The effect of non-coding variants on gene transcription in human blood cell types

Roman Kreuzhuber

Department of Haematology
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Trinity Hall   November 2018

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Roman Kreuzhuber
November 2018

Acknowledgements

I would like to thank my supervisors, Prof. Willem H. Ouwehand (University of Cambridge, UK) and Dr. Oliver Stegle (EMBL - European Bioinformatics Institute, UK), for their guidance in scientific matters. I am grateful for the trust and faith I was given by Prof. Ouwehand; his open support has given me confidence. I would like to thank Prof. Michel A. J. Georges (Université de Liège, Belgium), Prof. Julian C. Knight (University of Oxford, UK) and Prof. Sir Nilesh J. Samani (University of Leicester, UK) for their support in sharing eQTL data and for the discussions on different ways to analyse eQTL and Promotor Capture HiC data. Here I would like to highlight my great appreciation for Prof. Georges' agreement to share eQTL data prior to publication.

I want to thank Dr. Tadbir Bariana and Ziga Avsec for the fruitful collaborations of which the results are included in my thesis.

Especially, I would like to thank my partner, Carmen, for her love, strong moral support and patience. Her and our Dachshund Loki always managed to lighten my mood, especially in desperate times. I thank my friends here in Cambridge for all the good times and small adventures. Lastly, I would like to thank my parents, Anna and Wolfgang, whose steady encouragement and support have enabled me to get to this point.

Abstract

To understand complex genetic diseases it is necessary to study DNA, its transcription, translation and regulation thereof. In a mechanistic view diseases can be caused by alterations in the DNA sequence or by dysregulation of gene expression. To understand cell regulation, first connections have to be found between elements contributing to gene expression and its regulation.

During my PhD I have studied the regulatory effects of human genetic variation. To do so I have processed and analysed datasets measuring effects of genetic variants on transcript levels of genes, identified regulatory variants and put them in bigger biochemical and physiological context.

Expression quantitative trait locus (eQTL) studies identify genetically explainable gene expression variation in a tissue and potentially cell type specific manner. I have reprocessed and aggregated eQTL datasets of seven purified blood cell types that were generated by collaborators in different laboratories. I showed that the increased sample size enables the identification of additional associations, also with low-frequency variants, and I compared my cell type specific eQTL results to results from an eQTL study performed on whole blood without cell type purification.

Gene regulation is complex and relies on several layers of control, and only one of them is genetic background. To help understand genetic control, I compared my eQTL results across the seven cell types and highlighted cell type specific associations.

I put my eQTL results into bigger context, overlapping them with genomic regions, which are known to be important for gene regulation. Apart from the direct interpretation, eQTL results have been used as a tool to help improve the understanding of results from genome-wide association studies (GWAS) by means of colocalisation. I performed a colocalisation analysis and as an example I could show how a GWAS variant mechanistically exerts its effect on plateletcrit - a blood cell index studied in a recent GWAS (Astle et al., 2016). Finally, I drew a link between gene-regulation, three-dimensional chromatin structure and gene constraint against coding loss of function mutations.

Complementing association analyses like eQTL and GWAS, recent advances in machine learning can be used to predict *in silico* the effects of genetic variation. To facilitate the

application of published machine learning models on custom data and to predict biological effects of genetic variants, I have developed, in collaboration with a fellow PhD student, Ziga Avsec (Prof. Julien Gagneur's group, TU Munich, Germany), the software Kipoi (www.kipoi.org).

The software has been designed to facilitate sharing and re-use of trained machine learning models in genomics. Together with Ziga Avsec I have conceptualised and implemented core elements of the platform. My major contribution in this software project was the implementation of tools and features for the effect estimation of DNA variants.

Table of contents

Chapter 1

Introduction

Studying the association between DNA variation and disease risk has become an important instrument to improve the understanding of the genetic architecture of common diseases. About a decade ago this has become possible in a systematic way, when several groups pioneered the application of genome wide association studies (GWAS) to search for risk variants in a hypothesis-free manner. By now, thousands of genetically independent variants have been associated with the risks of many of the common diseases and biomedically relevant quantitative traits. Approximately 90% of the GWAS lead variants (the variant with the lowest p-value of association) are localised in the non-coding space and ample evidence has been gathered that most of these variants exert their effect by modifying gene regulation. The remaining variants, which are localised in the coding portion of the genome alter the function of the encoded proteins by changing their amino acid sequence or by introducing nonsense codons leading to premature stops or nonsense mediated decay of RNA. One of the main challenges of the post-GWAS era is to link risk and trait associated GWAS variants that localise in the non-coding space to their target genes in order to define the effects they exert on these genes. This requires a rigorous approach to identify the tissues and cell types in which the associated variants modify gene function. Therefore highlighting how variants influence the behaviour of tissues and cells.

In this thesis I have studied the gene transcription regulatory effects of human genetic variation. To do so I have analysed datasets in which genetic variants as well as RNA obtained from purified cells of human peripheral blood were measured in hundreds of individuals. I have combined the genotyping and gene expression data from four different studies to increase the power to identify expression quantitative trait loci (eQTLs). In a next step I have integrated the association results with epigenome reference maps. In the studies for the final part of my thesis I made use of recent advances in machine learning to predict *in silico* the effects of genetic variation on gene expression. To facilitate the application of published machine learning models on custom data and to predict biological effects of genetic variants I developed, in collaboration with a fellow PhD student, Ziga Avsec, the software *Kipoi* (www.kipoi.org).

1.1 Gene expression and regulation

Eukaryotic cells consist of a nucleus, a membrane, and organelles, which are necessary to compartmentalise cell function, enable the cell to proliferate, and perform its systemic function [1, p9]. Within the nucleus of a cell resides the genomic DNA.

1.1.1 Primary sequence

The primary structure of genomic DNA is a linear sequence of 3.2 billion nucleotides, organised in 23 chromosomes. The alphabet of nucleotides consists in four bases: adenine, cytosine, guanine, and thymine, which are commonly abbreviated by their initial letter (A, C, T, G). Parts of the human genome are genes that encode the amino-acid sequence of proteins. To produce amino acid sequences, DNA is transcribed by the RNA-polymerase II protein into messenger RNA (mRNA) [2, p448], which is subsequently translated into polypeptides or proteins by the ribosomes. The regions of the DNA that are transcribed into mRNA define the coding space of the human genome [2].

The coding space

Proteins are encoded by genes which are localised as linear blocks of DNA sequence. Eukaryotic genes consist of several sections: A transcription start site (*TSS*), untranslated regions, exons and introns. The core promoter of a gene is where transcription is initiated and where RNA polymerase II attaches to the DNA [2, p448]. At the TSS, RNA polymerase II opens the double stranded DNA and initiates transcription. Not all the sequence between the TSS and the stop codon is translated - *intronic* sequences are excluded from the mRNA by a mechanism called *splicing*. Alteration of the canonical genomic splice sites may modify the residual mRNA open reading frame sequence and the resulting protein [3, 4]. As a consequence the functionality of the protein product may be affected. Only a subset of what is transcribed into mRNA is translated into proteins - this fraction of each gene is termed coding. Transcribed regions that are not coding are *untranslated* regions (UTRs), which are found at either end of the gene.

The non-coding space

There are about 20,000 known protein coding genes in humans, of which the coding part occupies about 2% (~64Mb) of the entire genomic space. The rest of the genome is known as the non-coding space [5] of which the function is slowly being elucidated. Attempts to stratify the vast non-coding space have so far led to the identification of biochemical functionality of up to 80% of the human genome in at least one cell type [6].

