The genomic signature of trait-associated variants

Alida Sophie Dorothea Kindt

PhD
University of Edinburgh
2013

ABSTRACT

Genome-wide association studies have been used extensively to study hundreds of phenotypes and have determined thousands of associated SNPs whose underlying biology and causation is as yet largely unknown. Many previous studies attempted to clarify the causal biology by investigating overlaps of trait-associated variants with functional annotations, but lacked statistical rigor and examined incomplete subsets of available functional annotations. Additionally, it has been difficult to disentangle the relative contributions of different annotations that may show strong correlations with one another. In this thesis, we address these shortcomings and strengthen and extend the obtained results. Two methods, permutations and logistic regression, are applied in statistically rigorous analyses of genomic annotations and their observed enrichment or depletion of trait-associated SNPs. The genomic annotations range from genic regions and regulatory features to measures of conservation and aspects of chromatin structure. Logistic regressions in a number of trait-specific subsets identify genomic annotations influencing SNPs associated with both normal variation (e.g., eye or hair colour) and diseases, suggesting some generalities in the biological underpinnings of phenotypes. SNPs associated with phenotypes of the immune system are investigated and the results highlight the distinct aetiology for this subset. Despite the heterogeneity of the studied cancers, SNPs associated to different cancers are particularly enriched for conserved regions, unlike all other trait-subsets. Nonetheless, chromatin states are, perhaps surprisingly, among the most influential genomic annotations in all trait-subsets. Evolutionary conserved regions are rarely within the top genomic annotations despite their widespread use in prioritisation methods for follow-up studies. We identify a common set of enriched or depleted genomic annotations that significantly influence all traits, but also highlight trait-specific differences. These annotations may be used for the computational prioritisation of variants implicated in phenotypes of interest. The approaches developed for this thesis are further applied to studies of a specific human complex trait (height) and gene expression in atherosclerosis.

I declare that this thesis has been composed by myself and, except where otherwise stated, is entirely my own work.

……………………………………………

Alida Kindt
August 2013

Acknowledgements

I have many people to thank and much to be grateful for. Firstly, I would like to thank my supervisor, Chris Haley, for his guidance and support during my studies. His expertise and advice have been invaluable to this project and will be something I will always return to. Thank you also to my secondary supervisor, Colin Semple, for all his ideas and suggestions of how to advance this project further. I would also like to thank my third supervisor, Pau Navarro, for her advice and great attention to detail in all my analyses and manuscripts. I would also like to acknowledge funding from the MRC Capacity Building Studentship Scheme for this project and the conference trips.

I would like to acknowledge Teri Manolio and Lucia Hindorff for providing me with their original dataset for the replication of their method. I would also like to thank Tom Michoel, Hassan Foroughi Asl, and Johan Björkegren for access to the GWAS results analysing gene expression levels in the STAGE study, as well as their time spent on discussing the results.

There are many people I would like to thank for their help, company and discussions during my time at the HGU: I am very grateful to all the members of Chris Haley's group for discussions, Wenhua Wei and Mairead Bermingham for ongoing collaborations. I would like to particularly thank Athina Spiliopoulou for discussions on machine learning as well as lunches, coffees and laughs. Furthermore, I would like to thank all the members of the EvoGen group, especially Alison Meynert for her time and help with the project and the thesis. I would also like to thank my thesis committee, Martin Taylor and Paul McKeigue for their advice. I would like to acknowledge the MRC helpdesk for helping with issues concerning computing and maintaining the computer clusters used for all my analyses. Thank you to all of my colleagues from the BSA section for sharing the office, cakes, laughs and the coffee machine. Special thanks go to my fellow PhD students Emily Chambers and Dominik Glodzik for sharing the typical ups and downs of PhD life over the last four years. I would also like to thank Jenny

Huffman, for all lunches, coffees and cat-sitting as well as helpful discussions about data analysis. An extra special thank you to Lina Zgaga for all the laughs we had and introducing me to a whole new life.

I would not be where I am today without my family's continued support and belief in me. This next section will be in German to properly convey my thoughts to those to whom they apply. Lieber Papa, ich danke Dir sehr für alles, was Du mir ermöglicht hast und Deine Fürsorge. Mama, ich danke Dir, dass Du immer da warst für mich und, dass Du mich hast gehen lassen, obwohl es schwer gewesen sein muss. Danke für Euer Vertrauen, das Ihr mir gegeben habt und das mir wiederum grosses Selbstvertrauen gegeben hat! Dieses Buch ist handgrifflicher Beweis dafür. Laura, ich danke Dir sehr für die Zeit, die wir hier gemeinsam in Edinburgh hatten und Deine scheinbar grenzenlose Grosszügigkeit. Lucas, ich danke Dir für Deine Ermutigungen und Witze, die mich bei Laune hielten, wenn es mal nicht weiterging. Ich danke auch der Burg für die Möglichkeit hier zu sein, und bin sehr dankbar so viele gute und liebe Menschen in meiner Familie zu haben. Das gilt auch für Euch in Zürich und Dachau!

Last, but certainly not least I want to thank Vedran, the person without whom life would be infinitely more dreary and grey. Thank you for lighting up my life. I am looking forward to discovering what the future holds in store for us!

PUBLICATIONS

Kindt ASD, Navarro P, Semple CAM, Haley CS. *The genomic signature of trait-associated variants*. BMC Genomics. 2013. **14**:108
URL: http://www.biomedcentral.com/content/pdf/1471-2164-14-108.pdf

Chambers EV, Kindt AS, Semple CA. 2011. *Opening sequence: computational genomics in the era of high-throughput sequencing*. Genome Biol **12**: 310.
URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3334611/pdf/gb-2011-12-12-310.pdf

CONFERENCE PRESENTATIONS RELATED TO THIS THESIS

2012	Platform presentation: *The genomic signature of trait-associated SNPs,* IGMM Retreat, Edinburgh, UK
2012	Platform presentation: *The genomic signature of trait-associated SNPs,* Edinburgh Alliance of Complex Trait Genetics, Edinburgh, UK
2012	Poster: *The genomic signature of trait-associated SNPs*, Genome Informatics Conference, Cambridge, UK
2011	Poster: *Genome features underlying disease-associated SNPs*, Genome Informatics Conference, Cold Spring Harbor, USA
2010	Poster: *Genome features underlying trait-associated SNPs,* Genome Informatics Conference, Hinxton, UK

KEY ABBREVIATIONS

AIC – Akaike's Information Criterion

DNA – Deoxyribonucleic acid

CNV – Copy Number Variation

eQTL – Quantitative Trait Locus affecting gene expression

GWAS – Genome Wide Association Study

HCI – Higher Confidence Interval

Indels – Insertion/deletion event

LCI – Lower Confidence Interval

LD – Linkage Disequilibrium

MAF –Minor Allele Frequency

OR – Odds Ratio

ORF – Open Reading Frame

QTL – Quantitative Trait Locus

SNP – Single Nucleotide Polymorphism

TASPs – Trait-associated SNP Partners

TABs – Trait-associated Blocks

TSS – Transcription Start Site

TABLE OF CONTENTS

FIGURES

TABLES

1 INTRODUCTION

A policeman sees a drunken man searching for something under a streetlight and asks what the drunk has lost. He says he lost his keys and they both look under the streetlight together. After a few minutes the policeman asks if he is sure he lost them here, and the drunk replies, no, that he lost them in the park. The policeman asks why he is searching here, and the drunk replies, "This is where the light is." -- David H. Freedman (2010)

1.1 Phenotypes and heredity

Every living organism exhibits certain traits or phenotypes, which determine their ability to survive in any given environment. If the organism is successful in surviving and producing offspring, then these traits will get passed on to the next generation. This phenomenon is heredity and Gregor Johann Mendel, a 19th century Austrian monk, was the first to observe and describe heredity patterns of seven discreet traits, which existed in one of two forms, in garden pea plants (*Pisum sativum*) [1].

One of the traits he investigated was the colour of peas, which could be either yellow or green. In a crossing of two plants with yellow peas, which emerged from a cross of a plant with yellow peas and a plant with green peas, he observed a distinct 3:1 ratio of yellow to green peas. He was able to draw three conclusions based on the observed patterns. The first of these was that the determining factors of inheritance were “units”, which had been passed on to the descendants unchanged if the same trait was observed. The second was that each individual had two of these units, where one was obtained from each parent. The third conclusion was that, while the units were passed on, the trait did not have to be displayed. From these conclusions, two laws of inheritance were established much later [2], which are now called the Mendelian laws of inheritance. The Law of Segregation states that a parent only passes on one allele for any given trait, while the Law of Independent Assortment dictates that different pairs of alleles are passed on independently of each other to the descendants.

Mendel's work defined many of the terms used today. Mendel described two versions of the inherited units as alleles. Today, we know that these alleles describe different versions of genes or mutations. The terms heterozygous and homozygous define if an individual carries different or the same alleles, respectively. A recessive trait is only exhibited if the same alleles of a gene are inherited, while only one allele is necessary for a dominant trait as the state or version of the other is irrelevant in the development of the trait. Traits that follow the laws of inheritance are called Mendelian traits and while Mendel's conclusions were based on experiments in pea plants, the laws and conclusions are true for all diploid organisms, *i.e.,* organisms with two copies of genetic information in their cells.

1.2 Genomics and complex traits

Human DNA consists of ~3 billion nucleotides arranged into 23 chromosome pairs, which contain genes and regulatory elements controlling their expression. There are four nucleotides (Arginine, Thymine, Cytosine, and Guanine), which can be found in DNA, whose order encodes the information that determines the characteristics of the individual carrying the genes and regulatory elements. Expressed genes are transcribed into ribonucleic acid, or RNA, which is then translated into proteins. RNA contains three of the above-mentioned nucleotides and Uracil, which replaces Thymine.

The first working drafts of the human genome sequence were published in 2001 [3, 4] and announced finished in 2004 [5]. The sequencing was accomplished through a large international cooperation of many laboratories worldwide [3] and Celera [4, 6, 7], a privately owned company, and has been refined and improved several times. It is now in its 19th release [8] and its 37th release for the Genome Reference Consortium (GRC), with a 38th release planned for the end of summer of 2013 [9]. Each refinement improved the available reference genome but the working drafts or reference assemblies have never contained 100% of the genomic sequence due to sequencing and alignment problems particularly in repetitive regions [10].

Despite these problems, researchers were able to identify several types of variation between the sequences of individuals. These variations were either single base changes of the sequence, or insertions and deletions (Indels) of one or multiple bases, which can have effects on phenotypes. The effect can be caused by either changes within a gene's product (RNA or protein) or disruptions to a regulatory element controlling the expression of the gene. The mutations causing diseases such as Huntington's or sickle cell anaemia were identified within the protein coding part of a gene, either truncating or deforming the resulting protein and thus diminishing its function [11]. Mutations affecting the regulation of genes can also be detrimental and are therefore correspondingly selected against [12].

Mutations can occur during meiosis, mitosis or through exposure to mutational triggers such as UV radiation or exposure to certain chemicals. Base substitutions tend to affect only single base pairs, changing the identity of the local nucleotide. In rare events, it can affect more than one base pair during gene conversions. Deletions or insertions of nucleotides can vary in size from one to several nucleotides as a result of frame shift mutations or transposable elements. The mutational rate in humans across generations was recently estimated $\sim 1.20 \times 10^{-8}$ per base pair per generation [13-15], which is considered to be unexpectedly low [15]. Deleterious mutations, which either affect survival or reproduction of the carrier, are selected against, while those mutations offering an advantage can become fixed in the population. Selective pressures are determined by the environment and can be quite different between different populations. Mutations may therefore exhibit different allele frequencies in different populations or population groups, as selection pressures are different between populations depending on their environments.

This can cause stable genomic differences between groups of individuals within or across populations at certain points in the DNA called polymorphisms. A distinct and systematic difference in allele frequencies between populations is called population stratification. Admixture is the result of such two populations mixing, which results in new genetic lineages in a population. These are issues, which need to be taken into account in genetic studies, which are described later. Single nucleotide polymorphisms (SNPs) are DNA sequence variants where a single nucleotide has a different identity when compared to either other members of the same species or the paired chromosome. The different identities of a SNP are called alleles, where the majority of common SNPs have two alleles, although tri- or quadri-allelic SNPs can also be found. The minor allele of a SNP is defined as the allele with a frequency of less than 0.5 (minor allele frequency (MAF)) in a population or sample. To date (10th July 2013) 53,567,890 SNPs have been identified in the human genome, of which 38,072,522 have been validated according to the National Center for Biotechnology Information (NCBI) [16]. The genome contains highly specialized regions, which are associated with regulatory features or can contain a large number of genes. The distributions of genes, regulatory features and SNPs are far from homogenous with certain regions in the genome showing a higher SNP density than others [17, 18], with particular preference for areas outside of coding regions. An impediment to the reliable use of SNP data, or any DNA sequence data in further analyses, was the difficulty in genotyping certain areas of the genome. For instance, areas high in GC content are comparatively difficult to sequence or genotype accurately [19]. Furthermore, repetitive elements cause problems in the mapping and alignment of those sequences which hindered the identification of SNPs in these regions [20].

1.3 Genomic studies

Before nucleic acids were identified as the carriers of heritable traits, and the molecular structure of nucleic acids was identified [21, 22], researchers were continually observing phenotypes to further deduce conclusions about the molecule that determines our characteristics. They measured the frequency of alleles of genes by observing the frequency of traits in the population and calculated the expected and observed frequencies of certain traits to appear. They discovered a deviance in frequency of two alleles from the expected frequency as defined by the product of their individual frequencies, and termed it linkage disequilibrium (LD) [23]. LD is a measure of the genetic linkage between loci. It was hypothesized that as distance decreased alleles or genes would segregate together more often than expected by chance, so that a distance between the two loci could be estimated by calculating the number of recombinants in each cross. This greatly facilitated the identification of genetic components, as traits were analysed in relation to each other where the probability of segregation acted as a proxy for distance. With this knowledge, the first linkage or genetic maps were created in *Drosophila* ten years later [24], which aided discovering genes in that species. However, it was not until 1980 that linkage was used to create the first genetic maps in humans [25]. The genetic maps of human DNA sequence enabled the first linkage analyses, which identified disease-associated genomic regions solely through positional cloning or LD mapping [26]. Linkage analysis studies aim to associate genes to their locations in the genome and are performed by investigating genetic markers that are co-inherited with the analysed trait in related individuals. The rough location of the underlying genetics of the investigated trait is thereby identified and can be researched further to find the causal gene.

Linkage analyses or LD mapping were successful for Mendelian traits, *i.e.,* traits caused by single genes, which followed the Mendelian laws of inheritance. Among the first of the diseases, whose causal genes were successfully mapped was Huntington's disease. In 1993 [27] a mutation was identified in a single gene, hence named *Huntingtin*. More specifically, this mutation was the multiplication of a codon, within the coding region of that gene. The resulting protein functions differently depending on the number of repeats of a codon (CAG, coding for glutamate) where 36 or more codons result in Huntington's disease [28]. The number of codons within the gene is highly correlated with the age of onset and severity of the disease, where higher codon numbers cause more severe phenotypes. Parents with normal codon numbers can pass on a defect gene to their children due to a random multiplication event during meiosis. The clear separation between patient and healthy control and the large effect caused by the mutation in a single gene enabled the relatively easy identification of the causal gene.

However, not all genetic causes for traits were as detectable as the *Huntingtin* gene, as the majority of traits are not caused by a mutation in a single gene, which has a large effect on a phenotype. There are hundreds of examples of traits influencing behaviour or characteristics in humans that are not caused by a single gene with large effect [29]. Such complex traits can be caused by one or more genes acting either independently or in interactions with other genes, and/or the environment [29]. In complex traits, a mutation or genetic variant underlying certain the phenotype may also have only a very small effect on the trait making detection difficult. These genetic loci contributing towards complex traits are called quantitative trait loci (QTLs). The polygenic nature of complex traits means that they do not follow the Mendelian laws of inheritance, as the causal genes may not segregate together. This combined with the small effects of the causal variants means that it can be quite challenging to identify all genetic factors of complex traits [29-32], especially when limited by low-resolution methods such as linkage analyses.

The search for trait-associated genetic variants advanced substantially in the last decade with the implementation of genome-wide association studies (GWAS). These studies are designed to analyse the entire genome for regions associated to the trait under investigation [31] using an essentially blind approach and do not require a previously identified area of interest. GWAS test alleles of SNPs for associations with diseases or other measurable phenotypes returning a *P*-value of association, which reflects the strength of the association. The first GWAS were performed in 2005 and 2006 and identified SNPs with significantly different allele frequencies in healthy controls and patients with age-related macular degeneration [33, 34]. The era of GWAS truly began in 2007 with the Wellcome Trust Case Control Consortium (WTCCC), which analysed seven common diseases in 14,000 patients and 3,000 controls [35] and discovered several significantly associated SNPs. A SNP is said to be significantly associated with the trait, if its P-value of association was less than the commonly accepted genome-wide threshold of significance ($P \leq 5 \times 10^{-08}$), which is based on testing 1 million SNPs [36]. This P-value takes the number of performed tests into account to eliminate spurious associations caused by multiple testing. The success of a study, as measured by the number of identified significantly trait-associated variants, varied for different traits. This demonstrated two major difficulties encountered by GWAS performed since then. First, the cost of obtaining sufficiently large numbers of samples to detect variants of small effect can be prohibitive [31]. Second, certain traits may also be poorly defined, which makes the identification of the underlying genetics difficult. Bipolar affective disorder, for example, can be difficult to diagnose as it can be mistaken for other psychoses or unipolar depression with re-current episodes [37]. Other problems of GWAS include the lack of power to identify associations with rare SNPs and SNPs with small effect sizes. GWAS are also highly sensitive to admixture and population stratification, which could result in false positive associations as alleles segregated at different frequencies in different populations. The heritability estimate of any given trait may also influence the success of a GWAS, as high heritability estimates indicate a large genetic component. The assumption is that genome-wide association studies would be more successful at detecting associated genomic regions for these highly heritability traits than for traits which have a higher environmental component.

Despite all these issues, as of 25th June 2013 1,640 GWAS had been published, reporting a total of 10,876 trait-associated SNPs [38]. SNPs tested for association to a trait are said to be significantly associated with a trait, if they are found to co-occur with the trait more often than expected by chance. This is usually at a significance level of less than 0.05. However, since many hundreds of thousands of SNPs are analysed in each study, the significance threshold must be corrected for multiple testing to take into account spurious associations. One method often used to correct for multiple testing is the Bonferroni correction. An association P-value of $P \leq 0.05/n$, where n is the number of tests performed, *i.e.,* the number of SNPs on a genotyping array, is generally accepted as significant evidence that the SNP is associated to the disease [39].

According to some, GWAS produced relatively little understanding of the underlying biology [40, 41], and has even been characterized as a waste of money by some of the supporters for effectively the same reasons [31]. Regardless of the validity of these criticisms, the fact is that GWAS have greatly advanced our understanding of the genetics of complex traits [31], increasing our knowledge by a factor only previously matched by epidemiological studies [42].

When GWAS were first carried out it was hoped that trait-associated regions contained functional elements, which could explain some, if not all, of a trait's observed heritability [31]. Heritability is a calculated estimate of the genetic proportion of traits and is the ratio of the variance of a trait across generations and the total phenotypic variance in a population, which is the interaction between genetics and environment [43]. The identified variants have explained little of the estimated total heritability of the analysed traits, a phenomenon dubbed "missing heritability" [36, 38, 44]. Studies have since suggested that the heritability is not missing but that researchers either do not know how to look for it or that it was estimated incorrectly [30, 45, 46]. Several methods have been published which improve the heritability estimates by employing methods that analyse more than just the associated variants [30, 36, 46-48]. It is therefore possible that the "missing heritability" is not missing at all, but that it is calculated wrongly after the results are obtained. Furthermore, it is possible that the causal variant is not only not included on the genotyping array used for the study, but that the represented variants are not in full LD with the causal variant. Other reasons for the apparently missing heritability could be, as of yet unaccounted for, epistatic effects, epigenetic factors or use of models that only investigate significant SNPs rather than all effect sizes.

The genotypes of the SNPs used for GWAS are obtained using genotyping array technologies, which allow the simultaneous analysis of a large number of SNPs. Most GWAS use commercially available and standard genotyping arrays, which measure between 300,000 and 1.8 million genetic markers at the same time, although earlier arrays were less dense. The entire genome can be examined using small numbers of SNPs in comparison to the number present within the entire genome because of LD. The underlying assumption of GWAS is that the analysed SNPs are either the causal mutation or in LD with the causal mutation and can therefore 'tag' the causal mutation. Those tag SNPs, which are in LD with a large number of SNPs and are therefore the most informative SNPs, are included on the arrays to produce maximum coverage of the genome with a minimum number of genotyped variants [49]. The different genotype array producers use different criteria to choose the SNPs that were incorporated into arrays. For example, the Illumina HumanHap300 array included an intentional bias towards non-synonymous variants [50], due to their high impact on protein function. After performing quality control on the genotyped SNPs, badly genotyped or missing markers are often imputed to boost the numbers of analysed markers [51].

Imputation is a method, which is used to infer the genotypes of untyped SNPs based on the identity of genotyped SNPs and a reference population [52]. This adds statistical power to GWAS by adding more SNPs to the analysis, which can then be tested for an association with analysed traits in the same way as genotyped SNPs. Several algorithms, such as IMPUTE2 [53], MACH [54] or BEAGLE [55] to name but a few, are available for imputations. Meta-analyses routinely use imputation to combine different studies, which may have used different genotyping arrays, to infer the genotypes of untyped SNPs in the used studies. The imputed SNPs can be tested for association in the same way as the genotyped SNPs are. While imputations greatly boost the number of variants available for analysis, they can take a very long time to perform. It is additionally possible that imputations introduce errors into the analysis if the wrong reference genome is used. However, if the reference genome is matched appropriately to the population under investigation, the results can be quite reliable. As most studies now use imputations to boost the numbers of SNPs available for analysis [38], any study investigating trait-associated variants will have to consider SNPs resulting from imputations.

Several study designs are available for GWAS. One common design is the cohort study, where one group of people is analysed for a common quantitative trait, to identify a common genetic factor. A study using this design led to the discovery of the involvement of a urate transporter gene in gout [56]. Another popular design is the patient ('case') and control group study, where SNPs are investigated for a difference in allele frequencies in the two groups. In these types of studies, it is important to match the sampled populations properly to take population stratification and admixture into account. If these are not considered properly, spurious associations will result from the analysis, which are only indicative of a recent mixing of the populations than real trait-associations. The above mentioned WTCCC study in 2007 [35] was a case-control study which analysed seven different traits. Usually a discrete trait is analysed (*i.e.,* a trait which simply is present or not, for instance the presence or lack of the defining symptoms of the disease under investigation), although continuous traits such as e.g., height can also be examined [57]. The majority of the trait-associated variants identified so far are common alleles, which have allele frequencies greater than 0.05, with modest effect sizes on the trait [58]. This directly reflects the biases in the variants that were chosen for the genotyping arrays towards those with common alleles. These were selected, as the minor alleles will be present in a larger number of people, increasing the power to detect a given size of genetic effect for a fixed sample size. Furthermore, the cost of studies using the number of people needed to analyse small effects has so far been prohibitively large. However, this will likely change as the cost to genotype and sequence the genome is ever-decreasing.

Despite the shortcomings of GWAS, when results are produced it can have profound impact on our understanding of biology. For instance, when the underlying biology was verified through follow-up studies, the results were sometimes surprising. For example, the genes associated with multiple sclerosis were all found to have an autoimmune role rather than a neurodegenerative role as previously hypothesized [59]. The new biology introduced by GWAS has therefore shed light onto the aetiology of the investigated traits. The identification of new candidate treatment options as a consequence of GWAS, perhaps in the form of novel drug targets, is the best-case scenario and but so far has happened only rarely [31]. While we know only 10-20% of genetic variability contributing to certain diseases, we know 10-20% more of the underlying genetics than we did five years ago [31]. GWAS have therefore contributed substantially to our understanding of complex traits and diseases [31].

GWAS identifies SNPs that are only associated to traits, which does not imply that the associations really are the mutations causing the trait. A distinction therefore needs to be made between trait-associated variants and trait causing variants. A trait-associated variant will highlight areas of interest for follow-up studies, while causal mutations will clarify the aetiology of traits. The causal mutations are known for only a fraction of the 10,876 reported trait-associations. This is at least partly due to the majority of the trait-associated variants lying outside coding regions [50] and only a small fraction of the trait-associations are near or within genes [60]. This was surprising when first discovered and is contributing to the bottleneck in elucidating the molecular processes and pathways underlying these associations [36, 38, 50, 61] and hence in gaining new biological knowledge. Experiments identifying the causal underlying biology for confirmed associations are expensive and time-consuming. There has therefore been much interest in computational prioritisation of candidate variants, both to accelerate the search for causal variants, and to provide insights into the biology underlying disease states [62-

65]. Although confirmed trait-associated SNPs will most often not be the causal variants, the surrounding genomic regions in LD with associated SNPs are expected to contain causal variants with biological function. This triggers the question: what do the trait-associations coincide with, if not with coding regions?

1.4 Function of genome

Back in 1972 it was hypothesized that 6% of the genome was within coding regions and what remained was defined as "junk DNA" [66], *i.e.,* DNA that had no function as it was not under selection. Junk DNA has since been a highly disputed concept. However, the fact remains that a very small proportion of DNA is coding. The latest estimate of the number of coding genes in the human genome stood at 20,806 according to ENSEMBL (date of access: 13 May 2013), which corresponds to about 1.5% of the genome, a quarter of what was previously hypothesized. In the meantime, it was realized that the remaining 98.5% harbours important functional elements such as non-coding enhancers, silencers and promoters. Yet, not all the non-coding DNA has function associated with it. So, the right question is: how much of the genome is functional?

The recent ENCODE consortium estimated that ~80% of the genome is 'functional' in the sense that it possesses some biochemical activity [67], thereby effectively eliminating the term junk DNA. This percentage was highly disputed the moment it was announced [68, 69] and continues to be challenged [70]. Specifically, Graur *et al.* [71] declared that the ENCODE consortium did not estimate function correctly, blatantly refused to look at the evidence in front of them and grabbed a good sounding number out of the hat. The ENCODE authors themselves gave different functional genomic estimates ranging between 20%-80%, depending on which author was asked [71]. The death of the term "junk DNA" was therefore contestable. While the debate on how much of the genome contains functional elements is not yet resolved, it is indisputable that a wide range of functional elements has been identified over the years.

There are many regulatory and functional elements in the genome. Here, we separate them into three main classes. The first class includes some of the many identified genic annotations in the genome, such as gained stop codons 5' and 3' UTRs and pseudo genes as well as genes but also regulatory regions like eQTLs [72]. Previous studies had shown that distance to transcriptional start site (TSS) was very important for predicting eQTLs [72], so in chapter 5 onwards, we also included this quantitative variable in our logistic regression modelling in this class. The second class consists of other candidates for functional elements such as regions, which are highly conserved across multiple species or transcription factor binding sites [73, 74], enhancers identified through conserved sequences [75] and insulators. They are suspected to be important during development and contribute to the accurate function of the cell and have also been used as a proxy for functional elements [64, 76].

Yet a third class of potential functional elements is derived from a range of dynamic chromatin features that are associated with biological functions like promoters, enhancers, silencers or heterochromatin states. The epigenome is the set of heritable features that can alter gene expression or the cellular phenotype independently of the DNA sequence itself [77]. Epigenomic features include DNA methylations, histone modifications and the binding of transcription factors [77]. DNA methylations most commonly occur on a cytosine nucleotide when a guanine residue follows it on the same DNA strand. DNA methylations are necessary for the correct function of a cell, as aberrant methylation is highly associated with cancer [78, 79]. Certain amino acid residues within histone proteins, around which DNA molecules are wrapped for safe storage in the cell, are the targets of a variety of different biochemical modifications. The identity and location of the modification is strongly associated with their function. Presence of acetylations, for example, is almost always associated with transcriptionally active regions in the genome, whereas absence of these modifications may indicate inactive regions [80]. Methylations can be associated with repressed or active regions depending on their location on the histone protein [80]. Epigenomic mechanisms are not only crucial for cell differentiation during development but also seem to respond to environmental stimuli, such as diet [81]. A landmark paper investigated epigenomic modifications and their distribution across nine cell lines and identified a reproducible pattern of histone modifications for 15 genomic features [82]. These features ranged from promoters and enhancers to insulators and repressed regions, and showed varying degrees in strength of function as well as significant odds ratios for enrichment of specific trait associated variants. However, they did not investigate all available trait-associated variants.

1.5 Future outlook

Trait-associated variants are assumed to highlight the functional variants, which cause the investigated traits. The number of these associated variants is set to increase not only through more powerful GWA studies but also through next generation sequencing. In order to gain any insights into what drives these associations we will have to investigate where these trait-associated variants lie and what causes the associations. However, gathering the data is only the first part and in order to validate the associations, follow-up studies need to be performed. The cost of these in time and money is restrictive limiting the number of results that can be analysed. Given that trait-association data is going to continue to accumulate, ways of prioritizing the associated variants are going to become ever more important. We have therefore developed the project for this thesis, the scope of which we outline next.

1.6 Scope of this thesis

This thesis deals with the number and nature of the functional elements found to be overlapping with trait-associated variants, which were the results of GWA studies investigating a large number of traits. These SNPs may not be the causal mutations themselves, but it is the assumption that they are nonetheless in association with the underlying causal mutation, which might not be an identified SNP [36]. The introduction of this thesis sets out to define the most important concepts used in this study. Chapter 2 describes the SNP sets and genomic feature datasets we use throughout the thesis. In Chapters 3, 4 and 5 we explore three methods for the identification of sets of genomic annotations that were either significantly enriched or depleted for trait-associated variants in humans. First we investigate a sampling based method employed by Hindorff *et al.* [50] in Chapter 3. We then analyse a permutation-based method and compare the permutation results to the sampling results in Chapter 4. The permutations were investigated as an alternative to the sampling method, in the hope it would allow a more rapid, comprehensive, and statistically rigorous analysis. Chapter 5 details the third method we applied. This was logistic regression, a method that allowed estimates of the relative contributions of all genomic annotations to trait-association status, thereby eliminating redundant information. Here, we also included an additional annotation, distance to TSS, as suggested by a reviewer of our paper (following [72]). In Chapter 6 we apply two methods, permutations and regression, to two differently obtained sets of SNPs to see if the methods are adaptable. The first of these datasets was the result of the Stockholm Atherosclerosis Gene Expression (STAGE) study aimed at identifying eQTLs and contained 26,546 SNPs [83]. The second dataset was obtained from the Genetic Investigation of Anthropometric Traits (GIANT) consortium and contained *P*-values of association for ~2.5 million SNPs investigated for associations with height [57].

2 MATERIALS AND METHODS

2.1 SNP datasets

2.1.1 Hindorff SNPs

This data consisted of 465 unique trait-associated SNPs, with associations significant at the Bonferroni corrected genome-wide significance threshold of 5 × 10^{-8}. The significant SNPs originated from 151 of the 237 published studies until December 2008 [50]. This dataset of Hindorff SNPs was used only in the original paper in 2009. We attempted to reconstruct this set of trait-associated SNPs. All reported studies used at least one of the several commercially available genotyping arrays by different companies, and most studies ended up with ~2.5 million imputed SNPs. Imputation of unknown SNPs using a reference population is a cost and time efficient method for increasing the number of SNPs available for analyses [84].

2.1.2 Reconstructed Hindorff SNPs

The original GWAS catalogue, detailing all GWAS performed until 31st December 2008, was obtained through personal communication with the authors [50]. This GWAS catalogue reported 1,104 SNPs identified as trait-associated at various levels of significance in 237 studies for 165 different traits. Of these, 476 SNPs were unique, trait-associated SNPs significant at 5 × 10^{-8} for 95 traits identified in 151 studies. The final analysed dataset included 468 SNPs, as SNPs from the Y-chromosome or non-assigned chromosomes were removed from the data. It was impossible to identify the exact set of 465 SNPs used for the analysis by Hindorff *et al.*, as following their method to identify the significantly trait-associated SNPs resulted in a number of trait-associated SNPs that was far lower than they originally reported. It is possible that further steps were either not published or that their methodology was not clear. However, several approaches at replicating the data were attempted. The dataset analysed here was the closest approximation.

2.1.3 Significant SNPs (2011)

The National Human Genome Research Institute (NHGRI) GWAS catalogue [85] reports genome-wide association studies if they meet their inclusion criteria. The first one is that any included study must analyse at least 100,000 SNPs before the quality control was applied. Secondly, the catalogue only reports SNPs with *P*-values of association of < 1.00×10^{-5} in the total analysed population, which includes any initial and replication studies. There is no exclusion criterion based on minimum sample size, but the curators did exclude any studies that focussed only on SNPs in candidate regions. Trait-associated SNPs significant at the genome-wide significance threshold ($P \leq 5 \times 10^{-8}$) were extracted from the NHGRI GWAS catalogue [85], downloaded on 25 August 2011. The catalogue reported 5,800 associations from 764 studies in total. After the removal of SNPs on the Y-chromosome or non-assigned chromosomes 1,974 were significantly associated in 576 studies and 1,909 of these significantly trait-associated SNPs were analysed. The remainder either had rare allele frequencies in the study populations and were not present in the HapMap CEU II reference data, or were lost due to updated rs numbers when the positions were updated to build 37.

2.1.4 Suggestive SNPs (2011)

Suggestive SNPs (2011) consisted of the suggestively trait-associated SNPs derived from the NHGRI GWAS catalogue, which was accessed 25 August 2011 [86]. The significantly associated SNPs were also extracted from this set, as detailed in the paragraph above. The suggestively trait-associated SNP set was defined as SNPs with association *P*-values between 5×10^{-8} and 5×10^{-5}. SNPs that were located on either the Y-chromosome or unassigned to chromosomes were removed from all analyses. SNPs in the suggestively associated SNP set found to be in LD ($r^2 > 0.9$) with significant SNPs were removed from the dataset, resulting in 2,410 unique rs numbers from 412 studies present in the data. The remainder either had rare allele frequencies in the study populations, and were not present in the HapMap CEU II reference data, or were lost due to updated rs numbers when the positions were updated to build 37.

2.1.5 Significant SNPs (2013)

The NHGRI catalogue was accessed again on 18 January 2013 and the most recent version of the GWAS catalogue was downloaded, which incorporated all trait-associated SNPs known to that date. This dataset contained 3,421 unique SNPs that were associated with 492 individual traits identified in 929 studies. The total number of SNPs present in the background list of SNPs was 3,283 SNPs. As before, the remainder either had rare allele frequencies in the study populations, or were not present in the HapMap CEU II reference data, or were lost due to updated rs numbers when the positions were updated to build 37.

2.1.6 Significant SNPs (Difference)

This dataset consisted of the difference of Significant SNPs (2013) and Significant SNPs (2011). This set was analysed to see the impact of only the new variants when compared to the older variants. It comprised 1,477 significantly trait-associated SNPs that were identified in the period between 25 August 2011 and 18 January 2013. Three Japanese studies [87-89] that were part of Significant SNPs (2011) had been removed from the NHGRI catalogue by the time Significant SNPs (2013) was downloaded. No official reason has been found for the removal of the SNPs. However, this increased the total amount of analysed SNPs by 153 in the Significant SNPs (Difference). This set was analysed as the SNPs were expected to have a slightly different distribution to all SNPs from 2013, as the 2011 SNPs contributed to the overall set. The difference of the sets was expected to have a slightly different distribution due to the design of the newer GWAS, which tend to analyse comparatively larger populations.

2.1.7 Trait-subsets

The traits (phenotypes) associated with the Significant SNPs (2013) were divided into four subsets: Cancer traits, immune-related traits, general disease traits and normal variation traits. This was previously not possible, as the number of SNPs within the subsets was prohibitive for a reliable result. The SNPs that overlapped between the disease category and the normal variation category were classified as disease SNPs. Please refer to Section 9.1 for an overall view of all traits in each subset. We downloaded data from the Genetic Association Database (GaD, [90]) which has a detailed breakdown of traits into categories. The traits were separated into subsets using the information from the GaD. For those traits, which were not listed in GaD, we searched across the publicly available data online to sort them into their respective subsets according to the results of that investigation.

2.1.8 STAGE eQTLs

One dataset used for analysis was an eQTL dataset generated by the Stockholm Atherosclerosis Gene Expression (STAGE) study [83]. This study investigated gene expression levels in seven tissue types of 147 coronary artery disease patients eligible for coronary artery bypass grafting and/or carotid atherectomy. Tissue biopsies were extracted from atherosclerotic arterial wall (*n* = 68), internal mammary artery (*n* = 79), liver (*n* = 77), whole blood (*n* = 102) and subcutaneous (*n* = 63) and visceral fat (*n* = 88). The gene expression levels, as measured by RNA levels in the biopsies, were treated as traits in a correlation study with the SNP genotypes of the 109 patients, which had sufficient levels of DNA. Multiple testing was corrected for using false-discovery rate in each individual tissue. These SNPs were located within 1 Mb of 6,450 genes whose gene expression was used as the investigated trait. The total number of SNPs, identified as significantly associated with gene expression levels were referred to as eSNPs, was 29,530. Collaborators at the Karolinski Institute, Sweden, performed all the gene expression analyses. Here, we analysed 26,546 SNPs of 29,530 SNPs. The remainder were lost due to updated rs numbers when the positions were updated to build 37. The data could be separated into different subsets according to the tissue whose gene expression correlated with the genotyped SNPs.

2.1.9 GIANT SNPs

The Genetic Investigation of Anthropometric Traits (GIANT) consortium is an international genome-wide association meta-analysis consortium that focused on the identification of loci affecting measures for human body size and shape. This consortium has made datasets available to the public, including the meta-analyses for three traits: Height, BMI and Waist/Hip ratio adjusted for BMI. The height dataset was downloaded and analysed to identify genomic signature patterns of associated SNPs, and consisted of 2,469,635 SNPs with a range of association *P*-values [57].

2.1.10 Defining linkage disequilibrium partners

Linkage disequilibrium (LD) underlies the design and success of GWAS. The design of GWAS was based on genotyping arrays, which could capture a substantial proportion of genomic variation [91]. LD between two alleles is defined as the deviation (denoted *D*) of the observed frequency of two combined loci from the expected, where the expected frequency of two loci is the product of their allele frequencies [43]. While *D* is easy to calculate, it is highly dependent on allele frequencies, so usually *D'* is calculated which takes the allele frequencies into account. An alternate measure to *D'* is r^2, which is defined as the square of the correlation coefficient between pairs of loci. This measure also takes allele frequencies into account. It was the optimal choice for our work, as it is a commonly used measure and Hindorff *et al.* also used r^2 in their analysis [50]. SNPs in LD were referred to as LD partners and were important for the analysis of the underlying genomic structure of phenotypes. LD partners were characterized as SNPs from the HapMap CEU II data that were in LD above the chosen cut-off threshold ($r^2 > 0.9$) with a trait-associated or sampled SNP [92, 93]. Since only a fraction of the known SNPs were included on genotyping arrays, it was unlikely that causal mutations were genotyped. However, it is assumed that they were in LD with associated SNPs [50]. The LD threshold of $r^2 > 0.9$, a highly stringent threshold [50], was chosen in compliance with previous literature [50]. This cut-off point was also chosen for all our analyses, unless otherwise stated. The HapMap CEU II data on LD (release #24, phase I and II, http://hapmap.ncbi.nlm.nih.gov/downloads/ld_data/2009-02_rel24/) between SNPs was used to define the LD partners of all analysed SNPs (trait-associated and non-associated). HapMap CEU II data contained information on LD calculated for pairs of SNPs up to 250 Kb apart from each other [92]. This resulted in theoretical LD blocks of up to 500 Kb long for any one SNP. An additional cut-off point of $r^2 > 0.7$ was analysed to investigate reducing the LD threshold, capturing more LD SNP partners and therefore potentially more causal variants but also noise.

2.1.11 Scoring LD blocks and definition of odds ratios

The applied scoring system for the calculation of depletion/enrichment odds ratios was binary. An LD block of trait-associated SNPs and its partners was defined as overlapping with a particular genomic feature if at least one of the SNP variants in that block coincided with the genomic feature. Multiple hits within an LD block were not counted. Sample and permutation SNPs were treated the same as the trait-associated SNPs to enable a solid comparison between the expected and the observed data.

Odds ratios were calculated to enable comparisons with previous studies [50], where an odds ratio was defined as shown below. Here, the observed data are the number of overlaps of trait-associated SNPs and the expected data are the mean number overlaps of the background data.

$$\frac{(\text{Overlaps Observed Data}) * (\text{Non - Overlaps Expected Data})}{(\text{Non - Overlaps Observed Data}) * (\text{Overlaps Expected Data})}$$

2.1.12 Two-tailed two-sample t-test

In order to test for a significant difference between two odds-ratios of datasets, which had no common SNPs, a two-tailed two-sample t-test was applied assuming unequal variances of the two compared SNP sets. The test used the natural logarithm of the odds ratios, the natural logarithm of the standard error of the odds ratio, and the number of SNPs per analysed set to calculate a *P*-value for the difference of the odds ratios. The total number of analysed SNPs divided by two determined the degrees of freedom. The *P*-value was corrected for multiple testing using the Bonferroni correction, *i.e.,* for the number of annotations that were analysed. The *P*-value was significant if it was below the adjusted threshold of significance. The annotations, for which the difference of odds ratios was significant, were identified using a red star in all graphs.

2.2 Genomic annotations

2.2.1 Genome build and sources

Details on the genomic annotations sources are included in the corresponding paragraphs. All genomic annotations were downloaded in hg18, where available. The hg18 build was chosen, as more annotations were available for that build at the beginning of the project, than either the previous (hg17) or the later one (hg19). If they were not available in hg18, the UCSC liftOver tool [94] was used to transfer the annotated regions into hg18. The majority of the genome annotations were downloaded from the UCSC genome browser and were publicly available at the time of download. Other sources included the ENSEMBL webpage, which allowed selective download of a number of variations with specified biological functions. These were only available for download in hg19 and were transferred to hg18. The remainder of the sources were laboratory web pages which had made their data available online. If necessary, they were converted into the appropriate build to ensure the correct relative map positions.

2.2.2 Categories of genomic annotations

The final analysed data included 58 genomic features for which the genome was annotated. These were separated into three major categories to enable appropriate representation of the different underlying biology. The genomic annotations chosen for analysis were similar to, if not the same as, the 20 annotations previously published in the analysis we are replicating [50]. However, some of the annotations were no longer publicly available at the time of download (e.g., the regions under accelerated rates of substitution in the human genome). For these annotations, we used either approximations of the annotations, or, based on the significance of the odds ratios calculated by Hindorff *et al.*, were not included in this analysis. Of the 20 published annotations, 14 were downloaded and results obtained in theses categories were compared to the published results.

Genomic annotations that were added to the genomic annotation set included a range of conserved regions, areas with signs of purifying selection, and a large number of distinct and replicable histone modification patterns. The latter replaced a large number of individual histone modifications that were used as proxy for the different underlying biological functions.

A detailed description of the 58 genomic annotations and their sources are outlined below according to the category they belonged into. The three categories were genic and regulatory features, conserved and regulatory regions, and chromatin states.

2.2.2.1 Genic and regulatory features

This category contained all genomic annotations that were within genes, defined by their proximity to genes, or identified through sequence analyses as a regulatory element of transcription factor binding site. The text reflects the order of the genomic annotations as they are shown on the corresponding graphs in each figure.

The region upstream of the transcription start site (TSS) of a gene has strong literature evidence of containing putative promoters. In order to analyse these regions and identify long-range *vs.* short-range regulatory elements, two distances upstream of the TSS were analysed: **1 Kb** and **5 Kb upstream of TSS**. These two genomic annotations were derived from the RefSeq dataset downloaded from the UCSC table browser (accessed 15th November 2010), which details the position of the TSS and the strand on which the gene is found [95].

CpG islands are areas in the genome with a large proportion (larger than expected by chance) than expected by chance of unmethylated cytosines followed immediately by a guanine, where only a phosphate group separates the two. The unmethylated state of a single CpG is rare and will only be present if there is selective pressure to keep it unmethylated [96]. The methylated cytosine tends to turn into thymines due to spontaneous de-amination. The CpG islands are associated with promoter functions and – in vertebrates – in particular with housekeeping genes [97]. This annotation set was also downloaded on 15th November 2010 from the UCSC table browser.

The **ORegAnno** annotation reports regulatory regions, regulatory polymorphisms and transcription factor binding sites and was downloaded on 15th November 2010 from the UCSC table browser. It originates from the Open Regulatory Annotation database, which is an online repository that is publicly curated containing information validated through experiments [73, 98].

The Vertebrate Genome Annotation (vega) database contains frequently manually annotated regions with information on protein-coding genomic regions as well as pseudo genes and immunoglobulin segments. These were divided into the **vegaGenes** annotation and the **vega PseudoGenes** annotation, which were downloaded from the UCSC table browser on 15th November 2010.
The **OMIM genes** and **OMIM morbid regions** are no longer publicly available, but were available at the time of download. The Online Mendelian Inheritance in Man is a continuously updated catalogue of human genes and genetic disorders which incorporates all genes and genomic regions that have been identified through experiments. The OMIM morbid map shows the cytogenetic locations of specific diseases identified by previous studies [99].

The **Exons** annotation was derived from the RefSeq gene annotation [95], at the same time as the 1 Kb and 5 Kb upstream TSS annotations were created. This category included all exons possible through different splicing to take all isoforms into account. A number of SNP annotations, *i.e.,* **intronic, non-synonymous, synonymous, intergenic, splice sites** and sites in the **3' and 5'UTRs**, were extracted from the dbSNP 129 dataset [16], accessed on 20 January 2011. The non-synonymous and synonymous SNPs were combined to create the **coding SNPs** annotation. Non-synonymous SNPs resulting in **gained** or **lost stop codons** were downloaded from the ENSEMBL webpage.

The locations of **RNA genes**, which are genes that were expressed but were not coding for proteins and pseudo genes [100, 101], were downloaded from the UCSC Table browser on 25 November 2010. The genomic locations which were known to correspond to **totally intronic non-coding RNAs** were downloaded from the RNA database on 15 November 2010 [102]. The regulatory target sites for conserved mammalian microRNA families in the 3'UTRs of RefSeq Genes were predicted by an algorithm called **TargetScanS** [103-105]. These sites were downloaded from the UCSC Table Browser on 15 November 2010.

The **eQTLs** are defined as SNPs that have been associated to variation in gene expression levels. The SNPs represented in this annotation were downloaded from the eQTL web browser [106] and originate from a number of different studies [72, 107-110].

DNase Clusters represent DNase hypersensitive areas assayed in a large collection of cell types and have been shown to play a major role in human traits [111]. The dataset was downloaded from the UCSC Table Browser on 15 November 2010 [112].

Insulators in the human genome are necessary boundaries between different areas of the genome, which are translated or silenced. Genomic locations for human insulators were downloaded using the ENSEMBL biomart on 15 November 2010.

SNP sites found **within mature microRNAs** were downloaded using the ENSEMBL biomart to investigate if trait-associated SNPs can be preferentially found within microRNAs. The sites were downloaded on 15 November 2010.

Regions with sequences of at least 15 perfect di-nucleotide and tri-nucleotide repeats are likely to be useful as microsatellite markers and are usually highly polymorphic between populations [113]. This annotation was downloaded from the UCSC Table browser on 15 November 2010.

2.2.2.2 Conserved regions and evolutionary signatures

The conserved and evolutionary signatures category was chosen to represent genomic locations, which were either conserved between different species, or have been shown to be under selective pressures. The **evofold** annotation corresponded to RNA secondary structure predictions made with the evofold program. This program compared multiple-sequence alignments to identify conserved functional RNA structures [114]. This annotation was downloaded on 15 November 2010 from the UCSC Table browser.

The identity and genomic locations of 16,529 high-confidence orthologues showing tested for positive selection were downloaded from the UCSC Table browser on 15 November 2010. The high-confidence orthologues were from a multiple mammal alignment using the genome assemblies of human (hg18), chimp (panTro2), macaque (rheMac2), mouse (mm8), rat (rn4), and dog (canFam2). These genes were analysed in different evolutionary lineages to investigate mammalian positive selection [115]. However, here only the orthologues were used without restriction to the **positive selection** score.

A set of **enhancers** identified through a number of computational and experimental analyses to identify possible enhancers in human and mice were downloaded on 15 November 2010 from the VISTA enhancer browser [116].

Exapted repeats are conserved non-exonic sites that have been deposited by mobile sites (repeats) in a process called exaptation. These repeats are possible distal enhancers and were downloaded from the UCSC Table browser on 15 November 2010 [117-121].

Predicted *cis*-acting regulatory modules are presented in the **PREMOD** genomic annotations, which were downloaded from the PREMOD database [122, 123]. The location and score of **transcription factor binding sites** that are conserved in a human/mouse/rat alignment, based on computational predictions, were downloaded from the UCSC Table browser on 15 November

2010. The data were generated using the Transfac Matrix and Factor databases on Biobase, while Matt Weirauch and Brian Raney at the University of California at Santa Cruz created the track for the UCSC Table browser. Regions significant for purifying selection with respect to mutations involving sequence insertions and deletions have been implicated as possibly identifying long intergenic non-coding RNAs, which may have an impact on phenotypes [124]. This dataset, referred to as **Indels**, was downloaded on 15 November 2010 from the UCSC Table browser.

Conserved sites, identified using a number of different species alignments, may have a larger than expected chance of containing trait-associated variants, as conserved sites are thought to have an important biological function. The phastCons program was used for the different species alignments [76]. The species used for the **17 species alignment** were human (March 2006 (NCBI36/hg18), hg18), chimp (November 2003, panTro1), macaque (January 2006, rheMac2), mouse (February 2006, mm8), rat (November 2004, rn4), rabbit (May 2005, oryCun1), dog (May 2005, canFam2), cow (March 2005, bosTau2), armadillo (May 2005, dasNov1), elephant (May 2005, loxAfr1), tenrec (July 2005, echTel1), opossum (January 2006, monDom4), chicken (February 2004, galGal2), frog (October 2004, xenTro1), zebrafish (May 2005, danRer3), tetraodon (February 2004, tetNig1) and fugu (August 2002, fr1).

The **28 species alignment** includes all species from the 17 species alignment, six of which use updated sequences, and 11 new species. The updated species are for the chimp (March 2006, panTro2), cow (August 2006, bosTau3), chicken (May 2006, galGal3), frog (August 2005, xenTro2), fugu (October 2004, fr2) and zebrafish (March 2006, danRer4). Five of the new species are high-coverage (5-8.5X) assemblies of horse (February 2007, equCab1), platypus (March 2007, ornAna1), lizard (February 2007, anoCar1), stickleback (February 2006, gasAcu1) and medaka (Apr 2006, oryLat1), while the remaining six were low-coverage assemblies (2X) from bush baby (December 2006, otoGar1), tree shrew (December 2006, tupBel1), guinea pig (October 2005, cavPor2), hedgehog (June 2006, eriEur1), common shrew (June 2006, sorAra1), and cat (March 2006, felCat3). These 28 species were aligned and a subgroup of the species was used to identify elements conserved between placental mammals. The **placental mammals** excluded the sequences of the opossum, platypus, chicken, lizard, and frog as well as the following fish: tetraodon, fugu, stickleback, medaka and zebrafish [125] from the 28 species set above.

The **44 species alignment** is composed of the 28 species listed above and an additional 16 new ones. Eight of the 28 previous species were updated. The new assemblies are mouse (July 2007, mm9), cow (October 2007, bosTau4), guinea pig (February 2008, cavPor3), horse (Sep 2007, equCab2), elephant (July 2008, loxArr2), zebrafish (July 2007, danRer5), and medaka (October 2005, oryLat2). The orang-utan (July 2007, ponAbe2) and zebra finch (July 2007, danRer5) were high-coverage (5-8.5X) assemblies, and gorilla (October 2008, gorGor1), marmoset (June 2007, calJac1), tarsier (August 2008, tarSyr1), mouse lemur (June 2003, micMur1), kangaroo rat (July 2008, dipOrd1), squirrel (February 2008, February 2008, speTri1), pika (July 2008, ochPri2), mega bat (July 2008, pteVam1), micro bat (March 2006, myoLuc1), dolphin (February 2008, turTru1), alpaca (July 2008, viPac1), sloth (July 2008, choHof1), rock hyrax (July 2008, proCap1), and lamprey (March 2007, petMar1). The subsets were obtained from the total alignments of the 44 species from which some species were selected. The **placental mammals** included the sloth, armadillo, tenrec, rock hyrax, elephant, common shrew, hedgehog, mega bat, micro bat, dog, cat, horse, cow, dolphin, alpaca, pika, rabbit, squirrel, guinea pig, kangaroo rat, rat, mouse, tree shrew, bush baby, mouse lemur, tarsier, marmoset, rhesus monkey, orang-utan, gorilla, chimp and human. The **primates'** subset included the latter species from bush baby to human.

Overlaps of any of the above conserved, regulatory or genic sites were removed from the intergenic SNP set. This created a **negative** genomic feature, which was depleted of any regulatory elements, irrespective of chromatin states or histone modifications.

2.2.2.3 Chromatin states and histone modifications

Two higher order structures of chromatin were included in the analysis, which were established in a karyotypically normal lymphoblastoid cell line, GM06990 [126]. These chromatin states were downloaded on 15 November 2010 from the GO website (accession number: 19815776). The data identified two regions chromatin regions with different interaction patterns. The chromatin in a more "open" and more accessible state was associated with active genes and transcription patterns, while chromatin in a more "closed" conformation was associated with inactive genes. These were included in the analysis with the names **Open Chromatin** and **Closed Chromatin**, respectively. Additionally to these higher order structures, the lower order structures were also analysed. Initially, these consisted of a large number of individual histone modifications in different cell lines. However, the identification of 15 different replicable histone modification patterns in nine different cell lines made the set of individual modifications obsolete. These 15 different patterns were associated with underlying 15 biological functions in the genome [82]. The authors used nine different cell lines, out of which we used the GM12878 cell line to ease comparison with the open and closed chromatin states. The **biological functions associated with the histone modifications** were promoter, enhancer and insulator activities. The former two could be separated into different subcategories, which for the promoters were active, weak and inactive/poised. The active, weak and poised promoter labels were highly interchangeable between different cell lines, so that overall these labels pointed to regions with transcribed genes. The same can be said for strong and weak enhancers and transcribed regions. However, the identity of these regions only changed within their classes, so that it can be said that these regions tend to preserve their regulatory potential, still retaining their biological functions [82]. The enhancer and repetitive regions showed different positional enrichment along transcripts, where some elements acted on distal or rather more proximal genes [82]. These histone modifications were downloaded from the UCSC Table browser on 09 June 2011 ("wgEncodeBroadHmmGm12878HMM").

2.3 Distribution of genomic annotations

Table 2-1 shows the descriptive statistics of the 58 genomic annotations. The distribution of the genomic annotations was important for the analysis, as a low number of annotated SNPs caused odds ratios and confidence intervals with values of infinity. The table includes information on coverage of the annotations in number of sites in nucleotides (*i.e.,* genome coverage) and SNPs (*i.e.,* SNPs genotyped or imputed in the GWA studies surveyed), and the mean length of each annotation in the entire genome. Four (within mature miRNA, splice sites, lost stops and microsatellites) of the annotations had a very low coverage of the SNPs and were excluded from all graphs, as the odds ratios were undefined due to a division by zero. The mean allele frequency of 647,776 SNPs identified in the Human Genome Diversity Project overlapping with the annotations is also shown in Table 2-1. The MAFs are very stable across the annotations, so that a possible selection effect is unlikely to have influenced the results.

Table 2-1 – Summary statistics of the three classes of genomic annotations

This table shows the number of annotation sites (Sites), their mean length in base pairs (Mean Length (bp)), the percentage of nucleotides coinciding with them (Nucleotides (%)), the percentage of SNPs coinciding with them (SNPs (%)) and the mean minor allele frequency of SNPs in the HGDP overlapping with annotations (Mean MAF.).

Annotations	Sites	Mean Length (bp)	Nucleotides (%)	SNPs (%)	Mean MAF
TSS 1 Kb upstream	22624	1069.60	0.79	3.40	0.23
TSS 5 Kb upstream	20592	5533.29	3.70	8.83	0.23
CpG Islands	27458	764.28	0.68	1.76	0.23
ORegAnno	17903	627.89	0.37	1.97	0.23
Vega Genes	14651	64881.51	30.90	37.89	0.22
OMIM Genes	12307	64852.28	25.90	32.74	0.22
OMIM Morbid Regions	2532	69311.51	5.70	7.82	0.22
Exons	212325	245.31	1.69	7.14	0.22
Intronic SNPs	5125999	1.76	0.29	43.61	0.22
Non-Syn. SNPs	117692	55.54	0.21	3.73	0.22
Coding SNPs	186247	35.59	0.22	5.74	0.22
Syn. SNPs	72933	1.50	3.54×10^{-3}	2.99	0.13
Gained Stops	4186	1.01	1.37×10^{-4}	0.05	0.20
3'UTR	131649	1.63	0.01	3.81	0.22
5'UTR	27693	1.41	1.27×10^{-3}	0.94	0.27
RNA Genes	6936	132.54	0.03	0.25	0.23
ncRNA	890	15355.22	0.44	0.83	0.22
TS miRNA	40648	7.69	0.01	0.03	0.23
eQTLs	68619	1.00	2.23×10^{-3}	4.50	0.25
Vega PseudoGenes	6999	3094.04	0.70	1.78	0.22
Intergenic SNPs	8250331	1.73	0.46	63.71	0.22
DNase Clusters	969313	243.90	7.67	31.25	0.23
Insulators (sequence)	25546	1095.15	0.91	4.29	0.23
Within miRNA	395	1.09	1.39×10^{-5}	0.00	0.23
Splice Sites	1718	4.44	2.48×10^{-4}	0.04	0.23
Lost Stops	278	1.02	9.19×10^{-6}	0.02	0.25
Microsatellites	40186	40.56	0.05	0.08	0.25
Evofold	47244	38.81	0.06	0.21	0.23
Pos. Sel. Genes	16384	39030.69	20.80	28.43	0.22
Enhancers	1295	1526.59	0.06	0.32	0.22
Exapted Repeats	10400	99.50	0.03	0.29	0.22
PREMOD	122979	482.55	1.93	10.39	0.23
tfbsConsSites	2345848	16.45	1.25	8.48	0.23
Indels	2596839	82.23	6.93	32.61	0.23
17 spc. algmt	2201980	66.24	4.74	21.74	0.23
28 spc. algmt, plc.mmls	2028316	54.65	3.60	17.23	0.23
28 spc. algmt	2873612	48.33	4.51	20.30	0.23
44 spc. algmt	4846954	29.02	4.57	20.52	0.23
44 spc. algmt, plc.mmls	3945677	31.31	4.01	18.75	0.23
44 spc. algmt, prim	806524	150.28	3.93	17.85	0.22
Negative (sequence)	5277572	509.71	97.90	54.81	0.22
Open Chromatin	13843	99999.00	44.90	46.77	0.22
Closed Chromatin	13469	99999.00	43.70	52.95	0.22
Active promoter	15279	1440.81	0.72	2.82	0.23
Weak promoter	35076	568.26	0.65	2.99	0.23
Inactive/poised promoter	5265	891.55	0.15	0.35	0.22
Strong enhancer (proximal)	25486	964.20	0.80	3.13	0.23
Strong enhancer (distal)	38612	621.47	0.78	3.47	0.23
Weak/poised enhancer (proximal)	69144	388.98	0.87	4.37	0.23
Weak/poised enhancer (distal)	109526	555.25	1.97	8.14	0.23
Insulator	33311	468.99	0.51	3.36	0.23
Transcriptional transition	16223	1223.81	0.65	2.51	0.22
Transcriptional elongation	26473	5975.22	5.14	9.13	0.22
Weak transcribed	82235	3671.65	9.80	16.51	0.22
Polycomb repressed	25483	3524.60	2.92	7.29	0.22
Heterochrom; low signal	10530	891.55	0.31	81.79	0.22
Repetitive/CNV (proximal)	8033	627.54	0.16	0.24	0.24
Repetitive/CNV (distal)	6122	452.66	0.09	0.20	0.25

3 GETTING IT RIGHT: REPLICATION OF A PREVIOUS STUDY

3.1 Introduction

As mentioned in the introduction chapter of this thesis, most trait-associated GWAS hits are found outside of genic and usually coincide with genomic regions whose functions have not yet been identified. This hinders the identification of the causal underlying biology for the majority of GWAS hits [36, 127], as a target for follow-up studies is not immediately obvious. An investigation into the genomic environment of trait-associated SNPs and their LD partners was therefore warranted to aid the understanding of trait-associated variants. The question as to which genomic features underlie trait-associated SNPs more often (or less often) than expected by chance arises when GWAS hits are investigated. The answer to that question is important for future research, as it could be used in prediction mechanisms for trait-associated variants.

In 2009 Hindorff *et al.* published a study to answer the above question by investigating genomic regions for enrichment or depletion of trait-associated SNPs to identify potential aetiological mechanisms [50]. The 20 analysed genomic regions or features were mainly genic and mutually non-exclusive annotations, *i.e.,* the annotations were coinciding with each other. The results showed three annotations with significant odds ratios of enrichment and depletion of trait-associated variants. Non-synonymous SNPs and regions 1 Kb upstream of a transcription start site (TSS) were significantly enriched, while intergenic SNPs were significantly depleted for trait-associated SNPs. These results were obtained by creating 100 samples of non-associated SNPs, which closely matched the genotype array composition of the observed data, *i.e.,* the trait-associated SNPs, as we explain shortly. The majority of genotyping arrays were designed with a specific purpose in mind. The Illumina HumanHap 300 genotyping array was enriched for non-synonymous SNPs and targeted mainly common SNPs [50]. This inherent ascertainment bias present within all arrays could cause problems in the sampling analysis, if GWAS hits were to be compared with non-associated SNPs. If, for example, one were to choose sample

SNPs from the Illumina HumanHap 300 array only and the SNPs originated from a different platform, one would erroneously find relative depletion in the non-synonymous SNPs in comparison. This bias of the genotyping arrays and any imputed SNPs in the original trait-associated dataset had to be considered by any method aimed at obtaining random samples. This sampling with taking account of the genotyping array composition produced a background, or null-distribution, of expected data to which the observed data were compared and odds ratios of enrichment or depletion of the trait-associated SNPs were calculated.

Hindorff *et al.* showed that trait-associated SNPs had a distinct distribution in the investigated genomic regions with significant results in three of 20 annotations. However, since the study's publication in 2009 many more GWAS have been performed identifying many new trait-associations. An additional study into the new associations was therefore warranted. We began with a replication of the results with the original data, which was needed to compare the original results with the results from the more recent dataset. We also included additional genomic annotations to investigate a broader range of regulatory regions. Epigenetic modifications were also included, as they have been shown to contribute to stress-related phenotypes such as e.g., cancer or diabetes [77, 128, 129].

Here, we detail the steps taken to replicate the data, methods, and results of the Hindorff *et al.* study. We then performed an investigation with a larger set of SNPs, which were a more recent version of trait-associated variants from the NHGRI catalogue of GWAS results [86]. The results of the two SNP sets, the replicated set and the more recent set with more variants, were compared to the results obtained by Hindorff *et al.* We additionally expanded the investigation with more genomic features. Our results show that the sampling method could be reproduced thereby validating the way we performed the sampling method. This validation was necessary for an appropriate comparison of the sampling results with results obtained by a novel method discussed later.

