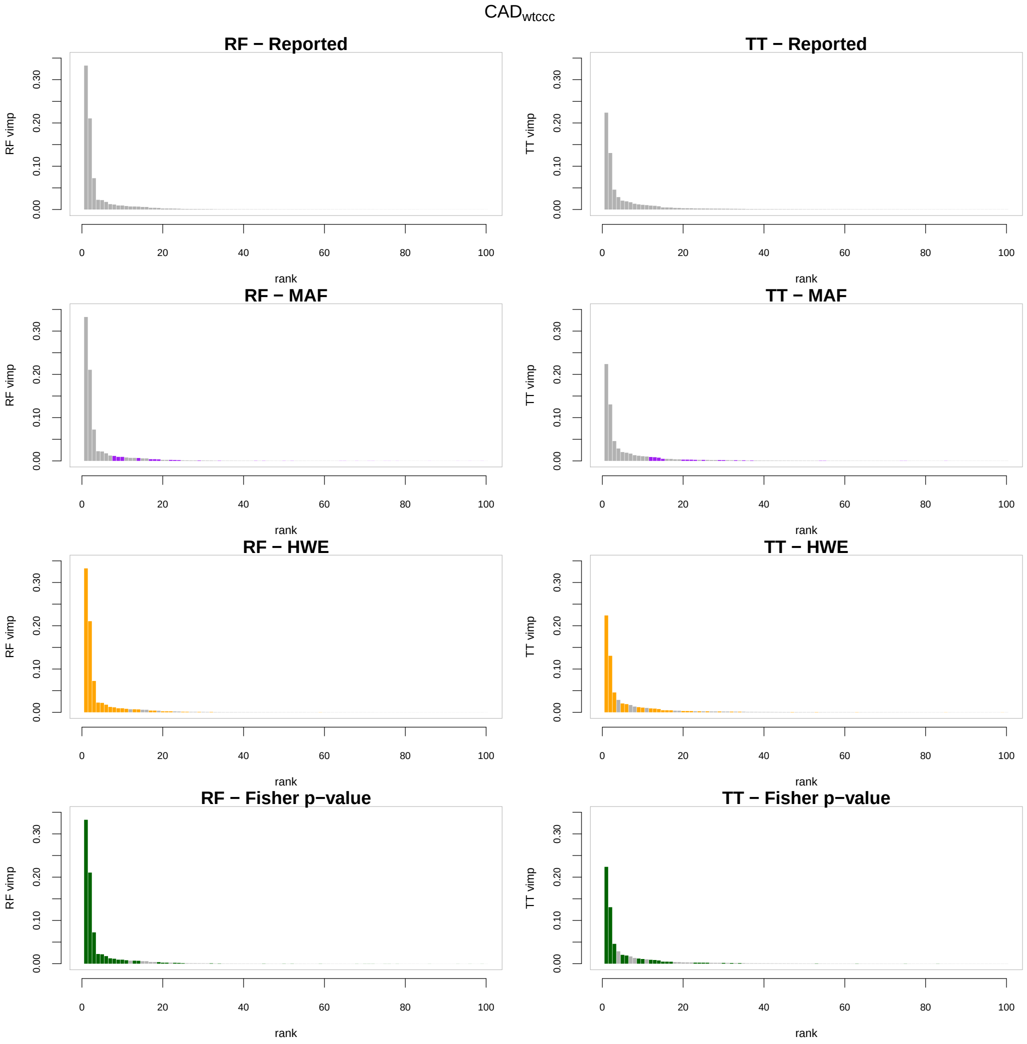


FIGURE 7.5 The first 100 variables according to the tree based importance rankings for CAD_{wtccc}. The horizontal axis corresponds to the ranks and the vertical axis to the variable importances. In the first column variables are ordered according to random forests variable importances and in the second column they are ordered according to the T-Trees variable importances. In the first row, red highlights the strongly associated reported regions. In the second row, purple corresponds to rare variants (MAF < 0.05). In the third row, orange highlights SNPs deviating from HWE and in the last row, green represents markers with a low Fisher p-value ($< 10^{-6}$).

Random Forests

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
10	30.99	31.06	6	rs3122348(*LOC645954*)$^{(1)}$	0.032	$2.85 \cdot 10^{-5}$	1	$5.57 \cdot 10^{-88}$	$8.72 \cdot 10^{-2}$ (1)
23	5.25	5.43	7	rs16983713	0.234	$5.21 \cdot 10^{-2}$	$8.83 \cdot 10^{-1}$	$6.16 \cdot 10^{-3}$	$3.82 \cdot 10^{-4}$ (57)
23	6.69	7.22	8	rs941172 (*HDHD1*)	0.244	$8.99 \cdot 10^{-1}$	$9.44 \cdot 10^{-1}$	$2.57 \cdot 10^{-2}$	$2.94 \cdot 10^{-4}$ (87)
23	8.82	8.83	2	rs17269753	0.224	$1.56 \cdot 10^{-1}$	$6.38 \cdot 10^{-1}$	$1.70 \cdot 10^{-2}$	$2.46 \cdot 10^{-4}$ (120)
23	11.37	11.39	2	rs5979435 (*ARHGAP6*)	0.189	$2.70 \cdot 10^{-2}$	$4.41 \cdot 10^{-1}$	$1.60 \cdot 10^{-2}$	$4.84 \cdot 10^{-4}$ (44)
23	14.72	14.85	6	rs5978704	0.492	$1.08 \cdot 10^{-1}$	$8.78 \cdot 10^{-1}$	$3.65 \cdot 10^{-3}$	$8.20 \cdot 10^{-4}$ (18)
23	15.42	15.43	2	SNP_A-1907375	0.182	$1.75 \cdot 10^{-1}$	$2.41 \cdot 10^{-1}$	$3.74 \cdot 10^{-2}$	$3.94 \cdot 10^{-4}$ (55)
23	21.73	21.76	5	rs12688591 (*SMS*)	0.296	$6.27 \cdot 10^{-1}$	$8.81 \cdot 10^{-2}$	$2.19 \cdot 10^{-1}$	$4.98 \cdot 10^{-4}$ (42)
23	21.92	22.02	4	rs12216932 (*PHEX*)	0.374	$2.41 \cdot 10^{-1}$	$8.86 \cdot 10^{-2}$	$1.91 \cdot 10^{-2}$	$5.85 \cdot 10^{-4}$ (30)
23	24.04	24.23	2	rs17312220	0.021	$2.15 \cdot 10^{-1}$	1	$7.42 \cdot 10^{-2}$	$4.01 \cdot 10^{-4}$ (51)
23	26.41	26.86	6	rs5944611	0.329	$1.42 \cdot 10^{-1}$	$2.15 \cdot 10^{-1}$	$3.92 \cdot 10^{-2}$	$6.83 \cdot 10^{-4}$ (24)
23	30.75	30.78	2	rs17315366	0.098	$7.88 \cdot 10^{-1}$	1	$2.88 \cdot 10^{-2}$	$3.64 \cdot 10^{-4}$ (63)
23	33.06	33.15	2	rs2057142	0.394	$3.03 \cdot 10^{-1}$	$3.96 \cdot 10^{-2}$	$3.44 \cdot 10^{-2}$	$3.27 \cdot 10^{-4}$ (73)
23	34.69	34.98	5	rs3128091	0.212	$7.58 \cdot 10^{-1}$	$5.82 \cdot 10^{-1}$	$3.68 \cdot 10^{-2}$	$3.04 \cdot 10^{-4}$ (81)
23	39.46	39.47	2	rs3002415	0.301	$5.59 \cdot 10^{-1}$	$9.51 \cdot 10^{-1}$	$7.63 \cdot 10^{-3}$	$8.41 \cdot 10^{-4}$ (17)
23	47.36	47.47	3	rs11091213 (*ZNF81*)	0.456	$2.70 \cdot 10^{-1}$	$8.70 \cdot 10^{-2}$	$7.73 \cdot 10^{-2}$	$2.11 \cdot 10^{-4}$ (148)
23	67.85	68.12	4	rs443731	0.423	$1.08 \cdot 10^{-1}$	$6.73 \cdot 10^{-1}$	$6.23 \cdot 10^{-3}$	$5.77 \cdot 10^{-4}$ (33)
23	68.97	68.98	5	rs5936814 (*EDA*)	0.388	$7.53 \cdot 10^{-1}$	$4.46 \cdot 10^{-1}$	$2.41 \cdot 10^{-3}$	$8.11 \cdot 10^{-4}$ (20)
23	90.18	90.20	5	rs2038452	0.229	$6.77 \cdot 10^{-1}$	$3.77 \cdot 10^{-1}$	$6.90 \cdot 10^{-2}$	$2.68 \cdot 10^{-4}$ (106)
23	95.09	95.14	2	rs2808726	0.124	$3.32 \cdot 10^{-1}$	$7.40 \cdot 10^{-1}$	$5.36 \cdot 10^{-1}$	$2.83 \cdot 10^{-4}$ (94)
23	97.74	97.85	3	rs5921205	0.287	$5.47 \cdot 10^{-1}$	$2.61 \cdot 10^{-1}$	$2.84 \cdot 10^{-1}$	$2.60 \cdot 10^{-4}$ (109)
23	99.55	99.71	2	rs12841456	0.390	$6.73 \cdot 10^{-1}$	$6.25 \cdot 10^{-1}$	$7.08 \cdot 10^{-2}$	$2.14 \cdot 10^{-4}$ (143)
23	116.77	116.78	2	rs1338512	0.292	$1.48 \cdot 10^{-1}$	$3.21 \cdot 10^{-1}$	$8.25 \cdot 10^{-2}$	$2.39 \cdot 10^{-4}$ (126)
23	117.52	117.56	2	rs5910392 (*DOCK11*)	0.145	$5.43 \cdot 10^{-1}$	$1.72 \cdot 10^{-1}$	$3.01 \cdot 10^{-2}$	$9.27 \cdot 10^{-4}$ (15)
23	124.24	124.26	2	rs3101156	0.404	$4.17 \cdot 10^{-1}$	$9.57 \cdot 10^{-1}$	$1.16 \cdot 10^{-3}$	$8.15 \cdot 10^{-4}$ (19)
23	125.62	125.93	2	rs204359	0.067	$8.71 \cdot 10^{-2}$	1	$1.85 \cdot 10^{-1}$	$3.07 \cdot 10^{-4}$ (79)
23	129.75	130.23	5	rs6529475	0.180	$1.31 \cdot 10^{-1}$	$1.18 \cdot 10^{-1}$	$1.11 \cdot 10^{-1}$	$1.29 \cdot 10^{-3}$ (11)
23	130.80	131.03	5	rs5933109	0.248	$4.65 \cdot 10^{-1}$	$3.11 \cdot 10^{-1}$	$4.48 \cdot 10^{-2}$	$4.73 \cdot 10^{-4}$ (46)
23	135.09	135.14	2	SNP_A-2065339	0.283	$4.09 \cdot 10^{-1}$	1	$2.26 \cdot 10^{-2}$	$2.45 \cdot 10^{-4}$ (122)
23	136.57	136.64	2	rs1342044	0.221	1	$4.96 \cdot 10^{-1}$	$7.95 \cdot 10^{-2}$	$4.82 \cdot 10^{-4}$ (45)
23	144.27	144.28	2	rs5965859	0.441	$4.44 \cdot 10^{-3}$	$2.50 \cdot 10^{-1}$	$1.37 \cdot 10^{-1}$	$3.58 \cdot 10^{-4}$ (64)
23	146.37	146.39	2	rs1076616	0.090	$4.08 \cdot 10^{-1}$	$8.83 \cdot 10^{-1}$	$7.90 \cdot 10^{-1}$	$3.74 \cdot 10^{-4}$ (60)
23	147.39	147.47	3	rs241127 (*AFF2*)	0.217	$5.54 \cdot 10^{-1}$	$5.93 \cdot 10^{-1}$	$5.91 \cdot 10^{-2}$	$2.59 \cdot 10^{-4}$ (110)
23	149.86	149.86	2	rs12845940	0.393	$7.51 \cdot 10^{-1}$	$5.13 \cdot 10^{-1}$	$2.02 \cdot 10^{-2}$	$3.92 \cdot 10^{-4}$ (56)
23	151.10	151.12	2	SNP_A-1999621	0.105	$5.40 \cdot 10^{-2}$	$4.45 \cdot 10^{-2}$	$1.79 \cdot 10^{-2}$	$3.98 \cdot 10^{-4}$ (53)

T-Trees

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	52.95	53.03	5	rs505444(*ZYG11B*)$^{(3)}$	0.423	$1.16 \cdot 10^{-1}$	$3.72 \cdot 10^{-4}$	$5.18 \cdot 10^{-2}$	$6.01 \cdot 10^{-4}$ (21)
1	185.16	185.18	4	rs6683655	0.436	$1.53 \cdot 10^{-1}$	$1.79 \cdot 10^{-1}$	$7.24 \cdot 10^{-1}$	$1.08 \cdot 10^{-3}$ (16)
2	5.37	5.37	2	rs1453783$^{(3)}$	0.373	$6.62 \cdot 10^{-1}$	$1.62 \cdot 10^{-2}$	$5.63 \cdot 10^{-1}$	$2.24 \cdot 10^{-4}$ (76)
3	97.61	97.67	8	rs326296	0.446	$1.04 \cdot 10^{-6}$	$6.68 \cdot 10^{-2}$	$2.05 \cdot 10^{-5}$	$6.75 \cdot 10^{-3}$ (4)
3	163.64	163.67	4	rs7434223	0.416	$6.37 \cdot 10^{-1}$	$5.73 \cdot 10^{-1}$	$2.94 \cdot 10^{-2}$	$5.46 \cdot 10^{-4}$ (24)
4	55.92	55.93	2	rs10014689	0.021	$4.98 \cdot 10^{-1}$	$6.49 \cdot 10^{-1}$	$3.45 \cdot 10^{-1}$	$1.46 \cdot 10^{-4}$ (154)
4	117.36	117.39	2	rs814376	0.392	$5.05 \cdot 10^{-1}$	$3.56 \cdot 10^{-3}$	$7.18 \cdot 10^{-1}$	$2.12 \cdot 10^{-4}$ (81)
6	32.87	32.88	6	rs2621384	0.384	$4.11 \cdot 10^{-1}$	$9.69 \cdot 10^{-1}$	$2.99 \cdot 10^{-1}$	$2.46 \cdot 10^{-3}$ (7)
7	37.11	37.12	2	rs2717992 (*ELMO1*)	0.446	$1.55 \cdot 10^{-1}$	$1.10 \cdot 10^{-1}$	$3.61 \cdot 10^{-1}$	$2.23 \cdot 10^{-4}$ (77)
7	50.31	50.31	3	rs11575535(*DDC*)$^{(1)}$	0.019	$8.14 \cdot 10^{-2}$	1	$1.08 \cdot 10^{-1}$	$4.21 \cdot 10^{-4}$ (33)
8	87.57	87.58	2	rs7460439 (*FAM82B*)	0.448	$8.54 \cdot 10^{-1}$	$7.65 \cdot 10^{-1}$	$1.76 \cdot 10^{-1}$	$2.31 \cdot 10^{-4}$ (73)
9	72.35	72.36	2	rs10121866	0.456	$9.27 \cdot 10^{-1}$	1	$1.92 \cdot 10^{-1}$	$1.77 \cdot 10^{-4}$ (112)
9	107.23	107.23	2	rs12343115	0.006	1	1	$1.08 \cdot 10^{-10}$	$2.43 \cdot 10^{-4}$ (66)
10	30.95	31.06	12	rs3122348(*LOC645954*)$^{(1)}$	0.032	$2.85 \cdot 10^{-5}$	1	$5.57 \cdot 10^{-88}$	$4.02 \cdot 10^{-2}$ (1)
10	113.20	113.22	2	rs1914139	0.427	$3.51 \cdot 10^{-1}$	$1.23 \cdot 10^{-1}$	$1.55 \cdot 10^{-1}$	$4.55 \cdot 10^{-4}$ (29)
10	125.23	125.25	4	rs913525	0.450	$8.54 \cdot 10^{-1}$	$5.79 \cdot 10^{-1}$	$3.19 \cdot 10^{-1}$	$4.35 \cdot 10^{-4}$ (31)
11	11.10	11.11	3	rs7103691	0.487	$6.49 \cdot 10^{-1}$	$3.59 \cdot 10^{-1}$	1	$6.06 \cdot 10^{-4}$ (20)
11	12.48	12.48	2	rs7126366 (*PARVA*)	0.004	1	1	$4.58 \cdot 10^{-7}$	$2.54 \cdot 10^{-4}$ (59)
14	103.34	103.36	3	rs17791722 (*PPP1R13B*)	0.332	$5.40 \cdot 10^{-1}$	$6.19 \cdot 10^{-1}$	$7.26 \cdot 10^{-1}$	$2.34 \cdot 10^{-4}$ (71)
22	35.93	35.99	8	rs10212068$^{(3)}$	0.028	1	$2.64 \cdot 10^{-3}$	$1.06 \cdot 10^{-57}$	$2.00 \cdot 10^{-2}$ (3)
22	44.07	44.07	3	rs5764698 (*SMC1B*)	0.458	$8.91 \cdot 10^{-1}$	$4.17 \cdot 10^{-1}$	$2.91 \cdot 10^{-1}$	$4.23 \cdot 10^{-4}$ (32)
23	2.70	2.96	3	rs2124012 (*ARSF*)	0.362	$9.12 \cdot 10^{-1}$	$1.71 \cdot 10^{-1}$	$1.58 \cdot 10^{-1}$	$2.10 \cdot 10^{-4}$ (83)
23	3.81	3.81	2	rs5916413	0.470	$1.08 \cdot 10^{-1}$	$1.80 \cdot 10^{-1}$	$4.90 \cdot 10^{-1}$	$2.72 \cdot 10^{-4}$ (52)
23	4.38	4.50	2	rs5916649	0.492	$1.33 \cdot 10^{-1}$	$3.28 \cdot 10^{-1}$	$5.57 \cdot 10^{-1}$	$1.32 \cdot 10^{-4}$ (187)
23	13.43	13.43	2	rs4830882 (*TCEANC*)	0.473	$1.32 \cdot 10^{-1}$	$6.43 \cdot 10^{-1}$	$7.02 \cdot 10^{-1}$	$1.39 \cdot 10^{-4}$ (172)
23	14.72	14.78	3	rs4240155	0.494	$5.74 \cdot 10^{-2}$	$8.37 \cdot 10^{-1}$	$6.89 \cdot 10^{-3}$	$2.62 \cdot 10^{-4}$ (57)
23	26.83	26.90	3	rs1898744	0.483	$7.19 \cdot 10^{-2}$	$5.10 \cdot 10^{-2}$	$9.19 \cdot 10^{-1}$	$3.38 \cdot 10^{-4}$ (41)
23	33.63	33.80	4	rs1948804	0.388	$2.14 \cdot 10^{-1}$	$1.80 \cdot 10^{-3}$	$9.58 \cdot 10^{-1}$	$2.65 \cdot 10^{-4}$ (53)
23	35.42	35.46	3	rs5928946	0.458	$6.16 \cdot 10^{-1}$	$1.13 \cdot 10^{-2}$	$7.01 \cdot 10^{-1}$	$2.25 \cdot 10^{-4}$ (75)
23	45.93	45.93	2	rs851234	0.425	$2.62 \cdot 10^{-1}$	$1.12 \cdot 10^{-1}$	$5.19 \cdot 10^{-1}$	$1.97 \cdot 10^{-4}$ (89)
23	87.94	89.23	5	rs4545257	0.397	$8.11 \cdot 10^{-2}$	$4.81 \cdot 10^{-2}$	$3.90 \cdot 10^{-1}$	$1.92 \cdot 10^{-4}$ (93)
23	97.64	97.87	6	rs2498864	0.454	$5.38 \cdot 10^{-2}$	$1.08 \cdot 10^{-1}$	$8.98 \cdot 10^{-1}$	$1.90 \cdot 10^{-4}$ (98)
23	100.06	100.09	2	rs12557159	0.416	$4.30 \cdot 10^{-3}$	$2.64 \cdot 10^{-1}$	$5.01 \cdot 10^{-1}$	$3.04 \cdot 10^{-4}$ (45)
23	111.98	111.98	2	rs4829520	0.444	$1.81 \cdot 10^{-1}$	$5.69 \cdot 10^{-1}$	$3.76 \cdot 10^{-2}$	$2.01 \cdot 10^{-4}$ (86)
23	115.40	115.74	4	rs6645482	0.364	$6.86 \cdot 10^{-3}$	$2.00 \cdot 10^{-1}$	$8.74 \cdot 10^{-1}$	$2.34 \cdot 10^{-4}$ (72)
23	117.93	117.93	3	rs2278954 (*LONRF3*)	0.362	$2.58 \cdot 10^{-3}$	$4.00 \cdot 10^{-2}$	$3.68 \cdot 10^{-1}$	$3.92 \cdot 10^{-4}$ (38)
23	118.71	118.79	2	rs1858934 (*NDUFA1, RNF113A*)	0.498	$5.03 \cdot 10^{-3}$	1	1	$1.56 \cdot 10^{-4}$ (138)
23	139.60	139.61	2	rs4824960	0.457	$1.03 \cdot 10^{-1}$	1	$1.25 \cdot 10^{-1}$	$1.71 \cdot 10^{-4}$ (120)
23	144.27	144.41	6	rs5965859	0.441	$4.44 \cdot 10^{-3}$	$2.50 \cdot 10^{-1}$	$1.37 \cdot 10^{-1}$	$3.15 \cdot 10^{-4}$ (42)
23	145.03	145.06	2	rs6525652	0.423	$1.55 \cdot 10^{-1}$	$1.42 \cdot 10^{-1}$	$4.09 \cdot 10^{-1}$	$1.51 \cdot 10^{-4}$ (144)
9	22.07	22.07	1	rs6475606	0.492	$7.48 \cdot 10^{-1}$	$8.44 \cdot 10^{-2}$	$2.22 \cdot 10^{-14}$	$3.00 \cdot 10^{-4}$ (46)

TABLE 7.7 CAD_{qc}: lists of regions identified by the Random Forests and the T-Trees methods.

Random Forests

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
4	17.85	17.88	6	rs1553460	0.313	$2.34 \cdot 10^{-101}$	$2.88 \cdot 10^{-5}$	$1.78 \cdot 10^{-27}$	$6.99 \cdot 10^{-3}$ (13)
4	139.61	139.63	3	rs890447 (*LINC00499*)	0.052	$2.70 \cdot 10^{-205}$	$2.22 \cdot 10^{-2}$	$2.64 \cdot 10^{-12}$	$1.24 \cdot 10^{-2}$ (7)
5	117.03	117.03	2	rs2416472	0.320	$1.49 \cdot 10^{-24}$	$9.35 \cdot 10^{-1}$	$4.36 \cdot 10^{-10}$	$3.02 \cdot 10^{-4}$ (59)
8	19.40	19.48	8	rs17480050 (*CSGALNACT1*)	0.118	0	1	0	$7.24 \cdot 10^{-2}$ (3)
8	95.06	95.06	2	rs6989092	0.017	1	$7.27 \cdot 10^{-1}$	$3.60 \cdot 10^{-35}$	$3.61 \cdot 10^{-3}$ (19)
8	139.11	139.11	3	rs16908145 (*FLJ45872*)	0.014	$2.71 \cdot 10^{-3}$	$4.02 \cdot 10^{-1}$	$2.70 \cdot 10^{-21}$	$2.13 \cdot 10^{-3}$ (23)
10	118.22	118.28	16	rs7906587 (*PNLIPRP3*)	0.078	$1.04 \cdot 10^{-285}$	$4.33 \cdot 10^{-3}$	$1.00 \cdot 10^{-28}$	$2.17 \cdot 10^{-2}$ (5)
11	129.30	129.32	2	rs2241572 (*PRDM10*)	0.344	$1.84 \cdot 10^{-313}$	1	0	$3.33 \cdot 10^{-1}$ (1)
16	11.28	11.31	5	rs11640295	0.284	$2.71 \cdot 10^{-40}$	$1.45 \cdot 10^{-4}$	$5.07 \cdot 10^{-9}$	$8.96 \cdot 10^{-4}$ (32)
17	62.46	62.65	9	rs3785579 (*CACNG1*)	0.277	0	$2.12 \cdot 10^{-2}$	0	$2.11 \cdot 10^{-1}$ (2)
18	25.70	25.71	2	rs1595963	0.039	$1.62 \cdot 10^{-184}$	$2.72 \cdot 10^{-1}$	$2.07 \cdot 10^{-14}$	$9.27 \cdot 10^{-3}$ (10)
18	33.34	33.36	5	rs4799934 (*CELF4*)	0.078	$4.83 \cdot 10^{-288}$	$2.83 \cdot 10^{-3}$	$1.90 \cdot 10^{-26}$	$2.24 \cdot 10^{-2}$ (4)
19	19.15	19.21	3	rs11671119 (*MEF2BNB-MEF2B*)	0.045	$2.63 \cdot 10^{-201}$	$1.12 \cdot 10^{-1}$	$9.27 \cdot 10^{-18}$	$1.15 \cdot 10^{-2}$ (8)
22	16.89	16.89	2	rs4819660 (*FLJ41941*)	0.233	$8.46 \cdot 10^{-78}$	$6.85 \cdot 10^{-7}$	$5.10 \cdot 10^{-9}$	$2.32 \cdot 10^{-3}$ (21)
23	2.68	2.75	2	rs311152	0.355	$9.20 \cdot 10^{-1}$	$5.99 \cdot 10^{-1}$	$1.13 \cdot 10^{-14}$	$2.25 \cdot 10^{-4}$ (70)
23	5.25	5.36	4	rs4826780	0.430	$6.80 \cdot 10^{-1}$	$4.29 \cdot 10^{-1}$	$1.59 \cdot 10^{-2}$	$9.26 \cdot 10^{-5}$ (127)
23	6.69	7.14	7	rs6530079	0.272	$5.48 \cdot 10^{-1}$	$4.36 \cdot 10^{-1}$	$1.51 \cdot 10^{-3}$	$9.80 \cdot 10^{-5}$ (121)
23	11.37	11.39	2	rs5979435 (*ARHGAP6*)	0.188	$1.79 \cdot 10^{-2}$	$2.97 \cdot 10^{-1}$	$1.40 \cdot 10^{-2}$	$9.90 \cdot 10^{-5}$ (116)
23	14.73	14.78	4	rs4240155	0.494	$6.98 \cdot 10^{-2}$	$7.17 \cdot 10^{-1}$	$8.32 \cdot 10^{-3}$	$1.42 \cdot 10^{-4}$ (94)
23	15.42	15.43	2	SNP_A-1907375	0.182	$1.69 \cdot 10^{-1}$	$1.71 \cdot 10^{-1}$	$2.83 \cdot 10^{-2}$	$1.36 \cdot 10^{-4}$ (98)
23	21.73	21.77	5	rs12688591 (*SMS*)	0.295	$6.27 \cdot 10^{-1}$	$4.84 \cdot 10^{-2}$	$1.68 \cdot 10^{-1}$	$1.70 \cdot 10^{-4}$ (84)
23	21.92	22.02	3	rs12216932 (*PHEX*)	0.375	$2.40 \cdot 10^{-1}$	$7.61 \cdot 10^{-2}$	$1.60 \cdot 10^{-2}$	$1.61 \cdot 10^{-4}$ (88)
23	22.49	22.49	2	rs3935727 (*LOC100873065*)	0.437	$4.74 \cdot 10^{-1}$	$1.01 \cdot 10^{-1}$	$4.40 \cdot 10^{-2}$	$1.18 \cdot 10^{-4}$ (106)
23	26.58	26.75	3	rs5944611	0.330	$1.72 \cdot 10^{-1}$	$2.35 \cdot 10^{-1}$	$3.59 \cdot 10^{-2}$	$1.74 \cdot 10^{-4}$ (83)
23	30.75	30.78	2	rs17315366	0.099	$7.89 \cdot 10^{-1}$	1	$4.86 \cdot 10^{-2}$	$8.66 \cdot 10^{-5}$ (140)
23	39.46	39.47	2	rs3002415	0.303	$5.58 \cdot 10^{-1}$	$7.57 \cdot 10^{-1}$	$1.11 \cdot 10^{-2}$	$2.26 \cdot 10^{-4}$ (69)
23	47.39	47.47	4	rs5953113 (*ZNF81*)	0.456	$3.17 \cdot 10^{-1}$	$1.17 \cdot 10^{-1}$	$4.47 \cdot 10^{-2}$	$9.19 \cdot 10^{-5}$ (130)
23	67.85	68.29	5	rs241388	0.417	$6.64 \cdot 10^{-2}$	$4.53 \cdot 10^{-1}$	$7.78 \cdot 10^{-3}$	$1.81 \cdot 10^{-4}$ (81)
23	68.97	68.98	5	rs5936814 (*EDA*)	0.387	$8.33 \cdot 10^{-1}$	$3.23 \cdot 10^{-1}$	$1.46 \cdot 10^{-3}$	$1.67 \cdot 10^{-4}$ (85)
23	95.09	95.14	2	rs2808726	0.123	$3.32 \cdot 10^{-1}$	$5.72 \cdot 10^{-1}$	$5.07 \cdot 10^{-1}$	$9.01 \cdot 10^{-5}$ (135)
23	116.77	116.78	2	rs1338512	0.292	$1.48 \cdot 10^{-1}$	$1.88 \cdot 10^{-1}$	$6.31 \cdot 10^{-2}$	$1.12 \cdot 10^{-4}$ (110)
23	117.52	117.56	2	rs5910392 (*DOCK11*)	0.146	$5.44 \cdot 10^{-1}$	$1.71 \cdot 10^{-1}$	$3.22 \cdot 10^{-2}$	$2.43 \cdot 10^{-4}$ (66)
23	124.24	124.26	2	rs3101156	0.404	$3.61 \cdot 10^{-1}$	1	$1.57 \cdot 10^{-3}$	$2.05 \cdot 10^{-4}$ (73)
23	125.62	125.64	2	rs7473865	0.048	$6.39 \cdot 10^{-3}$	$4.56 \cdot 10^{-1}$	$7.43 \cdot 10^{-10}$	$4.02 \cdot 10^{-4}$ (45)
23	129.75	130.19	4	rs6529475	0.179	$1.28 \cdot 10^{-1}$	$1.57 \cdot 10^{-1}$	$8.86 \cdot 10^{-2}$	$3.54 \cdot 10^{-4}$ (49)
23	130.80	131.03	3	rs5933109	0.249	$4.66 \cdot 10^{-1}$	$2.48 \cdot 10^{-1}$	$7.59 \cdot 10^{-2}$	$1.07 \cdot 10^{-4}$ (114)
23	144.27	144.28	2	rs5965859	0.440	$2.24 \cdot 10^{-3}$	$2.06 \cdot 10^{-1}$	$1.42 \cdot 10^{-1}$	$6.61 \cdot 10^{-5}$ (177)
23	146.37	146.39	2	rs1076616	0.089	$4.02 \cdot 10^{-1}$	$7.64 \cdot 10^{-1}$	$8.22 \cdot 10^{-1}$	$7.50 \cdot 10^{-5}$ (155)
23	149.86	149.86	2	rs12845940	0.393	$8.32 \cdot 10^{-1}$	$5.09 \cdot 10^{-1}$	$1.70 \cdot 10^{-2}$	$8.90 \cdot 10^{-5}$ (137)
23	151.10	151.12	2	SNP_A-1999621	0.106	$5.53 \cdot 10^{-2}$	$4.44 \cdot 10^{-2}$	$1.59 \cdot 10^{-2}$	$1.19 \cdot 10^{-4}$ (104)

T-Trees

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	86.95	86.98	3	rs1208054	0.217	$7.39 \cdot 10^{-1}$	$5.10 \cdot 10^{-1}$	$6.14 \cdot 10^{-1}$	$8.44 \cdot 10^{-4}$ (42)
2	25.31	25.36	3	rs2164411	0.156	$6.97 \cdot 10^{-20}$	$1.95 \cdot 10^{-1}$	$1.20 \cdot 10^{-2}$	$6.95 \cdot 10^{-4}$ (47)
3	53.24	53.27	6	rs7628245 (*TKT*)	0.178	$1.18 \cdot 10^{-72}$	$6.43 \cdot 10^{-5}$	$2.96 \cdot 10^{-8}$	$4.72 \cdot 10^{-3}$ (16)
3	97.63	97.67	4	SNP_A-4302324	0.082	$5.05 \cdot 10^{-1}$	1	$2.07 \cdot 10^{-2}$	$3.62 \cdot 10^{-3}$ (19)
3	121.83	121.83	2	rs804974 (*HGD*)	0.208	$5.73 \cdot 10^{-1}$	$8.68 \cdot 10^{-1}$	$4.13 \cdot 10^{-1}$	$4.23 \cdot 10^{-4}$ (67)
3	173.37	173.37	2	rs7653441 (*FNDC3B*)	0.017	$2.56 \cdot 10^{-76}$	1	$3.30 \cdot 10^{-1}$	$1.95 \cdot 10^{-3}$ (29)
4	17.85	17.90	9	rs1553460	0.313	$2.34 \cdot 10^{-101}$	$2.88 \cdot 10^{-5}$	$1.78 \cdot 10^{-27}$	$4.56 \cdot 10^{-3}$ (17)
4	32.67	32.73	7	rs10022638	0.301	$4.16 \cdot 10^{-1}$	$4.30 \cdot 10^{-1}$	$9.28 \cdot 10^{-1}$	$7.61 \cdot 10^{-4}$ (44)
4	73.28	73.30	2	rs9884478 (*NPFFR2*)	0.013	$6.78 \cdot 10^{-23}$	1	$5.88 \cdot 10^{-2}$	$2.86 \cdot 10^{-4}$ (85)
4	139.61	139.63	4	rs890447 (*LINC00499*)	0.052	$2.70 \cdot 10^{-205}$	$2.22 \cdot 10^{-2}$	$2.64 \cdot 10^{-12}$	$1.19 \cdot 10^{-2}$ (9)
4	186.09	186.11	5	rs13126272 (*ACSL1*)	0.346	$2.61 \cdot 10^{-68}$	$3.02 \cdot 10^{-2}$	$1.92 \cdot 10^{-9}$	$1.41 \cdot 10^{-3}$ (34)
5	117.02	117.06	8	rs17411921	0.344	$6.88 \cdot 10^{-1}$	$3.66 \cdot 10^{-1}$	$3.82 \cdot 10^{-1}$	$3.71 \cdot 10^{-3}$ (18)
5	172.87	172.90	4	rs17076079	0.097	$6.14 \cdot 10^{-146}$	$7.53 \cdot 10^{-7}$	$4.69 \cdot 10^{-17}$	$1.09 \cdot 10^{-2}$ (10)
6	18.18	18.19	3	rs4072775	0.350	$1.04 \cdot 10^{-5}$	$2.71 \cdot 10^{-6}$	$3.27 \cdot 10^{-1}$	$2.32 \cdot 10^{-4}$ (99)
8	19.40	19.48	12	rs17480050 (*CSGALNACT1*)	0.118	0	1	0	$4.59 \cdot 10^{-2}$ (3)
8	34.82	34.93	2	rs16883114	0.027	$4.29 \cdot 10^{-112}$	$6.37 \cdot 10^{-1}$	$1.66 \cdot 10^{-5}$	$2.93 \cdot 10^{-3}$ (21)
8	95.06	95.06	3	rs6989092	0.017	1	$7.27 \cdot 10^{-1}$	$3.60 \cdot 10^{-35}$	$2.33 \cdot 10^{-3}$ (25)
8	119.71	119.74	2	rs16891338 (*SAMD12-AS1*)	0.024	$4.52 \cdot 10^{-94}$	$4.02 \cdot 10^{-1}$	$2.17 \cdot 10^{-1}$	$2.97 \cdot 10^{-3}$ (20)
8	139.11	139.11	3	rs16908145 (*FLJ45872*)	0.014	$2.71 \cdot 10^{-3}$	$4.02 \cdot 10^{-1}$	$2.70 \cdot 10^{-21}$	$2.44 \cdot 10^{-3}$ (23)
10	118.18	118.27	6	rs7906587 (*PNLIPRP3*)	0.078	$1.04 \cdot 10^{-285}$	$4.33 \cdot 10^{-3}$	$1.00 \cdot 10^{-28}$	$1.89 \cdot 10^{-2}$ (6)
11	14.14	14.16	2	rs10832215	0.427	$9.63 \cdot 10^{-1}$	$8.50 \cdot 10^{-1}$	$3.45 \cdot 10^{-1}$	$1.28 \cdot 10^{-4}$ (159)
11	129.27	129.35	7	rs2241572 (*PRDM10*)	0.344	$1.84 \cdot 10^{-313}$	1	0	$2.24 \cdot 10^{-1}$ (1)
12	94.31	94.32	3	rs2769432	0.116	$4.37 \cdot 10^{-1}$	$5.24 \cdot 10^{-1}$	$7.69 \cdot 10^{-1}$	$2.98 \cdot 10^{-4}$ (82)
13	98.34	98.35	3	rs12430163 (*DOCK9*)	0.121	$9.15 \cdot 10^{-1}$	$5.40 \cdot 10^{-1}$	$7.99 \cdot 10^{-1}$	$7.78 \cdot 10^{-4}$ (43)
16	11.29	11.30	5	rs11640295	0.284	$2.71 \cdot 10^{-40}$	$1.45 \cdot 10^{-4}$	$5.07 \cdot 10^{-9}$	$2.39 \cdot 10^{-3}$ (24)
16	79.96	79.98	4	rs16955238	0.027	$8.76 \cdot 10^{-124}$	$6.46 \cdot 10^{-1}$	$1.75 \cdot 10^{-5}$	$2.84 \cdot 10^{-3}$ (22)
17	4.99	5.05	2	rs2641263 (*LOC100130950, SCIMP*)	0.357	$8.01 \cdot 10^{-1}$	$9.36 \cdot 10^{-1}$	$9.54 \cdot 10^{-2}$	$3.25 \cdot 10^{-4}$ (76)
17	62.45	62.65	12	rs3785579 (*CACNG1*)	0.277	0	$2.12 \cdot 10^{-2}$	0	$1.31 \cdot 10^{-1}$ (2)
18	25.70	25.71	2	rs1595963	0.039	$1.62 \cdot 10^{-184}$	$2.72 \cdot 10^{-1}$	$2.07 \cdot 10^{-14}$	$8.53 \cdot 10^{-3}$ (13)
18	33.34	33.37	6	rs4799934 (*CELF4*)	0.078	$4.83 \cdot 10^{-288}$	$2.83 \cdot 10^{-3}$	$1.90 \cdot 10^{-26}$	$2.04 \cdot 10^{-2}$ (5)
19	19.15	19.21	3	rs11671119 (*MEF2BNB-MEF2B*)	0.045	$2.63 \cdot 10^{-201}$	$1.12 \cdot 10^{-1}$	$9.27 \cdot 10^{-18}$	$9.01 \cdot 10^{-3}$ (12)
22	16.89	16.89	3	rs4819660 (*FLJ41941*)	0.233	$8.46 \cdot 10^{-78}$	$6.85 \cdot 10^{-7}$	$5.10 \cdot 10^{-9}$	$1.58 \cdot 10^{-3}$ (32)
23	1.80	1.81	2	rs6588810	0.016	$5.06 \cdot 10^{-3}$	$2.22 \cdot 10^{-5}$	$2.33 \cdot 10^{-8}$	$1.84 \cdot 10^{-3}$ (30)
23	14.73	14.78	2	rs6527137	0.496	$6.98 \cdot 10^{-2}$	$7.56 \cdot 10^{-1}$	$1.12 \cdot 10^{-2}$	$1.32 \cdot 10^{-4}$ (152)
23	33.63	33.71	2	rs11095312	0.423	$1.03 \cdot 10^{-1}$	$6.87 \cdot 10^{-4}$	$6.97 \cdot 10^{-1}$	$1.78 \cdot 10^{-4}$ (115)
23	35.42	35.46	4	rs11095405	0.462	$1.84 \cdot 10^{-1}$	$2.38 \cdot 10^{-1}$	$4.99 \cdot 10^{-1}$	$1.63 \cdot 10^{-4}$ (123)
23	97.74	97.78	2	rs241863	0.483	$6.87 \cdot 10^{-2}$	$5.27 \cdot 10^{-2}$	$2.92 \cdot 10^{-1}$	$1.16 \cdot 10^{-4}$ (173)
23	100.06	100.09	2	rs7050888 (*TRMT2B*)	0.474	$2.23 \cdot 10^{-1}$	$2.98 \cdot 10^{-1}$	$1.37 \cdot 10^{-1}$	$1.54 \cdot 10^{-4}$ (127)
23	117.93	117.93	3	rs2290514 (*LONRF3*)	0.361	$3.27 \cdot 10^{-3}$	$5.63 \cdot 10^{-2}$	$3.64 \cdot 10^{-1}$	$1.49 \cdot 10^{-4}$ (133)
23	144.27	144.28	2	rs5919854	0.440	$6.98 \cdot 10^{-4}$	$3.68 \cdot 10^{-1}$	$1.55 \cdot 10^{-1}$	$2.82 \cdot 10^{-4}$ (88)
10	85.11	85.11	1	rs11198290	0.061	$9.60 \cdot 10^{-4}$	$1.47 \cdot 10^{-5}$	$8.62 \cdot 10^{-1}$	$1.41 \cdot 10^{-4}$ (137)

TABLE 7.8 CAD_{wtccc}: lists of regions identified by the Random Forests and the T-Trees methods.