The non-coding space contains centromeric regions (about 2% of the human genome [7, 8]), non-coding genes, telomeres [9], scaffold attachment regions and regulatory regions [10]. These regions may be overlapping in genomic space or in function in the same way as exonic regions of one gene may act as expression regulatory regions for others [11]. Together, the regulatory annotation of the genome which is information that is biologically stored alongside the primary DNA sequence is termed the *epigenetic* profile.

To functionally characterise regulatory regions, the epigenetic profile is analysed in conjunction with the binding of transcription factors. These orchestrate transcription and therefore also control cell properties and cell type specific mechanisms.

Transcription factors Transcription factors (TFs) are proteins which are involved in the regulation of the expression of genes. They can act by facilitating or inhibiting the recruitment of RNA polymerase II [12]. There are an estimated 1,400 genes encoding TFs [13]. TFs bind directly to DNA and can recruit cofactors to form complexes for expression regulation [12]. Therefore, studying where TFs bind in the genome helps to identify regions of relevance for gene expression regulation.

Chromatin immuno-precipitation followed by sequencing (ChIP-seq) [14, 15] is a technique to query protein occupancy, such as TF binding and histone modification (see next paragraph), along the genome. ChIP entails the chemical cross-linking of DNA and proteins, followed by isolation and lysing of the cell nuclei. DNA is then fragmented and an antibody specific to the protein or protein modification of interest is used to precipitate cross-linked complexes. After reversal of cross-linking the DNA bound by the precipitated protein is isolated and made into a library that is then sequenced. A typical TF, like GATA1 binds at around 4,700 unique sites [16].

Accessible chromatin DNA is wrapped around complexes of proteins - called histones - forming a structure called *nucleosomes* [17–19]. A nucleosome consists of eight histones and takes up 147 bases of linear DNA. Histones have a globular shape except for their unstructured amino terminal ends ('tails') [20]. The amino acid chain of the tails can be chemically modified by e.g. methylases and acetylases and the presence of some of those modifications were found to indicate expression regulatory genomic regions (see paragraph: *Epigenetic profile*). Modifications of the histone tails are capable of orchestrating the unravelling of chromatin to control the accessibility of DNA [20]. Accessibility of the DNA sequence defines how well TFs can bind to a specific stretch of DNA *in vivo*. For a TF to bind it is therefore a prerequisite that the potential binding site (defined by nucleotide sequence) lies in accessible chromatin.

ATAC-seq (assay for transposase-accessible chromatin using sequencing) [21] is a technology to experimentally identify accessible chromatin regions, which are nucleosome-depleted. Briefly, the assay uses enzymes (transposases) that insert sequencing adapters into DNA; it does so preferentially in accessible chromatin. After PCR amplification only DNA fragments with adapters are sequenced, enabling the identification of accessible genomic regions.

Epigenetic profile As mentioned previously, histones can condense chromatin as well as they can make parts of it accessible. Selected modifications of the N-terminal tails of histones indicate regulatory relevant regions, known as enhancers, promoters, or repressors [20]. Histone tail modifications associated with enhancers and promoters are combinations of methylations and acetylations of lysins 4, 9, and 27 of histone 3, and a modified histone 2 [22]. The modifications of the amino acids of the N-terminal histone tails is dynamically performed by specialised enzymes, amongst which methyl-transferases and kinases are the most specific. Kouzarides assumes that therefore *methylation is the most characterized modification to date* [20].

Enhancers for example can be identified by a monomethylation of lysin 4 of histon 3 (H3K4me1), and if they are in an 'active' state if they also carry a acetylation of lysin 27 of histone 3 (H3K27ac) [23].

Enhancers Enhancers are genomic regions that have been shown to facilitate or increase expression of near-by as well as distant genes [24]. Therefore, the distant enhancer is brought into spatial proximity by establishing the required three dimensional structure of the chromatin [25]. This conformation may be established by the binding of structural proteins, as well as TFs and RNA fragments. The resulting loops, together with chromatin accessibility, influence gene expression by controlling the spatial density of regulatory regions, bound TFs and other activating co-factors that modulate recruitment of RNA polymerase II [24, 26].

Enhancers are marked by acetylation of Lysine 27 of Histone 3 (H3K27ac), together with monomethylation of Lysine 4 of Histone 3 (H3K4me1). Clustering of H3K27ac ChIP-seq read density has been shown to identify particularly expression-enhancing regions of the genome, termed superenhancers [27, 28]. In my thesis I will be using both enhancer and superenhancer regions to annotate eQTL variants.

1.1.2 3D chromatin conformation

Mechanistically, transcription factor binding is attributed an important role in gene expression regulation. Since this regulation acts through direct or indirect recruitment of RNA polymerase II, physical proximity between the regulatory genomic region that binds TFs and the promoter of the gene is one of the prerequisites for the control of gene expression. Enhancers can act over long distances and their mode of action has been explained by 3D chromatin structure, which brings linearly distant genomic regions into physical proximity [24].

There are many methods that try to elucidate the 3D structure of the human genome. Methods like microscopy and fluorescence in-situ hybridisation (*FISH*) can only measure few different genomic loci at low resolution [29]. Chromatin conformation capture [30] techniques use cross-linking of DNA and protein, as well as sequencing for the measurement of chromatin interactions. They therefore detect DNA-DNA interactions established by proteins.

In *HiC* (Fig. **1.1**) and *Promoter Cacpture HiC* (PCHiC) cross-linking of DNA and protein at their binding sites is performed to maintain the proximity between interactors. PCHiC relies on the technique of HiC (Fig. **1.1**), but enriches for sequence fragments overlapping gene promoters [31]. The enrichment is established by a hybridisation technique that was adapted from whole exome sequencing: HiC libraries are hybridised to RNA baits which had been designed for promoter regions and only the hybridised fragments are subsequently sequenced [31]. Using this technique every measured interaction starts at a promoter-associated fragment (*bait*) and ends at a promoter interacting region (PIR). PIRs can either be another promoter associated fragment (bait) or at any other fragment of the genome. In chapter section **4.4** I will explain how data from PCHiC experiments were used to map eQTL variants to PIRs.

1.2 Genetic variation

Diploid organisms, such as humans, have two sets of chromosomes, one is inherited from their mother and one from their father. In humans chromosomes 1 to 22 are termed autosomes, and every human has two copies of them. Therefore, variants in all autosomal chromosomes, can either be homozygous - two copies of one allele - or heterozygous - two different alleles

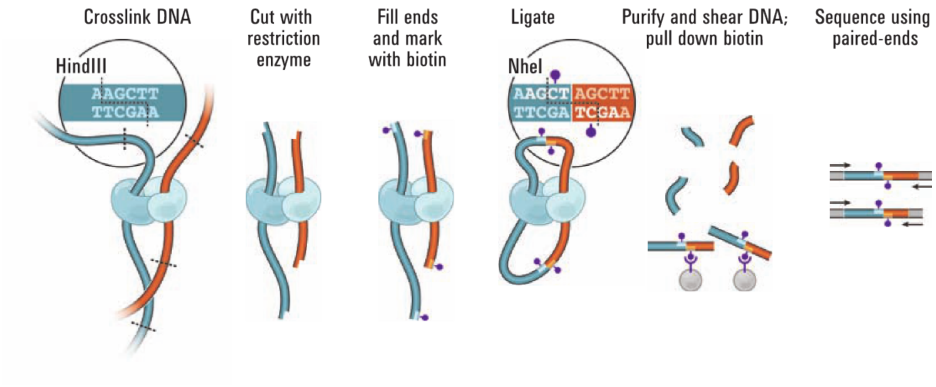

Crosslink DNA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
Purify and shear DNA; pull down biotin
Sequence using paired-ends
HindIII
AAGCTT
TTCGAA
NheI
AAGCT AGCTT
TTCGA TCGAA

Fig. 1.1 HiC is a chromatin conformation capture technique. After cross-linking, DNA is digested by a restriction enzyme (HindIII). Then ends of the two linked fragments are filled with biotin markers, which are then ligated. After shearing of the ligated fragments, fractions containing biotin are purified and paired-end sequenced. This way it is possible to map the junctions back to the human genome and estimate interaction frequencies between observed regions of the genome [30]. Image taken with permission from [32].

in one position. For bi-allelic variants, individuals in a population can therefore have one of three possible genotypes: homozygous major (two copies of the more common allele in the population), heterozygous (one copy of both allele), or homozygous minor (two copies of the less common allele in the population). This disambiguation is fundamental for association studies (section **1.2.1**, page **8**, and page **8**).