3.2 Method

3.2.1 Sampling genotyping arrays

The study we aimed to replicate was published in 2009 by Hindorff *et al.* [50], who used a sampling method to analyse the distribution of trait-associated variants in 20 genomic annotations. The method obtained sample sets of SNPs of equal size to the set of trait-associated SNPs represented on genotyping platforms. We used weighted groups based on the manufacturer(s) of the SNP platform(s) and the HapMap CEU II data to draw the samples, rather than on individual genotyping arrays, as that information was often unavailable. The numbers of SNPs drawn from each manufacturer group were proportional to the number of SNPs observed in the real data. Groups were established representing the union of varied combinations of genotyping arrays and imputed data. These groups were randomly sampled in the proportions of the observed data. Multiple entries of individual SNPs were possible and were not removed. This takes into account the greater chance of a SNP to be identified as a trait-associated SNP, if it was present on more than one genotyping array.

3.2.2 Odds ratios

Odds ratios, confidence intervals and *P*-values of significance for the observed results were calculated using the oddsratio.wald function from the epitools R package [130] of the statistical program R version 2.12.1 [131]. This function calculated the odds ratios by comparing unconditional maximum likelihoods of the observed value compared with the mean number of expected hits. Odds ratios of enrichment/depletion were calculated by comparing overlaps between genomic features and real trait-associated data with overlaps of SNPs determined by chance alone. The *P*-values were defined as significant when below the Bonferroni-corrected significance threshold, which in our case was calculated for 58 independent variables ($P \leq 8.62 \times 10^{-4}$). The analysed genomic annotations were not independent from each other, which means that the Bonferroni corrected *P*-value is conservative.

3.3 Results

3.3.1 Preliminary work

3.3.1.1 Study populations

In order to ascertain the correct population for the establishment of the LD blocks, an investigation was undertaken into which population was used the most often in GWAS. Hindorff *et al.* already showed that populations from European descent were the most numerous. However, for completeness we also reinvestigated this. Three populations were analysed: European, Asian, and African. The European category contained all studies specifying European populations (e.g., Croatian or Scottish) or those defining their study population as Caucasian or white. The Asian category consisted of studies with several populations with Asian background, such as Malaysian, Thai, Chinese or Japanese. The African category included populations such as Ghanaians or populations with African ancestry, e.g., African Americans. Since Reconstructed Hindorff SNPs were obtained from the authors of the original paper, it was expected that its structure was as published. In the Significant SNPs (2011) dataset 374 of 576 studies specified their study population in either the title or the sample descriptions.

Histograms of the risk allele frequencies in the entire dataset (A), the European (B), Asian (C), and African (D) populations respectively are shown in Figure 3-1. The histogram of the European population (green) matches that of the risk allele frequencies of all significant trait-associated SNPs the best.

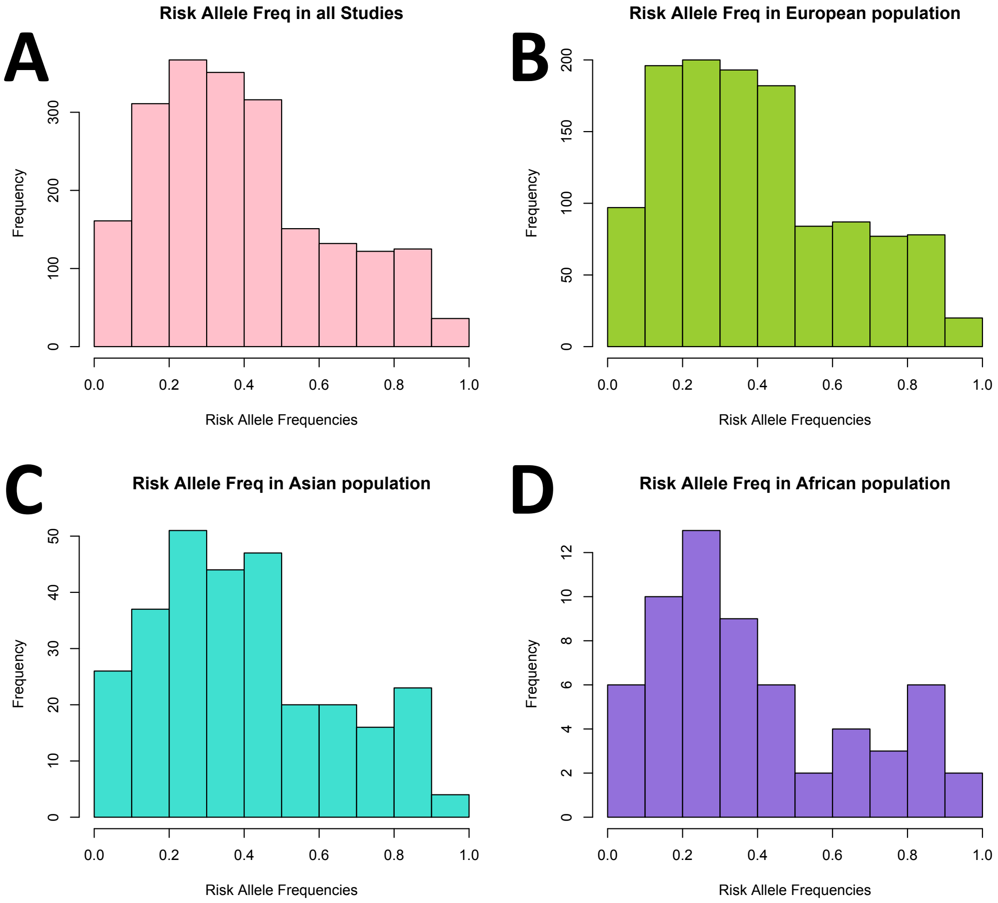


Figure 3-1 – Risk allele frequencies in published studies
The risk allele frequencies of all reported trait-associated SNPs are shown in the panel A (pink). Their distribution closely matches the distribution of the risk allele frequencies of trait-associated SNPs from GWAS specifying their study population as either European or white (green, B). While the Asian (turquoise, C) and African (purple, D) populations show similar trends, the numbers of observed variants are much smaller.

Additionally, the Euler diagram in Figure 3-2 showing the number of studies for each population and the overlap between all studies indicates that the majority of studies specified the use of a European study population. The study population for Significant SNPs (2013) was also the European population, as an Euler diagram of the studies with specified populations showed the same proportions as in Significant SNPs (2011). The definitions for the populations were the same as outlined previously. Significant SNPs (2013) contains many more studies than Significant SNPs (2011), and they are equally distributed across all populations.

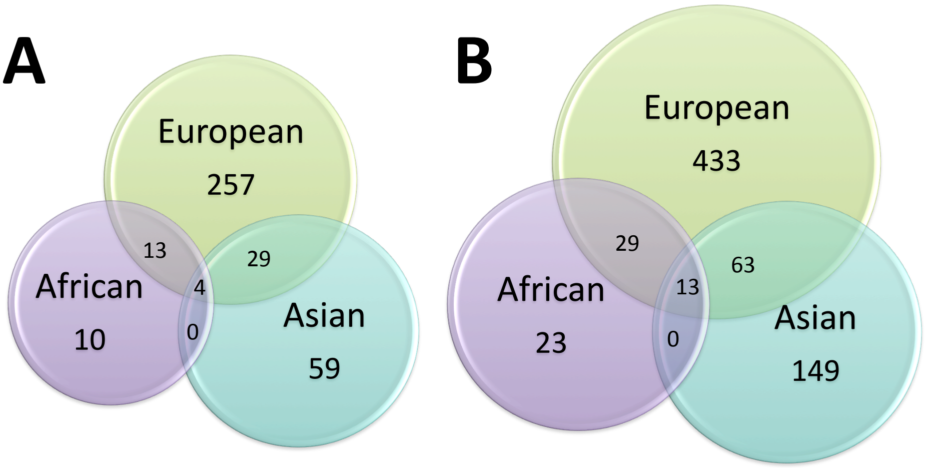


Figure 3-2 – Study populations of Significant SNPs (2011) and Significant SNPs (2013)
A) A total of 374 of 576 unique studies in the large dataset, which specified the study population (202 unspecified). There were four studies, which analysed all three populations but zero studies, which compared Asian populations with African populations. B) The total number of studies specifying the study populations was 710. The number of studies using European or African populations has increased by almost two-fold when compared to the data from 2011. However, the number of studies using Asian populations has more than tripled. The majority of these studies were performed in the Chinese population (data not shown). Of the 929 unique studies, 219 did not specify the study population in either the sample descriptions or the study title.

The results of this investigation led to the use of the HapMap CEPH LD data (CEU, release #24) in this study to establish LD partners of trait-associated or sample SNPs, and to make up the groups for the sampling of the genotyping arrays mixed with imputation results. The CEU data are obtained from Utah residents with ancestry from Northern and Western Europe [132]. Additionally, the use of this data is in concordance with published studies. The African population is known to be genetically the most diverse population with reportedly the smallest LD blocks [133].

3.3.1.2 LD partners vs. LD Blocks

There were two ways of creating LD blocks surrounding the analysed SNPs, which were available for our study. One way was using the Trait-Associated SNP Partners (TASPs), while the other way was using the Trait-Associated Blocks (TABs). The thesis was done using TASPs, as this analysis investigated only LD partners, while TABs analysed all nucleotides enclosed by the furthest LD partners of a SNP. Both ways had their benefits and drawbacks. The benefits of the TABs method were that the regions of the genome between two SNPs were included in the analysis, while the TASPs analysis only investigated those SNPs for which LD was calculated and the LD passed the cut-off point off. TABs would be advantageous for the analysis of sparse genomic annotations that did not often coincide with SNPs and even less often with trait-associated variants. The use of TASPs missed out those annotations that did not overlap with SNPs and furthermore did not allow the analysis of regions, which were difficult to genotype or sequence. However, it did guarantee that all analysed SNPs were in LD at the required threshold. This was not guaranteed in the TABs analysis, as LD varies across distances and with allele frequencies. Additionally, the value of LD was unknown for the regions between two SNPs. Figure 3-3 shows a cartoon-like representation of the same stretch of DNA for both TABs and TASPs, highlighting the overlap of a genomic annotation with the TAB method, but not with the TASP method. It was therefore decided that TASPs should be used rather than TABs, since the LD threshold was known for all analysed variants and because the TABs analysis included more noise in the results.

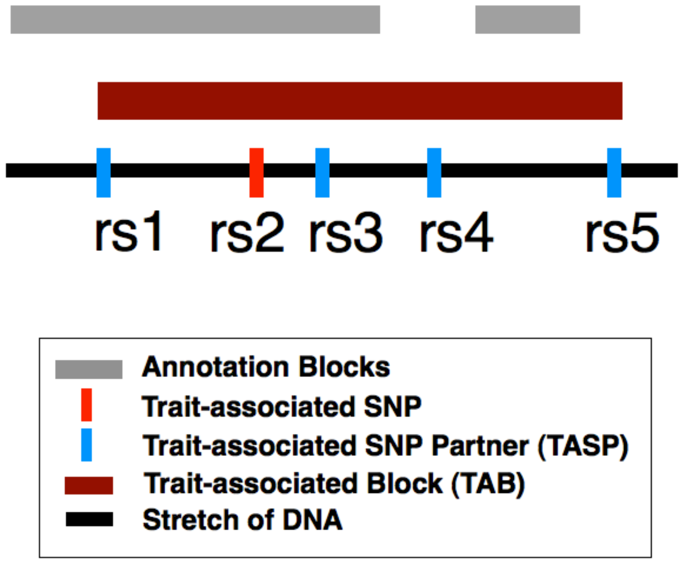


Figure 3-3 – Diagram of Trait-associated SNP partners (TASPs) and Trait-associated blocks (TABs) This diagram highlights the differences between Trait-associated SNP partners (TASPs) and Trait-associated blocks (TABs) for the genomic region. The block (brown) surrounding the trait-associated SNP (red) overlaps with both genomic annotation blocks (grey), however, only one block coincides with TASPs. The smaller annotation block coincides with the genomic region between rs4 and rs5 and is not counted as an overlap in the TASP analysis.

3.3.2 Replicating significant enrichment results

Odds ratios of enrichment/depletion of trait-associated SNPs were calculated for each genomic annotation. An odds ratio equal to unity indicated that trait-associated SNPs were as likely to coincide with the analysed genomic feature as non-associated SNPs. An odds ratio above unity indicated that the genomic feature was enriched for trait-associated SNPs, while odds ratios below unity were evidence for depletion. Figure 3-4 compared the published results with the results obtained for Reconstructed Hindorff SNPs. Our replication of the sampling method compared well with the published data. The trend of the enrichments and depletion were almost equivalent with enrichment in all of the genomic annotations, except for the intergenic SNPs. The odds ratio was not available for the TS annotation in the Reconstructed Hindorff SNPs, as none of the analysed SNPs overlapped with it. The observed correlation for the two datasets, once the not-available annotation was removed from both sets, was 0.84 with a *P*-value of 6.70 × 10^{-06}. This meant, that where the information was available, the two sets agreed well with each other.

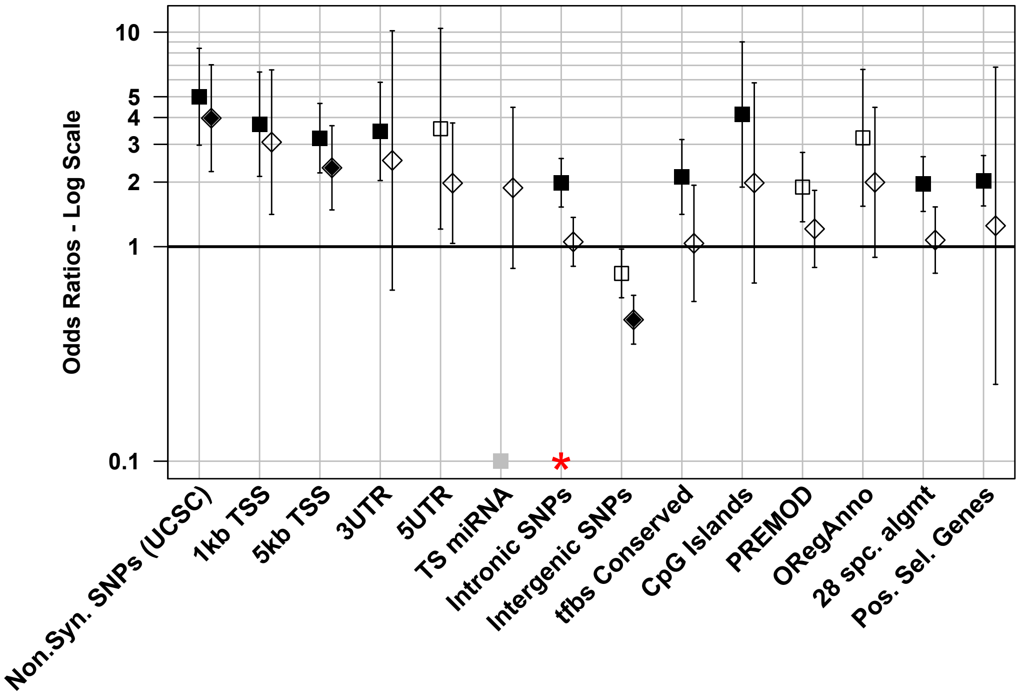


Figure 3-4 – Comparison of sampling results with published results
The odds ratios for trait-associated SNPs from Hindorff *et al.* (*n* = 465 SNPs; □) and our analysis using Reconstructed Hindorff SNPs (*n* = 468 SNPs; ◇) in selected genomic annotations are shown above. All results are displayed in odds ratios along with 95% confidence intervals, where solid symbols indicate significance at the Bonferroni corrected threshold. Odds ratios below or above one show depletions or enrichments respectively. Red stars (✱) at the bottom of the graph indicate significant differences between odds ratios. Grey symbols indicate that the odds ratio is undefined.

The differences in significant odds ratios might be explained through author-specific differences in the genomic annotation datasets, which were discussed in the Discussion sections of this chapter. Of the three significant results from the Hindorff *et al.* study, we have replicated two. These replicated results are for the non-synonymous SNPs annotation, which was the most significant and most enriched genomic annotation in the published data, and the 1 Kb upstream of TSS. The latter identified putative promoter regions, as most promoters were located upstream and in proximity to a TSS. The depletion observed by Hindorff *et al.* in the intergenic SNPs was replicated, although in our analysis, the odds ratio was no longer significant after correcting for multiple testing. In our analysis, we saw significant odds ratios of enrichment for seven more annotations. However, the 95% confidence intervals of the odds ratios of either study were overlapping with each other, indicating that they were not significantly different to each other.

The differences between sets were discussed in further detail in the Discussion of this chapter. Odds ratios could be interpreted in terms of fold enrichment, as the measures were almost identical with a significant correlation of 0.96 (Figure 3-5) when calculated for all annotations of Significant SNPs (2011). The *P*-value of the observed correlation was significant (8.33×10^{-42}).

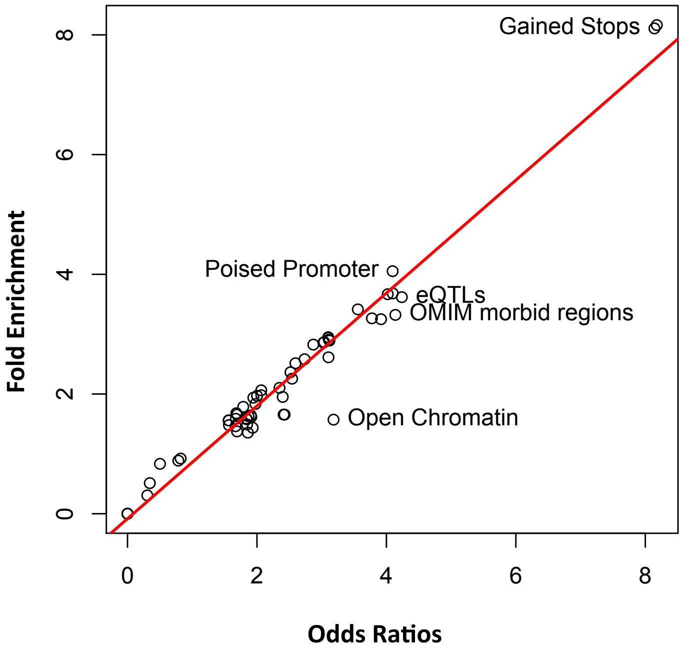


Figure 3-5 – Fold enrichment *vs.* odds ratios

The odds ratios and the fold enrichment calculated for Significant SNPs (2011) are plotted against each other. The red line indicates the line of best fit for the odds ratios and the fold enrichment. There is a strong and significant correlation between fold enrichment and odds ratios ($r^2 = 0.96$, *P*-value = 8.33×10^{-42}).

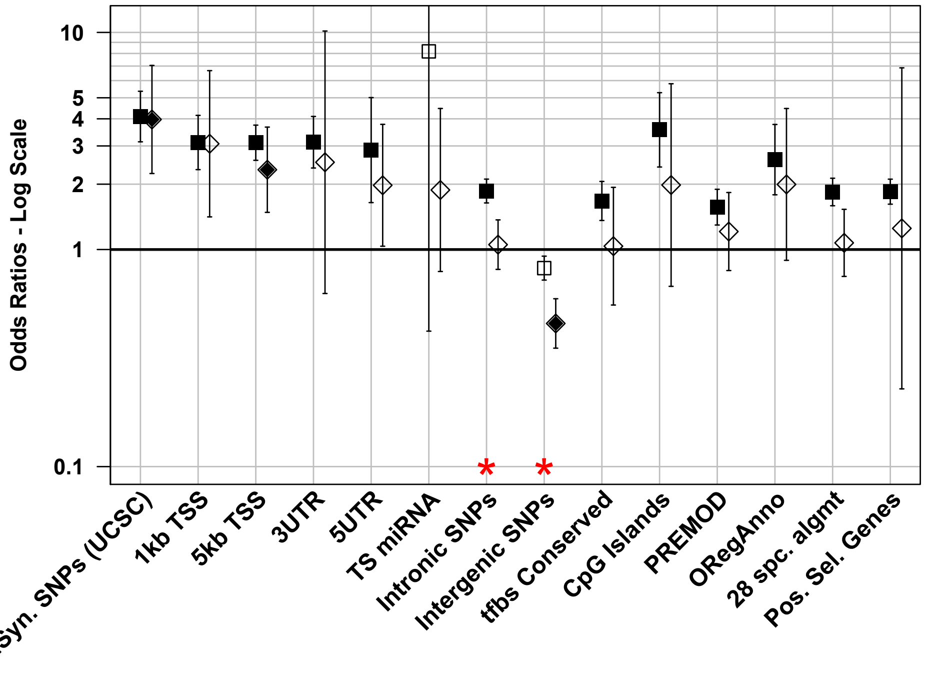


Figure 3-6 – Comparing published results with larger dataset
Enrichment of trait-associated SNPs in selected genomic annotations for Significant SNPs (2011) and Hindorff *et al.* (□,◇ respectively). All results are displayed in odds ratios along with 95% confidence intervals, where solid symbols indicate significance at the Bonferroni corrected threshold. Odds ratios below or above one show depletions or enrichments, respectively. A red star (*) at the bottom of the graph indicates significance at the Bonferroni corrected *P*-value.

Figure 3-6 shows the comparison of the results of the sampling strategy for the Hindorff results with Significant SNPs (2011) containing more SNPs than shown in Figure 3-4. This new set with more analysed SNPs is expected to have more statistical power to detect enrichment or depletion of associated SNPs. The coefficient of the regression of the odds ratios obtained by Hindorff *et al.* with the two analysed SNP sets is very high (Reconstructed Hindorff SNPs: 0.84, Significant SNPs (2011): 0.83) in the same annotations as above. The correlation for the odds ratios of the Hindorff set with Significant SNPs (2011) is significant with a *P*-value of 8.99 × 10^{-6}. We observed two genomic annotations in which the odds ratios differ significantly. The difference in the intronic SNPs is again significantly different, as in the comparison shown previously. However, the difference in the intergenic SNPs is unexpected. Since the trend of the odds ratios remained the same, the observed difference in significance of the odds ratios is most likely due to the decreased width in the confidence interval in the analysed genomic features.

3.3.3 Comparison of Reconstructed Hindorff SNPs and Significant SNPs (2011)

Of the 58 analysed annotations, four were too sparsely distributed in the genome to obtain any odds ratios. Summary statistics for the analysed annotations were calculated for all annotations (see Table 2-1). The table summarised the number of sites, the percentage of nucleotides covered in the analysed part of the genome, the percentage of SNPs covered in the analysed part of the genome, and the average length of the annotated sites in base pairs. Figure 3-7 shows the comparison of Reconstructed Hindorff SNPs with Significant SNPs (2011) and Table 3-1 and Table 3-2 present the results for the two sets, respectively. The three subcategories are displayed in three panels: The genic and regulatory regions are in the top panel, the conserved regions in the middle panel, and the chromatin states and histone modifications are included in the bottom panel. There were no significant differences between the two SNP datasets in any of the genomic annotations. The regression line of the available odds ratios of the two sets in all categories was 0.61 and significant (*P*-value = 5.16×10^{-12}). A number of annotations are significant for Significant SNPs (2011), which are not significant for Reconstructed Hindorff SNPs. The most prominent differences are the odds ratios for the synonymous SNPs and the 5'UTRs that were significant in our analysis.

SNPs in the HapMap CEU II data and all analysed genotyping arrays were analysed for an overlap of genomic annotations. There were a number of annotations, which rarely overlap with SNPs (see Table 2-1). These annotations overlapped with trait-associated SNPs at an even lower rate, as the trait-associated SNPs were a subset of the total set of analysed SNPs in the genome. The Significant SNPs (2011)(n = 1,909 SNPs) resulted in a defined odds ratio for these genomic annotations (e.g., TS miRNA or evofold), while Reconstructed Hindorff SNPs (n = 468) did not. The odds ratios for these sparse annotations were not significant and had large 95% confidence intervals (see Table 3-1 and Table 3-2).

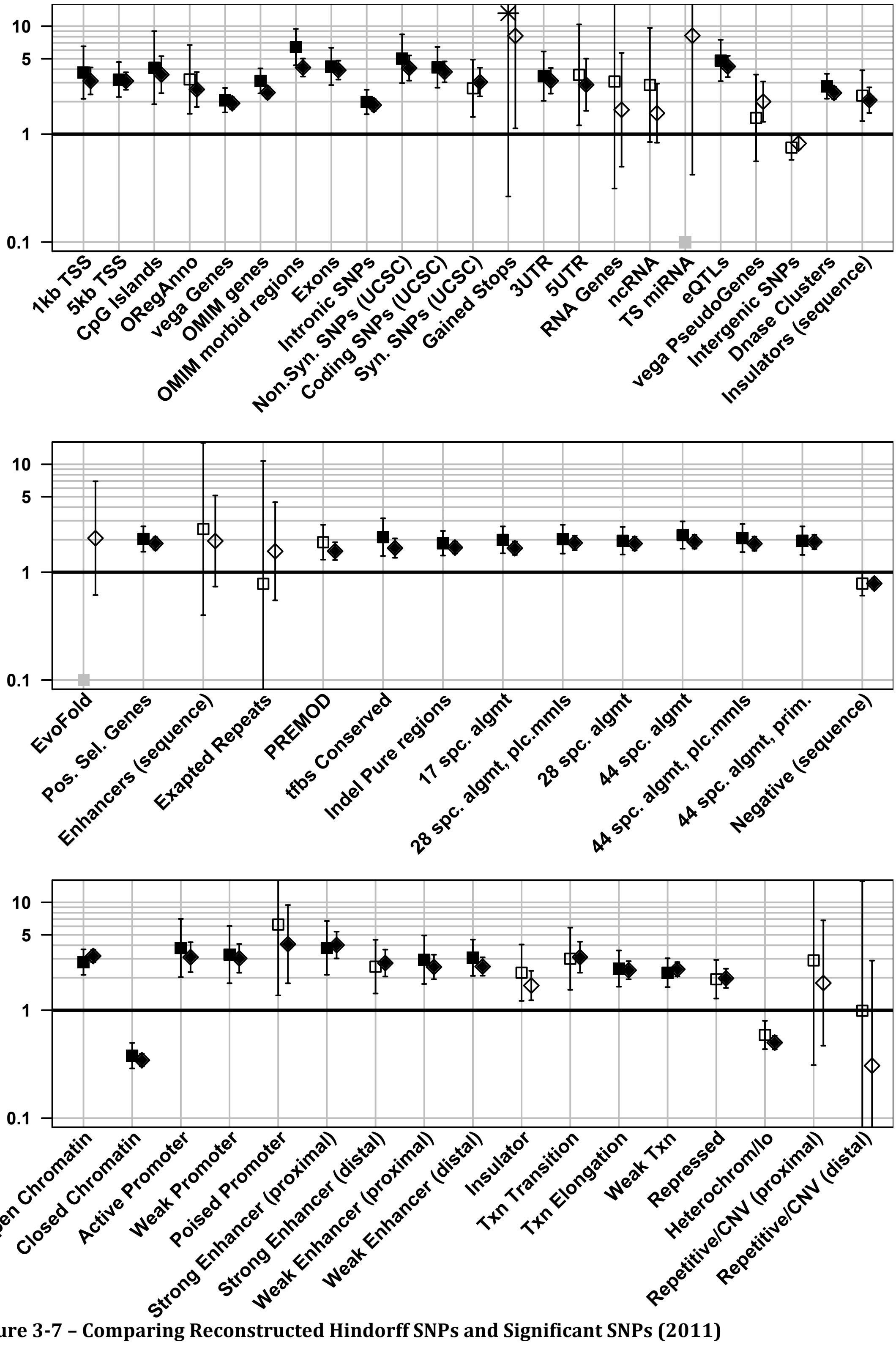


Figure 3-7 – Comparing Reconstructed Hindorff SNPs and Significant SNPs (2011)

The results of the sampling method for Reconstructed Hindorff SNPs (□/■) and Significant SNPs (2011) (◇/◆) are shown here. All results are displayed in odds ratios with 95% confidence intervals. Solid symbols are significant at the Bonferroni corrected significance threshold. Not available odds ratios (grey) and those with a value above the maximum of the graph are indicated (✱). Top: Genic and regulatory features. Middle: Conserved regions and evolutionary regions. Bottom: Chromatin states and histone modifications.

Table 3-1 – Sampling results for Reconstructed Hindorff SNPs

This table summarises the number of overlaps in the observed set (Real), the mean of the sample hits (Sample Mean), the calculated odds ratio and confidence interval (OR [LCI-HCI]) and the obtained *P*-value for each of the genomic annotations. Significant *P*-values in bold.

Annotation	Real	Sample Mean	*P*-value	OR [LCI - HCI]
1 Kb TSS	57	16.81	**1.35×10^{-06}**	3.72 [2.13-6.52]
5 Kb TSS	118	44.55	**3.09×10^{-10}**	3.20 [2.21-4.65]
CpG Islands	32	8.17	**1.33×10^{-04}**	4.13 [1.89-9.01]
ORegAnno	30	9.76	1.78×10^{-03}	3.22 [1.54-6.70]
vega Genes	260	176.53	**7.20×10^{-08}**	2.06 [1.59-2.68]
OMIM genes	283	154.13	**3.17×10^{-17}**	3.12 [2.38-4.07]
OMIM morbid regions	165	36.67	**2.43×10^{-25}**	6.41 [4.35-9.43]
Exons	122	35.90	**3.38×10^{-14}**	4.24 [2.85-6.32]
Intronic SNPs	286	206.77	**3.09×10^{-07}**	1.99 [1.53-2.58]
Non.Syn. SNPs (UCSC)	81	18.81	**2.76×10^{-11}**	5.00 [2.97-8.40]
Coding SNPs (UCSC)	101	29.03	**6.43×10^{-12}**	4.16 [2.69-6.43]
Syn. SNPs (UCSC)	38	15.05	1.62×10^{-03}	2.66 [1.44-4.90]
Gained Stops	6	0.16	3.08×10^{-02}	37.97 [0.26-5450.04]
3'UTR	61	19.51	**2.13×10^{-06}**	3.45 [2.03-5.84]
5'UTR	15	4.33	1.80×10^{-02}	3.55 [1.21-10.41]
RNA Genes	3	0.98	6.24×10^{-01}	3.07 [0.31-30.18]
ncRNA	10	3.55	1.76×10^{-01}	2.86 [0.85-9.65]
TS miRNA	0	0.13	1.00	0.00 [0.00-NA]
eQTLs	108	27.46	**1.92×10^{-14}**	4.81 [3.10-7.48]
vega PseudoGenes	11	7.84	6.44×10^{-01}	1.41 [0.56-3.56]
Intergenic SNPs	258	290.56	3.36×10^{-02}	0.75 [0.58-0.97]
DNase Clusters	274	157.75	**3.47×10^{-14}**	2.78 [2.13-3.62]
Insulators (sequence)	44	20.43	2.64×10^{-03}	2.27 [1.32-3.91]
Within miRNA	0	0.01	1.00	0.00 [0.00-NA]
Splice Sites	0	0.15	1.00	0.00 [0.00-NA]
Lost Stops	0	0.10	1.00	0.00 [0.00-NA]
Microsatellites	0	0.31	1.00	0.00 [0.00-NA]
EvoFold	0	1.17	1.00	0.00 [0.00-NA]
Pos. Sel. Genes	211	134.83	**3.52×10^{-07}**	2.03 [1.55-2.66]
Enhancers (sequence)	4	1.60	6.86×10^{-01}	2.51 [0.40-15.79]
Exapted Repeats	1	1.28	1.00	0.78 [0.06-10.71]
PREMOD	88	50.96	8.88×10^{-04}	1.90 [1.31-2.75]
tfbs Conserved	79	41.03	**2.71×10^{-04}**	2.11 [1.41-3.16]
Indels Pure regions	233	163.02	**4.79×10^{-06}**	1.85 [1.43-2.41]
17 spc. algmt	171	104.91	**2.94×10^{-06}**	1.99 [1.49-2.66]
28 spc. algmt plc.mmls	143	83.62	**8.94×10^{-06}**	2.02 [1.49-2.75]
28 spc. algmt	159	97.40	**7.17×10^{-06}**	1.96 [1.46-2.63]
44 spc. algmt	173	98.15	**8.42×10^{-08}**	2.21 [1.65-2.96]
44 spc. algmt plc.mmls	155	90.29	**1.76×10^{-06}**	2.07 [1.54-2.80]
44 spc. algmt prim.	145	87.21	**1.48×10^{-05}**	1.96 [1.45-2.66]
Negative (sequence)	225	253.42	7.74×10^{-02}	0.78 [0.61-1.01]
Open Chromatin	337	224.43	**5.72×10^{-14}**	2.79 [2.13-3.66]
Closed Chromatin	128	233.04	**2.20×10^{-12}**	0.38 [0.29-0.50]
Active Promoter	47	13.44	**6.33×10^{-06}**	3.78 [2.03-7.02]
Weak Promoter	44	14.38	**6.05×10^{-05}**	3.27 [1.78-6.02]
Poised Promoter	12	1.97	1.23×10^{-02}	6.23 [1.37-28.24]
Strong Enhancer (proximal)	55	15.92	**1.57×10^{-06}**	3.78 [2.13-6.71]
Strong Enhancer (distal)	42	17.52	1.92×10^{-03}	2.54 [1.43-4.50]
Weak Enhancer (proximal)	57	21.10	**2.62×10^{-05}**	2.94 [1.75-4.93]
Weak Enhancer (distal)	106	40.79	**5.55×10^{-09}**	3.07 [2.08-4.52]
Insulator	35	16.38	**8.93×10^{-03}**	2.23 [1.22-4.07]
Txn Transition	35	12.27	**8.11×10^{-04}**	3.00 [1.55-5.83]
Txn Elongation	93	43.29	**4.50×10^{-06}**	2.43 [1.65-3.58]
Weak Txn	150	81.75	**3.41×10^{-07}**	2.23 [1.64-3.03]
Repressed	71	39.53	2.30×10^{-03}	1.94 [1.28-2.93]
Heterochrom/low	331	376.14	**8.00×10^{-04}**	0.59 [0.44-0.80]
Repetitive/CNV (proximal)	3	1.04	6.24×10^{-01}	2.90 [0.31-27.05]
Repetitive/CNV (distal)	1	1.01	1.00	0.99 [0.06-15.77]

Table 3-2 – Sampling results for Significant SNPs (2011)

This table summarises the number of overlaps in the observed set (Real), the mean of the sample hits (Sample Mean), the calculated odds ratio and confidence interval (OR [LCI-HCI]) and the obtained *P*-value for each of the genomic annotations. Significant *P*-values in bold.

Annotation	Real	Sample Mean	*P*-value	OR [LCI - HCI]
1 Kb TSS	191	65.70	**3.46×10^{-16}**	3.11 [2.33-4.15]
5 Kb TSS	461	176.40	**1.07×10^{-35}**	3.10 [2.58-3.74]
CpG Islands	112	32.81	**1.30×10^{-11}**	3.56 [2.40-5.28]
ORegAnno	100	39.78	**2.58×10^{-07}**	2.59 [1.79-3.77]
vega Genes	1053	733.24	**1.59×10^{-24}**	1.93 [1.70-2.20]
OMIM genes	1067	644.30	**3.60×10^{-42}**	2.43 [2.13-2.76]
OMIM morbid regions	514	154.73	**2.31×10^{-54}**	4.14 [3.41-5.02]
Exons	452	139.10	**6.42×10^{-46}**	3.92 [3.20-4.79]
Intronic SNPs	1161	857.23	**4.31×10^{-22}**	1.86 [1.64-2.11]
Non.Syn. SNPs (UCSC)	266	72.27	**1.51×10^{-29}**	4.10 [3.13-5.36]
Coding SNPs (UCSC)	364	111.50	**2.55×10^{-36}**	3.78 [3.02-4.72]
Syn. SNPs (UCSC)	166	57.89	**6.25×10^{-14}**	3.04 [2.24-4.13]
Gained Stops	9	1.11	2.13×10^{-02}	8.14 [1.13-58.53]
3'UTR	212	73.29	**5.03×10^{-18}**	3.12 [2.37-4.10]
5'UTR	48	16.99	**1.34×10^{-04}**	2.87 [1.64-5.01]
RNA Genes	7	4.17	5.48×10^{-01}	1.68 [0.50-5.66]
ncRNA	25	16.05	2.09×10^{-01}	1.56 [0.83-2.94]
TS miRNA	4	0.49	1.25×10^{-01}	8.18 [0.42-158.95]
eQTLs	378	104.46	**2.84×10^{-42}**	4.24 [3.38-5.32]
vega PseudoGenes	63	31.99	1.69×10^{-03}	2.00 [1.30-3.08]
Intergenic SNPs	1114	1208.31	2.64×10^{-03}	0.82 [0.72-0.93]
DNase Clusters	1057	638.14	**1.71×10^{-41}**	2.41 [2.12-2.75]
Insulators (sequence)	163	82.35	**1.03×10^{-07}**	2.07 [1.57-2.72]
Within miRNA	0	0.04	1.00	0.00 [0.00-NA]
Splice Sites	0	0.78	1.00	0.00 [0.00-NA]
Lost Stops	0	0.39	1.00	0.00 [0.00-NA]
Microsatellites	0	1.31	1.00	0.00 [0.00-NA]
EvoFold	8	3.88	3.87×10^{-01}	2.07 [0.61-6.96]
Pos. Sel. Genes	831	557.46	**7.46×10^{-20}**	1.85 [1.62-2.11]
Enhancers (sequence)	12	6.19	2.37×10^{-01}	1.94 [0.74-5.14]
Exapted Repeats	9	5.77	6.07×10^{-01}	1.56 [0.55-4.45]
PREMOD	303	204.68	**3.80×10^{-06}**	1.57 [1.30-1.89]
tfbs Conserved	258	162.70	**1.15×10^{-06}**	1.67 [1.36-2.06]
Indels Pure regions	905	657.99	**1.08×10^{-15}**	1.69 [1.49-1.93]
17 spc. algmt	614	420.08	**2.54×10^{-12}**	1.67 [1.45-1.93]
28 spc. algmt plc.mmls	545	334.56	**1.04×10^{-15}**	1.87 [1.60-2.18]
28 spc. algmt	618	391.67	**1.83×10^{-16}**	1.84 [1.59-2.13]
44 spc. algmt	641	396.60	**1.19×10^{-18}**	1.91 [1.65-2.21]
44 spc. algmt plc.mmls	578	363.10	**1.06×10^{-15}**	1.84 [1.58-2.13]
44 spc. algmt prim.	573	348.80	**3.75×10^{-17}**	1.91 [1.64-2.22]
Negative (sequence)	932	1051.01	**1.73×10^{-04}**	0.79 [0.69-0.89]
Open Chromatin	1458	927.77	**2.26×10^{-67}**	3.18 [2.78-3.64]
Closed Chromatin	502	981.98	**1.54×10^{-56}**	0.34 [0.30-0.39]
Active Promoter	155	52.75	**2.15×10^{-13}**	3.10 [2.26-4.27]
Weak Promoter	164	57.36	**7.36×10^{-14}**	3.03 [2.23-4.12]
Poised Promoter	28	6.91	**4.81×10^{-04}**	4.10 [1.78-9.44]
Strong Enhancer (proximal)	231	62.99	**2.27×10^{-25}**	4.02 [3.02-5.35]
Strong Enhancer (distal)	176	68.18	**6.76×10^{-13}**	2.74 [2.05-3.65]
Weak Enhancer (proximal)	201	85.04	**8.52×10^{-13}**	2.52 [1.94-3.27]
Weak Enhancer (distal)	367	162.68	**9.54×10^{-22}**	2.54 [2.09-3.09]
Insulator	108	65.21	1.04×10^{-03}	1.69 [1.24-2.32]
Txn Transition	144	48.84	**1.46×10^{-12}**	3.10 [2.23-4.32]
Txn Elongation	360	171.23	**8.63×10^{-19}**	2.35 [1.93-2.85]
Weak Txn	631	323.36	**1.44×10^{-30}**	2.40 [2.06-2.79]
Repressed	289	157.59	**5.01×10^{-11}**	1.98 [1.61-2.43]
Heterochrom/low	1314	1576.96	**3.45×10^{-21}**	0.50 [0.43-0.58]
Repetitive/CNV (proximal)	6	3.36	5.07×10^{-01}	1.79 [0.47-6.81]
Repetitive/CNV (distal)	1	3.26	6.25×10^{-01}	0.31 [0.03-2.88]

3.4 Discussion

In this chapter, we represent the results of our attempt at replicating the sampling method published by Hindorff *et al.* [50] as closely as possible. We then expanded our analyses to incorporate more genomic annotations and more trait-associated SNPs.

3.4.1 Comparison of Hindorff results with Reconstructed Hindorff SNPs

The results of the Reconstructed Hindorff SNP set obtained in 14 annotations common to Hindorff *et al.* and this analysis were compared to the results from the original study [50]. In our analysis we obtained nine odds ratios of enrichment, which were significant at the significance threshold corrected for 14 annotations. The nine annotations with significant odds ratios contain two annotations, which replicated two of the significant results by Hindorff *et al.* [50]. These two annotations were the non-synonymous SNPs annotation and the 1 Kb upstream region of TSS annotation. The intergenic SNPs are depleted in our analysis, although the *P*-value was not significant after correcting for multiple testing.

Hindorff *et al.* [50] investigated the underlying genomic annotations of 465 significantly trait-associated SNPs and presented the results after taking account of hitchhiking effects caused by SNPs in LD with possibly deleterious non-synonymous variants. The non-synonymous variants were the most enriched signal and the authors tested if the odds ratios in the other genomic annotations were driven by non-synonymous variants that were in LD with the trait-associated variants. However, correcting the rest of the results for the top result, while not correcting the top result with the second highest enrichment signal is close to biasing the results in favour of the strongest signal and the difference between the corrected and not-corrected odds ratios are not significant [50]. The correction furthermore did not change the odds ratios significantly. For these reasons, we did not correct for hitchhiking effects in our analyses. Thus, all our results are compared to the non-corrected results by Hindorff *et al.* [50]. The differences between the published results and our analysis with the Reconstructed Hindorff SNPs could be due to author specific definitions of the analysed genomic features, where the algorithms and methods were not exactly matched. For example, different algorithms identify different targets for the miRNA binding sites and the authors did not specify which algorithms were chosen. Odds ratios of genomic annotations for which the definition was clear, such as non-synonymous SNPs, were very comparable.

Recall, the precise reconstruction of the SNP set used by Hindorff *et al.* was not entirely possible. Hindorff *et al.* corrected for the number of trait-associated SNPs in LD blocks, which is something we did not do. Overrepresentation of SNPs in LD blocks would occur if more than one SNP in a similar genomic region were found to be significantly associated to at least one trait. This overrepresentation of particular genomic regions could highlight important biological areas. We therefore did not reduce the number of trait-associated SNPs based on their location, other than removing the SNPs on non-assigned chromosomes or on the Y-chromosome.

3.4.2 Expansion of analysis to more annotations and trait-associated variants

Significant SNPs (2011), a dataset containing 1,909 SNPs, was analysed presenting a more recent set of GWAS hits than the set of SNPs from 2009. The results of the sampling method for Significant SNPs (2011) agreed with what may have been expected. The study gained statistical power by investigating more trait-associated SNPs, as measured by more results that are significant and a decreased confidence interval width in all genomic annotations, where an odds ratio was observed in the smaller dataset. Moreover, odds ratios were obtained in more annotations. For instance, we obtained odds ratios in the TS miRNA annotation, which did not obtain an odds ratio in the Reconstructed Hindorff SNPs. However, in this case, the corresponding confidence interval width was very large with a value of 158.53, as only four of the trait-associated SNPs overlapped with the annotation (see Table 3-2). On the other hand, all the significant results obtained in the smaller dataset were also observed in the larger set. Significant SNPs (2011) obtained significant enrichment results in 12 of 14 annotations analysed by Hindorff *et al.* Significant SNPs (2011) were additionally significantly different to the published results in the intergenic SNPs.

3.4.3 Reconstructed Hindorff SNPs *vs.* Significant SNPs (2011)

The odds ratios obtained for Significant SNPs (2011) were compared to the odds ratios obtained for Reconstructed Hindorff SNPs. Significant SNPs (2011) contained 1,909 SNPs, which included the 478 SNPs from the Reconstructed Hindorff SNPs. It was therefore expected that Significant SNPs (2011) obtained more significant odds ratios, as this dataset had more statistical power as more SNPs were analysed. This was observed in 10 genomic annotations, where the odds ratios for Significant SNPs (2011) were significant, but the odds ratios for the Reconstructed Hindorff SNPs were not. The Bonferroni corrected threshold changed from 3.57×10^{-3} to 8.62×10^{-5} with the inclusion of more genomic annotations. The odds ratios for seven of these 10 annotations were more moderate for the larger dataset, but significant due to a decrease of the 95% confidence interval width. Synonymous SNPs, transcriptional transition, and strong enhancers (distal) obtained higher odds ratios for Significant SNPs (2011). It is difficult to distinguish the effects of transcriptional elongation or synonymous SNPs from effects caused by genic regions of the genome, as one would define the other.

The sampling strategy has been reproduced to an extent that we are confident to use the results in comparison with a different method. We therefore continued our investigation into the genomic features underlying the trait-associated SNPs with a method developed by us for this purpose: chromosome-wide circular permutations.

4 CIRCULAR PERMUTATIONS

4.1 Introduction

The genomic features, such as the mentioned annotations, are not spread uniformly throughout the chromosomes of the human genome, but rather occur in clusters. This trend holds for the smallest changes in DNA to large genic sequences. For instance, entire chromosomes are packaged into different chromatin territories affecting gene transcription, DNA repair and replication [134, 135]. The same trend is reflected in the clustering of genes [134], functional elements [74] and transcription factor binding sites, which cluster around the transcription start sites (TSS) of genes [136]. The distribution and number of SNPs also vary between chromosomes and chromatin states. For example, heterochromatin contains a higher density of SNPs relative to more open regions [137], while conversely, causal SNPs are expected to appear more frequently in the latter. The uneven distribution of trait-associated SNPs [18] is an expected consequence of the non-random distribution of the other functional elements, and comprises additional information which can be used in further analyses.

The sampling method, introduced earlier, was designed to take certain types of known biases into account. The sampled data presented in the previous chapter were the variants present on genotyping arrays. As discussed in the introduction of the thesis, the variants included on the genotyping arrays were chosen according to different criteria. These criteria included the coverage of the genome through the selection of variants that gave information on many other SNPs ('tagging SNPs') or the overlap of variants with genes, and would therefore be biased in their representation of chromosomes and genic material. These biases were considered when the contributions of the different genotyping arrays were recreated in the drawn samples.

However, while the sampling method attempts to take the genotyping array bias into account, it also assumes that the SNPs are uniformly distributed through the genome. The subsets were chosen regardless of their genomic locations, which resulted in biased samples with regards to chromosomal representation. Some samples do not necessarily cover the entire genome, as chromosomes are over-/under-represented and could in extreme cases be completely missed out. We developed a method based on permutations that takes this representational bias and the non-uniform distribution of functional elements overlapping with SNPs into account. Each chromosome is analysed on its own returning the number of overlaps of SNPs and annotations for each chromosome, which are then added to obtain a genome wide result. This method therefore takes the chromosomal bias into account and preserves the local clustering of the real data. The chosen number of performed permutations (n = 20,000) allows the calculation of empirical P-values and confidence intervals. A fuller description of the method can be found on page 68 of this thesis.

4.2 Materials and method

4.2.1 Data structure

The permutations were restricted to chromosomes, which were analysed separately. Files representing all autosomes and the X chromosome were compiled, each listing the available SNPs according to their position on the chromosome. Each SNP row also contained information on any overlaps of a genomic annotation with either the SNP or its LD partners (see Chapter section 2.1.10 for definition). The lists contained a total of 3,840,944 variants incorporating all SNPs represented on 11 genotyping arrays and were HapMap CEU II SNPs. These lists were analysed individually and the results of the chromosomes per permutation were summed giving a genome wide result.

4.2.2 Circular permutation strategy

The structure of the data frames we had established attempted to recreate and preserve the internal structure of the genome. Our goal in particular was to maintain the number of SNPs between a SNP pair as a proxy of the non-random and non-uniform clustered distribution of SNPs in the genome. The permutation approach was designed to preserve this internal structure of the genome as each permutation maintained the clustered distribution of SNPs as trait-association status was shifted or permuted along the chromosome, while keeping the number of SNPs between trait-associated variants constant. The information for trait-associated GWAS hits was downloaded from the NHGRI webpage as previously stated (see Chapter section 2.1.3).

The permutations were performed per chromosome by shifting, or permuting, the trait-associated status along the chromosome, while the structure according to the SNPs on the chromosomes remained the same throughout the analyses. For each permutation a randomly generated number, drawn from a uniform distribution between one and the number of SNPs per analysed chromosome, was used to shift the status of trait-association along the chromosome. The shift was performed circularly, so that any variables exceeding the number of SNPs on the chromosome were pushed to the beginning of the chromosome. The trait-association status was therefore re-assigned to different variants, whose previous trait-association status was no longer relevant. The LD partners were then re-defined and overlaps of the newly labelled trait-association variants with annotations were counted as in the original trait-associated variant set. The permutations therefore sampled the chromosome in a controlled manner by maintaining the distance of SNPs as measured by number of SNPs, rather than nucleotides. This created a background distribution of SNPs and their LD partners that were as biased in the number of SNPs on a given chromosome and their distribution on that chromosome. Regions that are difficult to genotype would be less represented in both the real trait-associated variant set and the background set.

Figure 4-1 shows a cartoon-like illustration of the circularised chromosome in four scenarios: The observed data and three sequential permutations. This produced a population of 20,000 chromosomes, circularly permuted relative to the original chromosome, containing the same number of trait-associated SNPs and preserved the degree of genomic clustering observed in the original SNP datasets. The chromosome results were summed to give a genome-wide result.

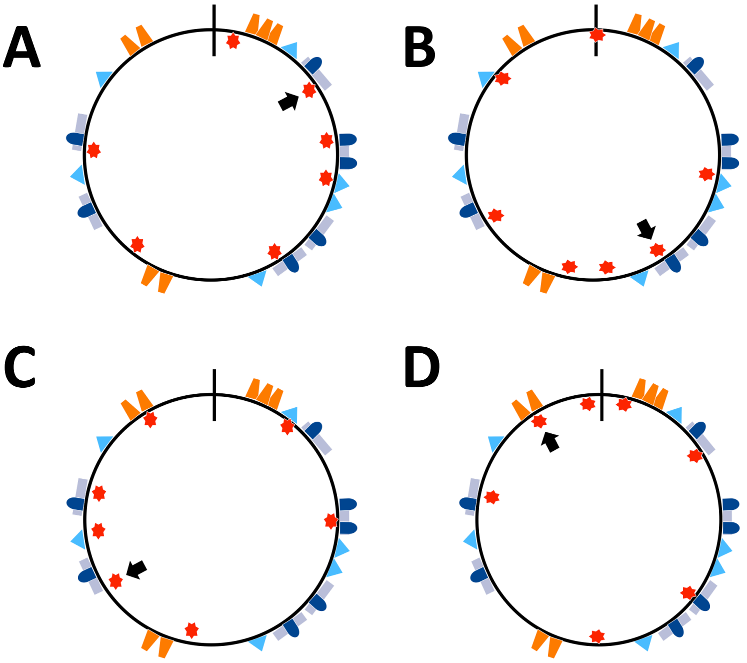


Figure 4-1 – Diagram of permutations on virtually circularised chromosome
The permutations were performed on a virtually circularised chromosome (black circle). The start and end of the chromosome are equivalent and are depicted by the vertical line at the top of the circle. The coloured symbols (orange cones, light-blue triangles, dark-blue arches and grey rectangles) represent different genomic features adopting a non-uniform distribution and showing a distinct clustering of the genomic annotations. The red stars depict the trait-associated SNPs. Non-associated SNPs are left off the diagram for clarity. The black arrow highlights the same trait-associated SNP in each chromosome. A) Observed data and starting position. B) First permutation of trait-associated SNPs by 90˚ in clockwise direction from the starting position. C) Second permutation of trait-associated SNPs by 180˚ in clockwise direction from the starting position. D) Third permutation of trait-associated SNPs by 270˚ in clockwise direction from the starting position. In practise the degree of rotation of SNPs versus annotation is randomly chosen.

4.2.3 Odds ratios and confidence intervals

The number of overlaps for the real trait-associated SNPs (observed) for each annotation and the mean number of overlaps of permuted SNPs (expected) were used to calculate odds ratios of enrichment or depletion (see equation below). The permuted overlaps were ranked according to the number of overlaps creating a discrete uniform distribution used to calculate the 95% confidence intervals. This ranking of the permutation overlaps allowed the calculation of the 2.5th and 97.5th percentile values, which were used to define the 95% confidence intervals of the data were observed. Since a two-sided hypothesis test was applied, *i.e.,* were there more or fewer trait-associated SNPs than expected by chance, the 2.5th (500th ranked position) and 97.5th (19,500th ranked position) percentile values represent the 95% boundaries. The calculation for the confidence intervals was the same as for the odds ratios (see paragraph 2.1.11), except that the 500th and 19,500th value replaced the mean number of overlaps of the expected data. The mean number of non-overlaps was defined as the difference of the total and the mean number of the overlaps. This gives a thorough representation of the underlying distribution of the overlaps between SNPs and genomic features, which can be easily displayed on the graphs. These empirically derived confidence intervals were not symmetric relative to the odds ratios on the log scale, in contrast to the confidence intervals of the sampling method. The latter were calculated as theoretical values using the standard errors of the odds ratios, which meant that they would always be symmetric relative to the odds ratios on a log scale.

If the divisor of the odds ratios was zero, the values for the confidence intervals or the odds ratios are undefined. In our case, this occurred if either the real number of non-overlaps was zero or the permuted number of overlaps was zero. As it was more unlikely that a mean of permuted hits was zero, the odds ratios were defined more often than confidence interval values, which were based on only one permutation value.

Confidence intervals, which were set to infinity by R, were artificially delimited to 30, a value beyond the range of the graphs to enable the plotting of these confidence intervals. The eight annotations, for which this occurred, are sparsely distributed (see Table 2-1) in the genome and were removed from any confidence interval width calculations (within miRNA, TS miRNA, evofold, splice sites, gained stops, lost stops, microsatellites and repetitive/CNV (distal)).

4.2.4 Calculating *P*-values

The *P*-value of the odds ratios for the permutations was the ratio of the number of permuted datasets that were equal to or more extreme than the observations in the observed trait-associated SNP set, and the total number of permutations. The lower bound of this empirically defined *P*-value was therefore 5×10^{-5} when only one in 20,000 permutations was equal to or more extreme than the number of overlaps in the observed data. If none of the permutations were equal to or more extreme than the observed data the *P*-value was $< 5 \times 10^{-5}$. The *P*-values are significant if they passed the Bonferroni corrected significance threshold corrected for testing 58 genomic annotations (8.60×10^{-4}), which is equivalent to 17 permutations being more extreme or equal to the observed data. Since the tested annotations were not independent from each other, this threshold is likely to be very stringent.

4.3 Results

4.3.1 Comparison of permutations and sampling in Significant SNPs (2011)

Figure 4-2 shows the results obtained by both the permutation and the sampling method for Significant SNPs (2011) with 1,909 significantly trait-associated SNPs. The three subcategories of the genomic features are shown in the three panels of the figure. The genic and regulatory features are presented in the top panel, the conserved and evolutionary figures are shown in the middle panel and the histone modifications and chromatin states are shown in the bottom panel. The majority of the genomic annotations were significantly enriched for trait-associated SNPs, while two annotations showed significant depletion (closed chromatin and heterochrom/lo). The large number of enrichment signals was also observed by the sampling method. Few differences existed between the results from the permutation and the sampling approaches, which we comment shortly. The obtained odds ratios correlated strongly with a correlation of 0.98 and a significant *P*-value ($P = 8.89 \times 10^{-51}$). The observed differences between the methods were significant at *P*-value ≤ 0.05.

There were three genomic annotations, where the odds ratio obtained by the permutation method was not significant, but was significant for the sampling method. The insulators (bottom panel) were significantly enriched for permutations but not the sampling method. The methods agreed on the remaining genomic annotations, 10 of which were not significantly enriched for or depleted of trait-associated SNPs. The two methods had very similar odds ratio in the OMIM morbid regions. These regions were considered a positive control, since they were defined as disease-associated regions by a number of different studies [99]. The annotation serving as a negative control, named negative (sequence), originated from the intergenic SNPs from which any overlapping conserved or genic regions were excluded. For this annotation, only the sampling method showed significant depletion, while the odds ratio of depletion obtained by the permutation method was not significant.

The lack of significant depletion in the negative annotation could be due to a number of reasons. One of these reasons was the difference in calculation for the odds ratios and *P*-values between the two methods. The odds ratios are almost identical between the methods (sampling = 0.79 [0.69-0.89]; permutations = 0.81 [0.71-0.94]) resulting from the number of overlaps in the expected dataset (1051 in sampling *vs.* 1029 in permutations). When the odds ratio and confidence intervals and *P*-value were calculated using the same formula as used for the sampling method the result was 0.82 [0.72-0.93] with a *P*-value of 1.69×10^{-3}. This *P*-value is not significant after accounting for multiple testing. Since the number of samples drawn and permutations performed differs substantially it is possible that the permutations capture more of the real distribution. It is also possible that the negative set was not defined properly or that an informative and as of yet unidentified annotation still overlaps with the negative annotation, but that would have impacted both methods equally.

The largest difference between the two methods is in confidence interval width, as shown above. The eight genomic annotations excluded from calculations for the mean odds ratio and confidence interval widths were within miRNA, TS miRNA, evofold, splice sites, gained stops, lost stops, microsatellites and repetitive/CNV (distal). The permutations obtained larger confidence intervals with an average of 2.06 in those annotations, where both methods obtained a defined confidence interval *i.e.,* one that was not set to infinity. The confidence interval width of the sampling method in the same annotations was 1.63. This was an indication that the permutation approach was generally more conservative, as expected, since the permutation approach was designed to take an appropriate account of non-random distributions of annotations and SNP locations. The eight genomic annotations, which obtained confidence intervals with values of infinity (denoted infinity in the tables), were sparsely distributed throughout the genome. A value of infinity was returned when the dividend of the odds ratio equation was too large when compared with the divisor (see paragraph "Scoring LD blocks and definition of odds ratios"). Among these genomic annotations were the four rare annotations mentioned previously (microsatellites, within microRNA, splice sites and lost stop codons). The other four genomic annotations were TS miRNA binding sites, gained stop codons, evofold, and distal repetitive/CNV elements. The coverage of these eight genomic features ranged between 0.00-0.21% annotated SNPs. The confidence intervals were very wide, illustrating the uncertainty of the results, which were based on a small number of coinciding SNPs. The number of overlaps for the Significant SNPs (2011) set and its permutations, the resulting odds ratios with confidence intervals and the *P*-value of significance are shown in Table 4-1.

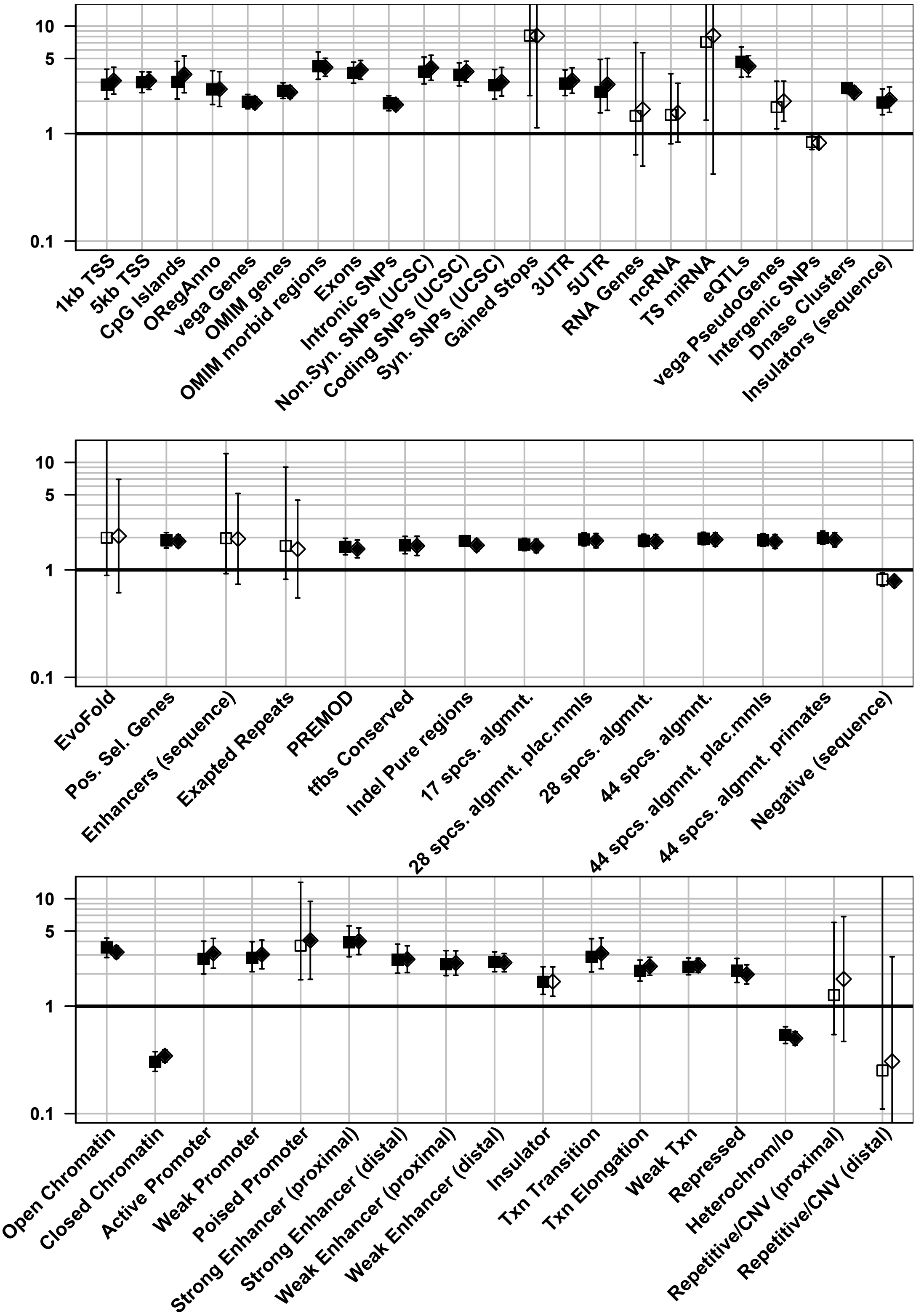


Figure 4-2 – Comparison of sampling *vs.* permutation results of Significant SNPs (2011)
There are few differences between the results for Significant SNPs (2011) (*n* = 1,909) with the permutation method (□) when compared with the sampling method (◇). All *P*-values are corrected for multiple testing for the analysed genomic annotations. Solid symbols indicate significance at the multiple-testing corrected threshold. Top: Genic and regulatory regions. Middle: Conserved regions and evolutionary signatures. Bottom: Chromatin states and histone modifications.

Table 4-1 – Permutation results for Significant SNPs (2011) at $r^2 > 0.9$ threshold

This table summarises the results for Significant SNPs (2011) (*n* = 1,909). The number of overlaps in the observed set (Real), the mean of the permuted hits (Permutation Mean), the calculated odds ratio and confidence interval (OR [LCI-HCI]) and the obtained *P*-value for each of the genomic annotations are shown below. Significant *P*-values in bold.