Crohn's disease

Again, we observed more rare variants in the first variables on the qc versions, while the markers deviating from HWE prevail over recent variants on the $wtccc$.

For Crohn's disease, as in the previous Chapter, we also included the 140 reported loci from [J$^+$13] in Tables 7.10, 7.11. As a reminder, these regions correspond to underlined marker ids and (*) corresponds to the nine reported regions while the coloured marker ids indicate regions being reported in [Wel07] Supplementary Information. Additionally, blue highlights important regions according to the T-Trees. To facilitate the comparison we also add a copy of the Chapter 6 CD_{ibd} table (Table 7.9).

A first comparison between Table 7.10 and Table 7.9 there is almost no difference between the two datasets. With the Random Forests, the same six loci were identified. But, most importantly, we notice the disappearance of the 35 markers located around rs11887827. Similarly, with the T-Trees, the main difference is characterised by the disappearance of the two "blue" loci and the presence of the "pathologic" rs10212068. That marker also appear as the most important one in the Random Forests variable ranking on the CD_{qc} dataset but it is isolated and thus not reported in the upper part of the Table. Firstly, that marker on chromosome 22 was not found in the CD_{ibd} version although none of our filters discarded it in the CD_{qc} alternative. Secondly, in the two "blue" regions 2p12 and 7q31, we noted the disappearance of rs11887827 and rs2107062 (although not reported as the most important that last marker is included in the region denoted by rs6947579 in 7q31) because of the HWE filter. Apparently, the loss of these two markers was enough to miss these regions. In other words, while trying to reproduce the QC filters, from one hand, we included markers in the qc version being excluded from the ibd and on the other hand, we removed markers from qc that were included in the ibd.

On CD_{qc}, with the T-Trees, a last notable difference is the presence of rs2157082, which share the same pathologic properties as rs10212068. It is removed by the WTCCC but in this case it allowed for the detection of a region reported in both [Wel07] Supplementary Information and [J$^+$13].

On CD_{wtccc}, with the Random Forest, most of the regions detected on CD_{qc} and CD_{ibd} are also detected. Markers strongly deviating from HWE appear in the top ranking. With the T-Trees, region 2p12 and 7q31 are detected but a few more from the reported strongly associated regions. An additional marker located in one of the 140 loci: rs12714959 is detected although it is not reported in the WTCCC publication. Also, rs11644392 and rs1553460 are also reported by [P$^+$12a] has "hub" SNPs.

Finally, we note the presence of a couple of markers located on chromosome X mostly with the T-Trees. Among these, we note the recurrent presence of markers located in the SMS gene (which is hypothesised to be linked to autoimmune diseases in [BLDP$^+$10]).

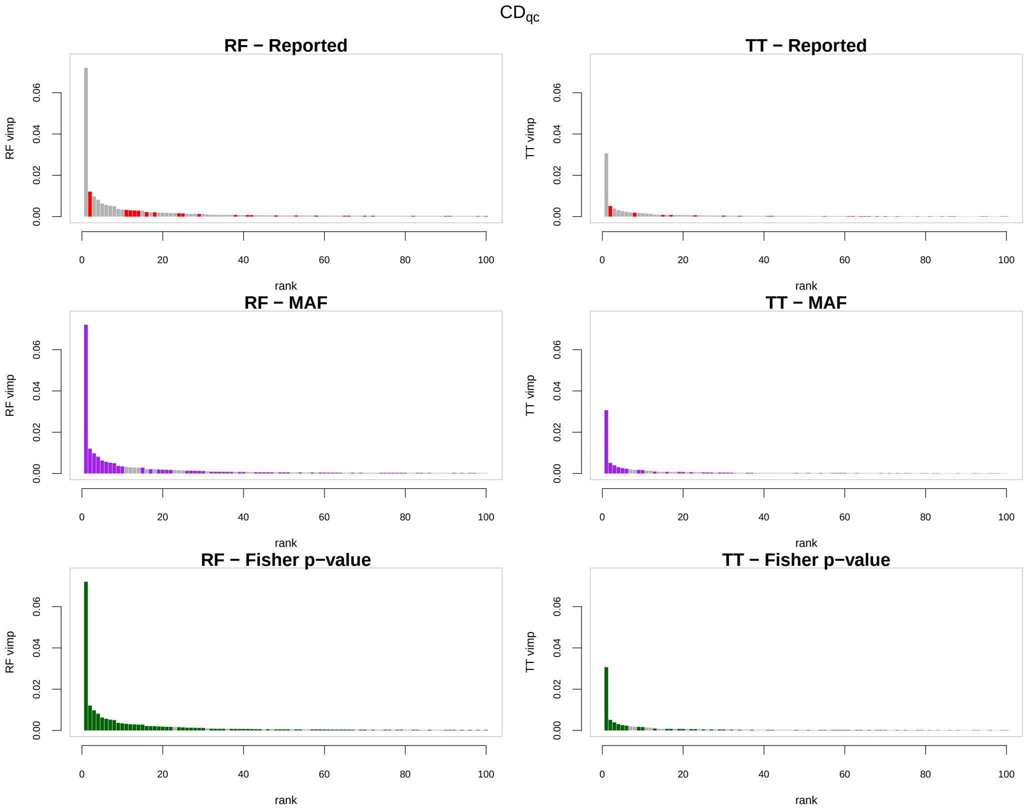


FIGURE 7.6 The first 100 variables according to the tree based importance rankings for CD_{qc}. The horizontal axis corresponds to the ranks and the vertical axis to the variable importances. In the first column variables are ordered according to random forests variable importances and in the second column they are ordered according to the T-Trees variable importances. In the first row, red highlights the reported strongly associated regions and orange the moderately associated regions. In the second row, purple corresponds to rare variants (MAF < 0.05). In the third row, green represents markers with a low Fisher p-value ($< 10^{-6}$).

Random Forests

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	67.31	67.46	10	rs11209026$^{(*)}$ (*IL23R*)	0.045	1	$3.52 \cdot 10^{-5}$	$8.24 \cdot 10^{-18}$	$1.40 \cdot 10^{-2}$ (1)
2	45.58	45.58	2	rs3755076	0.087	$2.65 \cdot 10^{-3}$	$6.02 \cdot 10^{-3}$	$5.18 \cdot 10^{-1}$	$5.30 \cdot 10^{-4}$ (48)
2	81.58	81.76	17	rs11887827	0.311	$1.54 \cdot 10^{-7}$	1	$2.42 \cdot 10^{-8}$	$1.27 \cdot 10^{-3}$ (20)
2	233.94	233.97	5	rs10210302$^{(*)}$ (*ATG16L1*)	0.452	$4.25 \cdot 10^{-1}$	$1.98 \cdot 10^{-2}$	$2.22 \cdot 10^{-13}$	$2.79 \cdot 10^{-3}$ (6)
3	49.43	49.68	6	rs11718165$^{(*)}$ (*BSN*)	0.295	$1.33 \cdot 10^{-2}$	$1.60 \cdot 10^{-2}$	$1.70 \cdot 10^{-6}$	$1.19 \cdot 10^{-3}$ (24)
4	114.61	114.62	2	rs17045935 (*ANK2*)	0.095	$2.38 \cdot 10^{-1}$	$1.07 \cdot 10^{-4}$	$5.28 \cdot 10^{-2}$	$6.45 \cdot 10^{-4}$ (39)
5	24.77	24.77	3	rs16893874	0.008	$2.61 \cdot 10^{-1}$	1	$3.18 \cdot 10^{-5}$	$3.32 \cdot 10^{-4}$ (80)
5	40.43	40.61	12	rs17234657$^{(*)}$	0.146	$4.18 \cdot 10^{-1}$	$3.51 \cdot 10^{-1}$	$1.72 \cdot 10^{-13}$	$2.26 \cdot 10^{-3}$ (10)
5	121.75	121.76	2	rs17149128 (*SNCAIP*)	0.122	$1.10 \cdot 10^{-2}$	$1.03 \cdot 10^{-2}$	$4.10 \cdot 10^{-1}$	$1.97 \cdot 10^{-4}$ (166)
5	150.21	150.31	4	rs931058	0.071	$5.64 \cdot 10^{-1}$	1	$1.53 \cdot 10^{-8}$	$5.83 \cdot 10^{-4}$ (44)
6	36.54	36.64	2	rs600382	0.001	1	1	$2.38 \cdot 10^{-5}$	$2.67 \cdot 10^{-4}$ (95)
8	129.88	129.96	4	rs10216909	0.003	1	1	$7.76 \cdot 10^{-5}$	$3.04 \cdot 10^{-4}$ (87)
10	65.96	65.96	2	rs16919914	0.080	$8.80 \cdot 10^{-2}$	$4.00 \cdot 10^{-4}$	$2.22 \cdot 10^{-1}$	$5.20 \cdot 10^{-4}$ (49)
11	130.84	130.84	2	rs1533339 (*NTM*)	0.005	1	1	$2.78 \cdot 10^{-4}$	$2.15 \cdot 10^{-4}$ (145)
16	49.30	49.32	4	rs2076756$^{(*)}$ (*NOD2*)	0.270	$4.50 \cdot 10^{-3}$	$7.62 \cdot 10^{-1}$	$3.95 \cdot 10^{-15}$	$3.88 \cdot 10^{-3}$ (4)
23	89.59	89.64	2	rs6522332	0.160	$3.10 \cdot 10^{-1}$	$5.50 \cdot 10^{-1}$	$3.23 \cdot 10^{-1}$	$2.08 \cdot 10^{-4}$ (155)
7	135.31	135.31	1	rs834771	0.151	$1.01 \cdot 10^{-1}$	$3.37 \cdot 10^{-2}$	$1.25 \cdot 10^{-3}$	$1.91 \cdot 10^{-4}$ (177)
8	77.90	77.90	1	rs10957818	0.024	$7.13 \cdot 10^{-1}$	$6.26 \cdot 10^{-1}$	$2.62 \cdot 10^{-5}$	$2.13 \cdot 10^{-4}$ (151)
14	77.10	77.10	1	rs4903604	0.227	$5.78 \cdot 10^{-3}$	$2.41 \cdot 10^{-2}$	$2.48 \cdot 10^{-3}$	$2.89 \cdot 10^{-4}$ (89)
18	12.77	12.77	1	rs2542151$^{(*)}$	0.180	$3.08 \cdot 10^{-1}$	$9.46 \cdot 10^{-1}$	$7.21 \cdot 10^{-8}$	$2.07 \cdot 10^{-4}$ (156)

T-Trees

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	3.25	3.26	2	rs12409315	0.077	$1.75 \cdot 10^{-1}$	$4.37 \cdot 10^{-3}$	$2.54 \cdot 10^{-3}$	$4.36 \cdot 10^{-4}$ (32)
1	67.31	67.46	10	rs11209026$^{(*)}$ (*IL23R*)	0.045	1	$3.52 \cdot 10^{-5}$	$8.24 \cdot 10^{-18}$	$5.23 \cdot 10^{-3}$ (5)
1	77.61	77.62	2	rs11162341	0.132	$4.01 \cdot 10^{-1}$	$3.78 \cdot 10^{-1}$	$8.99 \cdot 10^{-1}$	$2.28 \cdot 10^{-4}$ (57)
1	236.50	236.50	5	rs6677092 (*RPS7P5*)	0.373	$3.01 \cdot 10^{-6}$	1	$1.77 \cdot 10^{-4}$	$4.15 \cdot 10^{-4}$ (33)
2	81.58	81.85	35	rs11887827	0.311	$1.54 \cdot 10^{-7}$	1	$2.42 \cdot 10^{-8}$	$1.03 \cdot 10^{-2}$ (1)
2	143.22	143.28	2	SNP_A-2293058	0.003	1	1	$1.79 \cdot 10^{-5}$	$1.81 \cdot 10^{-4}$ (78)
2	233.94	233.97	5	rs10210302$^{(*)}$ (*ATG16L1*)	0.452	$4.25 \cdot 10^{-1}$	$1.98 \cdot 10^{-2}$	$2.22 \cdot 10^{-13}$	$3.07 \cdot 10^{-4}$ (48)
3	7.49	7.50	2	rs17047422	0.001	1	1	$3.45 \cdot 10^{-4}$	$1.91 \cdot 10^{-4}$ (73)
3	120.41	120.42	2	rs6774 (*B4GALT4*)	0.108	$2.50 \cdot 10^{-1}$	$7.10 \cdot 10^{-4}$	$1.39 \cdot 10^{-2}$	$3.41 \cdot 10^{-4}$ (43)
3	187.31	187.35	2	rs4686733	0.053	$6.50 \cdot 10^{-5}$	$1.12 \cdot 10^{-2}$	$3.65 \cdot 10^{-1}$	$1.39 \cdot 10^{-4}$ (93)
4	86.13	86.18	2	rs1872321	0.002	1	1	$6.88 \cdot 10^{-9}$	$1.19 \cdot 10^{-3}$ (17)
4	114.61	114.62	2	rs17045935 (*ANK2*)	0.095	$2.38 \cdot 10^{-1}$	$1.07 \cdot 10^{-4}$	$5.28 \cdot 10^{-2}$	$2.57 \cdot 10^{-4}$ (53)
4	178.27	178.28	3	rs1595154	0.002	1	1	$1.08 \cdot 10^{-7}$	$5.70 \cdot 10^{-4}$ (28)
5	40.43	40.53	10	rs17234657$^{(*)}$	0.146	$4.18 \cdot 10^{-1}$	$3.51 \cdot 10^{-1}$	$1.72 \cdot 10^{-13}$	$4.55 \cdot 10^{-4}$ (30)
6	21.33	21.35	2	rs16884693	0.004	1	1	$1.21 \cdot 10^{-3}$	$9.36 \cdot 10^{-5}$ (145)
6	129.84	129.84	3	rs2784899	0.260	$8.06 \cdot 10^{-1}$	$5.25 \cdot 10^{-1}$	$6.48 \cdot 10^{-2}$	$1.26 \cdot 10^{-4}$ (106)
7	35.37	35.37	2	rs10270692	0.066	1	$6.68 \cdot 10^{-1}$	$9.31 \cdot 10^{-2}$	$1.99 \cdot 10^{-4}$ (68)
7	125.13	125.16	9	rs6947579	0.317	$2.45 \cdot 10^{-1}$	$7.02 \cdot 10^{-1}$	$8.54 \cdot 10^{-1}$	$7.55 \cdot 10^{-3}$ (3)
8	129.90	129.92	2	rs10216909	0.003	1	1	$7.76 \cdot 10^{-5}$	$1.03 \cdot 10^{-4}$ (131)
10	38.31	38.38	2	rs11011417	0.001	1	1	$1.85 \cdot 10^{-5}$	$1.31 \cdot 10^{-4}$ (100)
11	14.16	14.16	2	rs9804490	0.459	$1.50 \cdot 10^{-10}$	$1.44 \cdot 10^{-6}$	$2.41 \cdot 10^{-5}$	$1.16 \cdot 10^{-4}$ (117)
12	42.78	42.80	2	rs11613902 (*TMEM117*)	0.099	$3.98 \cdot 10^{-7}$	$9.76 \cdot 10^{-2}$	$9.43 \cdot 10^{-1}$	$3.46 \cdot 10^{-4}$ (41)
14	84.39	84.43	4	rs10144260	0.008	1	1	$1.18 \cdot 10^{-9}$	$1.07 \cdot 10^{-3}$ (18)
14	104.47	104.53	2	rs2819467 (*C14orf79*)	0.011	$3.21 \cdot 10^{-1}$	1	$1.51 \cdot 10^{-3}$	$1.23 \cdot 10^{-4}$ (110)
16	49.30	49.31	3	rs2076756$^{(*)}$ (*NOD2*)	0.270	$4.50 \cdot 10^{-3}$	$7.62 \cdot 10^{-1}$	$3.95 \cdot 10^{-15}$	$6.43 \cdot 10^{-4}$ (25)
23	21.69	21.74	8	rs5904497 (*SMS*)	0.273	$5.97 \cdot 10^{-6}$	$1.50 \cdot 10^{-1}$	$4.41 \cdot 10^{-2}$	$3.26 \cdot 10^{-3}$ (9)
23	70.94	70.94	2	rs6624585 (*NHSL2*)	0.068	$7.76 \cdot 10^{-1}$	1	$2.69 \cdot 10^{-2}$	$2.24 \cdot 10^{-4}$ (58)
3	49.67	49.67	1	rs11718165$^{(*)}$ (*BSN*)	0.295	$1.33 \cdot 10^{-2}$	$1.60 \cdot 10^{-2}$	$1.70 \cdot 10^{-6}$	$7.93 \cdot 10^{-5}$ (159)
5	57.95	57.95	1	rs2279980	0.188	$8.16 \cdot 10^{-2}$	$7.99 \cdot 10^{-1}$	$6.19 \cdot 10^{-5}$	$7.03 \cdot 10^{-5}$ (182)
8	77.90	77.90	1	rs10957818	0.024	$7.13 \cdot 10^{-1}$	$6.26 \cdot 10^{-1}$	$2.62 \cdot 10^{-5}$	$1.06 \cdot 10^{-4}$ (126)
18	12.77	12.77	1	rs2542151$^{(*)}$	0.180	$3.08 \cdot 10^{-1}$	$9.46 \cdot 10^{-1}$	$7.21 \cdot 10^{-8}$	$9.35 \cdot 10^{-5}$ (146)

TABLE 7.9 CD_{ibd}: lists of regions identified by the Random Forests and the T-Trees methods.

Random Forests

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	67.31	67.46	11	rs11209026$^{(*)}$ (*IL23R*)	0.045	$6.24 \cdot 10^{-1}$	$3.59 \cdot 10^{-5}$	$3.71 \cdot 10^{-19}$	$1.21 \cdot 10^{-2}$ (2)
2	233.94	233.97	5	rs10210302$^{(*)}$ (*ATG16L1*)	0.451	$2.03 \cdot 10^{-1}$	$2.52 \cdot 10^{-2}$	$1.25 \cdot 10^{-14}$	$3.01 \cdot 10^{-3}$ (12)
3	7.49	7.50	2	rs9877337 (*GRM7*)	0.003	1	1	$1.34 \cdot 10^{-6}$	$2.84 \cdot 10^{-4}$ (94)
3	49.43	49.68	3	rs11718165$^{(*)}$ (*BSN*)	0.294	$1.88 \cdot 10^{-2}$	$2.41 \cdot 10^{-2}$	$2.55 \cdot 10^{-6}$	$4.83 \cdot 10^{-4}$ (58)
3	59.51	59.53	2	rs10510813	0.195	$1.36 \cdot 10^{-1}$	$1.78 \cdot 10^{-3}$	$7.49 \cdot 10^{-5}$	$3.04 \cdot 10^{-4}$ (89)
4	101.65	101.65	2	rs2903213	0.101	$9.86 \cdot 10^{-3}$	$2.25 \cdot 10^{-1}$	$1.40 \cdot 10^{-4}$	$2.59 \cdot 10^{-4}$ (109)
5	40.43	40.60	12	rs17234657$^{(*)}$	0.145	$2.76 \cdot 10^{-1}$	$3.12 \cdot 10^{-1}$	$5.23 \cdot 10^{-13}$	$1.50 \cdot 10^{-3}$ (25)
5	121.75	121.76	2	rs17149128 (*SNCAIP*)	0.124	$7.54 \cdot 10^{-4}$	$1.38 \cdot 10^{-2}$	$1.47 \cdot 10^{-1}$	$3.04 \cdot 10^{-4}$ (88)
5	150.24	150.31	2	rs931058	0.072	$8.91 \cdot 10^{-1}$	1	$1.49 \cdot 10^{-8}$	$2.66 \cdot 10^{-4}$ (103)
10	53.84	53.84	2	rs10824464	0.014	$5.55 \cdot 10^{-1}$	1	$3.98 \cdot 10^{-6}$	$2.10 \cdot 10^{-4}$ (139)
11	130.84	130.84	2	rs1533339 (*NTM*)	0.006	1	1	$2.61 \cdot 10^{-6}$	$2.29 \cdot 10^{-4}$ (121)
14	57.36	57.37	2	rs17093726 (*SLC35F4*)	0.119	$2.09 \cdot 10^{-4}$	1	$3.77 \cdot 10^{-3}$	$2.15 \cdot 10^{-4}$ (132)
16	49.30	49.36	8	rs2076756$^{(*)}$ (*NOD2*)	0.269	$3.61 \cdot 10^{-4}$	$7.63 \cdot 10^{-1}$	$1.72 \cdot 10^{-14}$	$3.22 \cdot 10^{-3}$ (11)
23	134.79	134.86	2	rs3761643	0.024	$6.26 \cdot 10^{-1}$	1	$9.39 \cdot 10^{-1}$	$6.15 \cdot 10^{-4}$ (45)
2	45.58	45.58	1	rs3755076	0.090	$3.66 \cdot 10^{-3}$	$1.91 \cdot 10^{-2}$	$9.08 \cdot 10^{-2}$	$2.20 \cdot 10^{-4}$ (127)
10	38.53	38.53	1	rs11011491	0.003	1	1	$4.58 \cdot 10^{-5}$	$1.64 \cdot 10^{-4}$ (194)
14	77.10	77.10	1	rs4903604	0.227	$9.38 \cdot 10^{-3}$	$3.29 \cdot 10^{-2}$	$2.82 \cdot 10^{-3}$	$2.18 \cdot 10^{-4}$ (128)
18	12.77	12.77	1	rs2542151$^{(*)}$	0.181	$1.45 \cdot 10^{-1}$	$8.41 \cdot 10^{-1}$	$1.46 \cdot 10^{-7}$	$1.94 \cdot 10^{-4}$ (155)

T-Trees

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	67.31	67.46	15	rs11209026$^{(*)}$ (*IL23R*)	0.045	$6.24 \cdot 10^{-1}$	$3.59 \cdot 10^{-5}$	$3.71 \cdot 10^{-19}$	$5.12 \cdot 10^{-3}$ (2)
1	236.47	236.50	5	rs6677092 (*RPS7P5*)	0.373	$2.56 \cdot 10^{-6}$	$9.38 \cdot 10^{-1}$	$3.01 \cdot 10^{-4}$	$4.76 \cdot 10^{-4}$ (28)
2	5.37	5.41	7	rs1453783$^{(3)}$	0.380	$4.07 \cdot 10^{-2}$	$1.62 \cdot 10^{-2}$	$1.19 \cdot 10^{-2}$	$1.42 \cdot 10^{-3}$ (11)
2	10.90	10.93	5	rs902133$^{(1)}$	0.004	1	1	$3.31 \cdot 10^{-11}$	$2.61 \cdot 10^{-3}$ (5)
2	233.94	233.97	5	rs10210302$^{(*)}$ (*ATG16L1*)	0.451	$2.03 \cdot 10^{-1}$	$2.52 \cdot 10^{-2}$	$1.25 \cdot 10^{-14}$	$3.00 \cdot 10^{-4}$ (42)
3	7.50	7.50	2	rs17047426 (*GRM7*)	0.002	1	1	$3.97 \cdot 10^{-6}$	$2.40 \cdot 10^{-4}$ (59)
3	120.41	120.42	2	rs6774 (*B4GALT4*)	0.109	$1.94 \cdot 10^{-1}$	$3.94 \cdot 10^{-4}$	$6.45 \cdot 10^{-3}$	$2.44 \cdot 10^{-4}$ (56)
3	133.77	133.78	2	rs3762678	0.002	1	1	$3.70 \cdot 10^{-7}$	$5.01 \cdot 10^{-4}$ (27)
3	187.31	187.35	3	rs4686733	0.055	$3.79 \cdot 10^{-6}$	$7.63 \cdot 10^{-3}$	$7.16 \cdot 10^{-1}$	$2.55 \cdot 10^{-4}$ (53)
4	56.95	56.96	3	rs4865080 (*KIAA1211*)	0.003	$1.14 \cdot 10^{-1}$	1	$3.27 \cdot 10^{-13}$	$1.71 \cdot 10^{-3}$ (9)
4	114.61	114.62	2	rs17045935 (*ANK2*)	0.096	$8.72 \cdot 10^{-2}$	$7.68 \cdot 10^{-5}$	$2.98 \cdot 10^{-2}$	$2.82 \cdot 10^{-4}$ (46)
4	147.08	147.09	2	rs17020598 (*ZNF827*)	0.003	1	1	$5.47 \cdot 10^{-6}$	$1.31 \cdot 10^{-4}$ (107)
4	178.27	178.28	3	rs1595154	0.003	1	1	$1.70 \cdot 10^{-10}$	$7.57 \cdot 10^{-4}$ (19)
5	40.43	40.53	6	rs17234657$^{(*)}$	0.145	$2.76 \cdot 10^{-1}$	$3.12 \cdot 10^{-1}$	$5.23 \cdot 10^{-13}$	$3.50 \cdot 10^{-4}$ (34)
5	113.07	113.07	2	rs6881153	0.062	$5.72 \cdot 10^{-1}$	$4.22 \cdot 10^{-1}$	$5.79 \cdot 10^{-1}$	$1.10 \cdot 10^{-4}$ (135)
6	32.85	32.87	9	rs2157082$^{(3)}$	0.425	$2.81 \cdot 10^{-1}$	$3.95 \cdot 10^{-4}$	$1.94 \cdot 10^{-2}$	$1.84 \cdot 10^{-3}$ (8)
6	63.43	63.44	2	rs7761764	0.031	$4.01 \cdot 10^{-1}$	$7.85 \cdot 10^{-2}$	$1.37 \cdot 10^{-1}$	$9.19 \cdot 10^{-5}$ (156)
7	35.37	35.37	2	rs12540326	0.139	1	$3.08 \cdot 10^{-1}$	$2.00 \cdot 10^{-1}$	$1.14 \cdot 10^{-4}$ (130)
8	30.27	30.33	5	rs7842024	0.003	1	1	$8.69 \cdot 10^{-12}$	$2.30 \cdot 10^{-3}$ (6)
8	32.06	32.09	3	rs16878847 (*NRG1*)	0.003	1	1	$8.31 \cdot 10^{-11}$	$1.64 \cdot 10^{-3}$ (10)
9	76.50	76.54	3	rs17062858 (*PRUNE2*)	0.012	1	1	$1.92 \cdot 10^{-4}$	$1.70 \cdot 10^{-4}$ (80)
9	110.09	110.10	2	rs3808888 (*TXN*)	0.002	$9.44 \cdot 10^{-4}$	1	$8.27 \cdot 10^{-9}$	$3.29 \cdot 10^{-4}$ (37)
10	131.95	131.97	2	rs7080464$^{(1)}$	0.030	$7.50 \cdot 10^{-2}$	$4.05 \cdot 10^{-1}$	$2.72 \cdot 10^{-7}$	$2.34 \cdot 10^{-4}$ (60)
11	37.28	37.44	4	rs12224887$^{(3)}$	0.014	$6.24 \cdot 10^{-1}$	1	$1.54 \cdot 10^{-8}$	$9.44 \cdot 10^{-4}$ (13)
12	42.76	42.78	2	rs11613902 (*TMEM117*)	0.099	$4.67 \cdot 10^{-8}$	$8.18 \cdot 10^{-2}$	$9.17 \cdot 10^{-1}$	$3.11 \cdot 10^{-4}$ (39)
13	93.28	93.31	2	rs12429608 (*GPC6, GPC6-AS2*)	0.003	$6.90 \cdot 10^{-2}$	1	$6.06 \cdot 10^{-7}$	$2.43 \cdot 10^{-4}$ (57)
14	83.00	83.04	5	rs10144243	0.006	1	1	$1.51 \cdot 10^{-13}$	$3.07 \cdot 10^{-3}$ (4)
14	104.47	104.53	2	rs2819467 (*C14orf79*)	0.011	$3.81 \cdot 10^{-1}$	1	$3.03 \cdot 10^{-4}$	$1.30 \cdot 10^{-4}$ (108)
15	80.18	80.21	4	rs16973411$^{(1)}$	0.003	1	1	$5.88 \cdot 10^{-14}$	$3.93 \cdot 10^{-3}$ (3)
16	49.30	49.32	3	rs2076756$^{(*)}$ (*NOD2*)	0.269	$3.61 \cdot 10^{-4}$	$7.63 \cdot 10^{-1}$	$1.72 \cdot 10^{-14}$	$6.34 \cdot 10^{-4}$ (23)
22	35.92	35.97	8	rs10212068$^{(3)}$	0.028	1	$2.60 \cdot 10^{-3}$	$3.20 \cdot 10^{-56}$	$3.07 \cdot 10^{-2}$ (1)
23	21.73	21.74	6	rs5904497 (*SMS*)	0.271	$1.73 \cdot 10^{-6}$	$1.19 \cdot 10^{-1}$	$3.54 \cdot 10^{-2}$	$2.01 \cdot 10^{-3}$ (7)
23	70.94	70.94	2	rs6624585 (*NHSL2*)	0.068	1	1	$5.13 \cdot 10^{-2}$	$2.73 \cdot 10^{-4}$ (49)
18	12.77	12.77	1	rs2542151$^{(*)}$	0.181	$1.45 \cdot 10^{-1}$	$8.41 \cdot 10^{-1}$	$1.46 \cdot 10^{-7}$	$8.25 \cdot 10^{-5}$ (170)

TABLE 7.10 CD_{qc}: lists of regions identified by the Random Forests and the T-Trees methods.

Random Forests

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	67.31	67.46	11	rs11209026$^{(*)}$ (*IL23R*)	0.045	1	$3.53 \cdot 10^{-5}$	$5.43 \cdot 10^{-18}$	$6.89 \cdot 10^{-3}$ (10)
1	117.16	117.18	9	rs12078461 (*PTGFRN*)	0.047	$2.51 \cdot 10^{-182}$	$4.91 \cdot 10^{-2}$	$7.19 \cdot 10^{-13}$	$4.95 \cdot 10^{-2}$ (2)
2	25.31	25.36	3	rs2164411	0.155	$2.85 \cdot 10^{-13}$	$1.95 \cdot 10^{-1}$	$4.82 \cdot 10^{-3}$	$5.32 \cdot 10^{-3}$ (13)
2	233.94	233.97	5	rs10210302$^{(*)}$ (*ATG16L1*)	0.451	$4.26 \cdot 10^{-1}$	$1.98 \cdot 10^{-2}$	$1.08 \cdot 10^{-13}$	$1.32 \cdot 10^{-3}$ (41)
3	16.45	16.46	2	rs9839841 (*RFTN1*)	0.195	$2.30 \cdot 10^{-6}$	$5.14 \cdot 10^{-1}$	$7.20 \cdot 10^{-13}$	$1.38 \cdot 10^{-3}$ (38)
3	49.43	49.68	3	rs11718165$^{(*)}$ (*BSN*)	0.295	$1.57 \cdot 10^{-2}$	$1.61 \cdot 10^{-2}$	$2.21 \cdot 10^{-6}$	$4.21 \cdot 10^{-4}$ (102)
4	16.37	16.48	14	rs157613 (*LDB2*)	0.082	$3.16 \cdot 10^{-253}$	$3.96 \cdot 10^{-6}$	$1.13 \cdot 10^{-14}$	$9.02 \cdot 10^{-2}$ (1)
4	17.73	17.93	16	rs1553460	0.315	$5.84 \cdot 10^{-93}$	$2.88 \cdot 10^{-5}$	$1.59 \cdot 10^{-31}$	$4.09 \cdot 10^{-2}$ (4)
4	158.43	158.43	3	rs17035797 (*GLRB*)	0.074	$1.38 \cdot 10^{-10}$	$9.06 \cdot 10^{-1}$	$1.29 \cdot 10^{-9}$	$5.16 \cdot 10^{-4}$ (93)
4	186.09	186.11	3	rs13126272 (*ACSL1*)	0.338	$5.00 \cdot 10^{-58}$	$3.02 \cdot 10^{-2}$	$3.65 \cdot 10^{-5}$	$6.24 \cdot 10^{-3}$ (11)
5	40.37	40.52	9	rs17234657$^{(*)}$	0.146	$4.18 \cdot 10^{-1}$	$3.52 \cdot 10^{-1}$	$2.37 \cdot 10^{-13}$	$7.16 \cdot 10^{-4}$ (68)
5	117.00	117.07	16	rs2416472	0.319	$2.40 \cdot 10^{-17}$	$9.35 \cdot 10^{-1}$	$7.19 \cdot 10^{-12}$	$4.55 \cdot 10^{-3}$ (14)
6	121.66	121.68	3	rs17083420 (*C6orf170*)	0.009	$3.45 \cdot 10^{-11}$	1	$7.68 \cdot 10^{-7}$	$1.52 \cdot 10^{-3}$ (36)
9	132.51	132.60	13	rs302925	0.475	$4.67 \cdot 10^{-20}$	$5.50 \cdot 10^{-1}$	$6.32 \cdot 10^{-6}$	$1.15 \cdot 10^{-3}$ (45)
10	125.67	125.67	3	rs7067790	0.391	$4.18 \cdot 10^{-22}$	$3.56 \cdot 10^{-1}$	$1.10 \cdot 10^{-7}$	$7.82 \cdot 10^{-3}$ (9)
11	113.02	113.31	17	rs17116117 (*HTR3B*)	0.049	$2.92 \cdot 10^{-4}$	$1.16 \cdot 10^{-1}$	$1.03 \cdot 10^{-23}$	$1.12 \cdot 10^{-2}$ (8)
14	97.06	97.07	6	rs234202	0.008	1	1	$1.21 \cdot 10^{-9}$	$1.41 \cdot 10^{-3}$ (37)
16	30.23	30.29	3	rs4471699 (*LOC595101*)	0.449	$2.31 \cdot 10^{-49}$	$6.03 \cdot 10^{-1}$	$4.64 \cdot 10^{-20}$	$3.27 \cdot 10^{-2}$ (5)
16	49.30	49.31	3	rs2076756$^{(*)}$ (*NOD2*)	0.270	$4.61 \cdot 10^{-3}$	$7.62 \cdot 10^{-1}$	$3.00 \cdot 10^{-15}$	$1.38 \cdot 10^{-3}$ (39)
2	45.58	45.58	1	rs3755076	0.087	$2.76 \cdot 10^{-3}$	$6.02 \cdot 10^{-3}$	$4.95 \cdot 10^{-1}$	$1.52 \cdot 10^{-4}$ (200)
6	32.79	32.79	1	rs3104404	0.148	$5.65 \cdot 10^{-2}$	$1.03 \cdot 10^{-1}$	$2.31 \cdot 10^{-6}$	$2.21 \cdot 10^{-4}$ (155)

T-Trees

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	67.31	67.42	5	rs11209026$^{(*)}$ (*IL23R*)	0.045	1	$3.53 \cdot 10^{-5}$	$5.43 \cdot 10^{-18}$	$2.55 \cdot 10^{-3}$ (28)
1	117.16	117.18	10	rs12078461 (*PTGFRN*)	0.047	$2.51 \cdot 10^{-182}$	$4.91 \cdot 10^{-2}$	$7.19 \cdot 10^{-13}$	$3.32 \cdot 10^{-2}$ (2)
1	214.81	214.86	6	rs1933641 (*RRP15*)	0.047	$3.76 \cdot 10^{-180}$	$7.90 \cdot 10^{-2}$	$7.59 \cdot 10^{-14}$	$3.17 \cdot 10^{-2}$ (3)
2	25.31	25.36	3	rs2164411	0.155	$2.85 \cdot 10^{-13}$	$1.95 \cdot 10^{-1}$	$4.82 \cdot 10^{-3}$	$1.53 \cdot 10^{-3}$ (41)
2	81.54	81.72	14	rs11887827	0.311	$1.66 \cdot 10^{-7}$	1	$2.19 \cdot 10^{-8}$	$4.37 \cdot 10^{-3}$ (16)
2	132.48	132.55	2	rs4080478	0.104	$2.69 \cdot 10^{-8}$	$1.44 \cdot 10^{-4}$	$3.64 \cdot 10^{-3}$	$9.92 \cdot 10^{-4}$ (62)
3	16.45	16.46	2	rs9839841 (*RFTN1*)	0.195	$2.30 \cdot 10^{-6}$	$5.14 \cdot 10^{-1}$	$7.20 \cdot 10^{-13}$	$4.83 \cdot 10^{-4}$ (93)
3	18.60	18.60	2	rs12714959	0.263	$1.74 \cdot 10^{-23}$	$7.82 \cdot 10^{-5}$	$5.06 \cdot 10^{-7}$	$1.46 \cdot 10^{-3}$ (44)
3	184.57	184.57	3	rs959880 (*MCF2L2*)	0.144	$4.85 \cdot 10^{-1}$	$2.65 \cdot 10^{-1}$	$4.64 \cdot 10^{-1}$	$1.14 \cdot 10^{-3}$ (56)
4	16.43	16.48	15	rs157613 (*LDB2*)	0.082	$3.16 \cdot 10^{-253}$	$3.96 \cdot 10^{-6}$	$1.13 \cdot 10^{-14}$	$6.11 \cdot 10^{-2}$ (1)
4	17.85	17.90	9	rs1553460	0.315	$5.84 \cdot 10^{-93}$	$2.88 \cdot 10^{-5}$	$1.59 \cdot 10^{-31}$	$1.96 \cdot 10^{-2}$ (4)
4	38.77	38.78	2	rs6816863	0.027	$4.18 \cdot 10^{-89}$	1	$2.86 \cdot 10^{-1}$	$1.25 \cdot 10^{-2}$ (9)
4	56.95	56.96	2	rs4865080 (*KIAA1211*)	0.002	1	1	$3.98 \cdot 10^{-10}$	$5.48 \cdot 10^{-4}$ (85)
4	158.43	158.43	3	rs17035814 (*GLRB*)	0.089	$1.05 \cdot 10^{-1}$	$6.44 \cdot 10^{-1}$	$7.35 \cdot 10^{-1}$	$2.52 \cdot 10^{-3}$ (29)
4	186.09	186.13	6	rs13126272 (*ACSL1*)	0.338	$5.00 \cdot 10^{-58}$	$3.02 \cdot 10^{-2}$	$3.65 \cdot 10^{-5}$	$5.33 \cdot 10^{-3}$ (15)
5	40.37	40.44	2	rs1186661$^{(*)}$	0.133	$3.80 \cdot 10^{-1}$	$2.76 \cdot 10^{-3}$	$1.55 \cdot 10^{-12}$	$2.27 \cdot 10^{-4}$ (140)
5	117.02	117.06	9	rs17411921	0.339	$2.19 \cdot 10^{-1}$	$3.66 \cdot 10^{-1}$	$7.87 \cdot 10^{-1}$	$1.83 \cdot 10^{-2}$ (5)
6	93.79	93.82	10	rs6454931	0.279	$2.16 \cdot 10^{-1}$	$8.90 \cdot 10^{-1}$	$5.84 \cdot 10^{-1}$	$3.58 \cdot 10^{-3}$ (19)
6	121.63	121.68	4	rs17083420 (*C6orf170*)	0.009	$3.45 \cdot 10^{-11}$	1	$7.68 \cdot 10^{-7}$	$2.62 \cdot 10^{-3}$ (26)
7	38.93	38.97	3	rs1525791 (*POU6F2*)	0.152	$1.22 \cdot 10^{-3}$	$6.26 \cdot 10^{-1}$	$2.14 \cdot 10^{-8}$	$6.11 \cdot 10^{-4}$ (79)
7	125.13	125.14	6	rs6947579	0.317	$2.24 \cdot 10^{-1}$	$7.34 \cdot 10^{-1}$	$8.19 \cdot 10^{-1}$	$3.41 \cdot 10^{-3}$ (21)
9	132.59	132.60	7	rs10901198 (*GTF3C4*)	0.119	$9.10 \cdot 10^{-1}$	$4.79 \cdot 10^{-1}$	$7.92 \cdot 10^{-1}$	$1.04 \cdot 10^{-2}$ (10)
10	10.32	10.32	2	rs2151595	0.068	1	$7.76 \cdot 10^{-1}$	$2.60 \cdot 10^{-1}$	$3.16 \cdot 10^{-4}$ (118)
10	125.66	125.67	4	rs769282	0.415	$6.57 \cdot 10^{-1}$	$4.94 \cdot 10^{-1}$	$7.95 \cdot 10^{-1}$	$5.50 \cdot 10^{-3}$ (14)
11	55.32	55.35	2	rs7951100	0.070	$2.58 \cdot 10^{-1}$	$3.97 \cdot 10^{-1}$	$6.75 \cdot 10^{-1}$	$3.03 \cdot 10^{-4}$ (121)
11	113.28	113.31	5	rs1176741 (*HTR3B*)	0.030	$6.45 \cdot 10^{-1}$	$1.19 \cdot 10^{-1}$	$3.50 \cdot 10^{-1}$	$1.66 \cdot 10^{-2}$ (6)
14	35.06	35.16	4	rs10483456 (*RALGAPA1*)	0.059	$5.04 \cdot 10^{-1}$	$5.22 \cdot 10^{-1}$	$2.81 \cdot 10^{-11}$	$1.29 \cdot 10^{-3}$ (49)
14	59.81	59.83	3	rs7154773 (*PPM1A*)	0.352	$2.10 \cdot 10^{-1}$	$1.96 \cdot 10^{-1}$	$6.84 \cdot 10^{-1}$	$6.87 \cdot 10^{-4}$ (72)
14	83.04	83.06	2	rs10144243	0.005	1	1	$2.25 \cdot 10^{-12}$	$1.24 \cdot 10^{-3}$ (51)
14	97.06	97.10	6	rs11846702	0.012	1	$3.33 \cdot 10^{-1}$	$4.42 \cdot 10^{-1}$	$2.76 \cdot 10^{-3}$ (25)
16	29.84	30.29	7	rs11644392 (*LOC595101*)	0.484	$6.99 \cdot 10^{-1}$	$6.83 \cdot 10^{-1}$	$2.20 \cdot 10^{-1}$	$1.60 \cdot 10^{-2}$ (7)
17	50.27	50.29	2	rs2934884	0.198	$4.06 \cdot 10^{-1}$	$4.16 \cdot 10^{-1}$	$9.15 \cdot 10^{-1}$	$7.33 \cdot 10^{-4}$ (70)
23	0.63	0.64	6	rs5988334	0.216	$2.36 \cdot 10^{-2}$	$2.06 \cdot 10^{-3}$	$1.13 \cdot 10^{-5}$	$1.36 \cdot 10^{-3}$ (47)
23	2.58	2.58	2	rs1419930	0.041	$5.51 \cdot 10^{-2}$	$5.71 \cdot 10^{-19}$	$2.33 \cdot 10^{-1}$	$1.11 \cdot 10^{-3}$ (57)
23	21.73	21.73	2	rs4824171 (*SMS*)	0.284	$6.52 \cdot 10^{-1}$	$2.14 \cdot 10^{-1}$	$3.22 \cdot 10^{-1}$	$8.84 \cdot 10^{-4}$ (66)
2	233.94	233.94	1	rs10210302$^{(*)}$ (*ATG16L1*)	0.451	$4.26 \cdot 10^{-1}$	$1.98 \cdot 10^{-2}$	$1.08 \cdot 10^{-13}$	$2.68 \cdot 10^{-4}$ (132)
16	49.31	49.31	1	rs2076756$^{(*)}$ (*NOD2*)	0.270	$4.61 \cdot 10^{-3}$	$7.62 \cdot 10^{-1}$	$3.00 \cdot 10^{-15}$	$4.29 \cdot 10^{-4}$ (105)

TABLE 7.11 CD_{wtccc}: lists of regions identified by the Random Forests and the T-Trees methods.