Every individual carries a unique combination of DNA variants. Most variants are inherited from ancestors, but new mutations occur randomly in every individual at an estimated rate of $0.5x10^{-9}bp^{-1}year^{-1}$ [33]. From their generation at random, new *alternative alleles* manifest over time at different genomic positions. Due to selection and other evolutionary effects some alleles become more frequent in a population. Measuring the frequency of occurrence of alternative alleles at every position of the genome gives the minor allele frequency (*MAF*) for every variant. MAF is therefore the frequency of the less common allele of a variant in a defined population. Genetic variants can have multiple alternative alleles and the detection of new alternative alleles depends on sample size of the study [34]. Most known variant positions in the human genome have only two alleles (are bi-allelic) [35].

Linkage disequilibrium (LD) is a manifestation of the genetic modes of inheritance. It measures whether two variants are inherited independently from each other or - if in linkage disequilibrium - two variants are inherited together more frequently than by chance due to their physical proximity on the chromosome. LD poses a central challenge in association studies as the similarity of inheritance of variants leads to correlation between alleles, complicating the identification of those variants which are the drivers of an observed phenotype. Most association tests therefore select one ’tag’ or ’lead’ variant for every cluster of highly correlated (high-LD) variants in order to simplify summarisation (more details see section **3.2.1**). Lead variants are not necessarily causal variants. Some prominent reasons for that are noise introduced by limited sample sizes, and the sparsity of genotyping data when genotyping arrays are used. An existing attempt to alleviate the limitation introduced by noise is bootstrapping which helps to find a set of plausible causal variants. Additionally, biological experiments such as functional assays may help to find causal variants, but also they have their own biases and caveats.

Coding and non-coding variants

Variants in the coding space - in exons of genes - have the potential to alter, truncate, or extend the amino acid sequence of the protein product. Therefore, non-synonymous (ns) coding variants - variants that cause changes in the amino acid sequence of a protein - have a potential to alter protein function and this may be contributing to the risk of disease. There are thousands of coding single nucleotide polymorphism (SNPs) in the genome of an individual and most of these do not alter protein function. Some variants may enhance or reduce function with common nsSNPs having subtle effects on protein function and rare variants may on occasions have profound effects on function. A prominent example of the latter are mutations in the BRCA1 gene, a locus in which some rare DNA variants are causal of an increased risk of breast cancer and ovarian cancer [36–38]. There are bioinformatics tools like *AnnoVar* [39] and *VeP* [40] to predict the functional consequences of coding nsSNPs. In this thesis a machine learning method has been implemented to predict the effect of non-coding variants on gene transcription.

Non-coding variants may lie inside genes, in introns or UTRs, or between genes. The prediction of their effects is inherently more complicated, compared to coding variants. The current theory suggests that a portion of the non-coding variants act via regulation of expression levels of genes by modifying the binding affinity of TFs, or inhibiting the correct chromatin conformation [41, 42]. These modes of action may affect the expression of multiple genes and pathways, but the links between a variant and the affected gene cannot trivially be established (see section **1.2.3**). As already reviewed, 90% of association variants identified by GWAS are non-coding [43]. Hence it is important to generate new experimental data and analysis methods to link the associated variants with high confidence to their target genes so that the GWAS results can be interpreted in the context of proteins and molecular pathways which may be amenable to therapeutic interventions.

1.2.1 Association studies

Association studies identify genetic variants that alter phenotypes by using sample cohorts. Measured phenotypes and genotypes from all individuals are therefore compared to find statistically significant effects. Associations are sought where the different alleles of a variant explain alterations in the phenotype measurements.

Genome wide association studies

The key aim of disease association studies is to identify genetic variants conferring risk of ill-health. Such studies are typically performed by using genotyping data from cohorts of individuals with and without disease. Until the advent of genome wide association studies (GWAS), the main approach for association studies for common diseases were based on a single candidate gene or often even a single DNA variant. These candidate genes had either been identified by studies of rodent models of human diseases or inferred as a putative candidate gene based on the results of cell biology and biochemical cell signaling studies.

Just over a decade ago the first large scale GWAS for six common diseases was performed [44]. The positive results of this and other early GWAS initiatives spurred on similar studies for many other common diseases and also for clinically relevant disease-related traits (e.g. lipid levels, blood pressure, kidney function parameters, blood cell indices). These initial studies were followed by international efforts to bring data together for the purpose of meta-analyses. These meta-analysis studies have now been replaced by GWAS in large prospective cohorts like the UK Biobank [45–49]. Thousands of genetically independent variants have by now been associated with the risk of common diseases [50] and the most recent GWAS for blood cell indices catalogued nearly 3,000 lead variants for 36 properties of red blood cells, platelets and leukocytes [51–55]. One of the major challenges for biology and medicine is the interpretation of GWAS results into mechanistic explanations of how variants exert their effects on the risk of diseases or the measured traits. An overview of some of those approaches can be found in section **1.2.3**.

Expression quantitative trait loci

Expression quantitative trait locus (eQTL) studies associate gene expression levels with genetic variation [56]. More precisely, transcript levels of genes as well as genotypes are measured for all participants in an eQTL study. For the association test the genotype is represented as non-reference allele count: in diploid organisms like humans the non-reference allele count can has values from 0 to 2 for every possible variant in the autosomal chromosomes.

One of the most common methods to perform the eQTL association test is by testing for a linear dependency between the non-reference allele count and the transcript level of the tested gene using data from all measured individuals [57–60]. The expression measurements can be confounded by relatedness, age, sex and environmental conditions. Given that eQTL studies require cohort sizes of at least hundreds of individuals (section **1.2.2**) it is not feasible to match samples for these co-factors and so they are usually accounted for as covariates in the linear model.

Cis- and trans- effects

Genetic effects can be classified into *cis*- and *trans*-effects. In eQTL studies cis-effects are seen as direct effects of genetic variation on gene expression and trans-effects as indirect ones. Conventionally, cis-effects are tested by selecting a genomic region (e.g. 1 Mb) (*cis-window*) around the gene body and testing for associations between variants in the defined region and the gene transcript level. The assumption that variants in the proximity of the gene promoter exert direct effects on gene expression is based on the working hypothesis that those variants affect the binding of transcription factors, which act in their physical proximity. Trans-effects are in practice identified as associations in which the variant lies outside the cis-window. With indirect effects the expression of the affected gene is altered by a different mechanism (e.g. biological pathway), which is associated, or at the least correlated with the alleles of the eQTL variant. This technique of the identification and disambiguation of cis- and trans-eQTLs neglects the long-range interaction of transcription factors with genes.