Annotation	Real	Permutation Mean	OR [LCI - HCI]	P-value
1 Kb TSS	191	71.53	2.86 [2.10-3.97]	**$< 5.00 \times 10^{-05}$**
5 Kb TSS	461	182.70	3.01 [2.41-3.76]	**$< 5.00 \times 10^{-05}$**
CpG Islands	112	38.42	3.03 [2.10-4.70]	**$< 5.00 \times 10^{-05}$**
ORegAnno	100	40.01	2.58 [1.86-3.85]	**$< 5.00 \times 10^{-05}$**
vega Genes	1053	731.82	1.98 [1.70-2.31]	**$< 5.00 \times 10^{-05}$**
OMIM genes	1067	640.94	2.51 [2.13-2.97]	**$< 5.00 \times 10^{-05}$**
OMIM morbid regions	514	152.78	4.24 [3.20-5.75]	**$< 5.00 \times 10^{-05}$**
Exons	452	148.72	3.67 [2.94-4.62]	**$< 5.00 \times 10^{-05}$**
Intronic SNPs	1161	854.46	1.92 [1.63-2.25]	**$< 5.00 \times 10^{-05}$**
Non.Syn. SNPs (UCSC)	266	78.45	3.78 [2.90-5.17]	**$< 5.00 \times 10^{-05}$**
Coding SNPs (UCSC)	364	119.72	3.52 [2.78-4.55]	**$< 5.00 \times 10^{-05}$**
Syn. SNPs (UCSC)	166	62.54	2.81 [2.10-3.94]	**$< 5.00 \times 10^{-05}$**
Gained Stops	9	1.11	8.17 [2.26-Infinity]	9.50×10^{-04}
3'UTR	212	78.20	2.92 [2.26-3.92]	**$< 5.00 \times 10^{-05}$**
5'UTR	48	19.92	2.45 [1.56-4.90]	**5.00×10^{-05}**
RNA Genes	7	4.80	1.46 [0.64-7.02]	2.25×10^{-01}
ncRNA	25	16.81	1.49 [0.80-3.61]	1.08×10^{-01}
TS miRNA	4	0.56	7.13 [1.33-Infinity]	9.95×10^{-03}
eQTLs	378	95.90	4.67 [3.35-6.39]	**$< 5.00 \times 10^{-05}$**
vega PseudoGenes	63	36.22	1.76 [1.11-3.07]	1.16×10^{-02}
Intergenic SNPs	1114	1196.85	0.83 [0.71-0.98]	1.11×10^{-02}
DNase Clusters	1057	609.66	2.64 [2.36-2.97]	**$< 5.00 \times 10^{-05}$**
Insulators (sequence)	163	87.60	1.94 [1.50-2.61]	**$< 5.00 \times 10^{-05}$**
Within miRNA	0	0.03	0.00 [0.00-NA]	**$< 5.00 \times 10^{-05}$**
Splice Sites	0	0.90	0.00 [0.00-NA]	**$< 5.00 \times 10^{-05}$**
Lost Stops	0	0.37	0.00 [0.00-NA]	**$< 5.00 \times 10^{-05}$**
Microsatellites	0	1.51	0.00 [0.00-NA]	**$< 5.00 \times 10^{-05}$**
EvoFold	8	4.03	1.99 [0.89-Infinity]	7.47×10^{-02}
Pos. Sel. Genes	831	555.76	1.88 [1.60-2.23]	**$< 5.00 \times 10^{-05}$**
Enhancers (sequence)	12	6.10	1.97 [0.92-12.07]	5.17×10^{-02}
Exapted Repeats	9	5.39	1.67 [0.82-9.04]	1.23×10^{-01}
PREMOD	303	197.22	1.64 [1.38-1.97]	**$< 5.00 \times 10^{-05}$**
tfbs Conserved	258	161.38	1.69 [1.41-2.05]	**$< 5.00 \times 10^{-05}$**
Indels Pure regions	905	624.32	1.85 [1.66-2.08]	**$< 5.00 \times 10^{-05}$**
17 spec. algmt	614	413.11	1.72 [1.51-1.97]	**$< 5.00 \times 10^{-05}$**
28 spec. algmt plc.mmls	545	328.96	1.92 [1.67-2.22]	**$< 5.00 \times 10^{-05}$**
28 spec. algmt	618	388.80	1.87 [1.65-2.15]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt	641	392.17	1.96 [1.72-2.24]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt plc.mmls	578	357.09	1.89 [1.65-2.17]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt prim.	573	339.39	1.98 [1.73-2.30]	**$< 5.00 \times 10^{-05}$**
Negative (sequence)	932	1029.98	0.81 [0.71-0.94]	1.95×10^{-03}
Open Chromatin	1458	915.13	3.51 [2.84-4.30]	**$< 5.00 \times 10^{-05}$**
Closed Chromatin	502	1030.96	0.30 [0.25-0.38]	**$< 5.00 \times 10^{-05}$**
Active Promoter	155	59.18	2.76 [1.99-4.03]	**$< 5.00 \times 10^{-05}$**
Weak Promoter	164	61.53	2.82 [2.09-3.98]	**$< 5.00 \times 10^{-05}$**
Poised Promoter	28	7.74	3.65 [1.76-14.19]	9.50×10^{-04}
Strong Enhancer (proximal)	231	64.59	3.93 [2.88-5.58]	**$< 5.00 \times 10^{-05}$**
Strong Enhancer (distal)	176	69.03	2.71 [2.03-3.78]	**$< 5.00 \times 10^{-05}$**
Weak Enhancer (proximal)	201	86.97	2.47 [1.92-3.29]	**$< 5.00 \times 10^{-05}$**
Weak Enhancer (distal)	367	161.86	2.57 [2.09-3.20]	**$< 5.00 \times 10^{-05}$**
Insulator	108	65.59	1.69 [1.29-2.32]	**1.50×10^{-04}**
Txn Transition	144	52.56	2.88 [2.08-4.24]	**$< 5.00 \times 10^{-05}$**
Txn Elongation	360	187.49	2.13 [1.71-2.69]	**$< 5.00 \times 10^{-05}$**
Weak Txn	631	333.95	2.33 [1.96-2.80]	**$< 5.00 \times 10^{-05}$**
Repressed	289	147.11	2.14 [1.66-2.78]	**$< 5.00 \times 10^{-05}$**
Heterochrom/lo	1314	1533.73	0.54 [0.45-0.64]	**$< 5.00 \times 10^{-05}$**
Repetitive/CNV (proximal)	6	4.73	1.27 [0.54-6.02]	6.65×10^{-01}
Repetitive/CNV (distal)	1	3.96	0.25 [0.11-Infinity]	3.02×10^{-02}

4.3.2 Comparison of different significant thresholds

There has been substantial interest in the nature of GWAS variants which showed ‘suggestive’ levels of significance (*i.e.,* SNPs with *P*-values = 5 × 10^{-5} - 5 × 10^{-8}), as they are believed to contain many true positives with modest effect size [138]. If that were correct, we would expect similarities in the functional enrichment patterns of these variants and the significantly associated variants. Our method can be used to test this hypothesis and the analysis of Suggestive SNPs (2011) was in agreement with this hypothesis. The majority of the genomic annotations are not significantly enriched or depleted of Suggestive SNPs (2011), which is unlike the Significant SNPs (2011). Moreover, those genomic annotations, which were significant for suggestively associated SNPs, obtained odds ratios that were less extreme than the odds ratios for the Significant SNPs (2011) in the same annotations.

Figure 4-3 shows the 14 enrichment and depletion results, where the suggestively associated SNPs obtained a significant odds ratio. The trends were similar to those observed for genome-wide significant SNPs, but with more moderate odds ratios. Of these 14 annotations, nine were from the genic annotation category, and five from the chromatin states. None of the annotations from the conserved annotation category obtained significant odds ratios for Suggestive SNPs (2011).

This lack of significant signal can also be observed Figure 4-4, which shows a comparison of all annotations. This figure shows a comparison of all of the genomic annotations and suggested that the general trend of all genomic annotations was similar to those observed for genome-wide significant SNPs. There were a few annotations, which were exceptions to the trend, where the suggestive SNPs had more extreme odds ratios. However, in those annotations the enrichment was not statistically significant for either the significant or the suggestive SNPs. The number of overlaps for Suggestive SNPs (2011), its permutation overlaps, the resulting odds ratio with confidence intervals and the *P*-value of significance for the odds ratio are shown in Table 4-2. A graph showing the correlation of the odds ratios obtained in the Suggestive SNPs (2011) and Significant SNPs (2011) was published in Kindt *et al.* [139]. The square of the correlation coefficient, r^2, of the regression line of the odds ratio obtained for the Suggestive SNPs (2011) onto the odds ratios obtained for the Significant SNPs (2011) was 0.81 with a *P*-value of 4.42×10^{-20}.

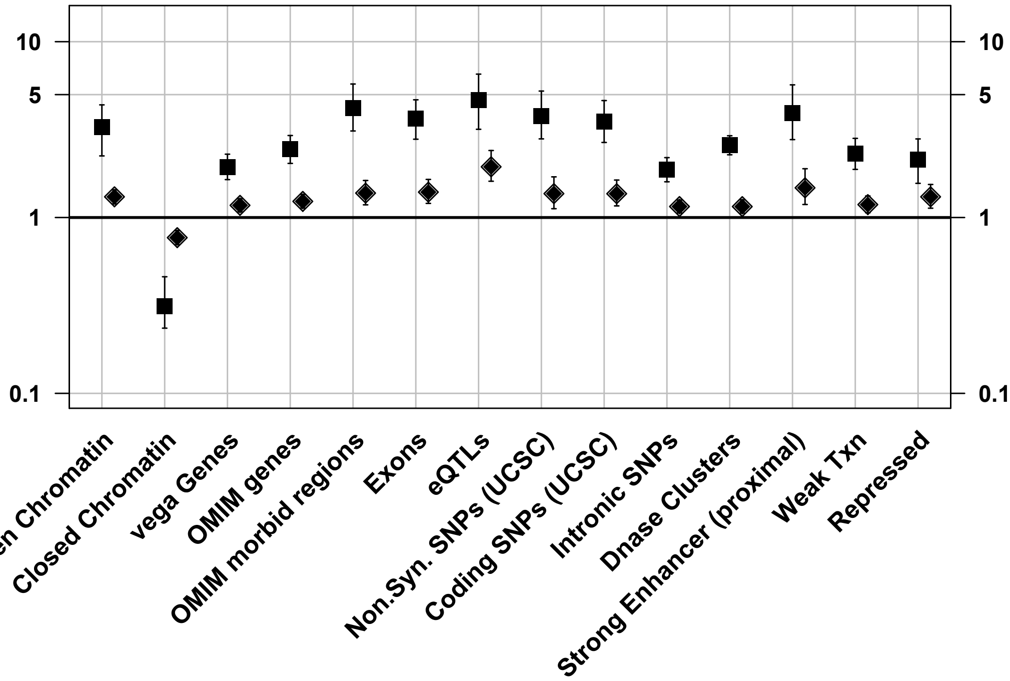


Figure 4-3 – Comparison of significant and suggestive variants in a subset of genomic annotations
The significantly trait-associated variants (Significant SNPs (2011); *n* = 1,909; □) and the variants with a suggestive *P*-value of association (5×10^{-5} > *P*-value > 5×10^{-8}; Suggestive SNPs (2011); *n* = 2,410); ◇) in the significant annotations for the Suggestive SNPs (2011) show similar trends. Suggestive SNPs (2011) show the same trend in enrichment/depletion but with more moderate odds ratios than Significant SNPs (2011).

The mean of the odds ratios for Significant SNPs (2011) in 50 of the analysed genomic annotations was 2.33. The mean of the odds ratios of the same genomic annotations in Suggestive SNPs (2011) was reduced to 1.13. The confidence intervals were also much narrower for Suggestive SNPs (2011) with a width of 0.56, while the confidence intervals were 2.06 for all 50 genomic annotations in Significant SNPs (2011), presumably because there it consisted of a larger number of SNPs. The confidence intervals were calculated without the eight genomic annotations previously mentioned (microsatellites, within microRNA, splice sites and lost stop codons, TS miRNA binding sites, gained stop codons, evofold, and distal repetitive/CNV elements).

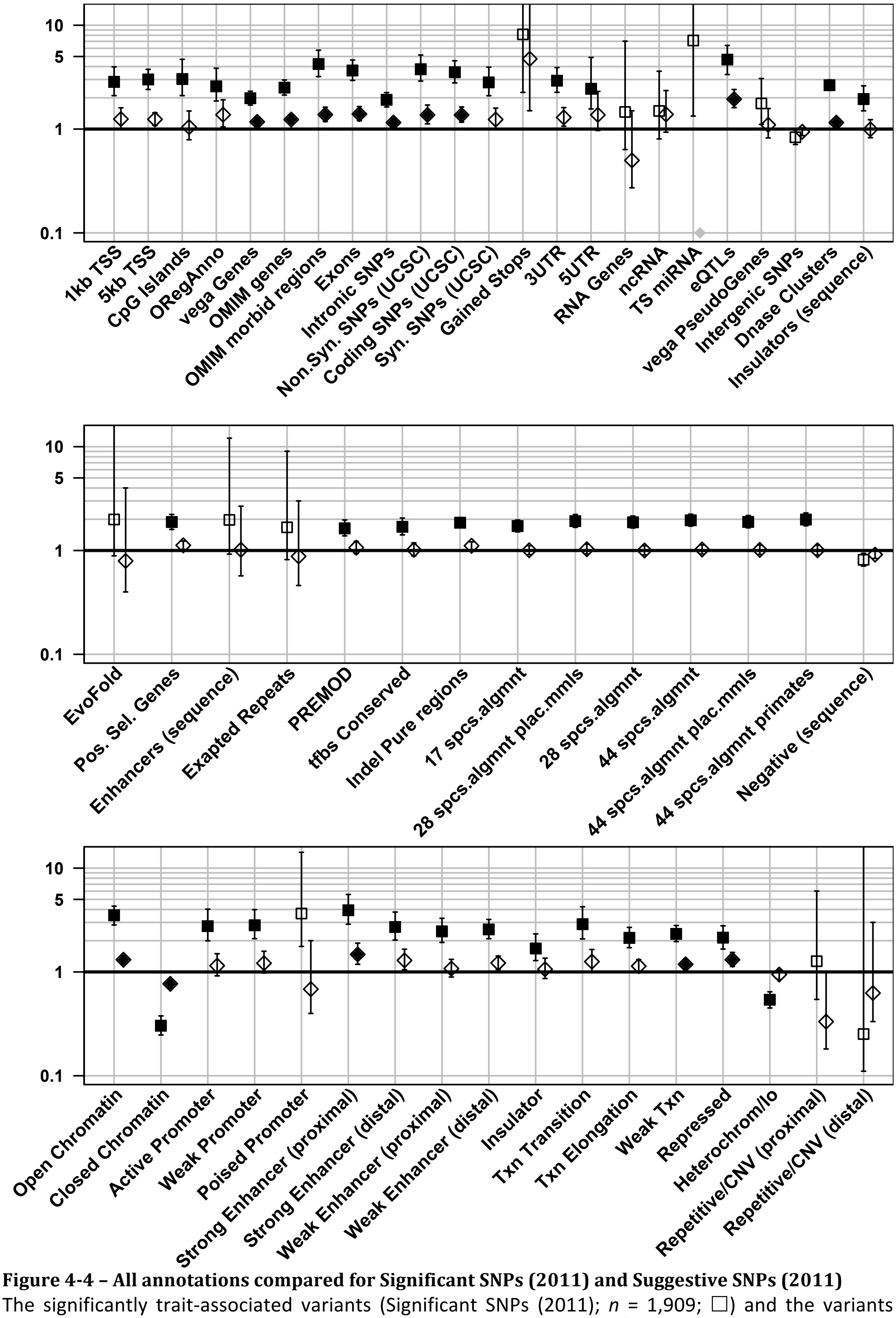


Figure 4-4 – All annotations compared for Significant SNPs (2011) and Suggestive SNPs (2011)

The significantly trait-associated variants (Significant SNPs (2011); *n* = 1,909; □) and the variants with a suggestive *P*-value of association (5×10^{-5} > *P*-value > 5×10^{-8}; Suggestive SNPs (2011); *n* = 2,410; ◇) in all genomic annotations show similar trends of enrichment and depletion. Solid symbols indicate significance at the multiple-testing corrected threshold. Top: Genic and regulatory regions. Middle: Conserved regions and evolutionary signatures. Bottom: Chromatin states and histone modifications.

Table 4-2 – Permutation results for Suggestive SNPs (2011)

This table summarises the results for Suggestive SNPs (2011), which contained 2,410 SNPs. the number of overlaps in the observed set (Real), the mean of the permuted hits (Permutation Mean), the calculated odds ratio and confidence interval (OR [LCI-HCI]) and the obtained *P*-value for each of the genomic annotations are shown below. Significant *P*-values in bold.

Annotation	Real	Permutation Mean	OR [LCI - HCI]]	P-value
1 Kb TSS	104	84.05	1.25 [1.01-1.60]	1.94×10^{-02}
5 Kb TSS	264	217.43	1.24 [1.08-1.44]	9.00×10^{-04}
CpG Islands	46	44.04	1.05 [0.79-1.49]	4.05×10^{-01}
ORegAnno	66	48.40	1.37 [1.05-1.91]	8.40×10^{-03}
vega Genes	1017	924.50	1.17 [1.08-1.28]	**5.00×10^{-05}**
OMIM genes	913	796.53	1.24 [1.13-1.35]	**$< 5.00 \times 10^{-05}$**
OMIM morbid regions	251	187.65	1.38 [1.18-1.62]	**$< 5.00 \times 10^{-05}$**
Exons	239	176.32	1.39 [1.20-1.65]	**$< 5.00 \times 10^{-05}$**
Intronic SNPs	1146	1059.80	1.15 [1.06-1.26]	**3.00×10^{-04}**
Non.Syn. SNPs (UCSC)	125	92.71	1.37 [1.12-1.70]	**7.00×10^{-04}**
Coding SNPs (UCSC)	190	142.10	1.37 [1.16-1.63]	**5.00×10^{-05}**
Syn. SNPs (UCSC)	91	74.17	1.24 [1.00-1.59]	3.09×10^{-02}
Gained Stops	6	1.27	4.74 [1.50-Infinity]	2.10×10^{-03}
3'UTR	120	93.83	1.29 [1.07-1.61]	4.75×10^{-03}
5'UTR	32	23.45	1.37 [0.97-2.30]	5.71×10^{-02}
RNA Genes	3	6.00	0.50 [0.27-1.50]	6.21×10^{-02}
ncRNA	28	20.28	1.38 [0.93-2.35]	6.73×10^{-02}
TS miRNA	0	0.72	0.00 [0.00-NA]	**$< 5.00 \times 10^{-05}$**
eQTLs	210	112.84	1.94 [1.61-2.40]	**$< 5.00 \times 10^{-05}$**
vega PseudoGenes	47	42.87	1.10 [0.82-1.58]	2.85×10^{-01}
Intergenic SNPs	1494	1527.22	0.94 [0.86-1.03]	8.68×10^{-02}
DNase Clusters	835	758.58	1.15 [1.06-1.26]	**3.00×10^{-04}**
Insulators (sequence)	105	105.58	0.99 [0.83-1.23]	4.61×10^{-01}
Within miRNA	0	0.04	0.00 [0.00-NA]	**$< 5.00 \times 10^{-05}$**
Splice Sites	3	1.05	2.85 [1.00-Infinity]	9.18×10^{-02}
Lost Stops	0	0.42	0.00 [0.00-NA]	**$< 5.00 \times 10^{-05}$**
Microsatellites	5	1.85	2.70 [1.00-Infinity]	4.13×10^{-02}
EvoFold	4	5.04	0.79 [0.40-4.00]	2.60×10^{-01}
Pos. Sel. Genes	750	691.65	1.12 [1.03-1.23]	6.65×10^{-03}
Enhancers (sequence)	8	7.87	1.02 [0.57-2.67]	4.72×10^{-01}
Exapted Repeats	6	6.83	0.88 [0.46-3.00]	3.25×10^{-01}
PREMOD	266	251.41	1.07 [0.94-1.22]	1.78×10^{-01}
tfbs Conserved	208	204.86	1.02 [0.89-1.18]	4.23×10^{-01}
Indel Pure regions	845	788.00	1.11 [1.02-1.21]	7.15×10^{-03}
17 spec. algmt	526	523.52	1.01 [0.91-1.11]	4.62×10^{-01}
28 spec. algmt plc.mmls	426	415.98	1.03 [0.93-1.15]	3.05×10^{-01}
28 spec. algmt	490	489.72	1.00 [0.91-1.11]	4.99×10^{-01}
44 spec. algmt	504	494.81	1.02 [0.93-1.13]	3.33×10^{-01}
44 spec. algmt plc.mmls	457	452.02	1.01 [0.92-1.13]	4.10×10^{-01}
44 spec. algmt prim.	433	431.29	1.00 [0.91-1.12]	4.69×10^{-01}
Negative (sequence)	1252	1306.89	0.91 [0.84-0.99]	1.53×10^{-02}
Open Chromatin	1308	1144.60	1.31 [1.19-1.44]	**$< 5.00 \times 10^{-05}$**
Closed Chromatin	1128	1286.75	0.77 [0.70-0.84]	**$< 5.00 \times 10^{-05}$**
Active Promoter	80	69.96	1.15 [0.92-1.50]	1.28×10^{-01}
Weak Promoter	89	73.76	1.21 [0.98-1.58]	4.64×10^{-02}
Poised Promoter	6	8.78	0.68 [0.40-2.00]	1.35×10^{-01}
Strong Enhancer (proximal)	113	77.75	1.48 [1.19-1.89]	**1.50×10^{-04}**
Strong Enhancer (distal)	109	85.19	1.29 [1.05-1.66]	6.35×10^{-03}
Weak Enhancer (proximal)	115	107.07	1.08 [0.89-1.32]	2.32×10^{-01}
Weak Enhancer (distal)	239	199.47	1.22 [1.06-1.42]	3.25×10^{-03}
Insulator	86	81.21	1.06 [0.86-1.36]	3.05×10^{-01}
Txn Transition	78	62.47	1.26 [1.00-1.65]	3.25×10^{-02}
Txn Elongation	255	226.96	1.14 [1.00-1.31]	3.25×10^{-02}
Weak Txn	466	405.59	1.18 [1.06-1.33]	**8.50×10^{-04}**
Repressed	228	177.83	1.31 [1.13-1.54]	**1.00×10^{-04}**
Heterochrom/lo	1939	1960.31	0.94 [0.85-1.05]	1.39×10^{-01}
Repetitive/CNV (proximal)	2	6.01	0.33 [0.18-1.00]	1.86×10^{-02}
Repetitive/CNV (distal)	3	4.79	0.63 [0.33-3.00]	1.46×10^{-01}

4.3.3 Results of Significant SNPs (2011) *vs.* Significant SNPs (2013)

An increased number of analysed SNPs resulted in an increased number of significant genomic annotations (see Figure 4-5), with three more significant annotations in the Significant SNPs (2013). The most notable results are the significant depletions in the negative set (middle panel) and the significant enrichment in the poised promoters and gained stop codons (bottom and top panel). The negative set consisted of only those intergenic SNPs that did not overlap with any of the analysed genic or conserved regions. These findings highlighted that results with narrower empirical confidence intervals could be obtained with a larger dataset. The poised promoters also passed the significance threshold after correcting for multiple testing. Since the difference of the odds ratios was not significantly different (see all panels of Figure 4-5 and bottom panel of Figure 4-6), the change must be due to a decrease of confidence interval width. The mean for the confidence interval width has decreased to 1.36 in Significant SNPs (2013) from 2.06 in Significant SNPs (2011), while the odds ratio decreased only marginally from 2.33 to 2.25. This was representative of the increase in statistical power caused by the larger number of analysed trait-associated SNPs. A linear model of Significant SNPs (2013) *vs.* Significant SNPs (2011) had an r^2 value of 0.91 and a significant *P*-value of 2.34×10^{-31}. This implies that the results from the same dataset but different methods (sampling and permutation) were more similar, than when two overlapping datasets were analysed with the same method (0.98 *vs.* 0.91, respectively). Figure 4-6 includes two graphs, the first of which plots the odds ratios for Significant SNPs (2011) obtained by sampling against those obtained by the permutations. The second plot in Figure 4-6 shows the odds ratios for Significant SNPs (2011) and Significant SNPs (2013) obtained by the permutation methods. The number of overlaps for Significant SNPs (2013) and the results obtained by the permutations, the resulting odds ratio with confidence intervals and the *P*-value of significance for the odds ratios of each genomic annotation are presented in Table 4-3.

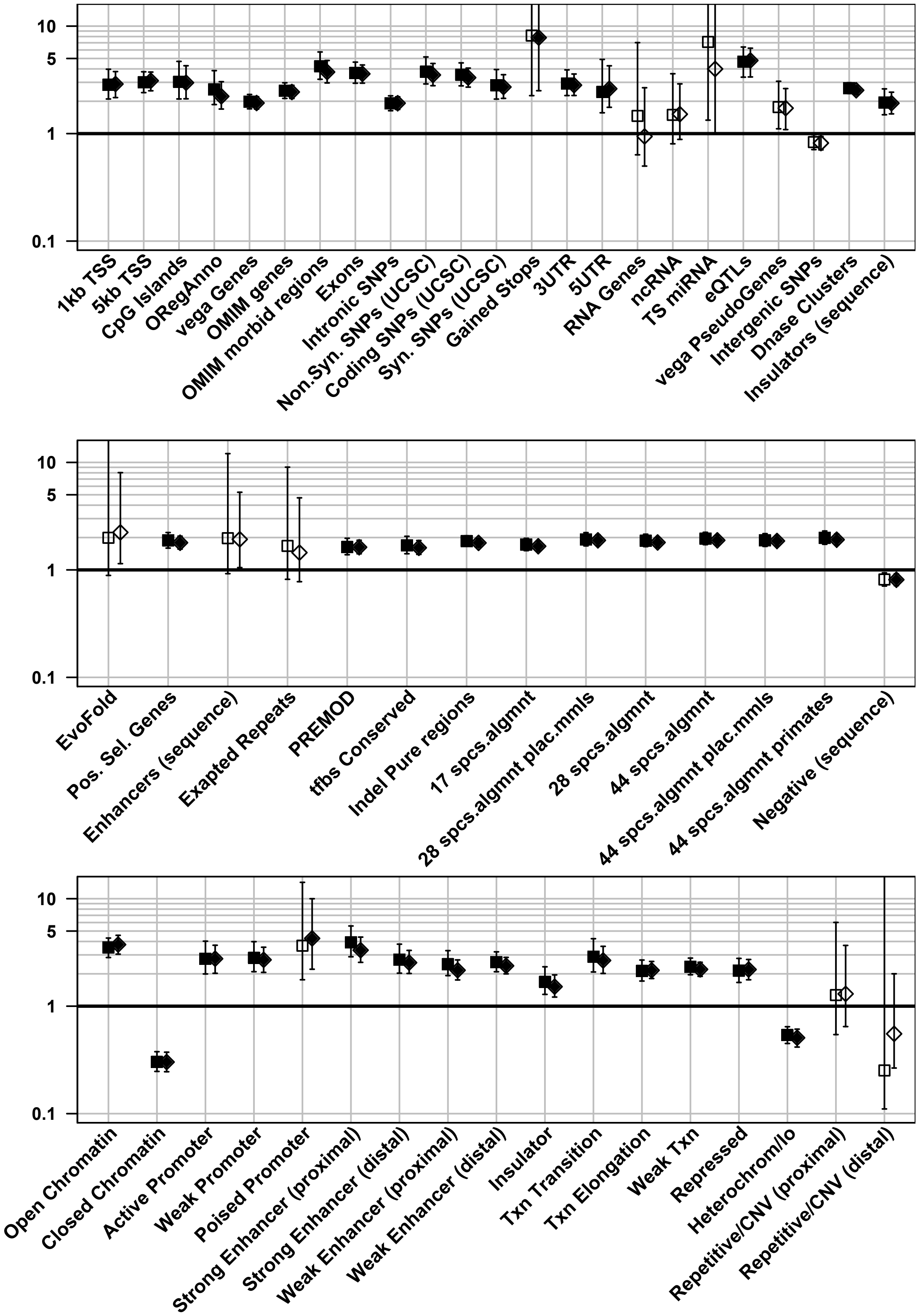
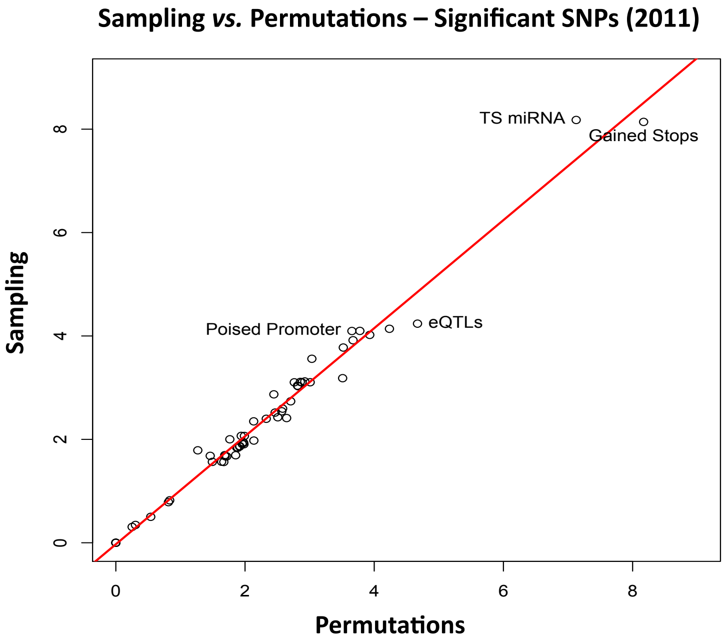


Figure 4-5 – Comparison of Significant SNPs (2011) and Significant SNPs (2013) in all annotations

The results of Significant SNPs (2011; *n* = 1,909; ☐/■) and Significant SNPs (2013; *n* = 3,283; (◇/◆) are very similar. The newer dataset obtains more significant results, probably due to the larger number of analysed SNPs. All *P*-values are corrected for multiple testing for the analysed genomic annotations and solid symbols indicate significance at that level. Top: Genic and regulatory regions. Middle: Conserved regions and evolutionary signatures. Bottom: Chromatin states and histone modifications.



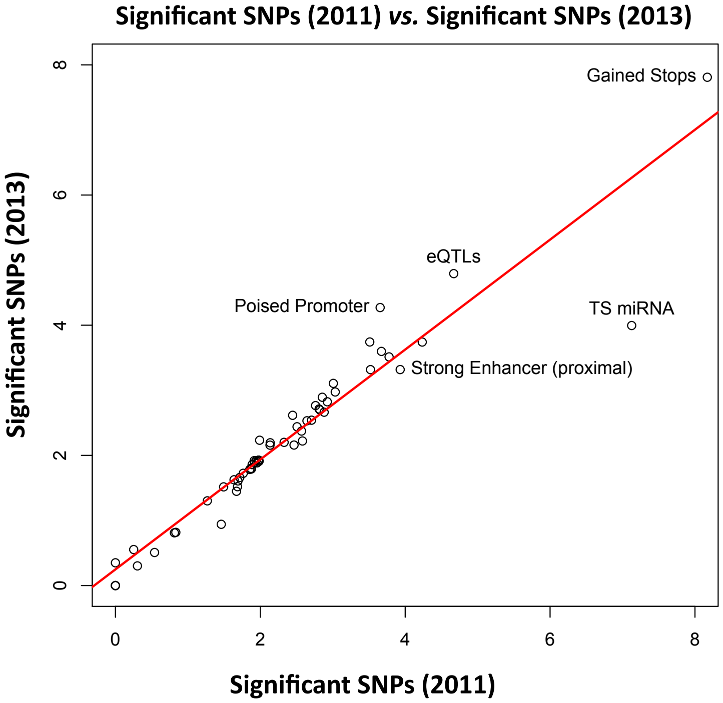


Figure 4-6 – Correlations between datasets and methods

Top: The correlation of odds ratios obtained by permutations and sampling in Significant SNPs (2011). A strong and significant correlation is observed between the two methods with a r^2 value of 0.98 and a *P*-value of 1.25×10^{-49}. Bottom: The correlation between two datasets analysed by permutations. Significant SNPs (2011) is a subset of Significant SNPs (2013). The correlation between the two datasets is strong and significant with an r^2 of 0.91 and a *P*-value of 2.17×10^{-31}. Correlations were determined by a linear regression of the x-axes values onto the y-axes values.

Table 4-3 – Permutation results for Significant SNPs (2013).

This table summarises the results for Significant SNPs (2013), which contained 3,283 SNPs. The number of overlaps in the observed set (Real), the mean of the permuted hits (Permutation Mean), the calculated odds ratio and confidence interval (OR [LCI-HCI]) and the obtained *P*-value for each of the genomic annotations are shown below. Significant *P*-values in bold.

Annotation	Real	Permutation Mean	OR [LCI - HCI]]	P-value
1 Kb TSS	345	128.07	2.89 [2.16-3.78]	**$< 5.00 \times 10^{-05}$**
5 Kb TSS	841	327.68	3.11 [2.50-3.74]	**$< 5.00 \times 10^{-05}$**
CpG Islands	196	68.62	2.97 [2.11-4.28]	**$< 5.00 \times 10^{-05}$**
ORegAnno	156	72.12	2.22 [1.69-3.04]	**$< 5.00 \times 10^{-05}$**
vega Genes	1874	1340.71	1.93 [1.69-2.21]	**$< 5.00 \times 10^{-05}$**
OMIM genes	1873	1157.37	2.44 [2.12-2.81]	**$< 5.00 \times 10^{-05}$**
OMIM morbid regions	837	275.12	3.74 [2.96-4.79]	**$< 5.00 \times 10^{-05}$**
Exons	793	267.00	3.60 [2.95-4.35]	**$< 5.00 \times 10^{-05}$**
Intronic SNPs	2070	1545.21	1.92 [1.68-2.20]	**$< 5.00 \times 10^{-05}$**
Non.Syn. SNPs (UCSC)	446	140.62	3.51 [2.79-4.49]	**$< 5.00 \times 10^{-05}$**
Coding SNPs (UCSC)	618	214.50	3.32 [2.71-4.07]	**$< 5.00 \times 10^{-05}$**
Syn. SNPs (UCSC)	288	112.37	2.71 [2.13-3.53]	**$< 5.00 \times 10^{-05}$**
Gained Stops	15	1.93	7.81 [2.51-Infinity]	**$< 5.00 \times 10^{-05}$**
3'UTR	369	140.92	2.82 [2.26-3.59]	**$< 5.00 \times 10^{-05}$**
5'UTR	92	35.79	2.62 [1.76-4.27]	**$< 5.00 \times 10^{-05}$**
RNA Genes	8	8.50	0.94 [0.05-2.67]	4.26×10^{-01}
ncRNA	46	30.48	1.52 [0.88-2.90]	6.18×10^{-02}
TS miRNA	4	1.00	4.00 [1.00-Infinity]	3.80×10^{-02}
eQTLs	684	170.89	4.79 [3.37-6.23]	**$< 5.00 \times 10^{-05}$**
vega PseudoGenes	113	66.48	1.72 [1.09-2.62]	1.38×10^{-02}
Intergenic SNPs	2005	2158.69	0.82 [0.71-0.94]	1.95×10^{-03}
DNase Clusters	1841	1100.58	2.53 [2.31-2.78]	**$< 5.00 \times 10^{-05}$**
Insulators (sequence)	290	157.94	1.92 [1.53-2.43]	**$< 5.00 \times 10^{-05}$**
Within miRNA	0	0.07	0.00 [0.00-NA]	**$< 5.00 \times 10^{-05}$**
Splice Sites	0	1.75	0.00 [0.00-NA]	**$< 5.00 \times 10^{-05}$**
Lost Stops	0	0.68	0.00 [0.00-NA]	**$< 5.00 \times 10^{-05}$**
Microsatellites	1	2.86	0.35 [0.12-Infinity]	1.05×10^{-01}
EvoFold	16	7.19	2.23 [1.14-8.03]	1.43×10^{-02}
Pos. Sel. Genes	1450	1005.77	1.79 [1.56-2.06]	**$< 5.00 \times 10^{-05}$**
Enhancers (sequence)	21	10.94	1.93 [1.05-5.28]	2.40×10^{-02}
Exapted Repeats	14	9.68	1.45 [0.78-4.68]	1.52×10^{-01}
PREMOD	543	356.76	1.63 [1.41-1.89]	**$< 5.00 \times 10^{-05}$**
tfbs Conserved	445	291.83	1.61 [1.39-1.88]	**$< 5.00 \times 10^{-05}$**
Indel Pure regions	1584	1126.87	1.78 [1.63-1.96]	**$< 5.00 \times 10^{-05}$**
17 spec. algmt	1075	744.36	1.66 [1.49-1.86]	**$< 5.00 \times 10^{-05}$**
28 spec. algmt plc.mmls	963	592.74	1.88 [1.68-2.12]	**$< 5.00 \times 10^{-05}$**
28 spec. algmt	1078	701.19	1.80 [1.62-2.01]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt	1119	706.72	1.89 [1.70-2.11]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt plc.mmls	1024	643.72	1.86 [1.66-2.09]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt prim.	1000	612.32	1.91 [1.71-2.16]	**$< 5.00 \times 10^{-05}$**
Negative (sequence)	1685	1856.23	0.81 [0.71-0.92]	**4.00×10^{-04}**
Open Chromatin	2594	1646.46	3.74 [3.05-4.56]	**$< 5.00 \times 10^{-05}$**
Closed Chromatin	931	1861.83	0.30 [0.25-0.37]	**$< 5.00 \times 10^{-05}$**
Active Promoter	277	105.89	2.76 [2.02-3.69]	**$< 5.00 \times 10^{-05}$**
Weak Promoter	282	110.46	2.70 [2.06-3.54]	**$< 5.00 \times 10^{-05}$**
Poised Promoter	59	14.01	4.27 [2.21-9.99]	**$< 5.00 \times 10^{-05}$**
Strong Enhancer (proximal)	358	116.78	3.32 [2.56-4.39]	**$< 5.00 \times 10^{-05}$**
Strong Enhancer (distal)	300	125.00	2.54 [2.02-3.30]	**$< 5.00 \times 10^{-05}$**
Weak Enhancer (proximal)	321	156.95	2.16 [1.75-2.69]	**$< 5.00 \times 10^{-05}$**
Weak Enhancer (distal)	619	292.71	2.37 [2.00-2.84]	**$< 5.00 \times 10^{-05}$**
Insulator	177	118.69	1.52 [1.22-1.95]	**3.00×10^{-04}**
Txn Transition	239	94.02	2.66 [2.02-3.60]	**$< 5.00 \times 10^{-05}$**
Txn Elongation	648	335.99	2.16 [1.81-2.61]	**$< 5.00 \times 10^{-05}$**
Weak Txn	1084	600.92	2.20 [1.89-2.56]	**$< 5.00 \times 10^{-05}$**
Repressed	532	265.97	2.19 [1.76-2.72]	**$< 5.00 \times 10^{-05}$**
Heterochrom/lo	2409	2772.18	0.51 [0.42-0.61]	**$< 5.00 \times 10^{-05}$**
Repetitive/CNV (proximal)	11	8.46	1.30 [0.65-3.68]	7.36×10^{-01}
Repetitive/CNV (distal)	4	7.22	0.55 [0.27-2.00]	1.15×10^{-01}

4.3.4 Results of Significant SNPs (2011) *vs.* Significant SNPs (Difference)

The regression of the permutation results onto the sampling results using the Significant SNPs (2011) in both analyses gave an r^2 value of 0.98. The results of the permutation method using Significant SNPs (2011) and Significant SNPs (2013) had a regression r^2 of 0.91. While both values are very high, an analysis of the difference between the sets was performed. This analysis focused on the trait-associated SNPs that appear in Significant SNPs (2013) only, *i.e.,* the most recent trait-associated variants. This set was termed Significant SNPs (Difference), contained 1,477 SNPs and was analysed the permutation method. The odds ratios of Significant SNPs (2011) and Significant SNPs (Difference) (see Figure 4-7) are no longer as similar as in the comparison between Significant SNPs (2011) and Significant SNPs (2013) (see Figure 4-5). There are three genomic annotations where the odds ratio is no longer significant for the recently identified trait-associated variants: open regulatory annotations (ORegAnno), regions associated with insulator activity and regions annotated as target sites for microRNAs (TSmiRNA) have an odds ratio of zero (*i.e.,* undefined). The mean of the odds ratios for Significant SNPs (Difference) in the 50 annotations previously outlined is 2.08 and lower than in either Significant SNPs (2013) (2.25) or Significant SNPs (2011) (2.33). Two additional annotations (RNA genes and repetitive/CNV (proximal) regions) obtain an undefined 95% confidence interval of (denoted "Infinity"). When these are removed, the mean of odds ratios increases to 2.14 and the mean of the confidence interval width is 1.94. There are two most plausible causes which could contribute to the drop in the value of the odds ratios: the lower number of trait-associated variants reduced the amount of signal, or the drop was caused by a slightly different distribution of the recent GWAS hits, or both. The results for the permutations of Significant SNPs (Difference) are presented in Table 4-4, which shows the number of overlaps in the real data and the mean number of permuted overlaps, the calculated odds ratios and their confidence intervals and their *P*-value.

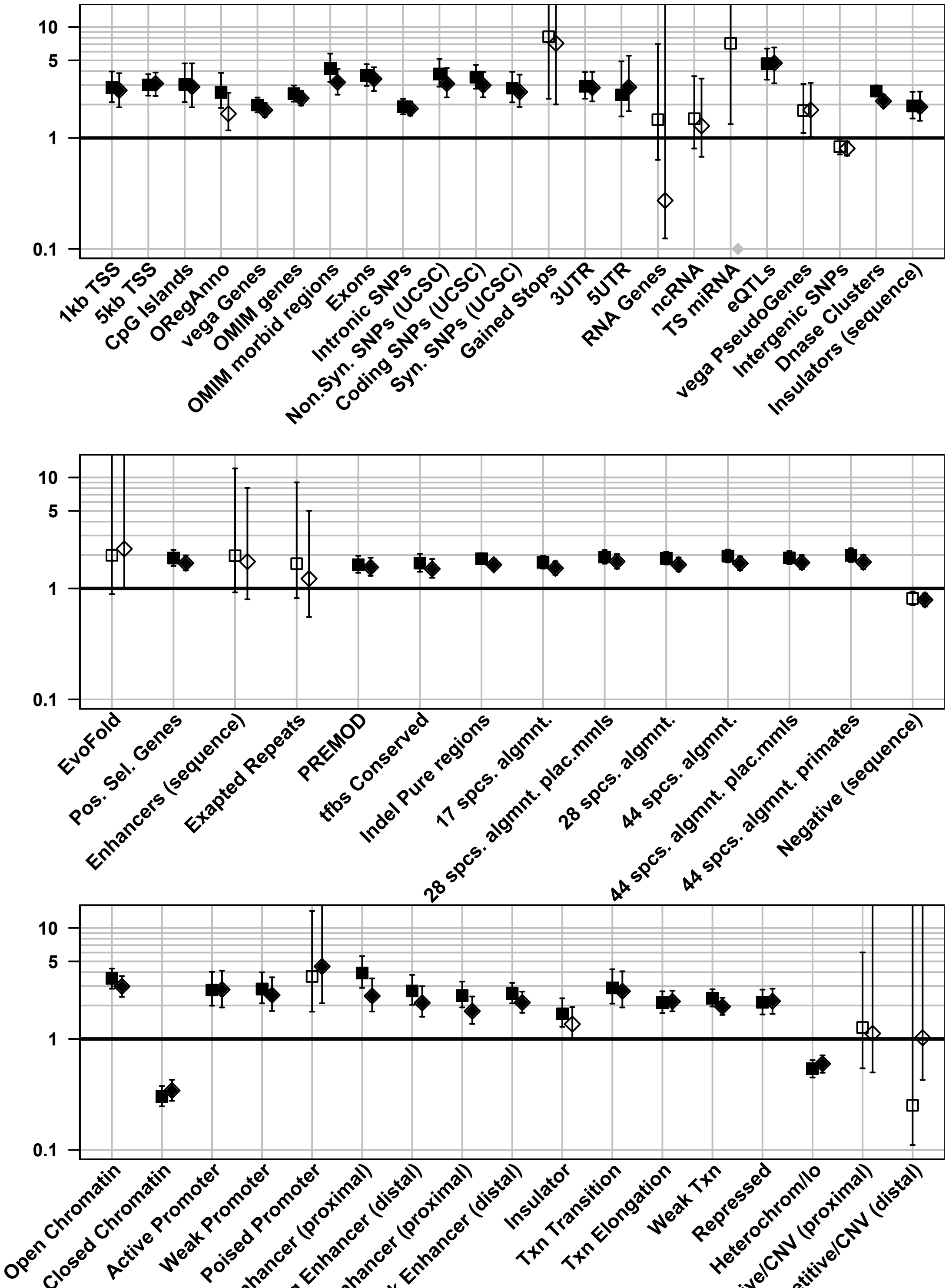


Figure 4-7 – Comparing Significant SNPs (2011) with Significant SNPs (Difference)

The results of Significant SNPs (2011; *n* = 1,909) and Significant SNPs (Difference; *n* = 1,477) (□,◇; respectively) show a number of differences. Significant SNPs (2013) have insignificant odds ratios in ORegAnno and Insulators and an unavailable odds ratio for TS miRNA in Significant SNPs (Difference). All *P*-values are corrected for multiple testing for the analysed genomic annotations and solid symbols indicate significance at that level. Top: Genic and regulatory regions. Middle: Conserved regions and evolutionary signatures. Bottom: Chromatin states and histone modifications.

The odds ratios of Significant SNPs (Difference) and Significant SNPs (2011) correlated weakly with each other (Figure 4-8). The r^2 of the regression of Significant SNPs (Difference) on Significant SNPs (2011) was 0.54, which was considerably less than all other comparisons, but still significant with a *P*-value of 2.94×10^{-11}. However, the information from this correlation was strongly influenced by the one change due to a genomic annotation, which was not significant in either dataset (TS miRNA). This annotation obtained an odds ratio of zero in the Significant SNPs (Difference), but reached a non-significant odds ratio of 7.13 in the Significant SNPs (2011). This could be due to a sampling bias, although the sizes of the datasets were fairly comparable (1,909 SNPs *vs.* 1,477 SNPs).

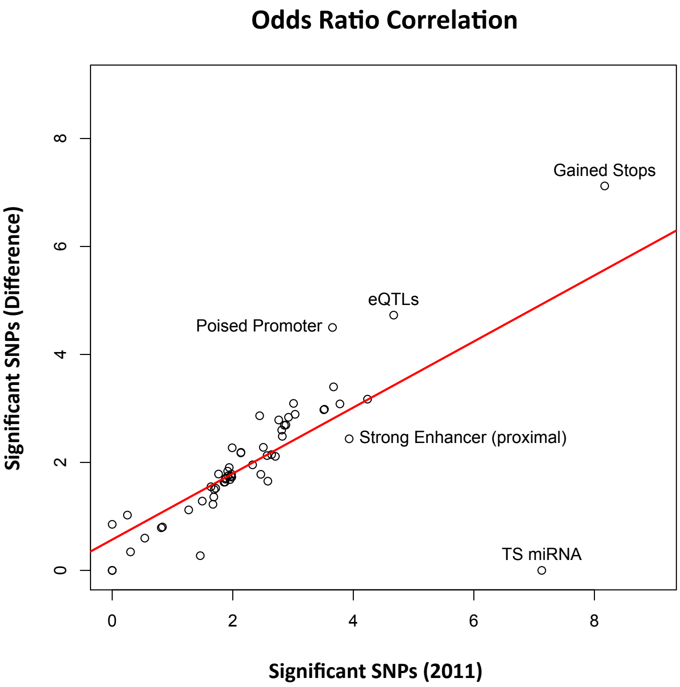


Figure 4-8 – Correlation of Significant SNPs (2011) with Significant SNPs (Difference)
The correlation of the odds ratios for Significant SNPs (2011) and Significant SNPs (Difference) is less significant and almost halved (r^2 = 0.54, *P*-value = 2.94×10^{-11}), when compared to the correlation of Significant SNPs (2011) with Significant SNPs (2013) (r^2 = 0.91, *P*-value = 2.17×10^{-31}). The most striking difference is in the TS miRNA annotation, which is zero in Significant SNPs (Difference), but in Significant SNPs (2011) it is 7.13, albeit not significantly enriched.

Table 4-4 – Permutation results for Significant SNPs (Difference)

This table summarises the results for Significant SNPs (Difference), which contained 1,477 SNPs. The number of overlaps in the observed set (Real), the mean of the permuted hits (Permutation Mean), the calculated odds ratio and confidence interval (OR [LCI-HCI]) and the obtained *P*-value for each of the genomic annotations are shown below. Significant *P*-values in bold.

Annotation	Real	Permutation Mean	OR [LCI-HCI]	P-value
1 Kb TSS	139	54.97	2.69 [1.89-3.83]	**$< 5.00 \times 10^{-05}$**
5 Kb TSS	362	140.52	3.09 [2.39-3.89]	**$< 5.00 \times 10^{-05}$**
CpG Islands	81	29.06	2.89 [1.89-4.70]	**$< 5.00 \times 10^{-05}$**
ORegAnno	50	30.67	1.65 [1.17-2.55]	1.50×10^{-03}
vega Genes	766	557.94	1.78 [1.55-2.06]	**$< 5.00 \times 10^{-05}$**
OMIM genes	784	491.47	2.28 [1.96-2.65]	**$< 5.00 \times 10^{-05}$**
OMIM morbid regions	315	116.44	3.17 [2.46-4.18]	**$< 5.00 \times 10^{-05}$**
Exons	327	114.17	3.40 [2.66-4.34]	**$< 5.00 \times 10^{-05}$**
Intronic SNPs	875	654.15	1.84 [1.59-2.13]	**$< 5.00 \times 10^{-05}$**
Non.Syn. SNPs (UCSC)	171	60.20	3.08 [2.32-4.27]	**$< 5.00 \times 10^{-05}$**
Coding SNPs (UCSC)	244	91.91	2.99 [2.32-3.92]	**$< 5.00 \times 10^{-05}$**
Syn. SNPs (UCSC)	119	48.18	2.60 [1.90-3.72]	**$< 5.00 \times 10^{-05}$**
Gained Stops	6	0.85	7.12 [2.00-Infinity]	1.10×10^{-03}
3'UTR	159	60.29	2.84 [2.14-3.93]	**$< 5.00 \times 10^{-05}$**
5'UTR	43	15.30	2.86 [1.74-5.51]	**$< 5.00 \times 10^{-05}$**
RNA Genes	1	3.66	0.27 [0.12-Infinity]	2.87×10^{-02}
ncRNA	17	13.28	1.28 [0.68-3.43]	2.20×10^{-01}
TS miRNA	0	0.45	0.00 [0.00-NA]	**$< 5.00 \times 10^{-05}$**
eQTLs	293	73.55	4.73 [3.11-6.53]	**$< 5.00 \times 10^{-05}$**
vega PseudoGenes	49	27.86	1.79 [1.02-3.13]	2.34×10^{-02}
Intergenic SNPs	845	921.66	0.80 [0.69-0.93]	1.40×10^{-03}
DNase Clusters	735	467.92	2.14 [1.90-2.43]	**$< 5.00 \times 10^{-05}$**
Insulators (sequence)	124	67.81	1.91 [1.43-2.62]	**$< 5.00 \times 10^{-05}$**
Within miRNA	0	0.03	0.00 [NA-NA]	**$< 5.00 \times 10^{-05}$**
Splice Sites	0	0.69	0.00 [0.00-NA]	**$< 5.00 \times 10^{-05}$**
Lost Stops	0	0.29	0.00 [0.00-NA]	**$< 5.00 \times 10^{-05}$**
Microsatellites	1	1.17	0.85 [0.25-Infinity]	3.28×10^{-01}
EvoFold	7	3.09	2.27 [1.00-Infinity]	4.87×10^{-02}
Pos. Sel. Genes	601	426.94	1.69 [1.45-1.99]	**$< 5.00 \times 10^{-05}$**
Enhancers (sequence)	8	4.59	1.75 [0.80-8.04]	1.10×10^{-01}
Exapted Repeats	5	4.09	1.22 [0.55-5.01]	3.85×10^{-01}
PREMOD	223	152.14	1.55 [1.30-1.89]	**$< 5.00 \times 10^{-05}$**
tfbs Conserved	179	124.53	1.50 [1.25-1.84]	**5.00×10^{-05}**
Indel Pure regions	649	479.55	1.63 [1.45-1.85]	**$< 5.00 \times 10^{-05}$**
17 spec. algmt	435	318.28	1.52 [1.33-1.76]	**$< 5.00 \times 10^{-05}$**
28 spec. algmt plc.mmls	392	253.14	1.75 [1.51-2.05]	**$< 5.00 \times 10^{-05}$**
28 spec. algmt	433	298.87	1.64 [1.43-1.90]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt	445	301.62	1.68 [1.47-1.95]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt plc.mmls	415	275.01	1.71 [1.48-2.00]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt prim.	400	261.50	1.73 [1.50-2.01]	**$< 5.00 \times 10^{-05}$**
Negative (sequence)	707	793.37	0.79 [0.69-0.90]	**2.50×10^{-04}**
Open Chromatin	1074	702.76	2.98 [2.40-3.69]	**$< 5.00 \times 10^{-05}$**
Closed Chromatin	416	785.57	0.34 [0.28-0.43]	**$< 5.00 \times 10^{-05}$**
Active Promoter	120	45.45	2.79 [1.92-4.13]	**$< 5.00 \times 10^{-05}$**
Weak Promoter	112	47.25	2.48 [1.78-3.59]	**$< 5.00 \times 10^{-05}$**
Poised Promoter	27	6.09	4.50 [2.10-27.49]	**5.00×10^{-05}**
Strong Enhancer (proximal)	116	49.91	2.44 [1.77-3.51]	**$< 5.00 \times 10^{-05}$**
Strong Enhancer (distal)	108	53.19	2.11 [1.59-2.99]	**$< 5.00 \times 10^{-05}$**
Weak Enhancer (proximal)	115	66.92	1.78 [1.37-2.41]	**$< 5.00 \times 10^{-05}$**
Weak Enhancer (distal)	242	124.69	2.13 [1.72-2.67]	**$< 5.00 \times 10^{-05}$**
Insulator	68	50.61	1.36 [1.02-1.93]	2.03×10^{-02}
Txn Transition	104	40.42	2.69 [1.92-4.07]	**$< 5.00 \times 10^{-05}$**
Txn Elongation	280	143.44	2.18 [1.78-2.72]	**$< 5.00 \times 10^{-05}$**
Weak Txn	430	256.67	1.96 [1.65-2.35]	**$< 5.00 \times 10^{-05}$**
Repressed	227	113.34	2.19 [1.68-2.83]	**$< 5.00 \times 10^{-05}$**
Heterochrom/lo	1041	1178.48	0.60 [0.50-0.71]	**$< 5.00 \times 10^{-05}$**
Repetitive/CNV (proximal)	4	3.57	1.12 [0.50-Infinity]	4.65×10^{-01}
Repetitive/CNV (distal)	3	2.93	1.02 [0.43-Infinity]	4.48×10^{-01}

4.3.5 Comparison of different LD thresholds in Significant SNPs (2011)

Recall, we had used r^2 abiding with standard notation to denote a measure of LD and the correlation computed by linear threshold models. In this thesis we will be using both concepts frequently, so to avoid confusion, we will be using r^2_{LD} and r^2_c to denote the LD threshold and correlation computed using linear regression models, respectively, in this section only. To continue the studies presented so far, an investigation into the influence of using different LD thresholds for the Significant SNPs (2011) was also performed, as an LD cut-off point of $r^2_{LD} > 0.9$ is considered rather stringent [50]. An additional LD threshold of $r^2_{LD} > 0.7$ was investigated to identify the effect of a lower threshold resulting in a larger number of LD partners. Since LD decays with increasing distance, the lower threshold would have a larger mean distance between trait-associated variants and their LD partners than the more stringent cut-off. Figure 4-9 shows the resulting odds ratios of permutations using the two thresholds for Significant SNPs (2011). There were seven genomic annotations, where the results of the analyses using different LD thresholds vary in significance, measured by the multiple testing corrected *P*-value obtained by the permutations. The lower threshold ($r^2_{LD} > 0.7$) obtained significance in four of these annotations (gained stop codons, TSmiRNA, intergenic SNPs, and negative) where the higher threshold ($r^2_{LD} > 0.9$) did not. The higher threshold was significant in three annotations (5'UTRs, insulators from the chromatin states, and weak transcription), where the lower was not. There was a strong and positive correlation ($r^2_c = 0.82$) of the odds ratios of the two LD thresholds with a significant *P*-value of 5.61×10^{-23} (Figure 4-10). The number of overlaps for the observed data, the mean number of overlaps for the permutations, the calculated odds ratios, their confidence intervals and the observed *P*-value of the analysis with the less stringent LD cut-of point ($r^2_{LD} > 0.7$) are shown in Table 4-5.

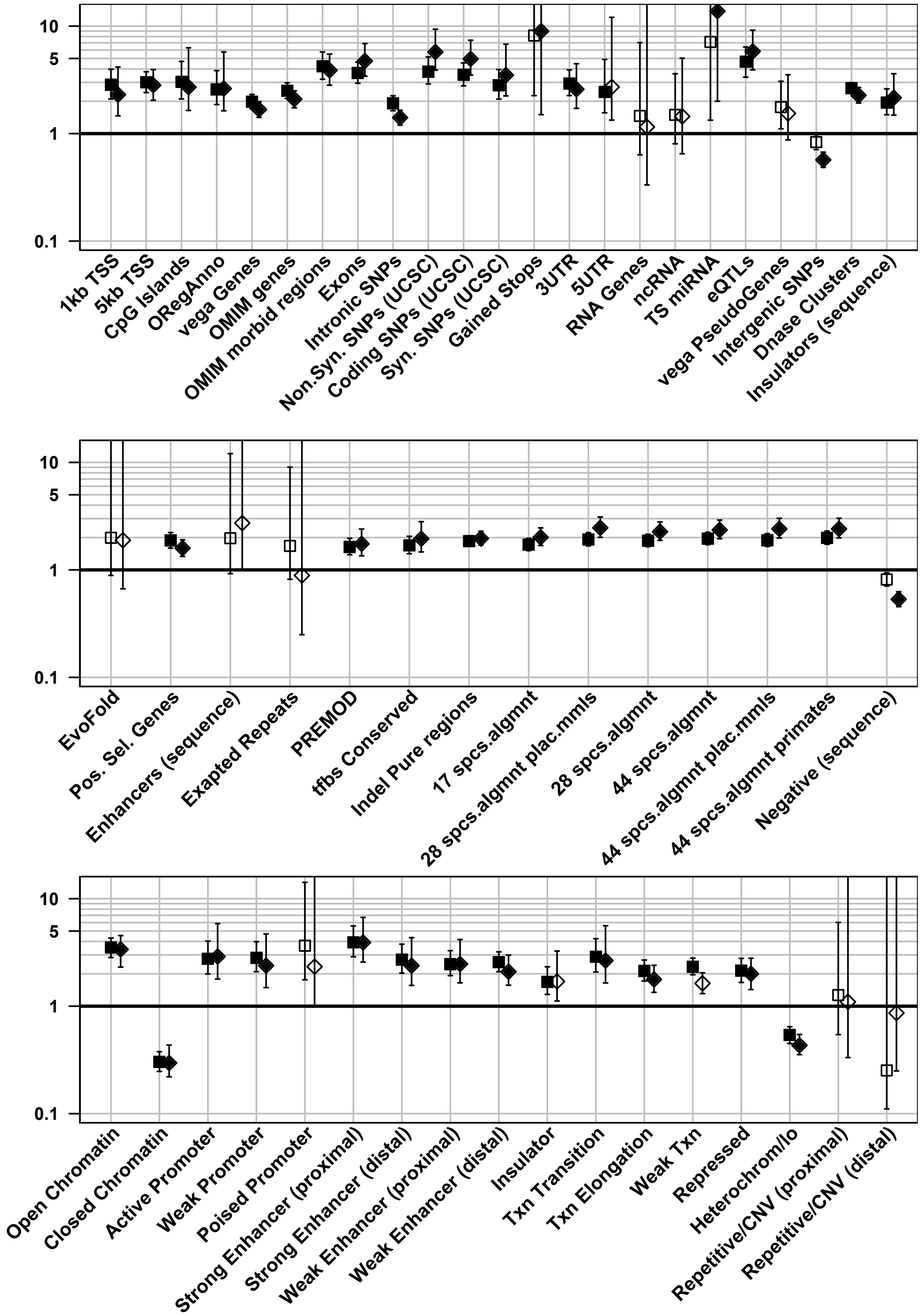


Figure 4-9 – Comparison of $r^2>0.9$ with $r^2>0.7$ in Significant SNPs (2011)

The results of the two LD thresholds ($r^2>0.9$ (□) and $r^2>0.7$ (◇)) for Significant SNPs (2011) show no significant differences. All *P*-values are corrected for multiple testing for the analysed genomic annotations and solid symbols indicate significance at that level. Top: Genic and regulatory regions. Middle: Conserved regions and evolutionary signatures. Bottom: Chromatin states and histone modifications.

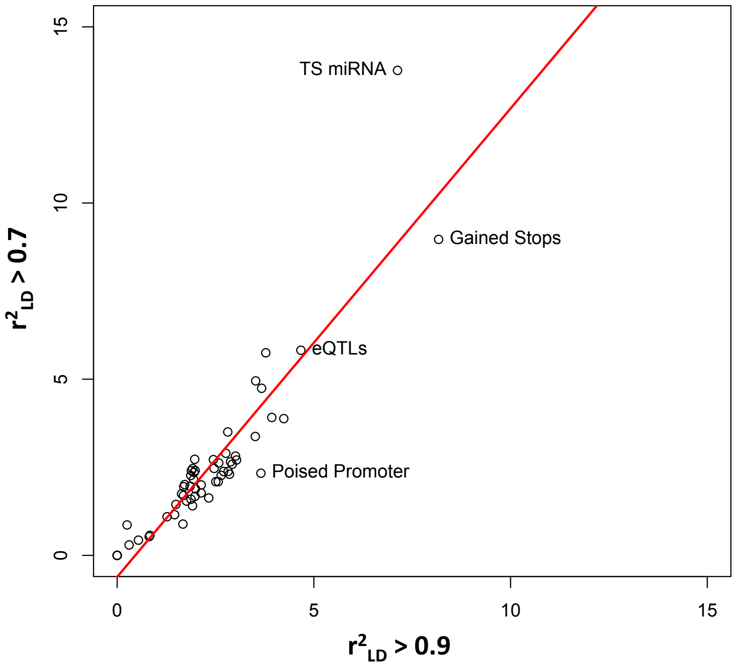


Figure 4-10 – Correlation of the odds ratios in the analyses of the LD partners of r^2_{LD}>0.9 and r^2_{LD}>0.7 in Significant SNPs (2011)

A strong positive correlation is observed for the odds ratios of the two different LD thresholds analysed in Significant SNPs (2011) (r^2_c = 0.82). The *P*-value of the correlation is highly significant with a value of 5.61 × 10^{-23}.

Table 4-5 – Permutations for Significant SNPs (2011) at $r^2_{LD} > 0.7$ LD threshold

This table summarises the results for Significant SNPs (2011) at a lower LD threshold ($r^2_{LD} > 0.7$). The number of overlaps in the observed set (Real), the mean of the permuted hits (Permutation Mean), the calculated odds ratio and confidence interval (OR [LCI-HCI]) and the obtained *P*-value for each of the genomic annotations are shown below. Significant *P*-values in bold.