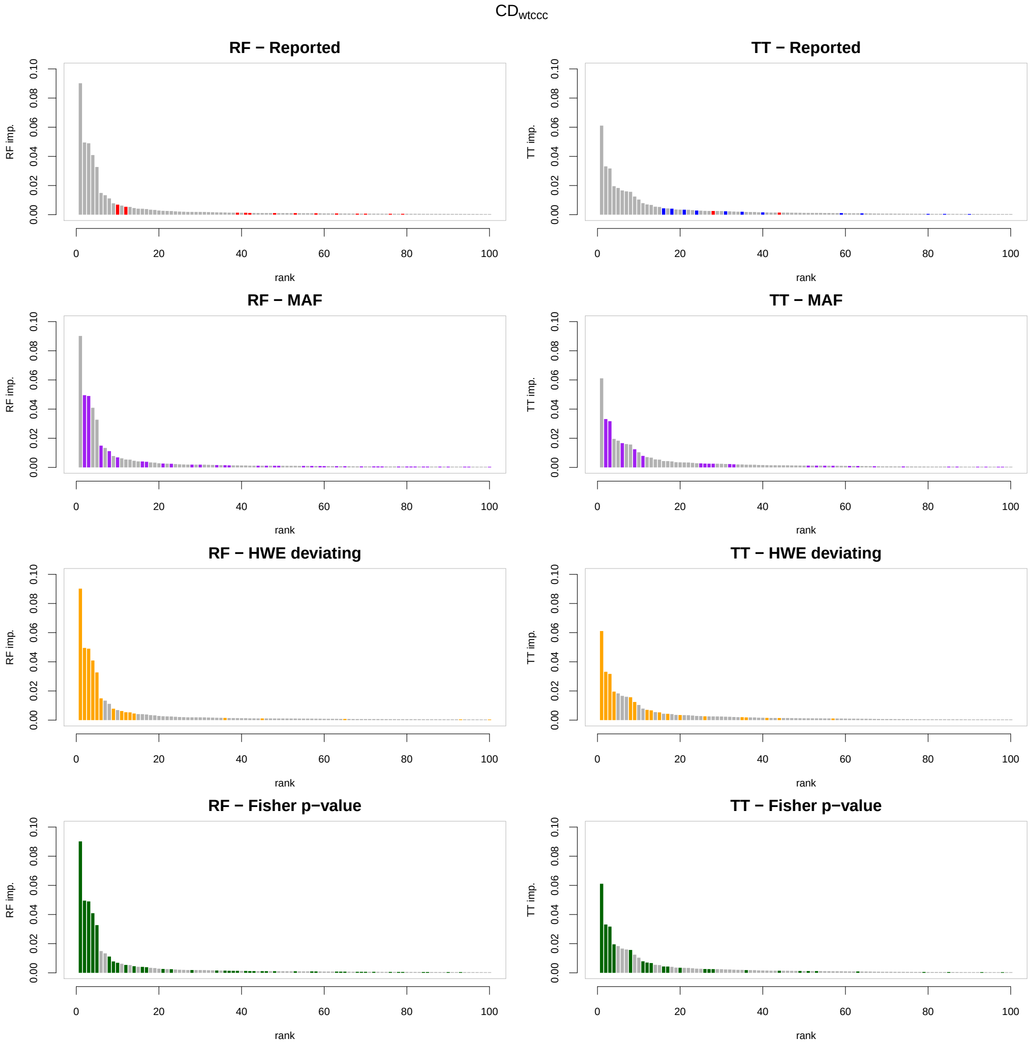


Figure 7.7 The first 100 variables according to the tree based importance rankings for CD_{wtccc}. The horizontal axis corresponds to the ranks and the vertical axis to the variable importances. In the first column variables are ordered according to random forests variable importances and in the second column they are ordered according to the T-Trees variable importances. In the first row, red highlights the nine reported regions and blue highlights two more regions mostly detected by tree based methods. In the second row, purple corresponds to rare variants (MAF < 0.05). In the third row, orange highlights SNPs deviating from HWE and in the last row, green represents markers with a low Fisher p-value ($< 10^{-6}$).

Hypertension

As for previous diseases, we obtain more rare variants in the top ranked variables on the qc version and more markers deviating from HWE on HT_{wtccc}. On HT_{qc}, we see `rs10212068` and `rs3122348` popping out again in the top ranked variables.

Interestingly, `rs6499937` located in `CSNK2A2` is detected in our four experiments. That gene is reported in [YLLP11] as potentially linked to hypertension. Many genes are also detected by a multiple-maker analysis in [SFS^{+}11] and well identified (particularly on HT_{qc}) by the T-Trees. Among these: `rs1372662` (`ZFAT`), `rs10188442` (`GPR39`) and `rs200759` (`MACROD2`) were found. Most of the SNP pairs reported in [SFS^{+}11] are captured in the corresponding groups in our table. Note that the `ZFAT` region is also found in [FZ10]. In addition, on HT_{wtccc}, `rs10843660` is also reported by [P^{+}12a] as being a "hub".

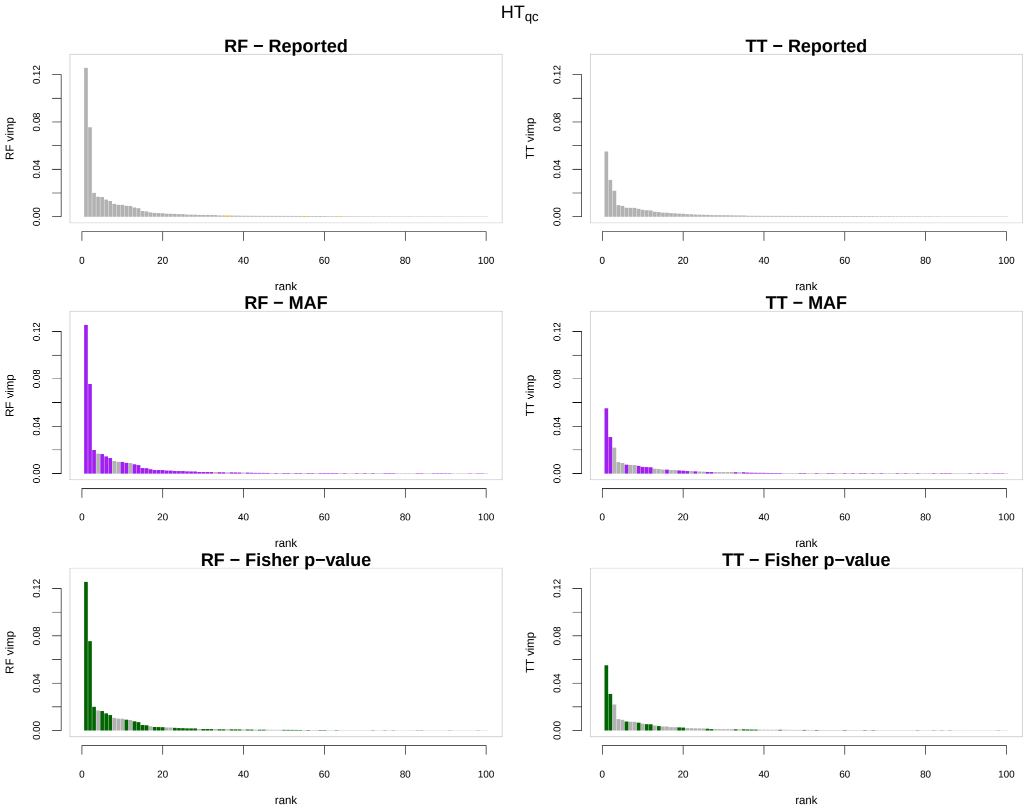


FIGURE 7.8 The first 100 variables according to the tree based importance rankings for HT_{qc}. The horizontal axis corresponds to the ranks and the vertical axis to the variable importances. In the first column variables are ordered according to random forests variable importances and in the second column they are ordered according to the T-Trees variable importances. In the first row, red highlights the reported strongly associated regions. In the second row, purple corresponds to rare variants (MAF < 0.05). In the third row, green represents markers with a low Fisher p-value ($< 10^{-6}$).

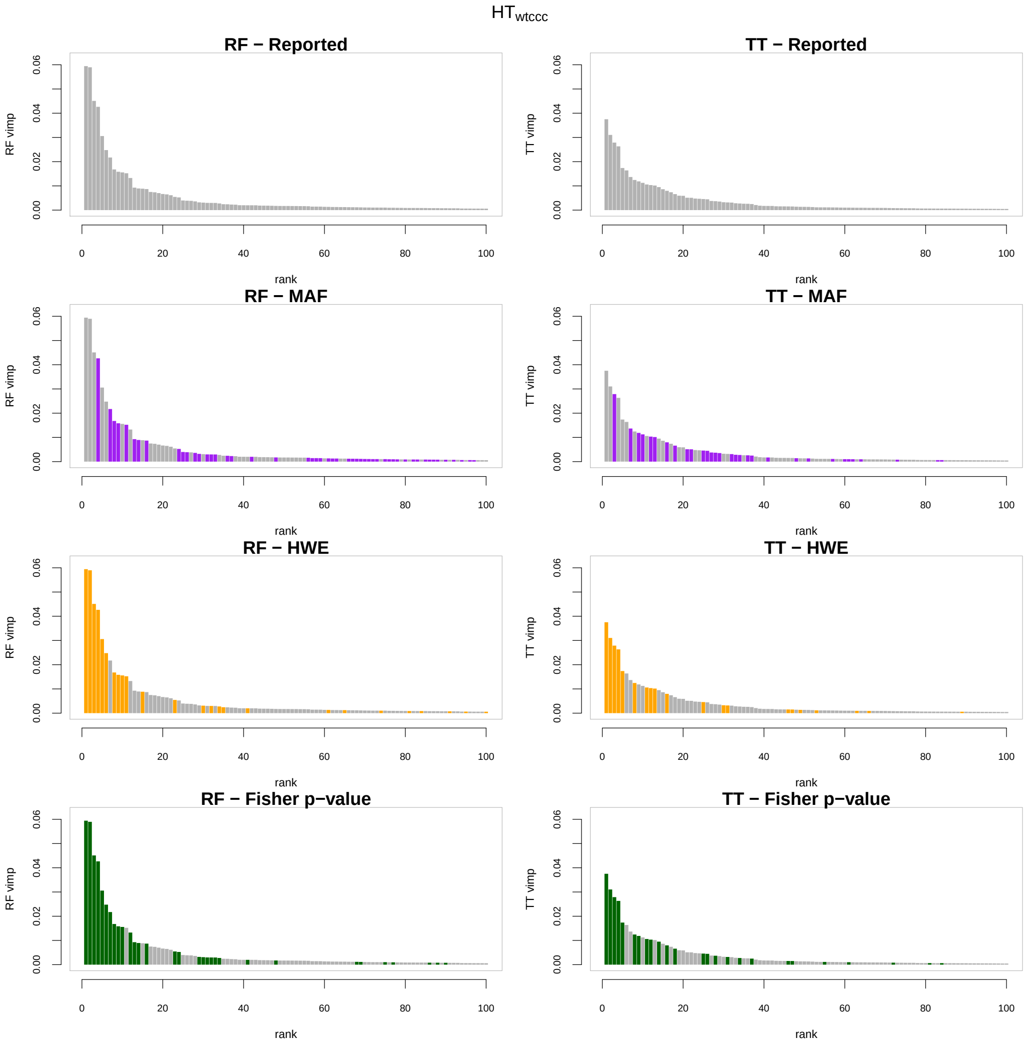


FIGURE 7.9 The first 100 variables according to the tree based importance rankings for HT_{wtccc}. The horizontal axis corresponds to the ranks and the vertical axis to the variable importances. In the first column variables are ordered according to random forests variable importances and in the second column they are ordered according to the T-Trees variable importances. In the first row, red highlights the strongly associated reported regions. In the second row, purple corresponds to rare variants (MAF < 0.05). In the third row, orange highlights SNPs deviating from HWE and in the last row, green represents markers with a low Fisher p-value ($< 10^{-6}$).

Random Forests

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
2	217.31	217.34	2	rs6733115	0.002	1	1	$8.63 \cdot 10^{-6}$	$4.71 \cdot 10^{-4}$ (59)
3	150.14	150.16	3	rs12163649	0.249	$7.21 \cdot 10^{-2}$	$4.86 \cdot 10^{-2}$	$2.90 \cdot 10^{-3}$	$1.96 \cdot 10^{-4}$ (95)
3	154.52	154.52	2	rs10513440	0.197	$1.70 \cdot 10^{-1}$	$2.71 \cdot 10^{-3}$	$1.71 \cdot 10^{-2}$	$1.58 \cdot 10^{-4}$ (116)
6	89.37	89.38	2	rs2610769	0.148	$1.70 \cdot 10^{-1}$	$2.69 \cdot 10^{-3}$	$7.26 \cdot 10^{-2}$	$1.31 \cdot 10^{-4}$ (147)
6	167.32	167.32	2	rs2757045 (*RNASET2*)	0.103	$6.13 \cdot 10^{-3}$	$2.91 \cdot 10^{-1}$	$1.33 \cdot 10^{-2}$	$2.36 \cdot 10^{-4}$ (82)
7	93.00	93.01	2	rs2519602	0.149	$6.60 \cdot 10^{-3}$	$1.55 \cdot 10^{-1}$	$3.26 \cdot 10^{-2}$	$1.30 \cdot 10^{-4}$ (149)
9	117.69	117.81	5	rs488101	0.363	$5.56 \cdot 10^{-6}$	$5.87 \cdot 10^{-1}$	$4.70 \cdot 10^{-7}$	$5.60 \cdot 10^{-4}$ (52)
10	30.99	31.06	9	rs3122348(*LOC645954*)$^{(1)}$	0.034	$4.73 \cdot 10^{-6}$	1	$6.12 \cdot 10^{-96}$	$1.26 \cdot 10^{-1}$ (1)
10	58.39	58.41	2	rs2393191	0.016	$3.10 \cdot 10^{-1}$	$5.71 \cdot 10^{-1}$	$2.46 \cdot 10^{-1}$	$2.57 \cdot 10^{-3}$ (21)
10	131.95	132.00	3	rs7080464$^{(1)}$	0.031	$4.70 \cdot 10^{-2}$	$4.05 \cdot 10^{-1}$	$5.07 \cdot 10^{-8}$	$1.27 \cdot 10^{-3}$ (32)
11	133.33	133.34	2	rs493888	0.120	$8.14 \cdot 10^{-6}$	$3.92 \cdot 10^{-1}$	$3.92 \cdot 10^{-1}$	$2.42 \cdot 10^{-4}$ (80)
14	60.77	60.77	2	rs4902035	0.096	$7.15 \cdot 10^{-3}$	$1.01 \cdot 10^{-4}$	$5.53 \cdot 10^{-1}$	$3.56 \cdot 10^{-4}$ (64)
15	63.11	63.12	2	rs2414869 (*MTFMT*)	0.120	$1.20 \cdot 10^{-3}$	$1.72 \cdot 10^{-2}$	$2.54 \cdot 10^{-1}$	$2.65 \cdot 10^{-4}$ (73)
16	56.75	56.79	2	rs6499937 (*CSNK2A2*)	0.006	1	1	$1.30 \cdot 10^{-8}$	$2.84 \cdot 10^{-3}$ (20)
23	24.36	24.44	2	SNP_A-1998393	0.011	$4.37 \cdot 10^{-4}$	$5.58 \cdot 10^{-2}$	$3.78 \cdot 10^{-7}$	$9.71 \cdot 10^{-4}$ (37)
23	70.69	70.96	6	rs5951179 (*NHSL2*)	0.041	$2.57 \cdot 10^{-1}$	$7.93 \cdot 10^{-1}$	$2.62 \cdot 10^{-4}$	$10.00 \cdot 10^{-3}$ (10)
23	74.34	74.47	6	rs5938070 (*ZDHHC15*)	0.025	1	$6.38 \cdot 10^{-1}$	$8.33 \cdot 10^{-7}$	$7.77 \cdot 10^{-3}$ (13)
2	212.20	212.20	1	rs6435632	0.307	$4.12 \cdot 10^{-3}$	$1.25 \cdot 10^{-2}$	$2.39 \cdot 10^{-2}$	$9.89 \cdot 10^{-5}$ (199)
6	99.62	99.62	1	rs1884184	0.215	$3.16 \cdot 10^{-1}$	$3.98 \cdot 10^{-3}$	$3.07 \cdot 10^{-4}$	$1.44 \cdot 10^{-4}$ (135)
8	123.90	123.90	1	rs10095188	0.176	$8.35 \cdot 10^{-2}$	$1.74 \cdot 10^{-5}$	$1.29 \cdot 10^{-2}$	$4.96 \cdot 10^{-4}$ (55)
10	10.32	10.32	1	rs2895065	0.094	$4.74 \cdot 10^{-2}$	$5.08 \cdot 10^{-3}$	$7.83 \cdot 10^{-2}$	$1.08 \cdot 10^{-4}$ (184)
13	60.33	60.33	1	rs167272	0.095	$5.15 \cdot 10^{-3}$	$1.27 \cdot 10^{-3}$	$2.19 \cdot 10^{-1}$	$9.77 \cdot 10^{-4}$ (36)

T-Trees

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	118.80	118.81	3	rs12134541	0.478	$5.56 \cdot 10^{-1}$	$8.54 \cdot 10^{-1}$	$6.96 \cdot 10^{-1}$	$2.95 \cdot 10^{-4}$ (77)
1	149.24	149.24	2	rs1923496	0.332	$7.54 \cdot 10^{-1}$	$2.34 \cdot 10^{-1}$	$1.24 \cdot 10^{-1}$	$1.55 \cdot 10^{-4}$ (126)
1	235.59	235.59	2	rs16837761	0.165	$8.80 \cdot 10^{-2}$	$3.85 \cdot 10^{-1}$	$6.78 \cdot 10^{-1}$	$1.41 \cdot 10^{-4}$ (139)
2	10.90	10.93	4	rs902133$^{(1)}$	0.006	1	1	$1.41 \cdot 10^{-16}$	$5.24 \cdot 10^{-3}$ (12)
2	17.91	17.91	2	rs13394205	0.004	1	$1.14 \cdot 10^{-1}$	$6.53 \cdot 10^{-9}$	$4.37 \cdot 10^{-4}$ (60)
2	133.02	133.02	4	rs10188442 (*GPR39*)	0.204	$1.65 \cdot 10^{-1}$	1	$7.40 \cdot 10^{-1}$	$6.00 \cdot 10^{-4}$ (46)
3	6.84	6.85	2	rs17046143	0.003	1	1	$3.69 \cdot 10^{-11}$	$2.50 \cdot 10^{-3}$ (20)
3	45.11	45.16	6	rs4683064(*CDCP1*)$^{(1)}$	0.005	1	1	$4.59 \cdot 10^{-16}$	$5.39 \cdot 10^{-3}$ (11)
3	66.60	66.60	2	rs3845903 (*LRIG1*)	0.006	1	1	$3.74 \cdot 10^{-9}$	$5 \cdot 10^{-4}$ (53)
3	147.40	147.40	3	rs16858040 (*PLSCR4*)	0.002	1	1	$1.94 \cdot 10^{-8}$	$7.67 \cdot 10^{-4}$ (38)
4	23.64	23.67	3	rs615604	0.155	$1.61 \cdot 10^{-1}$	$7.30 \cdot 10^{-1}$	$4.96 \cdot 10^{-1}$	$3.41 \cdot 10^{-3}$ (15)
4	116.10	116.16	6	rs7666328(*NDST4*)$^{(3)}$	0.038	$8.57 \cdot 10^{-2}$	$2.82 \cdot 10^{-3}$	$2.60 \cdot 10^{-14}$	$2.62 \cdot 10^{-3}$ (19)
5	36.46	36.46	3	rs12515142	0.140	$6.30 \cdot 10^{-1}$	$5.99 \cdot 10^{-1}$	$4.06 \cdot 10^{-1}$	$2.23 \cdot 10^{-4}$ (92)
6	9.59	9.62	6	rs9357438	0.043	$4.69 \cdot 10^{-2}$	$4.37 \cdot 10^{-2}$	$6.48 \cdot 10^{-1}$	$5.80 \cdot 10^{-3}$ (10)
6	32.86	32.87	4	rs2621382	0.448	$3.36 \cdot 10^{-1}$	$8.53 \cdot 10^{-1}$	$2.15 \cdot 10^{-1}$	$5.82 \cdot 10^{-4}$ (48)
8	15.70	15.70	3	rs2604383$^{(1)}$	0.027	$1.68 \cdot 10^{-1}$	$4.06 \cdot 10^{-1}$	$1.17 \cdot 10^{-3}$	$1.41 \cdot 10^{-4}$ (138)
8	120.42	120.43	4	rs2469997	0.187	$3.35 \cdot 10^{-1}$	$9.04 \cdot 10^{-1}$	$7.92 \cdot 10^{-1}$	$1.16 \cdot 10^{-3}$ (30)
8	135.58	135.67	10	rs1372662 (*ZFAT*)	0.333	$6.85 \cdot 10^{-1}$	$3.42 \cdot 10^{-1}$	$8.79 \cdot 10^{-1}$	$7.48 \cdot 10^{-3}$ (7)
9	117.77	117.81	10	rs2458879	0.386	$3.88 \cdot 10^{-1}$	$8.47 \cdot 10^{-1}$	$4.46 \cdot 10^{-1}$	$9.08 \cdot 10^{-3}$ (5)
10	14.82	14.82	4	rs2601749 (*FAM107B*)	0.401	$7.80 \cdot 10^{-1}$	$1.15 \cdot 10^{-2}$	$3.49 \cdot 10^{-2}$	$2.40 \cdot 10^{-4}$ (87)
10	30.97	31.06	15	rs3122348(*LOC645954*)$^{(1)}$	0.034	$4.73 \cdot 10^{-6}$	1	$6.12 \cdot 10^{-96}$	$5.51 \cdot 10^{-2}$ (1)
10	58.18	58.29	10	rs11005510	0.010	1	1	$1.27 \cdot 10^{-20}$	$7.59 \cdot 10^{-3}$ (6)
10	58.39	58.41	2	rs16909905	0.012	1	$3.53 \cdot 10^{-1}$	$9.25 \cdot 10^{-1}$	$6.72 \cdot 10^{-4}$ (42)
10	131.97	132.00	2	rs7918047	0.058	$9.05 \cdot 10^{-2}$	$4.04 \cdot 10^{-1}$	$8.61 \cdot 10^{-1}$	$1.69 \cdot 10^{-3}$ (25)
11	27.01	27.04	5	rs16916476 (*BBOX1*)	0.005	1	1	$9.29 \cdot 10^{-17}$	$6.64 \cdot 10^{-3}$ (9)
16	56.73	56.75	4	rs6499937 (*CSNK2A2*)	0.006	1	1	$1.30 \cdot 10^{-8}$	$1.05 \cdot 10^{-3}$ (33)
16	59.66	59.67	4	rs2133803	0.459	$6.17 \cdot 10^{-1}$	$1.20 \cdot 10^{-1}$	$1.42 \cdot 10^{-1}$	$1.92 \cdot 10^{-4}$ (108)
17	70.50	70.58	3	rs1873598 (*CDR2L*)	0.002	1	1	$5.42 \cdot 10^{-8}$	$6.61 \cdot 10^{-4}$ (44)
19	22.69	22.70	2	rs12980129	0.028	$5.14 \cdot 10^{-1}$	1	$3.35 \cdot 10^{-4}$	$1.43 \cdot 10^{-4}$ (136)
20	15.55	15.55	2	rs200759 (*MACROD2*)	0.113	$8.23 \cdot 10^{-1}$	$1.69 \cdot 10^{-1}$	$8.71 \cdot 10^{-1}$	$4.22 \cdot 10^{-4}$ (61)
22	35.92	35.99	8	rs10212068$^{(3)}$	0.028	1	$2.60 \cdot 10^{-3}$	$7.43 \cdot 10^{-62}$	$3.10 \cdot 10^{-2}$ (2)
23	24.36	24.44	2	SNP_A-1998393	0.011	$4.37 \cdot 10^{-4}$	$5.58 \cdot 10^{-2}$	$3.78 \cdot 10^{-7}$	$8.15 \cdot 10^{-4}$ (37)
23	70.86	70.96	8	rs5951179 (*NHSL2*)	0.041	$2.57 \cdot 10^{-1}$	$7.93 \cdot 10^{-1}$	$2.62 \cdot 10^{-4}$	$3.40 \cdot 10^{-3}$ (16)
23	74.40	74.45	4	rs4892579 (*ZDHHC15*)	0.027	$3.72 \cdot 10^{-1}$	$6.38 \cdot 10^{-1}$	$3.33 \cdot 10^{-3}$	$7.56 \cdot 10^{-4}$ (39)
13	60.33	60.33	1	rs167272	0.095	$5.15 \cdot 10^{-3}$	$1.27 \cdot 10^{-3}$	$2.19 \cdot 10^{-1}$	$1.08 \cdot 10^{-4}$ (167)
14	60.77	60.77	1	rs4902035	0.096	$7.15 \cdot 10^{-3}$	$1.01 \cdot 10^{-4}$	$5.53 \cdot 10^{-1}$	$1.47 \cdot 10^{-4}$ (132)

TABLE 7.12 HT_{qc}: lists of regions identified by the Random Forests and the T-Trees methods.

Random Forests

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	219.90	219.97	6	rs825148	0.053	$2.41 \cdot 10^{-210}$	$1.38 \cdot 10^{-2}$	$5.21 \cdot 10^{-13}$	$4.51 \cdot 10^{-2}$ (3)
2	96.53	96.54	3	rs1870340	0.027	$3.16 \cdot 10^{-122}$	$6.30 \cdot 10^{-1}$	$4.89 \cdot 10^{-8}$	$1.58 \cdot 10^{-2}$ (9)
2	127.16	127.17	2	rs935019 (*GYPC*)	0.261	$5.86 \cdot 10^{-13}$	$8.87 \cdot 10^{-1}$	$9.61 \cdot 10^{-3}$	$1.03 \cdot 10^{-3}$ (74)
4	6.06	6.08	11	rs16837871	0.148	$8.62 \cdot 10^{-23}$	$2.84 \cdot 10^{-1}$	$1.75 \cdot 10^{-28}$	$1.56 \cdot 10^{-2}$ (10)
4	17.85	17.92	11	rs1553460	0.310	$1.42 \cdot 10^{-94}$	$2.88 \cdot 10^{-5}$	$2.42 \cdot 10^{-24}$	$2.48 \cdot 10^{-2}$ (6)
4	141.58	141.62	5	rs6840033 (*LOC100129858, SCOC*)	0.206	$4.64 \cdot 10^{-19}$	$9.16 \cdot 10^{-1}$	$2.35 \cdot 10^{-11}$	$3.09 \cdot 10^{-3}$ (30)
6	107.25	107.26	2	rs10499044	0.105	$2.37 \cdot 10^{-18}$	$1.97 \cdot 10^{-3}$	$6.60 \cdot 10^{-13}$	$2.76 \cdot 10^{-3}$ (34)
7	120.71	120.72	16	rs1528356	0.066	$7.10 \cdot 10^{-264}$	$4.00 \cdot 10^{-3}$	$3.20 \cdot 10^{-22}$	$5.94 \cdot 10^{-2}$ (1)
8	15.34	15.35	2	rs7837736	0.122	$2.92 \cdot 10^{-75}$	$1.45 \cdot 10^{-1}$	$1.35 \cdot 10^{-102}$	$5.90 \cdot 10^{-2}$ (2)
9	1.79	1.81	4	rs17797701	0.014	$1.99 \cdot 10^{-9}$	$4.03 \cdot 10^{-1}$	$1.43 \cdot 10^{-20}$	$9.29 \cdot 10^{-3}$ (13)
10	58.39	58.41	2	rs2393191	0.016	$3.11 \cdot 10^{-1}$	$5.76 \cdot 10^{-1}$	$2.46 \cdot 10^{-1}$	$7.46 \cdot 10^{-4}$ (92)
11	113.02	113.31	21	rs17116117 (*HTR3B*)	0.053	$9.74 \cdot 10^{-6}$	$1.16 \cdot 10^{-1}$	$1.16 \cdot 10^{-31}$	$1.33 \cdot 10^{-2}$ (12)
12	30.25	30.28	11	rs10843660	0.380	$2.74 \cdot 10^{-65}$	$3.54 \cdot 10^{-2}$	$9.18 \cdot 10^{-35}$	$3.06 \cdot 10^{-2}$ (5)
13	90.79	90.83	4	rs17667894	0.020	1	$3.78 \cdot 10^{-1}$	$2.82 \cdot 10^{-42}$	$2.17 \cdot 10^{-2}$ (7)
15	77.69	77.69	2	rs2865199	0.013	1	$6.26 \cdot 10^{-1}$	$2.35 \cdot 10^{-13}$	$3.20 \cdot 10^{-3}$ (29)
16	56.75	56.79	2	rs6499937 (*CSNK2A2*)	0.006	1	1	$2.20 \cdot 10^{-8}$	$1.16 \cdot 10^{-3}$ (68)
16	79.96	79.98	5	rs16955238	0.027	$4.82 \cdot 10^{-114}$	$6.46 \cdot 10^{-1}$	$2.45 \cdot 10^{-5}$	$1.52 \cdot 10^{-2}$ (11)
17	17.27	17.27	2	SNP_A-1948953	0.279	$2.84 \cdot 10^{-26}$	$5.18 \cdot 10^{-3}$	$2.10 \cdot 10^{-6}$	$5.45 \cdot 10^{-3}$ (23)
22	27.70	27.75	6	rs8137391 (*ZNRF3, ZNRF3-AS1*)	0.010	1	1	$1.49 \cdot 10^{-19}$	$8.68 \cdot 10^{-3}$ (16)
23	70.95	70.96	2	rs5951179 (*NHSL2*)	0.040	$2.55 \cdot 10^{-1}$	1	$3.44 \cdot 10^{-4}$	$3.88 \cdot 10^{-3}$ (26)
8	123.90	123.90	1	rs10095188	0.175	$7.09 \cdot 10^{-2}$	$3.01 \cdot 10^{-5}$	$3.17 \cdot 10^{-2}$	$1.31 \cdot 10^{-4}$ (164)
13	60.33	60.33	1	rs167272	0.095	$5.08 \cdot 10^{-3}$	$1.68 \cdot 10^{-3}$	$1.91 \cdot 10^{-1}$	$2.17 \cdot 10^{-4}$ (138)
14	60.77	60.77	1	rs4902035	0.096	$6.77 \cdot 10^{-3}$	$1.38 \cdot 10^{-4}$	$5.05 \cdot 10^{-1}$	$1.39 \cdot 10^{-4}$ (160)

T-Trees

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	80.57	80.57	6	rs1896250	0.397	$6.97 \cdot 10^{-8}$	$9.37 \cdot 10^{-1}$	$8.26 \cdot 10^{-5}$	$1.14 \cdot 10^{-3}$ (54)
1	219.91	219.97	5	rs825148	0.053	$2.41 \cdot 10^{-210}$	$1.38 \cdot 10^{-2}$	$5.21 \cdot 10^{-13}$	$3.10 \cdot 10^{-2}$ (2)
2	25.31	25.36	3	rs2164411	0.161	$3.20 \cdot 10^{-7}$	$1.95 \cdot 10^{-1}$	$4.75 \cdot 10^{-1}$	$4.17 \cdot 10^{-4}$ (107)
2	96.53	96.54	3	rs1870340	0.027	$3.16 \cdot 10^{-122}$	$6.30 \cdot 10^{-1}$	$4.89 \cdot 10^{-8}$	$7.97 \cdot 10^{-3}$ (16)
2	127.16	127.17	4	rs935019 (*GYPC*)	0.261	$5.86 \cdot 10^{-13}$	$8.87 \cdot 10^{-1}$	$9.61 \cdot 10^{-3}$	$9.25 \cdot 10^{-4}$ (66)
2	213.89	213.89	2	rs12694298	0.105	$6.20 \cdot 10^{-1}$	$2.79 \cdot 10^{-1}$	$7.59 \cdot 10^{-1}$	$2.48 \cdot 10^{-4}$ (160)
3	121.76	121.83	3	rs804974 (*HGD*)	0.215	$9.48 \cdot 10^{-1}$	$8.68 \cdot 10^{-1}$	$2.00 \cdot 10^{-1}$	$1.58 \cdot 10^{-3}$ (43)
4	6.04	6.06	6	rs16837871	0.148	$8.62 \cdot 10^{-23}$	$2.84 \cdot 10^{-1}$	$1.75 \cdot 10^{-28}$	$1.06 \cdot 10^{-2}$ (11)
4	17.85	17.90	9	rs1553460	0.310	$1.42 \cdot 10^{-94}$	$2.88 \cdot 10^{-5}$	$2.42 \cdot 10^{-24}$	$1.24 \cdot 10^{-2}$ (8)
4	23.66	23.67	3	rs615604	0.155	$1.57 \cdot 10^{-1}$	$8.35 \cdot 10^{-1}$	$4.24 \cdot 10^{-1}$	$7.99 \cdot 10^{-4}$ (74)
4	141.58	141.62	8	rs6840033 (*LOC100129858, SCOC*)	0.206	$4.64 \cdot 10^{-19}$	$9.16 \cdot 10^{-1}$	$2.35 \cdot 10^{-11}$	$3.18 \cdot 10^{-3}$ (31)
6	18.17	18.19	4	rs4072775	0.348	$3.03 \cdot 10^{-7}$	$2.71 \cdot 10^{-6}$	$9.05 \cdot 10^{-2}$	$1.15 \cdot 10^{-3}$ (53)
6	91.73	91.73	2	rs6903505	0.405	$3.73 \cdot 10^{-1}$	$7.59 \cdot 10^{-1}$	$7.68 \cdot 10^{-1}$	$2.32 \cdot 10^{-4}$ (167)
6	99.33	99.36	5	rs4131463	0.050	$5.27 \cdot 10^{-202}$	$1.82 \cdot 10^{-1}$	$7.38 \cdot 10^{-14}$	$2.79 \cdot 10^{-2}$ (3)
6	107.25	107.26	3	rs10499044	0.105	$2.37 \cdot 10^{-18}$	$1.97 \cdot 10^{-3}$	$6.60 \cdot 10^{-13}$	$1.49 \cdot 10^{-3}$ (47)
7	120.71	120.72	14	rs1528356	0.066	$7.10 \cdot 10^{-264}$	$4.00 \cdot 10^{-3}$	$3.20 \cdot 10^{-22}$	$3.75 \cdot 10^{-2}$ (1)
8	15.34	15.37	6	rs7837736	0.122	$2.92 \cdot 10^{-75}$	$1.45 \cdot 10^{-1}$	$1.35 \cdot 10^{-102}$	$2.63 \cdot 10^{-2}$ (4)
8	95.09	95.09	2	rs3018857	0.065	$8.46 \cdot 10^{-1}$	$6.57 \cdot 10^{-1}$	$4.03 \cdot 10^{-1}$	$5.55 \cdot 10^{-4}$ (92)
8	114.39	114.43	2	rs7012271 (*CSMD3*)	0.316	$9.17 \cdot 10^{-1}$	$9.32 \cdot 10^{-1}$	$9.12 \cdot 10^{-1}$	$6.21 \cdot 10^{-4}$ (85)
8	120.42	120.42	2	rs2469997	0.187	$2.66 \cdot 10^{-1}$	$8.55 \cdot 10^{-1}$	$7.70 \cdot 10^{-1}$	$7.68 \cdot 10^{-4}$ (75)
8	135.63	135.65	7	rs1372662 (*ZFAT*)	0.333	$5.76 \cdot 10^{-1}$	$3.39 \cdot 10^{-1}$	$8.61 \cdot 10^{-1}$	$4.78 \cdot 10^{-3}$ (23)
9	1.79	1.81	4	rs17797701	0.014	$1.99 \cdot 10^{-9}$	$4.03 \cdot 10^{-1}$	$1.43 \cdot 10^{-20}$	$6.60 \cdot 10^{-3}$ (18)
9	106.56	106.56	3	rs2035783	0.473	$3.63 \cdot 10^{-3}$	$9.15 \cdot 10^{-2}$	$2.99 \cdot 10^{-3}$	$2.86 \cdot 10^{-4}$ (142)
9	117.79	117.81	3	rs2151370	0.379	$7.35 \cdot 10^{-1}$	$8.45 \cdot 10^{-1}$	$1.87 \cdot 10^{-1}$	$2.63 \cdot 10^{-3}$ (35)
10	58.20	58.21	3	rs11005510	0.010	1	1	$1.23 \cdot 10^{-20}$	$3.68 \cdot 10^{-3}$ (28)
10	58.39	58.41	2	rs2393191	0.016	$3.11 \cdot 10^{-1}$	$5.76 \cdot 10^{-1}$	$2.46 \cdot 10^{-1}$	$2.72 \cdot 10^{-4}$ (152)
11	113.28	113.31	6	rs1176741 (*HTR3B*)	0.031	$2.59 \cdot 10^{-1}$	$1.19 \cdot 10^{-1}$	$8.13 \cdot 10^{-1}$	$1.37 \cdot 10^{-2}$ (7)
12	30.24	30.26	9	rs10843660	0.380	$2.74 \cdot 10^{-65}$	$3.54 \cdot 10^{-2}$	$9.18 \cdot 10^{-35}$	$1.74 \cdot 10^{-2}$ (5)
12	68.67	68.67	2	rs10879068	0.300	$3.12 \cdot 10^{-1}$	$9.29 \cdot 10^{-1}$	$2.24 \cdot 10^{-1}$	$5.58 \cdot 10^{-4}$ (91)
12	124.76	124.76	2	rs16919463	0.118	$8.30 \cdot 10^{-1}$	$7.87 \cdot 10^{-1}$	$5.42 \cdot 10^{-1}$	$1.54 \cdot 10^{-3}$ (44)
13	90.79	90.83	6	rs17667894	0.020	1	$3.78 \cdot 10^{-1}$	$2.82 \cdot 10^{-42}$	$1.19 \cdot 10^{-2}$ (9)
15	77.69	77.69	2	rs16971150	0.018	$4.58 \cdot 10^{-1}$	$6.26 \cdot 10^{-1}$	$7.01 \cdot 10^{-1}$	$5.11 \cdot 10^{-3}$ (21)
16	56.75	56.75	3	rs6499937 (*CSNK2A2*)	0.006	1	1	$2.20 \cdot 10^{-8}$	$4.67 \cdot 10^{-4}$ (101)
16	79.96	79.98	6	rs16955238	0.027	$4.82 \cdot 10^{-114}$	$6.46 \cdot 10^{-1}$	$2.45 \cdot 10^{-5}$	$1.01 \cdot 10^{-2}$ (13)
17	17.27	17.27	2	SNP_A-1948953	0.279	$2.84 \cdot 10^{-26}$	$5.18 \cdot 10^{-3}$	$2.10 \cdot 10^{-6}$	$1.52 \cdot 10^{-3}$ (46)
17	50.27	50.29	2	rs2934884	0.196	$1.05 \cdot 10^{-1}$	$4.16 \cdot 10^{-1}$	$3.48 \cdot 10^{-1}$	$5.87 \cdot 10^{-4}$ (90)
18	43.91	43.93	3	rs8085875 (*ZBTB7C*)	0.174	$3.52 \cdot 10^{-1}$	$7.95 \cdot 10^{-1}$	$4.30 \cdot 10^{-1}$	$9.23 \cdot 10^{-4}$ (67)
22	27.75	27.75	2	rs16986990 (*ZNRF3, ZNRF3-AS1*)	0.015	1	1	$5.57 \cdot 10^{-1}$	$5.08 \cdot 10^{-3}$ (22)
23	2.58	2.58	2	rs1419930	0.050	$5.36 \cdot 10^{-1}$	$5.71 \cdot 10^{-19}$	$1.22 \cdot 10^{-7}$	$4.57 \cdot 10^{-3}$ (25)
23	70.95	70.96	2	rs5951179 (*NHSL2*)	0.040	$2.55 \cdot 10^{-1}$	1	$3.44 \cdot 10^{-4}$	$1.42 \cdot 10^{-3}$ (48)
23	74.42	74.43	2	rs4892579 (*ZDHHC15*)	0.028	$3.74 \cdot 10^{-1}$	$6.40 \cdot 10^{-1}$	$2.65 \cdot 10^{-3}$	$2.16 \cdot 10^{-4}$ (178)
10	10.32	10.32	1	rs1333834	0.064	$5.64 \cdot 10^{-3}$	$2.64 \cdot 10^{-6}$	$3.99 \cdot 10^{-1}$	$3.76 \cdot 10^{-4}$ (114)

TABLE 7.13 HT_{wtccc}: lists of regions identified by the Random Forests and the T-Trees methods.