Attempts have been made to incorporate additional data in eQTL analyses to improve variant prioritisation and characterisation [61, 62]. Neither of those tackle the issue of distinguishing direct associations from indirect ones. Cis-effects may be identified by allele specific expression analyses [63], where genotype phasing is used to identify which allele causes increased or reduced expression in every individual. Alternatively, biological perturbation experiments can be used to identify direct effects.

1.2.2 Minimum sample size for association studies

In order to detect associations between traits and common genetic variation (minor allele frequency (MAF) > 0.05) samples from a minimum number genetically homogenous individuals have to be tested. The sample sizes necessary for association studies depend on the strength of the effects of the variants on expression levels and the frequency of the tested variants in the population [64–66]. In eQTL studies it is possible to use fewer samples, because traits are better defined by gene expression rather than systemic phenotypes or disease versus non-disease status. Furthermore, observed effect sizes in eQTLs are generally higher than in GWAS [64]. In eQTL studies it is therefore possible to use sample sizes of 200 individuals, depending also on the number of co-regulated genes and the number of variants affecting gene expression [67].

1.2.3 Colocalisation

Results from GWAS reveal associations of thousands of variants with traits and common diseases [50]. More than 90% of them are localised in the non-coding part of the genome [43], making it far more challenging to identify the mechanistic link causing an association compared to causal coding variants which alter the amino acid sequence of a protein.

For a better biological understanding of these associations in the non-coding space, variants are linked to genes by various approaches: the simplest approach is by annotating the associated variant with the nearest gene [68]. This carries a high risk of erroneous assignments, as illustrated by Smemo et al. and Claussnitzer et al. for the FTO locus [69, 70]. Here, a body-weight associated variant, rs1421085, localises in an enhancer element that lies in the first intron of the FTO gene, but was found to regulate expression of the gene IRX3, located at a distance of 0.5Mb from the associated variant [69]. Before this long-range association was found, studies had been conducted trying to define the molecular mechanism by which FTO explains the GWAS trait, until Smemo et al. [69] showed in mouse models that the transcription factor Irx3 is the true mediator of the body-weight effect. These findings were later replicated in human primary adipocytes using CRISPR-Cas9 [70].

In agreement with the findings for FTO, approaches that leverage pathway analyses revealed that genes affected by regulatory variants are often up to 2 Mb away from each other [71]. Alternatively, information about three dimensional (3D) chromatin conformation has been used to link variants to genes [28], however conformation and physical interaction between distant DNA element does not necessarily convey function. Therefore, results from eQTL studies have been used to annotate variants with associated genes, based on empirical evidence at cell type resolution. This approach holds the promise to highlight the genes which are controlled by GWAS variants and therefore affect the studied disease or trait.

Another approach to functionally link non-coding GWAS variants to their target gene is by using statistical methods to determine putative colocalisation with eQTL variants. This approach has been successfully applied to link GWAS variants to genes as well as tissues and cell types for diseases like schizophrenia, immune-mediated diseases, including type 1 diabetes and inflammatory bowel disease and also for quantitative traits such as height and body mass index [72, 61, 73–80].

In my thesis I have bioinformatically processed and analysed four eQTL studies, which I then colocalised with a GWAS that associated 36 blood cell indices of 173,480 European individuals from the UK Biobank and INTERVAL cohorts [45] (section **4.3**).

1.3 The haematopoietic system

The haematopoietic system is a paradigm for understanding the biology of stem cells, cell proliferation, and diseases [81]. Blood was the first tissue in which a severe inherited disorder named sickle cell anaemia, could be explained at the molecular level [82]. Furthermore, haematopoietic stem cells are the most studied and best characterised tissue-specific stem cells [83]. Owing to their accessibility they are clinically highly relevant since the most common transplantation in medicine is the transplantation of stem cells obtained from peripheral blood or the bone marrow [84]. Together, this shows the importance of blood in biology and genomic research as well as for many diseases, including the plethora of immune disorders.

The generation of all blood cells, haematopoiesis, starts from a multipotent haematopoietic stem cell. The stem cells are rare and can mainly be found in the bone marrow at a rate of one stem cell in $3x10^6$ cells [85]. They have the potential to differentiate and to self-renew at the point of cell division in order to maintain cell population size. The differentiation process is summarised in the heamatopoietic tree (Fig. **1.2**), which illustrates the different branch points at which omnipotent progenitor cells commit to the myeloid and lymphoid lineages [81, 86]. The different states of the blood cell progenitors have been mainly defined by the use of monoclonal antibodies against cell surface markers. Most of the antibodies used for characterisation are against Cluster of Differentiation (CD) antigens. When purifying the different blood cells by cell sorting monoclonal antibodies against the CD antigens are used to select the desired fraction of cells. Depending on the properties of the respective cells positive or negative selection can yield desirable levels of purity and intactness of cells. For monocytes for example positive CD14 selection is used commonly to achieve high levels of purity [87–89, 79, 74]. The same is true for CD4 and CD8 T lymphocytes, CD19 B lymphocytes, and also for neutrophils positive selection by CD15 is a recommended purification procedure [90]. Platelets can be purified from platelet rich plasma by depleting for leukocytes using negative selection with CD45 microbeads [91]. During my PhD thesis I worked with data generated from these cell types following the cell type purification methods mentioned above.

The haematopoietic tree displays the intermediate stages that blood progenitor cells undergo until they have reached the final stages of precursor cells which generate the different types of blood cells present in the peripheral blood. The first step in the process of stem cell differentiation involves the loss of the potential to self-renew, which is followed by the gain of lineage commitment necessary to fulfil the tasks of the fully lineage committed precursor cells.

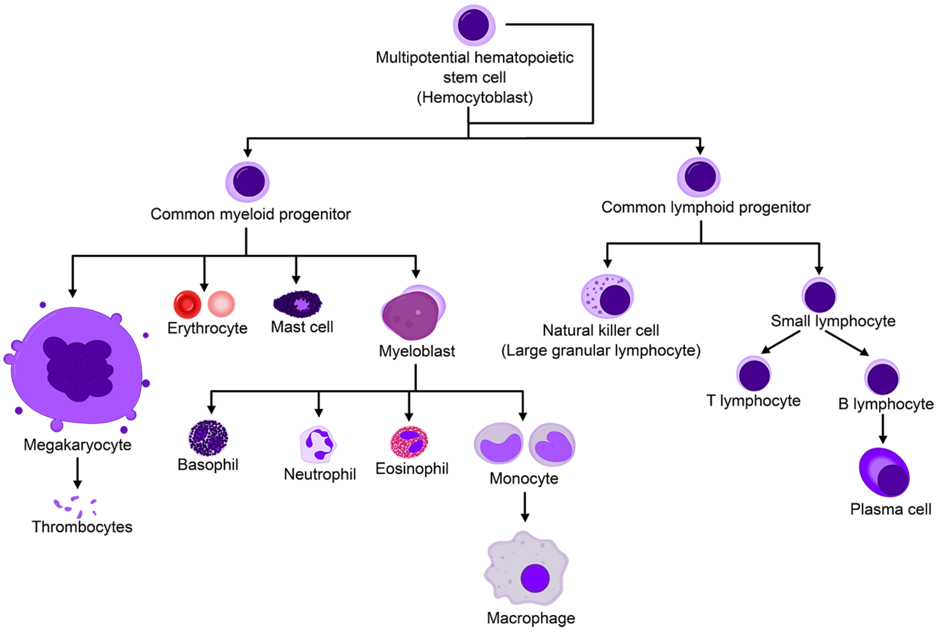

Multipotential hematopoietic stem cell (Hemocytoblast)
Common myeloid progenitor
Common lymphoid progenitor
Megakaryocyte
Erythrocyte
Mast cell
Myeloblast
Natural killer cell (Large granular lymphocyte)
Small lymphocyte
T lymphocyte
B lymphocyte
Basophil
Neutrophil
Eosinophil
Monocyte
Plasma cell
Thrombocytes
Macrophage

Fig. 1.2 The haematopoietic tree. Image taken from [92].