Annotation	**Real**	**Permutation Mean**	**OR [LCI-HCI]**	**P-value**
1 Kb TSS	45	19.80	2.30 [1.46-4.17]	**4.50×10^{-04}**
5 Kb TSS	203	77.33	2.82 [2.04-3.94]	**5.00×10^{-05}**
CpG Islands	31	11.57	2.71 [1.64-6.29]	**1.00×10^{-04}**
ORegAnno	34	13.11	2.62 [1.63-5.75]	**5.00×10^{-05}**
vega Genes	905	667.80	1.68 [1.41-1.98]	**$< 5.00 \times 10^{-05}$**
OMIM genes	903	573.34	2.09 [1.74-2.50]	**$< 5.00 \times 10^{-05}$**
OMIM morbid regions	423	130.43	3.88 [2.82-5.50]	**$< 5.00 \times 10^{-05}$**
Exons	223	51.81	4.74 [3.42-6.88]	**$< 5.00 \times 10^{-05}$**
Intronic SNPs	931	770.03	1.41 [1.20-1.65]	**1.50×10^{-04}**
Non.Syn. SNPs (UCSC)	140	25.92	5.75 [3.90-9.36]	**$< 5.00 \times 10^{-05}$**
Coding SNPs (UCSC)	183	40.02	4.95 [3.51-7.39]	**$< 5.00 \times 10^{-05}$**
Syn. SNPs (UCSC)	53	15.44	3.50 [2.24-6.79]	**$< 5.00 \times 10^{-05}$**
Gained Stops	3	0.34	8.97 [1.50-Infinity]	**4.50×10^{-04}**
3'UTR	61	24.06	2.59 [1.72-4.47]	**$< 5.00 \times 10^{-05}$**
5'UTR	12	4.43	2.72 [1.34-12.07]	1.15×10^{-03}
RNA Genes	1	0.87	1.15 [0.33-Infinity]	2.17×10^{-01}
ncRNA	15	10.40	1.45 [0.65-5.03]	1.44×10^{-01}
TS miRNA	2	0.15	13.76 [2.00-Infinity]	**7.00×10^{-04}**
eQTLs	222	42.20	5.82 [3.92-9.17]	**$< 5.00 \times 10^{-05}$**
vega PseudoGenes	21	13.69	1.54 [0.87-3.53]	6.20×10^{-02}
Intergenic SNPs	823	1089.91	0.57 [0.49-0.67]	**5.00×10^{-05}**
DNase Clusters	472	241.34	2.27 [1.93-2.69]	**$< 5.00 \times 10^{-05}$**
Insulators (sequence)	53	24.95	2.16 [1.49-3.61]	**1.50×10^{-04}**
Within miRNA	0	0.01	0.00 [NA-NA]	8.90×10^{-03}
Splice Sites	0	0.25	0.00 [0.00-NA]	2.20×10^{-01}
Lost Stops	0	0.05	0.00 [0.00-NA]	5.22×10^{-02}
Microsatellites	0	0.31	0.00 [0.00-NA]	2.70×10^{-01}
EvoFold	2	1.06	1.89 [0.67-Infinity]	9.40×10^{-02}
Pos. Sel. Genes	643	461.62	1.59 [1.34-1.90]	**$< 5.00 \times 10^{-05}$**
Enhancers (sequence)	5	1.84	2.73 [1.00-Infinity]	1.64×10^{-02}
Exapted Repeats	1	1.13	0.89 [0.25-Infinity]	3.09×10^{-01}
PREMOD	100	58.53	1.75 [1.35-2.40]	**$< 5.00 \times 10^{-05}$**
tfbs Conserved	74	38.71	1.95 [1.47-2.81]	**$< 5.00 \times 10^{-05}$**
Indel Pure regions	407	231.63	1.96 [1.70-2.29]	**$< 5.00 \times 10^{-05}$**
17 spec. algmt	237	125.79	2.01 [1.69-2.46]	**$< 5.00 \times 10^{-05}$**
28 spec. algmt plc.mmls	223	97.27	2.46 [2.01-3.10]	**$< 5.00 \times 10^{-05}$**
28 spec. algmt	246	116.97	2.27 [1.88-2.79]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt	257	118.45	2.35 [1.95-2.91]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt plc.mmls	238	106.74	2.40 [1.98-3.02]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt prim.	233	104.25	2.41 [1.98-3.02]	**$< 5.00 \times 10^{-05}$**
Negative (sequence)	572	850.14	0.53 [0.46-0.63]	**$< 5.00 \times 10^{-05}$**
Open Chromatin	1429	894.69	3.38 [2.31-4.54]	**$< 5.00 \times 10^{-05}$**
Closed Chromatin	477	1010.41	0.30 [0.22-0.43]	**$< 5.00 \times 10^{-05}$**
Active Promoter	46	16.13	2.90 [1.79-5.87]	**$< 5.00 \times 10^{-05}$**
Weak Promoter	37	15.70	2.38 [1.49-4.70]	**1.00×10^{-04}**
Poised Promoter	6	2.58	2.33 [1.00-Infinity]	2.26×10^{-02}
Strong Enhancer (proximal)	90	23.83	3.91 [2.57-6.70]	**$< 5.00 \times 10^{-05}$**
Strong Enhancer (distal)	51	21.75	2.38 [1.56-4.34]	**4.00×10^{-04}**
Weak Enhancer (proximal)	57	23.52	2.47 [1.65-4.17]	**$< 5.00 \times 10^{-05}$**
Weak Enhancer (distal)	112	55.21	2.09 [1.57-2.99]	**5.00×10^{-05}**
Insulator	29	17.16	1.70 [1.12-3.26]	4.95×10^{-03}
Txn Transition	44	16.81	2.66 [1.64-5.61]	**5.00×10^{-05}**
Txn Elongation	190	112.05	1.77 [1.34-2.40]	**7.50×10^{-04}**
Weak Txn	306	200.10	1.63 [1.31-2.04]	1.00×10^{-03}
Repressed	156	81.35	2.00 [1.43-2.79]	**$< 5.00 \times 10^{-05}$**
Heterochrom/lo	1067	1423.58	0.43 [0.35-0.54]	**2.50×10^{-04}**
Repetitive/CNV (proximal)	2	1.82	1.10 [0.33-Infinity]	2.72×10^{-01}
Repetitive/CNV (distal)	1	1.16	0.86 [0.25-Infinity]	3.17×10^{-01}

4.4 Discussion

In this chapter, we presented the enrichment and depletion results of different datasets of trait-associated SNPs in genomic regions annotated for 58 different functional genomic features using two different methods. Here, we discuss obtained results.

4.4.1 Permutations *vs.* sampling

4.4.1.1 Computational requirements

We have developed and used a novel permutation method, which took the internal genomic structure into account by preserving the distance between trait-associated SNPs in each permutation. The results obtained by the permutations were compared to the results obtained by the sampling approach, which does not depend on the internal structure, discussed in Chapter 3. The permutation method kept not only the number of SNPs per chromosome but also the relative distance of these SNPs fixed in every permutation. However, the process of performing the permutation analyses is quite time-consuming. The permutations were performed 20,000 times, which took almost three days to complete for the entire genome. The 20,000 permutations meant that the *P*-value of significance was delimited at 5×10^{-5}. To increase the significance by one order of magnitude to 5×10^{-6} would have meant increasing the permutation number, and therefore the computational time, by an order of magnitude, which would have been 30 days. In our analyses, this increase in time was not justifiable for a change in significance of only one order of magnitude.

The sampling method is comparatively faster than the permutation method. It was divided into two steps: the preparation of the samples and the analysis itself. The analysis itself took less than an hour, but the establishing of mutually exclusive samples and the establishing of their LD partners required a day. However, it was only 100 samples that were analysed, as the study we attempted to replicate also used 100 samples [50]. These 100 samples were not enough to establish defined confidence intervals that were based on the distribution of the expected results, as was done for the permutation method. The empirical confidence intervals of the permutation method gave a more conservative measure of enrichment or depletion as their average width was generally wider than the average of the theoretical confidence interval width for the sampling method.

4.4.1.2 Genomic annotations enriched for trait-associated SNPs

Overall, we observed significant enrichment significant GWAS hits in genic annotations and several features associated with particular chromatin states with both methods. The enrichment in genic annotations had been well documented in previous studies [50, 65], while there had been evidence for enrichment of trait-associated SNPs in regions with distinct chromatin structures [82].

There were some differences between the significant odds ratios obtained by the two methods. Among these was the negative set, which was created as an approximation to a negative control (Figure 4-2, middle panel). As mentioned before, the Bonferroni corrected *P*-value threshold was 8.62×10^{-4}. The *P*-values of the results had to be less than that threshold to be significant. The permutation method obtained an odds ratio of 0.81 [0.71-0.94] (*P*-value = 1.95×10^{-3}), while the sampling method obtained an odds ratio of 0.79 [0.69-0.89] with a *P*-value of 1.73×10^{-4}. The difference was marginal, but the *P*-value for the permutations was not significant after correcting for multiple testing using the Bonferroni correction, but the odds ratio obtained by sampling was. The poised promoters annotations were also not significantly enriched in the permutation results of Significant SNPs (2011). The chromatin states associated with insulating activity, however, were significant for the permutations, but not for the sampling method. Several of the sparsely annotated genomic features resulted in large confidence intervals on the estimated odds ratio by either method, confirming the difficulties in obtaining results.

The observed differences in significance were due to the different methods used to calculate the *P*-values and the confidence intervals. The theoretically determined *P*-values, which were a widely used asymptotic approximation (see page 50 of this thesis) used in the sampling method, were compared with those determined empirically in the circular permutations method (see page 71 of this thesis). Theoretical values were used for the sampling method since they were necessarily based on a limited number of random samples, and such limitations did not apply to the permutation approach. The confidence intervals derived by permutation were generally more conservative (*i.e.,* larger) than those for the sampling approach were. This is because the permutation confidence intervals were based on the empirically derived confidence intervals rather than the calculated ones. The empirical confidence intervals were chosen to represent the distribution of the overlaps obtained from the permutations not only in the tables, but also on the graphs. They clearly show where the 95% of the obtained permutation overlaps were.

4.4.2 Modest functional enrichment in Suggestive SNPs (2011)

There has been substantial interest in trait-associated variants with modest associations of *P*-values, as they are suspected to contain real positive associations. We therefore analysed SNPs with more moderate *P*-values that did not pass the genome-wide significance threshold ($5 \times 10^{-5} > P\text{-value} > 5 \times 10^{-8}$) to determine if their distribution was similar to the distribution of significant GWAS hits. Suggestive SNPs (2011) showed similar results to Significant SNPs (2011) but with odds ratios that were less extreme than the odds ratios for Significant SNPs (2011). This result was consistent with the suggestively associated SNPs being a mixture of false positives (which we would expect to have no bias towards particular annotations) and true associations, whose effects were not of sufficient magnitude to show genome-wide significance. These true positives would be expected to have the same bias towards particular genomic features as trait-associated SNPs attaining genome-wide significance [65].

The observed modest enrichment of functional elements in SNPs with not significant *P*-values of association is most likely due to reporting bias of associated SNPs in the literature. While significantly trait-associated SNPs are documented consistently, suggestive associations often remained unreported since they are generally assumed to contribute less to our understanding of the underlying biology. Additionally, the NHGRI GWAS catalogue only incorporated result variants with association levels starting at 5 × 10^{-5}, where the more commonly accepted level for suggestively associated SNPs is 5 × 10^{-4} [36, 50]. This meant that the significantly associated SNP set was likely to be a more comprehensive and complete SNP set of true associations, despite containing a smaller number of SNPs. The similarity of enrichment trends between Significant SNPs (2011) and Suggestive SNPs (2011) were encouraging. The results might be of use in follow-up studies identifying true associations from suggestively trait-associated variants by focusing the search on areas, which were enriched for significantly trait-associated SNPs. The prediction of functional suggestive variants may be improved by possibly investigating only those SNPs that were overlapping with multiple annotations. However, care should be taken in choosing the multiple overlapping annotations given that some annotations are overlapping by definition, e.g., exons and genes. An analysis to estimate the proportion of true positives in the set of suggestive variants could be undertaken by investigating which of the suggestive variants were replicated in the SNP set of 2013. However, this estimate would probably be biased as, as mentioned above, the suggestive SNPs set is likely incomplete.

4.4.3 Analysis of trait-associated variants identified since 2011

The Significant SNPs (2011) and Significant SNPs (2013) did not show significant differences in the analyses, despite the more recent set containing almost twice as many variants. This was likely due to the large overlap between the two sets. The Significant SNPs (Difference) showed three significant differences with an apparent shift of trait-associated SNPs from transcriptional elongation regions to regions 5 Kb upstream of transcription start sites.

Significant SNPs (2013) resulted in a larger number of genomic features with significant enrichment for trait-associated variants arising from the larger number of analysed variants. The most prominent difference between the datasets was the smaller confidence interval for the newer set. However, this was also expected, as a larger number of variants would have provided more statistical power for the analysis, which was reflected in the reduced 95% confidence interval widths.

We wanted to investigate the new additions to the NHGRI catalogue to compare them with the set of SNPs from 2011. The analysis showed that the results for Significant SNPs (2011) influenced the Significant SNPs (2013) results. We observed three genomic annotations, which changed significance. These were the ORegAnno annotation, gained stop codons, and the chromatin states associated with insulators. The odds ratios of predicted binding sites of miRNA (TS miRNA) and RNA genes dropped to odds ratios indicating depletion, although these odds ratios did not reach significance. A quick investigation into the effect sizes in the data showed that Significant SNPs (Difference) contained 11 SNPs with an odds ratio higher than 50, while Significant SNPs (2011) only had three. These 11 variants were unlikely to significantly change the distribution of the data, but they indicated that variants with possibly different mechanisms were identified. The mean risk allele frequencies of the trait-associated variants in Significant SNPs (2011), Significant SNPs (2013), and Significant SNPs (Difference) remained constant (0.39-0.40). It can therefore not be said, that the newer GWAS identified rare variants with large effects or more eQTLs. As mentioned in the Methods section, odds ratios were undefined when zero overlaps were observed in the trait-associated SNP sets. This could have been a sign of severe depletion in those genomic annotations where the odds ratio was not available. However, it was not possible to accurately gain insight into this, as the overlap of these genomic annotations with all SNPs was relatively low (see Table 2-1). The calculations of odds ratios, as done here, were known to perform poorly when small sample sizes were analysed [140], so a greater accuracy would only be gained once more information became available.

4.4.4 Comparison of different LD thresholds

The LD threshold for SNPs segregating with the associated variant in all datasets was set to $r^2_{LD} > 0.9$. However, for completeness we analysed an additional, lower LD threshold for Significant SNPs (2011) to $r^2_{LD} > 0.7$. The results for the lower LD threshold originated from more variants, as a larger number of SNPs segregated with the trait-associated variant at the lower threshold. Additionally, these variants were located at a greater distance away from the trait-associated variant, as LD decays with distance: LD between two variants decreases as the distance between them increases [141]. The results for the threshold of 0.7 were compared with the threshold of 0.9. The increased number of LD partners did result in more odds ratios with significant *P*-values in the genic regions. However, there were less odds ratios reaching significance in the chromatin states associated with different regulatory states. The number of genomic annotations for which the higher confidence interval bound was not available increased to 13 in the 0.7 set with an additional five annotations (RNA genes, enhancers (sequence), exapted repeats, poised promoters and repetitive/CNV (proximal)). The confidence interval widths were calculated for the remaining 45 genomic annotations. This increase in annotations for which the higher confidence interval was not available suggested that the higher number of analysed variants introduced more noise.

We argued that the larger confidence intervals observed in the comparison between the permutations *vs.* the sampling method were a more conservative measure. This is because the confidence intervals are directly related to the overlaps seen in the permutations rather than a theoretically calculated interval. Furthermore, they are not defined to be symmetrically distributed around the odds ratio unlike the theoretical values, so that the underlying distribution of the permutation overlaps is immediately evident from the result graphs. In the lower $r^2_{\ LD}$ threshold analysis more annotations had infinite confidence interval limits resulting from very large differences between the overlaps in the permutations and real dataset. This is likely due to the extra noise added by the additionally analysed variants that were included in the data when the lower r^2_{LD} threshold was analysed. The confidence intervals used for the permutations therefore give a good idea of the underlying distribution of the expected data.

4.4.5 Conclusion

The majority of the functional annotations showed enrichment for trait-associated SNPs of all analysed datasets. Some of the genomic annotations were defined as overlapping with at least one other genomic annotation due to their location. For example, non-synonymous SNPs are defined as SNPs within coding regions, which can change the resulting protein sequence. These SNPs will overlap with regions defined as genes in the gene datasets and possibly with functional elements associated with different chromatin states. These genomic annotations are therefore not mutually exclusive, though examining them all may provide extra information. Such dependencies among annotations make drawing conclusions from these results difficult. Additional analyses in later chapters examine solutions to this problem, and investigate the relative effect of the individual genomic annotation using a stepwise regression approach.

5 LOGISTIC REGRESSION

5.1 Introduction

As we have shown in the previous chapter, the methods discussed so far used to analyse enrichment or depletion in individual genomic annotations discussed so far gave encouraging results. However, the analysis of individual genomic annotations could give rise to false positive enrichment signals when two or more annotations are overlapping. In this case it is impossible to distinguish, which of the overlapping genomic annotations are causing the observed signal. A different method, logistic regression, was therefore applied to identify those genomic annotations with the most influence on trait association status. Logistic regression produces results based on an information criterion calculated for each genomic annotation and determines the impact of single or multiple independent variables. These variables are presented simultaneously and the analysis predicts if the variables are associated with one of the two states of the independent variable (0 or 1) more often than expected by chance [142, 143].

Logistic regression has previously been used to identify effects contributing to certain traits additionally to the analysed SNPs. Such effects could be gender, age or diet, which can affect a trait differently [144, 145]. Logistic regression has also been used to investigate different models which could be more appropriate to analyse GWAS [146]. All of the mentioned logistic regression analyses were performed using genotypes of individuals to analyse the effect of SNPs on a trait. In contrast to more traditional regression analyses investigating the association of SNP alleles with certain traits, we used the information on whether or not a SNP overlapped with a particular annotation. The aim of our research was, as in the sampling and permutation chapter, to identify genomic annotations that showed any evidence for enrichment or depletion of trait-associated SNPs. A conceptually similar approach was applied in a study investigating the location of eQTLs within the genome to see if eQTLs were most likely to coincide with transcription factor binding sites [147]. We further decided to explore distance to TSS in greater detail following [72], so we analysed the number of nucleotides as a quantitative variable. This means that the distance to TSS was not binary, but a continuous measure.

This chapter focuses on how the logistic regression method is applied to our available datasets. In all logistic regression models, we use genomic annotations as variables in the model, where the total information carried by a genomic annotation is the number of linkage disequilibrium (LD) blocks it overlaps with. First, we performed a univariate logistic regression to compare the results with the permutation results. The univariate approach is comparable to the permutation and sampling analyses, as only one genomic annotation is analysed at one time without information on other annotations. The results of the individual annotation analysis are compared with the results from the permutation analysis to identify possible biases in either approach.

We used a stepwise approach to identify multivariate logistic regression models where the smallest set of genomic annotations explains the maximum amount of information. The variables modelled are the genomic annotations, as in the univariate model. The identification of a smaller set of the entire set of analysed annotations was achieved by a stepwise approach explained later in the method section, which calculated the amount of information carried by a set of genomic annotations at each step. A decision was made to either include or remove a genomic annotation based on the amount of information each annotation carried. The information criterion used for this analysis was the Akaike's Information Criterion (AIC) [148], which calculates the information carried by a multiple variable model. All genomic annotations, which coincided with each other and therefore did not provide additional information, were removed from the model. This method identified the influence of the genomic annotations relative to and in combination with each other. The results of these analyses could be used further to calculate a prioritization score for newly discovered GWAS variants, which could help choose trait-associated SNPs for follow-up studies.

The result of the analysis was a defined set of genomic annotations deemed most important and informative, as judged by the decrease in the AIC value during the analysis and the *P*-value of the annotation in the final model. This set of genomic annotations was influential in explaining trait-association status of a broad range of phenotypes allowing drawing of general conclusions. However, this did not allow trait-specific conclusions. We therefore additionally investigated a number of trait-specific subsets. The traits were divided using data from the GaD database, into disease traits (e.g., schizophrenia or diabetes) or normal variation traits (e.g., height or eye colour), and immune traits or cancer traits. The association of the SNPs to these traits defined the following subsets: Disease SNPs, Normal Variation SNPs, Immune SNPs and Cancer SNPs, respectively. The stepwise logistic regression approach was applied to these datasets, resulting in models containing different genomic annotations with different effects. The effects were calculated as the weight or estimate of the annotation in the model and were used to calculate their odds ratios and confidence intervals. We identified a common set of genomic annotations influential to both Normal Variation SNPs and Disease SNPs. The Immune SNPs and Non-immune SNPs showed the most significant differences in the obtained odds of the same genomic annotations, while the Cancer SNPs had the least number of influential annotations. We discuss these findings in the discussion section of this chapter.

5.2 Method

Logistic regression was applied to model genomic annotations as variables that influenced the trait-association status of SNPs, where the models could include either a single variable or multiple variables. The single variable or univariate analysis only included the individual genomic annotation under investigation, onto which trait-association status was regressed. All regression analyses were performed using the function glm available in R with the option family=binomial("logit"). The stepwise analysis determining the most influential genomic annotations was performed using the stepAIC function available in the R package 'MASS' [149].

The trait-association status, or dependent variable, was binary (one and zero, where trait-association was coded as one). The genomic annotations were used as independent explanatory variables onto which the trait-association status was regressed. The independent variables contained information on the presence or absence of the annotated feature at a given location (coded as one and zero, respectively), taking into account the analysed SNPs and their LD partners. All analyses used the SNPs and genomic annotations described in the permutation chapter (Chapter 4) combined into one list of 3,840,944 SNPs. The summary of each final model included the estimated coefficients, their standard errors, the β-coefficients and the P-values of each variable (genotyping array – see section 5.2 – or genomic annotation) in the model. The β-coefficient is defined as the ratio of the estimate and its standard error and is used to calculate the significance as a P-value. The P-values in the multivariate models were not corrected for multiple testing. The values for the intercepts are only shown in the tables to show the complete model. The calculations for the odds ratio and confidence intervals for the genomic annotations were the standard calculations: the exponent of the estimate for the odds ratio and the exponents of (Estimates ± 1.96* Standard Error of Estimate) for the confidence intervals.

5.2.1 Stepwise multivariate logistic regression

All genotyping arrays were included as explanatory variables in the ‘Base model’, because genotyping arrays influence trait-association status, as discussed later. A Base model was needed as a starting point for the stepwise logistic regression and is shown below. This model was fixed and the variables included in this model were not subject to the ex- or inclusion of the step-wise approach.

Base Model: Trait-association Status ~ Affymetrix_250k_Nsp + Affymetrix_250k_Sty + Affymetrix_5.0 + Affymetrix_6.0 + Affymetrix_10k + Affymetrix_50k.1 + Affymetrix_50k.2 + Illumina_300 + Illumina_550 + Illumina_650 + Perlegen

If a stepwise regression analysis only focused on the inclusion of informative variables, the direction specification would be "forward". The "backward" direction focuses only on the exclusion of non-informative variables, or annotations. However, both of these directions have their drawbacks and benefits and so the "both" direction was used, which analysed the possible exclusion and inclusion of annotations at every step and combines the benefits of the other two directions without including their drawbacks. This ex- or inclusion of annotations in the model was based on a reduction of the Akaike's Information Criterion (AIC), which we will define shortly. The additional annotations that were considered for the ex- or inclusion were specified in the experimental model shown below. The genotyping arrays were not considered for the process, as they were included in the Base model. The results were robust to the annotation order that were presented in the experimental model, as the "both" direction started with the Base model and only added the informative annotations.

Experimental model: Trait-association Status ~ Affymetrix_250k_Nsp + Affymetrix_250k_Sty + Affymetrix_5.0 + Affymetrix_6.0 + Affymetrix_10k + Affymetrix_50k.1 + Affymetrix_50k.2 + Illumina_300 + Illumina_550 + Illumina_650 + Perlegen + 1 Kb TSS + 5 Kb TSS + CpG Islands + ORegAnno + vega Genes + Exons + Intronic SNPs + Non.Syn. SNPs (UCSC) + Coding SNPs (UCSC) + Syn. SNPs UCSC) + Gained Stops + 3'UTR + 5'UTR + RNA Genes + ncRNA + TS miRNA + eQTLs + vega PseudoGenes + Intergenic SNPs + DNase Clusters + Insulators (sequence) + EvoFold + Pos. Sel. Genes + Enhancers (sequence) + Exapted Repeats + PREMOD + tfbs Conserved + Indel Pure regions + 17 spec. algmt + 28 spec. algmt plc.mmls + 28 spec. algmt + 44 spec. algmt + 44 spec. algmt plc.mmls + 44 spec. algmt prim. + Negative (sequence) + Open Chromatin + Closed Chromatin + Active Promoter + Weak Promoter + Poised Promoter + Strong Enhancer (proximal) + Strong Enhancer (distal) + Weak Enhancer (proximal) + Weak Enhancer (distal) + Insulator + Txn Transition + Txn Elongation + Weak Txn + Repressed + Heterochrom/lo + Repetitive/CNV (proximal) + Repetitive/CNV (distal)

All analyses were performed in R with the available R package MASS [149]. The full code for the stepAIC function is stepAIC(direction="both", scope = list (upper = taspsfull, lower = empty), empty), where "taspsfull" is the Experimental model containing all genomic annotations and "empty" is the Base model containing only the genotyping arrays. The AIC was calculated to maximise the amount of variance explained by a model with the minimal number of genomic annotations by penalising any increase in the number of included variables. The AIC is defined as $AIC = 2k - 2\ln(L)$, where k is the number of parameters in the model and L is the maximised value of the likelihood function for the estimated model, if a parameter were to be included. The stepwise logistic regression method is an iterative process. This meant that after each in- or exclusion of a genomic annotation, the AIC was recalculated for each of the genomic annotations to evaluate the next step. This could be the inclusion of an additional annotation, or the exclusion of an annotation already in the model. The stepwise logistic regression process halted, when the AIC increased rather than decreased with additional variables. The final model of the logistic regression was the one with the smallest AIC value, resulting in a model balancing the maximum amount of information explained by the minimum number of variables.

As mentioned in the previous chapters, we had included the annotation of the OMIM morbid regions, which are defined as trait-associated regions of the genome as a positive control for GWAS hits. We did not include them in the logistic regression analysis, as they would skew the final model. This would occur because of the penalizing of any additional variables that might add extra information. Another annotation excluded from the analysis was the OMIM genes, as we had two annotations defining genes and kept only one dataset to avoid knowingly using redundant information in the logistic regression analysis.

5.2.2 Pseudo-r^2 values

In order to evaluate a regression model, a 'goodness-of-fit' parameter is assessed. In linear regression, this parameter is the r^2 value, which is defined as

$r^2 = 1 - \frac{\sum_{i-1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i-1}^{N} (y_i - \bar{y}_i)^2}$. In this equation, *N* is the number of observations within the model and *y* is its dependent variable, $\bar{y}$ is the mean of all the values in the dependent variable and $\hat{y}$ is the value predicted by the model [150, 151]. The r^2 value of linear regression can be interpreted in three different ways: First, it can be seen as the amount of variation the model explains [150, 151]. Second, it can be seen as an improvement of the full model over the null model. The third interpretation gives this value its name, as it is the square of the correlation between the model's predicted values and the actual analysed values [150, 151]. Logistic regression models do not have an r^2 value as such, and there is no generally agreed upon analogous value [152]. However, a value, which can have a similar interpretation to the r^2 value of linear regression, can be estimated in a variety of ways and is called a pseudo-r^2 value. Here, we chose the McKelvey and Zavoina's pseudo-r^2. The McKelvey and Zavoina's pseudo-r^2 was defined as the ratio of the variance of a predicted continuous latent variable, which is underlying the binary dependent variable (here: trait-association status), and the sum of that variance and an estimated error. This error variance is assumed to be $\pi^2/3$ in logistic models [150, 151, 153]. The McKelvey and Zavoina's pseudo-r^2 is to be interpreted as the amount of variation explained in a model. The R package descr (function LogRegR2) [154] was used to calculate the pseudo-r^2 values. The code for this function can be found in the Appendix (see page 222).

5.3 Results

5.3.1 Significant SNPs (2011)

5.3.1.1 Univariate regression vs. permutations

The Significant SNPs (2011) dataset was described previously (see page 32 for further information). Figure 5-1 compares the results obtained by the univariate logistic regression and the permutations for Significant SNPs (2011). Overall, the two methods obtained very similar odds ratios. A standard paired t-test showed that the odds ratios were not significantly different between the methods (*P*-value = 0.31, mean of the differences = 0.12, 95% confidence intervals = -0.12 – 0.36). When testing the standard errors of the odds ratios, the t-test was also not significant (*P*-value = 0.08, mean of the differences = -5.25, 95% confidence intervals = -11.22 – 0.72). The t-test was used on the results from 54 of 58 annotations, as the standard errors for four annotations (within miRNA, splice sites, lost stops, and microsatellites) were undefined. As mentioned before, undefined confidence intervals arose when the value at the 500th rank of the permutation overlaps was zero. The correlation coefficient of the odds ratios obtained by both methods was very high at 0.93. The r^2 of the linear regression of the odds ratios from the logistic regression onto the odds ratios from the permutations was 0.87 and a significant *P*-value of 1.03×10^{-26}. The odds ratios are shown in Figure 5-1, where the highest, albeit not significant, odds ratios in both analyses were obtained for TS miRNA and gained stop codons. The permutations obtained significant odds ratios twice, when the univariate logistic regression did not (5'UTRs (top panel in Figure 5-2) and insulators (bottom panel in Figure 5-2)). The logistic regression had three significant odds ratios, where the permutations did not reach significance (gained stops, intergenic SNPs (top panel) and negative (middle panel)). The results of the univariate logistic regression are shown in Table 5-1 listing the estimate, its standard error, the β-coefficient (defined as the ratio of the estimate and its standard error), the calculated odds ratio and its confidence interval and the *P*-value of each analysed annotation.

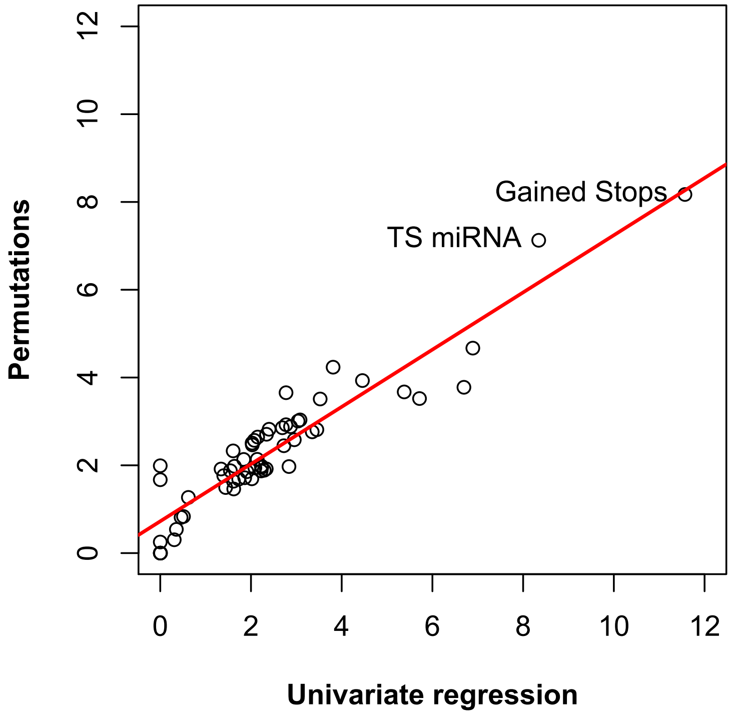


Figure 5-1 – Odds ratios of univariate logistic regression *vs.* permutations
The odds ratios obtained for the Significant SNPs (2011) by a univariate regression model and the odds ratios of the permutations are plotted in this figure. The odds ratios obtained by these two methods correlated well (correlation = 0.93). The adjusted r^2 of the regression shown in this figure is 0.87 with a significant *P*-value of 1.03×10^{-26}.

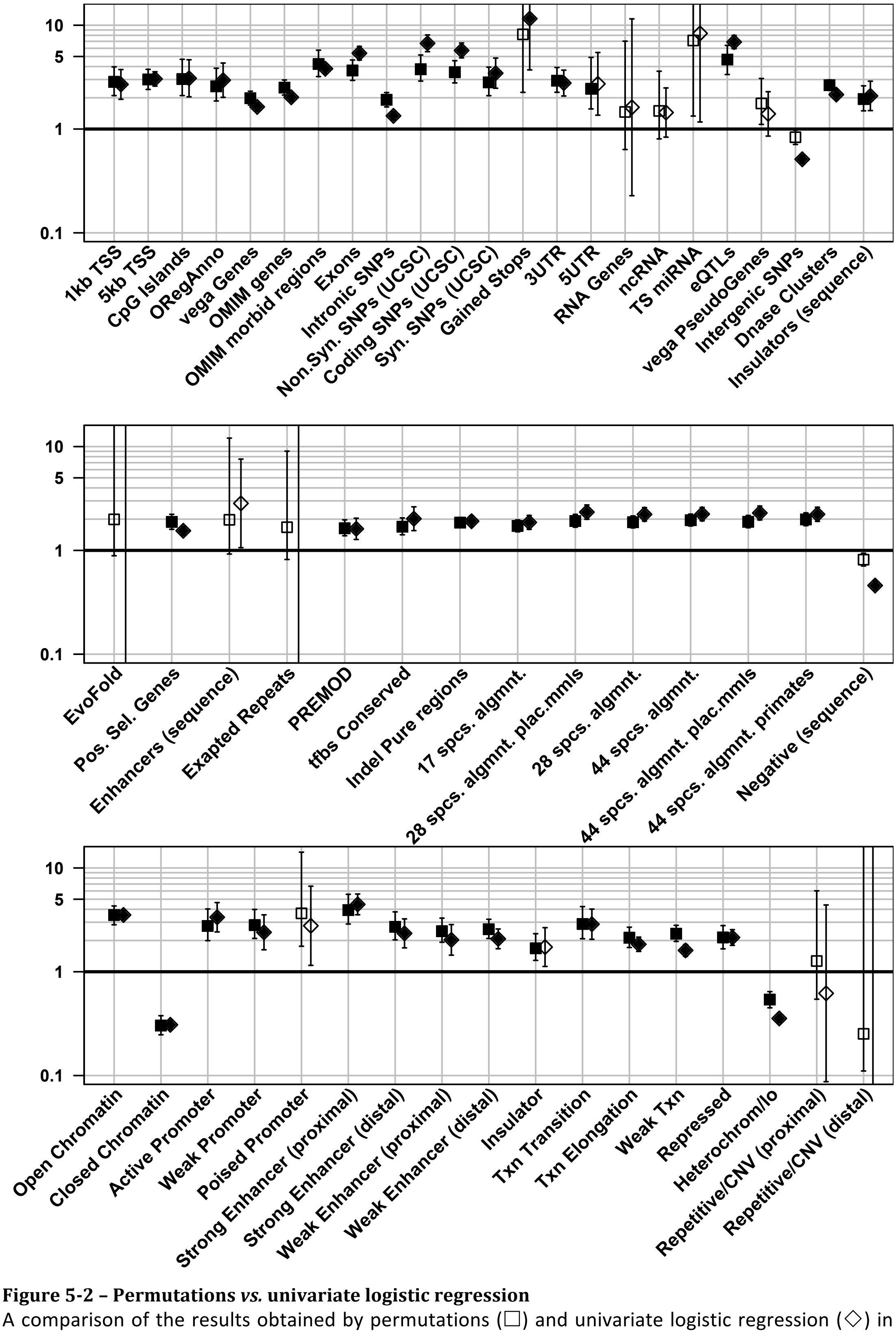


Figure 5-2 – Permutations *vs.* univariate logistic regression

A comparison of the results obtained by permutations (□) and univariate logistic regression (◇) in Significant SNPs (2011). All *P*-values were corrected for multiple testing for the analysed genomic annotations and solid symbols indicated significance at that level. Top: Genic and regulatory regions. Middle: Conserved regions and evolutionary signatures. Bottom: Chromatin states and histone modifications.

Table 5-1 – Univariate logistic regression for Significant SNPs (2011)

The results for the univariate logistic regression for the Significant SNPs (2011) are shown below. The table presents the estimate of the effect, its standard error, the β-coefficient, the calculated odds ratio with its confidence interval, and the *P*-value of the estimate for each of the individual genomic annotations. Significant *P*-values in bold.

Annotation	Estimate	Std. Error	β	OR [LCI-HCI]	*P*-value
1 Kb TSS	0.99	0.17	5.88	2.69 [1.93-3.74]	**4.13×10^{-09}**
5 Kb TSS	1.11	0.08	13.71	3.04 [2.59-3.56]	**8.91×10^{-43}**
CpG Islands	1.13	0.21	5.36	3.08 [2.04-4.65]	**8.26×10^{-08}**
ORegAnno	1.08	0.19	5.59	2.96 [2.02-4.32]	**2.32×10^{-08}**
vega Genes	0.49	0.05	10.70	1.63 [1.49-1.79]	**1.06×10^{-26}**
OMIM genes	0.71	0.05	15.33	2.02 [1.85-2.22]	**5.19×10^{-53}**
OMIM morbid regions	1.34	0.06	23.79	3.81 [3.41-4.25]	**3.92×10^{-125}**
Exons	1.68	0.08	21.77	5.38 [4.62-6.26]	**4.27×10^{-105}**
Intronic SNPs	0.29	0.05	6.37	1.34 [1.22-1.47]	**1.87×10^{-10}**
Non.Syn. SNPs (UCSC)	1.90	0.09	20.13	6.69 [5.56-8.06]	**3.83×10^{-90}**
Coding SNPs (UCSC)	1.74	0.08	20.72	5.72 [4.85-6.74]	**2.49×10^{-95}**
Syn. SNPs (UCSC)	1.24	0.17	7.26	3.45 [2.47-4.83]	**3.94×10^{-13}**
Gained Stops	2.45	0.58	4.22	11.57 [3.72-36.01]	**2.39×10^{-05}**
3'UTR	1.02	0.15	6.97	2.77 [2.08-3.69]	**3.18×10^{-12}**
5'UTR	1.00	0.35	2.83	2.73 [1.36-5.46]	4.69×10^{-03}
RNA Genes	0.48	1.00	0.48	1.62 [0.23-11.51]	6.30×10^{-01}
ncRNA	0.37	0.28	1.31	1.44 [0.84-2.49]	1.89×10^{-01}
TS miRNA	2.12	1.00	2.12	8.34 [1.17-59.50]	3.43×10^{-02}
eQTLs	1.93	0.08	25.62	6.89 [5.94-7.99]	**9.26×10^{-145}**
vega PseudoGenes	0.34	0.25	1.34	1.40 [0.86-2.29]	1.81×10^{-01}
Intergenic SNPs	-0.67	0.05	-14.40	0.51 [0.47-0.56]	**5.05×10^{-47}**
DNase Clusters	0.76	0.06	13.07	2.15 [1.91-2.41]	**4.87×10^{-39}**
Insulators (sequence)	0.74	0.17	4.43	2.09 [1.51-2.89]	**9.54×10^{-06}**
Within miRNA	-4.96	83.85	-0.06	0.01 [0.00-1.66 × 10^{69}]	9.53×10^{-01}
Splice Sites	-7.96	74.86	-0.11	0.00 [0.00-1.83 × 10^{60}]	9.15×10^{-01}
Lost Stops	-6.96	101.93	-0.07	0.00 [0.00-5.52 × 10^{83}]	9.46×10^{-01}
Microsatellites	-7.96	68.01	-0.12	0.00 [0.00-2.70 × 10^{54}]	9.07×10^{-01}
EvoFold	-9.96	98.69	-0.10	0.00 [0.00-4.78 × 10^{79}]	9.20×10^{-01}
Pos. Sel. Genes	0.44	0.05	8.84	1.55 [1.40-1.70]	**9.27×10^{-19}**
Enhancers (sequence)	1.04	0.50	2.08	2.84 [1.06-7.57]	3.73×10^{-02}
Exapted Repeats	-9.96	97.16	-0.10	0.00 [0.00-2.39 × 10^{78}]	9.18×10^{-01}
PREMOD	0.48	0.12	3.98	1.61 [1.27-2.04]	**6.94×10^{-05}**
tfbs Conserved	0.70	0.13	5.22	2.02 [1.55-2.63]	**1.78×10^{-07}**
Indels Pure regions	0.65	0.06	10.60	1.91 [1.70-2.15]	**2.83×10^{-26}**
17 specs. algmt.	0.62	0.08	7.85	1.86 [1.59-2.17]	**4.01×10^{-15}**
28 specs. algmt. plac. mmls	0.85	0.08	10.47	2.34 [1.99-2.74]	**1.16×10^{-25}**
28 specs. algmt.	0.80	0.08	10.45	2.22 [1.91-2.58]	**1.45×10^{-25}**
44 specs. algmt.	0.81	0.08	10.63	2.24 [1.93-2.60]	**2.17×10^{-26}**
44 specs. algmt. plac. mmls	0.83	0.08	10.59	2.29 [1.97-2.67]	**3.39×10^{-26}**
44 specs. algmt. primates	0.80	0.08	10.04	2.23 [1.91-2.61]	**9.81×10^{-24}**
Negative (sequence)	-0.78	0.05	-14.96	0.46 [0.41-0.51]	**1.36×10^{-50}**
Open Chromatin	1.26	0.05	23.95	3.53 [3.18-3.91]	**1.01×10^{-126}**
Closed Chromatin	-1.18	0.05	-22.20	0.31 [0.28-0.34]	**3.62×10^{-109}**
Active Promoter	1.21	0.17	7.28	3.35 [2.42-4.64]	**3.42×10^{-13}**
Weak Promoter	0.88	0.20	4.43	2.40 [1.63-3.53]	**9.45×10^{-06}**
Poised Promoter	1.02	0.45	2.28	2.77 [1.15-6.67]	2.28×10^{-02}
Strong Enhancer (proximal)	1.49	0.12	12.75	4.46 [3.54-5.61]	**2.97×10^{-37}**
Strong Enhancer (distal)	0.85	0.16	5.19	2.34 [1.70-3.23]	**2.08×10^{-07}**
Weak Enhancer (proximal)	0.71	0.17	4.09	2.03 [1.44-2.85]	**4.39×10^{-05}**
Weak Enhancer (distal)	0.73	0.11	6.53	2.07 [1.67-2.58]	**6.41×10^{-11}**
Insulator	0.55	0.22	2.49	1.73 [1.12-2.66]	1.26×10^{-02}
Txn Transition	1.05	0.17	6.18	2.87 [2.05-4.01]	**6.48×10^{-10}**
Txn Elongation	0.61	0.08	7.54	1.84 [1.57-2.16]	**4.70×10^{-14}**
Weak Txn	0.47	0.07	7.08	1.61 [1.41-1.83]	**1.45×10^{-12}**
Repressed	0.76	0.09	8.56	2.14 [1.80-2.54]	**1.10×10^{-17}**
Heterochrom/low	-1.04	0.05	-22.60	0.36 [0.32-0.39]	**4.75×10^{-113}**
Repetitive/CNV (proximal)	-0.48	1.00	-0.48	0.62 [0.09-4.40]	6.32×10^{-01}
Repetitive/CNV (distal)	-9.96	90.03	-0.11	0.00 [0.00-2.04 × 10^{72}]	9.12×10^{-01}

5.3.1.2 Multiple variables model

We hypothesized that the prior probability of a SNP found to be trait-associated would change, if it were included on one or more genotyping arrays. A multivariate model showed that genotyping arrays explained a non-negligible amount of information (pseudo-$r^2 = 0.14$), which supported our hypothesis. The genotyping arrays were therefore included in every following analysis. The model including all genotyping arrays was used as a baseline ("Base model") in all analyses. Any additional genomic annotations were added to this model, if they reduced the AIC (see page 102). The genotyping arrays were not removed from the model, as they were part of the Base model and therefore fixed.

As mentioned in the Methods section of this chapter, the stepwise analysis finished when no further variable could be added that explained extra information without incurring a penalty effect for the additional variable (see page 102). The variables analysed here, were the different genomic annotations. The final model contained all genotyping arrays, as outlined above, and all genomic annotations that added non-redundant information in explaining trait-associated variants, as defined by the stepwise procedure based on the change in AIC. The model that was returned at the end of the analysis was called the Final Model, and included the genotyping arrays and the subset of the annotations returned by the stepwise procedure. Figure 5-3 showed the significant genomic annotations of the Final Model for the Significant SNPs (2011). The odds ratio of a genomic annotation was calculated as the exponent of the estimate of the genomic annotation in the model. Its 95% confidence intervals were calculated as the exponents of the sum/difference of the estimate and the product of its standard error and 1.96, which is the approximate value of the 97.5 percentile point of the normal distribution (exp (Estimate ± 1.96*Standard Error)).

The genomic annotations shown in all figures in this chapter were ranked according to decreasing significance in the model, with the most significant annotations on the left. Four of the five significantly depleted genomic annotations were previously significantly enriched in all analyses investigating genomic annotations individually. These annotations were transcriptional elongation, synonymous SNPs, active promoters and 5'UTRs. In a model, which includes several annotations, the estimates are the effects of those annotations accounting for the fact that other annotations are already included in the model. These four annotations were therefore relatively depleted of trait-associated SNPs when compared to the other genomic annotations and once they were included in the model.

The most significant genomic annotations found to influence trait-association status in the dataset of Significant SNPs (2011) were open chromatin, eQTLs, and DNase clusters. These findings will to some extent be discussed at the end of this chapter, but to a greater detail in the Discussion chapter (Chapter 7). The results are also shown in Table 5-2. The McKelvey and Zavoina's pseudo-r^2 value of the Base model for the Significant SNPs (2011) was 0.14 and 0.23 for the final model. The genotyping arrays were not included in the figure below, but were a part of the final model.

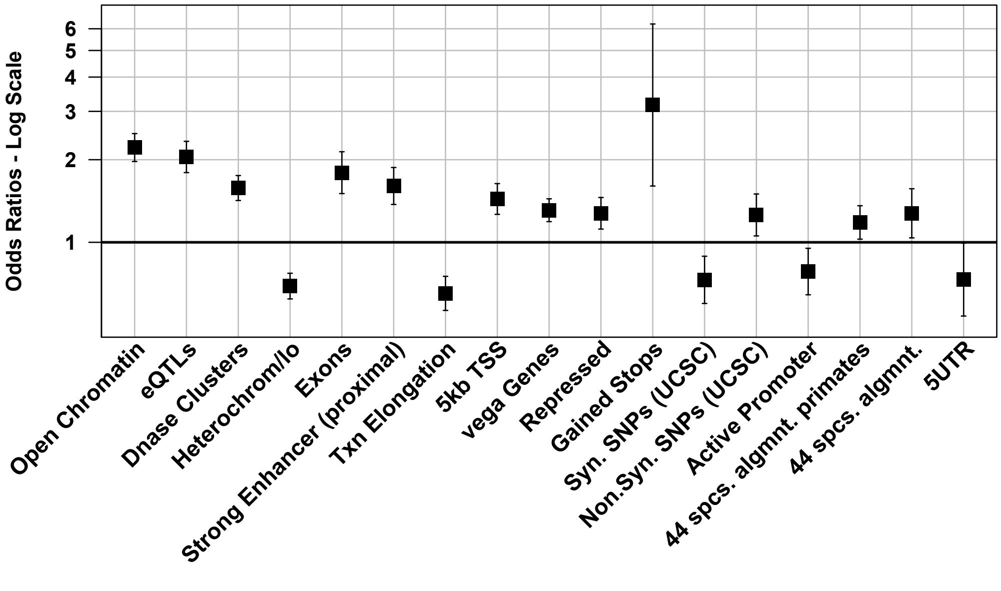


Figure 5-3 – Odds ratios of the final multivariate model for Significant SNPs (2011)
Odds ratios of the significant genomic annotations for the final multivariate model for Significant SNPs (2011), sorted in decreasing significance in the model. The above figure demonstrated that high odds ratio values did not imply a higher significance in the model.

Table 5-2 – Stepwise logistic regression results for Significant SNPs (2011) without Distance to TSS The results for the multivariate model using Significant SNPs (2011) without distance to TSS included in the model are shown below. The table presents the estimate of the effect, its standard error, the β-coefficient, the calculated odds ratio with its confidence interval, and the *P*-value of the estimate in the final model for each of the included genomic annotations in the final model. Significant *P*-values in bold.

Annotation	Estimate	Std. Error	β	OR [LCI-HCI]	*P*-value
Intercept	-9.26	0.08	-116.06	0.00 [0.00-0.00]	0.00×10^{00}
Affymetrix_250k_Nsp	0.33	0.16	2.10	1.40 [1.02-1.91]	**3.59×10^{-02}**
Affymetrix_250k_Sty	0.49	0.15	3.25	1.64 [1.22-2.21]	**1.16×10^{-03}**
Affymetrix_5.0	0.12	0.15	0.79	1.12 [0.84-1.51]	4.31×10^{-01}
Affymetrix_6.0	0.25	0.07	3.55	1.29 [1.12-1.49]	**3.80×10^{-04}**
Affymetrix_10k	0.50	0.30	1.67	1.66 [0.92-2.99]	9.44×10^{-02}
Affymetrix_50k.1	-0.06	0.14	-0.43	0.94 [0.71-1.25]	6.64×10^{-01}
Affymetrix_50k.2	-0.04	0.14	-0.30	0.96 [0.72-1.27]	7.63×10^{-01}
Illumina_300	0.85	0.07	11.44	2.34 [2.02-2.70]	**2.68×10^{-30}**
Illumina_550	1.44	0.20	7.29	4.22 [2.87-6.22]	**3.09×10^{-13}**
Illumina_650	-0.10	0.19	-0.54	0.90 [0.62-1.31]	5.91×10^{-01}
Perlegen	0.21	0.06	3.72	1.23 [1.10-1.38]	**1.99×10^{-04}**
Open Chromatin	0.79	0.06	13.31	2.21 [1.97-2.49]	**2.00×10^{-40}**
Exons	0.58	0.09	6.52	1.79 [1.50-2.14]	**6.94×10^{-11}**
DNase Clusters	0.46	0.05	8.52	1.58 [1.42-1.75]	**1.58×10^{-17}**
eQTLs	0.72	0.07	10.70	2.05 [1.79-2.33]	**1.01×10^{-26}**
Strong Enhancer (proximal)	0.47	0.08	5.98	1.60 [1.37-1.87]	**2.29×10^{-09}**
vega Genes	0.27	0.05	5.51	1.31 [1.19-1.44]	**3.59×10^{-08}**
Repressed	0.24	0.07	3.60	1.28 [1.12-1.46]	**3.19×10^{-04}**
Heterochrom/low	-0.37	0.05	-6.69	0.69 [0.62-0.77]	**2.26×10^{-11}**
Txn Elongation	-0.43	0.07	-5.87	0.65 [0.56-0.75]	**4.26×10^{-09}**
5 Kb TSS	0.36	0.07	5.53	1.44 [1.26-1.64]	**3.26×10^{-08}**
Non.Syn. SNPs (UCSC)	0.23	0.09	2.56	1.26 [1.05-1.50]	**1.06×10^{-02}**
44 specs. algmt. primates	0.17	0.07	2.33	1.18 [1.03-1.36]	**2.00×10^{-02}**
Active Promoter	-0.25	0.10	-2.47	0.78 [0.64-0.95]	**1.37×10^{-02}**
Gained Stops	1.15	0.35	3.32	3.16 [1.60-6.25]	**8.95×10^{-04}**
Syn. SNPs (UCSC)	-0.32	0.10	-3.12	0.73 [0.60-0.89]	**1.83×10^{-03}**
Indels Pure regions	0.11	0.06	1.76	1.11 [0.99-1.25]	**7.79×10^{-02}**
5’UTR	-0.31	0.16	-2.01	0.73 [0.54-0.99]	**4.50×10^{-02}**
Repetitive/CNV (distal)	-1.41	1.00	-1.41	0.24 [0.03-1.73]	1.58×10^{-01}
Poised Promoter	0.37	0.20	1.87	1.44 [0.98-2.12]	6.14×10^{-02}
Insulator	-0.18	0.10	-1.80	0.83 [0.68-1.02]	7.24×10^{-02}
Enhancers (sequence)	0.53	0.29	1.80	1.69 [0.95-3.00]	7.22×10^{-02}
Weak Enhancer (distal)	0.10	0.06	1.52	1.10 [0.97-1.25]	1.28×10^{-01}
44 specs. algmt.	0.24	0.11	2.31	1.28 [1.04-1.57]	2.07×10^{-02}
44 specs. algmt. plac. mmls	-0.20	0.11	-1.80	0.82 [0.66-1.02]	7.14×10^{-02}
TS miRNA	0.83	0.51	1.61	2.29 [0.83-6.28]	1.08×10^{-01}

As mentioned in the Introduction, we decided to explore distance from the trait-associated variant to the nearest transcription start site (TSS) to a greater detail, so we added a quantitative variable to the analyses: “Distance to TSS”. Previous analyses had suggested that the majority of eQTLs were within a 20 Kb window from the nearest TSS [72] and that the majority of trait-associated variants were eQTLs [110]. An additional analysis was performed using Significant SNPs (2011) to determine if the variable explained a significant amount of variation in trait-association status. Figure 5-4 shows the odds ratios obtained in the model for Significant SNPs (2011), which included the distance to TSS annotation. The five annotations, which were depleted for Significant SNPs (2011) in the model without distance to TSS, still have odds ratios of depletion. The odds ratios did not change significantly between the models with a correlation coefficient of 0.99 between the odds ratios of the common annotations of the two models. The model with the distance to TSS further included an additional five annotations (negative (sequence), splice sites, PREMOD, microsatellites, and lost stop codons), which were now adding extra information to the model. The annotation, which was not included in the model with distance to TSS but was included in the model without distance to TSS was the TS miRNA annotation. The biggest difference between the models was the change in the pseudo-r^2 value of the Final Model, which increased from 0.23 for the model without distance to TSS to 0.42 for the model including distance. The pseudo-r^2 therefore almost doubled with inclusion of the distance to TSS. The distance to TSS annotation obtained a very small estimate in the model, indicating that as the distance between a TSS and a SNP increases, the odds that this SNP is a trait-associated SNP decreases. However, the distance is measured in single bases rather than kilo bases, so the effect of increasing the distance would be very small. Distance to TSS was, however, very significant, which means that while the effect is small, it is very important in explaining trait-association status. Since the inclusion of the distance to TSS had such a large impact on the pseudo-r^2 value of the Significant SNPs (2011) model, the rest of the regression models were all performed with the inclusion of this annotation and only those models will be discussed in the remainder of the thesis.

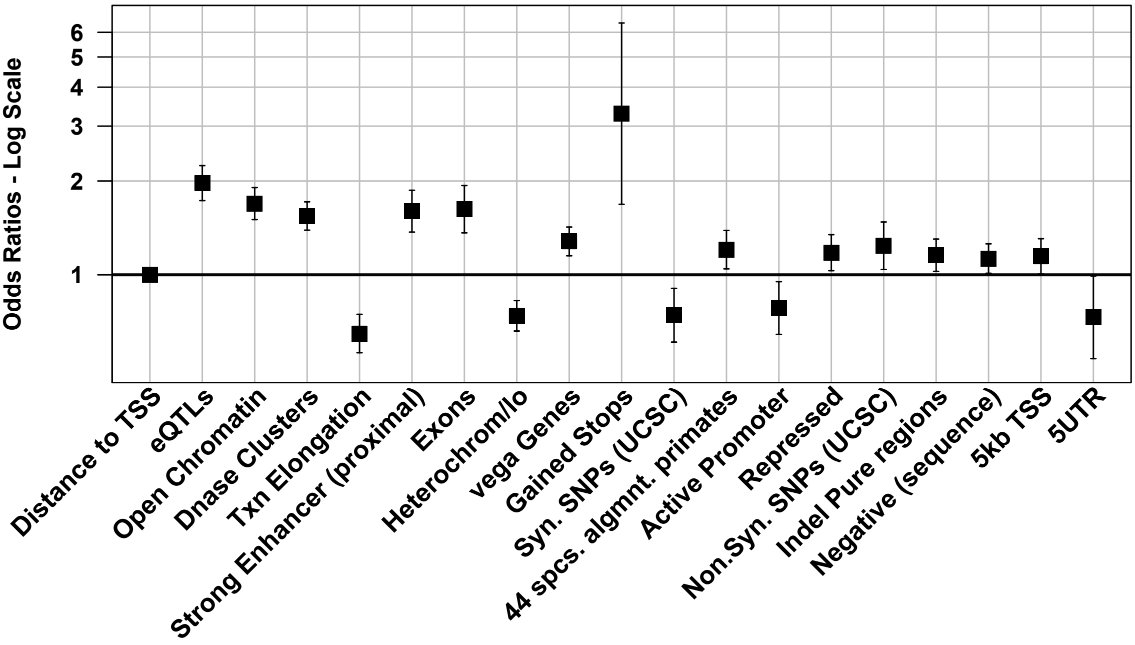


Figure 5-4 – Odds ratios of the final multivariate model for Significant SNPs (2011) including Distance to TSS

Odds ratios of the significant genomic annotations for the final multivariate model for Significant SNPs (2011), sorted in decreasing significance in the model including distance to TSS. Only the significant results are shown in this graph.

Table 5-3 – Stepwise logistic regression results for Significant SNPs (2011) with Distance to TSS
The results for the multivariate model using Significant SNPs (SNPs) with distance to TSS are shown below. The table presents the estimate of the effect, its standard error, the β-coefficient, the calculated odds ratio with its confidence interval, and the *P*-value of the estimate in the final model for each of the included genomic annotations in the final model. Significant *P*-values in bold.

Annotation	Estimate	Std. Error	β	OR [LCI-HCI]	*P*-value
Intercept	-8.79	0.09	-100.04	0.00 [0.00-0.00]	$0.00 \times 10^{+00}$
Affymetrix_250k_Nsp	0.35	0.16	2.17	1.41 [1.03-1.93]	**2.96×10^{-02}**
Affymetrix_250k_Sty	0.48	0.15	3.17	1.62 [1.20-2.18]	**1.53×10^{-03}**
Affymetrix_5.0	0.12	0.15	0.82	1.13 [0.84-1.51]	4.11×10^{-01}
Affymetrix_6.0	0.25	0.07	3.48	1.28 [1.11-1.48]	**5.10×10^{-04}**
Affymetrix_10k	0.50	0.30	1.66	1.65 [0.91-2.98]	9.67×10^{-02}
Affymetrix_50k.1	-0.04	0.14	-0.25	0.96 [0.73-1.28]	8.02×10^{-01}
Affymetrix_50k.2	-0.01	0.14	-0.08	0.99 [0.74-1.31]	9.37×10^{-01}
Illumina_300	0.84	0.07	11.30	2.31 [2.00-2.67]	**1.33×10^{-29}**
Illumina_550	1.43	0.20	7.25	4.18 [2.84-6.16]	**4.11×10^{-13}**
Illumina_650	-0.10	0.19	-0.52	0.91[0.63-1.31]	6.00×10^{-01}
Perlegen	0.20	0.06	3.57	1.22 [1.09-1.37]	**3.63×10^{-04}**
Distance to TSS	0.00	0.00	-12.18	1.00 [1.00-1.00]	**4.15×10^{-34}**
DNase Clusters	0.43	0.05	8.14	1.54 [1.39-1.71]	**4.04×10^{-16}**
eQTLs	0.68	0.07	10.28	1.97 [1.73-2.24]	**9.05×10^{-25}**
Open Chromatin	0.53	0.06	8.72	1.69 [1.50-1.91]	**2.68×10^{-18}**
Exons	0.48	0.09	5.44	1.62 [1.36-1.93]	**5.41×10^{-08}**
Strong Enhancer (proximal)	0.47	0.08	5.97	1.60 [1.37-1.87]	**2.44×10^{-09}**
Txn Elongation	-0.43	0.07	-5.99	0.65 [0.56-0.75]	**2.06×10^{-09}**
44 specs. algmt. primates	0.19	0.07	2.57	1.20 [1.05-1.39]	**1.02×10^{-02}**
Heterochrom/lo	-0.30	0.06	-5.27	0.74 [0.66-0.83]	**1.34×10^{-07}**
vega Genes	0.25	0.05	4.55	1.28 [1.15-1.42]	**5.25×10^{-06}**
Indels Pure regions	0.14	0.06	2.38	1.16 [1.03-1.30]	**1.74×10^{-02}**
Syn. SNPs (UCSC)	-0.30	0.10	-2.95	0.74 [0.61-0.90]	**3.19×10^{-03}**
Repressed	0.16	0.07	2.41	1.18 [1.03-1.35]	**1.60×10^{-02}**
Gained Stops	1.19	0.34	3.48	3.29 [1.68-6.43]	**4.93×10^{-04}**
Negative (sequence)	0.12	0.06	2.17	1.13 [1.01-1.26]	**3.03×10^{-02}**
Non.Syn. SNPs (UCSC)	0.21	0.09	2.40	1.24 [1.04-1.48]	**1.63×10^{-02}**
Active Promoter	-0.25	0.10	-2.47	0.78 [0.64-0.95]	**1.35×10^{-02}**
5 Kb TSS	0.14	0.07	2.04	1.15 [1.01-1.31]	**4.12×10^{-02}**
Splice Sites	-11.80	151.99	-0.08	0.00 [0.00-1.78 × 10^{+124}]	9.38×10^{-01}
PREMOD	0.14	0.07	1.89	1.15 [0.99-1.32]	5.92×10^{-02}
5'UTR	-0.31	0.16	-2.01	0.73 [0.54-0.99]	**4.42×10^{-02}**
Microsatellites	-10.99	109.51	-0.10	0.00 [0.00-2.79 × 10^{+88}]	9.20×10^{-01}
Insulator	-0.20	0.10	-1.92	0.82 [0.67-1.00]	5.52×10^{-02}
Repetitive/CNV (distal)	-1.38	1.00	-1.37	0.25 [0.04-1.80]	1.69×10^{-01}
Enhancers (sequence)	0.55	0.29	1.86	1.73 [0.97-3.07]	6.31×10^{-02}
44 specs. algmt.	0.25	0.11	2.35	1.28 [1.04-1.57]	**1.86×10^{-02}**
44 specs. algmt. plac. mmls	-0.19	0.11	-1.70	0.83 [0.67-1.03]	8.91×10^{-02}
Poised Promoter	0.31	0.20	1.57	1.36 [0.93-2.00]	1.16×10^{-01}
Weak Enhancer (distal)	0.10	0.06	1.48	1.10 [0.97-1.25]	1.38×10^{-01}
Lost Stops	-11.95	235.66	-0.05	0.00 [0.00-2.56 × 10^{+195}]	9.60×10^{-01}

5.3.2 Suggestive SNPs (2011)

Additionally to significantly associated SNPs we investigated SNPs with *P*-values of association that did not pass the genome-wide significance threshold. This dataset presumably contains a mixture of spurious associations and true signals, which did not pass the threshold due to insufficient sample size in their GWA study. The stepwise logistic regression analysis resulted in not only fewer genomic annotations included in the model for Suggestive SNPs (2011), but the included genomic annotations also obtained odds ratios with less extreme values when compared to the results obtained for Significant SNPs (2011). The results for the Suggestive SNPs (2011) are included in Table 5-4. Figure 5-5 shows the eight genomic annotations, which were significant in the Suggestive SNPs (2011) analysis. Four of these annotations obtained significantly different odds ratios. These four were eQTLs, exons, open chromatin, and distance to TSS. The McKelvey and Zavoina's pseudo-r^2 value of the Base model for the Suggestive SNPs (2011) was 0.16, and 0.18 for the final model. For suggestive SNPs the amount of variance explained by the genomic annotations was therefore very little in comparison with that added by the genotyping arrays.