Rheumatoid arthritis

For this disease, the WTCCC reports a large region named MHC (major histocompatibility complex) which encapsulates a family of genes implicated in autoimmune diseases. In our four experiments, that region is well detected, since many markers in our four top rankings appear to be located on chromosome 6 in the MHC region. There are more of them in the 200 first on RA_{qc} and also more of them with the Random Forests. In addition, rs6679677 reported in [Wel07] supplementary information is also detected in the 100 first variables in 3 experiments out of four.

We also notice the presence of many SNPs located on chromosome X especially on the qc version. Like the CAD dataset, there is a gender disproportion in the RA dataset (about 500 males and 1500 females).

And again, we note the presence of rs3122348 and rs10212068 in the qc version. And a SNP: rs3785579 reported in [P+12a] as a "hub" in the RA_{wtccc} version.

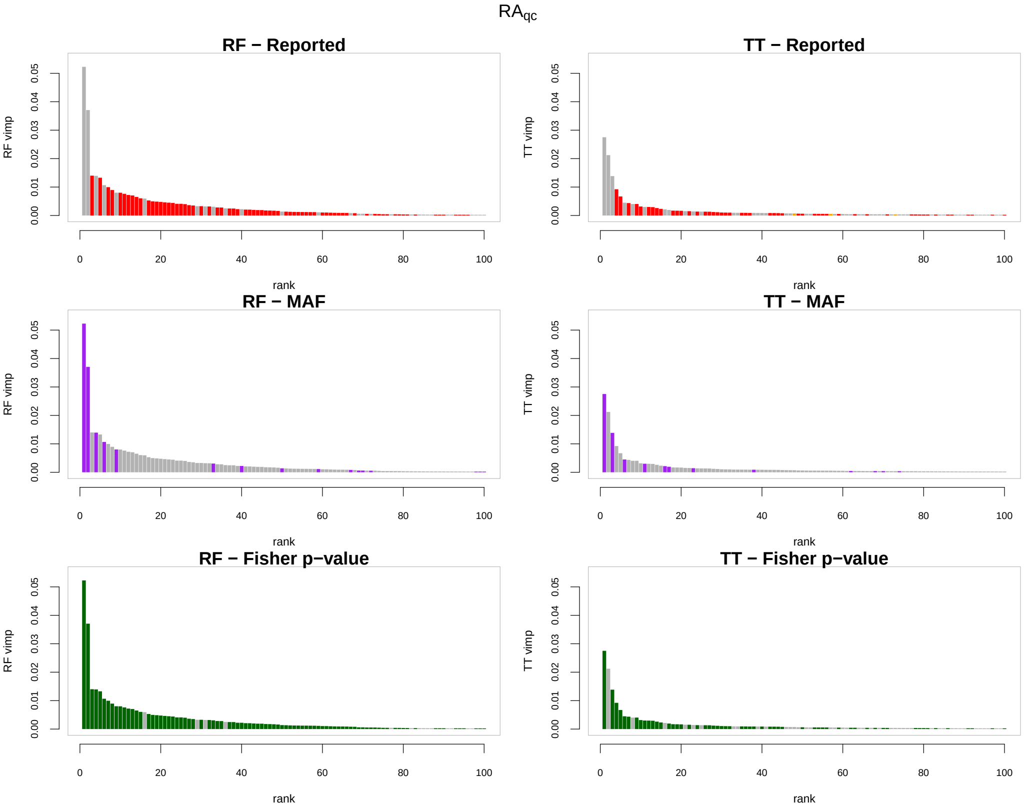


FIGURE 7.10 The first 100 variables according to the tree based importance rankings for RA_{qc}. The horizontal axis corresponds to the ranks and the vertical axis to the variable importances. In the first column variables are ordered according to random forests variable importances and in the second column they are ordered according to the T-Trees variable importances. In the first row, red highlights the reported strongly associated regions. In the second row, purple corresponds to rare variants (MAF < 0.05). In the third row, green represents markers with a low Fisher p-value ($< 10^{-6}$).

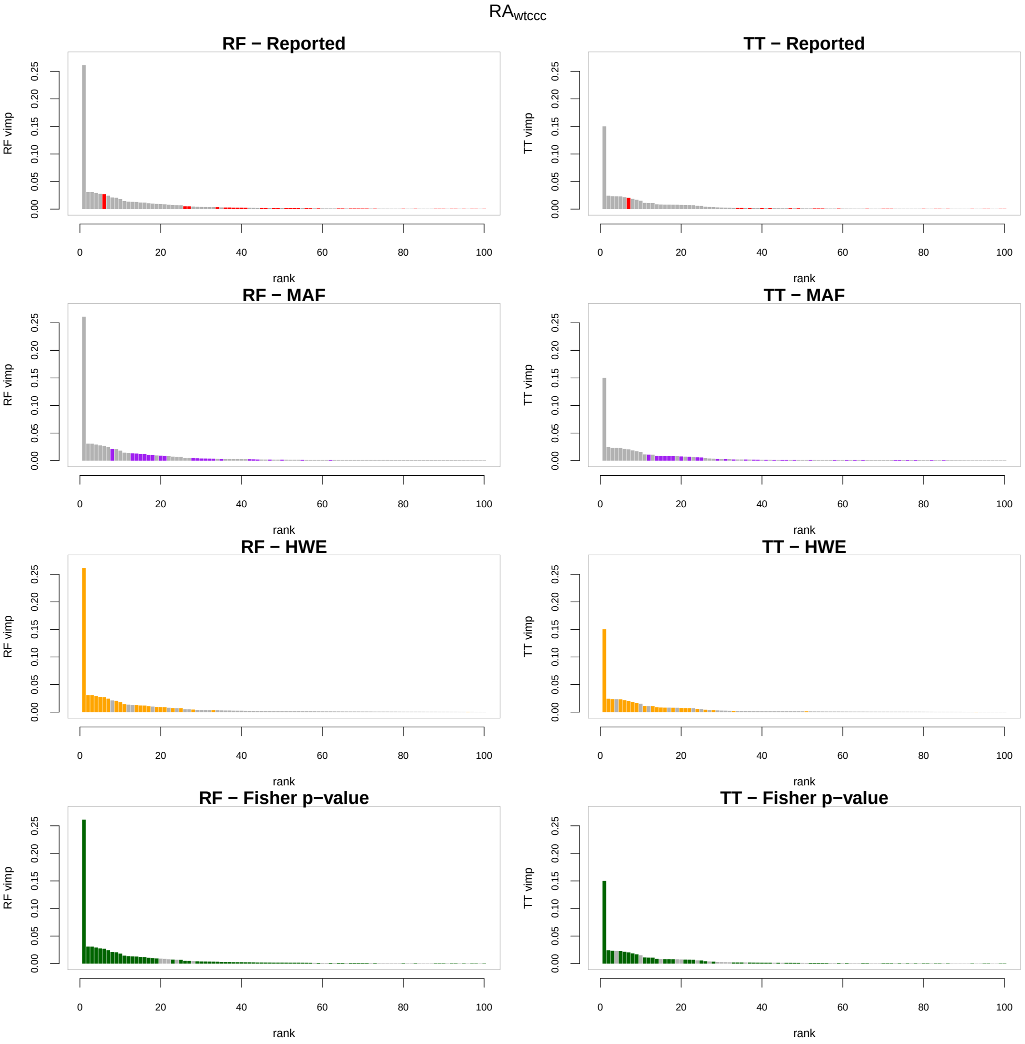


FIGURE 7.11 The first 100 variables according to the tree based importance rankings for RA_{wtccc}. The horizontal axis corresponds to the ranks and the vertical axis to the variable importances. In the first column variables are ordered according to random forests variable importances and in the second column they are ordered according to the T-Trees variable importances. In the first row, red highlights the strongly associated reported regions. In the second row, purple corresponds to rare variants (MAF < 0.05). In the third row, orange highlights SNPs deviating from HWE and in the last row, green represents markers with a low Fisher p-value ($< 10^{-6}$).

Random Forests

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	113.89	114.02	2	rs6679677	0.124	$5.01 \cdot 10^{-2}$	$2.41 \cdot 10^{-1}$	$5.47 \cdot 10^{-26}$	$3.63 \cdot 10^{-3}$ (27)
5	143.36	143.37	2	rs160619$^{(1)}$	0.014	1	$4.05 \cdot 10^{-1}$	$2.78 \cdot 10^{-27}$	$1.07 \cdot 10^{-2}$ (6)
6	31.70	32.82	94	rs6457620	0.433	$5.62 \cdot 10^{-1}$	$5.33 \cdot 10^{-1}$	$1.53 \cdot 10^{-83}$	$1.40 \cdot 10^{-2}$ (3)
10	30.99	31.06	6	rs3122348(*LOC645954*)$^{(1)}$	0.029	$1.65 \cdot 10^{-4}$	1	$3.64 \cdot 10^{-78}$	$5.23 \cdot 10^{-2}$ (1)
23	3.02	3.02	2	rs5982644 (*ARSF*)	0.082	$7.13 \cdot 10^{-2}$	$3.89 \cdot 10^{-1}$	$3.62 \cdot 10^{-1}$	$1.39 \cdot 10^{-4}$ (137)
23	9.08	9.08	2	rs16985421	0.190	$4.41 \cdot 10^{-1}$	$1.49 \cdot 10^{-2}$	$6.22 \cdot 10^{-1}$	$1.21 \cdot 10^{-4}$ (156)
23	9.95	10.04	3	rs7053877	0.127	$3.95 \cdot 10^{-1}$	$9.29 \cdot 10^{-2}$	$3.92 \cdot 10^{-1}$	$1.27 \cdot 10^{-4}$ (146)
23	12.71	12.77	2	rs11797883	0.117	1	$6.18 \cdot 10^{-1}$	$1.93 \cdot 10^{-2}$	$4.29 \cdot 10^{-4}$ (77)
23	14.23	14.50	5	rs5934184 (*GLRA2*)	0.239	$4.13 \cdot 10^{-1}$	$5.79 \cdot 10^{-1}$	$4.65 \cdot 10^{-3}$	$3.48 \cdot 10^{-4}$ (82)
23	20.21	20.40	3	rs5950315	0.242	$9.43 \cdot 10^{-1}$	1	$1.43 \cdot 10^{-1}$	$2.31 \cdot 10^{-4}$ (104)
23	22.09	22.10	2	rs4824185 (*LOC100873065*)	0.222	$9.38 \cdot 10^{-1}$	1	$1.94 \cdot 10^{-1}$	$1.29 \cdot 10^{-4}$ (142)
23	35.97	36.03	3	rs17273161	0.104	$5.25 \cdot 10^{-1}$	$6.78 \cdot 10^{-1}$	$1.49 \cdot 10^{-3}$	$1.56 \cdot 10^{-4}$ (127)
23	41.73	42.01	2	rs5918362	0.133	$3.45 \cdot 10^{-1}$	$9.15 \cdot 10^{-1}$	$2.27 \cdot 10^{-2}$	$1.50 \cdot 10^{-4}$ (130)
23	44.87	45.31	2	rs1883678	0.041	$6.99 \cdot 10^{-2}$	$4.80 \cdot 10^{-1}$	$8.19 \cdot 10^{-1}$	$1.65 \cdot 10^{-4}$ (122)
23	122.51	122.53	2	rs6648534 (*THOC2*)	0.283	$8.36 \cdot 10^{-2}$	$2.65 \cdot 10^{-1}$	$2.78 \cdot 10^{-2}$	$1.28 \cdot 10^{-4}$ (144)
23	135.96	135.96	3	rs1930218	0.241	$1.02 \cdot 10^{-1}$	$4.43 \cdot 10^{-2}$	$4.87 \cdot 10^{-2}$	$3.30 \cdot 10^{-4}$ (84)
6	31.35	31.35	1	rs3132486	0.427	$2.09 \cdot 10^{-1}$	$2.82 \cdot 10^{-1}$	$2.12 \cdot 10^{-14}$	$1.74 \cdot 10^{-4}$ (121)
6	33.07	33.07	1	rs3128947	0.215	$2.33 \cdot 10^{-1}$	$3.36 \cdot 10^{-1}$	$9.57 \cdot 10^{-16}$	$1.25 \cdot 10^{-4}$ (147)
21	41.43	41.43	1	rs2837960	0.179	$7.53 \cdot 10^{-4}$	$1.87 \cdot 10^{-4}$	$3.70 \cdot 10^{-2}$	$2.54 \cdot 10^{-4}$ (96)

T-Trees

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	153.90	153.96	4	rs16838195	0.010	$6.28 \cdot 10^{-1}$	1	$8.69 \cdot 10^{-34}$	$4.50 \cdot 10^{-3}$ (6)
2	10.90	10.91	2	rs902133$^{(1)}$	0.006	1	1	$8.41 \cdot 10^{-17}$	$8.99 \cdot 10^{-4}$ (38)
2	232.64	232.71	2	rs10460436	0.021	1	1	$7.75 \cdot 10^{-1}$	$2.12 \cdot 10^{-4}$ (113)
3	7.26	7.30	7	rs1605705 (*GRM7*)	0.178	$4.14 \cdot 10^{-1}$	$3.00 \cdot 10^{-1}$	$1.48 \cdot 10^{-2}$	$4.04 \cdot 10^{-3}$ (8)
3	141.43	141.44	4	rs9856460 (*CLSTN2*)	0.451	$4.92 \cdot 10^{-2}$	$8.53 \cdot 10^{-1}$	$2.81 \cdot 10^{-1}$	$1.54 \cdot 10^{-3}$ (21)
4	182.12	182.13	4	rs4532278	0.209	1	$5.44 \cdot 10^{-1}$	$7.03 \cdot 10^{-1}$	$7.24 \cdot 10^{-4}$ (48)
5	143.36	143.37	2	rs160619$^{(1)}$	0.014	1	$4.05 \cdot 10^{-1}$	$2.78 \cdot 10^{-27}$	$3.01 \cdot 10^{-3}$ (11)
5	150.15	150.21	6	rs4246045$^{(3)}$	0.142	$1.01 \cdot 10^{-1}$	$5.99 \cdot 10^{-5}$	$2.12 \cdot 10^{-3}$	$7.33 \cdot 10^{-4}$ (47)
6	31.73	32.80	79	rs6457617	0.434	$5.98 \cdot 10^{-1}$	$5.33 \cdot 10^{-1}$	$2.10 \cdot 10^{-83}$	$9.22 \cdot 10^{-3}$ (4)
6	32.87	32.87	2	rs2157082$^{(3)}$	0.443	$8.90 \cdot 10^{-1}$	$3.95 \cdot 10^{-4}$	$4.80 \cdot 10^{-2}$	$3.03 \cdot 10^{-4}$ (87)
8	107.27	107.29	3	rs16874204	0.043	$2.58 \cdot 10^{-1}$	$8.23 \cdot 10^{-1}$	$9.19 \cdot 10^{-1}$	$2.15 \cdot 10^{-3}$ (16)
10	30.97	31.04	11	rs3122348(*LOC645954*)$^{(1)}$	0.029	$1.65 \cdot 10^{-4}$	1	$3.64 \cdot 10^{-78}$	$2.75 \cdot 10^{-2}$ (1)
10	114.07	114.08	2	rs4256909 (*GUCY2GP*)	0.083	$2.68 \cdot 10^{-3}$	$4.49 \cdot 10^{-1}$	$7.08 \cdot 10^{-2}$	$5.17 \cdot 10^{-4}$ (60)
10	131.97	132.00	3	rs7080464$^{(1)}$	0.014	1	$4.05 \cdot 10^{-1}$	$8.35 \cdot 10^{-24}$	$1.94 \cdot 10^{-3}$ (17)
12	42.00	42.01	2	rs6582454	0.464	$3.86 \cdot 10^{-1}$	$9.41 \cdot 10^{-1}$	$2.64 \cdot 10^{-1}$	$1.54 \cdot 10^{-4}$ (134)
15	51.49	51.49	4	rs1711029	0.043	$5.47 \cdot 10^{-1}$	$1.22 \cdot 10^{-1}$	$2.25 \cdot 10^{-11}$	$4.05 \cdot 10^{-4}$ (70)
15	71.02	71.02	3	rs4777568	0.028	1	$2.91 \cdot 10^{-1}$	$8.50 \cdot 10^{-1}$	$1.29 \cdot 10^{-4}$ (154)
16	51.30	51.31	4	rs4238755	0.257	$6.36 \cdot 10^{-1}$	$1.01 \cdot 10^{-1}$	$6.20 \cdot 10^{-1}$	$1.36 \cdot 10^{-3}$ (25)
22	35.97	36.02	9	rs10212068$^{(3)}$	0.028	1	$2.64 \cdot 10^{-3}$	$1.05 \cdot 10^{-61}$	$1.39 \cdot 10^{-2}$ (3)
23	9.28	9.34	3	rs2521413 (*TBL1X*)	0.401	$8.29 \cdot 10^{-1}$	$2.53 \cdot 10^{-5}$	$1.13 \cdot 10^{-2}$	$2.77 \cdot 10^{-4}$ (93)
23	32.12	32.13	2	rs5927971 (*DMD*)	0.421	$1.31 \cdot 10^{-2}$	$3.98 \cdot 10^{-1}$	$3.54 \cdot 10^{-1}$	$1.83 \cdot 10^{-4}$ (121)
23	45.31	45.31	2	rs1883678	0.041	$6.99 \cdot 10^{-2}$	$4.80 \cdot 10^{-1}$	$8.19 \cdot 10^{-1}$	$1.49 \cdot 10^{-4}$ (139)
23	90.14	90.14	2	rs932574	0.438	$1.10 \cdot 10^{-1}$	$8.92 \cdot 10^{-3}$	$6.46 \cdot 10^{-1}$	$2.38 \cdot 10^{-4}$ (105)
23	96.83	96.85	3	rs2497903	0.484	$5.78 \cdot 10^{-2}$	$8.06 \cdot 10^{-2}$	$8.32 \cdot 10^{-2}$	$2.67 \cdot 10^{-4}$ (96)
23	99.46	99.48	2	rs5920824 (*PCDH19*)	0.385	$1.24 \cdot 10^{-1}$	$3.25 \cdot 10^{-2}$	$1.07 \cdot 10^{-1}$	$1.45 \cdot 10^{-4}$ (143)
1	114.02	114.02	1	rs6679677	0.124	$5.01 \cdot 10^{-2}$	$2.41 \cdot 10^{-1}$	$5.47 \cdot 10^{-26}$	$6.33 \cdot 10^{-4}$ (50)

TABLE 7.14 RA_{qc}: lists of regions identified by the Random Forests and the T-Trees methods.

Random Forests

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	15.21	15.24	4	rs7539166 (*TMEM51*)	0.067	$2.27 \cdot 10^{-243}$	$6.30 \cdot 10^{-3}$	$2.39 \cdot 10^{-14}$	$2.06 \cdot 10^{-2}$ (9)
1	113.89	114.02	2	rs6679677	0.124	$1.23 \cdot 10^{-2}$	$2.41 \cdot 10^{-1}$	$1.12 \cdot 10^{-24}$	$8.16 \cdot 10^{-4}$ (80)
2	105.54	105.55	3	rs13425033	0.069	$8.86 \cdot 10^{-264}$	$2.60 \cdot 10^{-3}$	$1.40 \cdot 10^{-25}$	$2.45 \cdot 10^{-2}$ (7)
2	223.66	223.73	7	rs1440065	0.072	$2.00 \cdot 10^{-274}$	$9.17 \cdot 10^{-3}$	$5.40 \cdot 10^{-26}$	$2.77 \cdot 10^{-2}$ (5)
3	71.93	71.94	3	rs17665418	0.084	$8.50 \cdot 10^{-285}$	$1.58 \cdot 10^{-5}$	$8.54 \cdot 10^{-21}$	$3.09 \cdot 10^{-2}$ (2)
4	17.85	17.87	3	rs1553460	0.308	$1.28 \cdot 10^{-104}$	$2.88 \cdot 10^{-5}$	$1.16 \cdot 10^{-22}$	$7.29 \cdot 10^{-3}$ (23)
6	31.17	31.22	3	rs4959053 (*PSORS1C1*)	0.074	$7.30 \cdot 10^{-273}$	$3.03 \cdot 10^{-4}$	$1.12 \cdot 10^{-22}$	$2.70 \cdot 10^{-2}$ (6)
6	31.73	32.79	75	rs6457620	0.435	$7.47 \cdot 10^{-1}$	$6.05 \cdot 10^{-1}$	$1.10 \cdot 10^{-80}$	$5.17 \cdot 10^{-3}$ (26)
7	66.40	66.45	7	rs4718582	0.056	$2.34 \cdot 10^{-2}$	$3.99 \cdot 10^{-1}$	$3.06 \cdot 10^{-75}$	$1.36 \cdot 10^{-2}$ (12)
7	157.71	157.73	6	rs7789415 (*PTPRN2*)	0.043	1	$2.55 \cdot 10^{-6}$	$2.45 \cdot 10^{-66}$	$1.29 \cdot 10^{-2}$ (14)
9	24.52	24.61	7	rs16908561	0.041	$1.22 \cdot 10^{-159}$	$4.91 \cdot 10^{-2}$	$1.35 \cdot 10^{-4}$	$8.75 \cdot 10^{-3}$ (21)
11	93.02	93.03	2	rs10501805	0.033	$1.06 \cdot 10^{-4}$	1	$1.73 \cdot 10^{-62}$	$1.01 \cdot 10^{-2}$ (18)
12	30.25	30.26	4	rs10843660	0.384	$2.24 \cdot 10^{-75}$	$3.54 \cdot 10^{-2}$	$4.35 \cdot 10^{-30}$	$7.01 \cdot 10^{-3}$ (25)
15	27.77	27.85	2	rs899848 (*TJP1*)	0.027	$5.43 \cdot 10^{-112}$	$6.33 \cdot 10^{-1}$	$1.51 \cdot 10^{-7}$	$3.61 \cdot 10^{-3}$ (33)
17	62.46	62.65	9	rs3785579 (*CACNG1*)	0.271	0	$2.12 \cdot 10^{-2}$	0	$2.61 \cdot 10^{-1}$ (1)
18	33.34	33.36	5	rs4799934 (*CELF4*)	0.078	$3.88 \cdot 10^{-286}$	$2.83 \cdot 10^{-3}$	$8.28 \cdot 10^{-27}$	$3.09 \cdot 10^{-2}$ (3)
23	1.02	1.80	3	rs6588810	0.015	$2.76 \cdot 10^{-6}$	$2.22 \cdot 10^{-5}$	$1.73 \cdot 10^{-8}$	$1.93 \cdot 10^{-3}$ (47)
23	14.23	14.50	5	rs5934184 (*GLRA2*)	0.239	$4.45 \cdot 10^{-1}$	$5.31 \cdot 10^{-1}$	$1.84 \cdot 10^{-3}$	$1.08 \cdot 10^{-4}$ (162)
6	31.28	31.28	1	rs9295961	0.062	$1.11 \cdot 10^{-1}$	$9.78 \cdot 10^{-2}$	$6.59 \cdot 10^{-1}$	$8.72 \cdot 10^{-5}$ (174)

T-Trees

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	15.23	15.24	3	rs7539166 (*TMEM51*)	0.067	$2.27 \cdot 10^{-243}$	$6.30 \cdot 10^{-3}$	$2.39 \cdot 10^{-14}$	$1.69 \cdot 10^{-2}$ (9)
1	87.97	87.97	2	rs4655988	0.237	$9.50 \cdot 10^{-1}$	$5.00 \cdot 10^{-1}$	$8.89 \cdot 10^{-2}$	$3.02 \cdot 10^{-4}$ (143)
1	205.19	205.19	2	rs17012953	0.019	$7.14 \cdot 10^{-2}$	1	$4.51 \cdot 10^{-35}$	$1.85 \cdot 10^{-3}$ (38)
1	209.90	209.92	2	rs1187768	0.310	$3.92 \cdot 10^{-1}$	$3.16 \cdot 10^{-1}$	$1.89 \cdot 10^{-1}$	$1.08 \cdot 10^{-3}$ (60)
2	105.54	105.55	3	rs13425033	0.069	$8.86 \cdot 10^{-264}$	$2.60 \cdot 10^{-3}$	$1.40 \cdot 10^{-25}$	$1.86 \cdot 10^{-2}$ (8)
2	126.13	126.16	4	rs12053356	0.313	$6.29 \cdot 10^{-1}$	$9.66 \cdot 10^{-1}$	$7.35 \cdot 10^{-1}$	$5.10 \cdot 10^{-4}$ (103)
2	178.51	178.51	6	rs3821009 (*PDE11A*)	0.080	$9.62 \cdot 10^{-155}$	$2.34 \cdot 10^{-5}$	$3.38 \cdot 10^{-15}$	$1.09 \cdot 10^{-2}$ (13)
2	223.68	223.70	2	rs1440065	0.072	$2.00 \cdot 10^{-274}$	$9.17 \cdot 10^{-3}$	$5.40 \cdot 10^{-26}$	$2.16 \cdot 10^{-2}$ (6)
3	7.27	7.30	3	rs1605705 (*GRM7*)	0.178	$1.99 \cdot 10^{-1}$	$2.95 \cdot 10^{-1}$	$8.96 \cdot 10^{-3}$	$7.59 \cdot 10^{-4}$ (79)
3	71.93	71.94	3	rs17665418	0.084	$8.50 \cdot 10^{-285}$	$1.58 \cdot 10^{-5}$	$8.54 \cdot 10^{-21}$	$2.34 \cdot 10^{-2}$ (3)
3	141.43	141.44	4	rs9856460 (*CLSTN2*)	0.448	$6.82 \cdot 10^{-2}$	$8.81 \cdot 10^{-1}$	$4.11 \cdot 10^{-1}$	$7.79 \cdot 10^{-4}$ (77)
4	17.87	17.90	6	rs1553460	0.308	$1.28 \cdot 10^{-104}$	$2.88 \cdot 10^{-5}$	$1.16 \cdot 10^{-22}$	$4.41 \cdot 10^{-3}$ (26)
4	32.72	32.73	3	rs10022638	0.306	$8.29 \cdot 10^{-1}$	$4.30 \cdot 10^{-1}$	$2.95 \cdot 10^{-1}$	$5.75 \cdot 10^{-4}$ (91)
4	186.09	186.11	4	rs13126272 (*ACSL1*)	0.335	$2.15 \cdot 10^{-65}$	$3.02 \cdot 10^{-2}$	$1.08 \cdot 10^{-3}$	$1.35 \cdot 10^{-3}$ (51)
6	31.21	31.22	3	rs4959053 (*PSORS1C1*)	0.074	$7.30 \cdot 10^{-273}$	$3.03 \cdot 10^{-4}$	$1.12 \cdot 10^{-22}$	$2.04 \cdot 10^{-2}$ (7)
6	32.17	32.79	41	rs9275572	0.361	$8.57 \cdot 10^{-1}$	$8.80 \cdot 10^{-1}$	$1.20 \cdot 10^{-60}$	$2.05 \cdot 10^{-3}$ (34)
6	133.96	133.99	6	rs2677822	0.182	$5.80 \cdot 10^{-1}$	$8.05 \cdot 10^{-1}$	$5.86 \cdot 10^{-1}$	$2.83 \cdot 10^{-3}$ (30)
7	66.40	66.42	6	rs4718582	0.056	$2.34 \cdot 10^{-2}$	$3.99 \cdot 10^{-1}$	$3.06 \cdot 10^{-75}$	$7.06 \cdot 10^{-3}$ (23)
7	82.59	82.62	3	rs12670243	0.044	$3.16 \cdot 10^{-1}$	$4.14 \cdot 10^{-1}$	$1.23 \cdot 10^{-32}$	$1.30 \cdot 10^{-3}$ (52)
7	121.03	121.04	2	rs10262109	0.036	$6.30 \cdot 10^{-4}$	1	$5.80 \cdot 10^{-75}$	$8.22 \cdot 10^{-3}$ (17)
7	157.73	157.73	5	rs7789415 (*PTPRN2*)	0.043	1	$2.55 \cdot 10^{-6}$	$2.45 \cdot 10^{-66}$	$8.08 \cdot 10^{-3}$ (18)
8	76.67	76.68	5	rs1449555	0.382	$1.79 \cdot 10^{-1}$	$3.52 \cdot 10^{-1}$	$6.68 \cdot 10^{-2}$	$8.65 \cdot 10^{-4}$ (69)
8	90.26	90.26	2	rs1483373	0.032	$2.50 \cdot 10^{-1}$	$2.45 \cdot 10^{-1}$	$5.92 \cdot 10^{-1}$	$3.43 \cdot 10^{-4}$ (135)
8	107.27	107.30	2	rs16874228	0.112	$4.73 \cdot 10^{-1}$	$4.67 \cdot 10^{-1}$	$4.26 \cdot 10^{-1}$	$2.99 \cdot 10^{-4}$ (144)
8	118.44	118.46	2	rs1948674	0.038	$6.04 \cdot 10^{-122}$	$3.32 \cdot 10^{-2}$	$6.97 \cdot 10^{-1}$	$6.00 \cdot 10^{-3}$ (24)
9	0.19	0.24	3	rs669980	0.348	$2.32 \cdot 10^{-85}$	$1.05 \cdot 10^{-1}$	$8.12 \cdot 10^{-15}$	$3.63 \cdot 10^{-3}$ (28)
9	24.59	24.60	2	rs16908561	0.041	$1.22 \cdot 10^{-159}$	$4.91 \cdot 10^{-2}$	$1.35 \cdot 10^{-4}$	$8.39 \cdot 10^{-3}$ (15)
10	53.94	53.97	4	rs1733720	0.042	$7.67 \cdot 10^{-1}$	1	$7.14 \cdot 10^{-1}$	$2.01 \cdot 10^{-3}$ (36)
10	55.33	55.36	2	rs2121526 (*PCDH15*)	0.080	$9.13 \cdot 10^{-279}$	$1.33 \cdot 10^{-4}$	$7.35 \cdot 10^{-27}$	$2.31 \cdot 10^{-2}$ (5)
11	93.02	93.03	3	rs10501805	0.033	$1.06 \cdot 10^{-4}$	1	$1.73 \cdot 10^{-62}$	$5.74 \cdot 10^{-3}$ (25)
12	30.25	30.26	4	rs10743704	0.299	$2.79 \cdot 10^{-3}$	$4.02 \cdot 10^{-1}$	$9.45 \cdot 10^{-1}$	$8.04 \cdot 10^{-3}$ (19)
12	57.73	57.74	3	rs2028720	0.268	$2.68 \cdot 10^{-1}$	$2.96 \cdot 10^{-1}$	$2.87 \cdot 10^{-1}$	$6.41 \cdot 10^{-4}$ (87)
14	51.63	51.63	3	rs2144977	0.159	$1.80 \cdot 10^{-3}$	$9.43 \cdot 10^{-1}$	$3.43 \cdot 10^{-2}$	$2.26 \cdot 10^{-4}$ (165)
14	76.20	76.21	2	rs17104722	0.029	$5.81 \cdot 10^{-2}$	1	$4.88 \cdot 10^{-27}$	$1.68 \cdot 10^{-3}$ (41)
15	27.85	27.86	2	rs899848 (*TJP1*)	0.027	$5.43 \cdot 10^{-112}$	$6.33 \cdot 10^{-1}$	$1.51 \cdot 10^{-7}$	$2.06 \cdot 10^{-3}$ (33)
15	83.99	84.00	2	rs16942813 (*AKAP13*)	0.052	$5.22 \cdot 10^{-206}$	$2.10 \cdot 10^{-2}$	$1.27 \cdot 10^{-14}$	$1.12 \cdot 10^{-2}$ (11)
17	62.45	62.64	10	rs3785579 (*CACNG1*)	0.271	0	$2.12 \cdot 10^{-2}$	0	$1.50 \cdot 10^{-1}$ (1)
17	76.21	76.23	4	rs7503807 (*RPTOR*)	0.439	$7.05 \cdot 10^{-1}$	$6.01 \cdot 10^{-1}$	$2.13 \cdot 10^{-1}$	$8.14 \cdot 10^{-4}$ (75)
18	33.34	33.36	5	rs4799934 (*CELF4*)	0.078	$3.88 \cdot 10^{-286}$	$2.83 \cdot 10^{-3}$	$8.28 \cdot 10^{-27}$	$2.44 \cdot 10^{-2}$ (2)
18	74.55	74.58	3	rs2941794	0.202	$4.51 \cdot 10^{-11}$	1	$2.80 \cdot 10^{-11}$	$4.72 \cdot 10^{-4}$ (110)
4	182.13	182.13	1	rs4532278	0.210	$9.44 \cdot 10^{-1}$	$4.36 \cdot 10^{-1}$	1	$1.88 \cdot 10^{-4}$ (194)
6	31.83	31.83	1	rs707939	0.368	$4.46 \cdot 10^{-1}$	$8.31 \cdot 10^{-1}$	$5.54 \cdot 10^{-33}$	$1.87 \cdot 10^{-4}$ (195)

TABLE 7.15 RA_{wtccc}: lists of regions identified by the Random Forests and the T-Trees methods.

Type 1 diabetes

The large region (human MHC) on chromosome 6 is also reported in [Wel07] as being associated with type 1 diabetes. Most of the signal identified by the tree-based methods on the two dataset versions is located in that region. More than 75% of the first 200 variables are located in the human MHC region. Additionally, **rs6679677** on chromosome 1p13 appears isolated but present among the 200 first variables in our four experiments. As suggested in [W+09], the removal of that strong signal captured by an important number of markers should allow for the identification of other associated regions. On $T1D_{qc}$, the last region detected by the T-Trees on chromosome 10 also contains a marker that has been removed by the WTCCC.

On $T1D_{wtccc}$, the tree-based methods also focused on other regions. These contain variables with a (really) strong deviation from HWE. Although these are extreme deviations, even the RF exploited many surrounding variables meaning that this departure is not an isolated batch effect. To our knowledge, no association has been reported in these regions (for information: **rs3805006** is located near **ITPR1** gene and **rs7078771** near **NRG3**. Several investigations [T+12a, vB+11, CvdLJJea11, vdL+07] found in these two regions important deletions linked to other diseases).

Note that **rs9273363** is also detected by the approach of [P+12a].