Overall, a healthy adult produces approximately 10^{11}-10^{12} new blood cells per day [93]. This output can be split along the different lineages of differentiation (Fig. **1.2**): neutrophils and monocytes belong to the myeloid branch [94, p3], which means they share common progenitor cells with erythroblasts and megakaryocytes, the bone-marrow-residing precursors of red cells and platelets, respectively [94, p111]. Lymphocytes belong to lymphoid branch and the myeloid and lymphoid branches bifurcate at the level of myeloid-lymphoid progenitor cells. The function of lymphocytes and myelocytes is substantially different: lymphocytes form the adaptive immune system and can learn to identify new pathogens by establishing a pool of T- and B-lymphocytes with memory of previous pathogen encounters. Cells from the myeloid branch are essential for oxygen and carbon dioxide transport (red cells), blood clotting (platelets), innate immunity (neutrophils, eosinophils, basophils) and phagocytosis (neutrophils and different types of monocyte-derived macrophages).

1.4 Machine learning models

In the course of my PhD studies I have used machine learning models for genomics. Together with a fellow PhD student, Ziga Avsec from the Gagneur group at the Technical University of Munich, Germany, we have identified that missing concepts for sharing and re-use of published models are a major hindrance for enhancement of existing and the development of new models. We have therefore developed a software platform that facilitates access to and use of complex machine learning models in the hope to foster research on the regulation of gene trasncription.

Machine learning is an area of artificial intelligence in which developers aim to build models that can describe and predict observed data. Many approaches have been made, including decision trees, nearest neighbour predictions, Bayesian models, and also neural networks. Generally, models can be categorised into supervised and unsupervised machine learning models.

Supervised machine learning models are used for problems where labelled data is available. In such cases for every input sample there is one class label, value, or other measure available that describes the input sample. The model is then designed and trained to produce the most accurate label predictions for given inputs [95].

Unsupervised machine learning models are used when no labels are available for model input samples. Clustering algorithms are an example of solutions for such problems. In these cases, the machine learning algorithm is used to identify structure in the data without further information but the input samples themselves [95].

In my thesis I will only be using supervised machine learning models as the presented problems involve labelled data.

1.4.1 Neural networks

The terms artificial neural network, neural network, and deep learning have been used to describe a kind of machine learning model that was designed in resemblance to the basic structure and function of the human brain [96].

The basic nerve cell - the neuron - has a soma and an axon. For a model of the brain the main functionally relevant property of the neuron is its activation. Neuron activation depends on the input it receives from synapses - junctions between its soma and axons of other neurons. The activation of a neuron depends on the sum of its inputs and on an internal threshold which is inherent to the cell. The activation (output) of the neuron is transmitted in the axon. From the early understanding of how neurons work and interact mathematical models of the networks of neurons have been created as early as 1943. McCulloch et al. [97] formulated an idea of how learning could work in a neural network. They described the possibility that information could be encoded by different activation states of the neurons in the network.

After this atomic model of biological neural networks, Rosenblatt, in 1957, developed a conceptual idea of how capabilities of how the human brain may be replicated by a machine - the *perceptron* [98]. The idea was published in a report to the Cornell Aeronautical Laboratory, Inc. and captures important ideas, concepts, and caveats of modern artificial neural networks. The perceptron was a concept that Rosenblatt envisioned to be implemented for example as an electromechanical device. It would react on the *'phenomenological world'* recognising patterns and objects irrespective of the angle or aspect of their appearance and it would identify patterns directly rather than searching through a pre-defined set of recognisable patterns stored in memory. The perceptron should recognise similarities or identities across optical, electrical, tonal, etc. inputs, in close resemblance to the brain.

Interestingly, Rosenblatt defined the (photo) perceptron as a black box with a video camera at the input and printer or signal lights on the output. The perceptron would have to learn to give the same output for images that belong to the same abstract class - for example three dimensional objects displayed from different angles. Rosenblatt defined learning as presenting random input samples that belong to the different classes and 'forcing' the corresponding correct output. Predictions should then resemble the learned associations. Further on, he predicted that outputs of such a model would have a probability of being wrong, which would not imply malfunctioning, but would be caused by using his statistical approach of learning. He postulates that this probability of false predictions could be reduced with additional training examples. All of this is conceptually true for state of the art machine learning models.

The perceptron (black box) would have a layered and therefore hierarchical structure with a sensory unit (now called: *input layer*), an association system (now called: *hidden layer[s]*), and a response system (now called: *output layer*). The sensory unit would be a raster (now called: *input data matrix*) and the association system would consist of multiple units. There would be one or more connections between all elements of the input raster and the units of the association system. Every unit of the association system (hidden layer) would calculate a weighted sum over its connected inputs, where the weights may be positive or negative. Additionally, the association system would have a threshold value to which the weighted sum would be compared to produce an output. The output of the association system would be

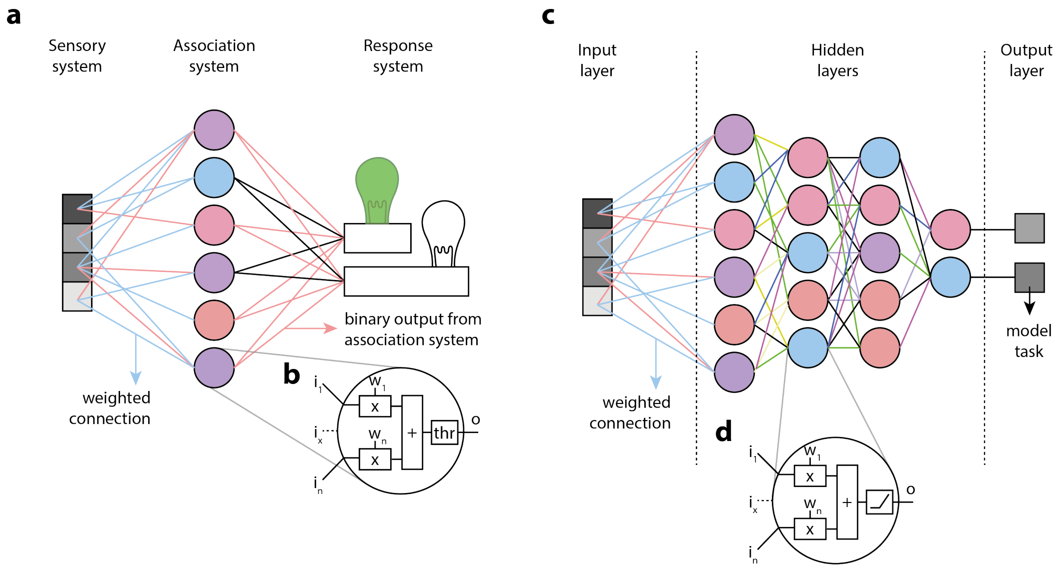


Fig. 1.3 Schema of a neural network. Fig. **1.3a** shows the basic components of a perceptron as defined by Rosenblatt [98]. The individual units of the association system are depicted as coloured circles. Fig. **1.3b** shows the schematic of one unit of the association system: it accepts multiple inputs, multiplies every input with a factor (*weight*) that is specific to every input and then the sum of all those values is calculated. Finally the weighted sum is compared to a threshold value and a binary output is returned. Fig. **1.3c** the basic schematic of a neural network, with a symbolic number of hidden layers and connections. Model output for every model task can be any numerical value. Fig. **1.3d** the basic neuron is identical to Rosenblatt's idea, except that the threshold operation is replaced by an activation function, which in most cases is non-linear.

linked to the response system. The response system would have few units (e.g. signal lights), which were triggered based on the outputs of the association system.