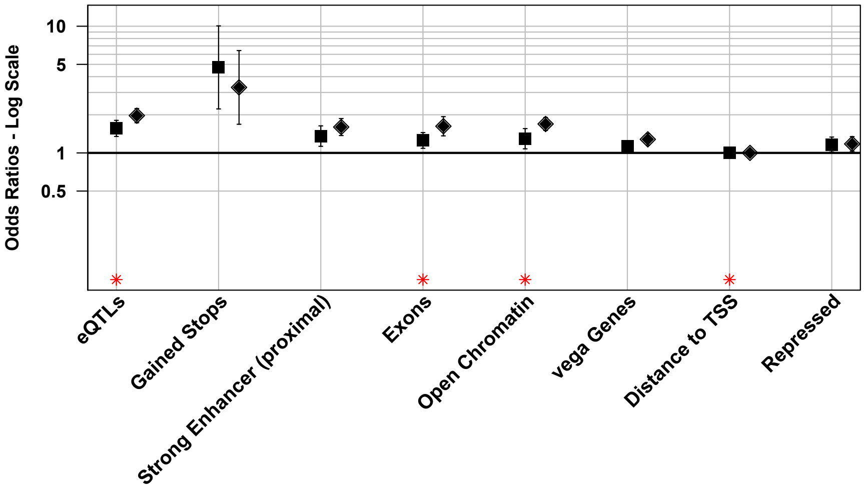


Figure 5-5 – Genomic annotations for Significant SNPs *vs.* Suggestive SNPs (2011)
Odds ratios for all significant genomic annotations in the Suggestive SNPs (2011) model present in the Significant SNPs (2011) sorted after significance for the suggestive SNPs. Suggestive SNPs (2011) are shown as □ and Significant SNPs (2011) are shown as ◇. Solid symbols indicate significance at *P*-value ≤ 0.05. Red stars (✳) indicate significant differences between the two datasets.

Two histograms of the frequency of the Significant SNPs (2011) and Suggestive SNPs (2011) are shown in Figure 5-6 and Figure 5-7, where the former shows the region up to 20 Kb and the latter shows the frequencies of the variants further away. Both of these figures were adapted from a supplementary figure from Kindt *et al.* published in 2013 [139]. These figures show the distribution of trait-associated variants and highlight what the Distance to TSS odds ratio already indicated. The majority of the significantly trait-associated variants was close to a transcription start site and outnumber the suggestively trait-associated variants up until a distance of < 11 Kb. The suggestively associated variants were most often located further away, as seen in Figure 5-7.

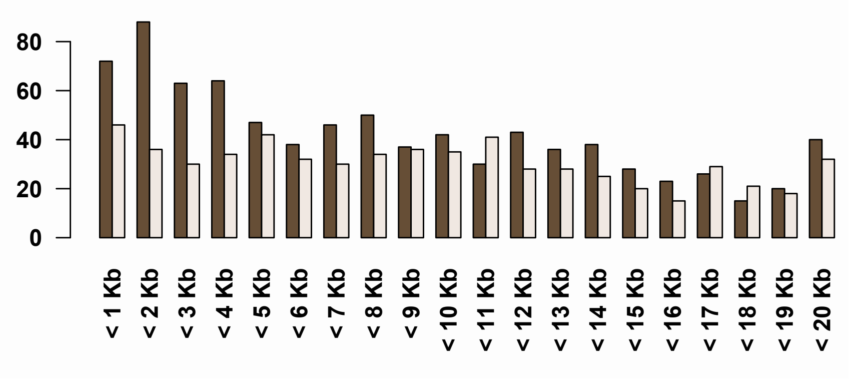


Figure 5-6 – Histogram of distance to TSS of significant and suggestive variants (< 20 Kb)
Significant SNPs (2011) are more frequent in the areas closer to the TSS. This trend is most obvious in the first 10 windows, but is prevalent throughout. Significant SNPs (2011) shown in filled bars, Suggestive SNPs (2011) shown in open bars.

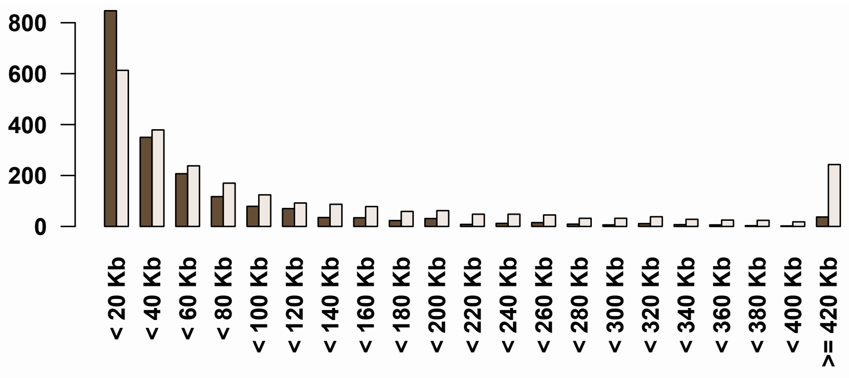


Figure 5-7 - Histogram of distance to TSS of significant and suggestive variants (≤ 420 Kb)
Significant SNPs (2011) are much more frequent regions up to 20 Kb away from the TSS. However, the Suggestive SNPs (2011) are more frequent in all other frequencies. Significant SNPs (2011) shown in filled bars, Suggestive SNPs (2011) shown in open bars.

Table 5-4 – Stepwise logistic regression results for Suggestive SNPs (2011) with Distance to TSS

This table lists the results for Suggestive SNPs (2011) for the multivariate model including the distance to TSS. The estimate of the effect, its standard error, the β-coefficient, the calculated odds ratio with its confidence interval, and the *P*-value of the estimate in the final model for each of the included genomic annotations in the final model are shown below. Significant *P*-values in bold.

Annotation	Estimate	Std. Error	β	OR [LCI-HCI]	*P*-value
Intercept	-8.58	0.10	-85.31	0.00 [0.00-0.00]	$0.00 \times 10^{+00}$
Affymetrix_250k_Nsp	0.74	0.12	5.98	2.09 [1.64-2.66]	**2.21×10^{-09}**
Affymetrix_250k_Sty	0.92	0.12	7.80	2.50 [1.99-3.15]	**6.12×10^{-15}**
Affymetrix_5.0	-0.11	0.11	-0.99	0.89 [0.71-1.12]	3.21×10^{-01}
Affymetrix_6.0	0.15	0.06	2.43	1.16 [1.03-1.32]	**1.51×10^{-02}**
Affymetrix_10k	0.25	0.22	1.14	1.28 [0.84-1.95]	2.53×10^{-01}
Affymetrix_50k.1	0.96	0.08	11.99	2.62 [2.24-3.06]	**3.82×10^{-33}**
Affymetrix_50k.2	0.71	0.09	7.70	2.02 [1.69-2.42]	**1.35×10^{-14}**
Illumina_300	0.75	0.06	11.96	2.12 [1.87-2.39]	**6.02×10^{-33}**
Illumina_550	1.43	0.17	8.43	4.19 [3.00-5.85]	**3.56×10^{-17}**
Illumina_650	-0.03	0.16	-0.17	0.97 [0.71-1.34]	8.65×10^{-01}
Perlegen	0.48	0.05	10.65	1.62 [1.49-1.78]	**1.83×10^{-26}**
eQTLs	0.45	0.07	5.96	1.56 [1.35-1.81]	**2.50×10^{-09}**
Open Chromatin	0.26	0.09	2.75	1.29 [1.08-1.56]	**5.91×10^{-03}**
Exons	0.23	0.07	3.11	1.25 [1.09-1.45]	**1.88×10^{-03}**
Gained Stops	1.55	0.39	4.04	4.73 [2.23-10.07]	**5.40×10^{-05}**
Strong Enhancer (proximal)	0.31	0.09	3.23	1.36 [1.13-1.64]	**1.24×10^{-03}**
Distance to TSS	0.00	0.00	-2.45	1.00 [1.00-1.00]	**1.42×10^{-02}**
vega Genes	0.12	0.05	2.63	1.13 [1.03-1.23]	**8.64×10^{-03}**
Repressed	0.15	0.07	2.20	1.16 [1.02-1.33]	**2.75×10^{-02}**
Microsatellites	1.05	0.45	2.33	2.86 [1.18-6.89]	**1.96×10^{-02}**
Poised Promoter	-0.67	0.41	-1.63	0.51 [0.23-1.14]	1.02×10^{-01}
Closed Chromatin	0.16	0.09	1.68	1.17 [0.97-1.40]	9.21×10^{-02}
Pos. Sel. Genes	-0.08	0.05	-1.53	0.93 [0.84-1.02]	1.25×10^{-01}
TS miRNA	-9.52	66.24	-0.14	0.00 [0.00-$1.77 \times 10^{+52}$]	8.86×10^{-01}
Insulators (sequence)	-0.15	0.10	-1.47	0.86 [0.71-1.05]	1.43×10^{-01}

5.3.3 Significant SNPs (2013)

The more recent catalogue of trait-associated SNPs, Significant SNPs (2013), was analysed as five different sets. First, it was analysed as a complete set of trait-associated SNPs incorporating associations across a broad range of phenotypes. The total set was then separated into subsets depending on different trait-categories to allow trait-specific results and conclusions. The traits in the subsets are listed in the Appendix (page 203). The results for the final model for Significant SNPs (2013) including Distance to TSS are shown in Table 5-5. There were no significant differences between the odds ratios of the Significant SNPs (2011) and Significant SNPs (2013). The odds ratios of the annotations, which were present in both models, are shown in Figure 5-6. While there were no differences in the odds ratios in the common annotations, the two models varied in some of the included annotations. Therefore for Significant SNPs (2011) further contained microsatellites, repetitive/CNV (distal), and 44 species alignment with placental mammals, which were not significant in the model. The Significant SNPs (2013) model also contained RNA genes, positively selected genes, and conserved sites from the 17 species alignment, all of which were significant. The weak enhancer (proximal) and the weak transcription regions were not significant in the model. Those annotations, which were not significant in the model, were included as they explained additional variation, albeit not significantly when compared to all other annotations.

For Significant SNPs (2013) there were 32 included annotations in the Final Model, of which 24 had a standard error of less than two and were significant at ≥0.05 (Figure 5-9). Distance to TSS has an odds ratio of one with a *P*-value of 9.01×10^{-52}, but is in fact depleted with an odds ratio of 0.999. This annotation is a quantitative annotation, as the distance to TSS was included as a linear variable. As discussed above, a change in the distance to TSS by one unit would be the change in one nucleotide, so would have a very small effect. The effect returned by the model is negative for distance to TSS, so with increasing distance the likelihood that a SNP is trait-associated decreases. The most significant annotations in the model are eQTLs, distance to TSS, open chromatin, DNase clusters and exons. The rest of the annotations have *P*-values that are an order of magnitude larger than these four very significant annotations.

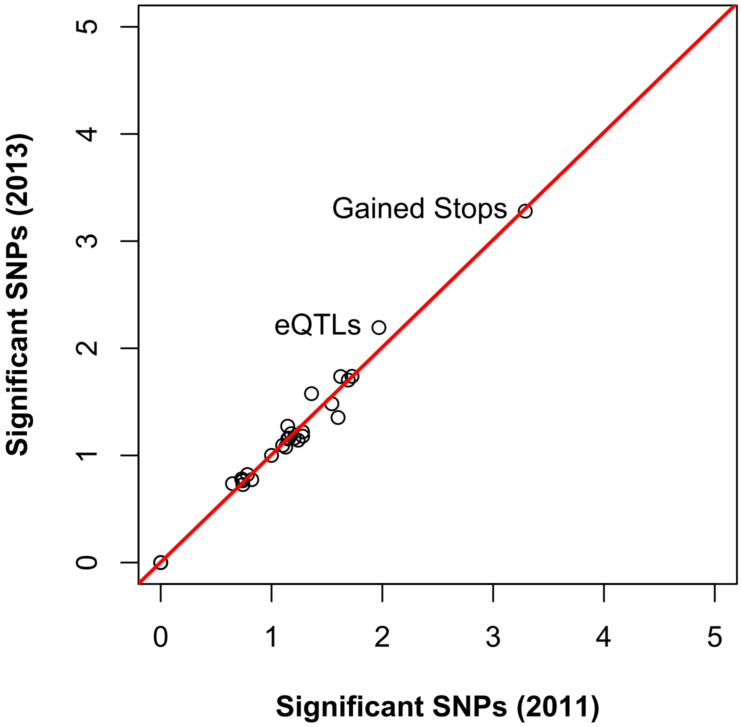


Figure 5-8 – Correlation of odds ratios for Significant SNPs (2011) and Significant SNPs (2013)
The odds ratios of the common genomic annotations for Significant SNPs (2011) and Significant SNPs (2013) agreed very well with each other. The r^2 of the regression line of the results for Significant SNPs (2011) onto Significant SNPs (2013) was 0.97 with a *P*-value of 3.37×10^{-22}.

Additionally to the annotation of "Distance to TSS" there were eight more annotations showing depletion of significantly trait-associated SNPs. The genomic annotations depleted in the dataset of Significant SNPs (2013) are regions with chromatin states associated with heterochromatin/low transcription, transcriptional elongation, insulators, and active promoters and genic regions annotated as synonymous SNPs, RNA genes, 5'UTRs, and regions conserved in a 17 species alignment. The McKelvey and Zavoina's pseudo-r^2 values were 0.12 for the genotyping arrays only, and 0.36 for the model including the Distance to TSS.

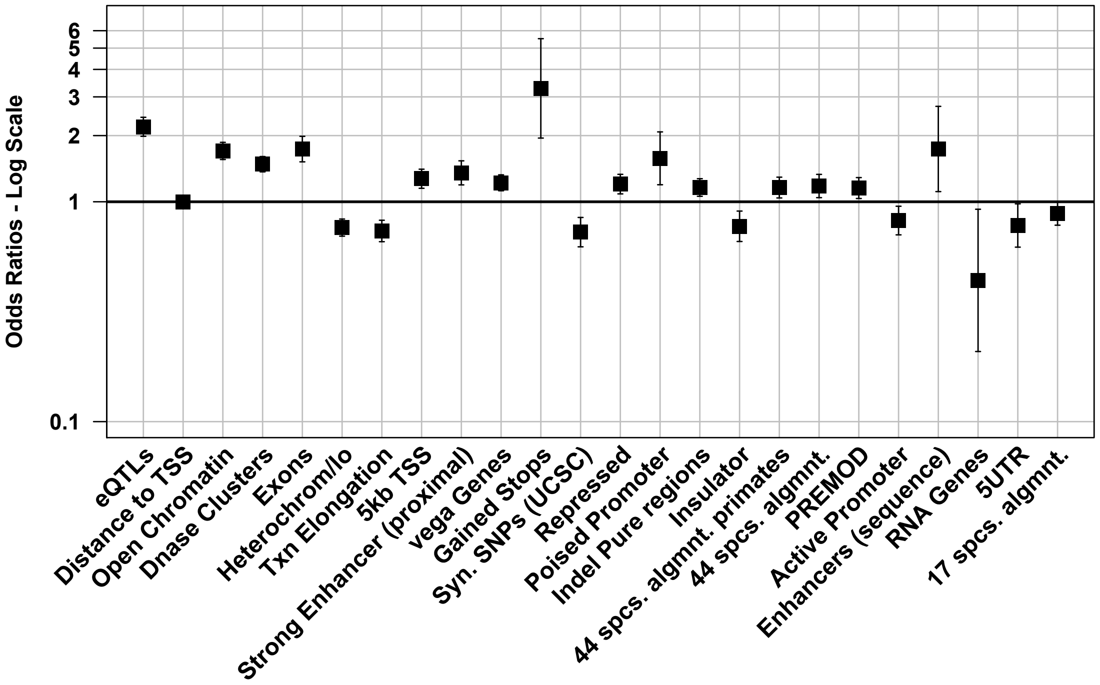


Figure 5-9 – Significant genomic annotations for Significant SNPs (2013)
Odds ratios for all significant genomic annotations in the Significant SNPs (2013) model sorted after significance. A total of 24 genomic annotations were significant and had a standard error of less than two.

Table 5-5 – Stepwise logistic regression results for Significant SNPs (2013)

The results for the multivariate model using Significant SNPs (2013) are shown below. The table presents the estimate of the effect, its standard error, the β-coefficient, the calculated odds ratio with its confidence interval, and the *P*-value of the estimate in the final model for each of the included genomic annotations in the final model. Significant *P*-valus in bold.

Annotation	Estimate	Std. Error	β	OR [LCI-HCI]	*P*-value
Intercept	-8.08	0.07	-121.06	0.00 [0.00-0.00]	$0.00 \times 10^{+00}$
Affymetrix_250k_Nsp	0.36	0.12	3.00	1.43 [1.13-1.81]	**2.73×10^{-03}**
Affymetrix_250k_Sty	0.38	0.12	3.25	1.46 [1.16-1.82]	**1.14×10^{-03}**
Affymetrix_5.0	0.04	0.11	0.34	1.04 [0.83-1.30]	7.31×10^{-01}
Affymetrix_6.0	0.37	0.05	6.95	1.45 [1.30-1.61]	**3.71×10^{-12}**
Affymetrix_10k	0.54	0.25	2.18	1.72 [1.06-2.81]	**2.90×10^{-02}**
Affymetrix_50k.1	-0.08	0.12	-0.67	0.92 [0.74-1.16]	5.00×10^{-01}
Affymetrix_50k.2	-0.17	0.12	-1.37	0.84 [0.66-1.08]	1.71×10^{-01}
Illumina_300	0.66	0.06	11.47	1.94 [1.73-2.17]	**1.94×10^{-30}**
Illumina_550	0.97	0.13	7.33	2.65 [2.04-3.44]	**2.23×10^{-13}**
Illumina_650	0.21	0.13	1.67	1.23 [0.96-1.58]	9.48×10^{-02}
Perlegen	0.26	0.04	6.11	1.30 [1.20-1.42]	**1.03×10^{-09}**
Distance to TSS	0.00	0.00	-15.14	1.00 [1.00-1.00]	**9.01×10^{-52}**
eQTLs	0.79	0.05	15.48	2.19 [1.99-2.42]	**4.81×10^{-54}**
DNase Clusters	0.39	0.04	9.62	1.48 [1.37-1.61]	**6.70×10^{-22}**
Open Chromatin	0.53	0.05	11.62	1.70 [1.56-1.86]	**3.17×10^{-31}**
Exons	0.55	0.07	8.10	1.74 [1.52-1.98]	**5.71×10^{-16}**
Repressed	0.18	0.05	3.54	1.20 [1.09-1.33]	**3.98×10^{-04}**
Indels Pure regions	0.15	0.05	3.19	1.16 [1.06-1.27]	**1.45×10^{-03}**
Strong Enhancer (proximal)	0.30	0.06	4.71	1.35 [1.19-1.54]	**2.44×10^{-06}**
Txn Elongation	-0.31	0.06	-5.34	0.74 [0.66-0.82]	**9.52×10^{-08}**
Heterochrom/lo	-0.27	0.05	-5.92	0.76 [0.70-0.83]	**3.28×10^{-09}**
5 Kb TSS	0.24	0.05	4.73	1.27 [1.15-1.41]	**2.26×10^{-06}**
Syn. SNPs (UCSC)	-0.32	0.08	-4.06	0.73 [0.62-0.85]	**5.00×10^{-05}**
44 specs. algmt. primates	0.15	0.06	2.68	1.16 [1.04-1.29]	**7.38×10^{-03}**
Gained Stops	1.19	0.27	4.47	3.28 [1.95-5.52]	**7.74×10^{-06}**
vega Genes	0.20	0.04	4.67	1.22 [1.12-1.33]	**2.96×10^{-06}**
Insulator	-0.26	0.08	-3.16	0.77 [0.66-0.91]	**1.59×10^{-03}**
Poised Promoter	0.46	0.14	3.22	1.58 [1.20-2.08]	**1.26×10^{-03}**
Active Promoter	-0.20	0.08	-2.57	0.82 [0.71-0.95]	**1.01×10^{-02}**
Splice Sites	-11.35	93.25	-0.12	0.00 [0.00-$2.82 \times 10^{+74}$]	9.03×10^{-01}
PREMOD	0.14	0.06	2.57	1.15 [1.03-1.29]	**1.01×10^{-02}**
5’UTR	-0.25	0.12	-2.15	0.78 [0.62-0.98]	**3.17×10^{-02}**
RNA Genes	-0.82	0.38	-2.17	0.44 [0.21-0.92]	**3.03×10^{-02}**
Enhancers (sequence)	0.55	0.23	2.43	1.74 [1.11-2.72]	**1.52×10^{-02}**
Pos. Sel. Genes	-0.08	0.04	-1.78	0.93 [0.85-1.01]	7.50×10^{-02}
44 specs. algmt.	0.17	0.06	2.66	1.18 [1.04-1.33]	**7.88×10^{-03}**
17 specs. algmt.	-0.12	0.06	-2.03	0.88 [0.78-1.00]	**4.27×10^{-02}**
Non.Syn. SNPs (UCSC)	0.13	0.07	1.91	1.14 [1.00-1.31]	5.64×10^{-02}
Lost Stops	-11.51	143.43	-0.08	0.00 [0.00-$1.23 \times 10^{+117}$]	9.36×10^{-01}
Weak Enhancer (proximal)	-0.12	0.07	-1.83	0.89 [0.78-1.01]	6.76×10^{-02}
Negative (sequence)	0.07	0.04	1.71	1.08 [0.99-1.17]	8.75×10^{-02}
Weak Enhancer (distal)	0.09	0.05	1.71	1.09 [0.99-1.21]	8.70×10^{-02}
Weak Txn	-0.08	0.05	-1.63	0.93 [0.84-1.02]	1.03×10^{-01}

5.3.4 Subsets of Significant SNPs (2013)

The dataset of Significant SNPs (2013) consisted of 3,283 SNPs, which is a sufficient number of SNPs that allowed splitting the set into several subsets divided on the basis of phenotype groupings. This splitting of the dataset allowed exploring if different phenotype classes were affected preferentially by different annotations. The Significant SNPs (Difference) were analysed as well as four SNP sets associated to four specific trait categories (disease traits, normal variation traits, immunity traits, and cancer traits). The trait-association of the SNPs to four trait-categories defined the SNP subsets and were analysed to compare different trait classes. In particular, the comparison of Normal Variation and Disease SNPs, which were two mutually exclusive datasets. The Immune SNPs and the Cancer SNPs were compared to Non-immune and Non-cancer SNPs. The traits in the subsets are listed in the Appendix (page 203).

5.3.4.1 Significant SNPs (Difference)

In order to analyse only the newest SNPs, we investigated the Significant SNPs (Difference) set, which contained only those SNPs that were present in Significant SNPs (2013) but not in Significant SNPs (2011). These SNPs were analysed in a multiple logistic regression model and the results were compared with the results for the Significant SNPs (2011). The 17 annotations common to both regression models showed only three significant differences. The transcriptional elongation regions were less depleted of Significant SNPs (Difference) than for Significant SNPs (2011) while the 5 Kb regions upstream of transcription start sites were more enriched for the Significant SNPs (Difference). The distance to TSS was also significantly different, as judged by their *P*-values obtained from a t-test. Figure 5-8 shows the genomic annotations common to both final models and presents the numerical results of the Significant SNPs (Difference). Table 5-6 lists the estimates, standard errors, β-coefficients (ratio of estimate over standard errors), odds ratios and confidence intervals and the *P*-values of all the genomic annotations, genotyping arrays and intercept of the final model for Significant SNPs (Difference). Table 5-3 contains the results for the Significant SNPs (2011).

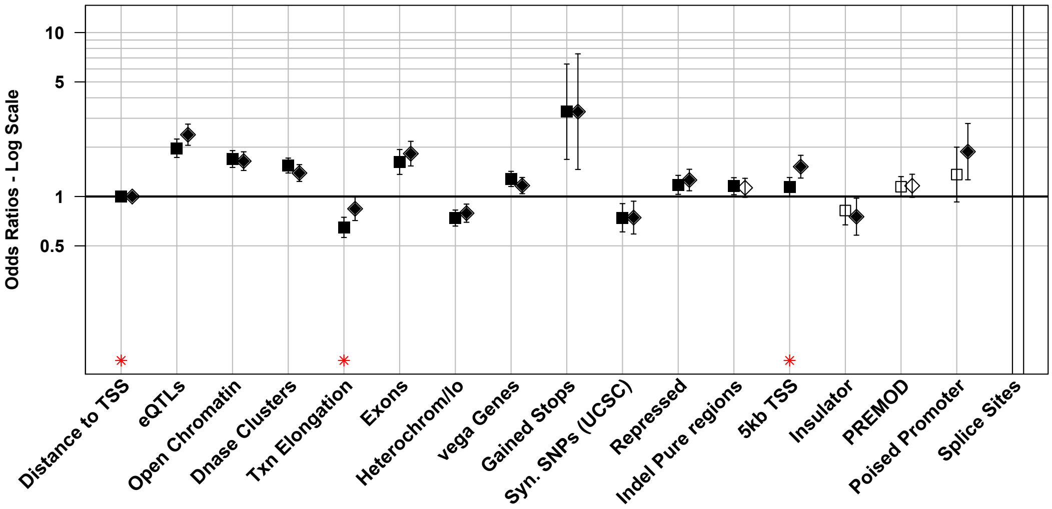


Figure 5-10 – Common genomic annotations for Significant SNPs (2011) and Significant SNPs (Difference)
Odds ratios for all common genomic annotations in the Significant SNPs (2011) and Significant SNPs (Difference) model sorted after significance. Significant SNPs (2011) are shown as □ and Significant SNPs (Difference) are shown as ◇. Solid symbols indicate significance at *P*-value ≤ 0.05. Red stars (✳) indicate significant differences between the two datasets.

Table 5-6 – Stepwise logistic regression results for Significant SNPs (Difference)

The results for the multivariate model using Significant SNPs (Difference) are shown below. The table presents the estimate of the effect, its standard error, the β-coefficient, the calculated odds ratio with its confidence interval, and the *P*-value of the estimate in the final model for each of the included genomic annotations in the final model. Significant *P*-values in bold.

Annotation	**Estimate**	**Std. Error**	β	**OR [LCI-HCI]**	***P*-value**
Intercept	-8.69	0.09	-92.44	0.00 [0.00-0.00]	$0.00 \times 10^{+00}$
Affymetrix_250k_Nsp	0.35	0.18	1.94	1.41 [1.00-2.01]	5.23×10^{-02}
Affymetrix_250k_Sty	0.22	0.17	1.25	1.24 [0.88-1.75]	2.12×10^{-01}
Affymetrix_5.0	-0.06	0.17	-0.34	0.94 [0.68-1.32]	7.35×10^{-01}
Affymetrix_6.0	0.48	0.08	6.19	1.61 [1.39-1.87]	**6.01×10^{-10}**
Affymetrix_10k	0.44	0.44	1.00	1.55 [0.66-3.65]	3.19×10^{-01}
Affymetrix_50k.1	-0.12	0.19	-0.64	0.89 [0.61-1.28]	5.24×10^{-01}
Affymetrix_50k.2	-0.34	0.21	-1.61	0.71 [0.47-1.08]	1.08×10^{-01}
Illumina_300	0.35	0.09	3.96	1.41 [1.19-1.68]	**7.60×10^{-05}**
Illumina_550	0.63	0.17	3.64	1.88 [1.34-2.64]	**2.73×10^{-04}**
Illumina_650	0.47	0.16	2.90	1.60 [1.16-2.19]	**3.72×10^{-03}**
Perlegen	0.33	0.07	5.10	1.39 [1.23-1.58]	**3.37×10^{-07}**
Distance to TSS	0.00	0.00	-9.39	1.00 [1.00-1.00]	**6.13×10^{-21}**
eQTLs	0.87	0.08	11.43	2.38 [2.05-2.76]	**2.94×10^{-30}**
Exons	0.60	0.09	6.81	1.83 [1.54-2.17]	**9.47×10^{-12}**
Open Chromatin	0.50	0.07	7.42	1.64 [1.44-1.87]	**1.14×10^{-13}**
DNase Clusters	0.33	0.06	5.48	1.39 [1.24-1.56]	**4.37×10^{-08}**
Repressed	0.23	0.08	3.00	1.26 [1.08-1.47]	**2.68×10^{-03}**
5 Kb TSS	0.42	0.08	5.13	1.52 [1.30-1.78]	**2.84×10^{-07}**
Insulator	-0.28	0.13	-2.14	0.75 [0.58-0.98]	**3.22×10^{-02}**
Poised Promoter	0.63	0.20	3.13	1.88 [1.26-2.79]	**1.77×10^{-03}**
RNA Genes	-1.88	1.00	-1.88	0.15 [0.02-1.08]	6.00×10^{-02}
Weak Enhancer (proximal)	-0.23	0.10	-2.23	0.79 [0.65-0.97]	**2.58×10^{-02}**
Heterochrom/lo	-0.23	0.07	-3.60	0.79 [0.70-0.90]	**3.19×10^{-04}**
Indels Pure regions	0.12	0.07	1.80	1.13 [0.99-1.29]	7.25×10^{-02}
Syn. SNPs (UCSC)	-0.30	0.12	-2.52	0.74 [0.59-0.94]	**1.17×10^{-02}**
Txn Elongation	-0.17	0.08	-2.05	0.84 [0.71-0.99]	**4.08×10^{-02}**
Gained Stops	1.19	0.41	2.88	3.29 [1.46-7.42]	**4.02×10^{-03}**
28 specs. algmt. plac. mmls	0.14	0.08	1.81	1.15 [0.99-1.34]	7.10×10^{-02}
vega Genes	0.15	0.06	2.64	1.17 [1.04-1.31]	**8.26×10^{-03}**
1 Kb TSS	-0.21	0.11	-1.86	0.81 [0.66-1.01]	6.27×10^{-02}
Weak Txn	-0.11	0.07	-1.58	0.90 [0.78-1.03]	1.15×10^{-01}
PREMOD	0.15	0.08	1.82	1.16 [0.99-1.37]	6.86×10^{-02}
Splice Sites	-11.51	155.31	-0.07	0.00 [0.00-$1.59 \times 10^{+127}$]	9.41×10^{-01}
TS miRNA	-11.85	181.48	-0.07	0.00 [0.00-$2.14 \times 10^{+149}$]	9.48×10^{-01}
Pos. Sel. Genes	-0.09	0.06	-1.48	0.91 [0.80-1.03]	1.40×10^{-01}
ORegAnno	-0.22	0.15	-1.42	0.81 [0.60-1.09]	1.55×10^{-01}

5.3.4.2 Immune SNPs

The final model for the Immune SNPs contained 39 genomic annotations, of which 20 were significant at $P \leq 0.05$ (see Figure 5-8). The final model obtained in the Immune SNPs was run on the complementary set of Non-immune SNPs to enable comparisons between the two sets. Table 5-7 and Table 5-8 list the results of the final models for the Immune SNPs and the Non-immune SNPs, respectively, showing the estimates, standard errors, β-coefficients, odds ratios and confidence intervals and the P-values of all the genomic annotations, genotyping arrays and intercept. The nine genomic annotations that are significantly depleted of immune-associated SNPs are mainly chromatin states or conserved regions with the exception of 5'UTRs. The annotations are ranked in order of decreasing significance distance to TSS, positively selected genes, regions associated with chromatin states indicative of heterochromatin/low transcription, transcriptional elongation, conserved transcription factor binding sites, open regulatory annotations, closed chromatin, 5'UTRs, and sites found to be conserved in a 28 species alignment.

The 11 annotations that are enriched for immune-associated SNPs are in order of decreasing significance eQTLs, open chromatin, strong enhancer (proximal), exons, vega genes, DNase clusters, 5 Kb TSS, repressed chromatin states, weak enhancers (distal), TS miRNA binding sites, and vega pseudo genes. The strong enhancer (proximal), eQTLs, DNase clusters, and open chromatin regions were the annotations that were enriched for trait-associated SNPs in the majority of the analyses. The strong enhancer (proximal) annotation was the most enriched annotation in the analysis for Significant SNPs (2011).

There are 13 genomic annotations, which have statistically significantly different odds ratios for Immune and Non-immune SNPs. The conserved regions (positively selected genes, conserved transcription factor binding sites, and conserved sites from a 28 species alignment) are significantly depleted of Immune SNPs, while Non-immune SNPs are either enriched in these regions or do not show significant enrichment/depletion. The ORegAnno annotation contains, amongst other regulatory annotations, transcription factor binding sites, so the Immune SNPs are consistently depleted in these sites. The other significant differences are in eQTLs, strong enhancer (proximal) regions, vega genes, distance to TSS, Heterochromatin/low transcription regions, 5 Kb regions upstream the TSS, closed chromatin, open regulatory annotations, vega pseudo genes, and strong enhancer (distal) regions.

The McKelvey and Zavoina's pseudo-r^2 value of the final model including the distance to TSS was 0.45 for the Immune SNPs, and for the Base model was 0.18. The pseudo-r^2 values for the empty Non-immune SNP model was 0.11, and for the final model 0.35.

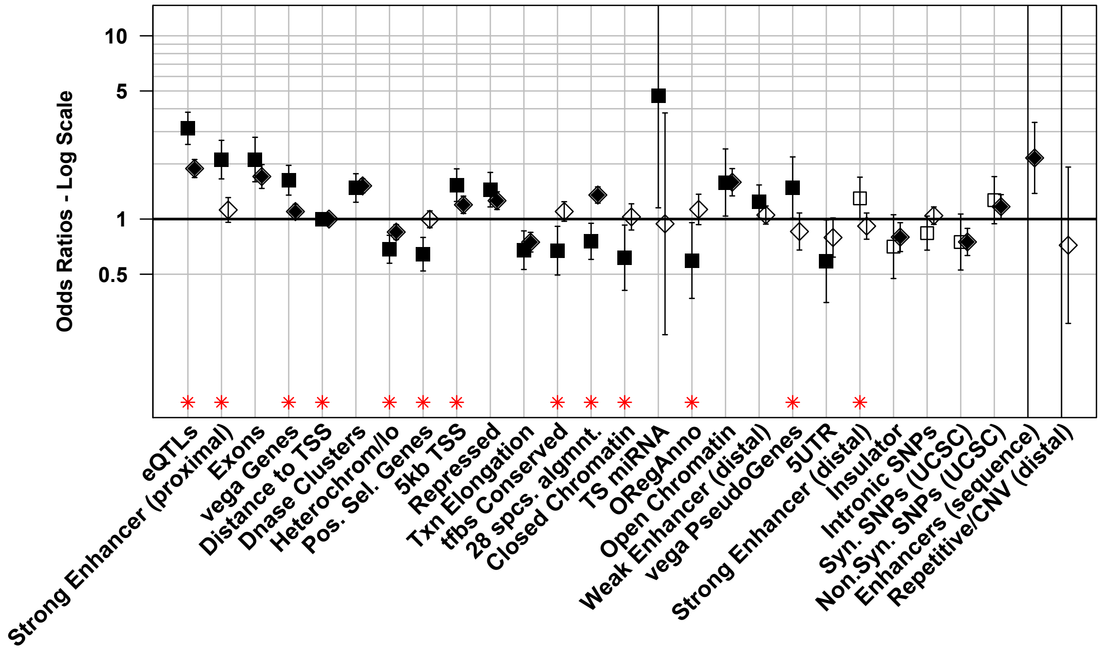


Figure 5-11 – Effect of genomic annotations in Immune *vs.* Non-immune SNPs

Odds ratios for all genomic annotations in the Immune SNPs model sorted after significance. Immune SNPs are shown as □ and Non-immune SNPs are shown as ◇. Solid symbols indicate significance at *P*-value ≤ 0.05. Red stars (✳) indicate significant differences between the two datasets.

Table 5-7 – Stepwise logistic regression results for Immune SNPs

The results for the multivariate model using the Immune SNPs are shown below. The table presents the estimate of the effect, its standard error, the β-coefficient, the calculated odds ratio with its confidence interval, and the *P*-value of the estimate in the final model for each of the included genomic annotations in the final model. Significant *P*-values in bold.

Annotation	Estimate	Std. Error	β	OR [LCI-HCI]	*P*-value
Intercept	-9.82	0.24	-41.21	0.00 [0.00-0.00]	$0.00 \times 10^{+00}$
Affymetrix_250k_Nsp	0.59	0.26	2.31	1.81 [1.09-3.00]	**2.08×10^{-02}**
Affymetrix_250k_Sty	0.58	0.24	2.36	1.78 [1.10-2.87]	**1.83×10^{-02}**
Affymetrix_5.0	-0.17	0.24	-0.70	0.84 [0.53-1.35]	4.81×10^{-01}
Affymetrix_6.0	0.23	0.12	1.91	1.26 [0.99-1.59]	5.58×10^{-02}
Affymetrix_10k	-0.93	1.02	-0.91	0.39 [0.05-2.91]	3.61×10^{-01}
Affymetrix_50k.1	-0.12	0.26	-0.44	0.89 [0.53-1.49]	6.61×10^{-01}
Affymetrix_50k.2	-0.03	0.26	-0.11	0.97 [0.58-1.62]	9.16×10^{-01}
Illumina_300	0.54	0.11	4.84	1.72 [1.38-2.14]	**1.30×10^{-06}**
Illumina_550	0.99	0.26	3.79	2.69 [1.61-4.48]	**1.50×10^{-04}**
Illumina_650	0.95	0.25	3.77	2.58 [1.57-4.22]	**1.66×10^{-04}**
Perlegen	0.09	0.10	0.97	1.10 [0.91-1.33]	3.31×10^{-01}
eQTLs	1.14	0.10	11.05	3.13 [2.55-3.83]	**2.29×10^{-28}**
Open Chromatin	0.46	0.22	2.13	1.58 [1.04-2.41]	**3.29×10^{-02}**
Strong Enhancer (proximal)	0.75	0.12	6.04	2.11 [1.66-2.69]	**1.51×10^{-09}**
Distance to TSS	0.00	0.00	-4.32	1.00 [1.00-1.00]	**1.56×10^{-05}**
Exons	0.75	0.14	5.25	2.11 [1.60-2.79]	**1.54×10^{-07}**
Pos. Sel. Genes	-0.44	0.11	-4.11	0.64 [0.52-0.79]	**4.02×10^{-05}**
vega Genes	0.49	0.10	5.12	1.63 [1.35-1.96]	**3.09×10^{-07}**
DNase Clusters	0.39	0.09	4.27	1.48 [1.24-1.77]	**1.95×10^{-05}**
Heterochrom/lo	-0.38	0.09	-4.24	0.68 [0.57-0.82]	**2.22×10^{-05}**
Txn Elongation	-0.39	0.12	-3.16	0.68 [0.53-0.86]	**1.56×10^{-03}**
5 Kb TSS	0.43	0.10	4.07	1.53 [1.25-1.88]	**4.72×10^{-05}**
tfbs Conserved	-0.40	0.16	-2.55	0.67 [0.49-0.91]	**1.09×10^{-02}**
ORegAnno	-0.52	0.24	-2.15	0.59 [0.37-0.96]	**3.19×10^{-02}**
Repressed	0.37	0.11	3.34	1.45 [1.16-1.80]	**8.41×10^{-04}**
Closed Chromatin	-0.49	0.21	-2.32	0.62 [0.41-0.93]	**2.03×10^{-02}**
Weak Enhancer (distal)	0.22	0.11	2.07	1.25 [1.01-1.54]	**3.89×10^{-02}**
5’UTR	-0.53	0.26	-2.00	0.59 [0.35-0.99]	**4.51×10^{-02}**
Syn. SNPs (UCSC)	-0.29	0.18	-1.61	0.75 [0.53-1.07]	1.08×10^{-01}
Strong Enhancer (distal)	0.26	0.14	1.92	1.30 [0.99-1.69]	5.52×10^{-02}
TS miRNA	1.55	0.72	2.16	4.71 [1.15-19.29]	**3.10×10^{-02}**
Insulator	-0.35	0.20	-1.69	0.71 [0.47-1.06]	9.02×10^{-02}
Repetitive/CNV (distal)	-11.63	173.69	-0.07	0.00 [0.00-6.30 × 10^{+142}]	9.47×10^{-01}
Enhancers (sequence)	-11.19	141.87	-0.08	0.00 [0.00-7.94 × 10^{+115}]	9.37×10^{-01}
Intronic SNPs	-0.18	0.11	-1.62	0.84 [0.68-1.04]	1.05×10^{-01}
vega PseudoGenes	0.40	0.20	2.01	1.49 [1.01-2.18]	**4.43×10^{-02}**
28 specs. algmt.	-0.28	0.12	-2.42	0.76 [0.60-0.95]	**1.55×10^{-02}**
Non.Syn. SNPs (UCSC)	0.24	0.15	1.57	1.27 [0.94-1.70]	1.16×10^{-01}

Table 5-8 – Stepwise logistic regression results for Non-immune SNPs

The results for the multivariate model using the Non-immune SNPs are shown below. The table presents the estimate of the effect, its standard error, the β-coefficient, the calculated odds ratio with its confidence interval, and the *P*-value of the estimate in the final model for each of the included genomic annotations in the final model.

Annotation	Estimate	Std. Error	β	OR [LCI-HCI]	*P*-value
Intercept	-8.20	0.10	-81.54	0.00 [0.00-0.00]	$0.00 \times 10^{+00}$
Affymetrix_250k_Nsp	0.30	0.14	2.24	1.36 [1.04-1.77]	**2.49×10^{-02}**
Affymetrix_250k_Sty	0.33	0.13	2.50	1.39 [1.07-1.79]	**1.25×10^{-02}**
Affymetrix_5.0	0.09	0.13	0.74	1.10 [0.86-1.41]	4.59×10^{-01}
Affymetrix_6.0	0.40	0.06	6.79	1.50 [1.33-1.68]	**1.12×10^{-11}**
Affymetrix_10k	0.75	0.26	2.87	2.11 [1.27-3.51]	**4.12×10^{-03}**
Affymetrix_50k.1	-0.05	0.13	-0.37	0.95 [0.74-1.23]	7.11×10^{-01}
Affymetrix_50k.2	-0.19	0.14	-1.38	0.82 [0.63-1.08]	1.68×10^{-01}
Illumina_300	0.70	0.07	10.36	2.00 [1.76-2.29]	**3.77×10^{-25}**
Illumina_550	0.97	0.15	6.26	2.63 [1.94-3.56]	**3.97×10^{-10}**
Illumina_650	0.03	0.15	0.20	1.03 [0.77-1.37]	8.41×10^{-01}
Perlegen	0.30	0.05	6.20	1.35 [1.23-1.48]	**5.75×10^{-10}**
eQTLs	0.63	0.06	10.96	1.89 [1.68-2.11]	**5.76×10^{-28}**
Open Chromatin	0.46	0.09	5.28	1.59 [1.34-1.89]	**1.27×10^{-07}**
Strong Enhancer (proximal)	0.11	0.08	1.44	1.12 [0.96-1.31]	1.49×10^{-01}
Distance to TSS	0.00	0.00	-14.21	1.00 [1.00-1.00]	**7.64×10^{-46}**
Exons	0.53	0.08	6.99	1.71 [1.47-1.98]	**2.69×10^{-12}**
Pos. Sel. Genes	0.00	0.05	-0.06	1.00 [0.90-1.11]	9.52×10^{-01}
vega Genes	0.09	0.05	2.04	1.10 [1.00-1.20]	**4.18×10^{-02}**
DNase Clusters	0.42	0.04	9.30	1.52 [1.39-1.66]	**1.46×10^{-20}**
Heterochrom/lo	-0.16	0.05	-3.39	0.85 [0.77-0.93]	**6.94×10^{-04}**
Txn Elongation	-0.29	0.06	-4.62	0.75 [0.66-0.85]	**3.93×10^{-06}**
5 Kb TSS	0.18	0.06	3.23	1.20 [1.07-1.33]	**1.24×10^{-03}**
tfbs Conserved	0.10	0.06	1.52	1.10 [0.97-1.24]	1.30×10^{-01}
ORegAnno	0.12	0.10	1.24	1.13 [0.93-1.37]	2.14×10^{-01}
Repressed	0.23	0.06	4.06	1.26 [1.13-1.41]	**4.80×10^{-05}**
Closed Chromatin	0.03	0.08	0.32	1.03 [0.87-1.21]	7.48×10^{-01}
Weak Enhancer (distal)	0.05	0.06	0.87	1.05 [0.94-1.18]	3.86×10^{-01}
5'UTR	-0.23	0.13	-1.85	0.79 [0.62-1.01]	6.47×10^{-02}
Syn. SNPs (UCSC)	-0.29	0.09	-3.31	0.75 [0.63-0.89]	**9.35×10^{-04}**
Strong Enhancer (distal)	-0.09	0.08	-1.05	0.92 [0.78-1.08]	2.92×10^{-01}
TS miRNA	-0.06	0.71	-0.08	0.94 [0.23-3.79]	9.33×10^{-01}
Insulator	-0.23	0.09	-2.45	0.80 [0.66-0.96]	1.41×10^{-02}
Repetitive/CNV (distal)	-0.33	0.50	-0.66	0.72 [0.27-1.92]	5.12×10^{-01}
Enhancers (sequence)	0.77	0.23	3.37	2.16 [1.38-3.37]	**7.41×10^{-04}**
Intronic SNPs	0.04	0.06	0.76	1.04 [0.94-1.16]	4.45×10^{-01}
vega PseudoGenes	-0.16	0.12	-1.32	0.86 [0.68-1.08]	1.88×10^{-01}
28 specs. algmt.	0.30	0.05	5.81	1.35 [1.22-1.50]	**6.35×10^{-09}**
Non.Syn. SNPs (UCSC)	0.16	0.08	2.03	1.17 [1.01-1.36]	**4.23×10^{-02}**

5.3.4.3 Cancer SNPs

The results for the stepwise logistic regression of Cancer SNPs are shown in Figure 5-9 and are also presented in Table 5-9. The Final Model in the Cancer SNPs was used to analyse the Non-cancer SNPs (see Table 5-10). A heterogeneous disease would be unlikely to have similar pathways to a phenotype. Since cancer is a very heterogeneous disease classification and the Cancer SNP set contained a small number of associated SNPs (268 SNPs), any result would have been encouraging. Despite these problems, we have obtained 15 annotations influencing cancer association, five of which were significantly different between Cancer and Non-cancer SNPs: Exons, conserved regions in primates in a 44 species alignment, coding SNPs, weak enhancers (proximal) regions, and intronic SNPs. The Cancer SNPs had a higher odds ratio in the conserved regions and weak enhancers than the Non-cancer SNPs suggesting a more important role of conserved regions in cancer aetiology. While the exons are significantly more enriched for Cancer SNPs than Non-cancer SNPs, the coding SNPs and the introns are significantly depleted for Cancer SNPs. This seems contradictory, but it is not since the annotations were analysed relative to each other. The McKelvey and Zavoina's pseudo-r^2 value for the Final Model of the Cancer SNPs was 0.41, which meant that the included genomic annotations improved the Base model (base model: 0.20). The McKelvey and Zavoina's pseudo-r^2 value for the Base model of the Non-cancer SNPs was 0.11 and for the full model was 0.37.

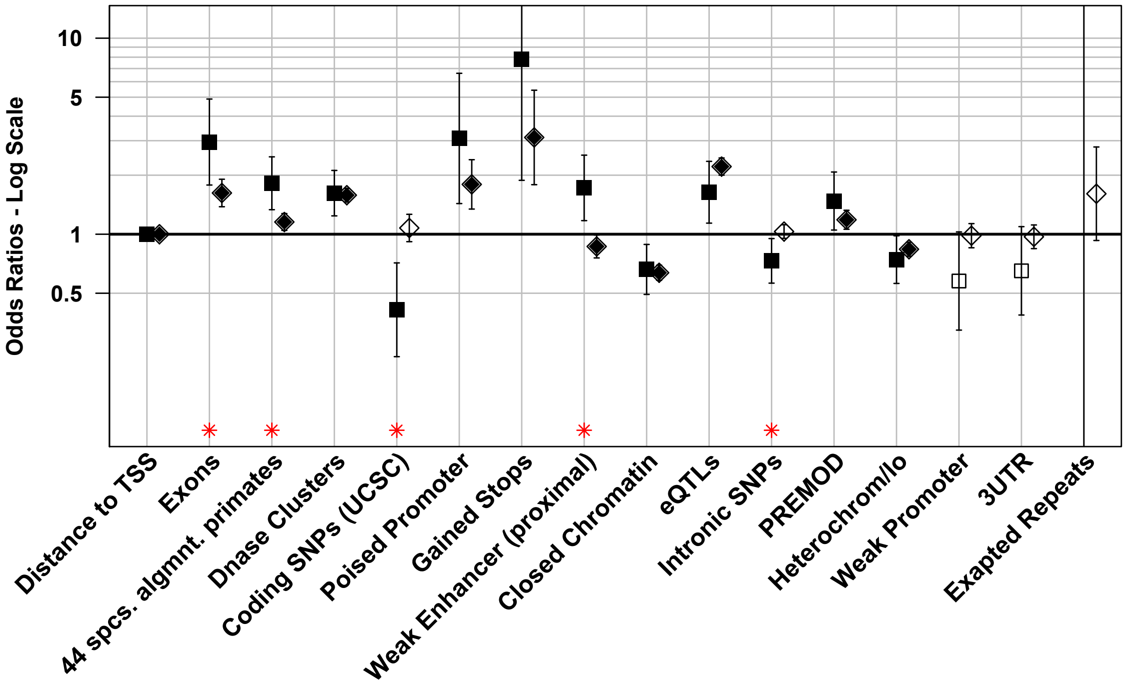


Figure 5-12 – Cancer SNPs *vs.* Non-cancer SNPs

Odds ratios for all genomic annotations in the Cancer SNPs model sorted after significance. Cancer SNPs are shown as □ and Non-cancer SNPs are shown as ◇. Solid symbols indicate significance at *P*-value ≤ 0.05. Red stars (✳) indicate significant differences between the two datasets.

Table 5-9 – Stepwise logistic regression results for Cancer SNPs

The results for the multivariate model using the Cancer SNPs are shown below. The table presents the estimate of the effect, its standard error, the β-coefficient, the calculated odds ratio with its confidence interval, and the *P*-value of the estimate in the final model for each of the included genomic annotations in the final model. Significant *P*-values in bold.

Annotation	Estimate	Std. Error	β	OR [LCI-HCI]	*P*-value
Intercept	-10.44	0.19	-55.11	0.00 [0.00-0.00]	$0.00 \times 10^{+00}$
Affymetrix_250k_Nsp	0.24	0.41	0.58	1.27 [0.57-2.82]	5.60×10^{-01}
Affymetrix_250k_Sty	0.57	0.38	1.51	1.77 [0.84-3.74]	1.32×10^{-01}
Affymetrix_5.0	-0.07	0.37	-0.19	0.93 [0.45-1.94]	8.52×10^{-01}
Affymetrix_6.0	0.50	0.18	2.85	1.65 [1.17-2.32]	**4.41×10^{-03}**
Affymetrix_10k	1.45	0.79	1.83	4.25 [0.90-20.03]	6.72×10^{-02}
Affymetrix_50k.1	0.04	0.33	0.11	1.04 [0.55-1.96]	9.13×10^{-01}
Affymetrix_50k.2	-1.52	0.65	-2.35	0.22 [0.06-0.78]	**1.86×10^{-02}**
Illumina_300	1.38	0.20	6.76	3.99 [2.67-5.96]	**1.35×10^{-11}**
Illumina_550	1.55	0.57	2.73	4.69 [1.55-14.24]	**6.35×10^{-03}**
Illumina_650	-0.08	0.54	-0.15	0.92 [0.32-2.65]	8.78×10^{-01}
Perlegen	0.26	0.14	1.78	1.29 [0.98-1.71]	7.43×10^{-02}
Distance to TSS	0.00	0.00	-4.29	1.00 [1.00-1.00]	**1.75×10^{-05}**
44 specs. algmt. primates	0.60	0.16	3.78	1.82 [1.33-2.48]	**1.57×10^{-04}**
DNase Clusters	0.48	0.14	3.54	1.62 [1.24-2.11]	**3.97×10^{-04}**
Closed Chromatin	-0.41	0.15	-2.76	0.66 [0.49-0.89]	**5.74×10^{-03}**
Poised Promoter	1.13	0.39	2.88	3.08 [1.43-6.62]	**3.94×10^{-03}**
Weak Enhancer (proximal)	0.54	0.20	2.77	1.72 [1.17-2.53]	**5.54×10^{-03}**
eQTLs	0.49	0.19	2.66	1.64 [1.14-2.35]	**7.74×10^{-03}**
Intronic SNPs	-0.31	0.13	-2.34	0.73 [0.56-0.95]	**1.92×10^{-02}**
Gained Stops	2.06	0.73	2.83	7.82 [1.88-32.47]	**4.63×10^{-03}**
PREMOD	0.39	0.17	2.24	1.48 [1.05-2.08]	**2.50×10^{-02}**
Heterochrom/lo	-0.30	0.14	-2.09	0.74 [0.56-0.98]	**3.67×10^{-02}**
Exons	1.08	0.26	4.20	2.95 [1.78-4.89]	**2.64×10^{-05}**
Coding SNPs (UCSC)	-0.89	0.28	-3.16	0.41 [0.24-0.71]	**1.58×10^{-03}**
Weak Promoter	-0.55	0.29	-1.87	0.58 [0.32-1.03]	6.17×10^{-02}
3'UTR	-0.43	0.26	-1.62	0.65 [0.39-1.09]	1.04×10^{-01}
Exapted Repeats	-12.08	245.23	-0.05	0.00 [0.00-$3.13 \times 10^{+203}$]	9.61×10^{-01}

Table 5-10 – Stepwise logistic regression results for Non-cancer SNPs
The results for the multivariate model using the Non-cancer SNPs are shown below. The table presents the estimate of the effect, its standard error, the β-coefficient, the calculated odds ratio with its confidence interval, and the *P*-value of the estimate in the final model for each of the included genomic annotations in the final model. Significant *P*-values in bold.

Annotation	**Estimate**	**Std. Error**	β	**OR [LCI-HCI]**	***P*-value**
Intercept	-7.54	0.05	-141.53	0.00 [0.00-0.00]	$0.00 \times 10^{+00}$
Affymetrix_250k_Nsp	0.38	0.13	3.06	1.47 [1.15-1.88]	**2.21×10^{-03}**
Affymetrix_250k_Sty	0.38	0.12	3.11	1.46 [1.15-1.85]	**1.88×10^{-03}**
Affymetrix_5.0	0.05	0.12	0.43	1.05 [0.83-1.33]	6.67×10^{-01}
Affymetrix_6.0	0.35	0.06	6.22	1.41 [1.27-1.58]	**4.87×10^{-10}**
Affymetrix_10k	0.48	0.26	1.81	1.61 [0.96-2.69]	7.06×10^{-02}
Affymetrix_50k.1	-0.08	0.12	-0.68	0.92 [0.72-1.17]	4.98×10^{-01}
Affymetrix_50k.2	-0.06	0.13	-0.51	0.94 [0.73-1.20]	6.08×10^{-01}
Illumina_300	0.59	0.06	9.83	1.81 [1.61-2.03]	**8.01×10^{-23}**
Illumina_550	0.93	0.14	6.82	2.54 [1.94-3.31]	**9.34×10^{-12}**
Illumina_650	0.23	0.13	1.81	1.26 [0.98-1.63]	6.97×10^{-02}
Perlegen	0.27	0.05	5.89	1.31 [1.20-1.43]	**3.75×10^{-09}**
Distance to TSS	0.00	0.00	-15.98	1.00 [1.00-1.00]	**1.69×10^{-57}**
44 specs. algmt. primates	0.14	0.05	2.84	1.15 [1.05-1.27]	**4.47×10^{-03}**
DNase Clusters	0.46	0.04	11.20	1.58 [1.46-1.71]	**3.88×10^{-29}**
Closed Chromatin	-0.45	0.05	-9.98	0.64 [0.58-0.70]	**1.79×10^{-23}**
Poised Promoter	0.59	0.15	3.96	1.80 [1.34-2.40]	**7.57×10^{-05}**
Weak Enhancer (proximal)	-0.14	0.07	-2.09	0.87 [0.76-0.99]	**3.62×10^{-02}**
eQTLs	0.79	0.05	15.45	2.21 [2.00-2.45]	**8.12×10^{-54}**
Intronic SNPs	0.03	0.04	0.77	1.03 [0.95-1.12]	4.43×10^{-01}
Gained Stops	1.14	0.28	4.02	3.12 [1.79-5.43]	**5.92×10^{-05}**
PREMOD	0.17	0.06	2.98	1.18 [1.06-1.32]	**2.89×10^{-03}**
Heterochrom/lo	-0.18	0.04	-4.17	0.84 [0.77-0.91]	**3.03×10^{-05}**
Exons	0.48	0.08	5.88	1.62 [1.38-1.91]	**4.16×10^{-09}**
Coding SNPs (UCSC)	0.07	0.08	0.89	1.08 [0.92-1.26]	3.73×10^{-01}
Weak Promoter	-0.02	0.07	-0.24	0.98 [0.85-1.13]	8.10×10^{-01}
3'UTR	-0.03	0.07	-0.43	0.97 [0.84-1.11]	6.70×10^{-01}
Exapted Repeats	0.48	0.28	1.70	1.61 [0.93-2.79]	8.98×10^{-02}

5.3.4.4 Normal Variation SNPs vs. Disease SNPs

Another partitioning of the Significant SNPs (2013) was undertaken to divide the total number of significantly trait-associated SNPs into SNPs associated with Normal Variation traits (see page 211) and Disease traits (see page 218). The comparison between SNPs associated with Normal Variation traits, such as height, eye or hair colour, and SNPs associated with Diseases showed that there was a common set of genomic annotations that influenced trait-association status. The results of the Final Model for the Normal Variation SNPs are shown in Table 5-11 and the results for the Disease SNPs are listed in Table 5-12.

Figure 5-13 shows these common genomic annotations. The Disease SNPs were significantly different to the Normal Variation SNPs in exons and 3'UTRs.

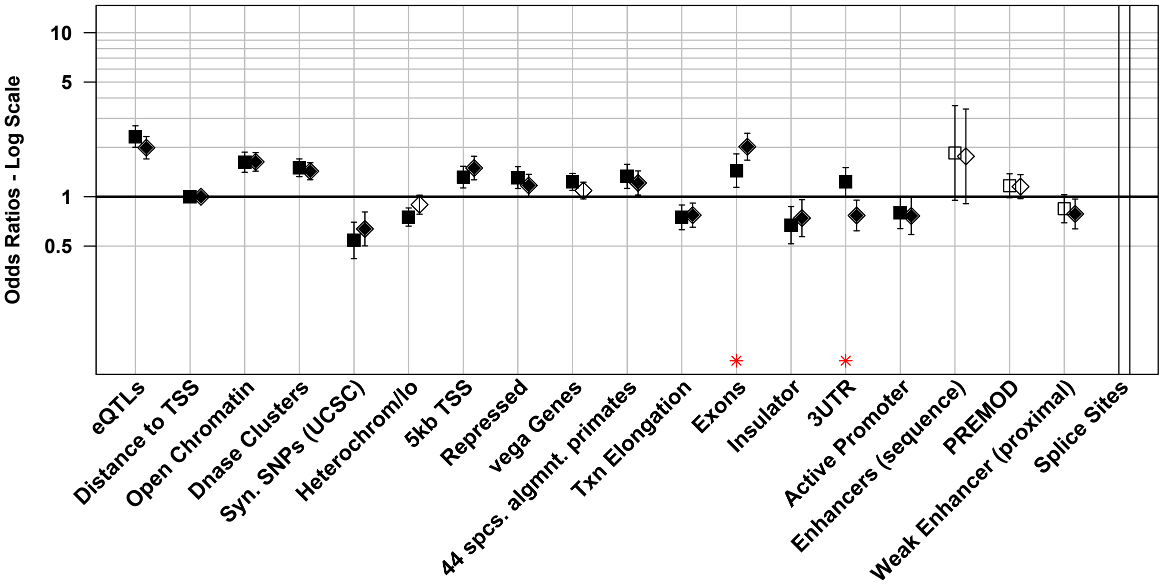


Figure 5-13 – Normal Variation SNPs *vs.* Disease SNPs
Odds ratios for all common genomic annotations in the Disease SNPs and Normal Variation SNPs models sorted after significance in the Disease model. Disease SNPs are shown as □ and Normal Variation SNPs are shown as ◇. Solid symbols indicate significance at *P*-value ≤ 0.05. The red stars (✳) indicate significant differences between the two datasets.

The McKelvey and Zavoina's pseudo-r^2 value for the final model determined for the Normal Variation SNPs was 0.38, while for the Base model it was 0.08. The Base model for the Disease SNPs had a McKelvey and Zavoina's pseudo-r^2 value of 0.14 and for the final model it was 0.39.

Table 5-11 – Stepwise logistic regression results for Normal Variation SNPs

The results for the multivariate model using the Normal Variation SNPs are shown below. The table presents the estimate of the effect, its standard error, the β-coefficient, the calculated odds ratio with its confidence interval, and the *P*-value of the estimate in the final model for each of the included genomic annotations in the final model. Significant *P*-values in bold.

Annotation	Estimate	Std. Error	β	OR [LCI-HCI]	*P*-value
Intercept	-8.67	0.10	-88.45	0.00 [0.00-0.00]	$0.00 \times 10^{+00}$
Affymetrix_250k_Nsp	0.34	0.19	1.78	1.40 [0.97-2.03]	7.52×10^{-02}
Affymetrix_250k_Sty	0.18	0.19	0.97	1.20 [0.83-1.72]	3.33×10^{-01}
Affymetrix_5.0	0.06	0.18	0.33	1.06 [0.75-1.51]	7.41×10^{-01}
Affymetrix_6.0	0.34	0.08	4.16	1.40 [1.20-1.65]	**3.14×10^{-05}**
Affymetrix_10k	0.92	0.34	2.71	2.50 [1.29-4.85]	**6.63×10^{-03}**
Affymetrix_50k.1	0.09	0.18	0.50	1.09 [0.77-1.54]	6.19×10^{-01}
Affymetrix_50k.2	-0.13	0.20	-0.65	0.88 [0.60-1.29]	5.16×10^{-01}
Illumina_300	0.45	0.09	4.83	1.56 [1.30-1.88]	**1.34×10^{-06}**
Illumina_550	0.68	0.19	3.60	1.98 [1.36-2.87]	**3.21×10^{-04}**
Illumina_650	0.26	0.18	1.46	1.29 [0.92-1.83]	1.45×10^{-01}
Perlegen	0.39	0.07	5.91	1.48 [1.30-1.68]	**3.44×10^{-09}**
Distance to TSS	0.00	0.00	-10.19	1.00 [1.00-1.00]	**2.09×10^{-24}**
Exons	0.70	0.10	7.28	2.02 [1.67-2.44]	**3.27×10^{-13}**
eQTLs	0.69	0.08	8.57	1.99 [1.70-2.33]	**1.02×10^{-17}**
Open Chromatin	0.49	0.07	7.32	1.63 [1.43-1.86]	**2.44×10^{-13}**
DNase Clusters	0.36	0.06	5.85	1.43 [1.27-1.61]	**4.79×10^{-09}**
44 specs. algmt.	0.47	0.13	3.68	1.59 [1.24-2.04]	**2.35×10^{-04}**
RNA Genes	-12.19	104.08	-0.12	0.00 [0.00-2.00 × 10^{+83}]	9.07×10^{-01}
Syn. SNPs (UCSC)	-0.45	0.12	-3.72	0.64 [0.50-0.81]	**1.99×10^{-04}**
Intergenic SNPs	-0.16	0.06	-2.59	0.85 [0.75-0.96]	**9.74×10^{-03}**
Txn Elongation	-0.26	0.09	-2.99	0.77 [0.65-0.91]	**2.76×10^{-03}**
5 Kb TSS	0.40	0.08	4.77	1.50 [1.27-1.76]	**1.82×10^{-06}**
1 Kb TSS	-0.32	0.12	-2.59	0.73 [0.57-0.93]	**9.72×10^{-03}**
Indels Pure regions	0.16	0.07	2.26	1.18 [1.02-1.35]	**2.37×10^{-02}**
Weak Enhancer (proximal)	-0.24	0.11	-2.27	0.79 [0.64-0.97]	**2.34×10^{-02}**
Repressed	0.16	0.08	1.98	1.17 [1.00-1.37]	**4.77×10^{-02}**
3'UTR	-0.26	0.11	-2.40	0.77 [0.62-0.95]	**1.65×10^{-02}**
Insulator	-0.30	0.13	-2.27	0.74 [0.57-0.96]	**2.31×10^{-02}**
ORegAnno	0.25	0.13	1.90	1.29 [0.99-1.68]	5.69×10^{-02}
44 specs. algmt. primates	0.19	0.09	2.24	1.21 [1.02-1.44	**2.54×10^{-02}**
44 specs. algmt. plac. mmls	-0.22	0.13	-1.70	0.81 [0.63-1.03]	8.91×10^{-02}
5'UTR	-0.34	0.19	-1.80	0.71 [0.49-1.03]	7.19×10^{-02}
CpG Islands	0.34	0.14	2.47	1.40 [1.07-1.83]	**1.37×10^{-02}**
Active Promoter	-0.27	0.13	-2.03	0.76 [0.59-0.99	**4.27×10^{-02}**
Enhancers (sequence)	0.57	0.34	1.67	1.76 [0.91-3.42]	9.47×10^{-02}
Splice Sites	-12.41	258.30	-0.05	0.00 [0.00-3.02 × 10^{+214}]	9.62×10^{-01}
Heterochrom/lo	-0.11	0.07	-1.65	0.89 [0.78-1.02]	9.98×10^{-02}
17 specs. algmt.	-0.15	0.09	-1.65	0.86 [0.71-1.03]	9.87×10^{-02}
PREMOD	0.14	0.09	1.64	1.15 [0.97-1.36]	1.01×10^{-01}
vega Genes	0.09	0.06	1.44	1.09 [0.97-1.23]	1.50×10^{-01}

Table 5-12 – Stepwise logistic regression results for Disease SNPs

The results for the multivariate model using the Disease SNPs are shown below. The table presents the estimate of the effect, its standard error, the β-coefficient, the calculated odds ratio with its confidence interval, and the *P*-value of the estimate in the final model for each of the included genomic annotations in the final model. Significant *P*-values in bold.