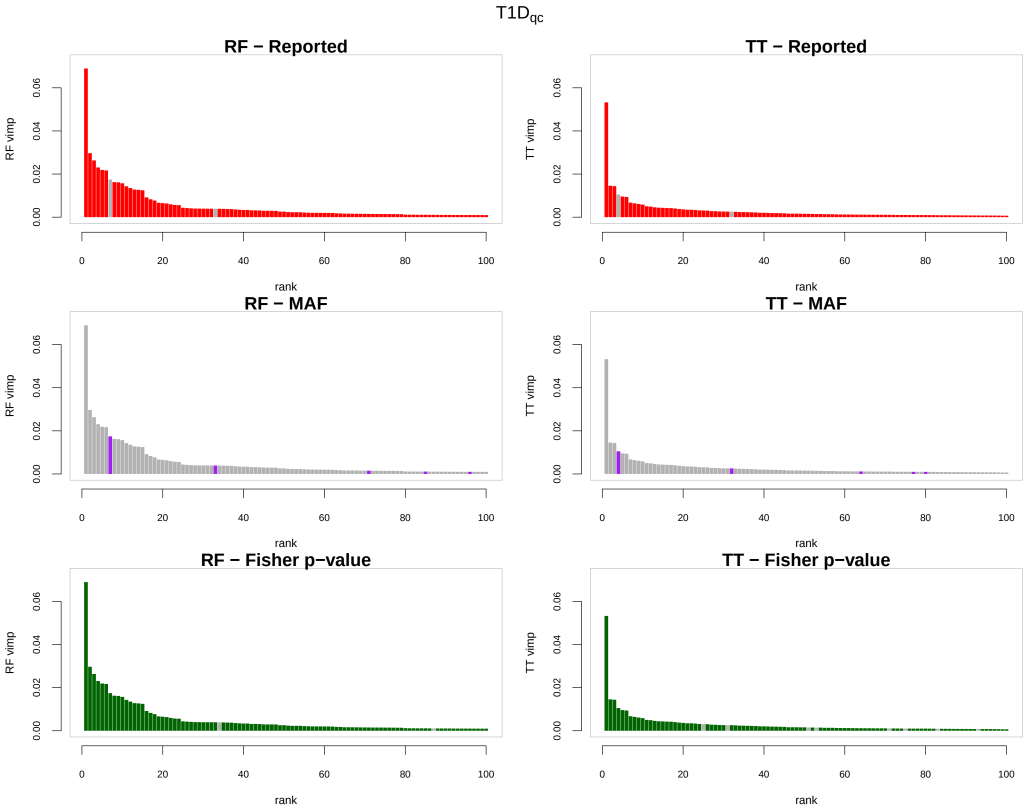


FIGURE 7.12 The first 100 variables according to the tree based importance rankings for $T1D_{qc}$. The horizontal axis corresponds to the ranks and the vertical axis to the variable importances. In the first column variables are ordered according to random forests variable importances and in the second column they are ordered according to the T-Trees variable importances. In the first row, red highlights the reported strongly associated regions. In the second row, purple corresponds to rare variants (MAF < 0.05). In the third row, green represents markers with a low Fisher p-value ($< 10^{-6}$).

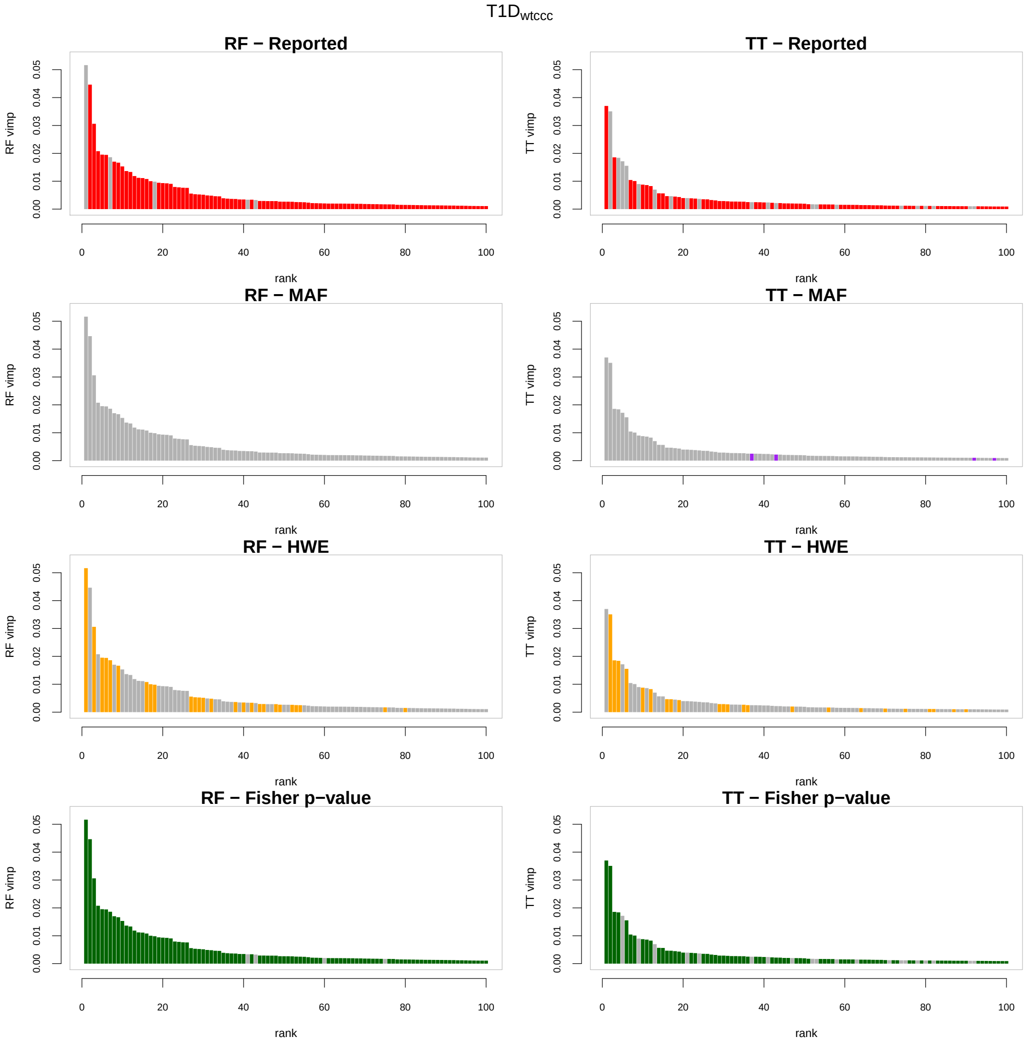


FIGURE 7.13 The first 100 variables according to the tree based importance rankings for $T1D_{wtccc}$. The horizontal axis corresponds to the ranks and the vertical axis to the variable importances. In the first column variables are ordered according to random forests variable importances and in the second column they are ordered according to the T-Trees variable importances. In the first row, red highlights the strongly associated reported regions. In the second row, purple corresponds to rare variants (MAF < 0.05). In the third row, orange highlights SNPs deviating from HWE and in the last row, green represents markers with a low Fisher p-value ($< 10^{-6}$).

Random Forests

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
6	30.47	30.47	2	rs3094703	0.172	$3.91 \cdot 10^{-2}$	$4.85 \cdot 10^{-1}$	$1.91 \cdot 10^{-29}$	$4.17 \cdot 10^{-4}$ (166)
6	30.87	30.90	2	rs6930444	0.030	$7.97 \cdot 10^{-1}$	$6.27 \cdot 10^{-1}$	$2.45 \cdot 10^{-14}$	$5.13 \cdot 10^{-4}$ (144)
6	31.17	31.20	2	rs3130544	0.188	$6.20 \cdot 10^{-2}$	$2.75 \cdot 10^{-1}$	$3.79 \cdot 10^{-35}$	$5.53 \cdot 10^{-4}$ (135)
6	31.35	31.35	2	rs3132486	0.399	$2.92 \cdot 10^{-1}$	$2.82 \cdot 10^{-1}$	$3.15 \cdot 10^{-48}$	$9.54 \cdot 10^{-4}$ (103)
6	31.44	33.06	185	rs9273363	0.467	$5.05 \cdot 10^{-3}$	$6.34 \cdot 10^{-1}$	0	$6.90 \cdot 10^{-2}$ (1)
1	114.02	114.02	1	rs6679677	0.125	$3.39 \cdot 10^{-1}$	$2.41 \cdot 10^{-1}$	$2.72 \cdot 10^{-27}$	$1.04 \cdot 10^{-3}$ (91)

T-Trees

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
6	30.45	30.47	2	rs3094050	0.172	$3.38 \cdot 10^{-2}$	$4.86 \cdot 10^{-1}$	$1.20 \cdot 10^{-29}$	$3.14 \cdot 10^{-4}$ (174)
6	30.87	30.90	3	rs6930444	0.030	$7.97 \cdot 10^{-1}$	$6.27 \cdot 10^{-1}$	$2.45 \cdot 10^{-14}$	$3.85 \cdot 10^{-4}$ (148)
6	31.17	31.25	4	rs3130544	0.188	$6.20 \cdot 10^{-2}$	$2.75 \cdot 10^{-1}$	$3.79 \cdot 10^{-35}$	$8.94 \cdot 10^{-4}$ (84)
6	31.35	31.35	2	rs3132486	0.399	$2.92 \cdot 10^{-1}$	$2.82 \cdot 10^{-1}$	$3.15 \cdot 10^{-48}$	$1.16 \cdot 10^{-3}$ (65)
6	31.44	33.06	179	rs9273363	0.467	$5.05 \cdot 10^{-3}$	$6.34 \cdot 10^{-1}$	0	$5.33 \cdot 10^{-2}$ (1)
10	131.97	132.00	2	rs7080464$^{(1)}$	0.014	1	$4.05 \cdot 10^{-1}$	$1.04 \cdot 10^{-27}$	$2.58 \cdot 10^{-3}$ (32)
1	114.02	114.02	1	rs6679677	0.125	$3.39 \cdot 10^{-1}$	$2.41 \cdot 10^{-1}$	$2.72 \cdot 10^{-27}$	$9.95 \cdot 10^{-4}$ (76)
6	30.71	30.71	1	rs2394390	0.037	$5.18 \cdot 10^{-1}$	1	$1.93 \cdot 10^{-14}$	$2.57 \cdot 10^{-4}$ (198)
6	31.02	31.02	1	rs3132581	0.191	$2.90 \cdot 10^{-3}$	$7.32 \cdot 10^{-1}$	$1.57 \cdot 10^{-22}$	$2.67 \cdot 10^{-4}$ (187)

TABLE 7.16 $T1D_{qc}$: lists of regions identified by the Random Forests and the T-Trees methods.

Random Forests

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
3	4.78	4.78	7	rs3805006 (*ITPR1*)	0.378	$7.50 \cdot 10^{-175}$	$4.27 \cdot 10^{-2}$	$7.16 \cdot 10^{-96}$	$5.16 \cdot 10^{-2}$ (1)
6	31.35	31.35	2	rs2524067	0.218	$5.39 \cdot 10^{-1}$	$5.65 \cdot 10^{-1}$	$1.02 \cdot 10^{-35}$	$4.24 \cdot 10^{-4}$ (188)
6	31.45	31.46	6	rs2853986	0.170	$3.42 \cdot 10^{-2}$	$4.47 \cdot 10^{-1}$	$2.97 \cdot 10^{-41}$	$8.30 \cdot 10^{-4}$ (130)
6	31.55	33.06	164	rs9273363	0.467	$5.83 \cdot 10^{-3}$	$6.01 \cdot 10^{-1}$	0	$4.46 \cdot 10^{-2}$ (2)
10	84.68	84.72	10	rs7078771 (*NRG3*)	0.076	$3.31 \cdot 10^{-257}$	$5.46 \cdot 10^{-5}$	$5.42 \cdot 10^{-15}$	$1.86 \cdot 10^{-2}$ (7)
13	82.21	82.33	5	rs4254200	0.468	$3.58 \cdot 10^{-132}$	$5.95 \cdot 10^{-1}$	$6.20 \cdot 10^{-29}$	$9.84 \cdot 10^{-3}$ (18)
1	114.02	114.02	1	rs6679677	0.126	$3.77 \cdot 10^{-1}$	$2.41 \cdot 10^{-1}$	$2.43 \cdot 10^{-26}$	$5.01 \cdot 10^{-4}$ (172)
6	30.87	30.87	1	rs6930444	0.030	$7.95 \cdot 10^{-1}$	$6.26 \cdot 10^{-1}$	$1.92 \cdot 10^{-13}$	$4.89 \cdot 10^{-4}$ (177)
6	31.17	31.17	1	rs3130544	0.189	$5.24 \cdot 10^{-2}$	$2.74 \cdot 10^{-1}$	$6.74 \cdot 10^{-34}$	$4.38 \cdot 10^{-4}$ (186)

T-Trees

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
3	4.78	4.78	9	rs3805006 (*ITPR1*)	0.378	$7.50 \cdot 10^{-175}$	$4.27 \cdot 10^{-2}$	$7.16 \cdot 10^{-96}$	$3.51 \cdot 10^{-2}$ (2)
3	146.99	146.99	5	rs6440373	0.464	$1.15 \cdot 10^{-61}$	$1.67 \cdot 10^{-2}$	$9.13 \cdot 10^{-11}$	$4.64 \cdot 10^{-3}$ (17)
6	31.35	31.35	2	rs3132486	0.399	$2.89 \cdot 10^{-1}$	$2.96 \cdot 10^{-1}$	$4.15 \cdot 10^{-47}$	$8.37 \cdot 10^{-4}$ (111)
6	31.45	33.06	156	rs9273363	0.467	$5.83 \cdot 10^{-3}$	$6.01 \cdot 10^{-1}$	0	$3.70 \cdot 10^{-2}$ (1)
10	84.67	84.68	3	rs7078771 (*NRG3*)	0.076	$3.31 \cdot 10^{-257}$	$5.46 \cdot 10^{-5}$	$5.42 \cdot 10^{-15}$	$1.84 \cdot 10^{-2}$ (4)
11	113.31	113.31	5	rs1176741 (*HTR3B*)	0.029	$3.80 \cdot 10^{-1}$	$1.19 \cdot 10^{-1}$	$1.12 \cdot 10^{-1}$	$2.49 \cdot 10^{-3}$ (37)
13	82.20	82.31	7	rs7332105	0.424	$9.63 \cdot 10^{-1}$	$8.21 \cdot 10^{-1}$	$3.07 \cdot 10^{-1}$	$1.72 \cdot 10^{-2}$ (5)
16	28.48	28.50	2	rs9924471 (*CCDC101*)	0.169	$1.74 \cdot 10^{-16}$	$4.69 \cdot 10^{-1}$	$1.40 \cdot 10^{-8}$	$6.02 \cdot 10^{-4}$ (135)
17	69.45	69.46	2	rs12103453	0.393	$2.66 \cdot 10^{-24}$	$6.06 \cdot 10^{-1}$	$8.71 \cdot 10^{-9}$	$7.31 \cdot 10^{-4}$ (120)
1	114.02	114.02	1	rs6679677	0.126	$3.77 \cdot 10^{-1}$	$2.41 \cdot 10^{-1}$	$2.43 \cdot 10^{-26}$	$4.12 \cdot 10^{-4}$ (182)
6	30.87	30.87	1	rs6930444	0.030	$7.95 \cdot 10^{-1}$	$6.26 \cdot 10^{-1}$	$1.92 \cdot 10^{-13}$	$3.58 \cdot 10^{-4}$ (194)
6	31.17	31.17	1	rs3130544	0.189	$5.24 \cdot 10^{-2}$	$2.74 \cdot 10^{-1}$	$6.74 \cdot 10^{-34}$	$6.25 \cdot 10^{-4}$ (130)
6	31.25	31.25	1	rs887464	0.482	$6.00 \cdot 10^{-4}$	$7.38 \cdot 10^{-3}$	$1.30 \cdot 10^{-30}$	$4.24 \cdot 10^{-4}$ (179)

TABLE 7.17 $T1D_{wtccc}$: lists of regions identified by the Random Forests and the T-Trees methods.

Type 2 diabetes

On Figures 7.14 and 7.15, Random Forests on $T2D_{qc}$ detected a few reported regions in the 100 first variables. In the three other experiments, a few such variables were selected but at much lower ranks. Again, in the absence of strong deviation from HWE, rare variants are preferentially selected on $T2D_{qc}$ while many SNPs with low HWE p-values are found in the 10 first variables on $T2D_{wtccc}$.

We notice that FAT3 gene (represented by rs10501796 and rs10501795) is selected in the four experiments. The ZFAT region is reported in a multi marker analysis ([FZ10]) and is detected by the T-Trees in both dataset versions. Also, rs959880 located in MCF2L2 gene which is reported as associated with type 2 diabetes in [T+08]. Similarly, rs13126272 located in ACSL1 is reported in [Z+13].

Finally, on $T2D_{wtccc}$, we note the presence of rs7077039 which is also reported in [P+12a]. And, on $T2D_{qc}$, rs10212068 was selected again as the most important one by the T-Trees.

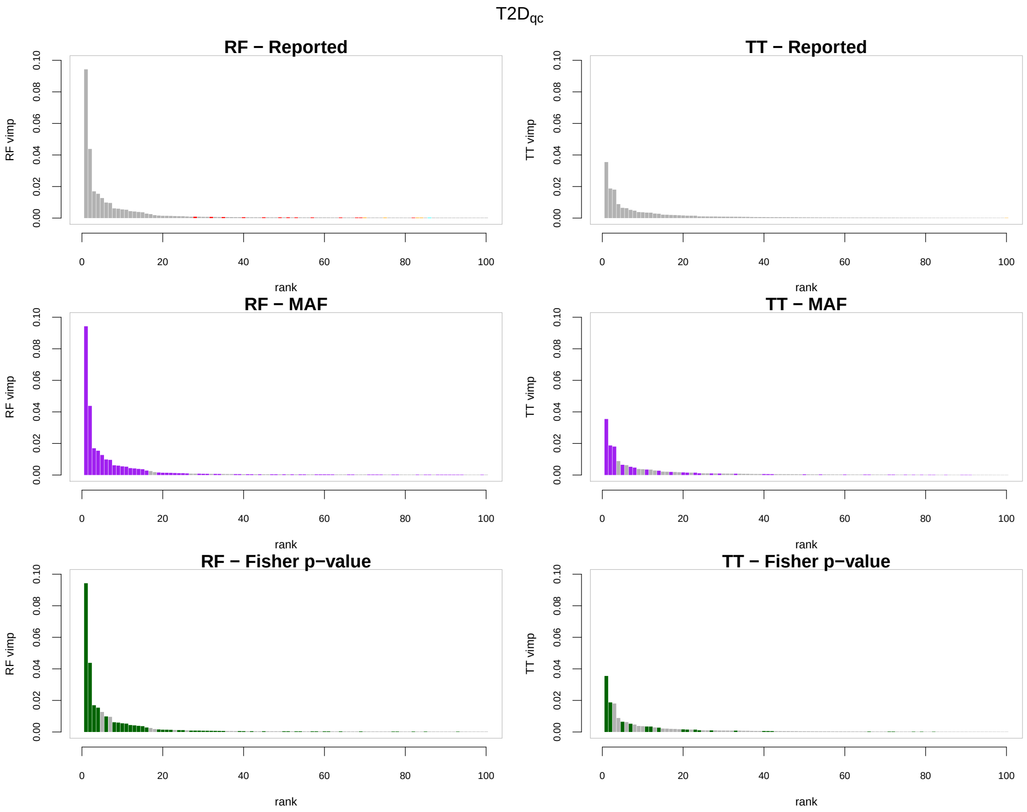


FIGURE 7.14 The first 100 variables according to the tree based importance rankings for $T2D_{qc}$. The horizontal axis corresponds to the ranks and the vertical axis to the variable importances. In the first column variables are ordered according to random forests variable importances and in the second column they are ordered according to the T-Trees variable importances. In the first row, red highlights the reported strongly associated regions. In the second row, purple corresponds to rare variants (MAF < 0.05). In the third row, green represents markers with a low Fisher p-value ($< 10^{-6}$).

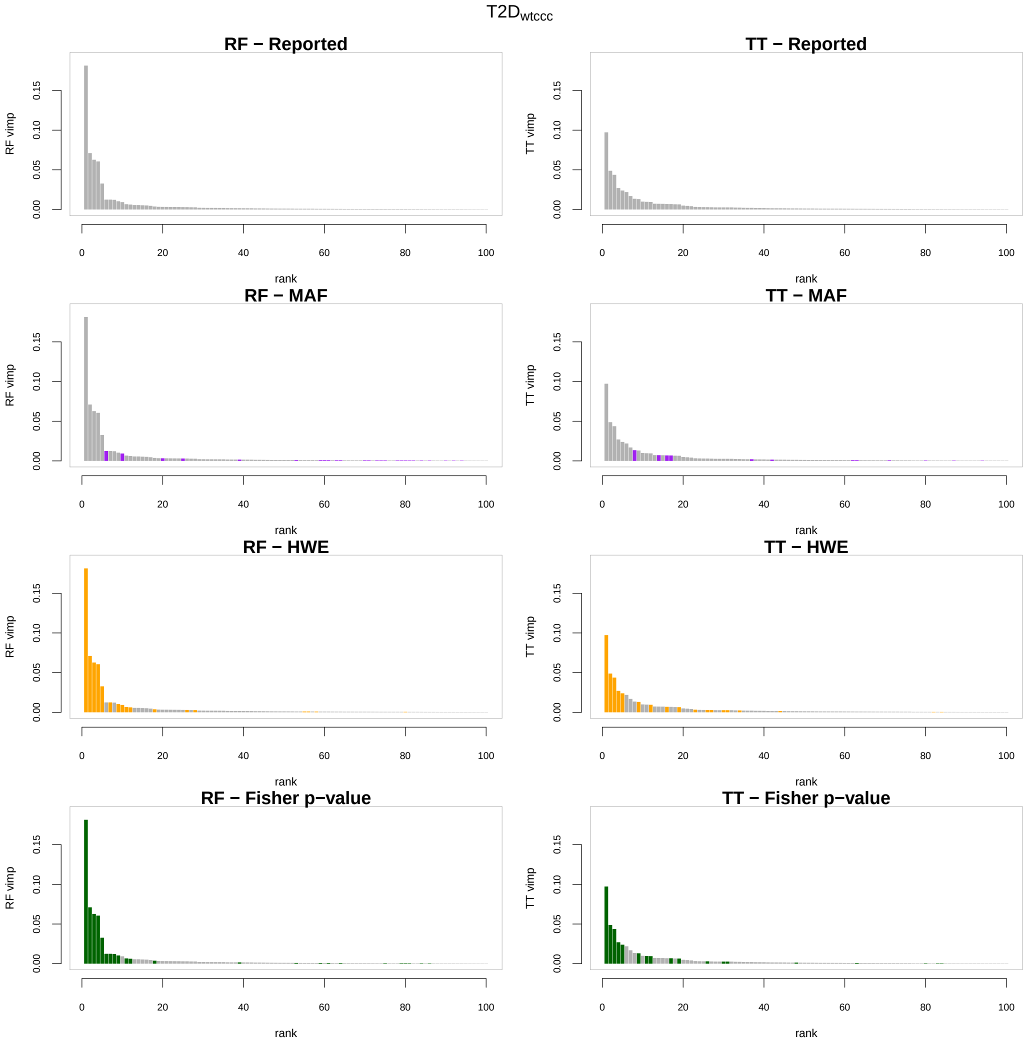


FIGURE 7.15 The first 100 variables according to the tree based importance rankings for $T2D_{wtccc}$. The horizontal axis corresponds to the ranks and the vertical axis to the variable importances. In the first column variables are ordered according to random forests variable importances and in the second column they are ordered according to the T-Trees variable importances. In the first row, red highlights the strongly associated reported regions. In the second row, purple corresponds to rare variants (MAF < 0.05). In the third row, orange highlights SNPs deviating from HWE and in the last row, green represents markers with a low Fisher p-value ($< 10^{-6}$).

Random Forests

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
3	154.52	154.52	2	rs10513440	0.201	$3.88 \cdot 10^{-1}$	$2.71 \cdot 10^{-3}$	$3.72 \cdot 10^{-4}$	$3.05 \cdot 10^{-4}$ (84)
5	10.65	10.66	2	rs17761026 (*ANKRD33B*)	0.111	$5.40 \cdot 10^{-5}$	1	$2.84 \cdot 10^{-3}$	$2.56 \cdot 10^{-4}$ (100)
5	82.93	82.97	2	rs6865544	0.010	$3.89 \cdot 10^{-1}$	1	$1.52 \cdot 10^{-5}$	$3.02 \cdot 10^{-4}$ (86)
6	20.75	20.83	2	rs9348440 (*CDKAL1*)	0.137	$3.63 \cdot 10^{-2}$	$8.16 \cdot 10^{-2}$	$1.80 \cdot 10^{-4}$	$4.51 \cdot 10^{-4}$ (49)
9	117.69	117.80	4	rs488101	0.363	$4.53 \cdot 10^{-8}$	$5.87 \cdot 10^{-1}$	$6.54 \cdot 10^{-7}$	$8.82 \cdot 10^{-4}$ (27)
10	28.62	28.64	2	rs11007003	0.068	$1.31 \cdot 10^{-2}$	$8.63 \cdot 10^{-2}$	$2.70 \cdot 10^{-2}$	$2.26 \cdot 10^{-4}$ (112)
10	84.07	84.08	2	rs11193797 (*NRG3*)	0.083	$9.13 \cdot 10^{-3}$	$6.65 \cdot 10^{-3}$	$4.34 \cdot 10^{-1}$	$2.49 \cdot 10^{-4}$ (102)
10	114.74	114.80	10	rs7077039 (*TCF7L2*)	0.483	1	$1.48 \cdot 10^{-1}$	$3.78 \cdot 10^{-12}$	$8.66 \cdot 10^{-4}$ (28)
11	16.84	16.88	3	rs392981 (*PLEKHA7*)	0.094	$1.62 \cdot 10^{-1}$	$9.96 \cdot 10^{-3}$	$1.29 \cdot 10^{-1}$	$1.89 \cdot 10^{-4}$ (131)
11	44.89	44.89	2	rs11038203 (*TSPAN18*)	0.005	1	1	$3.83 \cdot 10^{-9}$	$4.39 \cdot 10^{-3}$ (12)
11	92.07	92.09	3	rs10501796 (*FAT3*)	0.005	1	1	$4.28 \cdot 10^{-9}$	$3.88 \cdot 10^{-3}$ (14)
12	96.80	96.93	4	rs10492267	0.026	$9.73 \cdot 10^{-8}$	$8.07 \cdot 10^{-1}$	$7.56 \cdot 10^{-30}$	$4.38 \cdot 10^{-2}$ (2)
23	118.19	118.20	2	rs9988376	0.043	1	$7.54 \cdot 10^{-1}$	$6.40 \cdot 10^{-2}$	$1.48 \cdot 10^{-4}$ (171)
3	55.29	55.29	1	rs358806	0.199	$7.92 \cdot 10^{-6}$	$2.33 \cdot 10^{-2}$	$4.68 \cdot 10^{-1}$	$3.57 \cdot 10^{-4}$ (68)
3	150.03	150.03	1	rs16861027	0.034	$3.06 \cdot 10^{-2}$	$7.28 \cdot 10^{-1}$	$1.88 \cdot 10^{-5}$	$3.05 \cdot 10^{-4}$ (83)
5	122.49	122.49	1	rs6872465	0.018	$3.37 \cdot 10^{-1}$	$4.16 \cdot 10^{-1}$	$3.62 \cdot 10^{-5}$	$1.78 \cdot 10^{-4}$ (140)
8	98.43	98.43	1	rs2679765	0.169	$1.84 \cdot 10^{-2}$	$1.71 \cdot 10^{-2}$	$4.65 \cdot 10^{-3}$	$3.56 \cdot 10^{-4}$ (70)
12	18.47	18.47	1	rs12581163	0.171	$5.19 \cdot 10^{-3}$	$3.67 \cdot 10^{-3}$	$1.72 \cdot 10^{-1}$	$3.30 \cdot 10^{-4}$ (75)
16	52.37	52.37	1	rs8050136	0.422	$7.44 \cdot 10^{-2}$	$9.70 \cdot 10^{-1}$	$3.90 \cdot 10^{-8}$	$1.38 \cdot 10^{-4}$ (196)

T-Trees

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	18.82	18.84	4	rs2789336$^{(1)}$	0.006	1	1	$2.34 \cdot 10^{-11}$	$1.50 \cdot 10^{-3}$ (23)
1	30.67	30.68	6	rs12031413$^{(1)}$	0.004	1	1	$2.77 \cdot 10^{-15}$	$5.23 \cdot 10^{-3}$ (7)
1	86.93	86.98	4	rs1208054	0.220	$6.51 \cdot 10^{-1}$	$5.49 \cdot 10^{-1}$	$1.41 \cdot 10^{-1}$	$3.68 \cdot 10^{-3}$ (10)
1	164.79	164.87	9	rs275145(*GPR161*)$^{(1)}$	0.004	1	1	$3.02 \cdot 10^{-16}$	$6.49 \cdot 10^{-3}$ (5)
1	165.12	165.12	3	rs2300564 (*LOC100505918*)	0.405	$4.52 \cdot 10^{-1}$	$6.21 \cdot 10^{-1}$	$9.83 \cdot 10^{-1}$	$2.14 \cdot 10^{-4}$ (83)
2	5.37	5.39	6	rs1453783$^{(3)}$	0.374	$5.00 \cdot 10^{-1}$	$1.62 \cdot 10^{-2}$	$3.05 \cdot 10^{-1}$	$1.04 \cdot 10^{-3}$ (26)
2	10.90	10.91	3	rs902133$^{(1)}$	0.005	1	1	$2.82 \cdot 10^{-12}$	$3.47 \cdot 10^{-3}$ (11)
2	53.60	53.61	3	rs903228	0.058	$7.28 \cdot 10^{-1}$	$5.61 \cdot 10^{-1}$	$3.29 \cdot 10^{-5}$	$2.67 \cdot 10^{-4}$ (74)
3	7.92	7.93	6	rs10510375	0.004	1	1	$4.47 \cdot 10^{-7}$	$5.82 \cdot 10^{-4}$ (41)
3	66.59	66.60	2	rs3845903 (*LRIG1*)	0.006	1	1	$4.54 \cdot 10^{-11}$	$1.01 \cdot 10^{-3}$ (27)
3	154.52	154.56	2	rs10513440	0.201	$3.88 \cdot 10^{-1}$	$2.71 \cdot 10^{-3}$	$3.72 \cdot 10^{-4}$	$1.66 \cdot 10^{-4}$ (100)
3	163.64	163.70	5	rs9858104	0.421	$2.62 \cdot 10^{-1}$	$3.89 \cdot 10^{-1}$	$1.50 \cdot 10^{-1}$	$5.02 \cdot 10^{-4}$ (48)
4	1.11	1.12	4	rs6826705	0.298	$2.30 \cdot 10^{-1}$	$1.61 \cdot 10^{-1}$	$8.21 \cdot 10^{-1}$	$1.06 \cdot 10^{-3}$ (25)
4	33.74	33.78	3	rs10517298	0.090	$2.76 \cdot 10^{-1}$	$4.89 \cdot 10^{-1}$	$4.23 \cdot 10^{-1}$	$1.67 \cdot 10^{-4}$ (99)
4	116.14	116.14	2	rs7666328(*NDST4*)$^{(3)}$	0.033	$1.80 \cdot 10^{-1}$	$2.82 \cdot 10^{-3}$	$2.48 \cdot 10^{-7}$	$3.23 \cdot 10^{-4}$ (66)
5	14.68	14.70	2	rs153822	0.049	$1.53 \cdot 10^{-1}$	$1.36 \cdot 10^{-2}$	$5.03 \cdot 10^{-1}$	$1.17 \cdot 10^{-4}$ (127)
6	32.85	32.87	10	rs17429127	0.067	$7.15 \cdot 10^{-1}$	$5.53 \cdot 10^{-2}$	1	$2.88 \cdot 10^{-3}$ (13)
7	47.52	47.52	4	rs7792409	0.220	$7.45 \cdot 10^{-1}$	$3.58 \cdot 10^{-1}$	$3.19 \cdot 10^{-1}$	$6.41 \cdot 10^{-4}$ (38)
7	50.31	50.31	2	rs11575518 (*DDC*)	0.015	1	1	$6.72 \cdot 10^{-1}$	$8.38 \cdot 10^{-5}$ (177)
8	32.07	32.09	3	rs16878847 (*NRG1*)	0.002	1	1	$1.46 \cdot 10^{-9}$	$1.52 \cdot 10^{-3}$ (21)
8	37.16	37.16	2	rs7826024	0.005	1	1	$1.61 \cdot 10^{-1}$	$4.57 \cdot 10^{-4}$ (50)
8	135.58	135.65	12	rs1372662 (*ZFAT*)	0.332	$6.83 \cdot 10^{-1}$	$3.22 \cdot 10^{-1}$	$9.30 \cdot 10^{-1}$	$8.85 \cdot 10^{-3}$ (4)
9	117.78	117.82	6	rs2151370	0.377	$6.61 \cdot 10^{-1}$	$7.86 \cdot 10^{-1}$	$7.41 \cdot 10^{-2}$	$6.27 \cdot 10^{-3}$ (6)
10	77.12	77.12	2	rs7082404	0.004	1	1	$6.98 \cdot 10^{-7}$	$2.37 \cdot 10^{-4}$ (79)
10	114.74	114.78	3	rs7077039 (*TCF7L2*)	0.483	1	$1.48 \cdot 10^{-1}$	$3.78 \cdot 10^{-12}$	$8.50 \cdot 10^{-5}$ (174)
11	44.89	44.89	3	rs11038203 (*TSPAN18*)	0.005	1	1	$3.83 \cdot 10^{-9}$	$1.08 \cdot 10^{-3}$ (24)
11	92.07	92.09	8	rs10501795 (*FAT3*)	0.009	1	1	$5.12 \cdot 10^{-1}$	$4.71 \cdot 10^{-3}$ (8)
12	10.27	10.28	2	rs2900385	0.019	1	$6.25 \cdot 10^{-1}$	$6.00 \cdot 10^{-1}$	$3.26 \cdot 10^{-4}$ (65)
12	96.80	96.93	11	rs10492267	0.026	$9.73 \cdot 10^{-8}$	$8.07 \cdot 10^{-1}$	$7.56 \cdot 10^{-30}$	$1.87 \cdot 10^{-2}$ (2)
15	42.65	42.75	2	rs2277610 (*SPG11*)	0.006	1	$1.01 \cdot 10^{-1}$	1	$3.40 \cdot 10^{-4}$ (60)
17	14.28	14.29	2	rs7221370$^{(1)}$	0.003	1	1	$1.39 \cdot 10^{-6}$	$5.59 \cdot 10^{-4}$ (42)
18	27.45	27.45	2	rs974676	0.349	$7.66 \cdot 10^{-1}$	$1.70 \cdot 10^{-1}$	$7.95 \cdot 10^{-1}$	$1.40 \cdot 10^{-4}$ (111)
18	41.08	41.12	3	rs7235815(*SLC14A2*)$^{(1)}$	0.007	1	1	$3.29 \cdot 10^{-3}$	$2.30 \cdot 10^{-4}$ (81)
22	35.97	36.02	9	rs10212068$^{(3)}$	0.028	1	$2.64 \cdot 10^{-3}$	$8.47 \cdot 10^{-62}$	$3.55 \cdot 10^{-2}$ (1)
23	40.10	40.10	2	rs952836	0.328	$1.59 \cdot 10^{-4}$	$6.41 \cdot 10^{-2}$	$3.04 \cdot 10^{-2}$	$2.00 \cdot 10^{-4}$ (86)
5	82.93	82.93	1	rs6865544	0.010	$3.89 \cdot 10^{-1}$	1	$1.52 \cdot 10^{-5}$	$9.56 \cdot 10^{-5}$ (158)

TABLE 7.18 $T2D_{qc}$: lists of regions identified by the Random Forests and the T-Trees methods.