Setting technical details and the exact setup of the different layers aside, Rosenblatt's perceptron is a remarkable summary of the basic concepts and the limitations of neural network models. In Fig. **1.3** I show a visualisation of the basic elements suggested by Rosenblatt.

Based on these foundations, empirical experiments, and theory, the current neural network models have evolved. They use the basic concept of a layered approach as suggested by Rosenblatt. An input layer, which is a numerical matrix representation of the input data is connected to one or a stack of *hidden* layers. Finally, an output layer, which in most cases has less units than the input layer, is connected to the last hidden layer (Fig. **1.3**). Networks that have many of those hidden layers are known as 'deep neural networks', their use has coined the term 'deep learning'. The computational units of hidden layers are artificial neurons. An artificial neuron is designed in analogy to the biological neuron: it calculates a weighted sum over the input values, where every weight for each input is a constant property of the neuron; the weighted sum is then converted to the output by a non-linear activation function. The activation function is the generalisation of thresholding that produces binary output (suggested for the perceptron). The architecture of a neural network defines the connections of every neuron of one layer to the outputs of the previous layer. Different basic architectures will be presented further on.

Training

The weights of all neurons of a network contain all the information that has been learned from training data. An untrained neural network is initialised with random weights. During training, weights are continuously updated minimising a *loss function*. The loss function can be interpreted to measure the distance between the current state of the model and the ideal state of the model in an arbitrary scale. In most supervised settings the loss function measures the distance between output values calculated (predicted) by the current model state and the desired model output (model *task*) using training examples. The calculated loss is then used to update the model weights. This is most commonly done by stochastic gradient descent, where the most influential weights for a prediction are modified the most, taking the calculated loss into account. Training is an iterative process, which involves prediction followed by weight updates. It is finished when the training objective is reached, for example if the loss cannot be reduced any further within a certain number of training iterations.

The state of the final trained model is therefore a function of the initial weight values prior to training, the model architecture, the training examples and their order, the loss function, the definition of the condition that defines when training is complete and other factors. This indicates that practically it is almost impossible to replicate the exact same model by re-training a model with the same initial weights.

Architectures

Neural network architecture, as well as their loss function, are parameters determining how well machine learning models can capture and describe a given dataset. There are three basic and popular layer types that are important design blocks for neural network architecture (Fig. **1.4**):

Fully connected layer In the fully connected layer all inputs of all neurons are connected to all outputs from the previous layer. Additionally, all weights of all layers are independent from each other and have to be learned individually. This type of layer is ’expensive’ as it

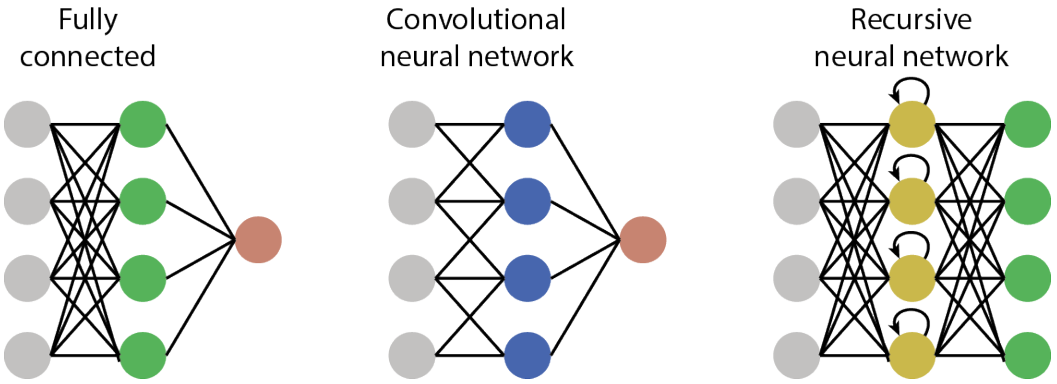


Fig. 1.4 Schematic depiction of three basic neural network architectures. Grey dots indicate elements of the input layer, green dots are neurons of fully connected layers, yellow dots are the more complicated neurons of a recurrent layer, red dots are the model output.

requires the most memory and also because it has the most parameters to learn. The more parameters that have to be learned from data, the longer the model has to be trained and the more training examples are necessary. Fully connected layers can usually be found towards the output of the model, where information from all parts of the layer input are relevant and where input data has been condensed and compressed by previous layers.

Convolutional layer In convolutional layers every neuron only has a limited number of inputs and 'slides' along the input dimensions. Therefore the number of weights is drastically reduced, which reduces training time, memory efficiency and necessity of input data. Convolutional layers perform a task that in the field of image segmentation is known as *edge detection*. These layers should be used where data-points that lie close together are similar, or where those neighbouring data-points can be summarised as a higher-order structure or pattern. Therefore first layers of deep learning models for genomics are mostly convolutional layers as they are capable to identify patterns, such as DNA motifs.

Recursive layer In recursive layers the elements are complex neurons, which don't only depend on the current input, but also on their previous activation state. They therefore have memory, so that contextual information can be stored within each neuron. This property makes them especially interesting for sequential data. They are popular in language modelling, handwriting recognition, speech recognition, etc. [99]

Convolutional neural network Convolutional neural networks (*CNN*s) are networks that contain convolutional layers and may also contain fully connected layers. CNNs are feedforward networks, which means that no memory components are available and the output only depends on one input sample.

Recursive neural network Recursive neural networks (RNNs) contain at least one recursive layer and may contain fully connected layers. RNNs retain information, therefore the output depends on the current, and previous input samples, and their order. RNNs and CNNs can be combined into convolutional recurrent neural networks (CRNNs) by the stacking layers accordingly.

1.4.2 Software frameworks for machine learning

Machine learning applies statistical models to complex datasets. Often, the best fitting algorithm has to be found for a dataset, making it necessary to train multiple models of different kinds. The algorithms of the models remain mostly the same, irrespective of the dataset. Therefore software packages - frameworks - have been developed to help train and apply the models to data. Apart from facilitating the application, frameworks simplify sharing of trained models. Additionally, frameworks usually implement many different kinds of machine learning algorithms, which facilitates the comparison of models trained on one dataset.

Neural networks are computationally very expensive during training and prediction, because for every neuron in the network multiple multiplications and additions have to be performed. This task can be parallelised very well, which lead to using graphical processing units (GPUs) rather than the central processing unit (CPU) of a computer. GPUs are similar to CPUs, but driven by the matrix operations necessary for computer graphics, they are highly optimised to perform multiplication and addition in a highly efficient and parallel way. The use of GPUs instead of CPUs for efficient calculations has enabled the recent successes of deep learning. This comes at the price of having to use more technically challenging code in order to access and use the power of GPUs. Driven by the success, deep learning has become heavily used in many fields and industries and therefore also many different software platforms have been developed for deep learning. The setup of these frameworks is intrinsically more difficult than the setup of other machine learning frameworks as access to GPU hardware needs to be handled by specialised software libraries. Software installation managers like *conda* and *brew* help with the correct setup of those deep learning frameworks as they enable efficient installation of software with all its dependencies even for computer users without administrative privileges.

Deep learning frameworks bring building blocks like different kinds of hidden layers in a ready-to-use format and offer well-tested implementations of training algorithms. This enables the developer to focus on the design of the network architecture rather than the correct implementation.