Annotation	Estimate	Std. Error	β	OR [LCI-HCI]	*P*-value
Intercept	-9.03	0.10	-89.78	0.00 [0.00-0.00]	$0.00 \times 10^{+00}$
Affymetrix_250k_Nsp	0.36	0.18	2.01	1.44 [1.01-2.05]	**4.49×10^{-02}**
Affymetrix_250k_Sty	0.55	0.17	3.22	1.74 [1.24-2.43]	**1.27×10^{-03}**
Affymetrix_5.0	0.02	0.17	0.15	1.03 [0.74-1.42]	8.82×10^{-01}
Affymetrix_6.0	0.43	0.08	5.18	1.53 [1.30-1.80]	**2.24×10^{-07}**
Affymetrix_10k	0.44	0.41	1.09	1.56 [0.70-3.46]	2.75×10^{-01}
Affymetrix_50k.1	-0.36	0.20	-1.81	0.70 [0.47-1.03]	7.06×10^{-02}
Affymetrix_50k.2	-0.25	0.19	-1.31	0.78 [0.53-1.13]	1.90×10^{-01}
Illumina_300	0.77	0.09	8.69	2.15 [1.81-2.55]	**3.59×10^{-18}**
Illumina_550	1.18	0.22	5.38	3.26 [2.12-5.01]	**7.52×10^{-08}**
Illumina_650	0.12	0.21	0.59	1.13 [0.75-1.71]	5.53×10^{-01}
Perlegen	0.20	0.07	3.05	1.23 [1.08-1.40]	**2.33×10^{-03}**
Distance to TSS	0.00	0.00	-9.51	1.00 [1.00-1.00]	**1.88×10^{-21}**
eQTLs	0.84	0.08	10.99	2.33 [2.00-2.71]	**4.50×10^{-28}**
DNase Clusters	0.41	0.06	6.40	1.50 [1.33-1.70]	**1.55×10^{-10}**
Open Chromatin	0.48	0.07	6.72	1.62 [1.41-1.87]	**1.84×10^{-11}**
Exons	0.37	0.12	3.06	1.44 [1.14-1.82]	**2.22×10^{-03}**
Strong Enhancer (proximal)	0.48	0.09	5.12	1.61 [1.34-1.94]	**3.00×10^{-07}**
44 specs. algmt. primates	0.29	0.09	3.31	1.33 [1.12-1.58]	**9.30×10^{-04}**
Syn. SNPs (UCSC)	-0.61	0.13	-4.71	0.54 [0.42-0.70]	**2.48×10^{-06}**
Gained Stops	1.65	0.33	4.96	5.23 [2.72-10.05]	**6.93×10^{-07}**
Repressed	0.27	0.08	3.43	1.31 [1.12-1.53]	**6.11×10^{-04}**
Heterochrom/lo	-0.29	0.07	-4.39	0.75 [0.66-0.85]	**1.15×10^{-05}**
Txn Elongation	-0.29	0.09	-3.28	0.75 [0.63-0.89]	**1.05×10^{-03}**
5 Kb TSS	0.28	0.08	3.55	1.32 [1.13-1.53]	**3.85×10^{-04}**
Insulator	-0.40	0.13	-2.99	0.67 [0.52-0.87]	**2.76×10^{-03}**
vega Genes	0.21	0.06	3.42	1.23 [1.09-1.39]	**6.33×10^{-04}**
Pos. Sel. Genes	-0.22	0.07	-3.24	0.80 [0.70-0.92]	**1.18×10^{-03}**
Coding SNPs (UCSC)	0.37	0.13	2.92	1.45 [1.13-1.86]	**3.51×10^{-03}**
Weak Enhancer (distal)	0.20	0.08	2.55	1.22 [1.05-1.41]	**1.09×10^{-02}**
Insulators (sequence)	-0.22	0.11	-2.04	0.80 [0.64-0.99]	**4.10×10^{-02}**
3'UTR	0.21	0.10	2.01	1.23 [1.01-1.50]	**4.42×10^{-02}**
28 specs. algmt. plac. mmls	0.31	0.13	2.42	1.36 [1.06-1.74]	**1.56×10^{-02}**
28 specs. algmt.	-0.23	0.12	-1.97	0.79 [0.63-1.00]	**4.91×10^{-02}**
Active Promoter	-0.22	0.11	-1.97	0.80 [0.64-1.00]	**4.89×10^{-02}**
PREMOD	0.15	0.09	1.79	1.17 [0.99-1.38]	7.28×10^{-02}
Splice Sites	-11.53	149.85	-0.08	0.00 [0.00-3.55 × 10^{+122}]	9.39×10^{-01}
Exapted Repeats	0.67	0.36	1.87	1.95 [0.97-3.94]	6.18×10^{-02}
Enhancers (sequence)	0.61	0.34	1.81	1.85 [0.95-3.59]	7.01×10^{-02}
Weak Enhancer (proximal)	-0.17	0.10	-1.67	0.85 [0.69-1.03]	9.55×10^{-02}
Microsatellites	-10.65	109.66	-0.10	0.00 [0.00-5.27 × 10^{+88}]	9.23×10^{-01}
TS miRNA	0.90	0.52	1.72	2.47 [0.88-6.90]	8.52×10^{-02}

5.3.5 Pseudo-r^2 values

As mentioned in the methods section of this chapter, we used the McKelvey and Zavoina's pseudo-r^2 value. The McKelvey and Zavoina's pseudo-r^2 value has two interpretations: the first and third approach of the linear regression r^2 outlined in the methods section of this chapter [150, 151, 153]. First, it can be understood as the proportion of variance explained by a model. It can also be seen as a square of the correlation between the model's predicted values for the dependent variable and its actual values. The dataset with the highest pseudo-r^2 value was the Immune SNPs with a value of 0.45. The base models containing only the genotyping arrays ranged between 0.08 and 0.20 depending on the dataset under investigation.

Table 5-13 – McKelvey and Zavoina's pseudo-r^2 values for different SNP sets
This table presents the McKelvey and Zavoina's pseudo-r^2 values calculated for the different logistic regression models calculated for the Base model containing only the genotyping arrays (Genotyping arrays only) and the Final Model, which was returned as the final model (Final model).

Dataset	Genotyping arrays only	Final model
Significant SNPs (2011)	0.14	0.42
Suggestive SNPs (2011)	0.16	0.18
Significant SNPs (Difference)	0.12	0.31
Significant SNPs (2013)	0.12	0.36
Immune SNPs	0.18	0.45
Non-immune SNPs	0.11	0.35
Cancer SNPs	0.20	0.41
Non-cancer SNPs	0.11	0.37
Normal Variation SNPs	0.08	0.38
Disease SNPs	0.14	0.39

5.4 Discussion

In this chapter we analysed univariate and multivariate logistic regression models to analyse the distribution of trait-associated variants. We analysed eight different datasets, which consisted of two thresholds of significance, a more recent set of trait-associated variants and the analysis of trait-specific subsets. The multivariate analysis obtained different models for the trait-subsets, which allowed the trait-specific conclusions.

The univariate regression was performed to allow comparisons between the sampling and permutation methods, which could be called univariate analyses as they were analysing individual annotations. The analysis of the univariate regression returned odds ratios and standard errors, which were not significantly different to the results returned by the permutations. This suggested that univariate logistic regression might be a useful method of analysing individual annotations, as they were performed significantly faster (one day for 100 samples *vs.* three days for 20,000 permutations *vs.* two hours for the univariate regression).

The iterative stepwise logistic regression approach in- or excluded genomic annotations according to the information they added to the model. This was calculated as the Akaike's Information Criterion (AIC), which was defined in the Methods (see page 102). Once a genomic annotation was included, it could have a positive or negative influence on trait-association status, which was indicative of enrichment for or depletion of trait-associated variants, respectively. Genomic annotations with 100% overlap or non-overlap with trait-associated SNPs were deemed more influential or informative than those annotations that occurred with trait-associated variants only 50% of the time.

The stepwise logistic regression of the multivariate model was performed in order to remove the redundant information and balanced the information carried by each genomic annotation against the included number of genomic annotations. The model showed the relative influence of the annotations to each other, once all other genomic annotations and genotyping arrays were taken into account. This can clearly be seen in Figure 5-3 and Table 5-2, where five genomic annotations were significantly depleted, which were enriched in the individual annotation analyses. Four of these (transcriptional elongation, synonymous SNPs, active promoters, and 5'UTRs) were enriched in all of the methods analysing the genomic annotations individually. These four annotations, which were depleted in the multivariate model, are all coinciding with or are close to coding regions of the genome and could be overlapping with the distance to TSS annotation. The coding regions are possibly overrepresented in the analysis as the majority of the genomic annotations are annotating genes or genic regions, while only 1.5% of the genome is thought to be coding. However, since there is a bias in the ascertainment of the genomic annotation there is, as previously discussed, also a bias in the inclusion of the SNPs on the genotyping arrays with clear preferences towards the coding regions of the genome. It is also possible that the univariate methods overstated the importance of the analysed genomic annotation, as the influences of the other genomic annotations were not taken into account in the analysis. The latter is a more plausible explanation since the majority of the genomic annotations are not mutually exclusive and therefore share some of the information. These results were therefore not contradictory to previous findings but did highlight the necessity of analysing the genomic annotations together to add extra information into the model. A possible improvement to this model could be to analyse each individual annotation after the genotyping arrays were included in the model as a single variate combining the information carried by all genotyping arrays. Alternatively, an analysis of the genomic annotations without the distance to TSS could be performed to investigate its effect on these annotations. This was not done, as we wanted to compare the univariate analyses to each other and neither the permutation nor the sampling method allowed that particular step. However, it would add an additional step into the comparison process of univariate and multiple variate analyses and could be performed in future studies.

The results for the trait-associated SNPs with *P*-values that did not pass the genome-wide significance threshold (Suggestive SNPs (2011)) were very comparable to the results for the significantly trait-associated SNPs. The results were therefore consistent with the hypothesis that the dataset of suggestively trait-associated SNPs was a mixture of false positives and real trait-associated SNPs. The real trait-associations were, however, not of sufficient magnitude to reach significance at the genome-wide level. These real trait-associated variants would be expected to have the same bias towards particular genomic features as the significantly trait-associated variants [65]. False positives, however, would be expected to have a similar genomic annotation profile as non-associated variants.

GWAS identify many trait-associated variants with different *P*-values of trait-association, many of which do not pass the genome-wide significance threshold. It is therefore expected, that for every reported suggestive SNP there were many more that were not reported, since the general assumption is that suggestive associations provided less information on the trait. Additionally, the NHGRI catalogue reported genome-wide studies associations with association *P*-values starting from 5×10^{-5} while the more commonly accepted *P*-value for suggestively associated SNPs started at 5×10^{-4} [36, 85]. The Significant SNPs (2011) were therefore a more comprehensive dataset, despite it being smaller, and the conclusions drawn from its results were thus expected to be more informative. However, the similarity of enrichment and depletion trends between Significant SNPs (2011) and Suggestive SNPs (2011) were encouraging and may aid further research aimed at identifying true positives. The genomic annotations, which were shown to be important for significant associations could be used to calculate a prior probability of trait-association. Operatively, this in turn could be used to adjust the *P*-value of suggestive SNPs, which could lead to more SNPs that pass the chosen significance threshold.

The newly included annotation of distance to TSS was quantitative, rather than binary and including the annotation explained an additional 20%, as judged by the McKelvey and Zavoina's pseudo-r^2, of the observed variation in Significant SNPs (2011). While the new final model, obtained after including distance to TSS, did include four additional annotations only one of them was significant and they did not contribute greatly to the pseudo-r^2 value. It is likely, that the model included more annotations, as the distance to TSS annotation added noise, as well as signal. This meant that including the additional genomic annotations further aided in explaining the observed variation.

5.4.1 Immune SNPs *vs.* Non-immune SNPs

The aim of the comparison of Immune *vs.* Non-Immune SNPs was to identify a set of genomic annotations that were significantly different between the two datasets, thereby allowing the drawing of immune specific conclusions. We compared the Final Model obtained for the Immune SNPs with the estimates obtained for the Non-immune SNPs using the same genomic annotations. The odds ratios obtained for the two datasets differed in many genomic annotations, but the most striking was for the positively selected genes. These genic regions were high-confidence orthologues in a multi-species alignment [115], which were then tested for positive selection [115]. However, as mentioned in the Methods section, these regions were not restricted according to their positive selection score, so that the regions analysed here highlight highly conserved regions. They were larger than the other conserved elements included in the analyses and also included pseudo genes and were significantly depleted for Immune SNPs. Depletion in conserved sites was shown by the significant odds ratios observed in the conserved regions identified in a 28 species alignment. The conserved transcription factor binding sites were significantly depleted for Immune SNPs and significantly different from the Non-immune SNPs. The depletion in these conserved sites corroborated the observed significant depletion in the positively selected genes. It is assumed that they have an important function in the cell so they were conserved over time [64], so that their disruption would likely be deleterious. The eQTLs were significantly more enriched for Immune SNPs than Non-immune SNPs, which could highlight potential involvement of eQTLs in the immune response. A recent study discovered that both *cis*- and *trans*-eQTLs were involved in newly identified cell-type specific networks in the pathogenesis of autoimmune diseases [155]. However, it has generally been accepted that all trait-associated SNPs were more likely to be eQTLs, and the enrichment in the Non-immune SNPs was therefore not surprising [110]. The significant difference in depletion in the closed chromatin annotations was also reasonable, if not expected, as the annotation was obtained from a lymphoblastoid cell line [126]. It was therefore encouraging to see that Immune SNPs were depleted in chromatin states associated with closed and not transcribed genes in a cell line co-ordinating immune response. The same reasoning is true for the Heterochromatin/low transcription regions, which were obtained from a lymphoblastoid cell after a comparison of nine different cell lines.

5.4.2 Cancer SNPs *vs.* Non-cancer SNPs

Since only 268 SNPs were associated to cancers, a very heterogeneous disease classification in itself, the identification of any genomic annotations that significantly influence association status would be surprising and encouraging. When the results of the regression model for the Cancer associated SNPs were compared to the results for the Non-cancer SNPs, five genomic annotations were significantly different between the two sets. Exons, conserved sites in primates, weak enhancers (proximal) were significantly more enriched in Cancer SNPs than Non-cancer SNPs while coding SNPs and intronic SNPs had lower odds ratios for Cancer SNPs. As mentioned before, a disruption in conserved sites is likely to be deleterious.

The results for the Cancer SNPs showed significant enrichment in the exons, but significant depletion in the coding and intronic SNPs. As mentioned before, the stepwise logistic regression approach analysed the genomic annotations in relation with each other, which explains results, which may seem contradictory at first glance. While exons are enriched in comparison with the other genomic annotations, coding SNPs were significantly depleted in comparison when all other genomic annotations were taken into account. The weak enhancer annotation was significantly enriched for Cancer SNPs and significantly depleted for Non-cancer SNPs. This genomic annotation was identified as regions with a distinct combination of histone modifications that were repeatedly associated with enhancer regions across nine different human cell lines [82]. The method identified four different states of enhancers that were divided according to the strength of their gene regulation and the observed distance between enhancers and the expressed genes. While the distance between the enhancers and the genes did not change between cell lines, the strength of the regulations did [82]. The enhancers analysed here were weak enhancers in the GM12878 lymphoblastoid cell line, but could be strong enhancers in a different cell line. The analysis therefore indicates that Cancer SNPs could lie within strong enhancers.

5.4.3 Normal Variation SNPs *vs.* Disease SNPs

We identified the genomic annotations influential on the association status of SNPs associated with normal variation traits and compared them with the genomic annotations that were identified as important for disease-associated SNPs. The comparison highlighted that there were 19 genomic annotations that were influential for both datasets, but two had significantly different impact on trait-association status. The exons had a significantly higher impact on the SNPs associated with normal variation than with diseases. The 3'UTRs were significantly depleted for the Normal Variation SNPs, while they were significantly enriched for the Disease SNPs. The other 17 annotations had odds ratios that were very similar between the two SNP categories. The common annotations significant for both categories showed that there are some common underlying biological mechanisms for the disease- and the normal variation-associated SNPs.

5.4.4 Pseudo-r^2 values

The McKelvey and Zavoina's pseudo-r^2 value was designed to be analogous of two interpretations of the traditional r^2 value obtained by linear regression. The full models varied between 18-42% of explained variability, while the base models containing only the genotyping arrays varied between 11-20%. These values may have been higher if validated trait-causing mutations were analysed. Although the false positive rate has decreased since multiple stage testing approaches have been implemented [156] and meta-analyses are continuously being performed to analyse trait-association across different arrays and studies [57, 63, 144, 157, 158], false positives may still exist in the trait-associated SNP datasets. Additionally, false negatives will undoubtedly be included in the background data. False negatives occur when a study does not have the required sample size to detect variants with modest effects that did not reach the genome-wide significance threshold [32, 58]. As in any analysis, this lack of information about true associations will have impacted the results, although there is no way of knowing how much and how the results would have differed if there was a clear separation between true and false positives. There is additionally an imbalance of non-associated SNPs *vs.* trait-associated SNPs (~3.5 million *vs.* a few hundreds or thousands). This disproportionate amount of non-associated variants will have skewed the results of logistic regression to be better at explaining non-associated than associated data.

5.4.5 Method discussion and future work

The logistic regression method is highly affected by co-linearity of the independent variables, as it is assumed that the included variables are independent from each other. If complete co-linearity of variables/annotations occurred in the analyses, an error was printed and the analysis stopped. While co-linearity could have occurred at certain instances, only some of the genomic annotations would be perfectly correlated with each other. An example of almost perfect co-linearity would be the position of genes downloaded from more than one database. While the databases would have some differences, the majority of the covered bases by the annotated genes would overlap with each other. It was therefore prudent to remove one set of genes from the multivariate analyses, which we did by excluding all genes from the OMIM database. For future studies it could be worthwhile to assess the partial co-linearity of the annotations by possibly by looking at a similarity matrix between annotations or analysing the correlation between annotations using a Spearman's rho correlation. The gained values could be used as a threshold for the inclusion of genomic annotations in the experimental model, *i.e.*, into the analysis itself. Alternatively, the stability of the final multivariate models could be assessed by cross-validation. With respect to univariate models it must be remembered that the variance explained in univariate analyses will not be additive in multivariate analyses.

The genotyping arrays were included as independent factors rather than one single co-variate. If there has been an annotation bias in the selection of variants on an array i.e. non-synonymous SNPs, this could have given certain genotyping arrays an erroneously inflated weight thereby decreasing the effect of any included genomic annotation. A genotyping array that could have potentially have been affected by this is the Ilumina 300 array given the biased selection of variants. Fitting the number of genotyping arrays containing a particular SNP, rather than each array individually, may allow for the correction for multiple testing without biasing the annotation. This is also a possible step that could be taken in future analyses.

The results from the trait-subset analyses could possibly be used to guide study designs on a trait class basis, where the results of GWAS analysing certain trait-categories would be analysed differently or the SNPs selection biased towards those annotations types of annotation enrich in immune versus non-immune traits for example. The follow up of GWAS results could be guided according to the importance of the genomic annotations from the different subsets, as it was shown that Immune-associated SNPs have a completely different genomic signature than, e.g., Cancer SNPs. While these analyses are great for the analysis of future GWAS, it is unclear if they could guide next generation sequencing studies. It is potentially possible to guide whole genome sequencing studies, but it is less likely that this could guide exome studies. The results presented in this thesis do include genomic areas outside of coding regions, so for the results to be valid in exome sequencing studies it would be recommended to perform additional logistic regression analyses with only coding areas.

6 APPLICATION OF METHODS TO OTHER DATA

6.1 Introduction

As the permutation and stepwise regression methods have produced promising results, which confirmed previous study results and were intuitive with enrichment and depletion in expected genomic regions, we investigated the applicability of our methods to two different datasets. The dataset used for the permutation analysis was a set of SNPs associated to gene expression levels in the Stockholm atherosclerosis gene expression study (STAGE) [83], which investigated gene expression levels in seven tissues of a cohort of cardiovascular disease patients. RNA was extracted from tissue biopsies of skeletal muscle, atherosclerotic arterial wall, internal mammary artery, liver, subcutaneous and visceral fat, and whole blood in 147 patients. The result of the study was the identification of several thousand SNPs associated with changing levels of gene expression in the different tissues. We analysed significant associations, *i.e.,* SNPs with a *P*-value that were below the genome-wide significance threshold. Experiments validating the effect on gene expression were not performed, so they are only suspected to be eQTLs and were therefore termed eSNPs. The investigation of the distribution of these eSNPs was an opportunity to use different datasets for the permutation analysis, as well as providing a thorough examination of the genomic distribution of potential eQTLs. The analysis of a SNP set originating from a single study meant that the background distribution was not as heterogeneous and taking account of different genotyping platforms was no longer needed. In previous permutation analyses (see Chapter 4) we strived to analyse the entire background of SNPs, which were analysed for trait-association. Using just one study allowed the creation of a background consisting of all SNPs tested for association to gene expression levels. The background distribution for the eSNPs was therefore changed accordingly for the analysis. The aim of the investigation was to identify differences between the distributions of GWAS hits and eSNPs. Furthermore, differences between tissue-specific eSNPs and eSNPs shared across different tissues were also investigated. If the results show significant differences in enrichment/depletion patterns between GWAS hits and eSNPs, it could imply that GWAS hits and eSNPs affect different pathways leading to phenotypes.

We additionally wanted to apply a regression model to test the effect of using a continuous distribution of *P*-values rather than an artificially created binary variable for trait-association status. The use of a continuous distribution allowed the analysis of all association *P*-values. This included more information than the analysis of a binary variable, as all SNPs had a value associated with it, informing on more SNPs. For this analysis we used the *P*-values obtained by a meta-analysis investigating the genetic components of height published by the Genetic Investigation of Anthropometric Traits (GIANT) consortium [57]. The *P*-values were used rather than the effect sizes, which might have been more informative, as these were not available at the time when the GIANT data was downloaded. This meta-analysis combined a total of 61 studies investigating height in different populations providing an in-depth analysis of the underlying genetic factors of height. Since a continuous variable was used, we applied linear rather than logistic regression.

6.2 Methods

6.2.1 STAGE eSNPs

A total of 109 patients rather than 147 patients had sufficient DNA, so the final set of analysed individuals was 109. For the analyses of the eSNPs data the background distribution was the intersection of the SNPs present in the dataset used for the permutations and logistic regression in the previous chapters, and the SNPs genotyped with the GenomeWideSNP_6 Affymetrix array used in the STAGE study that passed quality control. The intersection of these two datasets was used to ensure that an appropriate background distribution was used, against which the observed data was compared. The background distribution had to consider all the SNPs, which were tested for an association to changing gene expression levels. This meant that only the SNPs that passed quality control represented the background distribution. This set of background SNPs was termed **eSNP background** and was used for all analyses and comparisons that were performed for this dataset.

In all analyses outlined below, the eSNPs with *P*-values of association below the genome-wide significance threshold (*P*-value ≤ 5×10^{-8}) were coded as one, and the non-significant SNPs were coded as zero. All eSNPs were analysed with their LD partners, *i.e.,* SNPs that were in linkage disequilibrium (LD; $r^2 > 0.9$) with the significantly associated eSNP. This was similar to the analysis of trait-associated SNPs in the previous chapters. The permutations, which were run to analyse the distribution of eSNPs in 58 genomic annotations, were also performed 20,000 times, like in the previous permutation analyses. The eSNPs were only analysed on the autosomes, therefore disregarding the sex chromosomes. Table 6-1 shows a summary of the total number of analysed SNPs and the number of significant eSNPs per chromosome. A two-sample two-sided t-test (see page 37 of this thesis), assuming unequal variances, was employed to test for significant differences between the obtained odds ratios in the enrichment/depletion permutation analysis. The *P*-value was only judged as significant if it passed the Bonferroni corrected significance threshold, which took the 58 analysed annotations into account.

Two comparisons of different SNP sets were carried out for the eSNPs dataset. The first comparison analysed differences between the distributions in the genome of significant eSNPs and GWAS hits. The GWAS hits used in this analysis were downloaded from the NHGRI GWAS catalogue on 4th October 2012. The data contained 969 GWAS hits which were analysed like the trait-associated GWAS hits in Chapter 4 with the exception of a different background distribution. The number of significant eSNPs was 29,530, but 26,546 eSNPs were available in the previously compiled background of SNPs, due to changes in rs numbers. The 26,546 eSNPs were compared to 969 GWAS hits. GWAS hits and eSNPs were analysed with LD partners ($r^2 > 0.9$).

Table 6-1 – Distribution of SNPs and eSNPs across chromosomes
The number of SNPs and different eSNP datasets were summarised across chromosomes. The sex chromosomes were not analysed in this study.

Chromosome	Total Number of SNPs	Number of eSNPs	Number of Shared SNPs	Number of Tissue-Specific SNPs
Chr1	36875	2553	588	1965
Chr2	29481	2234	478	1756
Chr3	33959	1471	379	1092
Chr4	28865	1275	210	1065
Chr5	31715	1586	384	1202
Chr6	31819	2377	820	1557
Chr7	26617	1153	217	936
Chr8	27845	1138	232	906
Chr9	22608	1038	238	800
Chr10	26671	1392	426	966
Chr11	24446	1663	498	1165
Chr12	23956	1456	367	1089
Chr13	18000	652	156	496
Chr14	15402	851	147	704
Chr15	14780	884	230	654
Chr16	15008	855	213	642
Chr17	11064	999	303	696
Chr18	14504	450	88	362
Chr19	6349	944	296	648
Chr20	12902	576	143	433
Chr21	6575	470	124	346
Chr22	6135	529	163	366

The second set of comparisons was performed to investigate differences between shared and tissue-specific eSNPs, in terms of distribution across annotations. The tissue-specific eSNPs were significantly associated with gene expression in only one tissue, while the shared eSNPs were significantly affecting gene expression in at least two tissues. Of the total of 26,546 eSNPs, 6,700 were identified as shared eSNPs and 19,846 were tissue-specific eSNPs. All analyses were performed, as mentioned above, with 20,000 permutations and the eSNP background.

6.2.2 GIANT SNPs

The Genetic Investigation of Anthropometric Traits (GIANT) consortium performed a meta-analysis on three different human traits (height, body mass index, and body mass index adjusted for hip to waist ratio. Here, we have only investigated the data analysed for association to height. The height meta-analysis applied two stages of testing and an additional family-based analysis investigating SNPs for an association with height. The first stage combined 46 studies, which in total consisted of 133,653 individuals of recent European ancestry. The meta-analysis was performed by imputing 2,834,208 SNPs using the HapMap CEU II reference population. The second stage analysed an additional 50,074 individuals from 15 further studies, which allowed the replication of 180 of 207 previously significantly associated genetic loci [57]. The association data are publicly available and contained information on the SNP name, the alleles of the SNP, and the frequency of the trait-increasing allele in the HapMap CEU II population (Link at which the data is available: http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files). Additionally, it listed the *P*-value of association obtained after the performed meta-analysis and after the results were corrected for an inflation of test statistics using genomic control. It also listed the number of times a SNP was observed within the population sample.

The previously established dataset used in the permutations and the logistic regression analyses contained 2,469,014 SNPs of the total number of SNPs tested for height-association (2,834,208 SNPs). A stepwise regression model was employed to establish the most influential genomic annotations for height, as measured by a reduction in the AIC value (see Methods chapter of this thesis and Chapter 5). However, the difference to the logistic regression employed in Chapter 5, is that here a linear regression was used, since the dependent variable was the negative logarithm of the association *P*-value instead of a binary trait-association status (trait-associated or not). The analysed variable was therefore continuous rather than binary, which meant that linear regression rather than logistic regression was the appropriate method for the analysis. As the dependent variable was continuous, the returned estimate of any analysed annotation in the model is to be interpreted as explaining unit changes in the negative logarithm of the *P*-value, rather than determining trait-association status. Since linear regression was employed, pseudo-r^2 values were not calculated, but an r^2 value was extracted from the summary of the analyses. The total number of analysed SNPs was 2,469,014, as SNPs that were not in our background data were excluded from the analyses. This ensured the selection of the appropriate background distribution, and that a SNP was compared with only those SNPs that were tested for height-association.

6.3 Results

6.3.1 STAGE eSNPs

The STAGE eSNPs were analysed in two different comparisons. The dataset was analysed as a total set (All eSNPs) and compared to GWAS hits. The second analysis was the partition of All eSNPs into two subsets according to the number of tissues, in which the eSNPs were associated to changes in gene expression levels; Shared eSNPs and Tissue-specific eSNPs.

6.3.1.1 GWAS hits vs. All eSNPs

The results of the permutation analyses using the STAGE eSNPs were compared to GWAS hits that were present on the same genotyping array as the SNPs tested for gene expression. The results are shown in Figure 6-1, where a red star indicates eight significant differences between GWAS hits and eSNPs. Table 6-2 and Table 6-3 present the numerical results for the GWAS hits and the eSNPs, respectively. The tables show the number of overlaps for each genomic annotation with the observed data and the mean number of overlaps obtained for the permuted data. They also show the calculated odds ratio and its confidence interval as well as the obtained *P*-value, which was determined by the number of permutations that had the same or more extreme number of overlaps as the observed data.

The most important results of this comparison are the significant differences in the eQTL annotation and the OMIM morbid regions, as these were the two positive controls for the different datasets. The results in these annotations, therefore, added confidence to the remaining results. The GWAS hits (OR = 3.68 [2.71-5.08]) obtained an odds ratio that was twice as high as the odds ratio obtained by the eSNPs (OR = 1.99 [1.70-2.33]) in the OMIM morbid regions. While the enrichment pattern was the opposite in the eQTL annotation (OR for GWAS hits = 3.28 [2.36-4.50]; OR for eSNPs = 8.91 [6.68-10.36]). The overall results were that eSNPs showed more extreme and significant odds ratios of enrichment than the GWAS hits. The only annotation, which had a significant result for the GWAS hits, but not the eSNPs, was the heterochromatin/low state. However, there were many examples of annotations, which were significant for the eSNPs but not the GWAS hits. The mean of the odds ratios of All eSNPs was 2.64, while for GWAS hits it was 2.01. The eight significant differences between the eSNPs and the GWAS hits are in the 1 Kb and 5 Kb regions upstream of TSS, the OMIM morbid regions, the intronic SNPs, the eQTL annotation, positively selected genes, and regions associated with transcriptional elongation and weak transcription. The three annotations, which had obtained the highest odds ratios for All eSNPs besides eQTLs, 1 Kb upstream of TSS (OR = 4.28 [3.34-4.99]), 5 Kb upstream of TSS (OR = 4.22 [3.39-4.75]) and regions associated with transcriptional elongation (OR = 4.16 [3.46-4.71]). Splice sites and microsatellites also obtained odds ratios with realistic confidence intervals (OR for splice sites = 2.94 [1.41-9.01]; OR for microsatellites = 0.94 [0.51-2.22]), which was the first time for both of the annotations in any of the permutation analyses. However, neither of the annotations passed the Bonferroni corrected significance threshold and were not included on the graphs.

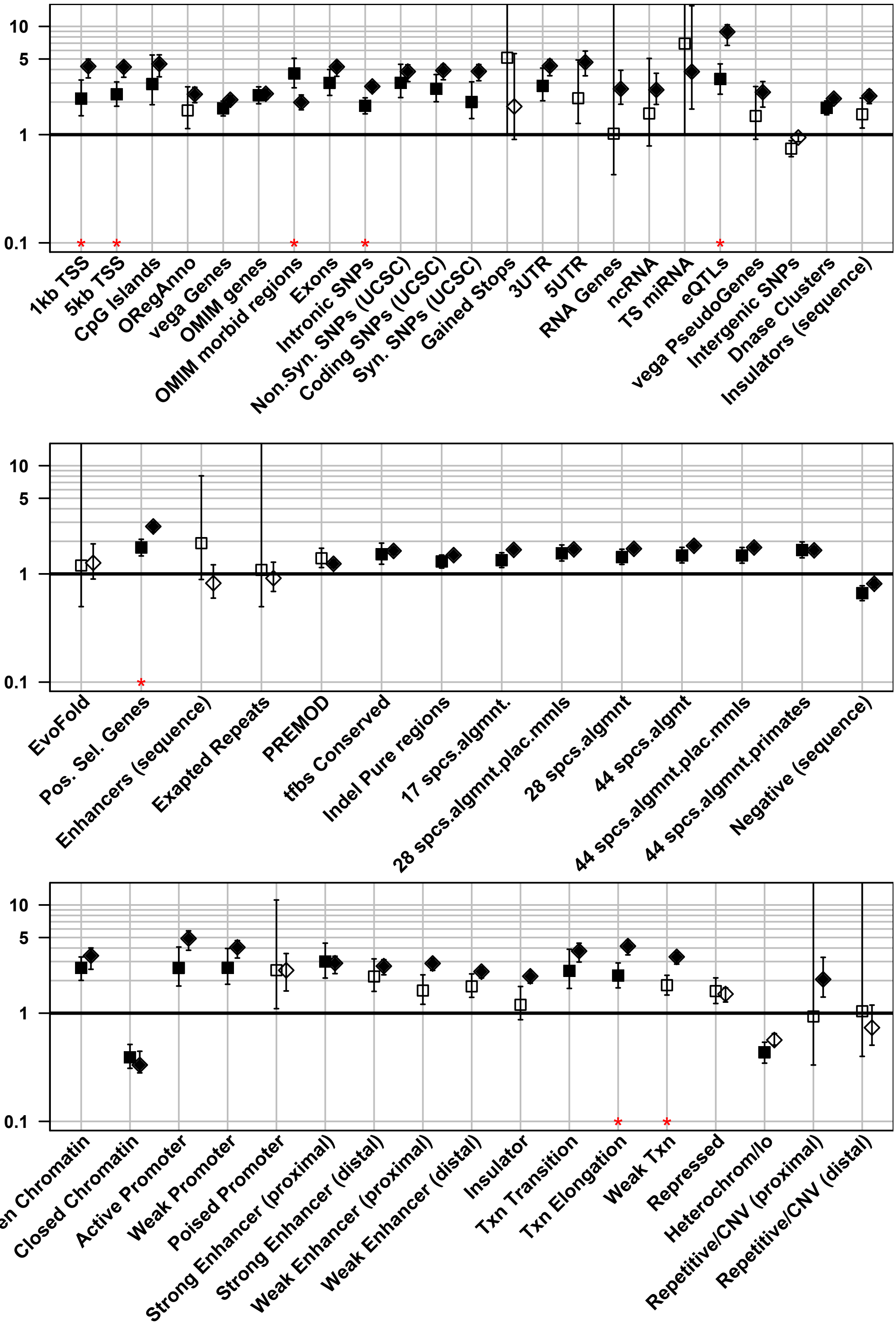


Figure 6-1 – Comparison of odds ratios obtained for GWAS hits and All eSNPs

A comparison of GWAS hits (n = 969, □) and All eSNPs (n = 26,546, ◇). Solid symbols indicate significance at multiple-test corrected P-value. Eight annotations are significantly different in their enrichment patterns. Red stars (✱) indicate significant differences. All P-values are corrected for multiple testing for the analysed genomic annotations. Top: Genic and regulatory regions. Middle: Conserved regions and evolutionary signatures. Bottom: Chromatin states and histone modifications.

Table 6-2 – Permutation results for the GWAS hits for the comparison with All STAGE eSNPs

This table summarises results for the GWAS hits showing the number of overlaps in the observed set (Real), the mean of the permuted hits (Permutation Mean), the calculated odds ratio and confidence interval (OR [LCI-HCI]) and the *P*-value for each of the annotations. Significant *P*-values in bold.

Annotation	Real	Permutation Mean	OR [LCI-HCI]	P-value
1 Kb TSS	87	42.55	2.15 [1.49-3.20]	**1.00×10^{-04}**
5 Kb TSS	223	108.82	2.36 [1.83-3.07]	**$< 5.00 \times 10^{-05}$**
CpG Islands	57	20.17	2.94 [1.89-5.44]	**5.00×10^{-05}**
ORegAnno	43	26.15	1.67 [1.14-2.77]	5.60×10^{-03}
vega Genes	515	381.57	1.75 [1.49-2.04]	**$< 5.00 \times 10^{-05}$**
OMIM genes	533	335.31	2.31 [1.93-2.77]	**$< 5.00 \times 10^{-05}$**
OMIM morbid regions	240	79.56	3.68 [2.71-5.08]	**$< 5.00 \times 10^{-05}$**
Exons	219	85.86	3.00 [2.30-4.00]	**$< 5.00 \times 10^{-05}$**
Intronic SNPs	599	452.83	1.85 [1.56-2.19]	**5.00×10^{-05}**
Non.Syn. SNPs (UCSC)	121	43.79	3.01 [2.20-4.47]	**$< 5.00 \times 10^{-05}$**
Coding SNPs (UCSC)	164	69.11	2.65 [2.01-3.59]	**$< 5.00 \times 10^{-05}$**
Syn. SNPs (UCSC)	73	37.98	2.00 [1.41-3.08]	**1.50×10^{-04}**
Gained Stops	3	0.59	5.14 [1.00-Infinity]	3.37×10^{-02}
3'UTR	123	47.59	2.82 [2.06-4.12]	**$< 5.00 \times 10^{-05}$**
5'UTR	24	11.21	2.17 [1.27-4.90]	3.75×10^{-03}
RNA Genes	3	2.94	1.02 [0.43-Infinity]	4.53×10^{-01}
ncRNA	15	9.62	1.57 [0.79-5.06]	1.05×10^{-01}
TS miRNA	2	0.29	6.94 [1.00-Infinity]	4.12×10^{-02}
eQTLs	234	85.85	3.28 [2.36-4.50]	**$< 5.00 \times 10^{-05}$**
vega PseudoGenes	30	20.38	1.49 [0.91-2.78]	5.93×10^{-02}
Intergenic SNPs	560	628.32	0.74 [0.63-0.88]	3.00×10^{-04}
DNase Clusters	534	396.42	1.77 [1.53-2.05]	**$< 5.00 \times 10^{-05}$**
Insulators (sequence)	83	55.54	1.54 [1.15-2.18]	2.25×10^{-03}
Within miRNA	0	0.01	0.00 [NA-NA]	**$< 5.00 \times 10^{-05}$**
Splice Sites	0	0.56	0.00 [0.00-NA]	**$< 5.00 \times 10^{-05}$**
Lost Stops	0	0.18	0.00 [0.00-NA]	**$< 5.00 \times 10^{-05}$**
Microsatellites	0	0.81	0.00 [0.00-NA]	**$< 5.00 \times 10^{-05}$**
EvoFold	3	2.51	1.20 [0.50-Infinity]	4.51×10^{-01}
Pos. Sel. Genes	424	298.20	1.75 [1.47-2.09]	**$< 5.00 \times 10^{-05}$**
Enhancers (sequence)	8	4.18	1.92 [0.89-8.06]	8.14×10^{-02}
Exapted Repeats	4	3.67	1.09 [0.50-Infinity]	4.90×10^{-01}
PREMOD	168	126.65	1.39 [1.15-1.73]	5.00×10^{-04}
tfbs Conserved	148	102.91	1.52 [1.23-1.92]	**$< 5.00 \times 10^{-05}$**
Indel Pure regions	464	401.07	1.30 [1.13-1.49]	**1.00×10^{-04}**
17 spec. algmt	313	254.92	1.34 [1.15-1.57]	**$< 5.00 \times 10^{-05}$**
28 spec. algmt plc.mmls	285	205.07	1.55 [1.32-1.85]	**$< 5.00 \times 10^{-05}$**
28 spec. algmt	310	239.39	1.43 [1.22-1.69]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt	321	242.99	1.48 [1.26-1.76]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt plc.mmls	297	222.86	1.48 [1.26-1.75]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt prim.	308	212.56	1.66 [1.41-1.97]	**$< 5.00 \times 10^{-05}$**
Negative (sequence)	456	554.32	0.66 [0.57-0.78]	**$< 5.00 \times 10^{-05}$**
Open Chromatin	701	483.75	2.62 [2.01-3.31]	**$< 5.00 \times 10^{-05}$**
Closed Chromatin	297	514.06	0.39 [0.31-0.51]	**$< 5.00 \times 10^{-05}$**
Active Promoter	84	33.95	2.61 [1.78-4.09]	**5.00×10^{-05}**
Weak Promoter	92	37.20	2.63 [1.85-3.96]	**$< 5.00 \times 10^{-05}$**
Poised Promoter	11	4.44	2.50 [1.10-11.11]	1.79×10^{-02}
Strong Enhancer (proximal)	113	40.84	3.00 [2.11-4.44]	**5.00×10^{-05}**
Strong Enhancer (distal)	92	44.44	2.18 [1.59-3.17]	3.50×10^{-04}
Weak Enhancer (proximal)	88	56.31	1.62 [1.21-2.26]	1.05×10^{-03}
Weak Enhancer (distal)	174	106.68	1.77 [1.40-2.31]	5.50×10^{-04}
Insulator	50	42.19	1.20 [0.87-1.76]	1.49×10^{-01}
Txn Transition	77	32.82	2.46 [1.69-3.90]	**$< 5.00 \times 10^{-05}$**
Txn Elongation	199	100.98	2.22 [1.71-2.91]	**$< 5.00 \times 10^{-05}$**
Weak Txn	302	193.53	1.81 [1.47-2.24]	6.50×10^{-04}
Repressed	148	98.16	1.60 [1.23-2.12]	8.50×10^{-04}
Heterochrom/low	667	809.90	0.43 [0.35-0.54]	**$< 5.00 \times 10^{-05}$**
Repetitive/CNV (proximal)	2	2.15	0.93 [0.33-Infinity]	3.87×10^{-01}
Repetitive/CNV (distal)	2	1.92	1.04 [0.40-Infinity]	4.45×10^{-01}

Table 6-3 – Permutation results for All STAGE eSNPs

This table summarises the results for All eSNPs showing the number of overlaps in the observed set (Real), the mean of the permuted hits (Permutation Mean), the calculated odds ratio and confidence interval (OR [LCI-HCI]) and the obtained *P*-value for each annotation. Significant *P*-values in bold.

Annotation	**Real**	**Permutation Mean**	**OR [LCI-HCI]**	**P-value**
1 Kb TSS	4346	1161.25	4.28 [3.34-4.99]	**$< 5.00 \times 10^{-05}$**
5 Kb TSS	9273	2996.71	4.22 [3.39-4.75]	**$< 5.00 \times 10^{-05}$**
CpG Islands	2327	554.79	4.50 [3.42-5.46]	**$< 5.00 \times 10^{-05}$**
ORegAnno	1649	723.23	2.36 [1.97-2.74]	**$< 5.00 \times 10^{-05}$**
vega Genes	15164	10300.26	2.10 [1.88-2.29]	**$< 5.00 \times 10^{-05}$**
OMIM genes	14813	9179.15	2.39 [2.12-2.62]	**$< 5.00 \times 10^{-05}$**
OMIM morbid regions	3950	2144.92	1.99 [1.70-2.33]	**$< 5.00 \times 10^{-05}$**
Exons	7756	2357.28	4.24 [3.46-4.76]	**$< 5.00 \times 10^{-05}$**
Intronic SNPs	18804	12356.07	2.79 [2.48-3.06]	**$< 5.00 \times 10^{-05}$**
Non.Syn. SNPs (UCSC)	4094	1207.13	3.83 [3.09-4.42]	**$< 5.00 \times 10^{-05}$**
Coding SNPs (UCSC)	6171	1902.31	3.92 [3.21-4.45]	**$< 5.00 \times 10^{-05}$**
Syn. SNPs (UCSC)	3610	1043.26	3.85 [3.15-4.44]	**$< 5.00 \times 10^{-05}$**
Gained Stops	28	15.38	1.82 [0.90-5.60]	5.64×10^{-02}
3'UTR	4854	1307.14	4.32 [3.50-4.94]	**$< 5.00 \times 10^{-05}$**
5'UTR	1387	309.87	4.67 [3.50-5.92]	**$< 5.00 \times 10^{-05}$**
RNA Genes	211	80.10	2.65 [1.91-3.93]	**$< 5.00 \times 10^{-05}$**
ncRNA	663	259.54	2.59 [1.90-3.69]	**$< 5.00 \times 10^{-05}$**
TS miRNA	31	8.13	3.82 [1.72-15.52]	**1.00×10^{-04}**
eQTLs	12324	2352.39	8.91 [6.68-10.36]	**$< 5.00 \times 10^{-05}$**
vega PseudoGenes	1266	528.07	2.47 [1.80-3.09]	**$< 5.00 \times 10^{-05}$**
Intergenic SNPs	16877	17246.25	0.94 [0.86-1.03]	8.88×10^{-02}
DNase Clusters	15833	10809.54	2.15 [1.99-2.27]	**$< 5.00 \times 10^{-05}$**
Insulators (sequence)	3218	1517.03	2.28 [1.93-2.58]	**$< 5.00 \times 10^{-05}$**
Within miRNA	2	0.24	8.50 [1.00-Infinity]	6.07×10^{-02}
Splice Sites	45	15.31	2.94 [1.41-9.01]	1.45×10^{-03}
Lost Stops	46	5.18	8.89 [3.29-Infinity]	**$< 5.00 \times 10^{-05}$**
Microsatellites	20	21.24	0.94 [0.51-2.22]	4.58×10^{-01}
EvoFold	87	68.86	1.26 [0.90-1.89]	9.64×10^{-02}
Pos. Sel. Genes	14598	8176.41	2.74 [2.44-3.01]	**$< 5.00 \times 10^{-05}$**
Enhancers (sequence)	91	110.84	0.82 [0.60-1.21]	1.53×10^{-01}
Exapted Repeats	91	99.72	0.91 [0.69-1.28]	2.81×10^{-01}
PREMOD	4173	3470.07	1.24 [1.15-1.34]	**$< 5.00 \times 10^{-05}$**
tfbs Conserved	4295	2807.51	1.63 [1.52-1.75]	**$< 5.00 \times 10^{-05}$**
Indel Pure regions	13566	10939.49	1.49 [1.42-1.57]	**$< 5.00 \times 10^{-05}$**
17 spec. algmt	9874	6943.60	1.67 [1.58-1.78]	**$< 5.00 \times 10^{-05}$**
28 spec. algmt plc.mmls	8235	5585.88	1.69 [1.58-1.81]	**$< 5.00 \times 10^{-05}$**
28 spec. algmt	9492	6526.30	1.71 [1.60-1.82]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt	10014	6630.26	1.82 [1.71-1.94]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt plc.mmls	9088	6071.62	1.76 [1.65-1.88]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt prim.	8373	5786.80	1.65 [1.54-1.77]	**$< 5.00 \times 10^{-05}$**
Negative (sequence)	13791	15165.28	0.81 [0.75-0.89]	**1.50×10^{-04}**
Open Chromatin	20526	13300.19	3.40 [2.54-4.00]	**$< 5.00 \times 10^{-05}$**
Closed Chromatin	7202	14028.49	0.33 [0.28-0.44]	**$< 5.00 \times 10^{-05}$**
Active Promoter	4013	931.43	4.90 [3.81-5.76]	**$< 5.00 \times 10^{-05}$**
Weak Promoter	3670	1009.57	4.06 [3.23-4.69]	**$< 5.00 \times 10^{-05}$**
Poised Promoter	286	115.62	2.49 [1.60-3.56]	2.50×10^{-04}
Strong Enhancer (proximal)	2992	1115.53	2.90 [2.32-3.36]	**$< 5.00 \times 10^{-05}$**
Strong Enhancer (distal)	3069	1216.51	2.72 [2.26-3.12]	**$< 5.00 \times 10^{-05}$**
Weak Enhancer (proximal)	3969	1526.03	2.88 [2.48-3.21]	**$< 5.00 \times 10^{-05}$**
Weak Enhancer (distal)	6079	2897.83	2.42 [2.11-2.67]	**$< 5.00 \times 10^{-05}$**
Insulator	2390	1146.15	2.19 [1.89-2.46]	**$< 5.00 \times 10^{-05}$**
Txn Transition	3064	892.80	3.75 [2.97-4.43]	**$< 5.00 \times 10^{-05}$**
Txn Elongation	8715	2788.14	4.16 [3.46-4.71]	**$< 5.00 \times 10^{-05}$**
Weak Txn	12020	5301.61	3.32 [2.84-3.66]	**$< 5.00 \times 10^{-05}$**
Repressed	3779	2639.21	1.50 [1.27-1.71]	1.30×10^{-03}
Heterochrom/low	19686	22180.76	0.56 [0.51-0.65]	2.50×10^{-04}
Repetitive/CNV (proximal)	121	59.07	2.05 [1.41-3.28]	**1.00×10^{-04}**
Repetitive/CNV (distal)	38	51.70	0.73 [0.51-1.19]	8.49×10^{-02}

6.3.1.2 Shared eSNPs vs. Tissue-specific eSNPs

The odds ratios of the Shared eSNPs and the Tissue-specific eSNPs obtained by permutations are shown in Table 6-4 and Table 6-5, which list the overlaps for each annotation, the mean number of overlaps in permutations, the calculated odds ratios and their confidence intervals and the obtained *P*-value. The results for the Shared eSNPs *vs.* Tissue-specific eSNPs are compared with each other in Figure 6-2, where a red star indicates significant differences between the two sets. The majority of the annotations show significantly different odds ratios for the two datasets. The Shared eSNPs have more extreme odds ratios in all annotations, where the odds ratios were significant at the Bonferroni corrected threshold. However, the trend between the two datasets is the same. Only one genomic annotation, the repetitive/CNV (distal) sites, has an odds ratio of enrichment for the Tissue-specific eSNPs when it was depleted in the Shared SNPs. The mean of the odds ratios for the Shared eSNPs is 4.49 and for the Tissue-specific eSNPs is 2.15. The Shared eSNPs therefore have on average a much higher odds ratios than the Tissue-specific eSNPs, All eSNPs or GWAS hits. The highest odds ratio was obtained for the eQTL annotation for the Shared eSNPs (OR = 27.99 [17.11-37.29]), which suggests that they contain a larger proportion of 'true' eQTLs. The likelihood of detecting tissue specific eQTLs is much lower than detecting eQTLs that are affecting several tissues (*i.e.*, the shared eSNPs), as tissue specific studies will have less power. It is therefore possible that the Shared eSNPs are more represented in the eQTL annotation as there was more power to detect them, rather than representing a greater proportion of true eQTLs. This will be discussed further in the Discussion section of this chapter.

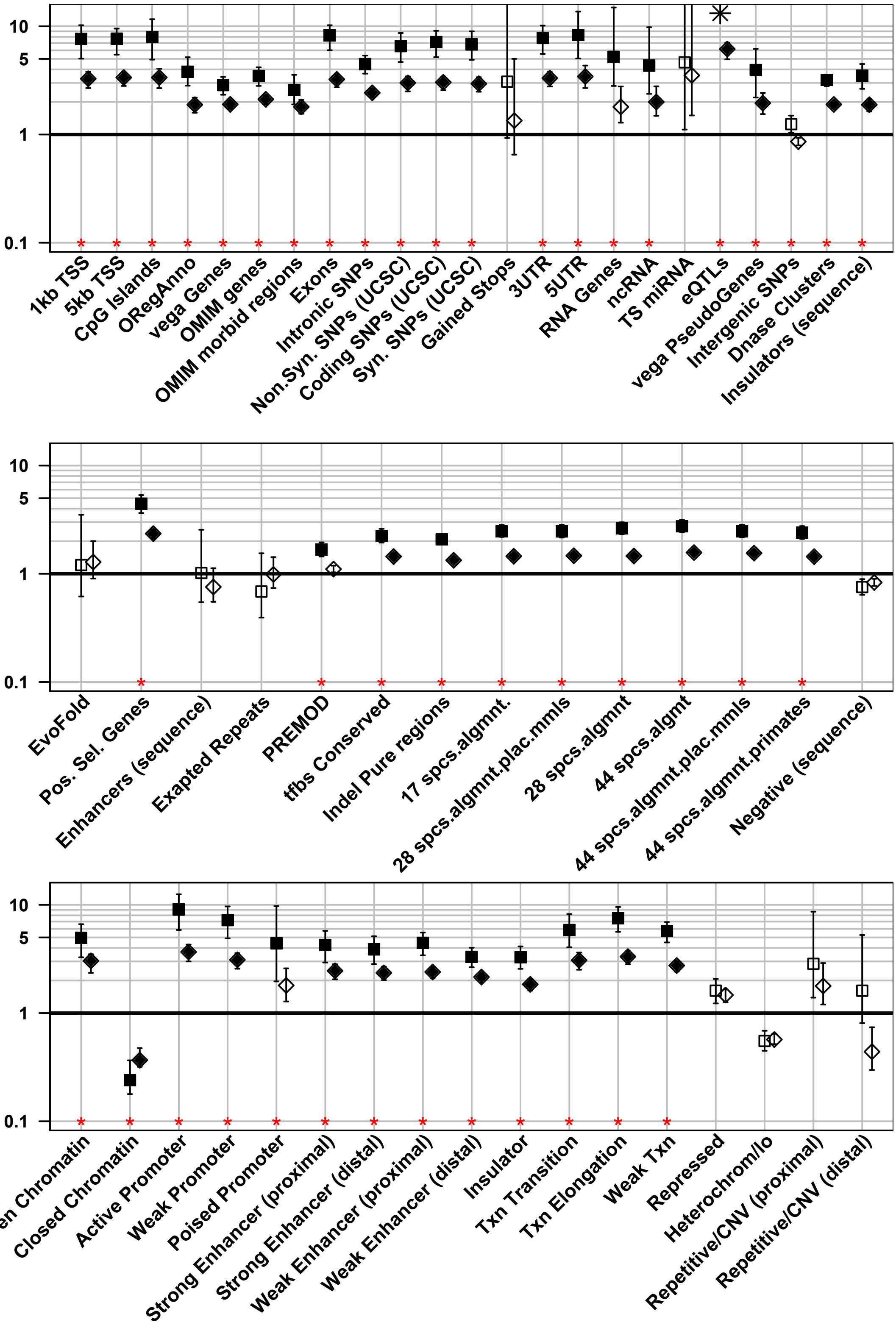


Figure 6-2 – Shared eSNPs *vs.* Tissue-specific eSNPs

A comparison of Shared eSNPs (*n* = 6,700, □) and Tissue-specific eSNPs (*n* = 19,846, ◇). Red stars (✱) indicate significant differences between the odds ratios. Solid symbols indicate significance at multiple corrected *P*-value. A black star (✳) indicates the odds ratio for eQTLs in the Shared eSNPs (27.99 [17.11-37.29]) data that is greater than the maximum of the graph. Top: Genic and regulatory regions. Middle: Conserved regions and evolutionary signatures. Bottom: Chromatin states and histone modifications.

Table 6-4 – Permutation results for Shared STAGE eSNPs

This table summarises the results for the Shared eSNPs showing the number of overlaps in the observed set (Real), the mean of the permuted hits (Permutation Mean), the odds ratio and confidence interval (OR [LCI-HCI]) and the *P*-value for each annotation. Significant *P*-values in bold.

Annotation	**Real**	**Permutation Mean**	**OR [LCI-HCI]**	**P-value**
1 Kb TSS	1787	303.63	7.66 [5.03-10.23]	**$< 5.00 \times 10^{-05}$**
5 Kb TSS	3372	779.75	7.69 [5.47-9.53]	**$< 5.00 \times 10^{-05}$**
CpG Islands	1021	148.04	7.96 [4.92-11.63]	**$< 5.00 \times 10^{-05}$**
ORegAnno	662	187.60	3.81 [2.83-5.18]	**$< 5.00 \times 10^{-05}$**
vega Genes	4324	2605.72	2.86 [2.33-3.41]	**$< 5.00 \times 10^{-05}$**
OMIM genes	4355	2333.71	3.47 [2.81-4.18]	**$< 5.00 \times 10^{-05}$**
OMIM morbid regions	1249	546.68	2.58 [1.90-3.56]	**$< 5.00 \times 10^{-05}$**
Exons	3040	614.27	8.23 [6.01-10.23]	**$< 5.00 \times 10^{-05}$**
Intronic SNPs	5346	3138.10	4.48 [3.65-5.37]	**$< 5.00 \times 10^{-05}$**
Non.Syn. SNPs (UCSC)	1637	314.88	6.56 [4.68-8.67]	**$< 5.00 \times 10^{-05}$**
Coding SNPs (UCSC)	2438	495.91	7.16 [5.19-9.08]	**$< 5.00 \times 10^{-05}$**
Syn. SNPs (UCSC)	1498	272.45	6.79 [4.90-8.99]	**$< 5.00 \times 10^{-05}$**
Gained Stops	13	4.23	3.08 [0.93-Infinity]	3.98×10^{-02}
3'UTR	1964	338.34	7.80 [5.56-10.15]	**$< 5.00 \times 10^{-05}$**
5'UTR	625	81.92	8.31 [5.04-13.68]	**$< 5.00 \times 10^{-05}$**
RNA Genes	103	19.97	5.22 [2.81-14.93]	**$< 5.00 \times 10^{-05}$**
ncRNA	283	67.59	4.33 [2.38-9.81]	**$< 5.00 \times 10^{-05}$**
TS miRNA	10	2.16	4.64 [1.11-Infinity]	1.83×10^{-02}
eQTLs	4946	613.22	27.99 [17.11-37.29]	**$< 5.00 \times 10^{-05}$**
vega PseudoGenes	526	141.69	3.94 [2.20-6.19]	**$< 5.00 \times 10^{-05}$**
Intergenic SNPs	4667	4341.70	1.25 [1.04-1.49]	7.95×10^{-03}
DNase Clusters	4614	2739.51	3.20 [2.80-3.55]	**$< 5.00 \times 10^{-05}$**
Insulators (sequence)	1196	390.87	3.51 [2.64-4.48]	**$< 5.00 \times 10^{-05}$**
Within miRNA	0	0.07	0.00 [0.00-NA]	**$< 5.00 \times 10^{-05}$**
Splice Sites	28	4.16	6.76 [1.87-Infinity]	1.10×10^{-03}
Lost Stops	21	1.39	15.21 [3.01-Infinity]	**$< 5.00 \times 10^{-05}$**
Microsatellites	7	5.63	1.24 [0.41-Infinity]	3.28×10^{-01}
EvoFold	21	17.44	1.20 [0.62-3.51]	2.93×10^{-01}
Pos. Sel. Genes	4465	2076.46	4.45 [3.65-5.34]	**$< 5.00 \times 10^{-05}$**
Enhancers (sequence)	28	27.51	1.02 [0.55-2.55]	4.54×10^{-01}
Exapted Repeats	17	24.75	0.69 [0.39-1.55]	1.53×10^{-01}
PREMOD	1341	871.68	1.67 [1.44-1.95]	**$< 5.00 \times 10^{-05}$**
tfbs Conserved	1398	707.01	2.24 [1.94-2.60]	**$< 5.00 \times 10^{-05}$**
Indel Pure regions	3971	2755.27	2.08 [1.89-2.33]	**$< 5.00 \times 10^{-05}$**
17 spec. algmt	3122	1746.02	2.48 [2.19-2.83]	**$< 5.00 \times 10^{-05}$**
28 spec. algmt plc.mmls	2648	1405.10	2.46 [2.16-2.83]	**$< 5.00 \times 10^{-05}$**
28 spec. algmt	3091	1646.63	2.63 [2.32-2.99]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt	3197	1670.35	2.75 [2.43-3.15]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt plc.mmls	2823	1527.32	2.47 [2.17-2.84]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt prim.	2678	1455.98	2.40 [2.10-2.76]	**$< 5.00 \times 10^{-05}$**
Negative (sequence)	3347	3816.17	0.75 [0.64-0.89]	2.50×10^{-04}
Open Chromatin	5574	3347.33	4.96 [3.28-6.62]	**$< 5.00 \times 10^{-05}$**
Closed Chromatin	1421	3544.92	0.24 [0.18-0.37]	**$< 5.00 \times 10^{-05}$**
Active Promoter	1699	242.57	9.04 [5.86-12.52]	**$< 5.00 \times 10^{-05}$**
Weak Promoter	1518	261.29	7.22 [4.90-9.67]	**$< 5.00 \times 10^{-05}$**
Poised Promoter	134	30.91	4.40 [1.96-9.75]	**5.00×10^{-05}**
Strong Enhancer (proximal)	1077	288.35	4.26 [2.93-5.75]	**$< 5.00 \times 10^{-05}$**
Strong Enhancer (distal)	1064	311.40	3.87 [2.84-5.13]	**$< 5.00 \times 10^{-05}$**
Weak Enhancer (proximal)	1450	391.03	4.46 [3.42-5.54]	**$< 5.00 \times 10^{-05}$**
Weak Enhancer (distal)	1954	741.28	3.31 [2.65-4.02]	**$< 5.00 \times 10^{-05}$**
Insulator	871	292.02	3.28 [2.56-4.13]	**$< 5.00 \times 10^{-05}$**
Txn Transition	1163	231.98	5.86 [4.05-8.22]	**$< 5.00 \times 10^{-05}$**
Txn Elongation	3180	718.51	7.52 [5.63-9.55]	**$< 5.00 \times 10^{-05}$**
Weak Txn	3984	1364.94	5.73 [4.49-6.93]	**$< 5.00 \times 10^{-05}$**
Repressed	1026	676.50	1.61 [1.23-2.07]	2.50×10^{-03}
Heterochrom/low	4897	5567.66	0.55 [0.45-0.69]	6.50×10^{-04}
Repetitive/CNV (proximal)	43	15.15	2.85 [1.39-8.65]	1.15×10^{-03}
Repetitive/CNV (distal)	21	13.06	1.61 [0.81-5.26]	1.04×10^{-01}

Table 6-5 – Permutation results for Tissue-specific STAGE eSNPs

This table summarises the results for the Tissue-specific eSNPs, including the number of overlaps in the observed set (Real), the mean of the permuted hits (Permutation Mean), the calculated odds ratio and confidence interval (OR [LCI-HCI]) and the obtained *P*-value for each of the genomic annotations. Significant *P*-values in bold.