Random Forests

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
2	25.31	25.36	3	rs2164411	0.153	$1.45 \cdot 10^{-17}$	$1.95 \cdot 10^{-1}$	$4.46 \cdot 10^{-4}$	$2.98 \cdot 10^{-3}$ (26)
4	6.06	6.08	9	rs16837871	0.150	$3.34 \cdot 10^{-26}$	$2.84 \cdot 10^{-1}$	$8.97 \cdot 10^{-25}$	$1.25 \cdot 10^{-2}$ (7)
4	186.09	186.11	4	rs13126272 (*ACSL1*)	0.340	$4.19 \cdot 10^{-43}$	$3.02 \cdot 10^{-2}$	$3.11 \cdot 10^{-6}$	$3.79 \cdot 10^{-3}$ (18)
5	34.78	34.79	2	rs2048646 (*RAI14*)	0.231	$3.77 \cdot 10^{-23}$	$1.67 \cdot 10^{-1}$	$3.84 \cdot 10^{-2}$	$2.77 \cdot 10^{-3}$ (28)
5	117.00	117.07	17	rs2416472	0.313	$2.21 \cdot 10^{-20}$	$9.35 \cdot 10^{-1}$	$2.06 \cdot 10^{-16}$	$6.69 \cdot 10^{-3}$ (11)
6	20.75	20.83	2	rs9465871 (*CDKAL1*)	0.194	$2.29 \cdot 10^{-1}$	$4.36 \cdot 10^{-2}$	$1.38 \cdot 10^{-6}$	$1.53 \cdot 10^{-4}$ (150)
6	46.01	46.03	8	rs3777582 (*CLIC5*)	0.082	$4.40 \cdot 10^{-295}$	$1.29 \cdot 10^{-4}$	$6.76 \cdot 10^{-30}$	$7.11 \cdot 10^{-2}$ (2)
6	107.25	107.26	2	rs10499044	0.099	$3.75 \cdot 10^{-13}$	$1.97 \cdot 10^{-3}$	$7.06 \cdot 10^{-22}$	$6.23 \cdot 10^{-3}$ (12)
7	136.22	136.37	6	rs1477523 (*LOC349160*)	0.053	$2.97 \cdot 10^{-189}$	$5.95 \cdot 10^{-3}$	$3.58 \cdot 10^{-8}$	$3.27 \cdot 10^{-2}$ (5)
8	17.51	17.52	2	rs2517202 (*PDGFRL*)	0.006	1	1	$2.21 \cdot 10^{-8}$	$5.93 \cdot 10^{-4}$ (79)
8	19.35	19.53	19	rs17480050 (*CSGALNACT1*)	0.114	0	1	0	$1.81 \cdot 10^{-1}$ (1)
9	117.69	117.79	2	rs488101	0.364	$3.77 \cdot 10^{-8}$	$6.12 \cdot 10^{-1}$	$6.93 \cdot 10^{-7}$	$2.15 \cdot 10^{-4}$ (128)
10	114.74	114.80	9	rs7077039 (*TCF7L2*)	0.484	1	$1.35 \cdot 10^{-1}$	$4.20 \cdot 10^{-12}$	$2.63 \cdot 10^{-4}$ (113)
11	10.16	10.19	5	rs11042656 (*SBF2*)	0.076	$8.28 \cdot 10^{-274}$	$5.07 \cdot 10^{-4}$	$8.38 \cdot 10^{-26}$	$6.06 \cdot 10^{-2}$ (4)
11	92.07	92.09	3	rs10501796 (*FAT3*)	0.005	1	1	$1.23 \cdot 10^{-8}$	$8.40 \cdot 10^{-4}$ (61)
11	113.02	113.31	19	rs17116117 (*HTR3B*)	0.053	$8.32 \cdot 10^{-5}$	$1.16 \cdot 10^{-1}$	$3.41 \cdot 10^{-31}$	$1.22 \cdot 10^{-2}$ (8)
12	96.85	96.93	3	rs10492267	0.026	$6.38 \cdot 10^{-8}$	$8.03 \cdot 10^{-1}$	$8.53 \cdot 10^{-30}$	$1.25 \cdot 10^{-2}$ (6)
16	58.89	58.90	2	rs9889057	0.360	$3.84 \cdot 10^{-32}$	$1.04 \cdot 10^{-1}$	$3.94 \cdot 10^{-5}$	$9.69 \cdot 10^{-4}$ (55)
21	26.95	27.00	17	rs226261	0.374	$1.10 \cdot 10^{-53}$	$9.03 \cdot 10^{-1}$	$6.75 \cdot 10^{-8}$	$1.04 \cdot 10^{-2}$ (9)
22	23.76	23.88	10	rs11705626 (*KIAA1671*)	0.081	$8.08 \cdot 10^{-28}$	$5.83 \cdot 10^{-2}$	$1.06 \cdot 10^{-96}$	$6.27 \cdot 10^{-2}$ (3)
23	0.49	0.64	2	rs5988334	0.216	$3.42 \cdot 10^{-3}$	$2.06 \cdot 10^{-3}$	$3.31 \cdot 10^{-6}$	$3.20 \cdot 10^{-4}$ (104)
3	55.29	55.29	1	rs358806	0.201	$7.67 \cdot 10^{-6}$	$2.00 \cdot 10^{-2}$	$4.82 \cdot 10^{-1}$	$9.38 \cdot 10^{-5}$ (200)
5	82.93	82.93	1	rs6865544	0.010	$3.93 \cdot 10^{-1}$	1	$1.28 \cdot 10^{-5}$	$1.66 \cdot 10^{-4}$ (144)
8	98.43	98.43	1	rs2679765	0.170	$2.28 \cdot 10^{-2}$	$2.36 \cdot 10^{-2}$	$2.39 \cdot 10^{-3}$	$1.38 \cdot 10^{-4}$ (158)
11	94.53	94.53	1	rs11021059	0.106	$8.26 \cdot 10^{-1}$	$2.98 \cdot 10^{-6}$	$1.86 \cdot 10^{-3}$	$1.11 \cdot 10^{-4}$ (179)
12	18.47	18.47	1	rs12581163	0.171	$3.97 \cdot 10^{-3}$	$5.05 \cdot 10^{-3}$	$1.17 \cdot 10^{-1}$	$1.93 \cdot 10^{-4}$ (134)

T-Trees

chr	start	end	size	rsid	MAF	HWE$_{case}$	HWE$_{control}$	p-value	importance
1	22.10	22.15	4	rs2473324 (*CDC42*)	0.069	1	$2.44 \cdot 10^{-1}$	$7.75 \cdot 10^{-1}$	$4.45 \cdot 10^{-4}$ (88)
1	75.68	75.68	2	rs13373826 (*SLC44A5*)	0.148	$1.01 \cdot 10^{-10}$	$4.32 \cdot 10^{-1}$	$1.43 \cdot 10^{-1}$	$5.25 \cdot 10^{-4}$ (82)
1	80.57	80.57	4	rs1896250	0.396	$8.04 \cdot 10^{-3}$	$9.37 \cdot 10^{-1}$	$4.19 \cdot 10^{-4}$	$3.59 \cdot 10^{-4}$ (101)
1	86.95	86.98	3	rs1208054	0.219	$4.34 \cdot 10^{-1}$	$5.10 \cdot 10^{-1}$	$1.75 \cdot 10^{-1}$	$1.30 \cdot 10^{-3}$ (49)
1	89.44	89.44	2	rs11587221 (*GBP5*)	0.421	$3.51 \cdot 10^{-1}$	$6.22 \cdot 10^{-1}$	$3.78 \cdot 10^{-1}$	$1.03 \cdot 10^{-3}$ (59)
2	25.31	25.36	3	rs2164411	0.153	$1.45 \cdot 10^{-17}$	$1.95 \cdot 10^{-1}$	$4.46 \cdot 10^{-4}$	$1.56 \cdot 10^{-3}$ (44)
3	184.57	184.57	3	rs959880 (*MCF2L2*)	0.146	$2.65 \cdot 10^{-2}$	$2.65 \cdot 10^{-1}$	$9.53 \cdot 10^{-1}$	$4.40 \cdot 10^{-4}$ (91)
4	6.02	6.06	8	rs16837871	0.150	$3.34 \cdot 10^{-26}$	$2.84 \cdot 10^{-1}$	$8.97 \cdot 10^{-25}$	$9.49 \cdot 10^{-3}$ (12)
4	186.09	186.11	5	rs13126272 (*ACSL1*)	0.340	$4.19 \cdot 10^{-43}$	$3.02 \cdot 10^{-2}$	$3.11 \cdot 10^{-6}$	$2.65 \cdot 10^{-3}$ (30)
5	34.78	34.79	2	rs334912 (*RAI14*)	0.457	$3.11 \cdot 10^{-1}$	$1.19 \cdot 10^{-1}$	$4.41 \cdot 10^{-1}$	$4.92 \cdot 10^{-3}$ (20)
5	117.02	117.07	10	rs17411921	0.336	1	$3.66 \cdot 10^{-1}$	$3.45 \cdot 10^{-1}$	$9.94 \cdot 10^{-3}$ (10)
6	46.01	46.02	6	rs3777582 (*CLIC5*)	0.082	$4.40 \cdot 10^{-295}$	$1.29 \cdot 10^{-4}$	$6.76 \cdot 10^{-30}$	$4.88 \cdot 10^{-2}$ (2)
6	93.79	93.81	7	rs503319	0.195	$4.33 \cdot 10^{-1}$	$8.58 \cdot 10^{-1}$	$2.49 \cdot 10^{-1}$	$2.67 \cdot 10^{-3}$ (28)
6	107.25	107.26	3	rs10499044	0.099	$3.75 \cdot 10^{-13}$	$1.97 \cdot 10^{-3}$	$7.06 \cdot 10^{-22}$	$2.96 \cdot 10^{-3}$ (26)
7	136.26	136.27	6	rs1477523 (*LOC349160*)	0.053	$2.97 \cdot 10^{-189}$	$5.95 \cdot 10^{-3}$	$3.58 \cdot 10^{-8}$	$2.39 \cdot 10^{-2}$ (5)
8	19.35	19.51	16	rs17480050 (*CSGALNACT1*)	0.114	0	1	0	$9.73 \cdot 10^{-2}$ (1)
8	37.16	37.16	2	rs7826024	0.005	1	1	$1.96 \cdot 10^{-1}$	$2.38 \cdot 10^{-4}$ (129)
8	55.10	55.23	6	rs11984645	0.142	$1.17 \cdot 10^{-8}$	1	$4.32 \cdot 10^{-1}$	$1.05 \cdot 10^{-3}$ (56)
8	135.58	135.65	9	rs1372662 (*ZFAT*)	0.333	$7.97 \cdot 10^{-1}$	$3.39 \cdot 10^{-1}$	$9.47 \cdot 10^{-1}$	$4.16 \cdot 10^{-3}$ (22)
9	117.79	117.81	3	rs2151370	0.377	$6.94 \cdot 10^{-1}$	$8.45 \cdot 10^{-1}$	$6.89 \cdot 10^{-2}$	$2.99 \cdot 10^{-3}$ (25)
11	10.12	10.19	6	rs11042656 (*SBF2*)	0.076	$8.28 \cdot 10^{-274}$	$5.07 \cdot 10^{-4}$	$8.38 \cdot 10^{-26}$	$4.37 \cdot 10^{-2}$ (3)
11	44.89	44.89	2	rs11038203 (*TSPAN18*)	0.005	1	1	$1.17 \cdot 10^{-8}$	$3.40 \cdot 10^{-4}$ (103)
11	55.32	55.35	2	rs17501618	0.054	$3.63 \cdot 10^{-3}$	$7.29 \cdot 10^{-4}$	$3.77 \cdot 10^{-1}$	$2.20 \cdot 10^{-4}$ (140)
11	92.09	92.09	2	rs10501795 (*FAT3*)	0.009	1	1	$5.04 \cdot 10^{-1}$	$1.60 \cdot 10^{-3}$ (42)
11	113.28	113.31	7	rs1176741 (*HTR3B*)	0.031	1	$1.19 \cdot 10^{-1}$	$6.75 \cdot 10^{-1}$	$1.35 \cdot 10^{-2}$ (8)
12	96.85	96.91	7	rs10860262	0.042	$7.79 \cdot 10^{-1}$	$3.50 \cdot 10^{-1}$	$7.18 \cdot 10^{-1}$	$7.21 \cdot 10^{-3}$ (14)
12	124.08	124.09	3	rs879993 (*AACS*)	0.137	$2.47 \cdot 10^{-18}$	$2.66 \cdot 10^{-3}$	$8.45 \cdot 10^{-2}$	$2.91 \cdot 10^{-4}$ (114)
14	59.81	59.83	3	rs7154773 (*PPM1A*)	0.358	$4.62 \cdot 10^{-1}$	$1.96 \cdot 10^{-1}$	$6.27 \cdot 10^{-2}$	$2.72 \cdot 10^{-4}$ (121)
14	86.39	86.39	2	rs1362719	0.449	$1.92 \cdot 10^{-21}$	$7.84 \cdot 10^{-5}$	$9.63 \cdot 10^{-6}$	$5.01 \cdot 10^{-4}$ (84)
15	21.57	21.58	3	rs17117531	0.022	$1.95 \cdot 10^{-92}$	$6.26 \cdot 10^{-1}$	$1.05 \cdot 10^{-2}$	$6.92 \cdot 10^{-3}$ (16)
15	32.00	32.03	6	rs597414 (*AVEN*)	0.274	$1.71 \cdot 10^{-15}$	$5.92 \cdot 10^{-2}$	$8.69 \cdot 10^{-3}$	$2.35 \cdot 10^{-3}$ (34)
16	58.89	58.90	2	rs10500428	0.351	$1.61 \cdot 10^{-1}$	$2.25 \cdot 10^{-1}$	$8.62 \cdot 10^{-1}$	$6.60 \cdot 10^{-3}$ (18)
17	40.43	40.47	3	rs9915259 (*DCAKD*)	0.207	$3.34 \cdot 10^{-1}$	$5.35 \cdot 10^{-1}$	$7.39 \cdot 10^{-1}$	$8.50 \cdot 10^{-4}$ (66)
17	50.27	50.29	2	rs2934884	0.196	$7.13 \cdot 10^{-1}$	$4.16 \cdot 10^{-1}$	$4.48 \cdot 10^{-1}$	$1.87 \cdot 10^{-4}$ (150)
21	26.95	27.00	13	rs2830322	0.191	$9.41 \cdot 10^{-1}$	$5.51 \cdot 10^{-1}$	$9.79 \cdot 10^{-1}$	$1.69 \cdot 10^{-2}$ (7)
22	23.72	23.88	13	rs11705626 (*KIAA1671*)	0.081	$8.08 \cdot 10^{-28}$	$5.83 \cdot 10^{-2}$	$1.06 \cdot 10^{-96}$	$2.70 \cdot 10^{-2}$ (4)

TABLE 7.19 $T2D_{wtccc}$: lists of regions identified by the Random Forests and the T-Trees methods.

7.4 Overall remarks

Although the studies reported in the present chapter are less extensive and detailed than our study of Chapter 6, we believe that they provide good insight on how the tree-based methods behave on real-life datasets in the GWAS field.

Most importantly, in terms of predictive power, our novel T-Trees method produced consistently better AUC values than the standard Random Forests on every dataset. These promising AUC increases validate our previous investigation of the Crohn's disease dataset. Taking into account the structured nature of the variables in the genomic context has proven to be profitable. Even though we saw with the Random Forests that the most important variables already appeared in groups, we believe it confirms that the treatment of several neighbouring variables at once as in the T-Trees method allows to take advantage over Random Forests type of methods. Indeed, in the Random Forests, the chance of exploiting several markers in a region are lower. Constructing the "right" cascade of nodes might take longer while the T-Trees directly focuses its efforts on grouped variables and their local structure.

The analysis of variable importances pointed out that some markers showing particularities were preferentially considered as important for both of the tree-based methods. For example, with "our" quality control filters, some variables that were excluded from the WTCCC datasets appeared in our top rankings. Most of these were rare variants with a missing rate $> 1\%$ or discriminant for the two sub-group of controls. For example, we noticed the recurrent presence of a marker (`rs10212068`) on all of our qc datasets. That marker was selected by both Random Forests and T-Trees, although it was not always reported in the Random Forests tables as it appeared alone.

On the other hand, for the WTCCC dataset versions, many markers strongly deviating from the Hardy-Weinberg equilibrium were spotted in the most important descriptor listings. Strangely, while this filter is commonly accepted has being good exclusion criteria, it is also disputed ([ZVSW10]). We discovered that when such variables were exploited in a forest, they were "followed" by many of their neighbours. In addition, these particularities allowed in some cases to detect signals there were not reported in the WTCCC study.

We also found out similarities with our findings to those of other investigations of the same datasets pointing to some of these (rare or deviating form HWE) variables. This might indicate that, while these markers are set aside and discarded from classical studies due to the lack of statistical power, they might reveal true associations with the use of different and maybe more powerful approaches.

Interestingly, we identified some markers located in genes reported posterior to the WTCCC genome-wide association study and in other datasets. Some of these markers are rare variants and might have been underestimated due to the lack of power of classical univariate approaches in the WTCCC study. Different sets of individuals may have increased the chance of identifying such regions as it might have changed the observed allele frequencies.

With the T-Trees, we also noticed some similarities with other research findings. These researches were focused either on finding epistatic interactions or detecting "super-allele" associations allowing them to spot loci that were unknown before. Although these finding were a few (and might be questionable), they might as well point that such approaches are currently underestimated in addition to reinforcing our primary motivations/intuitions that lead to the T-Trees approach.

Of course, we are aware that some of our findings obviously point to spurious associations, but still, these are statistical associations. Beside that, we also showed that the tree-based methods were able to spot many of the reported regions, and not only the ones that were discovered by the WTCCC study but a few more. Although we were not always able to link regions to the literature findings, we noticed in some cases a strong consistency among experiments involving the same phenotype. It has to be noted that we mostly discussed markers directly located in genes but some of the most important ones might also be located outside genes. They can still be in linkage disequilibrium with many others.

Finally, we mainly looked at the 100 first variables. In cases like $T1D$ or RA, where an important number of SNPs are expected to be associated, one might want to look further down and see what is found just below in the rankings or to remove the regions from the candidate attributes to discover what else is popping out in the top rankings.

Part III

Conclusion

Chapter 8

Closure

Below we first summarise our research pathway, and then we report our main findings, and finally we suggest further research and development directions.

8.1 Epitome

The initial objective of this thesis was to study the application of state-of-the-art tree-based supervised learning methods in the context of genome-wide association studies of complex diseases, with a twofold goal, namely, on the one hand, the inference of predictive models of the disease risk from available datasets, and, on the other hand, the identification of the genetic information contained in these datasets and that is useful for making these predictions. To this end, we have considered both existing methods (such as Random Forests, and Extra-Trees) and designed a novel algorithm (called T-Trees) which is tailored to the strong correlations among genetic markers stemming from the so-called genetic linkage. To study and compare these three algorithms, we have developed our own software package, and designed sound and reproducible empirical protocols, and applied them both on synthetic and on real datasets.

The organisation of this manuscript reflects the main steps of our research: in Part I, we analyse the state-of-the art in the application field of genome wide association studies and in supervised machine learning, and subsequently describe in details the three tree-based ensemble methods that we have implemented and applied in our research; in Part II, we report our empirical investigations, in three successive steps, namely i.) a preliminary study on simulated datasets yielding controlled conditions with known ground-truth and allowing for a first sanity check of the T-Trees methods, in ideal conditions; ii.) a detailed study on a given real-life dataset concerning Crohn's disease, where we try to understand the main features of the three different algorithms in terms of predictive accuracy and capability of identification of relevant genetic information, and their sensitivity with respect to various kinds of quality control procedures and algorithmic parameters; iii.) a systematic study, where we confirm, on 7 different datasets from the Welcome-Trust-Consortium, the main outcomes of our study on the Crohn's disease, while using default parameter settings.

In addition to the main scientific questions that we have addressed during our research, we have also devoted a significant amount of efforts to develop our simulation and supervised learning software and to find out relevant ways of presenting the information extracted by these methods from the datasets. Notice that many of the side-simulations that we carried out during our work are not reported in this thesis. For example, we did not report the non-conclusive results of our study of the average genetic distances within cases and control sub-cohorts, a study that was aimed at finding out whether sample "stratification" was present and responsible of some of the abnormally high AUC values that we observed. We also tried out different algorithmic versions of the supervised learning methods, before settling our final choices documented in this manuscript. In particular, during these latter investigations we found out that the kind of normalisation of the splitting score measures used in tree-based methods may have a major impact on their intrinsic properties (some of these results are reported in Appendix A), and hence led us to choose the least biased normalisation in order to carry out our tests in Part II of the thesis.

8.2 Main findings

- Overall, the paradigm of tree-based ensemble supervised learning methods that we investigated in this thesis constitutes an interesting approach in the GWAS context, due to their intrinsic algorithmic properties. Some of their core features have shown to be particularly well suited in this context, since they were able to effectively classify individuals given their genotypes, and at the same time provided information for ranking genetic markers in terms of their relevance in these predictive models.

- We are confident in the observations we made along this thesis, as they generalised and remained consistent across many datasets and experiences. In particular, we found out that these methods are potentially very sensitive to particular types of variables (rare variants, markers deviating from HWE), but we also observed that the overall decision tree forest behaviour stayed stable across the different datasets and experiences. Hence it is important to very well document the quality control procedures used to pre-process datasets when these latter are exploited by supervised learning methods.

- We also found out that the way the splitting score measure is normalised may have a major impact on the outcome of applying these tree-based ensemble methods in the context of genome-wide-association-studies. In particular, the classical normalisation (called gain-ratio) used by many researchers appeared to be highly biased towards the selection of markers with small minor allele frequencies, while the other normalisations are much more agnostic in this respect. We believe that this finding is an important one, since it links some internal "algorithmic details" (often neglected by researchers) to some of the main outcomes of these methods (i.e. the variable importances). Hence it is important to report in a very precise way all algorithmic details about the supervised learning methods used to exploit GWAS datasets.

- In terms of the predictive power, as assessed by our cross-validation protocol over multiple datasets, we found that the T-Trees method that we proposed in an attempt to take into account the block-wise linkage disequilibrium structure of GWAS, outperforms in a significant and consistent way both Random Forests and Extra-Trees. Hence the exploitation of the structure of input variable dependencies is a credible avenue for tailoring supervised learning methods to GWAS datasets.

- Our empirical studies with the T-Trees method allowed us to identify two novel susceptibility loci in the context of Crohn's disease, which according to our analysis are potentially relevant from the biological point of view, and certainly indicative that this method approaches the datasets in a quite different way than the classical state-of-the-art Random Forest types of methods. Hence, biologists might be interested by further analysing the actual relevance of these novel susceptibility regions and by using the T-Trees method to look at their own datasets.

8.3 Further work

With hindsight, our research certainly opens more new questions than it has been able to answer. In the following lines we will organise the discussion of these many open questions into three parts, namely first those more specifically related to tree-based supervised learning algorithm development, then those related to the extension of our framework to the broad context of the genetical dissection of diseases and other complex phenotypes, and finally those related to practical implementation and use of our algorithms.

8.3.1 Tree-based supervised learning methods

There are many possibilities for further enhancing the tree-based methods in order to better exploit the correlation structures among input variables of complex datasets.

In particular, we would like to explicitly take into account the observed correlation structure in a given dataset when defining the blocks of input variables exploited by the T-Trees method. This could be achieved by combining in some way unsupervised learning methods (like clustering and Bayesian networks) with the tree-based supervised learning techniques.

We would also like to extend our T-Trees method to other problem domains revealing similar geometrical correlation structures, such as time-series classification and image classification. Within this latter context, a comparison between our approach and the so-called 'Segment and Combine' paradigm already successfully used in time-series and image classification [GMW06] would be of great interest.

Finally, we believe that from a more theoretical point of view, it would be interesting to compare the two-level input variable handling approach used by our T-Trees method (and its possible variants) with the two-level structure of Group-LASSO-based supervised learning methods [MVDGB08, JOV09]. In particular, it would be of interest to compare these latter methods with our own algorithm on our datasets used in this thesis.

8.3.2 Genetical dissection of complex phenotypes by supervised learning

In this context, we believe that one first practical problem that should be tackled in a better way is to cope in a non ad-hoc way with missing information. In our research we have not worked on this aspect, but in our applications we had to fill in missing values in a rough way so as to be able to apply our algorithms. There should be more elegant ways to treat this aspect of missing values, and many of these have been reported in the literature, but not really investigated in the practice of GWAS.

Second, the methods that we have developed have focused on the study of binary (e.g. case/control) traits, while non-binary and even quantitative traits have also to be studied. The adaptation of our algorithms is more or less straightforward in principle; however, the empirical findings that we have reported have to be analysed carefully when generalising the approach to non-binary traits. Specially, in the context of complex diseases it is likely that the binary phenotype is only a very approximate representation of the information that is collected by physicians, and it is certainly of interest to hypothesise that such complex diseases are actually composed of many sub-syndromes, and if information were available about these sub-syndromes it could as well be exploited by supervised learning.

The fact that the individuals collected in GWAS studies can not in practice be assumed to be sampled in an i.i.d. fashion from a general population, notably because they are to a smaller or larger extent belonging to the same 'family', and because of artefacts introduced by sample collection constraints and measurement protocols, in particular due to the fact that environmental factors can not be fully eliminated by the experiment design, make the interpretation of the findings of multi-variable approaches difficult in comparison with the traditional uni-variate approaches. Further work has to be carried out in order to find out ways to cancel such 'sampling and measurement artefacts' when interpreting the notion of variable importance computed by supervised learning algorithms. Also, we recommend to collect all information about the cases and controls that could have some influence on the observed phenotype, and about the measurement protocols that could to some extent impact their recorded genetic information, so as to be then able to look at this information in conjunction with the genotypes, when analysing the datasets. The modern supervised learning methods are indeed able to exploit in the same way all sources of information, be they of genetic, of environmental, or of experimental nature.

Given the progress in high-throughput sequencing technology, we can foresee that in the near future we will have at our disposal large cohorts of fully sequenced patients and controls. How to leverage the supervised learning methods, and in particular the tree-based ensemble methods studied in our thesis, to such scenarios is certainly the major research direction for the future in order to enable the best synergy between computational and experimental methods in the pursue of the understanding of complex biology.

8.3.3 Software development and implementation

In spite of all the future research directions discussed above, we believe that the T-Trees method that we have developed, is in its current form already of practical interest, together with the Random Forests and the Extra-Trees.

Therefore, we believe that it would be of interest to pack our software modules in such a way that they could be easily exploited by other researchers, and automatically take advantage of the growing grid-computing environments.

To achieve this in the best way, we would need to collaborate with potential end-users and specialists in the grid-computing technology.

Appendices

Appendix A

Additional data related to Chapter 6

For the sake of readability and to keep the Chapter 6 uncluttered, we report here additional results gathered for the investigation conducted in Chapter 6.

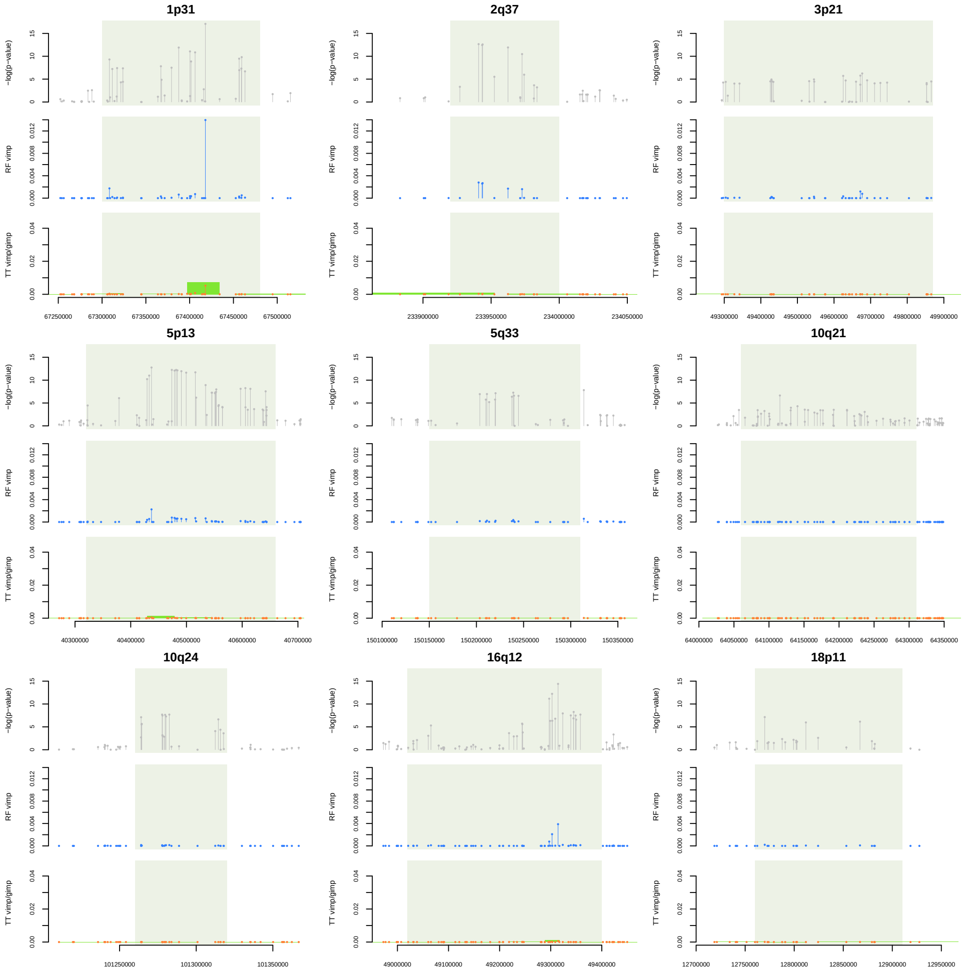


Figure A.1 The nine reported regions for Crohn's disease on CD_{ibd}. In grey, the exact Fisher test based p-values, in blue random forest variable importances ($K = 10000$, $T = 1000$ and $N_{min} = 250$), in orange the T-Trees variable importances while the green boxes denote the T-Trees group importances ($T = 1000$, $K = 1000$, $IC = 5$ and $N_{min} = 2000$). The light grey shaded boxes delimit the nine regions as reported.

#	Chr	Pos.	SNP	RF imp.	Fisher p-value	χ^2 p-value	MAF	f_{miss}	pmiss	0	1	2	HWE
1	4	16454556	rs157613	$9.02 \cdot 10^{-2}$	$1.13 \cdot 10^{-14}$	$3.93 \cdot 10^{-15}$	0.0821	0.0141	$6.08 \cdot 10^{-1}$	190	379	4051	$2.14 \cdot 10^{-121}$
								0.0133		0	376	2523	$3.96 \cdot 10^{-6}$
								0.0155		190	3	1528	$3.16 \cdot 10^{-253}$
2	1	117184091	rs12078461	$4.95 \cdot 10^{-2}$	$7.19 \cdot 10^{-13}$	$2.37 \cdot 10^{-13}$	0.0471	0.0100	$1.57 \cdot 10^{-11}$	116	205	4318	$2.14 \cdot 10^{-107}$
								0.002 38		0	204	2727	$4.91 \cdot 10^{-2}$
								0.0229		116	1	1591	$2.51 \cdot 10^{-182}$
3	1	214857356	rs1933641	$4.90 \cdot 10^{-2}$	$7.59 \cdot 10^{-14}$	$1.68 \cdot 10^{-14}$	0.0467	0.0196	$2.25 \cdot 10^{-17}$	115	199	4280	$8.68 \cdot 10^{-108}$
								0.006 13		0	198	2722	$7.90 \cdot 10^{-2}$
								0.0423		115	1	1558	$3.76 \cdot 10^{-180}$
4	4	17872130	rs1553460	$4.09 \cdot 10^{-2}$	$1.59 \cdot 10^{-31}$	$6.55 \cdot 10^{-32}$	0.315	0.0540	$3.05 \cdot 10^{-46}$	699	1395	2339	$1.58 \cdot 10^{-70}$
								0.0848		237	969	1483	$2.88 \cdot 10^{-5}$
								0.002 29		462	426	856	$5.84 \cdot 10^{-93}$
5	16	30227808	rs4471699	$3.27 \cdot 10^{-2}$	$4.64 \cdot 10^{-20}$	$5.03 \cdot 10^{-20}$	0.449	0.0374	$4.20 \cdot 10^{-26}$	774	2498	1239	$1.10 \cdot 10^{-15}$
								0.0143		672	1461	763	$6.03 \cdot 10^{-1}$
								0.0761		102	1037	476	$2.31 \cdot 10^{-49}$
6	4	38783368	rs6816863	$1.49 \cdot 10^{-2}$	$2.86 \cdot 10^{-1}$	$2.73 \cdot 10^{-1}$	0.0267	0.009 82	$4.14 \cdot 10^{-5}$	49	150	4441	$3.99 \cdot 10^{-48}$
								0.005 11		1	146	2776	1
								0.0177		48	4	1665	$4.18 \cdot 10^{-89}$
7	16	30235818	rs11644392	$1.34 \cdot 10^{-2}$	$2.20 \cdot 10^{-1}$	$2.14 \cdot 10^{-1}$	0.484	0.0147	$6.16 \cdot 10^{-1}$	1091	2287	1239	$5.76 \cdot 10^{-1}$
								0.0140		698	1437	762	$6.83 \cdot 10^{-1}$
								0.0160		393	850	477	$6.99 \cdot 10^{-1}$
8	11	113306801	rs17116117	$1.12 \cdot 10^{-2}$	$1.03 \cdot 10^{-23}$	$6.85 \cdot 10^{-25}$	0.0487	0.0226	$1.23 \cdot 10^{-28}$	1	444	4135	$3.11 \cdot 10^{-4}$
								0.004 08		0	183	2743	$1.16 \cdot 10^{-1}$
								0.0538		1	261	1392	$2.92 \cdot 10^{-4}$
9	10	125667027	rs7067790	$7.82 \cdot 10^{-3}$	$1.10 \cdot 10^{-7}$	$1.12 \cdot 10^{-7}$	0.391	0.0397	$2.00 \cdot 10^{-16}$	616	2287	1597	$7.51 \cdot 10^{-6}$
								0.0211		499	1369	1008	$3.56 \cdot 10^{-1}$
								0.0709		117	918	589	$4.18 \cdot 10^{-22}$
10	1	67417979	**rs11209026**	$6.89 \cdot 10^{-3}$	$5.43 \cdot 10^{-18}$	$1.45 \cdot 10^{-16}$	0.0450	0.006 62	$4.43 \cdot 10^{-3}$	0	419	4236	$1.02 \cdot 10^{-4}$
								0.009 19		0	342	2569	$3.53 \cdot 10^{-5}$
								0.002 29		0	77	1667	1
11	4	186107089	rs13126272	$6.24 \cdot 10^{-3}$	$3.65 \cdot 10^{-5}$	$3.39 \cdot 10^{-5}$	0.338	0.0578	$1.26 \cdot 10^{-28}$	685	1612	2118	$1.37 \cdot 10^{-33}$
								0.0844		302	1123	1265	$3.02 \cdot 10^{-2}$
								0.0132		383	489	853	$5.00 \cdot 10^{-58}$
12	3	18596095	rs12714959	$5.42 \cdot 10^{-3}$	$5.06 \cdot 10^{-7}$	$5.48 \cdot 10^{-7}$	0.263	0.0546	$3.88 \cdot 10^{-1}$	200	1926	2304	$5.21 \cdot 10^{-17}$
								0.0524		177	1208	1399	$7.82 \cdot 10^{-5}$
								0.0584		23	718	905	$1.74 \cdot 10^{-23}$
13	2	25364557	rs2164411	$5.33 \cdot 10^{-3}$	$4.82 \cdot 10^{-3}$	$4.75 \cdot 10^{-3}$	0.155	0.0305	$4.35 \cdot 10^{-2}$	89	1234	3220	$2.03 \cdot 10^{-2}$
								0.0266		86	764	2010	$1.95 \cdot 10^{-1}$
								0.0372		3	470	1210	$2.85 \cdot 10^{-13}$
14	5	117033845	rs2416472	$4.55 \cdot 10^{-3}$	$7.19 \cdot 10^{-12}$	$8.42 \cdot 10^{-12}$	0.319	0.008 96	$1.22 \cdot 10^{-6}$	551	1860	2233	$1.33 \cdot 10^{-7}$
								0.0136		345	1307	1246	$9.35 \cdot 10^{-1}$
								0.001 14		206	553	987	$2.40 \cdot 10^{-17}$
15	16	30293004	rs11863150	$4.13 \cdot 10^{-3}$	$6.41 \cdot 10^{-1}$	$6.32 \cdot 10^{-1}$	0.366	0.001 71	$2.72 \cdot 10^{-1}$	611	2198	1869	$3.95 \cdot 10^{-1}$
								0.002 38		379	1374	1178	$4.98 \cdot 10^{-1}$
								0.000 572		232	824	691	$6.07 \cdot 10^{-1}$
16	1	87862563	rs17130103	$4.08 \cdot 10^{-3}$	$1.10 \cdot 10^{-12}$	$2.34 \cdot 10^{-10}$	0.0101	0.0109	$3.54 \cdot 10^{-19}$	0	94	4541	1
								0.000 681		0	89	2847	1
								0.0280		0	5	1694	1
17	14	83044749	rs10144243	$3.87 \cdot 10^{-3}$	$2.25 \cdot 10^{-12}$	$1.31 \cdot 10^{-12}$	0.005 27	0.007 47	$4.34 \cdot 10^{-9}$	0	49	4602	1
								0.001 70		0	7	2926	1
								0.0172		0	42	1676	1
18	10	125667065	rs17680424	$3.44 \cdot 10^{-3}$	$9.83 \cdot 10^{-1}$	$9.74 \cdot 10^{-1}$	0.411	0.000 854	$6.32 \cdot 10^{-1}$	790	2266	1626	1
								0.000 681		500	1411	1025	$7.03 \cdot 10^{-1}$
								0.001 14		290	855	601	$6.57 \cdot 10^{-1}$
19	2	132482534	rs4080478	$3.27 \cdot 10^{-3}$	$3.64 \cdot 10^{-3}$	$3.35 \cdot 10^{-3}$	0.104	0.0305	$5.30 \cdot 10^{-2}$	59	829	3655	$1.31 \cdot 10^{-1}$
								0.0344		10	530	2297	$1.44 \cdot 10^{-4}$
								0.0240		49	299	1358	$2.69 \cdot 10^{-8}$
20	4	16453444	rs150260	$2.85 \cdot 10^{-3}$	$9.32 \cdot 10^{-2}$	$9.08 \cdot 10^{-2}$	0.0696	0.000 854	$6.32 \cdot 10^{-1}$	21	610	4051	$8.21 \cdot 10^{-1}$
								0.000 681		14	401	2521	$7.85 \cdot 10^{-1}$
								0.001 14		7	209	1530	1

TABLE A.1 The 20 first markers according to the random forest variable importances (denoted RF imp. in the gray shaded column) on the CD_{wtccc} datasets and the corresponding statistics. Green shaded cells refer to statistics related to controls only while red shaded cells refer to cases only statistics.

#	Chr	Pos.	SNP	**Fisher p-value**	χ^2 p-value	MAF	f_{miss}	pmiss	0	1	2	HWE
1	4	17872130	rs1553460	$1.59 \cdot 10^{-31}$	$6.55 \cdot 10^{-32}$	0.315	0.0540	$3.05 \cdot 10^{-46}$	699	1395	2339	$1.58 \cdot 10^{-70}$
							0.0848		237	969	1483	$2.88 \cdot 10^{-5}$
							0.002 29		462	426	856	$5.84 \cdot 10^{-93}$
2	11	113306801	rs17116117	$1.03 \cdot 10^{-23}$	$6.85 \cdot 10^{-25}$	0.0487	0.0226	$1.23 \cdot 10^{-28}$	1	444	4135	$3.11 \cdot 10^{-4}$
							0.004 08		0	183	2743	$1.16 \cdot 10^{-1}$
							0.0538		1	261	1392	$2.92 \cdot 10^{-4}$
3	16	30227808	rs4471699	$4.64 \cdot 10^{-20}$	$5.03 \cdot 10^{-20}$	0.449	0.0374	$4.20 \cdot 10^{-26}$	774	2498	1239	$1.10 \cdot 10^{-15}$
							0.0143		672	1461	763	$6.03 \cdot 10^{-1}$
							0.0761		102	1037	476	$2.31 \cdot 10^{-49}$
4	1	67417979	**rs11209026**	$5.43 \cdot 10^{-18}$	$1.45 \cdot 10^{-16}$	0.0450	0.006 62	$4.43 \cdot 10^{-3}$	0	419	4236	$1.02 \cdot 10^{-4}$
							0.009 19		0	342	2569	$3.53 \cdot 10^{-5}$
							0.002 29		0	77	1667	1
5	16	49314382	**rs2076756**	$3.00 \cdot 10^{-15}$	$1.76 \cdot 10^{-15}$	0.270	0.007 90	$1.67 \cdot 10^{-2}$	374	1762	2513	$9.18 \cdot 10^{-3}$
							0.005 45		174	1065	1683	$7.62 \cdot 10^{-1}$
							0.0120		200	697	830	$4.61 \cdot 10^{-3}$
6	4	16454556	rs157613	$1.13 \cdot 10^{-14}$	$3.93 \cdot 10^{-15}$	0.0821	0.0141	$6.08 \cdot 10^{-1}$	190	379	4051	$2.14 \cdot 10^{-121}$
							0.0133		0	376	2523	$3.96 \cdot 10^{-6}$
							0.0155		190	3	1528	$3.16 \cdot 10^{-253}$
7	1	214857356	rs1933641	$7.59 \cdot 10^{-14}$	$1.68 \cdot 10^{-14}$	0.0467	0.0196	$2.25 \cdot 10^{-17}$	115	199	4280	$8.68 \cdot 10^{-108}$
							0.006 13		0	198	2722	$7.90 \cdot 10^{-2}$
							0.0423		115	1	1558	$3.76 \cdot 10^{-180}$
8	2	233940839	**rs10210302**	$1.08 \cdot 10^{-13}$	$1.08 \cdot 10^{-13}$	0.451	0.000 427	$5.32 \cdot 10^{-1}$	936	2354	1394	$3.16 \cdot 10^{-1}$
							0.000 681		646	1530	760	$1.98 \cdot 10^{-2}$
							0		290	824	634	$4.26 \cdot 10^{-1}$
9	2	233943769	**rs6431654**	$1.24 \cdot 10^{-13}$	$1.29 \cdot 10^{-13}$	0.450	0.001 07	$6.57 \cdot 10^{-1}$	933	2350	1398	$3.45 \cdot 10^{-1}$
							0.001 36		645	1525	764	$2.65 \cdot 10^{-2}$
							0.000 572		288	825	634	$4.86 \cdot 10^{-1}$
10	4	38445177	rs17615966	$1.71 \cdot 10^{-13}$	$6.54 \cdot 10^{-13}$	0.003 66	0.008 32	$4.15 \cdot 10^{-3}$	16	2	4629	$3.57 \cdot 10^{-44}$
							0.0112		0	1	2904	1
							0.003 43		16	1	1725	$1.45 \cdot 10^{-38}$
11	2	233943448	**rs6752107**	$1.76 \cdot 10^{-13}$	$1.83 \cdot 10^{-13}$	0.451	0.000 427	$5.32 \cdot 10^{-1}$	937	2353	1394	$3.30 \cdot 10^{-1}$
							0.000 681		646	1529	761	$1.99 \cdot 10^{-2}$
							0		291	824	633	$4.26 \cdot 10^{-1}$
12	12	127945945	rs11060028	$1.95 \cdot 10^{-13}$	$6.98 \cdot 10^{-13}$	0.102	0.0668	$2.42 \cdot 10^{-34}$	30	836	3507	$8.28 \cdot 10^{-3}$
							0.0317		29	622	2194	$4.01 \cdot 10^{-2}$
							0.126		1	214	1313	$5.30 \cdot 10^{-3}$
13	5	40437266	**rs17234657**	$2.37 \cdot 10^{-13}$	$1.10 \cdot 10^{-13}$	0.146	0.001 92	$7.35 \cdot 10^{-1}$	113	1135	3429	$9.98 \cdot 10^{-2}$
							0.001 70		51	629	2253	$3.52 \cdot 10^{-1}$
							0.002 29		62	506	1176	$4.18 \cdot 10^{-1}$
14	1	67387537	**rs11805303**	$3.70 \cdot 10^{-13}$	$2.91 \cdot 10^{-13}$	0.345	0.001 28	1	589	2051	2040	$3.85 \cdot 10^{-2}$
							0.001 36		313	1236	1385	$1.36 \cdot 10^{-1}$
							0.001 14		276	815	655	$3.94 \cdot 10^{-1}$
15	2	233962410	**rs3828309**	$5.06 \cdot 10^{-13}$	$5.34 \cdot 10^{-13}$	0.452	0.002 99	$5.90 \cdot 10^{-1}$	935	2349	1388	$3.15 \cdot 10^{-1}$
							0.003 40		645	1522	761	$2.64 \cdot 10^{-2}$
							0.002 29		290	827	627	$5.51 \cdot 10^{-1}$
16	16	49302700	**rs2066843**	$5.46 \cdot 10^{-13}$	$3.58 \cdot 10^{-13}$	0.285	0.003 41	$1.27 \cdot 10^{-1}$	423	1817	2430	$2.01 \cdot 10^{-3}$
							0.002 38		206	1106	1619	$3.61 \cdot 10^{-1}$
							0.005 15		217	711	811	$1.98 \cdot 10^{-3}$
17	1	117184091	rs12078461	$7.19 \cdot 10^{-13}$	$2.37 \cdot 10^{-13}$	0.0471	0.0100	$1.57 \cdot 10^{-11}$	116	205	4318	$2.14 \cdot 10^{-107}$
							0.002 38		0	204	2727	$4.91 \cdot 10^{-2}$
							0.0229		116	1	1591	$2.51 \cdot 10^{-182}$
18	3	16454562	rs9839841	$7.20 \cdot 10^{-13}$	$4.09 \cdot 10^{-13}$	0.195	0.0414	$1.54 \cdot 10^{-20}$	136	1478	2878	$1.01 \cdot 10^{-3}$
							0.0201		80	831	1968	$5.14 \cdot 10^{-1}$
							0.0772		56	647	910	$2.30 \cdot 10^{-6}$
19	5	40473705	**rs9292777**	$7.53 \cdot 10^{-13}$	$8.86 \cdot 10^{-13}$	0.367	0.001 71	1	661	2107	1910	$4.07 \cdot 10^{-2}$
							0.001 70		468	1375	1090	$3.34 \cdot 10^{-1}$
							0.001 72		193	732	820	$1.24 \cdot 10^{-1}$
20	1	87862563	rs17130103	$1.10 \cdot 10^{-12}$	$2.34 \cdot 10^{-10}$	0.0101	0.0109	$3.54 \cdot 10^{-19}$	0	94	4541	1
							0.000 681		0	89	2847	1
							0.0280		0	5	1694	1

TABLE A.2 The 20 first markers according to the Fisher p-value on the CD_{wtccc} datasets and the corresponding statistics. Green shaded cells refer to statistics related to controls only while red shaded cells refer to cases only statistics.