Chapter 2

Methods

All methods that were used are elaborated in the results chapters. To avoid duplication this chapter serves as an overview over and as an index of the methods. It provides references to the individual sections of the results that describe methods and offers an alternative approach to find sections of the text independently from the main narrative.

2.1 eQTL analysis

2.1.1 Datasets

Expression quantitative trait locus (eQTL) datasets of purified human blood cells were acquired from collaborators (section **3.1**, page **25**). At the time when the data was shared two datasets were still unpublished. The eQTL dataset from the CEDAR project was kindly shared prior to the publication [79]. The BLUEPRINT Consortium platelet eQTL dataset was shared and is not yet published.

2.1.2 Data preprocessing

In order to facilitate the joint use of the eQTL datasets from the four studies a unified preprocessing pipeline was designed (section **3.2.1**, page **28**) and validated (section **3.2.1**, page **32**). The preprocessing pipeline also included the correction of the bead array probe annotation and estimation of minor allele frequencies and Hardy-Weinberg equilibrium. The aim of the unified preprocessing was to reduce technical noise and batch effects caused by study-specific preprocessing pipelines.

2.1.3 eQTL analysis

After preprocessing cis-eQTL analyses were performed (section **3.2.1**, page **30**) i) for the individual datasets and ii) after aggregating datasets across studies by cell type to maximise the number of samples.

As an alternative to the default cis-eQTL analyses, a way to use three dimensional interaction data (Promoter capture HIC) as a prior for eQTL testing was developed (section **4.4.2**, page **64**).

2.1.4 LD clumping to compare associations

Results of association test are often summarised by the lead variant, but depending on linkage disequilibrium (LD) between associated variants it may happen that lead variants found in different studies are part of the same association signal (explanation: section **3.2.1**, page **32**). To compare association signals across studies or across cell types variants were clumped based on pairwise LD (section **4.1**, page **48**).

2.1.5 Replication of association signals

In order to assess whether eQTL associations were present in a replication cohort the LD-clumping approach was extended to sub-threshold associations. For these the discovery-signal was genome wide significant, but the replication signal was only required to pass a lenient threshold of $p < 0.05$ (section **3.2.4**, page **40**).

2.1.6 Annotation of eQTL variants

eQTL results were annotated with different genetic features. An overlap was assumed if the variant lied within a genomic region that was tagged as:

- ATACseq peaks (section **4.2**, page **52**)
- H3K27ac peaks (section **4.3.1**, page **57**)

To assess properties of low-frequency eQTL associations, eQTL variants were annotated with UK10K [100] minor allele frequency data (section **3.2.3**, page **38**). Finally it was tested whether the effect size of eQTLs was related to the gene constraint score (pLI) (section **4.5**, page **75**).

2.1.7 Colocalisation with GWAS

Colocalisation - statistically linking GWAS variants to corresponding eQTLs associations - was performed using the Astle et al. study [45] (section **4.3**, page **54**).

2.1.8 Integration of PCHiC data

Preprocessing of PCHiC

PCHiC interaction data had to be re-annotated prior to analysis (section **4.4.1**, page **64**).

Enrichment of eQTLs in PCHiC PIRs

To assess the gene expression regulatory relevance of PCHiC the enrichment of eQTL associations in promoter interacting regions (PIRs) was assessed (section **4.4.3**, page **69**).

PCHiC promoter connectivity and transcript level variation

It was assessed whether the number of PCHiC connections of a promoter was related to transcript level variation (section **4.4.3**, page **69**). The calculation included a normalisation of transcript level variation to reduce mean expression level bias.

2.2 Genetic variant effect prediction using deep learning

2.2.1 Implementation of the software environment

A software platform for sharing trained machine learning models was developed to facilitate comparison and re-use of published models (section **5.2.1**, page **87**).

2.2.2 Variant effect prediction

Within the above framework algorithms for the prediction of variant effects based on model predictions were implemented (section **5.3**, page **92**).

2.2.3 Model and data visualisation

To facilitate the interpretability of model predictions and of variant effect predictions visualisation tools were integrated in the software platform (section **5.4**, page **98**).

Chapter 3

eQTL datasets

In the last decade high-throughput measurements of genotype and molecular traits have enabled association studies involving hundreds to thousands of individuals [88, 89, 87, 79, 101, 102, 74, 78]. One of these kinds of analyses identifies genetic variants that modulate gene expression - expression quantitative trait loci (eQTLs). eQTL studies identify genetically explainable gene expression variation in a tissue and potentially cell type specific manner [87, 77]. The results of these studies can be used to interpret and explain results from other studies such as genome wide association studies (GWAS) [103–105, 72, 106–108, 73, 109, 75].

In this chapter, I will discuss how I have reprocessed and integrated eQTL datasets of seven purified blood cell types that were generated by collaborators in different laboratories. I will show that the increased sample size enables the identification of additional associations, and I will compare my cell type specific eQTL results to results from a whole-blood eQTL study.

3.1 Available eQTL datasets

The expression datasets used in this thesis are all based on measurements made by the expression Bead Chip technology from the same vendor, namely Illumina. All but one of the studies used the same Chip version (Illumina HumanHT-12 v4.0).

	Expression	Genotypes
Cardiogenics	Illumina Human-Ref-8 v3 BeadChip	Human 610 Quad Custom array and Sentrix Human Custom 1.2M
CEDAR	Illumina HumanHT-12 v4.0 Expression BeadChip	Illumina Human OmniExpress 12 v1.0 BeadChip
WTCHG	Illumina HumanHT-12 v4.0 Expression BeadChip	Illumina Human OmniExpress 12 v1.0 BeadChip
BLUEPRINT (PLT)	Illumina HumanHT-12 v4.0 Expression BeadChip	Whole genome sequencing

Table 3.1 Available microarray-based eQTL datasets.

In terms of probe content the *Illumina Human-Ref-8 v3 BeadChip* is half of the *Illumina HumanHT-12 v4.0*, where 96% of probes of the former are also present on the latter. The original results from experiments from different laboratories were either retrieved from ArrayExpress [110] or from the laboratory which had performed the experiment. The details of the experiments are reviewed briefly.

Wellcome Trust Centre for Human Genetics (Julian Knight, University of Oxford, UK; WTCHG) Data was obtained from ArrayExpress [110] in raw and preprocessed format. Cell types included neutrophils [111], total B-cells [87], and monocytes [112]. Expression results using RNA from four states of monocytes were generated; non-activated or naive, 2 hours after activation with lipopolysaccharide (LPS), 24 hours after activation with LPS or with γ interferon. The corresponding genotype data of the same individuals was obtained from the EGA [113]. For the work reported in my thesis I have not used expression data from activated monocytes.

The CEDAR project (Michel Georges, Université de Liège, Belgium) For the CEDAR project expression data was generated using the RNA samples from six blood cell types (neutrophils, monocytes, CD4+ T lymphocytes, CD8+ T lymphocytes, CD19+ B lymphocytes and platelets) obtained from 300 healthy individuals. The preprocessed expression data was made available (the preprocessing was a joint effort of Julia Dmitrieva from the Georges group and myself) together with the corresponding genotyping data.

The Cardiogenics Consortium eQTL project (Nilesh Samani/Willem H Ouwehand, University of Leicester/University of Cambridge) The Cardiogenics consortium generated expression using RNA samples from monocytes and macrophages obtained by culturing of monocytes in the presence of DMEM for a period of up to 5 days [88, 89]. The consortium processed expression and genotyping data from 395 healthy individuals from the Cambridge NIHR BioResource [114] and 363 individuals with a previous diagnosis of coronary artery disease. Participants of the latter group were enrolled at hospitals in Leicester, Lubeck, Munich and Paris. Raw and preprocessed data were retrieved from a study archive at the University of Cambridge.