Annotation	Real	Permutation Mean	OR [LCI-HCI]	P-value
1 Kb TSS	2559	857.62	3.28 [2.69-3.80]	**$< 5.00 \times 10^{-05}$**
5 Kb TSS	5901	2216.96	3.36 [2.81-3.75]	**$< 5.00 \times 10^{-05}$**
CpG Islands	1306	406.75	3.37 [2.68-4.05]	**$< 5.00 \times 10^{-05}$**
ORegAnno	987	535.62	1.89 [1.59-2.19]	**$< 5.00 \times 10^{-05}$**
vega Genes	10840	7694.54	1.90 [1.73-2.05]	**$< 5.00 \times 10^{-05}$**
OMIM genes	10458	6845.43	2.12 [1.90-2.30]	**$< 5.00 \times 10^{-05}$**
OMIM morbid regions	2701	1598.24	1.80 [1.56-2.09]	**$< 5.00 \times 10^{-05}$**
Exons	4716	1743.01	3.24 [2.73-3.61]	**$< 5.00 \times 10^{-05}$**
Intronic SNPs	13458	9217.97	2.43 [2.19-2.64]	**$< 5.00 \times 10^{-05}$**
Non.Syn. SNPs (UCSC)	2457	892.25	3.00 [2.50-3.44]	**$< 5.00 \times 10^{-05}$**
Coding SNPs (UCSC)	3733	1406.40	3.04 [2.57-3.42]	**$< 5.00 \times 10^{-05}$**
Syn. SNPs (UCSC)	2112	770.81	2.95 [2.49-3.38]	**$< 5.00 \times 10^{-05}$**
Gained Stops	15	11.16	1.34 [0.65-5.00]	2.27×10^{-01}
3'UTR	2890	968.80	3.32 [2.78-3.78]	**$< 5.00 \times 10^{-05}$**
5'UTR	762	227.96	3.44 [2.69-4.34]	**$< 5.00 \times 10^{-05}$**
RNA Genes	108	60.13	1.80 [1.29-2.78]	4.00×10^{-04}
ncRNA	380	191.95	2.00 [1.49-2.79]	**5.00×10^{-05}**
TS miRNA	21	5.97	3.52 [1.50-21.02]	1.50×10^{-03}
eQTLs	7378	1739.17	6.16 [4.95-7.07]	**$< 5.00 \times 10^{-05}$**
vega PseudoGenes	740	386.37	1.95 [1.54-2.42]	**1.50×10^{-04}**
Intergenic SNPs	12210	12904.55	0.86 [0.79-0.93]	8.50×10^{-04}
DNase Clusters	11219	8070.03	1.90 [1.78-2.00]	**$< 5.00 \times 10^{-05}$**
Insulators (sequence)	2022	1126.16	1.89 [1.64-2.13]	**$< 5.00 \times 10^{-05}$**
Within miRNA	2	0.17	11.96 [1.00-Infinity]	3.42×10^{-02}
Splice Sites	17	11.15	1.52 [0.74-5.67]	1.40×10^{-01}
Lost Stops	25	3.80	6.59 [2.50-Infinity]	**$< 5.00 \times 10^{-05}$**
Microsatellites	13	15.62	0.83 [0.45-2.17]	3.21×10^{-01}
EvoFold	66	51.42	1.28 [0.90-2.00]	8.69×10^{-02}
Pos. Sel. Genes	10133	6099.94	2.35 [2.12-2.56]	**$< 5.00 \times 10^{-05}$**
Enhancers (sequence)	63	83.33	0.76 [0.55-1.13]	7.30×10^{-02}
Exapted Repeats	74	74.97	0.99 [0.74-1.42]	4.67×10^{-01}
PREMOD	2832	2598.39	1.10 [1.03-1.19]	$2.55 \times 10\text{-}^{03}$
tfbs Conserved	2897	2100.50	1.44 [1.35-1.55]	**$< 5.00 \times 10^{-05}$**
Indel Pure regions	9595	8184.22	1.33 [1.27-1.40]	**$< 5.00 \times 10^{-05}$**
17 spec. algmt	6752	5197.59	1.45 [1.38-1.54]	**$< 5.00 \times 10^{-05}$**
28 spec. algmt plc.mmls	5587	4180.78	1.47 [1.38-1.56]	**$< 5.00 \times 10^{-05}$**
28 spec. algmt	6401	4879.67	1.46 [1.38-1.55]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt	6817	4959.90	1.57 [1.48-1.66]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt plc.mmls	6265	4544.29	1.55 [1.46-1.65]	**$< 5.00 \times 10^{-05}$**
44 spec. algmt prim.	5695	4330.83	1.44 [1.35-1.53]	**$< 5.00 \times 10^{-05}$**
Negative (sequence)	10444	11349.11	0.83 [0.77-0.90]	4.00×10^{-04}
Open Chromatin	14952	9952.86	3.04 [2.35-3.50]	**$< 5.00 \times 10^{-05}$**
Closed Chromatin	5781	10483.57	0.37 [0.32-0.47]	**5.00×10^{-05}**
Active Promoter	2314	688.86	3.67 [3.00-4.29]	**$< 5.00 \times 10^{-05}$**
Weak Promoter	2152	748.29	3.10 [2.57-3.57]	**$< 5.00 \times 10^{-05}$**
Poised Promoter	152	84.70	1.80 [1.28-2.59]	2.50×10^{-03}
Strong Enhancer (proximal)	1915	827.18	2.46 [2.05-2.83]	**$< 5.00 \times 10^{-05}$**
Strong Enhancer (distal)	2005	905.11	2.35 [2.02-2.69]	**$< 5.00 \times 10^{-05}$**
Weak Enhancer (proximal)	2519	1135.00	2.40 [2.11-2.67]	**$< 5.00 \times 10^{-05}$**
Weak Enhancer (distal)	4125	2156.55	2.15 [1.91-2.36]	**$< 5.00 \times 10^{-05}$**
Insulator	1519	854.13	1.84 [1.62-2.07]	**$< 5.00 \times 10^{-05}$**
Txn Transition	1901	660.82	3.08 [2.52-3.62]	**$< 5.00 \times 10^{-05}$**
Txn Elongation	5535	2069.63	3.32 [2.83-3.74]	**$< 5.00 \times 10^{-05}$**
Weak Txn	8036	3936.66	2.75 [2.42-3.02]	**$< 5.00 \times 10^{-05}$**
Repressed	2753	1962.71	1.47 [1.26-1.65]	1.30×10^{-03}
Heterochrom/low	14789	16613.10	0.57 [0.52-0.64]	2.50×10^{-04}
Repetitive/CNV (proximal)	78	43.92	1.78 [1.20-2.90]	2.35×10^{-03}
Repetitive/CNV (distal)	17	38.64	0.44 [0.30-0.74]	1.35×10^{-03}

6.3.2 GIANT consortium data

The results of the linear regression performed using the negative logarithm of the *P*-value of association in the meta-analysis investigating height are shown in Figure 6-3. The final model contained 51 genomic annotations, which explained some of the variability towards the observed *P*-values of association for height. This is a very large number of annotations included in the model. Roughly speaking, the stepwise regression process eliminates those genomic annotations, which carry redundant information as explained in more detail earlier. The analysis for the GIANT dataset removed only three annotations (TS miRNA, CpG islands and exapted repeats) from the full model that were not informative, when other annotations were included in the model, while all the others carried additional information. The r^2 value of the final model was extracted from the summary with a value of 0.03 (0.027), meaning that most of the variation in –log(*P*-value of association) remained unexplained. The model, which contained only the genotyping arrays, had an r^2 value of 0.00 (0.0004). Of the 51 annotations included in the final model, 49 were significant at $P \leq 0.05$. The majority of these annotations had odds ratios with values indicative of enrichment, while 13 of the annotations had odds ratios that indicated depletion of height-associated SNPs. These 13 annotations were distance to TSS, closed chromatin, RNA genes, gained stops, coding SNPs, 1 Kb upstream of TSS, splice sites, ncRNA, exons, within miRNA, evofold regions, lost stops, and insulators. These regions were depleted of height-associated SNPs relative to the other 36 annotations that were enriched for height-associated SNPs and when these were included in the model. All the annotations had a very small effect on the negative logarithm of the *P*-value of association, as judged by their odds ratios. The highest odds ratio observed was for eQTLs (ORs = 1.24 [1.23-1.24]) and the lowest for within miRNA binding sites (ORs = 0.66 [0.54-0.81]). While the values of the odds ratios may be modest, the *P*-values of the annotations in the model are very significant (see Table 6-6). Height-associated SNPs overlapped preferentially with the region between the 1 Kb and 5 Kb region upstream of TSS, which could indicate a preference to a specific type of promoters. This was suggested by a combination of the negative β-coefficients (indicating relative depletion) of 1 Kb upstream of the TSS and the positive β-coefficients (indicating relative enrichment) of the 5 Kb upstream of the TSS. Additionally, the distance to TSS annotation obtained a very high negative β-coefficient, which indicated that SNPs closer to the TSS were more likely to be height-associated. The distance was calculated as the absolute value of the minimum distance between a SNP and its nearest TSS, which did not discriminate between SNPs up- or downstream of the nearest TSS. This separation could have been performed with the inclusion of an additional annotation and could be performed in future experiments. The genomic annotation with the highest impact on the *P*-value of height-associated SNPs was the eQTL annotation. The β-coefficient for that annotation was 82.58 and an odds ratio of 1.32 [1.32-1.33], the highest odds ratio in that analysis.

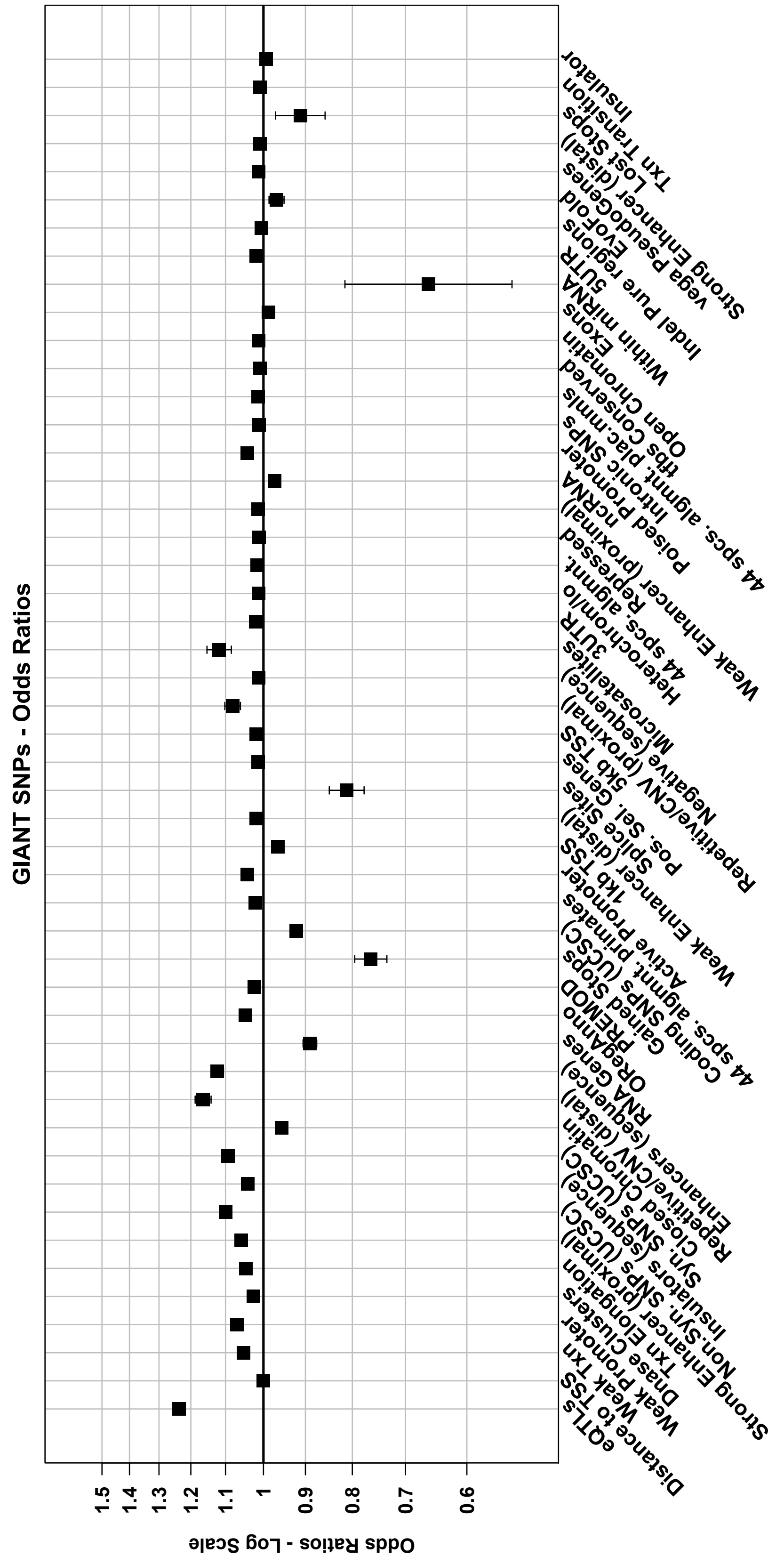


Figure 6-3 – Odds ratios for GIANT height data

This figure shows the significant odds ratios ($P \leq 0.05$) for linear regression in the data for height from the GIANT consortium ranked after increasing *P*-value of the annotation in the model.

Table 6-6 – GIANT height linear regression results

This table presents the results for the height-associated SNPs including the estimate of the effect, its standard error, the β-coefficient, the calculated odds ratio with its confidence interval, and the *P*-value of the estimate in the final model for each of the included genomic annotations in the final model. Significant *P*-values in bold.

Annotation	Estimate	Std. Error	β	OR [LCI-HCI]	*P*-value
Intercept	0.44	0.00	127.14	1.55 [1.54-1.57]	$0.00 \times 10^{+00}$
Affymetrix_250k_Nsp	0.00	0.00	0.43	1.00 [0.99-1.01]	6.71×10^{-01}
Affymetrix_250k_Sty	0.00	0.00	0.46	1.00 [0.99-1.01]	6.45×10^{-01}
Affymetrix_5.0	0.01	0.00	1.60	1.01 [1.00-1.01]	1.09×10^{-01}
Affymetrix_6.0	-0.01	0.00	-6.89	0.99 [0.99-0.99]	**5.46×10^{-12}**
Affymetrix_10k	0.00	0.01	0.45	1.00 [0.99-1.02]	6.56×10^{-01}
Affymetrix_50k.1	-0.01	0.00	-1.58	0.99 [0.99-1.00]	1.14×10^{-01}
Affymetrix_50k.2	0.01	0.00	2.07	1.01 [1.00-1.02]	**3.87×10^{-02}**
Illumina_300	0.02	0.00	7.05	1.02 [1.01-1.02]	**1.82×10^{-12}**
Illumina_550	0.01	0.00	3.85	1.01 [1.01-1.02]	**1.18×10^{-04}**
Illumina_650	-0.02	0.00	-4.69	0.98 [0.98-0.99]	**2.67×10^{-06}**
Perlegen	0.01	0.00	4.23	1.01 [1.00-1.01]	**2.35×10^{-05}**
eQTLs	0.21	0.00	95.08	1.24 [1.23-1.24]	$0.00 \times 10^{+00}$
Weak Txn	0.05	0.00	29.24	1.05 [1.05-1.05]	**7.47×10^{-188}**
Open Chromatin	0.01	0.00	4.31	1.01 [1.01-1.02]	**1.60×10^{-05}**
44 specs. algmt. plac. mmls	0.01	0.00	5.23	1.01 [1.01-1.02]	**1.72×10^{-07}**
Weak Promoter	0.07	0.00	23.04	1.07 [1.06-1.08]	**2.05×10^{-117}**
Distance to TSS	0.00	0.00	-33.66	1.00 [1.00-1.00]	**2.62×10^{-248}**
DNase Clusters	0.02	0.00	21.05	1.03 [1.02-1.03]	**2.28×10^{-98}**
Txn Elongation	0.04	0.00	19.49	1.05 [1.04-1.05]	**1.27×10^{-84}**
Strong Enhancer (proximal)	0.06	0.00	19.23	1.06 [1.05-1.06]	**2.06×10^{-82}**
Insulators (sequence)	0.04	0.00	17.59	1.04 [1.04-1.04]	**2.95×10^{-69}**
PREMOD	0.02	0.00	13.23	1.02 [1.02-1.03]	**5.93×10^{-40}**
ORegAnno	0.04	0.00	13.28	1.05 [1.04-1.05]	**3.22×10^{-40}**
44 specs. algmt. primates	0.02	0.00	12.40	1.02 [1.02-1.02]	**2.74×10^{-35}**
Repetitive/CNV (distal)	0.15	0.01	14.86	1.16 [1.14-1.19]	**6.32×10^{-50}**
Non.Syn. SNPs (UCSC)	0.10	0.01	18.03	1.10 [1.09-1.11]	**1.20×10^{-72}**
Enhancers (sequence)	0.12	0.01	14.83	1.12 [1.11-1.14]	**9.44×10^{-50}**
Active Promoter	0.04	0.00	12.38	1.04 [1.03-1.05]	**3.26×10^{-35}**
Closed Chromatin	-0.05	0.00	-16.52	0.95 [0.95-0.96]	**2.65×10^{-61}**
Weak Enhancer (distal)	0.02	0.00	9.85	1.02 [1.01-1.02]	**7.04×10^{-23}**
RNA Genes	-0.12	0.01	-13.36	0.89 [0.87-0.91]	**9.91×10^{-41}**
Gained Stops	-0.27	0.02	-13.09	0.76 [0.73-0.80]	**3.61×10^{-39}**
Syn. SNPs (UCSC)	0.09	0.01	17.52	1.09 [1.08-1.10]	**1.03×10^{-68}**
Coding SNPs (UCSC)	-0.08	0.01	-12.88	0.92 [0.91-0.93]	**5.98×10^{-38}**
Pos. Sel. Genes	0.01	0.00	8.02	1.01 [1.01-1.02]	**1.06×10^{-15}**
Negative (sequence)	0.01	0.00	7.16	1.01 [1.01-1.02]	**8.24×10^{-13}**
Splice Sites	-0.21	0.02	-9.38	0.81 [0.78-0.85]	**6.73×10^{-21}**
44 specs. algmt.	0.02	0.00	6.46	1.02 [1.01-1.02]	**1.08×10^{-10}**
Repetitive/CNV (proximal)	0.08	0.01	7.77	1.08 [1.06-1.10]	**8.10×10^{-15}**
Heterochrom/lo	0.01	0.00	6.83	1.01 [1.01-1.02]	**8.31×10^{-12}**
Repressed	0.01	0.00	6.17	1.01 [1.01-1.02]	**6.67×10^{-10}**
Microsatellites	0.11	0.02	7.13	1.12 [1.08-1.15]	**1.01×10^{-12}**
Intronic SNPs	0.01	0.00	5.33	1.01 [1.01-1.02]	**9.73×10^{-08}**
1 Kb TSS	-0.04	0.00	-11.16	0.96 [0.96-0.97]	**6.37×10^{-29}**

Annotation	Estimate	Std. Error	β	OR [LCI-HCI]	*P*-value
5 Kb TSS	0.02	0.00	7.94	1.02 [1.01-1.02]	**2.08×10^{-15}**
Weak Enhancer (proximal)	0.01	0.00	5.93	1.01 [1.01-1.02]	**2.95×10^{-09}**
3'UTR	0.02	0.00	6.99	1.02 [1.01-1.03]	**2.75×10^{-12}**
Poised Promoter	0.04	0.01	5.43	1.04 [1.03-1.06]	**5.71×10^{-08}**
ncRNA	-0.03	0.01	-5.54	0.97 [0.96-0.98]	**3.09×10^{-08}**
tfbs Conserved	0.01	0.00	4.86	1.01 [1.01-1.01]	**1.17×10^{-06}**
Indels Pure regions	0.00	0.00	3.79	1.00 [1.00-1.01]	**1.52×10^{-04}**
Within miRNA	-0.41	0.11	-3.87	0.66 [0.54-0.81]	**1.08×10^{-04}**
EvoFold	-0.03	0.01	-3.41	0.97 [0.95-0.99]	**6.40×10^{-04}**
Strong Enhancer (distal)	0.01	0.00	3.15	1.01 [1.00-1.01]	**1.65×10^{-03}**
vega PseudoGenes	0.01	0.00	3.34	1.01 [1.00-1.02]	**8.51×10^{-04}**
Exons	-0.01	0.00	-3.94	0.99 [0.98-0.99]	**8.08×10^{-05}**
5'UTR	0.02	0.00	3.79	1.02 [1.01-1.03]	**1.50×10^{-04}**
Lost Stops	-0.09	0.03	-2.92	0.91 [0.86-0.97]	**3.53×10^{-03}**
Txn Transition	0.01	0.00	2.69	1.01 [1.00-1.02]	**7.18×10^{-03}**
Insulator	-0.01	0.00	-2.58	0.99 [0.99-1.00]	**9.90×10^{-03}**
Intergenic SNPs	0.00	0.00	1.79	1.00 [1.00-1.01]	7.28×10^{-02}
vega Genes	0.00	0.00	1.79	1.00 [1.00-1.01]	7.29×10^{-02}

6.4 Discussion

6.4.1 STAGE eSNPs

The results of the four analysed datasets for the STAGE eSNPs differed substantially from each other. The means of the odds ratios for the different sets were 2.64 for All eSNPs, 2.01 for GWAS hits, 4.49 for the Shared eSNPs, and 2.15 for Tissue-specific SNPs. The odds ratios for All eSNPs are therefore much closer to the odds ratios for Tissue-specific eSNPs than Shared eSNPs. This could be expected, since there are more Tissue-specific eSNPs than Shared eSNPs the combined set would be more influenced by the larger subset. However, the results were even less similar to the GWAS hits than to each other. The Shared eSNPs had a very high mean of odds ratios in comparison with the others, while the GWAS hits had the lowest mean of odds ratios. The correlations of the odds ratios for the annotations between the SNP datasets were 0.90 for Shared *vs.* Tissue-specific eSNPs, 0.94 for Shared eSNPs *vs.* All eSNPs, and 0.99 for Tissue-specific eSNPs *vs.* All eSNPs. These correlations are very high and stand in quite a large contrast with the correlations of the odds ratios of the GWAS hits and the eSNPs sets: The correlation of GWAS *vs.* All eSNPs was 0.52, GWAS *vs.* Shared eSNPs was 0.39, and GWAS *vs.* Tissue-specific eSNPs was 0.57. This suggests that the eSNPs were more similar to each other than to GWAS hits, but GWAS hits were more like the Tissue-specific eSNPs than either Shared eSNPs or All eSNPs.

The poor correlations between the GWAS hits and the eSNPs may be due to the number of analysed SNPs. The total number of GWAS hits was 969, while the different eSNPs sets were 6,700, 19,846 and 26,546 SNPs. This may not be a fair comparison due to the large difference in numbers, which may have influenced the results, especially in the sparser genomic annotations like the splice sites. It could be speculated that the Shared eSNPs are a set of "truer" eQTLs. There has been more experimental evidence for them, as Shared eSNPs have been shown to influence gene expression in at least two different tissues. The association to gene expression has therefore been replicated in at least one other tissue. The set of Tissue-specific eSNPs is quite likely to consist of true positives and spurious associations, similar to the set of trait-associated SNPs with suggestive levels of trait-association. The trend of enrichments for the Shared eSNPs is followed by the Tissue-specific eSNPs. The copying of the trend of the odds ratios for the Tissue-specific eSNPs without reaching the same observed levels as the odds ratios in the Shared eSNP set, was very comparable to the trend of the results of the Suggestive SNPs (2011) following the results of the Significant SNPs (2011). However, here the majority of genomic annotations reached significance, which is likely due to the number of analysed SNPs, where the Tissue-specific eSNPs had almost three times more SNPs than the Suggestive SNPs (2011).

All eSNP datasets had more significantly enriched annotations than the GWAS hits dataset. The analysed eSNPs are only suspected to be eQTLs influencing gene expression levels, as no validating experiments were performed. If they were true eQTLs, they would be enriched in those regions where experimental evidence already exists for eQTLs, *i.e.,* regions shown to influence levels of gene expression. The genomic annotations that showed significantly different odds ratios of enrichment between GWAS and All eSNPs, where eSNPs had a higher odds ratio were: 1 Kb and 5 Kb regions upstream of transcription start sites (TSS), intronic SNPs, previously identified eQTLs, positively selected genes, transcriptional elongation and weak transcription. The regions upstream of the transcription start sites (TSS) are regions associated with putative promoters [50], and there has been a substantial amount of evidence for eQTLs clustering in those regions [72, 108, 147, 159, 160].

Further literature findings supporting the existence of real eQTLs in the eSNPs showed that exons were enriched over introns and that the preferred location of eQTLs was in or near target genes and transcribed regions [72]. We observed similar odds ratios in exons and exonic SNPs obtaining higher than intronic SNPs and significant differences in the regions associated with weak transcription and transcriptional elongation (see Table 6-3). Additionally, it has previously been observed that non-synonymous SNPs are preferentially not eQTLs, so the relative depletion in comparison with synonymous SNPs is an encouraging result [161, 162]. There was a strong and significant correlation in gene expression levels of positively selected regions in the Yoruba population and the number of eQTLs coinciding with these regions [163], which could explain the significant odds ratio of enrichment in the positively selected genes (All eSNPs OR = 2.74 [2.44-3.01], Shared eSNPs = 4.45 [3.65-5.34], Tissue-specific eSNPs = 2.35 [2.12-2.56]). Enrichment of eSNPs in these areas is therefore an indication that the eSNPs datasets contain real eQTLs, while the other enriched regions could have regulatory functions that were previously not shown.

Most of these regions either coincided with or were regulatory regions; so higher enrichment signals for eQTLs in comparison with GWAS were highly encouraging. The only genomic annotation, which had a higher odds ratio of enrichment for GWAS hits than for eSNPs, was OMIM morbid regions. As pointed out in previous chapters, these regions were the genomic locations of the underlying biology for a large variety of traits and were used as a proxy for a positive control for GWAS hits. The enrichment result of GWAS hits in the OMIM morbid regions and high enrichment signal of All eSNPs in the previously identified eQTL dataset were very intuitive and encouraging. We have therefore shown that there is not only significant enrichment of eSNPs in areas prone to harbour eQTLs, but also show that the distribution of the eSNPs is different to GWAS hits.

The enrichment of eSNPs in a number of genomic annotations previously associated with eQTLs suggests that the analysed datasets did contain real eQTLs. The investigation of gene expression levels was performed on a very small number of patients (n = 109), which could have introduced a large number of false (positive and negative) signals due to multiple testing and stochastic associations. The false discovery rate applied to identify the real associations may not have been stringent enough to distinguish between the real and false positive signal. However, the analysis of gene expression in different tissues did lend extra confidence to the accuracy of the results as more evidence of the association to gene expression levels was found. It may have benefitted the results and conclusions of the study to restrict the analysis to only those eSNPs, for which there was evidence of associations to gene expression in at least two tissues (*i.e.,* Shared eSNPs), as it would have reduced the number of false positives. Additionally, increasing the number of analysed patients or the inclusion of a control group of healthy people may also aid the discovery of real eQTLs.

6.4.2 GIANT height consortium

The SNPs from the GIANT consortium were the first analysed dataset for which the inclusion of the genotyping arrays did not contribute any additional information. The model, which contained only the genotyping arrays, had an r^2 value of 0.00 (0.0004), so the information added by the genotyping arrays to the final model was minimal. This was a surprising result, as the meta-analysis combined the results of 61 individual cohort or case-control studies, which all used different genotyping arrays [57]. It would have been expected that the genotyping arrays added more than 0.0004 to the model. However, it is possible that the effect was reduced to insignificant amounts, as all of the 61 studies performed imputations to ease comparison across the different studies. The results of the meta-analysis were all based on the imputed genotypes of the SNPs. The r^2 value of the final model was extracted from the summary with a value of 0.03 (0.027), meaning that most of the variation remained unexplained. The modest r^2 value of the final model suggests that the majority of the height-associated variation is either still not captured or is explained by a different genomic annotation not yet analysed. Half of the final genomic annotations with significant *P*-values had a negative estimated effect in the model. The dependent variable was the negative logarithm of the *P*-value of association to height of individual SNPs. Annotations with a negative estimate would indicate the relative negative influence on the logarithm of the *P*-value, *i.e.,* a reduction in significance. These annotations expectedly included regions associated with closed or repressed chromatin states, which are associated with little to no transcriptional activity.

6.4.3 Conclusion

In conclusion, we showed that eSNPs had a genomic signature, which was indicative for eQTL enrichment and was distinctly different from GWAS hits. The analysis also showed, that the method and the established dataset was adaptable to analyse results from different study types. The application of the linear regression to the SNPs associated to height highlighted another way of utilizing the built dataset for different studies. While the linear regression did not result in a large reduction of the analysed genomic annotations, it did highlight that a number of genomic annotations were influencing height-association. The result could either be due to the analysis of a continuous variable or could underline the genetic complexity of height. It needs to be kept in mind that the change of one unit in the dependent variable is associated with the change of one unit in the independent variable, which may only be a small increase in the continuous *P*-value of height-association, but significantly add towards explaining the observed variability in the dataset.

7 DISCUSSION

We wanted to investigate the distribution of trait associated SNPs in the genome with respect to a range of genomic annotations, which covered a range of regulatory features that were identified by experiments or computer algorithms. We expanded the work performed by a previous study [50], to address some flaws and investigate more annotations. We have presented a statistically rigorous analysis of enrichment or depletion of trait associated variants within 58 genomic annotations aimed at elucidating the question of what GWAS hits coincide with using novel techniques and data available from the NHGRI GWAS catalogue (http://www.genome.gov/gwastudies/). The methods we used for this investigation were sampling, permutations and regression, which are summarised below. We have further applied the methods presented to other data, in particular the results of a GWAS analysing gene expression levels and a meta-analysis investigating height. Here, we discuss related studies investigating the genomic context of trait-associated variants.

7.1 Summary of sampling method

Hindorff *et al.* performed a similar study [50] where a sampling method was used to analyse the distribution of the then-known GWAS hits in 20 genomic annotations. The analysed sampling method set out to create a null distribution against which the observed data could be compared by randomly drawing many (in our case 100) sets of SNPs. These sets of SNPs were drawn from genotyping arrays and imputed data used in the original set of GWA studies, matching the genotyping array composition of the observed trait-associated SNPs data (GWAS hits). It therefore compared the observed hit distribution to a background distribution to assess the significance of the results. This method had four problems associated with it. The first was that it was computationally intensive and took one day for the analysis of 100 samples. Second, information on genotyping platform used were not always available for all the GWA studies recorded in the NHGRI GWAS catalogue [86] on important details of the GWA studies curated, such as the genotyping array used in a given study. Third, the optimal number of samples needed to appropriately describe the background (null) distribution of SNPs is unclear and determined by computational constraints. Fourth, this method ignores the observed clustering of trait-associated SNPs in the genome, as such SNPs will often co-occur with regulatory and genic regions [164, 165]. It is therefore unlikely that the null distribution produced by the samples drawn would reflect this observed clustering, and this may be a source of artificial enrichments or depletions of trait-associated SNPs in some of the studied functional annotations. Despite being aware of these problems, the first analysis performed in this thesis was the replication of the study of Hindorff *et al.* [50]. This was done as a benchmark to compare the methodology we developed and summarise below. We extended the set of annotations used by Hindorff *et al.* to include more regulatory annotations, as well as several measures of conserved elements and regions with different chromatin states.

7.2 Summary of permutation method

In order to preserve the observed clustering of SNPs and functional annotations in the genome, we developed a method that explicitly preserves that structure and was based on chromosome-bound circular permutations. The method produces a null distribution consisting of the 20,000 circularly permuted genomes, which contained the same number of analysed SNPs per chromosome and respected the clustering of those SNPs and all functional elements in the genome. The 20,000 permutations were run in parallel on a locally maintained 256 CPU computer cluster and finished in three days for all chromosomes. In comparison, the sampling method took about a day for 100 samples. It will therefore likely take more than three days to analyse 20,000 sets of samples. Importantly, the permutation method is readily scalable to very large numbers of trait-associated SNPs and functional annotations. The results of the permutations were very comparable with the results from the sampling method with an r^2 of the regression of their odds ratios of 0.98. The confidence intervals of the odds ratios derived by permutation are generally slightly more conservative (*i.e.,* larger) than those from the sampling approach. This is because the empirically calculated confidence intervals use information on the underlying distribution of the permuted number of overlaps. We also showed that SNPs, which had suggestive levels of trait-association, had less extreme odds ratios of enrichment or depletion than significantly trait-associated SNPs. This result confirms the hypothesis that the suggestive SNPs consist of true positive signals, which would have the same distribution as the significantly associated variants, and false positive signals, which would be expected to lie outside trait-associated regions. In the analysis above one of the parameters was the LD threshold, which determines the number of SNPs that are analysed as LD partners, which we initially set at 0.9. A second LD threshold was analysed, which showed that including more LD partners did not establish a clearer picture of enrichment versus depletion of trait-associated variants. Furthermore, we analysed trait-associated variants from more recent GWAS and showed that they obtained less extreme odds ratios than the older data; see Significant SNPs (2011) compared with Significant SNPs (Difference) in Chapter 4. This means that the newer variants have a different distribution than the older variants.

7.3 Summary of regression models

The sampling and permutation methods analysed individual genomic annotations and are therefore called univariate analyses. We wanted to test all genomic annotations at the same time and remove all redundant information. For this we used a multivariate logistic regression model. However, in order to allow comparisons between the between the permutations and sampling methods and the multivariate model, we analysed individual genomic annotations using univariate logistic regression. This allowed comparisons between different univariate analyses and different regression methods. The weight of the annotation was an outcome from the analysis and can be interpreted as the natural logarithm of an odds ratio. The weight was based on the presence or absence of overlaps with trait-associated variants, but instead of analysing 100 samples or the results of 20,000 permutations all SNPs present in the background data were analysed. This method is exceptionally fast, finishing in less than two hours when run on the computer cluster mentioned above (Chapter 5). The null distribution of this method was the number of real non-associated SNPs, which were compared to the trait-associated SNPs.

While the univariate methods, examining annotations individually, may be very useful for investigating specific genomic annotations of interest, the overall conclusions of these analyses may be misleading (see Chapter 5). It is unclear from such results, which of the overlapping and often-interdependent genomic annotations are driving any observed enrichment. While it is difficult to identify the drivers of the enrichment with any of the analysed methods, the multivariate approach does at least show the effect of the correlations and highlights the most informative annotations. The multivariate analysis calculated an information criterion, the Akaike's Information Criterion (see Chapter 5), based upon which a decision was taken to include or exclude genomic annotations from the model. This was a comparatively slow method even when run on the computer cluster taking up to ten days to complete. However, the final model was determined by the amount of information each genomic annotation carried (Chapter 5) and the method had the added advantage that the effect of the genotyping array used was explicitly included in the model, unlike other methods. If a SNP was represented on several arrays, it had a higher chance of being trait-associated due to it being tested more often. This prior probability was taken into account by including the genotyping arrays into the model. A possible avenue that could be taken to extend this method is by including additional variables extra annotations or interactions between variables. However, this would further increase the running time of the method. The multivariate models highlighted those genomic annotations with relative depletion of trait-associated SNPs, which obtained odds ratios of enrichment in the univariate analyses. The relative depletion results from the comparison with other genomic annotations, which contain more SNPs than the depleted ones. This suggests that the univariate models may have overestimated the importance of the annotations, as other influences contributing to the observed enrichment were not taken into account, which is an important finding for future analyses.

7.4 Recommendation of methods

We have analysed three different methods of investigating the overlaps between trait-associated variants and genomic annotations. The three considered univariate analyses, which were sampling, permutations, and single variable logistic regression, produced remarkably similar and highly comparable results. However, if a recommendation were to be given as to which univariate method should be used in further analyses, it would be suggested to use the permutation method. This is due to the number of expected – or background – overlaps against which the real trait-associated variants were compared. The permutations generated 20,000 virtual genomes with the same number of variants analysed per chromosome producing highly robust results. In comparison, the sampling method relied upon 100 samples and while it focussed on the distribution of the variants across genotyping arrays, it did not control the number of analysed variants per chromosome. Furthermore, the sampling method was the slowest of the methods and took a total of three days for an analysis. While the logistic regression using single variables was the fastest of the three methods, it produced possibly inflated odds ratios. The results were produced by a direct comparison of overlaps and non-overlaps of the SNP sets and the background. However, the background here was the entire genome, which consisted of a very large number of non-associated variants (*i.e.*, a very large number of zeros). This disproportionate separation of the data could have inflated the importance of any annotation found to be overlapping with the variants.

If at all possible, however, it is recommended to use the multivariate analysis to analyse genomic annotations overlapping with associated variants. This is because the variants may overlap with more than one annotation at the time, which overestimates the importance of individual annotations. The multivariate analysis, however, takes that into account and removes redundant information resulting in a model containing the minimum number of genomic annotations explaining the maximum amount of variation. The sampling and permutation methods were not capable of that, so the logistic regression was used.

7.5 Summary of application to other data

The permutation and multivariate regression methods were respectively applied to two different datasets, a gene expression study of seven different tissues in myocardial infarction and a GWAS meta-analysis for height-associated SNPs, to demonstrate the practical application of the methods described in previous chapters (Chapter 6). The permutation method discovered that eSNPs, which were significant GWAS hits in an analysis investigating gene expression levels, are more enriched in promoter and regulatory regions than other GWAS hits. This implied that eSNPs may have a different mechanism to influence traits than GWAS hits, as would have been expected if they are, or contain a high number of, real eQTLs. It is possible that GWAS hits affect a given trait less subtly than eQTLs do, given that eQTLs influence traits by affecting gene expression levels. This effect could be quite small and therefore not immediately noticeable. GWAS hits, however, could act upon traits by disrupting coding or binding regions, which could affect phenotypes quite quickly. We also showed that Shared eSNPs (*i.e.,* eSNPs which were significant in more than two tissues) have more extreme odds ratios than either of the other analysed eSNPs datasets - All eSNPs or Tissue-specific eSNPs (see Chapter 6). A linear regression investigating height-associated SNPs showed that 51 of 54 genomic annotations jointly influenced the *P*-value of height-association. While this is a very high number of annotations, each annotation influenced the height-association only slightly with odds ratios ranging from 0.66 – 1.24.

7.6 Discussion of genomic annotations

The three categories of genomic annotations had different relative enrichment of trait-associated variants. The genic category included regions associated with genes and other regulatory elements, such as eQTLs. The majority of the annotations included in this category were enriched with trait-associated variants. The enrichment signal, however, could possibly be caused by the coinciding annotations. The genic category contained genomic annotations, which ranged from single nucleotides to full genes thereby introducing some heterogeneity into the dataset. However, the impact of the genic category is still larger than the conserved annotation category. The majority of the genic annotations were analysed previously by a number of authors in different studies, where DNase I hypersensitive sites, eQTLs, and distance from the TSS feature the most often [50, 65, 110, 111, 147, 166]. The genic annotations were also the ones that were included most frequently in the final model of multivariate analyses with the exception of the Cancer SNPs set.

The annotations included in the conserved region were slightly more comparable in the lengths of the annotated genomic blocks, but they were only modestly enriched in all analyses. The Immune SNPs and Cancer SNPs were the two SNP sets that were the exception to this, as conserved regions were significantly and consistently depleted in the Immune SNPs while the Cancer SNPs showed enrichment for these sites. A different study that also investigated phastCons sites was the study by Gaffney *et al.*, which also showed only modest enrichment of SNPs in these sites [147]. Hindorff *et al.* [50] also investigated conserved sites, but chose only the conserved sites across 28 species. The low enrichment in our study could be due to the use of all sites identified by phastCons rather than restricting the annotation to only those sites with high LOD scores. This could have reduced the odds ratios of the annotations, as it could be possible that those sites that are overlapping with non-associated variants have low scores and would have been removed from the dataset if a threshold had been applied. However, the obtained modest enrichment found by Gaffney *et al.* [147] corroborates our results. This annotation category, like the chromatin states, contained only data obtained computational analyses, but had very different enrichment and depletion signals than the chromatin states. It is therefore not possible to draw a line between the quality of data obtained by experimental or computational analyses.

The annotations in the third category, the chromatin states, were all identified by experimental data, which was refined through computational analyses. This category contained annotations with the largest annotated genomic blocks and could potentially be more robust to shifts in genomic positions between genomic reference maps. Furthermore, the enrichment signals obtained in the chromatin states category were very encouraging as there has been mounting evidence [82, 167, 168] that trait-associated variants preferentially lie in regions that can be identified through histone modifications by either analysing the histone modifications individually or as a pattern.

As mentioned above, none of the analysed annotations were restricted according to a threshold. This applied to data obtained from either computational analyses or results obtained through experimental analyses, like ChIP-seq. The lack of threshold could have resulted in false positive or false negative signals, depending on the over- or under-prediction of the annotation in the observed or expected datasets. For example, a genomic region that was falsely annotated and overlapped with real trait-associated variants but not with permuted variants could have obtained a higher odds ratio estimate. Alternatively, a genomic region that was falsely annotated but overlapped only with permuted variants could have caused a falsely reduced odds ratio. These are caveats in this analysis, which would have to be addressed if this study were to be repeated.

An additional caveat, which could have caused inflated odds ratios, is the inclusion of non-Caucasian studies into the analysis. While the percentage of studies in non-Caucasian populations is still quite small, it could have erroneously inflated the number of overlaps in annotations. This is due to the shorter LD blocks found in African population when compared to Caucasian populations. For future studies it would be recommended to remove the non-Caucasian studies from further analyses.

7.7 Other studies investigating functional annotations

Many recently published studies are focussing on the functional annotations of underlying trait-associated variants to inform future GWAS and population sequencing studies [65, 165, 169, 170]. The interest in regions annotated with functional elements is largely driven by the attempt to identify those genomic annotations enriched for GWAS hits and to aid the often-laborious search for causal variants. It is hoped that a separation of spurious associations from true associations can be attempted by prioritizing those regions that were previously associated with trait-associated variants. This could potentially increase the number of significantly trait-associated variants or highlight variants with a higher chance of giving a positive result. A study published in 2011 [65] suggested the empirical Bayes Factors for three annotations (*cis*-eQTLs, non-synonymous SNPs and promoter SNPs) for prioritisation algorithms to identify candidates for GWAS follow-up studies to be 4, 3 and 2, respectively. This implies that eQTLs are most usually enriched for trait-associated SNPs, which confirms the findings of a study in 2010, which found that trait-associated variants were most likely to be eQTLs [110]. A study in 2013 analysed a genomic inflation correction by estimating a genomic control from intergenic SNPs [165] while also identifying strong enrichment in 5'UTRs. Their observations of strong enrichment in regulatory genic elements agree well with our results, but they only analysed 10 annotations and did not incorporate functional regions annotated by chromatin states [165]. The Encyclopedia of DNA Elements (ENCODE) is a large international consortium, which aims to categorize DNA elements. The publication of an ENCODE study investigating the annotations of disease-associated SNPs showed that ENCODE data can be used for these an notational studies [74]. This finding was supported by the publication of a database detailing regulatory and rare SNPs [170]. An investigation also using multivariate logistic regression to identify the regulatory architecture of eQTLs by investigating transcription factor binding sites and conserved sites found a distinct enrichment of eQTLs in transcription factor binding sites, but observed only showed modest enrichment in conserved sites [147]. In this thesis, the conserved regions also showed modest odds ratios enrichment in the univariate and the final multivariate models, where included. This is despite the frequent use of conservation measures in variant prioritization methods [64], which implicitly suggests that they are important for trait-associated variants. But here it appears that other annotations are more influential. It appears that there are many studies investigating how to mine the existing GWAS data to leverage more information from them to aid in the dissecting of the genetics contributing to complex traits. While all studies to date have analysed different genomic annotations across a large spectrum of functional elements, a study with as many genomic annotations as analysed in this thesis had not been attempted. All of these studies suggested it was worthwhile to undertake a more comprehensive analysis of functional elements and to examine the distribution of GWAS hits within them.

7.8 Future work and developments and their impacts

There has been a substantial increase in the number of studies reporting on the annotations of trait-associated variants [167-170] since the original study by Hindorff *et al.* in 2009. Furthermore, there have been a large number of online tools investigating ways of easing the often-laborious process of annotating trait-associated variants. Among these are HaploReg [171] and RegulomeDB [170], which aim to aid researchers investigating the underlying genomic regions of trait-associated variants. This could potentially lead to a faster and cheaper way to discover the underlying biological causes of trait-associated variants. The future will likely see more online tools used to annotate variants, as there is an ever-increasing pressure on the society to not just identify associated variants. The tools may aid in the discovery of the real causes of the associations as they highlight which annotations are most often coinciding with trait-associated variants. The future will tell if this is right.

During the course of the research presented in this thesis, a number of new annotations were published. The ENCODE data was released over the past few years, locating a large number of transcription factor binding sites and other sites of functional annotations [172-174]. Before 2011 we had included a large number of histone modifications, which were identified in several cell lines. However, in 2011 a meta-analysis of histone modification patterns in nine cell lines showed that particular combinations (chromatin states) were associated with particular classes of functional regions [82]. Most of the final multivariate regression models contained at least one of these functional classes, where they were influential in determining trait-association status. One can only expect that the next few years will see more leaps in the quality and quantity of functional data resulting in better models and possible identification of highly explanatory annotations.

GWAS sample sizes are increasing to allow the identification of variants with minor allele frequencies (MAFs) of less than 1% [175] to be detected [31] and common variants with smaller effects. Their design will also have to include a wider range of different populations to analyse diseases common to specific populations and not focus on mainly Caucasian populations [176, 177]. Whole genome sequencing technologies have begun to uncover many novel variations [58, 178] including SNPs with low allele frequencies and 'private' mutations, seen in only one individual. This new variation will facilitate the discovery of genetic causes, as a large proportion of highly deleterious SNPs [179] are rare or private mutations [180]. These private mutations could aid the understanding of complex traits greatly, because they could have very large effects. The large effect would be strongly selected against, therefore keeping their allele frequencies low. The rarity of these mutations may not necessarily be matched by their abundance in certain pathways, *i.e.*, it could be that a large number of rare or private mutations disrupt the same pathway giving the same phenotype. The new information, which will be discovered, might also include structural variants, such as copy number variants (CNVs), or repetitive regions that have so far been difficult to assay and analyse. While the sequencing technologies will get better, the range of available functional genomic data will also be improved upon. These annotations may include retrotransposons and more RNA molecules, which are emerging as functional [173, 181, 182]. The emergence of more SNPs and new annotations will lead to a higher demand for predictive modelling of variant function, which will also feed back into the improvement of future models. The research presented in this thesis is highly adaptable and can be easily applied to data for new disease associated variants and new functional annotation. We analysed a broader range of functional annotations simultaneously than other studies to date furthermore investigated different trait-subsets. These allowed the drawing of more trait-specific conclusions, which in turn may feed into more specific disease risk predictive models and say something about the architecture of the traits. The genomic annotations from the multivariate models used here (Chapter 5) could therefore aid the calculation of the prior probability – or 'weighing' – of SNPs to identify those variants, which were more likely to affect phenotypes.

Better predictive models and the ever-decreasing cost of genotyping may eventually lead to more precise disease risk predictions for individuals. Companies such as 23andme (http://www.23andme.com) or deCODEme (http://www.decodeme.com) are already attempting to predict risk of certain diseases by investigating the genotypes of a number of SNPs. However, the results produced by the companies still vary greatly between companies as they use different predictive algorithms, different SNP sets, and different average population risks [183]. Truly accurate risk predictions may therefore be out of reach until we can thoroughly comprehend environmental influences, which affect genetic predisposition, and include them in our analyses. It might be possible that these predictions become more accurate, as more information, such as environmental influences, gets included in the analysis. Such environmental influences could, for example, be modelled using different histone modifications known to respond to outside stimuli. However, these are out of reach at present. In this thesis, we have shown that while different combinations of genomic annotations influence different trait-subsets, common regulatory features are present and most often underlie variants associated to a broad range of traits.

8 BIBLIOGRAPHY

1. Mendel JG: **Versuche über Pflanzenhybriden**. *Verhandlungen des naturforschenden Vereines in Brünn* 1866, **IV**:3-47.

2. Monaghan F, Corcos A: **On the origins of the Mendelian laws**. *J Hered* 1984, **75**(1):67-69.

3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E,

Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.

4. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A,

Zandieh A, Zhu X: **The sequence of the human genome**. *Science* 2001, **291**(5507):1304-1351.

5. Consortium IHGS: **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**(7011):931-945.

6. Venter JC: **A part of the human genome sequence**. *Science* 2003, **299**(5610):1183-1184.

7. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC: **The diploid genome sequence of an individual human**. *PLoS Biol* 2007, **5**(10):e254.

8. **E pluribus unum**. *Nat Methods* 2010, **7**(5):331.

9. **Human Genome Reference Consortium** [http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/]

10. Dolgin E: **Human genomics: The genome finishers**. *Nature* 2009, **462**(7275):843-845.

11. Parekh-Olmedo H, Kmiec EB: **Targeted nucleotide exchange in the CAG repeat region of the human HD gene**. *Biochem Biophys Res Commun* 2003, **310**(2):660-666.

12. Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey**. *Nucleic Acids Res* 2002, **30**(17):3894-3900.

13. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Wong WS, Sigurdsson G, Walters GB, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Magnusson OT, Thorsteinsdottir U, Stefansson K: **Rate of de novo mutations and the importance of father's age to disease risk.** *Nature* 2012, **488**(7412):471-475.

14. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, Walker MF, Ober GT, Teran NA, Song Y, El-Fishawy P, Murtha RC, Choi M, Overton JD, Bjornson RD, Carriero NJ, Meyer KA, Bilguvar K, Mane SM, Sestan N, Lifton RP, Gunel M, Roeder K, Geschwind DH, Devlin B, State MW: **De novo mutations revealed by whole-exome sequencing are strongly associated with autism**. *Nature* 2012, **485**(7397):237-241.

15. Scally A, Durbin R: **Revising the human mutation rate: implications for understanding human evolution**. *Nat Rev Genet* 2012, **13**(10):745-753.

16. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic Acids Res* 2001, **29**(1):308-311.

17. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**(7319):1061-1073.

18. Preuss C, Riemenschneider M, Wiedmann D, Stoll M: **Evolutionary dynamics of co-segregating gene clusters associated with complex diseases**. *PLoS One* 2012, **7**(5):e36205.

19. Kieleczawa J: **Fundamentals of sequencing of difficult templates--an overview**. *J Biomol Tech* 2006, **17**(3):207-217.

20. Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation sequencing: computational challenges and solutions**. *Nat Rev Genet* 2012, **13**(1):36-46.

21. Avery OT, Macleod CM, McCarty M: **Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii**. *J Exp Med* 1944, **79**(2):137-158.

22. Watson JD, Crick FH: **The structure of DNA**. *Cold Spring Harb Symp Quant Biol* 1953, **18**:123-131.

23. Morgan TH: **Random Segregation Versus Coupling in Mendelian Inheritance**. *Science* 1911, **34**(873):384.

24. Sturtevant AH: **Linkage Variation and Chromosome Maps**. *Proc Natl Acad Sci U S A* 1921, **7**(7):181-183.

25. Botstein D, White RL, Skolnick M, Davis RW: **Construction of a genetic linkage map in man using restriction fragment length polymorphisms**. *Am J Hum Genet* 1980, **32**(3):314-331.

26. Cichon S, Craddock N, Daly M, Faraone SV, Gejman PV, Kelsoe J, Lehner T, Levinson DF, Moran A, Sklar P, Sullivan PF: **Genomewide association studies: history, rationale, and prospects for psychiatric disorders**. *Am J Psychiatry* 2009, **166**(5):540-556.

27. MacDonald ME, Ambrose CM, Duyao MP, Myers HR, Lin C, Srinidi L, Barnes G, Taylor SA, James M, Groat N, MacFarlane H, Jenkins B, Anderson MA, Wexler NS, Gusella JF, Bates GP, Baxendale S, Hummerich

H, Kirby S, North M, Youngman S, Mott R, Zehetner G, Sedlacek Z, Poustka A, Frischauf AM, Lehrach H, Buckler AJ, Church D, Doucette-Stamm L, O'Donovan MC, Riba-Ramirer L, Shah M, Stanton VP, Strobel SA, Draths KM, Wales JL, Dervan P, Housman DE, Altherr M, Shiang R, Thompson L, Fielder T, Wasmuth JJ, Tagle D, Valdes J, Elmer L, Allard M, Castilla L, Swaroop M, Blanchard K, Collins FS, Snell R, Holloway T, Gillespie K, Datson N, Shaw D, Harper PS: **A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group**. *Cell* 1993, **72**(6):971-983.

28. Walker FO: **Huntington's disease**. *Lancet* 2007, **369**(9557):218-228.

29. Frazer KA, Murray SS, Schork NJ, Topol EJ: **Human genetic variation and its contribution to complex traits**. *Nat Rev Genet* 2009, **10**(4):241-251.

30. Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, Martin NG, Montgomery GW, Weedon MN, Loos RJ, Frayling TM, McCarthy MI, Hirschhorn JN, Goddard ME, Visscher PM: **Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits**. *Nat Genet* 2012, **44**(4):369-375, S361-363.

31. Visscher PM, Brown MA, McCarthy MI, Yang J: **Five years of GWAS discovery**. *Am J Hum Genet* 2012, **90**(1):7-24.

32. Cooper GM, Shendure J: **Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data**. *Nat Rev Genet* 2011, **12**(9):628-640.

33. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J: **Complement factor H polymorphism in age-related macular degeneration**. *Science* 2005, **308**(5720):385-389.

34. Dewan A, Liu M, Hartman S, Zhang SS, Liu DT, Zhao C, Tam PO, Chan WM, Lam DS, Snyder M, Barnstable C, Pang CP, Hoh J: **HTRA1 promoter polymorphism in wet age-related macular degeneration**. *Science* 2006, **314**(5801):989-992.

35. Consortium WTCC: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls**. *Nature* 2007, **447**(7145):661-678.

36. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL,

Mackay TF, McCarroll SA, Visscher PM: **Finding the missing heritability of complex diseases**. *Nature* 2009, **461**(7265):747-753.

37. Phillips ML, Kupfer DJ: **Bipolar disorder diagnosis: challenges and future directions**. *Lancet* 2013, **381**(9878):1663-1671.

38. **A Catalog of Published Genome-Wide Association Studies** [http://www.genome.gov/gwastudies/]

39. WTCCC: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls**. *Nature* 2007, **447**(7145):661-678.

40. Goldstein DB: **Common genetic variation and human traits**. *N Engl J Med* 2009, **360**(17):1696-1698.

41. Hirschhorn JN: **Genomewide association studies--illuminating biologic pathways**. *N Engl J Med* 2009, **360**(17):1699-1701.

42. Lee JH: **Importance of complex traits.** In: *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics.* 2006.

43. Falconer D, Mackay T: **Introduction to Quantitative Genetics**, 4 edn: Pearson, Prentice Hall; 1960.

44. Slatkin M: **Epigenetic inheritance and the missing heritability problem**. *Genetics* 2009, **182**(3):845-850.

45. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM: **Common SNPs explain a large proportion of the heritability for human height**. *Nat Genet* 2010, **42**(7):565-569.

46. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, de Andrade M, Feenstra B, Feingold E, Hayes MG, Hill WG, Landi MT, Alonso A, Lettre G, Lin P, Ling H, Lowe W, Mathias RA, Melbye M, Pugh E, Cornelis MC, Weir BS, Goddard ME, Visscher PM: **Genome partitioning of genetic variation for complex traits using common SNPs**. *Nat Genet* 2011, **43**(6):519-525.

47. Axenovich TI, Zorkoltseva IV, Belonogova NM, Struchalin MV, Kirichenko AV, Kayser M, Oostra BA, van Duijn CM, Aulchenko YS: **Linkage analysis of adult height in a large pedigree from a Dutch genetically isolated population**. *Hum Genet* 2009, **126**(3):457-471.

48. McEvoy BP, Visscher PM: **Genetics of human height**. *Econ Hum Biol* 2009, **7**(3):294-306.

49. Peiffer DA, Gunderson KL: **Design of tag SNP whole genome genotyping arrays**. *Methods Mol Biol* 2009, **529**:51-61.

50. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits**. *Proc Natl Acad Sci U S A* 2009, **106**(23):9362-9367.

51. Spencer CC, Su Z, Donnelly P, Marchini J: **Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip**. *PLoS Genet* 2009, **5**(5):e1000477.

52. Hao K, Chudin E, McElwee J, Schadt EE: **Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies**. *BMC Genet* 2009, **10**:27.

53. Howie BN, Donnelly P, Marchini J: **A flexible and accurate genotype imputation method for the next generation of genome-wide association studies**. *PLoS Genet* 2009, **5**(6):e1000529.

54. Li Y, Abecasis G: **Mach 1.0: Rapid Haplotype Reconstruction and Missing Genotype Inference**. *Annu Rev Genomics Hum Genet* 2006, **10**.

55. Browning BL, Browning SR: **A fast, powerful method for detecting identity by descent**. *Am J Hum Genet* 2011, **88**(2):173-182.

56. Vitart V, Biloglav Z, Hayward C, Janicijevic B, Smolej-Narancic N, Barac L, Pericic M, Klaric IM, Skaric-Juric T, Barbalic M, Polasek O, Kolcic I, Carothers A, Rudan P, Hastie N, Wright A, Campbell H, Rudan I: **3000 years of solitude: extreme differentiation in the island isolates of Dalmatia, Croatia**. *Eur J Hum Genet* 2006, **14**(4):478-487.

57. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segre AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, Qi L, Vernon Smith A, Magi R, Pastinen T, Liang L, Heid IM, Luan J, Thorleifsson G, Winkler TW, Goddard ME, Sin Lo K, Palmer C, Workalemahu T, Aulchenko YS, Johansson A, Zillikens MC, Feitosa MF, Esko T, Johnson T, Ketkar S, Kraft P, Mangino M, Prokopenko I, Absher D, Albrecht E, Ernst F, Glazer NL, Hayward C, Hottenga JJ, Jacobs KB, Knowles JW, Kutalik Z, Monda KL, Polasek O, Preuss M, Rayner NW, Robertson NR, Steinthorsdottir V, Tyrer JP, Voight BF, Wiklund F, Xu J, Zhao JH, Nyholt DR, Pellikka N, Perola M, Perry JR, Surakka I, Tammesoo ML, Altmaier EL, Amin N, Aspelund T, Bhangale T, Boucher G, Chasman DI, Chen C, Coin L, Cooper MN, Dixon AL, Gibson Q, Grundberg E, Hao K, Juhani Junttila M, Kaplan LM, Kettunen J, Konig IR, Kwan T, Lawrence RW, Levinson DF, Lorentzon M, McKnight B, Morris AP, Muller M, Suh Ngwa J, Purcell S, Rafelt S, Salem RM, Salvi E,

Sanna S, Shi J, Sovio U, Thompson JR, Turchin MC, Vandenput L, Verlaan DJ, Vitart V, White CC, Ziegler A, Almgren P, Balmforth AJ, Campbell H, Citterio L, De Grandi A, Dominiczak A, Duan J, Elliott P, Elosua R, Eriksson JG, Freimer NB, Geus EJ, Glorioso N, Haiqing S, Hartikainen AL, Havulinna AS, Hicks AA, Hui J, Igl W, Illig T, Jula A, Kajantie E, Kilpelainen TO, Koiranen M, Kolcic I, Koskinen S, Kovacs P, Laitinen J, Liu J, Lokki ML, Marusic A, Maschio A, Meitinger T, Mulas A, Pare G, Parker AN, Peden JF, Petersmann A, Pichler I, Pietilainen KH, Pouta A, Ridderstrale M, Rotter JI, Sambrook JG, Sanders AR, Schmidt CO, Sinisalo J, Smit JH, Stringham HM, Bragi Walters G, Widen E, Wild SH, Willemsen G, Zagato L, Zgaga L, Zitting P, Alavere H, Farrall M, McArdle WL, Nelis M, Peters MJ, Ripatti S, van Meurs JB, Aben KK, Ardlie KG, Beckmann JS, Beilby JP, Bergman RN, Bergmann S, Collins FS, Cusi D, den Heijer M, Eiriksdottir G, Gejman PV, Hall AS, Hamsten A, Huikuri HV, Iribarren C, Kahonen M, Kaprio J, Kathiresan S, Kiemeney L, Kocher T, Launer LJ, Lehtimaki T, Melander O, Mosley TH, Jr., Musk AW, Nieminen MS, O'Donnell CJ, Ohlsson C, Oostra B, Palmer LJ, Raitakari O, Ridker PM, Rioux JD, Rissanen A, Rivolta C, Schunkert H, Shuldiner AR, Siscovick DS, Stumvoll M, Tonjes A, Tuomilehto J, van Ommen GJ, Viikari J, Heath AC, Martin NG, Montgomery GW, Province MA, Kayser M, Arnold AM, Atwood LD, Boerwinkle E, Chanock SJ, Deloukas P, Gieger C, Gronberg H, Hall P, Hattersley AT, Hengstenberg C, Hoffman W, Lathrop GM, Salomaa V, Schreiber S, Uda M, Waterworth D, Wright AF, Assimes TL, Barroso I, Hofman A, Mohlke KL, Boomsma DI, Caulfield MJ, Cupples LA, Erdmann J, Fox CS, Gudnason V, Gyllensten U, Harris TB, Hayes RB, Jarvelin MR, Mooser V, Munroe PB, Ouwehand WH, Penninx BW, Pramstaller PP, Quertermous T, Rudan I, Samani NJ, Spector TD, Volzke H, Watkins H, Wilson JF, Groop LC, Haritunians T, Hu FB, Kaplan RC, Metspalu A, North KE, Schlessinger D, Wareham NJ, Hunter DJ, O'Connell JR, Strachan DP, Wichmann HE, Borecki IB, van Duijn CM, Schadt EE, Thorsteinsdottir U, Peltonen L, Uitterlinden AG, Visscher PM, Chatterjee N, Loos RJ, Boehnke M, McCarthy MI, Ingelsson E, Lindgren CM, Abecasis GR, Stefansson K, Frayling TM, Hirschhorn JN: **Hundreds of variants clustered in genomic loci and biological pathways affect human height**. *Nature* 2010, **467**(7317):832-838.

58. Cirulli ET, Goldstein DB: **Uncovering the roles of rare variants in common disease through whole-genome sequencing**. *Nat Rev Genet* 2010, **11**(6):415-425.

59. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, Moutsianas L, Dilthey A, Su Z, Freeman C, Hunt SE, Edkins S, Gray E, Booth DR, Potter SC, Goris A, Band G, Oturai AB, Strange A, Saarela J, Bellenguez C, Fontaine B, Gillman M, Hemmer B, Gwilliam R, Zipp F, Jayakumar A, Martin R, Leslie S, Hawkins S, Giannoulatou E, D'Alfonso S, Blackburn H, Martinelli Boneschi F, Liddle J, Harbo HF, Perez ML, Spurkland A, Waller MJ, Mycko MP, Ricketts M, Comabella M, Hammond N, Kockum I, McCann OT, Ban M, Whittaker P, Kemppinen A, Weston P,

Hawkins C, Widaa S, Zajicek J, Dronov S, Robertson N, Bumpstead SJ, Barcellos LF, Ravindrarajah R, Abraham R, Alfredsson L, Ardlie K, Aubin C, Baker A, Baker K, Baranzini SE, Bergamaschi L, Bergamaschi R, Bernstein A, Berthele A, Boggild M, Bradfield JP, Brassat D, Broadley SA, Buck D, Butzkueven H, Capra R, Carroll WM, Cavalla P, Celius EG, Cepok S, Chiavacci R, Clerget-Darpoux F, Clysters K, Comi G, Cossburn M, Cournu-Rebeix I, Cox MB, Cozen W, Cree BA, Cross AH, Cusi D, Daly MJ, Davis E, de Bakker PI, Debouverie M, D'Hooghe M B, Dixon K, Dobosi R, Dubois B, Ellinghaus D, Elovaara I, Esposito F, Fontenille C, Foote S, Franke A, Galimberti D, Ghezzi A, Glessner J, Gomez R, Gout O, Graham C, Grant SF, Guerini FR, Hakonarson H, Hall P, Hamsten A, Hartung HP, Heard RN, Heath S, Hobart J, Hoshi M, Infante-Duarte C, Ingram G, Ingram W, Islam T, Jagodic M, Kabesch M, Kermode AG, Kilpatrick TJ, Kim C, Klopp N, Koivisto K, Larsson M, Lathrop M, Lechner-Scott JS, Leone MA, Leppa V, Liljedahl U, Bomfim IL, Lincoln RR, Link J, Liu J, Lorentzen AR, Lupoli S, Macciardi F, Mack T, Marriott M, Martinelli V, Mason D, McCauley JL, Mentch F, Mero IL, Mihalova T, Montalban X, Mottershead J, Myhr KM, Naldi P, Ollier W, Page A, Palotie A, Pelletier J, Piccio L, Pickersgill T, Piehl F, Pobywajlo S, Quach HL, Ramsay PP, Reunanen M, Reynolds R, Rioux JD, Rodegher M, Roesner S, Rubio JP, Ruckert IM, Salvetti M, Salvi E, Santaniello A, Schaefer CA, Schreiber S, Schulze C, Scott RJ, Sellebjerg F, Selmaj KW, Sexton D, Shen L, Simms-Acuna B, Skidmore S, Sleiman PM, Smestad C, Sorensen PS, Sondergaard HB, Stankovich J, Strange RC, Sulonen AM, Sundqvist E, Syvanen AC, Taddeo F, Taylor B, Blackwell JM, Tienari P, Bramon E, Tourbah A, Brown MA, Tronczynska E, Casas JP, Tubridy N, Corvin A, Vickery J, Jankowski J, Villoslada P, Markus HS, Wang K, Mathew CG, Wason J, Palmer CN, Wichmann HE, Plomin R, Willoughby E, Rautanen A, Winkelmann J, Wittig M, Trembath RC, Yaouanq J, Viswanathan AC, Zhang H, Wood NW, Zuvich R, Deloukas P, Langford C, Duncanson A, Oksenberg JR, Pericak-Vance MA, Haines JL, Olsson T, Hillert J, Ivinson AJ, De Jager PL, Peltonen L, Stewart GJ, Hafler DA, Hauser SL, McVean G, Donnelly P, Compston A: **Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis**. *Nature* 2011, **476**(7359):214-219.

60. Juran BD, Lazaridis KN: **Genomics in the post-GWAS era**. *Semin Liver Dis* 2011, **31**(2):215-222.

61. Hakonarson H, Grant SF: **GWAS and its impact on elucidating the etiology of diabetes.** *Diabetes Metab Res Rev* 2011, **27**(7):685–696.

62. Sethumadhavan R, Doss CG, Rajasekaran R: **In silico searching for disease-associated functional DNA variants**. *Methods Mol Biol* 2011, **760**:239-250.

63. Cantor RM, Lange K, Sinsheimer JS: **Prioritizing GWAS results: A review of statistical methods and recommendations for their application.** *Am J Hum Genet* 2010, **86**(1):6-22.

64. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function**. *Nucleic Acids Res* 2003, **31**(13):3812-3814.

65. Knight J, Barnes MR, Breen G, Weale ME: **Using functional annotation for the empirical determination of Bayes Factors for genome-wide association study analysis**. *PLoS One* 2011, **6**(4):e14808.

66. Ohno S: **So much "junk" DNA in our genome**. *Brookhaven Symp Biol* 1972, **23**:366-370.

67. Pennisi E: **Genomics. ENCODE project writes eulogy for junk DNA**. *Science* 2012, **337**(6099):1159, 1161.

68. Eddy SR: **The C-value paradox, junk DNA and ENCODE**. *Curr Biol* 2012, **22**(21):R898-899.

69. Niu DK, Jiang L: **Can ENCODE tell us how much junk DNA we carry in our genome?** *Biochem Biophys Res Commun* 2013, **430**(4):1340-1343.

70. Green P, Ewing B: **Comment on "Evidence of abundant purifying selection in humans for recently acquired regulatory functions"**. *Science* 2013, **340**(6133):682; discussion 682.

71. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E: **On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE**. *Genome Biol Evol* 2013, **5**(3):578-590.

72. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK: **High-resolution mapping of expression-QTLs yields insight into human gene regulation**. *PLoS Genet* 2008, **4**(10):e1000214.

73. Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, Mahony S, Sleumer MC, Bilenky M, Haeussler M, Griffith M, Gallo SM, Giardine B, Hooghe B, Van Loo P, Blanco E, Ticoll A, Lithwick S, Portales-Casamar E, Donaldson IJ, Robertson G, Wadelius C, De Bleser P, Vlieghe D, Halfon MS, Wasserman W, Hardison R, Bergman CM, Jones SJ: **ORegAnno: an open-access community-driven resource for regulatory annotation**. *Nucleic Acids Res* 2008, **36**(Database issue):D107-113.

74. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, Rozowsky J, Birney E, Bickel P, Snyder M, Gerstein M: **Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors**. *Genome Biol* 2012, **13**(9):R48.

75. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Jr., Kinney JB, Kellis M, Lander ES, Mikkelsen TS:

Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* 2012, **30**(3):271-277.

76. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome Res* 2005, **15**(8):1034-1050.

77. Feinberg AP, Irizarry RA: **Evolution in Health and Medicine Sackler Colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease**. *Proc Natl Acad Sci U S A* 2010.

78. Ehrlich M: **DNA methylation in cancer: too much, but also too little**. *Oncogene* 2002, **21**(35):5400-5413.

79. von Kanel T, Huber AR: **DNA methylation analysis**. *Swiss Med Wkly* 2013, **143**:w13799.

80. Holliday R: **Epigenetics: a historical overview**. *Epigenetics* 2006, **1**(2):76-80.

81. Verma M: **Cancer control and prevention: nutrition and epigenetics**. *Curr Opin Clin Nutr Metab Care* 2013.

82. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE: **Mapping and analysis of chromatin state dynamics in nine human cell types**. *Nature* 2011, **473**(7345):43-49.

83. Hagg S, Skogsberg J, Lundstrom J, Noori P, Nilsson R, Zhong H, Maleki S, Shang MM, Brinne B, Bradshaw M, Bajic VB, Samnegard A, Silveira A, Kaplan LM, Gigante B, Leander K, de Faire U, Rosfors S, Lockowandt U, Liska J, Konrad P, Takolander R, Franco-Cereceda A, Schadt EE, Ivert T, Hamsten A, Tegner J, Bjorkegren J: **Multi-organ expression profiling uncovers a gene module in coronary artery disease involving transendothelial migration of leukocytes and LIM domain binding 2: the Stockholm Atherosclerosis Gene Expression (STAGE) study**. *PLoS Genet* 2009, **5**(12):e1000754.

84. Li Y, Willer C, Sanna S, Abecasis G: **Genotype imputation**. *Annu Rev Genomics Hum Genet* 2009, **10**:387-406.

85. **A Catalog of Published Genome-Wide Association Studies** [http://www.genome.gov/gwastudies/]

86. **A Catalog of Published Genome-Wide Association Studies** [http://www.genome.gov/gwastudies/]

87. Cui R, Kamatani Y, Takahashi A, Usami M, Hosono N, Kawaguchi T, Tsunoda T, Kamatani N, Kubo M, Nakamura Y, Matsuda K: **Functional variants in ADH1B and ALDH2 coupled with alcohol and smoking synergistically enhance esophageal cancer risk**. *Gastroenterology* 2009, **137**(5):1768-1775.

88. Kamatani Y, Matsuda K, Okada Y, Kubo M, Hosono N, Daigo Y, Nakamura Y, Kamatani N: **Genome-wide association study of hematological and biochemical traits in a Japanese population**. *Nat Genet* 2010, **42**(3):210-215.

89. Takeuchi F, Isono M, Nabika T, Katsuya T, Sugiyama T, Yamaguchi S, Kobayashi S, Ogihara T, Yamori Y, Fujioka A, Kato N: **Confirmation of ALDH2 as a Major locus of drinking behavior and of its variants regulating multiple metabolic phenotypes in a Japanese population**. *Circ J* 2011, **75**(4):911-918.

90. Zhang Y, De S, Garner JR, Smith K, Wang SA, Becker KG: **Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information**. *BMC Med Genomics* 2010, **3**:1.

91. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN: **Genome-wide association studies for complex traits: consensus, uncertainty and challenges**. *Nat Rev Genet* 2008, **9**(5):356-369.

92. HapMap: **A haplotype map of the human genome**. *Nature* 2005, **437**(7063):1299-1320.

93. HapMap: **The International HapMap Project**. *Nature* 2003, **426**(6968):789-796.

94. **UCSC Genome Browser Utilities: Batch Coordinate Conversion (liftOver).** [http://genome.ucsc.edu/cgi-bin/hgLiftOver]

95. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 2005, **33**(Database issue):D501-504.

96. Cocozza S, Akhtar MM, Miele G, Monticelli A: **CpG islands undermethylation in human genomic regions under selective pressure**. *PLoS One* 2011, **6**(8):e23156.

97. Gardiner-Garden M, Frommer M: **CpG islands in vertebrate genomes**. *J Mol Biol* 1987, **196**(2):261-282.

98. Montgomery SB, Griffith OL, Sleumer MC, Bergman CM, Bilenky M, Pleasance ED, Prychyna Y, Zhang X, Jones SJ: **ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation**. *Bioinformatics* 2006, **22**(5):637-640.

99. Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA: **Online Mendelian Inheritance in Man (OMIM)**. *Hum Mutat* 2000, **15**(1):57-61.

100. Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, Hayward DC, Ball EE, Degnan B, Muller P, Spring J, Srinivasan A, Fishman M, Finnerty J, Corbo J, Levine M, Leahy P, Davidson E, Ruvkun G: **Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA**. *Nature* 2000, **408**(6808):86-89.

101. Mourelatos Z, Dostie J, Paushkin S, Sharma A, Charroux B, Abel L, Rappsilber J, Mann M, Dreyfuss G: **miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs**. *Genes Dev* 2002, **16**(6):720-728.

102. Pang KC, Stephen S, Dinger ME, Engstrom PG, Lenhard B, Mattick JS: **RNAdb 2.0--an expanded database of mammalian non-coding RNAs**. *Nucleic Acids Res* 2007, **35**(Database issue):D178-182.

103. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets**. *Cell* 2005, **120**(1):15-20.

104. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB: **Prediction of mammalian microRNA targets**. *Cell* 2003, **115**(7):787-798.

105. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP: **MicroRNA targeting specificity in mammals: determinants beyond seed pairing**. *Mol Cell* 2007, **27**(1):91-105.

106. **eQTL resources** [http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/]

107. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M, Gagnebin M, Nisbett J, Deloukas P, Dermitzakis ET, Antonarakis SE: **Common regulatory variation impacts gene expression in a cell type-dependent manner**. *Science* 2009, **325**(5945):1246-1250.

108. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK: **Understanding**

mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010, **464**(7289):768-772.

109. Gilad Y, Rifkin SA, Pritchard JK: **Revealing the architecture of gene regulation: the promise of eQTL studies**. *Trends Genet* 2008, **24**(8):408-415.

110. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ: **Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS**. *PLoS Genet* 2010, **6**(4):e1000888.

111. Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, Stephens M, Gilad Y, Pritchard JK: **DNase I sensitivity QTLs are a major determinant of human expression variation**. *Nature* 2012, **482**(7385):390-394.

112. Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A, Weaver M, Shafer A, Lee K, Neri F, Humbert R, Singer MA, Richmond TA, Dorschner MO, McArthur M, Hawrylycz M, Green RD, Navas PA, Noble WS, Stamatoyannopoulos JA: **Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays**. *Nat Methods* 2006, **3**(7):511-518.

113. Benson G: **Tandem repeats finder: a program to analyze DNA sequences**. *Nucleic Acids Res* 1999, **27**(2):573-580.

114. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D: **Identification and classification of conserved RNA secondary structures in the human genome**. *PLoS Comput Biol* 2006, **2**(4):e33.

115. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A: **Patterns of positive selection in six Mammalian genomes**. *PLoS Genet* 2008, **4**(8):e1000144.

116. Visel A, Minovitsky S, Dubchak I, Pennacchio LA: **VISTA Enhancer Browser--a database of tissue-specific human enhancers**. *Nucleic Acids Res* 2007, **35**(Database issue):D88-92.

117. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D: **A distal enhancer and an ultraconserved exon are derived from a novel retroposon**. *Nature* 2006, **441**(7089):87-90.

118. Lowe CB, Bejerano G, Haussler D: **Thousands of human mobile element fragments undergo strong purifying selection near developmental genes**. *Proc Natl Acad Sci U S A* 2007, **104**(19):8005-8010.

119. Jurka J: **Repbase Update - a database and an electronic journal of repetitive elements**. *Trends in Genetics* 2000, **16**(9):418-420.

120. Gould SJ, Vrba ES: **Exaptation - a Missing Term in the Science of Form**. *Paleobiology* 1982, **8**(1):4-15.

121. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: **Human-mouse alignments with BLASTZ**. *Genome Res* 2003, **13**(1):103-107.

122. Blanchette M, Bataille AR, Chen XY, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, Coulombe B, Robert FO: **Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression**. *Genome Research* 2006, **16**(5):656-668.

123. Ferretti V, Poitras C, Bergeron D, Coulombe B, Robert F, Blanchette M: **PReMod: a database of genome-wide mammalian cis-regulatory module predictions**. *Nucleic Acids Res* 2007, **35**(Database issue):D122-126.

124. Lunter G, Ponting CP, Hein J: **Genome-wide identification of human functional DNA using a neutral indel model**. *PLoS Comput Biol* 2006, **2**(1):e5.

125. Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King DC, Baertsch R, Blankenberg D, Pond SLK, Nekrutenko A, Giardine B, Harris RS, Diekhans STM, Diekhans M, Pringle TH, Murphy WJ, Lesk A, Weinstock GM, Lindblad-Toh K, Gibbs RA, Lander ES, Siepel A, Haussler D, Kent WJ: **28-way vertebrate alignment and conservation track in the UCSC Genome Browser**. *Genome Research* 2007, **17**(12):1797-1808.

126. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome**. *Science* 2009, **326**(5950):289-293.

127. Tung YC, Yeo GS: **From GWAS to biology: lessons from FTO**. *Ann N Y Acad Sci* 2011, **1220**:162-171.

128. Javierre BM, Fernandez AF, Richter J, Al-Shahrour F, Martin-Subero JI, Rodriguez-Ubreva J, Berdasco M, Fraga MF, O'Hanlon TP, Rider LG, Jacinto FV, Lopez-Longo FJ, Dopazo J, Forn M, Peinado MA, Carreno L, Sawalha AH, Harley JB, Siebert R, Esteller M, Miller FW, Ballestar E: **Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus**. *Genome Res* 2009.

129. Johnstone SE, Baylin SB: **Stress and the epigenetic landscape: a link to the pathobiology of human diseases?** *Nat Rev Genet* 2010, **11**(11):806-812.

130. Aragon T: **epitools: Epidemiology Tools.** . In., R package version 0.5-6. edn; 2010.

131. Team RDC: **A language and environment for statistical computing.** . Vienna, Austria. ; 2010.

132. Project IH: **A haplotype map of the human genome**. *Nature* 2005, **437**(7063):1299-1320.

133. Huang L, Jakobsson M, Pemberton TJ, Ibrahim M, Nyambo T, Omar S, Pritchard JK, Tishkoff SA, Rosenberg NA: **Haplotype variation and genotype imputation in African populations**. *Genet Epidemiol* 2011, **35**(8):766-780.

134. Sproul D, Gilbert N, Bickmore WA: **The role of chromatin structure in regulating the expression of clustered genes**. *Nat Rev Genet* 2005, **6**(10):775-781.

135. Gilbert N, Gilchrist S, Bickmore WA: **Chromatin organization in the mammalian nucleus**. *Int Rev Cytol* 2005, **242**:283-336.

136. Koudritsky M, Domany E: **Positional distribution of human transcription factor binding sites**. *Nucleic Acids Res* 2008, **36**(21):6795-6805.

137. Prendergast JG, Campbell H, Gilbert N, Dunlop MG, Bickmore WA, Semple CA: **Chromatin structure and evolution in the human genome**. *BMC Evol Biol* 2007, **7**:72.

138. Stahl EA, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, Voight BF, Kraft P, Chen R, Kallberg HJ, Kurreeman FA, Kathiresan S, Wijmenga C, Gregersen PK, Alfredsson L, Siminovitch KA, Worthington J, de Bakker PI, Raychaudhuri S, Plenge RM: **Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis**. *Nat Genet* 2012, **44**(5):483-489.

139. Kindt AS, Navarro P, Semple CA, Haley CS: **The genomic signature of trait-associated variants**. *BMC Genomics* 2013, **14**(1):108.

140. Agresti A, Coull BA: **Order-restricted tests for stratified comparisons of binomial proportions**. *Biometrics* 1996, **52**(3):1103-1111.

141. Pritchard JK, Przeworski M: **Linkage disequilibrium in humans: models and data**. *Am J Hum Genet* 2001, **69**(1):1-14.

142. **Logistic Regression - Extension chapters on advanced techniques** [http://www.uk.sagepub.com/burns/website%20material/Chapter%2024%20-%20Logistic%20regression.pdf]

143. Burns RP, Burns R: **Business Research Methods and Statistics Using SPSS**: SAGE Publications Ltd; 2009.

144. Fesinmeyer MD, North KE, Ritchie MD, Lim U, Franceschini N, Wilkens LR, Gross MD, Buzkova P, Glenn K, Quibrera PM, Fernandez-Rhodes L, Li Q, Fowke JH, Li R, Carlson CS, Prentice RL, Kuller LH, Manson JE, Matise TC, Cole SA, Chen CT, Howard BV, Kolonel LN, Henderson BE, Monroe KR, Crawford DC, Hindorff LA, Buyske S, Haiman CA, Le Marchand L, Peters U: **Genetic Risk Factors for BMI and Obesity in an Ethnically Diverse Population: Results From the Population Architecture Using Genomics and Epidemiology (PAGE) Study**. *Obesity (Silver Spring)* 2012.

145. McQuillan R, Eklund N, Pirastu N, Kuningas M, McEvoy BP, Esko T, Corre T, Davies G, Kaakinen M, Lyytikainen LP, Kristiansson K, Havulinna AS, Gogele M, Vitart V, Tenesa A, Aulchenko Y, Hayward C, Johansson A, Boban M, Ulivi S, Robino A, Boraska V, Igl W, Wild SH, Zgaga L, Amin N, Theodoratou E, Polasek O, Girotto G, Lopez LM, Sala C, Lahti J, Laatikainen T, Prokopenko I, Kals M, Viikari J, Yang J, Pouta A, Estrada K, Hofman A, Freimer N, Martin NG, Kahonen M, Milani L, Heliovaara M, Vartiainen E, Raikkonen K, Masciullo C, Starr JM, Hicks AA, Esposito L, Kolcic I, Farrington SM, Oostra B, Zemunik T, Campbell H, Kirin M, Pehlic M, Faletra F, Porteous D, Pistis G, Widen E, Salomaa V, Koskinen S, Fischer K, Lehtimaki T, Heath A, McCarthy MI, Rivadeneira F, Montgomery GW, Tiemeier H, Hartikainen AL, Madden PA, d'Adamo P, Hastie ND, Gyllensten U, Wright AF, van Duijn CM, Dunlop M, Rudan I, Gasparini P, Pramstaller PP, Deary IJ, Toniolo D, Eriksson JG, Jula A, Raitakari OT, Metspalu A, Perola M, Jarvelin MR, Uitterlinden A, Visscher PM, Wilson JF: **Evidence of inbreeding depression on human height**. *PLoS Genet* 2012, **8**(7):e1002655.

146. Wason JM, Dudbridge F: **Comparison of multimarker logistic regression models, with application to a genomewide scan of schizophrenia**. *BMC Genet* 2010, **11**:80.

147. Gaffney DJ, Veyrieras JB, Degner JF, Roger PR, Pai AA, Crawford GE, Stephens M, Gilad Y, Pritchard JK: **Dissecting the regulatory architecture of gene expression QTLs**. *Genome Biol* 2012, **13**(1):R7.

148. Akaike T: **A new look at the statistical model identification**. *IEEE Transactions on Automatic Control* 1974, **19**(6):716-723.

149. Venables WN, Ripley BD: **Modern Applied Statistics with S.** , Fourth edn: Springer, New York; 2002.

150. Long JS: **Regression Models for Categorical and Limited Dependent Variables.**, vol. 7: Sage Publications.; 1997.

151. Long JS, Freese J: **Regression Models for Categorical Outcomes Using Stata.**, Second Edition. edn: Stata Press; 2005.

152. Cohen J, Cohen P, West SG, Aiken LS: **Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.**, Third edn: Routledge; 2002.

153. **Introduction to SAS** [http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm]

154. Aquino J, Enzmann D, Schwartz M, Jain N: **descr: Descriptive Statistics**. In., 0.9.7 edn; 2009: This package contains functions to describe weighted categorical variables and functions to faciliate the character encoding conversion of objects.

155. Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, Dilthey A, Ellis P, Langford C, Vannberg FO, Knight JC: **Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles**. *Nat Genet* 2012, **44**(5):502-510.

156. Petersen A, Spratt J, Tintle NL: **Incorporating prior knowledge to increase the power of genome-wide association studies**. *Methods Mol Biol* 2013, **1019**:519-541.

157. Bis JC, Kavousi M, Franceschini N, Isaacs A, Abecasis GR, Schminke U, Post WS, Smith AV, Cupples LA, Markus HS, Schmidt R, Huffman JE, Lehtimaki T, Baumert J, Munzel T, Heckbert SR, Dehghan A, North K, Oostra B, Bevan S, Stoegerer EM, Hayward C, Raitakari O, Meisinger C, Schillert A, Sanna S, Volzke H, Cheng YC, Thorsson B, Fox CS, Rice K, Rivadeneira F, Nambi V, Halperin E, Petrovic KE, Peltonen L, Wichmann HE, Schnabel RB, Dorr M, Parsa A, Aspelund T, Demissie S, Kathiresan S, Reilly MP, Taylor K, Uitterlinden A, Couper DJ, Sitzer M, Kahonen M, Illig T, Wild PS, Orru M, Ludemann J, Shuldiner AR, Eiriksdottir G, White CC, Rotter JI, Hofman A, Seissler J, Zeller T, Usala G, Ernst F, Launer LJ, D'Agostino RB, Sr., O'Leary DH, Ballantyne C, Thiery J, Ziegler A, Lakatta EG, Chilukoti RK, Harris TB, Wolf PA, Psaty BM, Polak JF, Li X, Rathmann W, Uda M, Boerwinkle E, Klopp N, Schmidt H, Wilson JF, Viikari J, Koenig W, Blankenberg S, Newman AB, Witteman J, Heiss G, Duijn C, Scuteri A, Homuth G, Mitchell BD, Gudnason V, O'Donnell CJ: **Meta-analysis of genome-wide association studies from the CHARGE consortium identifies common variants associated with carotid intima media thickness and plaque**. *Nat Genet* 2011, **43**(10):940-947.

158. Ho JE, Levy D, Rose L, Johnson AD, Ridker PM, Chasman DI: **Discovery and replication of novel blood pressure genetic loci in the Women's Genome Health Study**. *J Hypertens* 2011, **29**(1):62-69.

159. Dermitzakis ET, Stranger BE: **Genetic variation in human gene expression**. *Mamm Genome* 2006, **17**(6):503-508.

160. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavare S, Deloukas P, Dermitzakis ET: **Population genomics of human gene expression**. *Nat Genet* 2007, **39**(10):1217-1224.

161. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, Todd JA, Donnelly P, Barrett JC, Davison D, Easton D, Evans DM, Leung HT, Marchini JL, Morris AP, Spencer CC, Tobin MD, Attwood AP, Boorman JP, Cant B, Everson U, Hussey JM, Jolley JD, Knight AS, Koch K, Meech E, Nutland S, Prowse CV, Stevens HE, Taylor NC, Walters GR, Walker NM, Watkins NA, Winzer T, Jones RW, McArdle WL, Ring SM, Strachan DP, Pembrey M, Breen G, St Clair D, Caesar S, Gordon-Smith K, Jones L, Fraser C, Green EK, Grozeva D, Hamshere ML, Holmans PA, Jones IR, Kirov G, Moskivina V, Nikolov I, O'Donovan MC, Owen MJ, Collier DA, Elkin A, Farmer A, Williamson R, McGuffin P, Young AH, Ferrier IN, Ball SG, Balmforth AJ, Barrett JH, Bishop TD, Iles MM, Maqbool A, Yuldasheva N, Hall AS, Braund PS, Dixon RJ, Mangino M, Stevens S, Thompson JR, Bredin F, Tremelling M, Parkes M, Drummond H, Lees CW, Nimmo ER, Satsangi J, Fisher SA, Forbes A, Lewis CM, Onnie CM, Prescott NJ, Sanderson J, Matthew CG, Barbour J, Mohiuddin MK, Todhunter CE, Mansfield JC, Ahmad T, Cummings FR, Jewell DP, Webster J, Brown MJ, Lathrop MG, Connell J, Dominiczak A, Marcano CA, Burke B, Dobson R, Gungadoo J, Lee KL, Munroe PB, Newhouse SJ, Onipinla A, Wallace C, Xue M, Caulfield M, Farrall M, Barton A, Bruce IN, Donovan H, Eyre S, Gilbert PD, Hilder SL, Hinks AM, John SL, Potter C, Silman AJ, Symmons DP, Thomson W, Worthington J, Dunger DB, Widmer B, Frayling TM, Freathy RM, Lango H, Perry JR, Shields BM, Weedon MN, Hattersley AT, Hitman GA, Walker M, Elliott KS, Groves CJ, Lindgren CM, Rayner NW, Timpson NJ, Zeggini E, Newport M, Sirugo G, Lyons E, Vannberg F, Hill AV, Bradbury LA, Farrar C, Pointon JJ, Wordsworth P, Brown MA, Franklyn JA, Heward JM, Simmonds MJ, Gough SC, Seal S, Stratton MR, Rahman N, Ban M, Goris A, Sawcer SJ, Compston A, Conway D, Jallow M, Rockett KA, Bumpstead SJ, Chaney A, Downes K, Ghori MJ, Gwilliam R, Hunt SE, Inouye M, Keniry A, King E, McGinnis R, Potter S, Ravindrarajah R, Whittaker P, Widden C, Withers D, Cardin NJ, Ferreira T, Pereira-Gale J, Hallgrimsdo'ttir IB, Howie BN, Su Z, Teo YY, Vukcevic D, Bentley D, Mitchell SL, Newby PR, Brand OJ, Carr-Smith J, Pearce SH, Reveille JD, Zhou X, Sims AM, Dowling A, Taylor J, Doan T, Davis JC, Savage L, Ward MM, Learch TL, Weisman MH, Brown M: **Association scan of 14,500 nonsynonymous SNPs in**

four diseases identifies autoimmunity variants. *Nat Genet* 2007, **39**(11):1329-1337.

162. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M: **Mapping complex disease traits with global gene expression**. *Nat Rev Genet* 2009, **10**(3):184-194.

163. Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET, Pritchard JK: **Gene expression levels are a target of recent natural selection in the human genome**. *Mol Biol Evol* 2009, **26**(3):649-658.

164. Guo YJ, Jamison DC: **The distribution of SNPs in human gene regulatory regions**. *BMC Genomics* 2005, **6**.

165. Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF, Kelsoe JR, O'Donovan MC, Furberg H, Schork NJ, Andreassen OA, Dale AM: **All SNPs Are Not Created Equal: Genome-Wide Association Studies Reveal a Consistent Pattern of Enrichment among Functionally Annotated SNPs**. *PLoS Genet* 2013, **9**(4):e1003449.

166. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA: **Systematic localization of common disease-associated variation in regulatory DNA**. *Science* 2012, **337**(6099):1190-1195.

167. Trynka G, Raychaudhuri S: **Using chromatin marks to interpret and localize genetic associations to complex human traits and diseases**. *Curr Opin Genet Dev* 2013.

168. Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, Raychaudhuri S: **Chromatin marks identify critical cell types for fine mapping complex trait variants.** *Nat Genet* 2013, **45**(2):124-130.

169. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M: **Linking disease associations with regulatory information in the human genome**. *Genome Res* 2012, **22**(9):1748-1759.

170. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M: **Annotation of functional variation in personal genomes using RegulomeDB**. *Genome Res* 2012, **22**(9):1790-1797.

171. Ward LD, Kellis M: **HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of**

genetically linked variants. *Nucleic Acids Res* 2012, **40**(Database issue):D930-934.

172. Hudson RS, Yi M, Volfovsky N, Prueitt RL, Esposito D, Volinia S, Liu CG, Schetter AJ, Van Roosbroeck K, Stephens RM, Calin GA, Croce CM, Ambs S: **Transcription signatures encoded by ultraconserved genomic regions in human prostate cancer**. *Mol Cancer* 2013, **12**:13.

173. Hadjiargyrou M, Delihas N: **The Intertwining of Transposable Elements and Non-Coding RNAs**. *Int J Mol Sci* 2013, **14**(7):13307-13328.

174. Khatun J, Yu Y, Wrobel JA, Risk BA, Gunawardena HP, Secrest A, Spitzer WJ, Xie L, Wang L, Chen X, Giddings MC: **Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions**. *BMC Genomics* 2013, **14**:141.

175. Morris AP, Zeggini E: **An evaluation of statistical approaches to rare variant analysis in genetic association studies**. *Genet Epidemiol* 2010, **34**(2):188-193.

176. Shriner D, Adeyemo A, Ramos E, Chen G, Rotimi CN: **Mapping of disease-associated variants in admixed populations**. *Genome Biol* 2011, **12**(5):223.

177. Need AC, Goldstein DB: **Next generation disparities in human genomics: concerns and remedies**. *Trends Genet* 2009, **25**(11):489-494.

178. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: **An integrated map of genetic variation from 1,092 human genomes**. *Nature* 2012, **491**(7422):56-65.

179. Gibson G: **Rare and common variants: twenty arguments**. *Nat Rev Genet* 2011, **13**(2):135-145.

180. Thomson PA, Parla JS, McRae AF, Kramer M, Ramakrishnan K, Yao J, Soares DC, McCarthy S, Morris SW, Cardone L, Cass S, Ghiban E, Hennah W, Evans KL, Rebolini D, Millar JK, Harris SE, Starr JM, Macintyre DJ, McIntosh AM, Watson JD, Deary IJ, Visscher PM, Blackwood DH, McCombie WR, Porteous DJ: **708 Common and 2010 rare DISC1 locus variants identified in 1542 subjects: analysis for association with psychiatric disorder and cognitive traits**. *Mol Psychiatry* 2013.

181. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB: **Annotating non-coding regions of the genome**. *Nat Rev Genet* 2010, **11**(8):559-571.

182. Hancks DC, Kazazian HH, Jr.: **Active human retrotransposons: variation and disease**. *Curr Opin Genet Dev* 2012, **22**(3):191-203.

183. Kalf RR, Mihaescu R, Kundu S, de Knijff P, Green RC, Janssens AC: **Variations in predicted risks in personal genome testing for common complex diseases**. *Genet Med* 2013.

9 APPENDIX

9.1 Traits divided into four different subsets

9.1.1 Immune traits

- Activated partial thromboplastin time
- Acute anterior uveitis
- Acute chest syndrome asthma
- Acute graft-versus-host disease
- Acute renal allograft rejection
- Acute respiratory distress syndrome IL* production
- Addison's disease
- Aggressive periodontitis
- AIDS
- AIDS progression
- Airway hyperresponsiveness atopy
- Airway hyperresponsiveness, methacholine induced
- Allele frequency/ normal
- Allergic asthma
- Allergic bronchopulmonary aspergillosis abpa
- Allergic disease
- Allergic disease, ige -mediated
- Allergic diseases (bronchial asthma. Atopic dermatitis and/or food-related anaphylaxis)
- Allergic rhinitis
- Allergic rhinitis asthma
- Allergic rhinitis dermatitis and eczema fatty acid
- Allergic rhinitis IgE
- Allergies; common cold
- Allergy
- Allergy asthma
- Allergy dermatitis and eczema
- Allergy, latex; latex allergy
- Allergy, latex; latex allergy; pemphigoid, bullous
- Allogenic stem cell transplantation
- Allograft dysfunction, renal
- Allograft outcome
- Allograft rejection, heart
- Alopecia areata
- Altered CCR5 Expression or coreceptor function
- Alveolitis, extrinsic allergic
- Alzheimer`s disease
- ANCA positive patients
- Ankylosing spondylitis
- Annexin A5 antibodies
- Anti-cyclic citrullinated peptide antibodies rheumatoid arthritis
- Anti-GAD65 antibody

- Anti-islet autoantibodies diabetes, type 1
- Anti-neutrophil cytoplasmic antibodies kidney failure, chronic polyangiitis wegener's granulomatosis
- Anti-ro 52-KD autoantibodies
- Anti-ro autoantibodies
- Antibody formation crohn's disease ulcerative colitis
- Anticardiolipin antibody production lupus erythematosus
- Antineutrophil cytoplasmic antibody-associated vasculitis
- Antineutrophil cytoplasmic antibody; (ANCA)-associated vasculitis
- Antiphospholipid syndrome
- Aplastic anemia, acquired
- Apolipoprotein levels
- Arthritis
- Arthritis (juvenile idiopathic)
- Arthritis lupus erythematosus
- Aseptic abscesses crohn's disease inflammatory bowel disease
- Aspirin-induced asthma
- Aspirin-intolerant asthma
- Asthma
- Asthma (childhood onset)
- Asthma in combination with a variety of other diseases
- Atherosclerosis, coronary
- Atopic asthma
- Atopic asthma. BHR. Total IgE. SPT
- Atopic dermatitis
- Atopic eczema
- Atopy
- Atopy (IgE)
- Atopy (total & specific IgE)
- Atopy IgE urticaria, aspirin-intolerant
- Atopy vaccine response
- Atopy-susceptibility
- Atopy; dermatitis and eczema
- Atopy; IgE levels
- Atopy. Airway obstruction. BHR. Asthma
- Atopy. Asthma
- Atopy. Asthma. Netherton
- Atopy. BHR
- Atopy. Spige. Total IgE. Asthma. Atopic asthma
- Atopy/ asthma
- Autoimmunity
- Autologous mixed lymphocyte reaction
- Bee venom allergy
- Behcet's disease
- Beta cell autoimmunity
- B-cell function; diabetes, type 1
- Betacl osteocalcin
- BHR
- Birth weight bronchopulmonary dysplasia sepsis
- Blau syndrome
- Bone marrow transplantation
- Bronchial asthma

- Bronchial asthma (childhood & adult)
- Bronchial asthma (childhood only)
- Bronchial hyperreactivity
- Bronchial hyperresponsiveness
- Bronchopulmonary dysplasia respiratory distress syndrome, neonatal
- Bullous pemphigoid
- C-reactive protein
- Carotid atherosclerosis in HIV Infection
- CD14 expression
- CD14 levels
- Cedar pollinosis
- Celiac disease
- Celiac disease diabetes, type 1
- Celiac disease gluten intolerance
- Celiac disease lupus erythematosus rheumatoid arthritis
- Celiac disease; colitis
- Cerebral malaria
- Childhood asthma
- Childhood atopic asthma
- Childhood atopic asthma
- Childhood B-cell non-hodgkin's lymphoma
- Chinese ankylosing spondylitis patients
- Cholangitis, sclerosing
- Cholangitis, sclerosing crohn's disease inflammatory bowel disease
- Cholangitis, sclerosing crohn's disease ulcerative colitis
- Cholesterol, LDL; cholesterol, total; C-reactive protein; APOA2; APOB
- Chronic bronchitis
- Chronic hepatitis C infection
- Chronic immune thrombocytopenic purpura.
- Chronic nonproductive cough
- Chronic obstructive pulmonary disease
- Chronic pancreatitis
- Chronic periodontitis.
- Chronic progressive multiple sclerosis.
- Cirrhosis, biliary primary
- Cirrhosis, biliary primary hepatitis, autoimmune
- Coeliac disease.
- Collagen disease juvenile arthritis rheumatoid arthritis still's disease
- Common variable immunodeficiency
- Congenital thrombotic thrombocytopenic purpura
- Contact allergy
- Contact hypersensitivity
- Contact sensitisation
- COPD
- Coronary heart disease
- Crohn's disease
- Cryoglobulinemia
- Cutaneous neonatal lupus
- Cytokine lung function
- Cytokine release
- Cytokine release mortality
- Cytokine response to measles vaccine

- Cytokine synthesis
- Cytokines; tumor markers
- Decreased airway responsiveness
- Dengue shock syndrome
- Dermatitis and eczema
- Dermatitis herpetiformis.
- Dermatitis, atopic
- Dermatomyositis myopathy, idiopathic inflammatory polymyositis
- Dermatomyositis polymyositis
- Diabetes (gestational)
- DRS
- Early onset of multiple sclerosis.
- Early onset periodontitis
- Early onset psoriasis
- Early polyarthritis
- Early-onset periodontitis
- Eczema
- Eczema food allergy IgE
- Emphysema
- Eosinophil counts
- Eosinophilia
- Eosinophilic esophagitis (pediatric)
- Epithelial neutrophil activating peptide
- Epstein-barr virus
- Erythema nodosum
- Familial hemophagocytic lymphohistiocytosis
- Familial juvenile onset psoriasis
- Familial mediterranean fever
- FAS levels
- FEV1
- Food allergy
- Fuchs heterochromic cyclitis
- Glomerulonephritis, hepatitis B virus-associated
- Graft acceptance, liver
- Graft occlusion, atherosclerotic
- Graft rejection, liver
- Graft versus host disease
- Graves' disease
- Graves' hyperthyroidism
- Graves' ophthalmopathy
- H-thyroiditis
- Haemophilia with chronic synovitis
- Hashimoto's thryoiditis
- Hematology indices
- Hemophagocytic lymphohistiocytosis
- Henoch-schonlein purpura
- Hepatitis B
- Hepatitis B (viral clearance)
- Hepatitis C induced liver fibrosis
- Hepatitis type 1, autoimmune (AIH-1)
- Hepatitis type 2, autoimmune
- HIV

- HIV-1 control
- HIV-1 infection
- HLA-associated diseases
- HPV seropositivity
- Human T-cell lymphotropic virus type I associated myelopathy
- Hyper-IgE syndrome and severe eczema. Atopy
- Hyper-IgM syndrome
- Hyper-IgM syndrome can form oligomers and trigger CD40-mediated signals
- Hyperresponsiveness
- Hypothyroidism
- Hypothyroidism, autoimmune
- Hypothyroidism, goitrous juvenile autoimmune
- Idiopathic chronic pancreatitis
- Idiopathic inflammatory myopathies
- IgA
- IgA deficiency
- IgA deficiency and common variable immunodeficiency
- IgA nephropathy
- IgD
- IgE
- IgE grass sensitization
- IgE levels
- IgE response
- IGF-I
- IGF-I levels; IGFBP-3 levels
- IgG
- IgM
- IL-18 concentration physical functioning
- IL-18 lupus erythematosus
- IL-1beta
- IL-4
- IL18 expression level
- IL6 transcription
- Immune deficiency
- Immunoglobulin A deficiency
- Immunoglobulin A glomerulonephritis
- Immunology study
- Immunotherapy response
- Improved survival in sepsis
- Increased expression of the G gamma and A gamma globin
- Increased IgE
- Increased interleukin-10 (IL-10) plasma levels
- Inflammatory bowel disease
- Inflammation
- Inflammation oxidative stress
- Inflammatory biomarkers
- Inflammatory bowel disease
- Inflammatory disease
- Inflammatory markers
- Inflammatory myopathies
- Inflammatory response
- Inflammatory urogenital disease

- Insulin dependent diabetes
- Interferon response
- Interleukin-1 beta (IL-1 beta) synthesis capacity
- Irritable bowel syndrome
- Juvenile ankylosing spondylitis
- Juvenile arthritis
- Juvenile idiopathic arthritis
- Juvenile rheumatoid arthritis
- Kawasaki disease
- Kidney transplant
- Kidney transplant complications
- Kidney transplant complications; lipids
- Knee osteoarthritis
- Latex allergy
- Leprosy
- Leukemia virus type I
- Liver transplantation, immunosuppression after
- Lung function
- Lupus
- Lupus erythematosus
- Lupus nephritis
- Malaria
- Measles vaccine immunity
- Microscopic polyangiitis
- Microsomal epoxide hydrolase
- Mite-sensitive asthma
- Monocyte chemoattractant protein-1
- Morbidity mortality
- Multiple sclerosis
- Myasthenia gravis
- Myositis
- Narcolepsy
- Neonatal lupus
- Nephropathy, IgA
- Neutrophil immunodeficiency syndrome
- No exhalation
- Nocturnal asthma
- Osteoarthritis
- Osteomyelitis
- Otitis media
- Pancreatitis
- Peanut allergy
- Pemphigus
- Pemphigus foliaceus
- Pemphigus vulgaris
- Penicillins allergy
- Periodontal disease
- Periodontitis
- Physician diagnosed asthma
- Pityriasis rosea
- Plasma IL6 levels
- Plasminogen activator inhibitor type 1 levels (PAI-1)

- Pneumoconiosis
- Pollen allergy
- Pollen-induced allergic rhinitis
- Pollinosis, cedar
- Polymylagia rheumatica
- Polymyositis and dermatomyositis.
- Polyneuropathy vasculitis
- Postoperative systemic inflammatory reaction
- Postpartum thyroiditis
- Primary biliary cirrhosis
- Primary sclerosing cholangitis
- Primary sjogren's syndrome
- Psoriasis
- Reactive arthritis
- Reiter's syndrome
- Renal disease
- Renal transplant rejection
- Respiratory syncytial virus
- Response to endotoxin
- Retinopathy, diabetic; nephropathy in other diseases
- Rheumatic diseases
- Rheumatic fever
- Rheumatic heart disease
- Rheumatoid arthritis
- Rhinitis
- Rickets
- Rubella vaccine, cytokine response to
- Sarcoidosis
- Sarcoidosis
- Sarcoidosis tuberculosis
- Sarcoidosis uveitis
- Scleroderma
- Scleroderma; jaundice
- Sclerosing cholangitis and ulcerative colitis (combined)
- Sclerosis, systemic
- Semple rabies vaccine-induced autoimmune encephalomyelitis
- Sepsis
- Sepsis development or mortality
- Septic shock
- Serum IgE levels
- Severe asthma
- Severe chronic neutropenia
- Severe combined immunodeficiency
- Severe ulcerative colitis
- Silicosis
- Sinusitis
- Sjogren's syndrome
- SLE
- Soluble CD14 plasma levels
- Specific IgE
- Spondyloarthropathies
- SPT

- Staphylococcal infection
- Steroid-dependent asthma
- Steroid-requiring asthma in sedentary women
- Stevens-Johnson syndrome
- Still's disease
- Sulfasalazine, adverse effects of
- Syncytial virus bronchiolitis
- Systemic inflammatory response syndrome
- Systemic juvenile idiopathic arthritis
- Systemic lupus erythematosus
- Systemic scleroderma
- Systemic sclerosis
- Thimerosal sensitization
- Thrombosis, deep vein; Behcet's disease
- Thryoiditis, chronic lymphocytic
- Thyroid autoimmunity
- Thyroiditis, chronic lymphocytic
- Thyroiditis, Hashimoto's
- Thyrotoxic hypokalemic periodic paralysis
- TIgE
- TNF-Alpha
- Total IgE
- Total serum IgE
- Tropical calcific pancreatitis
- Tuberculosis
- Tumor necrosis factor receptor-associated periodic syndrome
- Type 1 diabetes
- Type 1 diabetes autoantibodies
- Type 1 diabetes nephropathy
- Type 2 diabetes
- Type 2 diabetes and other traits
- Ulcerative colitis
- Vaspin levels
- Vitiligo
- Vogt-Koyanagi-Harada's disease
- Wheeze
- White blood cell types
- X-linked lymphoproliferative disease
- X-linked severe combined immunodeficiency

9.1.2 Cancer traits

- Acute lymphoblastic leukemia (childhood)
- Basal cell carcinoma
- Basal cell carcinoma (cutaneous)
- Bladder cancer
- Breast cancer
- Breast cancer (male)
- Carcinoma
- Chronic lymphocytic leukemia
- Chronic myeloid leukemia
- Colorectal cancer

- Endometrial cancer
- Erectile dysfunction and prostate cancer treatment
- Esophageal cancer
- Esophageal cancer (alcohol interaction)
- Esophageal cancer (squamous cell)
- Esophageal cancer and gastric cancer
- Ewing sarcoma
- Follicular lymphoma
- Gastric cancer
- Glaucoma
- Glaucoma (exfoliation)
- Glaucoma (primary open-angle)
- Glioma
- Glioma (high-grade)
- Hepatocellular carcinoma
- Hodgkin's lymphoma
- Lung adenocarcinoma
- Lung cancer
- Melanoma
- Meningioma
- Multiple myeloma
- Myeloproliferative neoplasms
- Nasopharyngeal carcinoma
- Neuroblastoma
- Neuroblastoma (high-risk)
- Non-small cell lung cancer
- Ovarian cancer
- Pancreatic cancer
- Prostate cancer
- Renal cell carcinoma
- Testicular cancer
- Testicular germ cell cancer
- Testicular germ cell tumor
- Thyroid cancer
- Upper aerodigestive tract cancers
- Urinary bladder cancer
- Wilms tumor
- YKL-40 levels
- Multiple cancers (lung cancer and gastric cancer and squamous cell carcinoma)

9.1.3 Normal Variation traits

- Acenocoumarol maintenance dosage
- Activated partial thromboplastin time
- Adiponectin levels
- Aging
- Aging traits
- Alcohol and nictotine co-dependence
- Alcohol consumption
- Amyloid A levels
- Androgen levels

- Angiotensin-converting enzyme activity
- Ankle-brachial index
- Anthropometric traits
- Anticoagulant levels
- Antipsychotic drug-induced weight gain
- Aortic root size
- Aortic stiffness
- Arterial stiffness
- Aspartate aminotransferase
- Bilirubin levels
- Biochemical measures
- Birth weight
- Bitter taste response
- Black *vs.* Blond hair color
- Black *vs.* Red hair color
- Bleomycin sensitivity
- Blond *vs.* Brown hair color
- Blood lipid traits
- Blue *vs.* Brown eyes
- Blue *vs.* Green eyes
- Body mass (lean)
- Brain structure
- Breast size
- Burning and freckling
- Butyrylcholinesterase levels
- C4B binding protein levels
- Caffeine consumption
- Calcium levels
- Cannabis dependence
- Capecitabine sensitivity
- Cardiac hypertrophy
- Cardiac repolarization
- Cardiac structure and function
- Carotenoid and tocopherol levels
- Carotid intima media thickness
- CD4:CD8 lymphocyte ratio
- Central corneal thickness
- Cholelithiasis-related traits in sickle cell anemia
- Circulating cell-free DNA
- Coagulation factor levels
- Coffee consumption
- Cognitive decline
- Cognitive function
- Common traits (other)
- Complement C3 and C4 levels
- Corneal astigmatism
- Corneal curvature
- Corneal structure
- Cortical structure
- Cortical thickness
- Creatinine levels
- Cutaneous nevi

- Cystatin C
- D-dimer levels
- Dehydroepiandrosterone sulphate levels
- Dental caries
- Diastolic blood pressure
- Drinking behavior
- Drug-induced liver injury
- Drug-induced liver injury (amoxicillin-clavulanate)
- Drug-induced liver injury (flucloxacillin)
- E-selectin levels
- Electrocardiographic traits
- Electroencephalographic traits in alcoholism
- Eosinophil counts
- Epirubicin-induced leukopenia
- Erythrocyte sedimentation rate
- Exercise (leisure time)
- Exercise treadmill test traits
- Eye color
- Eye color traits
- F-cell distribution
- Facial morphology
- Factor VII
- Fasting glucose-related traits
- Fasting glucose-related traits (interaction with BMI)
- Fasting insulin-related traits
- Fasting insulin-related traits (interaction with BMI)
- Fasting plasma glucose
- Fetal hemoglobin levels
- Fibrinogen
- Folate pathway vitamin levels
- Freckles
- Freckling
- Gamma gluatamyl transferase levels
- Gamma glutamyl transpeptidase
- Glycated hemoglobin levels
- Hair color
- Hair morphology
- Handedness in dyslexia
- Haptoglobin levels
- HBA2 levels
- HDL cholesterol
- HDL cholesterol - triglycerides (HDLC-TG)
- Head circumference (infant)
- Heart failure
- Height
- Hematocrit
- Hematological and biochemical traits
- Hematological parameters
- Hematology traits
- Hemoglobin
- Hemostatic factors and hematological phenotypes
- Hepatitis B vaccine response

- Hepcidin levels
- Hippocampal volume
- Homocysteine levels
- HPV seropositivity
- Hypertension
- Hypertension risk in short sleep duration
- IFN-related cytopenia
- IgE levels
- IgG levels
- IgM
- Immune reponse to smallpox (secreted IFN-alpha)
- Immune reponse to smallpox (secreted IL-10)
- Immune reponse to smallpox (secreted IL-12p40)
- Immune reponse to smallpox (secreted IL-1beta)
- Immune reponse to smallpox (secreted IL-2)
- Immune reponse to smallpox (secreted TNF-alpha)
- Immune response to smallpox vaccine (IL-6)
- Immunoglobulin A
- Insulin-like growth factors
- Insulin-related traits
- Interleukin-18 levels
- Intracranial volume
- Intraocular pressure
- Iris characteristics
- Iris color
- Iron levels
- Iron status biomarkers
- Keloid
- Left ventricular mass
- Lentiform nucleus volume
- Lipid levels in hepatitis c treatment
- Lipid metabolism phenotypes
- Lipid traits
- Lipoprotein-associated phospholipase A2 activity and mass
- Liver enzyme levels
- Liver enzyme levels (alanine transaminase)
- Liver enzyme levels (alkaline phosphatase)
- Liver enzyme levels (gamma-glutamyl transferase)
- Longevity
- LP (A) levels
- Lumiracoxib-related liver injury
- Magnesium levels
- Major depressive disorder
- Male-pattern baldness
- Mammographic density
- Matrix metalloproteinase levels
- Mean corpuscular hemoglobin
- Mean corpuscular volume
- Mean platelet volume
- Menarche
- Menarche (age at onset)
- Menarche and menopause (age at onset)

- Menopause
- Menopause (age at onset)
- Metabolic traits
- Monocyte chemoattractant protein-1
- Morbidity-free survival
- MRI atrophy measures
- N-glycan levels
- Natriuretic peptide levels
- Neuranatomic and neurocognitive phenotypes
- Neutrophil count
- Nevirapine-induced rash
- Nicotine dependence
- Non-albumin protein levels
- Normalized brain volume
- Obesity
- Obesity and blood pressure
- Obesity-related traits
- Optic disc parameters
- Optic disc parameters
- Optic disc size
- Optic disc size (disc)
- Other erythrocyte phenotypes
- Other metabolic traits
- Pain
- Pericardial fat
- Permanent tooth development
- Personality dimensions
- Phospholipid levels (plasma)
- Phosphorus levels
- Phytosterol levels
- Plasma C4B binding protein levels
- Plasma carotenoid and tocopherol levels
- Plasma coagulation factors
- Plasma E-selectin levels
- Plasma eosinophil count
- Plasma homocysteine
- Plasma level of vitamin B12
- Plasma levels of liver enzymes
- Plasma levels of protein C
- Platelet aggregation
- Platelet counts
- PR interval
- Primary sclerosing cholangitis
- Primary tooth development (number of teeth)
- Primary tooth development (time to first tooth eruption)
- Progranulin levels
- Proinsulin levels
- Prostate-specific antigen levels
- Protein biomarker
- Protein quantitative trait loci
- Prothrombin time
- Pulmonary function

- Pulmonary function decline
- Pulmonary function measures, QT interval
- Quantitative traits
- Reasoning
- Recombination rate (females)
- Recombination rate (males)
- Red blood cell traits
- Red *vs.* Non-red hair color
- Refractive error
- Renal function and chronic kidney disease
- Renal function-related traits (bun)
- Renal function-related traits (EGRFCREA)
- Renal function-related traits (SCR)
- Renal function-related traits (urea)
- Resistin levels
- Response to antidepressants
- Response to antipsychotic therapy (extrapyramidal side effects)
- Response to antipsychotic treatment
- Response to citalopram treatment
- Response to clopidogrel therapy
- Response to fenofibrate
- Response to gemcitabine in pancreatic cancer
- Response to hepatitis C treatment
- Response to interferon beta therapy
- Response to metformin
- Response to statin therapy
- Response to statin therapy (LDL-C)
- Response to tocilizumab in rheumatoid arthritis
- Response to vitamin E supplementation
- Resting heart rate
- Retinal vascular caliber
- Retinol levels
- Ribavirin-induced anemia
- RR interval (heart rate)
- Select biomarker traits
- Serum albumin level
- Serum bilirubin levels
- Serum calcium
- Serum creatinine
- Serum dehydroepiandrosterone sulphate levels
- Serum IgE levels
- Serum iron levels
- Serum markers of iron status
- Serum metabolites
- Serum phosphorus levels
- Serum phytosterol levels
- Serum prostate-specific antigen levels
- Serum soluble E-selectin
- Serum total protein level
- Serum urate
- Serum uric acid
- Sex hormone-binding globulin levels

- Skin pigmentation
- Skin sensitivity to sun
- Sleepiness
- Smoking behavior
- Soluble E-selectin levels
- Soluble leptin receptor levels
- Soluble levels of adhesion molecules
- Speech perception in dyslexia
- Sphingolipid levels
- Systolic blood pressure
- T-tau
- Tanning
- Telomere length
- Testosterone levels
- Thyroid function
- Thyroid volume
- Triglycerides
- Triglycerides-blood pressure (TG-BP)
- Two-hour glucose challenge
- Urate levels
- Uric acid levels
- Urinary albumin excretion
- Urinary metabolites
- Vascular endothelial growth factor levels
- Vaspin levels
- Venous thromboembolism
- Ventricular conduction
- Vertical cup-disc ratio
- Visceral adipose tissue/subcutaneous adipose tissue ratio
- Visceral fat
- Vitamin B12 levels
- Vitamin D insufficiency
- Vitamin D levels
- Vitamin E levels
- Volumetric brain MRI
- Waist circumference
- Waist circumference - triglycerides (WC-TGS)
- Waist circumference and related phenotypes
- Waist-hip ratio
- Warfarin maintenance dose
- Weight
- White blood cell count
- White blood cell types
- White matter hyperintensity burden
- Working memory
- Wrist bone mass

9.1.4 Disease traits

- AB1-42
- Abdominal aortic aneurysm
- Acute lymphoblastic leukemia (childhood)
- Adiposity
- Age-related macular degeneration (wet)
- Age-related macular degeneration
- Age-related macular degeneration (CNV *vs.* GA)
- Age-related macular degeneration (CNV)
- Age-related macular degeneration (GA)
- Alcohol dependence
- Alopecia areata
- Alzheimer's disease
- Alzheimer's disease (age of onset)
- Alzheimer's disease (late onset)
- Alzheimer's disease biomarkers
- Amyotrophic lateral sclerosis
- Ankylosing spondylitis
- Arthritis (juvenile idiopathic)
- Asthma
- Atrial fibrillation/atrial flutter
- Attention deficit hyperactivity disorder
- Autism
- Barrett's esophagus
- Basal cell carcinoma
- Beta thalassemia/hemoglobin e disease
- Biliary atresia
- Biomedical quantitative traits
- Bipolar disorder
- Bipolar disorder and major depressive disorder (combined)
- Bipolar disorder and schizophrenia
- Bladder cancer
- Blood pressure
- Body mass in chronic obstructive pulmonary disease
- Body mass index
- Bone mineral density
- Bone mineral density (hip)
- Bone mineral density (spine)
- Breast cancer
- Breast cancer (male)
- C-reactive protein
- C-reactive protein and white blood cell count
- Cardiovascular disease risk factors
- Carotid atherosclerosis in HIV infection
- Celiac disease
- Celiac disease and rheumatoid arthritis
- Cholesterol (total)
- Chronic hepatitis C infection
- Chronic kidney disease
- Chronic kidney disease and serum creatinine levels

- Chronic lymphocytic leukemia
- Chronic myeloid leukemia
- Chronic obstructive pulmonary disease
- Cleft lip
- Colorectal cancer
- Conduct disorder (symptom count)
- Coronary artery calcification
- Coronary heart disease
- Creutzfeldt-Jakob disease
- Creutzfeldt-Jakob disease (variant)
- Crohn's disease
- Crohn's disease and celiac disease
- Crohn's disease and psoriasis
- Cystic fibrosis severity
- Diabetes (gestational)
- Diabetic retinopathy
- Diastolic blood pressure
- Dilated cardiomyopathy
- Disc degeneration (lumbar)
- Drinking behavior
- Duodenal ulcer
- Dupuytren's disease
- Electrocardiographic traits
- End-stage renal disease (non-diabetic)
- Endometrial cancer
- Endometriosis
- Eosinophilic esophagitis (pediatric)
- Epilepsy
- Epilepsy (generalized)
- Erectile dysfunction
- Erectile dysfunction and prostate cancer treatment
- Esophageal cancer
- Esophageal cancer (alcohol interaction)
- Esophageal cancer (squamous cell)
- Essential tremor
- Ewing sarcoma
- Fasting plasmaglucose
- Follicular lymphoma
- Fuchs's corneal dystrophy
- Gallstones
- Gamma glutamyltranspeptidase
- Gastric cancer
- Glaucoma
- Glaucoma (primary open-angle)
- Glioma
- Glioma (high-grade)
- Glomerulosclerosis
- Glycated hemoglobinlevels
- Graves' disease
- HDL cholesterol
- HDL cholesterol - triglycerides (HDLC-TG)
- Hematological and biochemical traits

- Hepatitis B
- Hepatocellular carcinoma
- Hippocampal atrophy
- Hirschsprung's disease
- Hodgkin's lymphoma
- Hypertension
- Hypertriglyceridemia
- Hypospadias
- Hypothyroidism
- Idiopathic pulmonary fibrosis
- IgA nephropathy
- IgE levels
- Infantile hypertrophic pyloric stenosis
- Inflammatory bowel disease
- Inflammatory bowel disease (early onset)
- Intracranial aneurysm
- Kawasaki disease
- Kidney stones
- Knee osteoarthritis
- LDL cholesterol
- Leprosy
- Lipid metabolism phenotypes
- Lipoprotein-associated phospholipase A2 activity and mass
- Liver enzyme levels (alanine transaminase)
- Liver enzyme levels (alkalinephosphatase)
- Liver enzyme levels (gamma-glutamyl transferase)
- Longevity
- Lung adenocarcinoma
- Lung cancer
- Major depressive disorder
- Major mood disorders
- Malaria
- Melanoma
- Menarche (age at onset)
- Meningioma
- Meningococcal disease
- Metabolic syndrome
- Metabolic syndrome (bivariate traits)
- Metabolic traits
- Metabolite levels
- Migraine
- Moyamoya disease
- Multiple cancers (lung cancer and gastric cancer and squamous cell carcinoma)
- Multiple myeloma
- Multiple sclerosis
- Myasthenia gravis
- Myeloproliferative neoplasms
- Myocardial infarction
- Myocardial infarction (early onset)
- Myopia (pathological)
- Narcolepsy
- Nephrolithiasis

- Nephropathy
- Nephropathy (idiopathic membranous)
- Neuroblastoma
- Neuroblastoma (high-risk)
- Non-alcoholic fatty liver disease histology (other)
- Non-obstructive azoospermia
- Non-small cell lung cancer
- Nonalcoholic fatty liver disease
- Obesity (early onset extreme)
- Obesity (extreme)
- Orofacial clefts
- Osteoarthritis
- Osteoporosis
- Otosclerosis
- Ovarian cancer
- Paget's disease
- Pancreatic cancer
- Panic disorder
- Parkinson's disease
- Periodontitis
- Phospholipid levels (plasma)
- Plasminogen activator inhibitor type 1 levels (PAI-1)
- Polycystic ovary syndrome
- Primary biliary cirrhosis
- Primary sclerosing cholangitis
- Progressive supranuclear palsy
- Proinsulin levels
- Prostate cancer
- Protein quantitative traitloci
- Psoriasis
- Psoriatic arthritis
- Pulmonary function
- Renal cell carcinoma
- Renal function and chronic kidney disease
- Restless legs syndrome
- Rheumatoid arthritis
- Sarcoidosis
- Schizophrenia
- Schizophrenia and bipolar disorder and depression (combined)
- Sclerosing cholangitis and ulcerative colitis (combined)
- Scoliosis
- Soluble E-selectin levels
- Soluble levels of adhesion molecules
- Stevens-johnson syndrome and toxic epidermal necrolysis (SJS-TEN)
- Stroke
- Stroke (ischemic)
- Sudden cardiac arrest
- Suicide attempts in bipolar disorder
- Systemic lupus erythematosus
- Systemic sclerosis
- Systolic blood pressure
- Temperament (bipolar disorder)

- Testicular cancer
- Testicular germ cell cancer
- Testicular germ cell tumor
- Thoracic aortic aneurysms and dissections
- Thyroid cancer
- Thyrotoxic hypokalemic periodic paralysis
- Tourette syndrome
- Triglycerides
- Tuberculosis
- Two-hour glucose challenge
- Type 1 diabetes
- Type 1 diabetes autoantibodies
- Type 1 diabetes nephropathy
- Type 2 diabetes
- Type 2 diabetes and othertraits
- Ulcerative colitis
- Urinary bladder cancer
- Uterine fibroids
- Vitiligo
- Wilms tumor
- YKL-40 levels
- Response to antipsychotic therapy (extrapyramidal side effects)

9.2 R code for LogReg2 model

This function was taken from the R package descr [154] and has been included here for reference.

```
LogRegR2 = function (model) {# version 2.0, 22-Jan-2012, Dirk Enzmann
 # Calculates multiple R² analogs (pseudo R²) of logistic regression:

if ((model$family$family != "binomial") | (model$family$link != "logit"))
{ stop('No logistic regression model, no pseudo R² computed\n') }

 n = dim(model$model)[1]
 Chi2 = model$null - model$dev
 Df = model$df.null - model$df.res
 p = 1-pchisq(Chi2,Df)

 lp = predict(model)
 var_lp = var(lp)

 RL2 = Chi2/model$null # also called McFaddens R²
 Cox = 1-exp(-Chi2/n) # Cox & Snell Index
 Nag = Cox/(1-exp(-model$null/n)) # Nagelkerke Index
 MZ = var_lp/(var_lp + pi^2/3) # McKelvey & Zavoina's R²

list('Chi2'=Chi2,'df'=Df,'p'=p,'RL2'=RL2,'CoxR2'=Cox,'NagelkerkeR2'=Nag,'McKelvey
_ZavoinaR2'=MZ)
}
```