#	Chr	Pos.	SNP	Fisher p-value	χ^2 p-value	MAF	f_{miss}	pmiss	0	1	2	HWE
1	1	67417979	**rs11209026**	$8.24 \cdot 10^{-18}$	$2.09 \cdot 10^{-16}$	0.0451	0.006 63	$4.45 \cdot 10^{-3}$	0	419	4226	$1.03 \cdot 10^{-4}$
							0.009 19		0	342	2568	$3.52 \cdot 10^{-5}$
							0.002 30		0	77	1658	1
2	16	49314382	**rs2076756**	$3.95 \cdot 10^{-15}$	$2.25 \cdot 10^{-15}$	0.270	0.007 91	$1.64 \cdot 10^{-2}$	373	1758	2508	$9.11 \cdot 10^{-3}$
							0.005 45		174	1065	1682	$7.62 \cdot 10^{-1}$
							0.0121		199	693	826	$4.50 \cdot 10^{-3}$
3	5	40437266	**rs17234657**	$1.72 \cdot 10^{-13}$	$8.09 \cdot 10^{-14}$	0.146	0.001 71	1	113	1132	3423	$9.90 \cdot 10^{-2}$
							0.001 70		51	628	2253	$3.51 \cdot 10^{-1}$
							0.001 73		62	504	1170	$4.18 \cdot 10^{-1}$
4	2	233940839	**rs10210302**	$2.22 \cdot 10^{-13}$	$2.31 \cdot 10^{-13}$	0.452	0.000 428	$5.33 \cdot 10^{-1}$	936	2349	1389	$3.30 \cdot 10^{-1}$
							0.000 681		646	1529	760	$1.98 \cdot 10^{-2}$
							0		290	820	629	$4.25 \cdot 10^{-1}$
5	2	233943769	**rs6431654**	$2.55 \cdot 10^{-13}$	$2.75 \cdot 10^{-13}$	0.451	0.001 07	$6.57 \cdot 10^{-1}$	933	2345	1393	$3.60 \cdot 10^{-1}$
							0.001 36		645	1524	764	$2.91 \cdot 10^{-2}$
							0.000 575		288	821	629	$4.85 \cdot 10^{-1}$
6	2	233943448	**rs6752107**	$3.61 \cdot 10^{-13}$	$3.89 \cdot 10^{-13}$	0.452	0.000 428	$5.33 \cdot 10^{-1}$	937	2348	1389	$3.45 \cdot 10^{-1}$
							0.000 681		646	1528	761	$2.19 \cdot 10^{-2}$
							0		291	820	628	$3.97 \cdot 10^{-1}$
7	5	40481438	**rs11957215**	$5.78 \cdot 10^{-13}$	$6.87 \cdot 10^{-13}$	0.299	0.001 71	$2.72 \cdot 10^{-1}$	436	1923	2309	$2.22 \cdot 10^{-1}$
							0.002 38		318	1272	1340	$5.28 \cdot 10^{-1}$
							0.000 575		118	651	969	$5.29 \cdot 10^{-1}$
8	5	40473705	**rs9292777**	$5.79 \cdot 10^{-13}$	$6.26 \cdot 10^{-13}$	0.366	0.001 71	1	659	2102	1907	$4.05 \cdot 10^{-2}$
							0.001 70		468	1374	1090	$3.15 \cdot 10^{-1}$
							0.001 73		191	728	817	$1.36 \cdot 10^{-1}$
9	16	49302700	**rs2066843**	$5.96 \cdot 10^{-13}$	$4.26 \cdot 10^{-13}$	0.285	0.003 42	$1.26 \cdot 10^{-1}$	422	1813	2425	$1.99 \cdot 10^{-3}$
							0.002 38		206	1106	1618	$3.61 \cdot 10^{-1}$
							0.005 18		216	707	807	$1.92 \cdot 10^{-3}$
10	5	40483754	**rs4957295**	$7.32 \cdot 10^{-13}$	$1.00 \cdot 10^{-12}$	0.301	0.001 71	$4.80 \cdot 10^{-1}$	442	1924	2302	$1.75 \cdot 10^{-1}$
							0.001 36		320	1277	1336	$5.85 \cdot 10^{-1}$
							0.002 30		122	647	966	$3.45 \cdot 10^{-1}$
11	5	40478626	**rs10213846**	$8.04 \cdot 10^{-13}$	$1.03 \cdot 10^{-12}$	0.299	0.000 855	$6.31 \cdot 10^{-1}$	436	1921	2315	$1.97 \cdot 10^{-1}$
							0.000 681		317	1273	1345	$5.56 \cdot 10^{-1}$
							0.001 15		119	648	970	$4.49 \cdot 10^{-1}$
12	5	40490831	**rs4957297**	$1.12 \cdot 10^{-12}$	$1.41 \cdot 10^{-12}$	0.299	0.001 28	$2.94 \cdot 10^{-2}$	438	1918	2314	$1.63 \cdot 10^{-1}$
							0.000 341		317	1274	1345	$5.56 \cdot 10^{-1}$
							0.002 88		121	644	969	$3.44 \cdot 10^{-1}$
13	2	233962410	**rs3828309**	$1.19 \cdot 10^{-12}$	$1.12 \cdot 10^{-12}$	0.452	0.002 99	$5.90 \cdot 10^{-1}$	935	2344	1383	$3.15 \cdot 10^{-1}$
							0.003 41		645	1521	761	$2.89 \cdot 10^{-2}$
							0.002 30		290	823	622	$5.18 \cdot 10^{-1}$
14	1	67387537	**rs11805303**	$1.24 \cdot 10^{-12}$	$1.02 \cdot 10^{-12}$	0.344	0.001 28	1	583	2049	2038	$5.58 \cdot 10^{-2}$
							0.001 36		313	1235	1385	$1.36 \cdot 10^{-1}$
							0.001 15		270	814	653	$5.45 \cdot 10^{-1}$
15	5	40515944	**rs6871834**	$1.96 \cdot 10^{-12}$	$2.32 \cdot 10^{-12}$	0.296	0.000 428	1	430	1909	2335	$1.61 \cdot 10^{-1}$
							0.000 341		309	1271	1356	$6.72 \cdot 10^{-1}$
							0.000 575		121	638	979	$2.28 \cdot 10^{-1}$
16	5	40499496	**rs4957300**	$2.30 \cdot 10^{-12}$	$2.67 \cdot 10^{-12}$	0.300	0.001 07	$1.65 \cdot 10^{-1}$	438	1922	2311	$1.85 \cdot 10^{-1}$
							0.001 70		316	1274	1342	$6.14 \cdot 10^{-1}$
							0		122	648	969	$3.46 \cdot 10^{-1}$
17	16	49297083	**rs17221417**	$6.88 \cdot 10^{-12}$	$5.53 \cdot 10^{-12}$	0.313	0.000 855	$6.31 \cdot 10^{-1}$	499	1925	2248	$4.72 \cdot 10^{-3}$
							0.000 681		256	1175	1504	$2.24 \cdot 10^{-1}$
							0.001 15		243	750	744	$1.60 \cdot 10^{-2}$
18	1	67400370	**rs10489629**	$8.65 \cdot 10^{-12}$	$9.15 \cdot 10^{-12}$	0.430	0.0182	$5.74 \cdot 10^{-1}$	859	2227	1505	$4.89 \cdot 10^{-1}$
							0.0191		614	1404	863	$3.48 \cdot 10^{-1}$
							0.0167		245	823	642	$5.06 \cdot 10^{-1}$
19	5	40433109	**rs16869934**	$1.00 \cdot 10^{-11}$	$1.28 \cdot 10^{-11}$	0.267	0.001 07	$3.68 \cdot 10^{-1}$	349	1796	2526	$2.32 \cdot 10^{-1}$
							0.000 681		252	1203	1480	$7.54 \cdot 10^{-1}$
							0.001 73		97	593	1046	$3.04 \cdot 10^{-1}$
20	1	67406223	**rs2201841**	$1.41 \cdot 10^{-11}$	$1.15 \cdot 10^{-11}$	0.345	0	1	564	2094	2018	$5.60 \cdot 10^{-1}$
							0		311	1251	1375	$2.89 \cdot 10^{-1}$
							0		253	843	643	$4.19 \cdot 10^{-1}$

Table A.3 The 20 first markers according to the Fisher p-value on the CD_{ibd} datasets and the corresponding statistics. Green shaded cells refer to statistics related to controls only while red shaded cells refer to cases only statistics.

Chr	Pos.	SNP	RF imp.	TT imp.	Fisher p-value	χ^2 p-value	MAF	f_{miss}	pmiss	0	1	2	HWE
2	81542598	rs11688716	$1 \cdot 10^{-6}$	$6 \cdot 10^{-6}$	$7.65 \cdot 10^{-3}$	$7.45 \cdot 10^{-3}$	0.274	0.007 47	$1.65 \cdot 10^{-1}$	358	1832	2461	$5.07 \cdot 10^{-1}$
								0.008 85		243	1165	1504	$4.12 \cdot 10^{-1}$
								0.005 15		115	667	957	1
2	81542876	rs10211262	$1 \cdot 10^{-6}$	$1 \cdot 10^{-6}$	$7.17 \cdot 10^{-1}$	$7.63 \cdot 10^{-1}$	0.000 748	0.001 71	$2.72 \cdot 10^{-1}$	0	7	4671	1
								0.002 38		0	4	2927	1
								0.000 572		0	3	1744	1
2	81543370	rs6706111	$3 \cdot 10^{-6}$	$1.86 \cdot 10^{-4}$	$4.74 \cdot 10^{-3}$	$4.73 \cdot 10^{-3}$	0.283	0.002 77	$7.78 \cdot 10^{-1}$	431	1784	2458	$5.38 \cdot 10^{-5}$
								0.003 06		289	1140	1500	$1.13 \cdot 10^{-3}$
								0.002 29		142	644	958	$2.33 \cdot 10^{-2}$
2	81550580	rs11692929	0.	$1.22 \cdot 10^{-4}$	$4.65 \cdot 10^{-1}$	$4.53 \cdot 10^{-1}$	0.450	0.001 49	$4.36 \cdot 10^{-1}$	949	2310	1420	$8.59 \cdot 10^{-1}$
								0.001 02		602	1453	880	$9.70 \cdot 10^{-1}$
								0.002 29		347	857	540	$8.46 \cdot 10^{-1}$
2	81577055	rs17020238	$5.60 \cdot 10^{-5}$	$3.89 \cdot 10^{-4}$	$9.76 \cdot 10^{-1}$	$9.51 \cdot 10^{-1}$	0.152	0.007 90	1	114	1185	3350	$4.59 \cdot 10^{-1}$
								0.007 83		67	753	2095	1
								0.008 01		47	432	1255	$1.91 \cdot 10^{-1}$
2	81577090	rs17020239	$9.30 \cdot 10^{-5}$	$2.34 \cdot 10^{-3}$	$9.76 \cdot 10^{-1}$	$9.57 \cdot 10^{-1}$	0.148	0.002 56	$7.71 \cdot 10^{-1}$	109	1166	3399	$4.51 \cdot 10^{-1}$
								0.002 38		63	741	2127	$9.42 \cdot 10^{-1}$
								0.002 86		46	425	1272	$1.54 \cdot 10^{-1}$
2	81577812	rs11887827	$3.88 \cdot 10^{-4}$	$4.37 \cdot 10^{-3}$	$2.19 \cdot 10^{-8}$	$2.42 \cdot 10^{-8}$	0.311	0.002 77	$8.63 \cdot 10^{-2}$	499	1907	2267	$1.32 \cdot 10^{-3}$
								0.001 70		322	1300	1311	1
								0.004 58		177	607	956	$1.66 \cdot 10^{-7}$
2	81579712	rs17020244	$7.10 \cdot 10^{-5}$	$1.60 \cdot 10^{-3}$	$9.76 \cdot 10^{-1}$	$9.75 \cdot 10^{-1}$	0.149	0.000 854	$6.32 \cdot 10^{-1}$	111	1171	3400	$3.87 \cdot 10^{-1}$
								0.000 681		65	743	2128	1
								0.001 14		46	428	1272	$1.85 \cdot 10^{-1}$
2	81581046	rs12623313	$9.20 \cdot 10^{-5}$	$2.85 \cdot 10^{-3}$	$9.52 \cdot 10^{-1}$	$9.39 \cdot 10^{-1}$	0.148	0.001 71	$3.71 \cdot 10^{-4}$	110	1168	3400	$4.18 \cdot 10^{-1}$
								0		65	743	2130	1
								0.004 58		45	425	1270	$1.84 \cdot 10^{-1}$
2	81581185	rs10520335	$7.30 \cdot 10^{-5}$	$4.91 \cdot 10^{-4}$	$9.52 \cdot 10^{-1}$	$9.54 \cdot 10^{-1}$	0.148	0.003 41	$3.09 \cdot 10^{-1}$	109	1167	3394	$4.52 \cdot 10^{-1}$
								0.002 72		63	742	2125	$9.42 \cdot 10^{-1}$
								0.004 58		46	425	1269	$1.55 \cdot 10^{-1}$
2	81585945	rs7593114	$6.30 \cdot 10^{-5}$	$4.11 \cdot 10^{-3}$	$1.44 \cdot 10^{-1}$	$1.41 \cdot 10^{-1}$	0.326	0.002 13	$4.65 \cdot 10^{-2}$	498	2053	2125	$9.47 \cdot 10^{-1}$
								0.001 02		323	1300	1312	$9.67 \cdot 10^{-1}$
								0.004 01		175	753	813	1
2	81586635	rs9646997	$5.60 \cdot 10^{-5}$	$9.25 \cdot 10^{-4}$	$8.81 \cdot 10^{-1}$	$8.82 \cdot 10^{-1}$	0.149	0.002 13	$4.65 \cdot 10^{-2}$	111	1168	3397	$3.86 \cdot 10^{-1}$
								0.001 02		63	744	2128	$8.84 \cdot 10^{-1}$
								0.004 01		48	424	1269	$8.91 \cdot 10^{-2}$
2	81587078	rs11126813	$5.80 \cdot 10^{-5}$	$4.80 \cdot 10^{-5}$	1	$9.73 \cdot 10^{-1}$	0.148	0.002 99	$1.65 \cdot 10^{-1}$	110	1167	3395	$4.18 \cdot 10^{-1}$
								0.002 04		64	743	2125	1
								0.004 58		46	424	1270	$1.54 \cdot 10^{-1}$
2	81593778	rs7570013	$1 \cdot 10^{-6}$	$4 \cdot 10^{-6}$	$5.93 \cdot 10^{-1}$	$5.73 \cdot 10^{-1}$	0.176	0.001 49	$4.36 \cdot 10^{-1}$	152	1338	3189	$4.19 \cdot 10^{-1}$
								0.001 02		97	846	1992	$5.27 \cdot 10^{-1}$
								0.002 29		55	492	1197	$6.15 \cdot 10^{-1}$

TABLE A.4 CD_{wtccc} variable importances results on chromosome 2 around a tree based only selected region. The 10 gray shaded rows correspond to the most important block according to the T-Trees method in that region.

Chr	Pos.	SNP	RF imp.	TT imp.	Fisher p-value	χ^2 p-value	MAF	f_{miss}	pmiss	0	1	2	HWE
2	81577055	rs17020238	$1.77 \cdot 10^{-4}$	$3.10 \cdot 10^{-5}$	$9.52 \cdot 10^{-1}$	$9.38 \cdot 10^{-1}$	0.152	0.007 91	1	114	1182	3343	$4.25 \cdot 10^{-1}$
								0.007 83		67	753	2094	1
								0.008 05		47	429	1249	$1.61 \cdot 10^{-1}$
2	81577090	rs17020239	$2.25 \cdot 10^{-4}$	$3.60 \cdot 10^{-5}$	$9.28 \cdot 10^{-1}$	$9.30 \cdot 10^{-1}$	0.148	0.002 14	$7.53 \cdot 10^{-1}$	109	1165	3392	$4.51 \cdot 10^{-1}$
								0.002 38		63	741	2126	$9.42 \cdot 10^{-1}$
								0.001 73		46	424	1266	$1.54 \cdot 10^{-1}$
2	81577812	rs11887827	$1.27 \cdot 10^{-3}$	$1.03 \cdot 10^{-2}$	$2.42 \cdot 10^{-8}$	$2.73 \cdot 10^{-8}$	0.311	0.002 14	$5.15 \cdot 10^{-1}$	499	1903	2264	$1.04 \cdot 10^{-3}$
								0.001 70		322	1299	1311	1
								0.002 88		177	604	953	$1.54 \cdot 10^{-7}$
2	81579712	rs17020244	$1.64 \cdot 10^{-4}$	$5.66 \cdot 10^{-3}$	$9.76 \cdot 10^{-1}$	$9.69 \cdot 10^{-1}$	0.149	0.000 642	1	111	1169	3393	$3.86 \cdot 10^{-1}$
								0.000 681		65	743	2127	1
								0.000 575		46	426	1266	$1.57 \cdot 10^{-1}$
2	81581046	rs12623313	$2.38 \cdot 10^{-4}$	$4.30 \cdot 10^{-3}$	1	$9.87 \cdot 10^{-1}$	0.149	0.001 07	$7.09 \cdot 10^{-3}$	110	1168	3393	$4.18 \cdot 10^{-1}$
								0		65	743	2129	1
								0.002 88		45	425	1264	$2.16 \cdot 10^{-1}$
2	81581185	rs10520335	$3.10 \cdot 10^{-4}$	$3.52 \cdot 10^{-3}$	$9.04 \cdot 10^{-1}$	$9.06 \cdot 10^{-1}$	0.149	0.002 78	1	109	1167	3387	$4.87 \cdot 10^{-1}$
								0.002 72		63	742	2124	$9.42 \cdot 10^{-1}$
								0.002 88		46	425	1263	$1.56 \cdot 10^{-1}$
2	81585945	rs7593114	$2.34 \cdot 10^{-4}$	$9.39 \cdot 10^{-3}$	$1.44 \cdot 10^{-1}$	$1.41 \cdot 10^{-1}$	0.326	0.002 14	$4.60 \cdot 10^{-2}$	498	2046	2122	$8.94 \cdot 10^{-1}$
								0.001 02		323	1299	1312	$9.67 \cdot 10^{-1}$
								0.004 03		175	747	810	$9.12 \cdot 10^{-1}$
2	81586635	rs9646997	$2.34 \cdot 10^{-4}$	$1.40 \cdot 10^{-3}$	$8.33 \cdot 10^{-1}$	$8.34 \cdot 10^{-1}$	0.149	0.001 50	$4.36 \cdot 10^{-1}$	111	1168	3390	$3.86 \cdot 10^{-1}$
								0.001 02		63	744	2127	$8.84 \cdot 10^{-1}$
								0.002 30		48	424	1263	$8.99 \cdot 10^{-2}$
2	81587078	rs11126813	$8.00 \cdot 10^{-5}$	$7.28 \cdot 10^{-4}$	$9.76 \cdot 10^{-1}$	$9.79 \cdot 10^{-1}$	0.149	0.002 35	$5.51 \cdot 10^{-1}$	110	1167	3388	$4.18 \cdot 10^{-1}$
								0.002 04		64	743	2124	1
								0.002 88		46	424	1264	$1.55 \cdot 10^{-1}$
2	81593778	rs7570013	0.	$1.34 \cdot 10^{-4}$	$5.54 \cdot 10^{-1}$	$5.45 \cdot 10^{-1}$	0.175	0.001 07	1	152	1333	3186	$3.89 \cdot 10^{-1}$
								0.001 02		97	845	1992	$5.26 \cdot 10^{-1}$
								0.001 15		55	488	1194	$5.56 \cdot 10^{-1}$
2	81594169	rs17020301	$5 \cdot 10^{-6}$	$2.80 \cdot 10^{-3}$	$3.89 \cdot 10^{-1}$	$3.79 \cdot 10^{-1}$	0.0812	0.001 28	$6.77 \cdot 10^{-1}$	30	698	3942	1
								0.001 02		16	433	2485	$6.14 \cdot 10^{-1}$
								0.001 73		14	265	1457	$6.40 \cdot 10^{-1}$
2	81599721	rs12613517	$1.72 \cdot 10^{-4}$	$2.14 \cdot 10^{-3}$	$9.04 \cdot 10^{-1}$	$9.05 \cdot 10^{-1}$	0.149	0.001 28	1	110	1170	3390	$4.53 \cdot 10^{-1}$
								0.001 36		64	743	2126	1
								0.001 15		46	427	1264	$1.86 \cdot 10^{-1}$
2	81600746	rs11689930	$2.22 \cdot 10^{-4}$	$2.23 \cdot 10^{-4}$	$9.52 \cdot 10^{-1}$	$9.39 \cdot 10^{-1}$	0.149	0.000 642	1	110	1172	3391	$4.53 \cdot 10^{-1}$
								0.000 681		65	743	2127	1
								0.000 575		45	429	1264	$2.56 \cdot 10^{-1}$
2	81601323	rs1052397	$4 \cdot 10^{-6}$	$2.13 \cdot 10^{-4}$	$7.13 \cdot 10^{-1}$	$6.93 \cdot 10^{-1}$	0.0748	0.0124	$1.71 \cdot 10^{-1}$	15	661	3942	$1.90 \cdot 10^{-2}$
								0.0106		9	412	2485	$7.64 \cdot 10^{-2}$
								0.0155		6	249	1457	$2.28 \cdot 10^{-1}$

TABLE A.5 CD_{ibd} variable importances results on chromosome 2 around a tree based only selected region. The 10 gray shaded rows correspond to the most important block according to the T-Trees method in that region.

Appendix B

Additional data related to Chapter 7

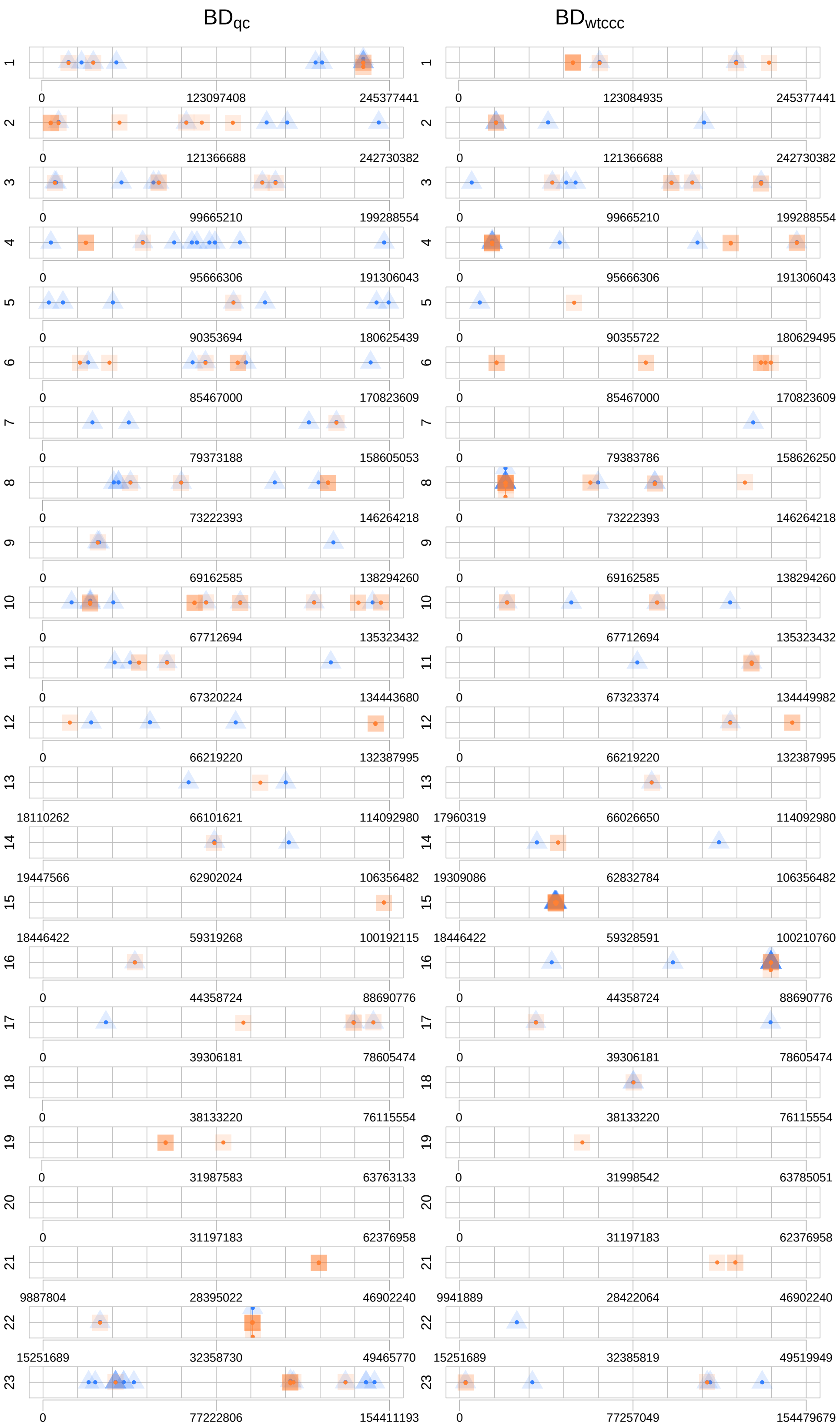


FIGURE B.1 *BD*: the position of the first 100 variables according to the tree based methods. Blue and triangle correspond to RF variable importances. Orange and square correspond to TT variable importances. As points overlap, opacity increases, indicating several hits concentrated in a region

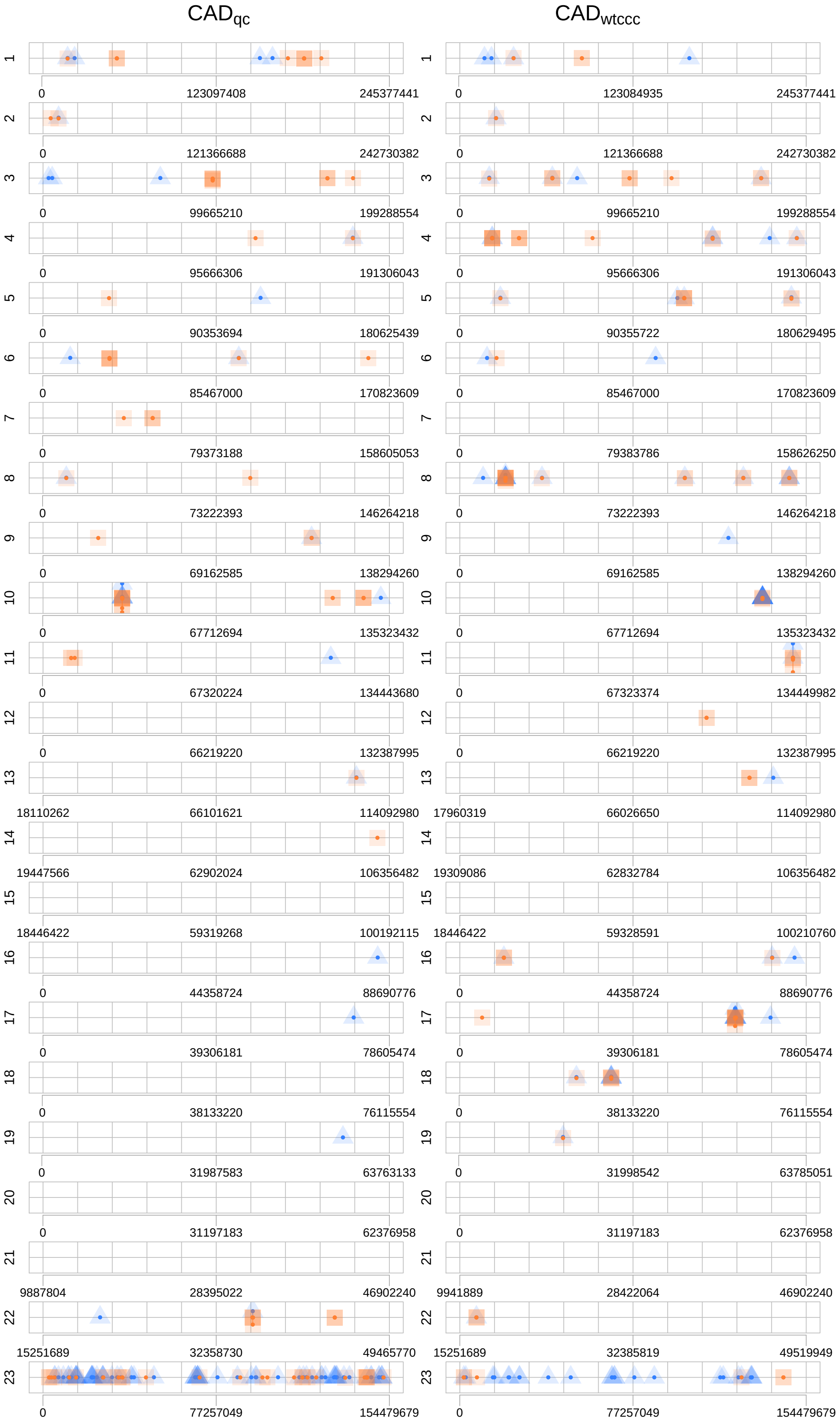


Figure B.2 *CAD*: the position of the first 100 variables according to the tree based methods. Blue and triangle correspond to RF variable importances. Orange and square correspond to TT variable importances. As points overlap, opacity increases, indicating several hits concentrated in a region

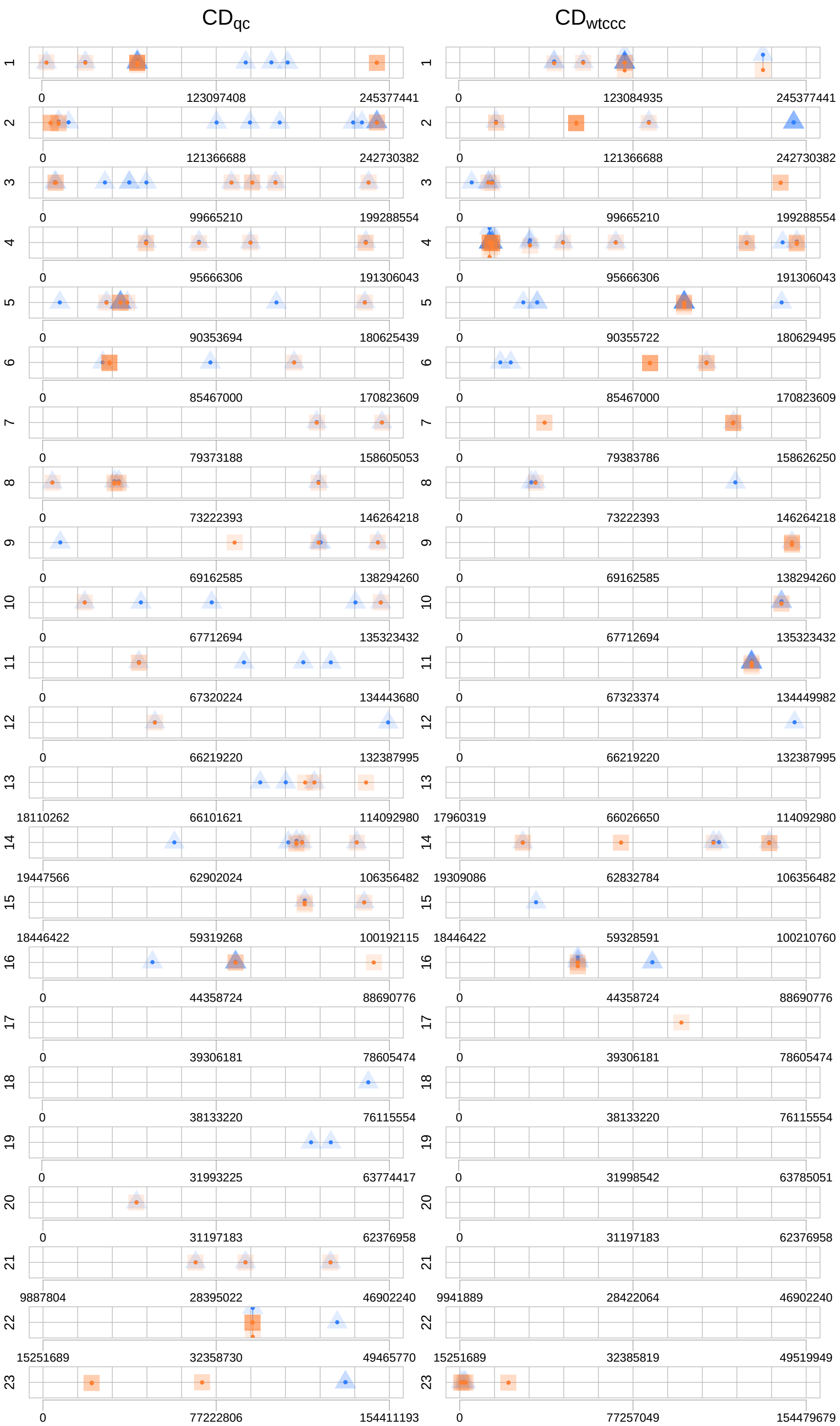


FIGURE B.3 *CD*: the position of the first 100 variables according to the tree based methods. Blue and triangle correspond to RF variable importances. Orange and square correspond to TT variable importances. As points overlap, opacity increases, indicating several hits concentrated in a region

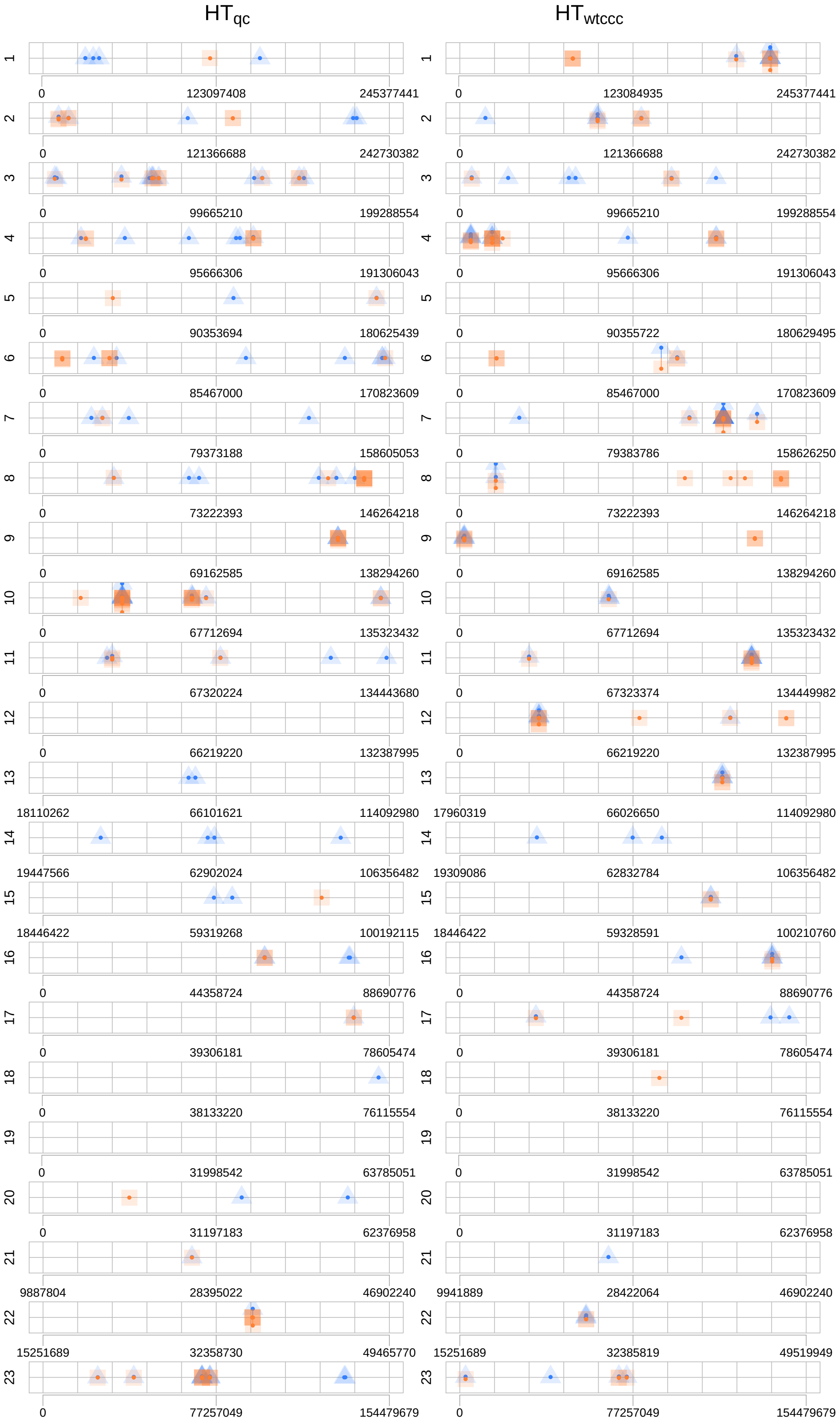


FIGURE B.4 *HT*: the position of the first 100 variables according to the tree based methods. Blue and triangle correspond to RF variable importances. Orange and square correspond to TT variable importances. As points overlap, opacity increases, indicating several hits concentrated in a region

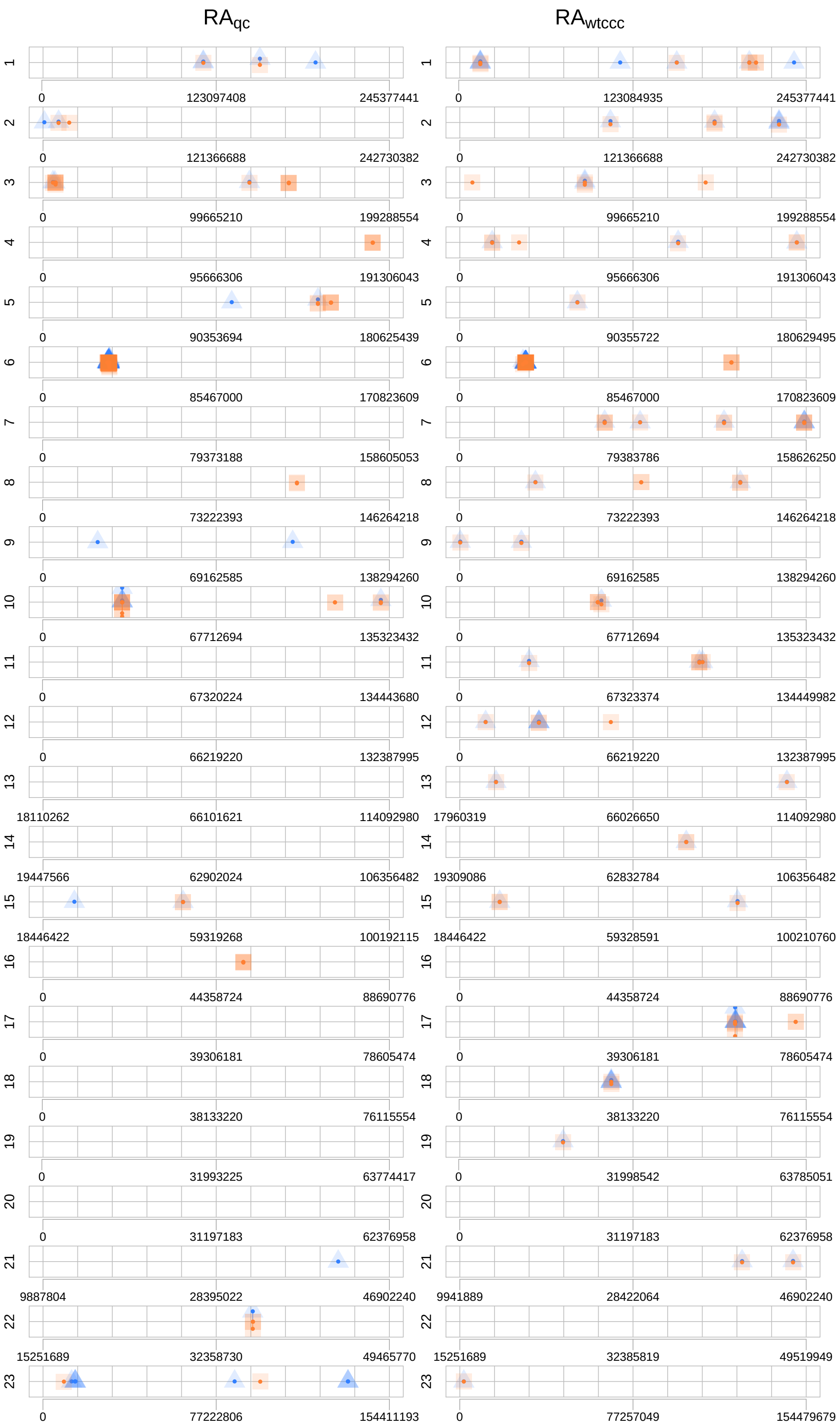


Figure B.5 *RA*: the position of the first 100 variables according to the tree based methods. Blue and triangle correspond to RF variable importances. Orange and square correspond to TT variable importances. As points overlap, opacity increases, indicating several hits concentrated in a region

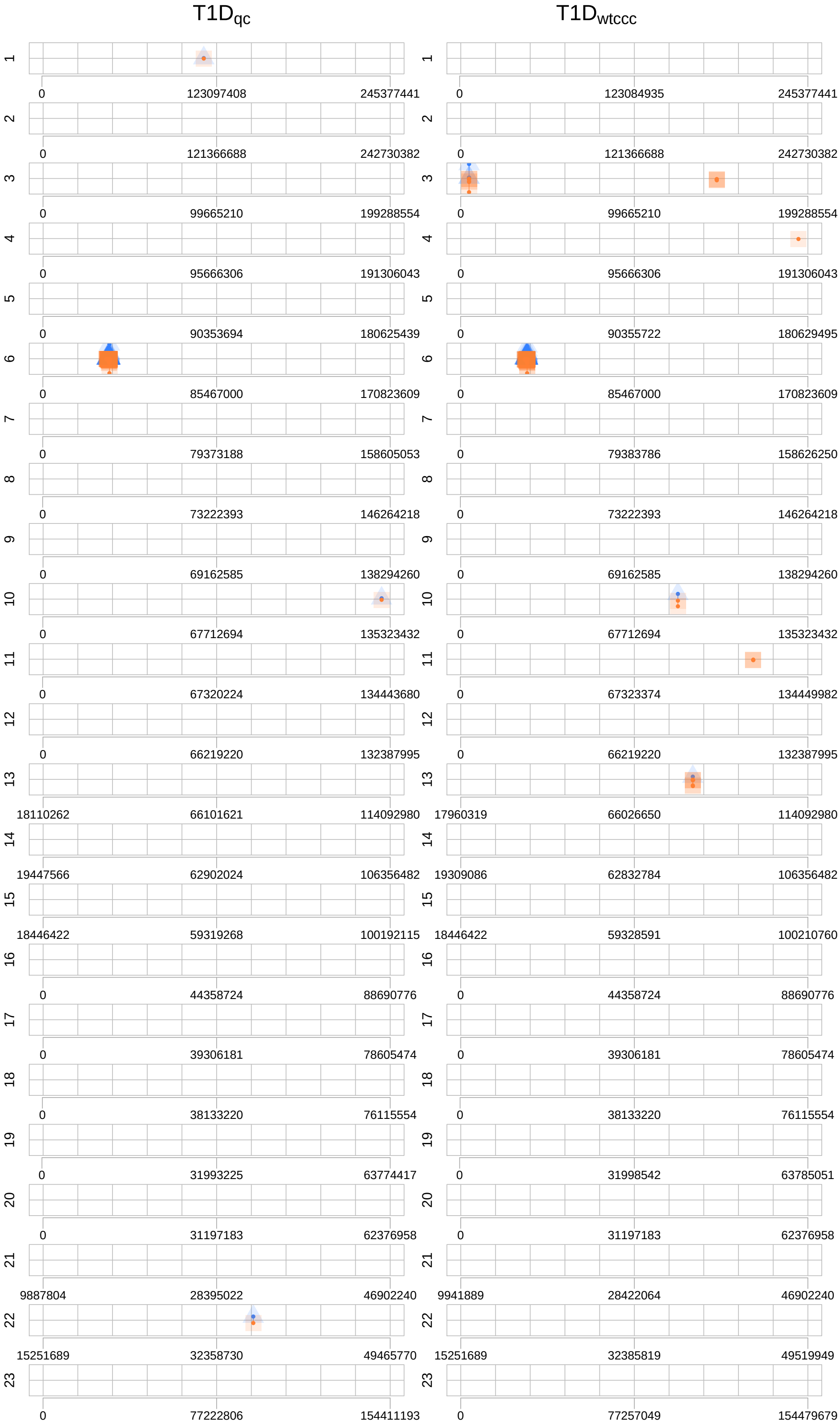


FIGURE B.6 $T1D$: the position of the first 100 variables according to the tree based methods. Blue and triangle correspond to RF variable importances. Orange and square correspond to TT variable importances. As points overlap, opacity increases, indicating several hits concentrated in a region

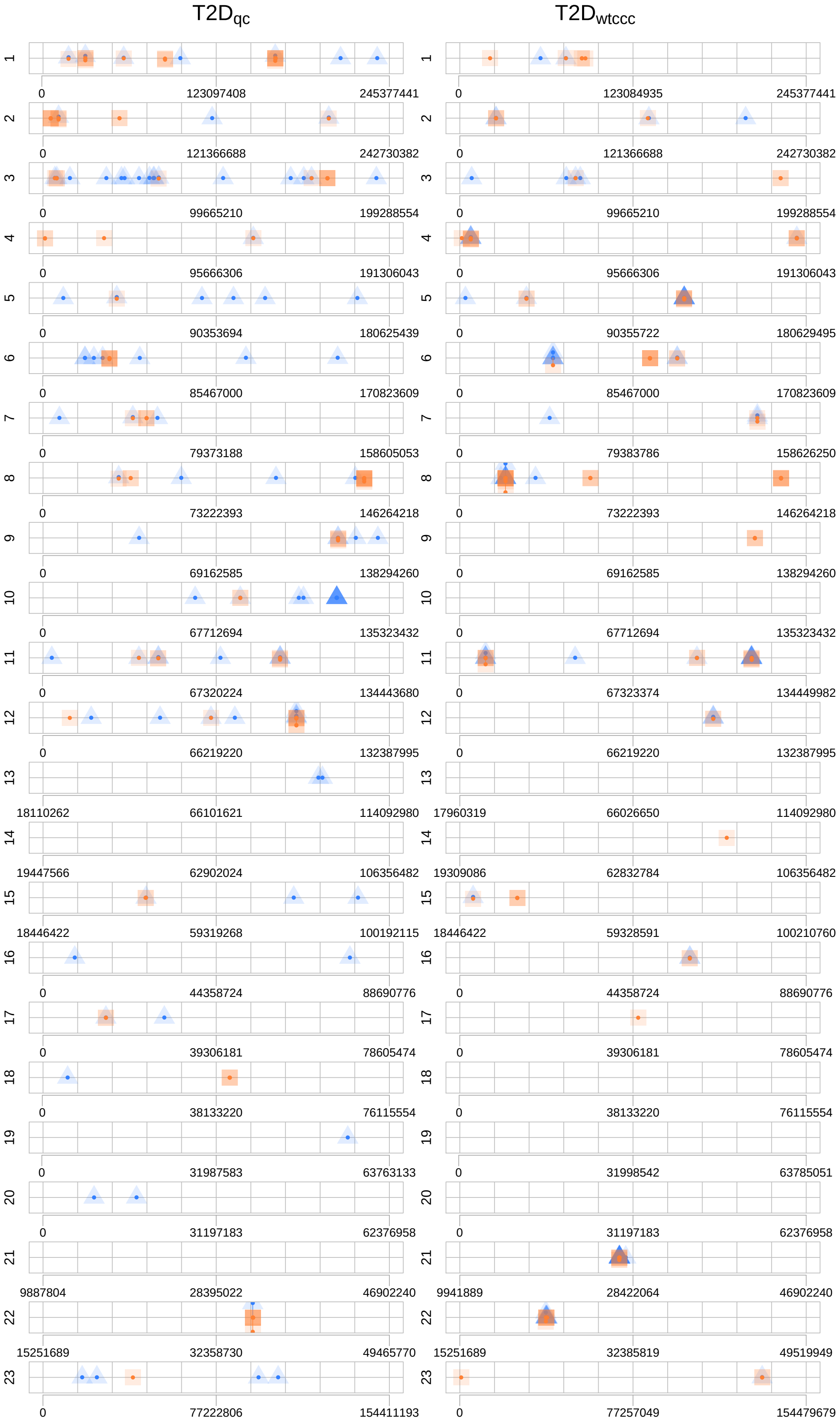


FIGURE B.7 *T2D*: the position of the first 100 variables according to the tree based methods. Blue and triangle correspond to RF variable importances. Orange and square correspond to TT variable importances. As points overlap, opacity increases, indicating several hits concentrated in a region

Appendix C

About the score measure

This section justifies our score measure choice. We applied the random forest (a 10-folds cross validation with $T = 100$ and no pruning) on CD_{wtccc} (resp. CD_{ibd}). Figure C.1 (resp. C.2) shows how the AUC evolve w.r.t. the value of K for the three different score measures: S_C^T, Q_C^T and I_C^T (cf. Chapter 4). We notice that S_C^T reaches the highest value. With I_C^T for identical values of K we reach lower AUCs than with S_C^T. Finally, the Q_C^T curve increases rapidly w.r.t. K but saturates around 0.8.

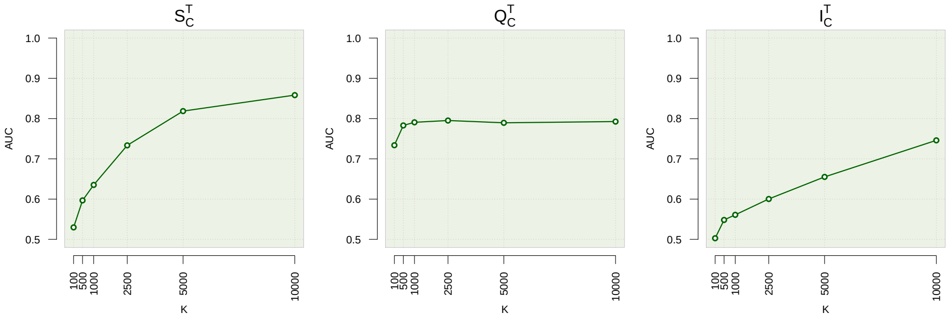


FIGURE C.1 CD_{wtccc}: the different AUC curve profiles we obtained with RF ($T = 100$) using three different score measures while increasing the K parameter.

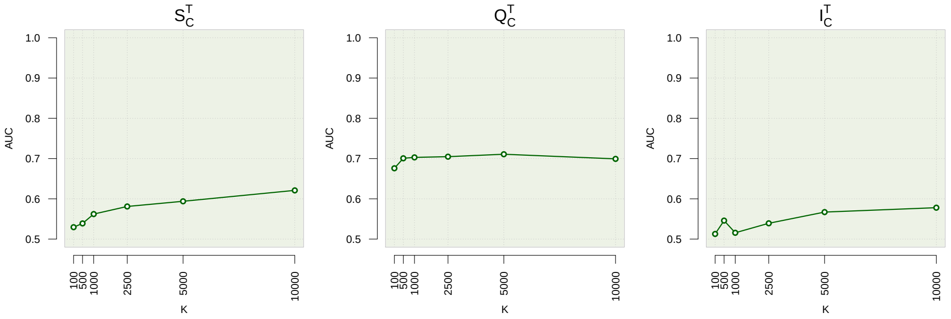


FIGURE C.2 CD_{ibd}: the different AUC curve profiles we obtained with RF ($T = 100$) using three different score measures while increasing the K parameter.

Also, Figure C.3 (resp. C.4) depicts the minor allele frequencies of the 1000 first markers according to each variable importance ranking obtained from forests using the three different score measures on CD_{wtccc} (resp. CD_{ibd}). Especially, with Q_C^T, the "end-cut" preference is a real problem. As we can see the most important variables are systematically the ones with a low MAF. It also explains why the AUC saturated with that score: the variable selection process is biased towards low MAF variables which forces the trees to exploit only a small subset of possible attributes and renders the method much less sensitive to an increase of K. On the other hand, we see for I_C^T that there is also a (less pronounced) bias but towards major allele frequencies variables this time. Finally, we see that with the S_C^T, the MAF values are more evenly distributed among top variables using the "adequately" normalised score measure.

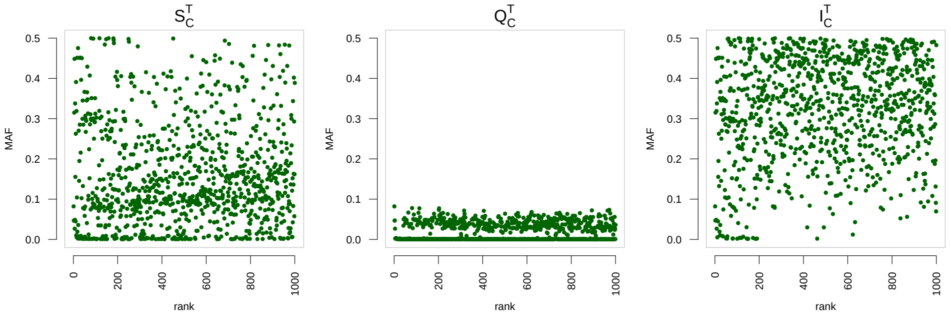


FIGURE C.3 CD_{wtccc}: Minor allele frequencies of the 1000 first variables according to three forests (RF, $K = 1000$, $T = 100$ and no pruning) using the three different score measures.

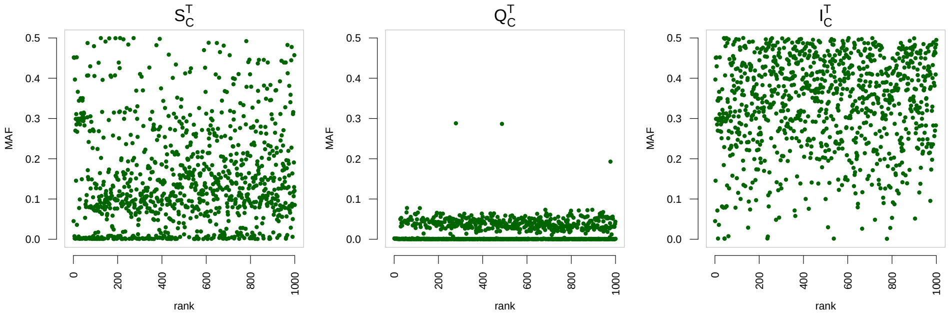


FIGURE C.4 CD_{ibd}: Minor allele frequencies of the 1000 first variables according to three forests (RF, $K = 1000$, $T = 100$ and no pruning) using the three different score measures.

Those two observations justify our choice of the S_C^T score measure as it leads to better predictions while not being biased towards any particular MAF value.

Bibliography

[A$^+$13] Ole A. Andreassen et al., *Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate*, PLoS Genet **9** (2013), no. 4, e1003455.

[B$^+$08] Jeffrey C Barrett et al., *Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease*, Nat Genet **40** (2008), no. 8, 955–62.

[B$^+$09] Manda Banerji et al., *Galaxy zoo: Reproducing galaxy morphologies via machine learning*, arXiv **astro-ph.CO** (2009), no. 1, 342–353.

[Bal06] David J Balding, *A tutorial on statistical methods for population association studies*, Nat Rev Genet **7** (2006), no. 10, 781–91.

[BBLBS12] Anne-Laure Boulesteix, Andreas Bender, Justo Lorenzo Bermejo, and Carolin Strobl, *Random forest gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations*, Brief Bioinform **13** (2012), no. 3, 292–304.

[BCB04] Viv Bewick, Liz Cheek, and Jonathan Ball, *Statistics review 8: Qualitative data - tests of association*, Critical Care **8** (2004), no. 1, 46–53.

[BDH$^+$03] Alexandre Bureau, Josee Dupuis, Brooke Hayward, Kathleen Falls, and Paul Van Eerdewegh, *Mapping complex traits using random forests*, BMC Genetics **4** (2003), no. Suppl 1, S64.

[BGHW08a] Vincent Botta, Pierre Geurts, Sarah Hansoul, and Louis Wehenkel, *Prediction of genetic risk of complex diseases by supervised learning*, Proc. of Benelearn 2008, the annual machine learning conference of Belgium and The Netherlands, May 2008, pp. 83–84.

[BGHW08b] ______, *Raw genotypes vs haplotype blocks for genome wide association studies by random forests*, Proc. of MLSB 2008, second workshop on Machine Learning in Systems Biology, 2008.

[BLDP$^+$10] Wesley H Brooks, Christelle Le Dantec, Jacques-Olivier Pers, Pierre Youinou, and Yves Renaudineau, *Epigenetics and autoimmunity.*, Journal of autoimmunity **34** (2010), no. 3, J207–19.

[BLGW13] Vincent Botta, Gilles Louppe, Pierre Geurts, and Louis Wehenkel, *T-trees: A novel tree-based approach for genome-wide association studies*, Submitted to Bioinformatics.

[Bre84] Leo Breiman, *Classification and regression trees*, Wadsworth Publishing, Jan 1984.

[Bre96] ______, *Bagging predictors*, Mach. Learn. **24** (1996), no. 2, 123–140.

[Bre01] ______, *Random forests*, Machine Learning **45** (2001), no. 1, 5–32.

[BSTB10] D Brinza, M Schultz, G Tesler, and V Bafna, *RAPID detection of gene-gene interactions in genome-wide association studies*, Bioinformatics **26** (2010), no. 22, 2856–2862.

[CG10] Elizabeth T Cirulli and David B Goldstein, *Uncovering the roles of rare variants in common disease through whole-genome sequencing.*, Nature reviews Genetics **11** (2010), no. 6, 415–425.

[CMFF10] Salvatore Catanese, Pasquale De Meo, Emilio Ferrara, and Giacomo Fiumara, *Analyzing the facebook friendship graph*, Proceedings of the 1st International Workshop on Mining the Future Internet (MIFI '10), 2010 (2010), 14–19.

[CUV+08] M L Calle, V Urrea, G Vellalta, N Malats, and K V Steen, *Improving strategies for detecting genetic patterns of disease susceptibility in association studies*, Statistics in medicine **27** (2008), no. 30, 6532–6546.

[CvdLJJea11] Marelli C, van de Leemput J, Johnson JO, and et al, *Sca15 due to large itpr1 deletions in a cohort of 333 white families with dominant ataxia*, Archives of Neurology **68** (2011), no. 5, 637–643.

[Die00] Thomas G Dietterich, *An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization*, Machine learning (2000), 139–157.

[E+08] Todd L. Edwards et al., *Generating linkage disequilibrium patterns in data simulations using genomeSIMLA*, Proceedings of the 6th European conference on Evolutionary computation, machine learning and data mining in bioinformatics (Berlin, Heidelberg), EvoBIO'08, Springer-Verlag, 2008, pp. 24–35.

[F+08] Manuel A R Ferreira et al., *Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder.*, Nature genetics **40** (2008), no. 9, 1056–1058.

[FCH+08] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, *LIBLINEAR: A Library for Large Linear Classification*, The Journal of Machine Learning Research **9** (2008), 1871–1874.

[Fri77] J.H. Friedman, *A recursive partitioning decision rule for nonparametric classification*, Computers, IEEE Transactions on **C-26** (1977), no. 4, 404–408.

[FS06] Ronen Feldman and James Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, December 2006.

[FZ10] Tao Feng and Xiaofeng Zhu, *Genome-wide searching of rare genetic variants in WTCCC data.*, Human Genetics **128** (2010), no. 3, 269–280.

[G+05a] Pierre Geurts et al., *Proteomic mass spectra classification using decision tree based ensemble methods*, Bioinformatics **21** (2005), no. 14, 3138–3145.

[G+05b] Elaine K Green et al., *Operation of the schizophrenia susceptibility gene, neuregulin 1, across traditional diagnostic boundaries to increase risk for bipolar disorder.*, Archives of general psychiatry **62** (2005), no. 6, 642–648.

[G+08a] Lyudmila Georgieva et al., *Support for neuregulin 1 as a susceptibility gene for bipolar disorder and schizophrenia.*, Biological psychiatry **64** (2008), no. 5, 419–427.

[G+08b] Maud M Gueders et al., *A novel formulation of inhaled doxycycline reduces allergen-induced inflammation, hyperresponsiveness and remodeling by matrix metalloproteinases and cytokines modulation in a mouse model of asthma*, Biochem Pharmacol **75** (2008), no. 2, 514–26.

[G+10] 1000 Genomes Project Consortium et al., *A map of human genome variation from population-scale sequencing.*, Nature **467** (2010), no. 7319, 1061–1073.

[Geu02] Pierre Geurts, *Contributions to decision tree induction: bias/variance tradeoff and time series classification*, Ph.D. thesis, University of Liège, Belgium, May 2002.

[GEW06] Pierre Geurts, Damien Ernst, and Louis Wehenkel, *Extremely randomized trees*, Machine Learning **36** (2006), no. 1, 3–42.

[GGS11] Torsten Gunther, Inka Gawenda, and Karl Schmid, *phenosim - a software to simulate phenotypes for testing in genome-wide association studies*, BMC Bioinformatics **12** (2011), no. 1, 265.

[GHCB10] Benjamin A Goldstein, Alan E Hubbard, Adele Cutler, and Lisa F Barcellos, *An application of random forests to a genome-wide association dataset: Methodological considerations & new findings*, BMC Genetics **11** (2010), no. 1, 49.

[GMW06] Pierre Geurts, Raphaël Marée, and Louis Wehenkel, *Segment and combine: a generic approach for supervised learning of invariant classifiers from topologically structured data*, Proceedings of the Machine Learning Conference of Belgium and The Netherlands (Benelearn), 2006, pp. 15–23.

[GRF11] Oscar González-Recio and Selma Forni, *Genome-wide prediction of discrete traits using bayesian regressions and machine learning*, Genetics Selection Evolution **43** (2011), 7.

[H$^+$06] A Geert Heidema et al., *The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases*, BMC Genetics **7** (2006), 23.

[HDM09] Bryan N Howie, Peter Donnelly, and Jonathan Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies*, PLoS Genet **5** (2009), no. 6, e1000529.

[HGV08] William G Hill, Michael E Goddard, and Peter M Visscher, *Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits*, PLoS Genetics **4** (2008), no. 2, e1000008.

[HJF69] EG Henrichon Jr and KS Fu, *A nonparametric partitioning procedure for pattern classification*, Computers (1969), 614–624.

[Ho98] Tin Kam Ho, *The random subspace method for constructing decision forests*, IEEE Trans. Pattern Anal. Mach. Intell. **20** (1998), no. 8, 832–844.

[HR76] Laurent Hyafil and Ronald L. Rivest, *Constructing optimal binary decision trees is np-complete*, Inf. Process. Lett. **5** (1976), no. 1, 15–17.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*, Springer, Jan 2009.

[Hun66] Earl B Hunt, *Experiments in induction*, 1st edition. ed., Academic Press, 1966.

[HV03] Susan M Hailpern and Paul F Visintainer, *Odds ratios and logistic regression: further examples of their use and interpretation*, interpretation **318** (2003), no. 134, 0.356.

[INTCI01] J P Ioannidis, E E Ntzani, T A Trikalinos, and D G Contopoulos-Ioannidis, *Replication validity of genetic association studies*, Nat Genet **29** (2001), no. 3, 306–9.

[J$^+$13] Luke Jostins et al., *Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease*, Nature **490** (2013), no. 7422, 119–124.

[JOV09] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert, *Group lasso with overlap and graph lasso*, Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009.), 2009.

[JTWF09] Rui Jiang, Wanwan Tang, Xuebing Wu, and Wenhui Fu, *A random forest approach to the detection of epistatic interactions in case-control studies*, BMC Bioinformatics **10** (2009), no. Suppl 1, S65.

[K$^+$03] Giulia C Kennedy et al., *Large-scale genotyping of complex DNA*, Nature Biotechnology **21** (2003), no. 10, 1233–1237.

[K$^+$11] Valeria V. Krzhizhanovskaya et al., *Flood early warning system: design, implementation and computational modules*, Procedia Computer Science **4** (2011), 106–115.

[KBR84] Igor Kononenko, Ivan Bratko, and Esidija Roskar, *Experiments in automatic learning of medical diagnostic rules*, Tech. report, Technical Report, Jozef Stefan Institute, Ljubljana, Yugoslavia, 1984.

[L$^+$10] Peng Lin et al., *A new statistic to evaluate imputation reliability*, PLoS ONE **5** (2010), no. 3, e9697.

[LCVK11] X Liu, R Cheng, M Verbitsky, and S Kisselev, *Genome-Wide association study identifies candidate genes for Parkinson's disease in an Ashkenazi Jewish population*, BMC Medical … **12** (2011), 104.

[LG12] Gilles Louppe and Pierre Geurts, *Ensembles on random patches*, ECML PKDD'12: Proceedings of the 2012 European conference on Machine Learning and Knowledge Discovery in Databases, Springer-Verlag, September 2012.

[LHSVE04] Kathryn Lunetta, L Brooke Hayward, Jonathan Segal, and Paul Van Eerdewegh, *Screening large-scale association study data: exploiting interactions using random forests*, BMC Genetics **5** (2004), no. 1, 32.

[LL08] Chun Li and Mingyao Li, *GWAsimulator*, Bioinformatics **24** (2008), no. 1, 140–142.

[lS02] Bernhard Sch lkopf and Alexander J Smola, *Learning With Kernels*, Support Vector Machines, Regularization, Optimization and Beyond, The MIT Press, 2002.

[M$^+$08] Mark I Mccarthy et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges*, Nat Rev Genet **9** (2008), no. 5, 356–369.

[M$^+$09] Teri A Manolio et al., *Finding the missing heritability of complex diseases*, Nature **461** (2009), no. 7265, 747–53.

[M$^+$10] K. Miclaus et al., *Variability in GWAS analysis: the impact of genotype calling algorithm inconsistencies.*, Pharmacogenomics J **10** (2010), no. 4, 324–35.

[Man10] Teri A Manolio, *Genome wide association studies and assessment of the risk of disease*, N Engl J Med **363** (2010), no. 2, 166–76.

[MAW10] Jason H Moore, Folkert W Asselbergs, and Scott M Williams, *Bioinformatics challenges for genome-wide association studies*, Bioinformatics **26** (2010), no. 4, 445–55.

[MGW07] Raphaël Marée, Pierre Geurts, and Louis Wehenkel, *Random subwindows and extremely randomized trees for image classification in cell biology*, BMC Cell Biology supplement on Workshop of Multiscale Biological Imaging, Data Mining and Informatics **8** (2007), no. S1, S2.

[MGW09] ______, *Content-based image retrieval by indexing random subwindows with randomized trees*, IPSJ Transactions on Computer Vision and Applications (open-access) **1** (2009), no. 1, 46–57.

[Min89] John Mingers, *An Empirical Comparison of Pruning Methods for Decision Tree Induction*, Machine Learning **4** (1989), no. 2, 227–243.

[MS63] JN Morgan and JA Sonquist, *Problems in the analysis of survey data, and a proposal*, Journal of the American Statistical (1963), 415–434.

[MVDGB08] Lukas Meier, Sara Van De Geer, and Peter Bühlmann, *The group lasso for logistic regression*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70** (2008), no. 1, 53–71.

[MWG13] Raphaël Marée, Louis Wehenkel, and Pierre Geurts, *Decision Forests for Computer Vision and Medical Image Analysis*, Springer London, London, 2013.

[MYC+09] Yan A Meng, Yi Yu, L Adrienne Cupples, Lindsay A Farrer, and Kathryn L Lunetta, *Performance of random forest when SNPs are in linkage disequilibrium*, BMC Bioinformatics **10** (2009), 78.

[NEW98] D M Nielsen, M G Ehm, and B S Weir, *Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus.*, American journal of human genetics **63** (1998), no. 5, 1531–1540.

[Nic11] Kristin K Nicodemus, *Letter to the Editor: On the stability and ranking of predictors from random forest variable importance measures*, Briefings in Bioinformatics **12** (2011), no. 4, 369–373.

[NM09] Kristin K Nicodemus and James D Malley, *Predictor correlation impacts machine learning algorithms: implications for genomic studies*, Bioinformatics **25** (2009), no. 15, 1884–90.

[NMSZ10] Kristin K Nicodemus, James D Malley, Carolin Strobl, and Andreas Ziegler, *The behaviour of random forest permutation-based variable importance measures under predictor correlation*, BMC Bioinformatics **11** (2010), no. 1, 110.

[P+00] P Pajukanta et al., *Two loci on chromosomes 2 and X for premature coronary heart disease identified in early- and late-settlement populations of Finland.*, American journal of human genetics **67** (2000), no. 6, 1481–1493.

[P+07] Shaun Purcell et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.*, American journal of human genetics **81** (2007), no. 3, 559–575.

[P+12a] Jittima Piriyapongsa et al., *iLOCi: a SNP interaction prioritization technique for detecting epistasis in genome-wide association studies*, BMC Genomics **13** (2012), no. Suppl 7, S2.

[P+12b] Zachary D. Pozun et al., *Optimizing transition states via kernel-based machine learning*, The Journal of Chemical Physics **136** (2012), no. 17, 174101.

[PRMN04] Tomaso Poggio, Ryan Rifkin, Sayan Mukherjee, and Partha Niyogi, *General conditions for predictivity in learning theory.*, Nature **428** (2004), no. 6981, 419–422.

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay, *Scikit-learn: Machine Learning in Python*, The Journal of Machine Learning Research **12** (2011), 2825–2830.

[Qui83] J. R. Quinlan, *Learning Efficient Classification Procedures and Their Application to Chess End-Games*, pp. 463–482, Morgan Kaufmann Publishers, Los Altos, CA, 1983.

[S$^+$06] Frank J Steemers et al., *Whole-genome genotyping with the single-base extension assay*, Nature Methods **3** (2006), no. 1, 31–33.

[S$^+$12] Fayaz Seifuddin et al., *Meta-analysis of genetic association studies on bipolar disorder*, American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics **159B** (2012), no. 5, 508–18.

[SBZH07] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn, *Bias in random forest variable importance measures: illustrations, sources and a solution*, BMC Bioinformatics **8** (2007), 25.

[SdS$^+$12] Márcio Gerhardt Soeiro-de Souza et al., *The impact of the CACNA1C risk allele on limbic structures and facial emotions recognition in bipolar disorder subjects and healthy controls.*, Journal of affective disorders **141** (2012), no. 1, 94–101.

[SFS$^+$11] Thomas P Slavin, Tao Feng, Audrey Schnell, Xiaofeng Zhu, and Robert C Elston, *Two-marker association tests yield new disease associations for coronary artery disease and hypertension.*, Human Genetics **130** (2011), no. 6, 725–733.

[Sha09] Barkur S Shastry, *SNPs: impact on gene function and phenotype.*, Methods in molecular biology (Clifton, N.J.) **578** (2009), 3–22.

[SMD11] Zhan Su, Jonathan Marchini, and Peter Donnelly, *HAPGEN2: simulation of multiple disease SNPs.*, Bioinformatics **27** (2011), no. 16, 2304–2305.

[Ste12] Kristel Van Steen, *Travelling the world of gene-gene interactions*, Briefings in Bioinformatics **13** (2012), no. 1, 1–19.

[T$^+$07] P A Thomson et al., *Association of Neuregulin 1 with schizophrenia and bipolar disorder in a second cohort from the Scottish population.*, Molecular psychiatry **12** (2007), no. 1, 94–104.

[T$^+$08] F Takeuchi et al., *Search for type 2 diabetes susceptibility genes on chromosomes 1q, 3q and 12q.*, Journal of human genetics **53** (2008), no. 4, 314–324.

[T$^+$11] Stephen Turner et al., *Quality control procedures for genome-wide association studies*, Curr Protoc Hum Genet **Chapter 1** (2011), Unit1.19.

[T$^+$12a] Clara Sze-Man Tang et al., *Genome-Wide Copy Number Analysis Uncovers a New HSCR Gene: NRG3*, PLoS Genet **8** (2012), no. 5, e1002687.

[T$^+$12b] Lanhua Tang et al., *Meta-analysis of association between PITX3 gene polymorphism and Parkinson's disease*, J Neurol Sci **317** (2012), no. 1-2, 80–6.

[T$^+$12c] W G Touw et al., *Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?*, Briefings in Bioinformatics **14** (2012), 315–326.

[The03] The International HapMap Consortium, *The international hapmap project.*, Nature **426** (2003), no. 6968, 789–796.

[TKTJ11] Y Takahashi, I Kou, A Takahashi, and TA Johnson, *A genome-wide association study identifies common variants near LBX1 associated with adolescent idiopathic scoliosis*, Nature genetics **43** (2011), no. 12, 1237–40.

[Tor01] Luís Torgo, *A Study on End-Cut Preference in Least Squares Regression Trees*, EPIA '01: Proceedings of the10th Portuguese Conference on Artificial Intelligence on Progress in Artificial Intelligence, Knowledge Extraction, Multi-agent Systems, Logic Programming and Constraint Solving, Springer-Verlag, December 2001.

[Vap98a] Vladimir Naumovich Vapnik, *Statistical learning theory*, Wiley-Interscience, September 1998.

[Vap98b] Vladimir Naumovich Vapnik, *Statistical learning theory*, Wiley-Interscience, Jan 1998.

[vB$^+$11] Bregje W M van Bon et al., *The phenotype of recurrent 10q22q23 deletions and duplications*, European Journal of Human Genetics **19** (2011), no. 4, 400–408.

[vdL$^+$07] Joyce van de Leemput et al., *Deletion at ITPR1 Underlies Ataxia in Mice and Spinocerebellar Ataxia 15 in Humans*, PLoS Genet **3** (2007), no. 6, e108.

[VSKZ09] Maren Vens, Arne Schillert, Inke R König, and Andreas Ziegler, *Look who is calling: a comparison of genotype calling algorithms*, BMC Proceedings **3 Suppl 7** (2009), S59.

[W$^+$09] Zhi Wei et al., *From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes*, PLoS Genet **5** (2009), no. 10, e1000678.

[W$^+$12] Stacey J Winham et al., *SNP interaction detection with random forests in high-dimensional genetic data*, BMC Bioinformatics **13** (2012), no. 1, 164.

[WA93] Louis Wehenkel and Vijay Akella, *A hybrid decision tree - neural network approach for power system dynamic security assessment*, Proceedings of the 4th International Symposium on Expert Systems Application to Power Systems, 1993, pp. 285–291.

[Weh96] Louis Wehenkel, *On uncertainty measures used for decision tree induction*, Proceedings of the International Congress on Information Processing and Management of Uncertainty in Knowledge based Systems, IPMU96 (Granada), 1996, pp. 413–418.

[Wel07] Wellcome Trust Case Control Consortium, *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls*, Nature **447** (2007), no. 7145, 661–78.

[YLLP11] Zeng-You Ye, De-Pei Li, Li Li, and Hui-Lin Pan, *Protein kinase CK2 increases glutamatergic input in the hypothalamus and sympathetic vasomotor tone in hypertension.*, The Journal of neuroscience : the official journal of the Society for Neuroscience **31** (2011), no. 22, 8271–8279.

[Z$^+$13] Xu Zhao et al., *Single-nucleotide polymorphisms inside microRNA target sites influence the susceptibility to type 2 diabetes.*, Journal of human genetics **58** (2013), no. 3, 135–141.

[Zha04] Tong Zhang, *Solving large scale linear prediction problems using stochastic gradient descent algorithms*, ICML '04: Proceedings of the twenty-first international conference on Machine learning, ACM, July 2004.

[ZHSL12] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander, *The mystery of missing heritability: Genetic interactions create phantom heritability.*, Proceedings of the National Academy of Sciences of the United States of America **109** (2012), no. 4, 1193–1198.

[ZL07] Yu Zhang and Jun S Liu, *Bayesian inference of epistatic interactions in case-control studies.*, Nat Genet **39** (2007), no. 9, 1167–1173.

[ZVSW10] Andreas Ziegler, Kristel Van Steen, and Stefan Wellek, *Investigating Hardy–Weinberg equilibrium in case–control or cohort studies or meta-analysis*, Breast Cancer Research and Treatment **128** (2010), no. 1, 197–201.