The BLUEPRINT Consortium eQTL project (Kate Downes, University of Cambridge) The BLUEPRINT consortium generated an eQTL dataset for monocytes, neutrophils and CD4 T-lymphocytes from 200 healthy individuals from the Cambridge NIHR BioResource for which RNA samples were analysed by RNA-seq [115, 74]. The same donations of blood used for the isolation of leukocytes were also used to recover platelets. RNA was obtained from the platelets using a validated protocol and applied to Illumina bead chip arrays to measure transcript levels. Since the integration of RNA-seq in the analyses in this thesis would introduce technical noise only the platelet dataset was used. Genotyping results were obtained by whole genome sequencing and were made available by the group of Nicole Soranzo at Wellcome Sanger Institute, Cambridge.

	Cardiogenics	CEDAR	WTCHG	BLUEPRINT (PLT)
Monocyte	758	300	432	
Monocyte + 2h LPS			432	
Monocyte + 24h LPS			432	
Monocyte + 24h IFN			432	
Macrophage M0	614			
Granulocyte Neutrophil		300	101	
Total CD4+ T-lymphocyte		300		
Total CD8+ T-lymphocyte		300		
Total CD19+ B-lymphocyte		300	283	
Platelet		268		156

Table 3.2 Microarray-based eQTL datasets used in this thesis. Number of individuals for whom genotyping and expression data were available are presented.

3.2 Merged analysis of datasets

Increased sample size in eQTL studies increases the number of identifiable associations in common variants and enables the identification of associations with low frequency variants. A typical approach to increase sample size is by meta analysis in which summary data from individual studies are combined. Alternatively, the raw input data can be reprocessed for a single joint analysis. A joint analysis, although requiring far more analysis time, preserves more power to detect associations compared to the meta-analysis approach [101]. Having access to the source datasets for the four eQTL studies (data generated on the same platform) provided a unique opportunity to design an analysis pipeline for the unified preprocessing of all available expression datasets.

3.2.1 Preprocessing expression data

Starting from the raw bead chip signal intensity data, I designed a preprocessing pipeline to fit all needs of individual datasets. The preprocessing was done in R version 3.2.1, the procedure was based on recommendations in the R package *lumi* [116], an approach that was previously also taken by Fairfax et al. [112]. Most of the following analysis steps were performed using the lumi package allowing the processing of the raw data in the same and consistent way:

- Log2 transformation
- Outlier detection based on distance from sample center. Threshold is scaled by median distance from center. (`detectOutliers` in lumi package)
- Removal of duplicate samples, by selecting replicates with highest median probe intensity
- Robust spline normalisation (RSN) within the batches, datasets and cell types. RSN from the lumi package was used.
- Batch correction with ComBat [117] within datasets and cell types. ComBat from the sva package was used (for details, see below).
- Only keep probes with significant expression p-value (Th: $p<0.01$) above background level in at least 2.5% of all samples.
- Calculate PEER (version 1.0) covariates using 10 factors and mean as to be added to the model during calculation. [118]

Key steps of the analysis pipeline will be dicussed in greater detail because the data processing included quality control measures, including the detection and removal of outliers. Principal component analysis (PCA) was used to identify samples with evidence for being an outlier in the processed and reprocessed data. Next, data was normalised using RSN. Several other normalisations were tested, such as simple scaling normalisation and quantile normalisation. Their performance was compared by reviewing expression level distributions and PCA across samples - RSN resulted in the most homogeneous outputs (data not shown). Batch correction was performed using ComBat [117] for datasets in which microarray expression data was processed in batches. Data of the four studies was produced in batches for at least one of the cell types. ComBat was executed with default parameters and without using covariates. After batch correction the residual batch effect was reviewed using PCA and expression value distribution. Finally, 10 covariates were estimated for every dataset independently using PEER. These were subsequently used as covariates in the association tests.

After successful preprocessing of the expression data, the expression bead chip probe annotation was corrected. The used chip contains measurement nucleotide probes of the length of 50 bp and every probe has a unique identifier assigned and tags a transcript of a human gene. The probe identifiers are assigned to gene names by the manufacturer, but the annotation delivered by Illumina was shown to be inaccurate [119]. Therefore probes were annotated with R package ReAnnotator [119]. ReAnnotator maps the probe sequences to an mRNA reference database and consequently assigns probes to transcripts. Apart from annotating probes, ReAnnotator also produces a mapping quality as output. I removed the annotation of 10,002 probes with a bad mapping quality. The probe sequences may correspond to RNA of a genomic region harbouring a common variant (MAF > 0.05 in 1000 Genomes). Transcripts containing the same allele as the probe sequence could then preferentially bind to that probe, thereby distorting the association signal of the given probe. Therefore 2,748 probes overlapping common variants were excluded from further analyses by removing their annotation.

Genotyping data

For three of the four datasets, Illumina genotyping microarrays were used to measure the genotypes of the DNA samples. These arrays were designed to query the genotype of a sample at multiple positions of the genome (Illumina Human OmniExpress 12 v1.0 BeadChip: 733,202 SNPs; Human 610 Quad Custom array and Sentrix Human Custom 1.2M: 1,115,839 SNPs and 80,128 CNVs). Based on linkage disequilibrium (LD) in the population, the density of these measurements can then be computationally increased (imputed) giving a more complete set of the genotypes of the individuals. To increase the density of the genotyping data, the Sanger Institute Imputation Service [120] has been used with the *Haplotype Reference Consortium release 1.1* (HRC) reference panel for imputation and EAGLE for pre-phasing, which are default settings. Phasing attempts to detected which alleles of a set of consecutive variants lie on the same chromosome, again by leveraging population level data. Pre-phasing is necessary for the pipeline setup of the Sanger Imputation Service.

The results after imputation contain information on the imputation quality of every variant. This measure, as well as Hardy-Weinberg equilibrium (HWE) p-value, and MAF, are important quality control measures for individual variants. The HWE p-value estimates the probability that the variant is inherited by random mating, and negligible mutation, migration, stratification, genetic drift and selection [121]. A threshold on the HWE p-value can be used to detect genotyping errors [121].

The MAF measures the relative occurrence of the less common allele of a variant in a population. The MAF of all measured variants in the dataset is obtained by counting alleles. Before testing the association of a variant with the expression of a gene, variants were reviewed of being imputed correctly (imputation score > 0.7), fulfilling HWE (HWE p-value $> 10^{-6}$) and for being common enough so that sufficient observations of gene expression for homozygous minor individuals are present (e.g. MAF > 0.05). When compiling the genotyping data, no variants were excluded, but all variants were annotated with these three metrics.

In the association test of the default eQTL analysis only variants were used that had an imputation score > 0.7, HWE p-value > $10^{-}6$, and a MAF > 0.05. It is explicitly stated if for an analysis variants not meeting one or several of these three metrics were used. The advantage of retaining variants that don't pass the default thresholds is that any variant that has been imputed can be tested on demand without rerunning data preprocessing procedures.

eQTL analysis

cis-eQTL analyses find associations where genomic variants in the proximity of a gene regulate the expression levels of the RNA generated from the respective gene. To identify these associations, only variants were tested that passed quality control measures and lie within the gene body or within 1 Mb windows flanking the 5' and 3' of the tested gene. The eQTL tests were performed using a linear mixed model (LIMIX version 0.8.5 [59]) taking population structure into account. Population structure within the samples was estimated by calculating the normalised covariance matrix of genotypes. Ten PEER factors were used as covariates for every dataset. Identification of independent association signals was performed by forward-selection - recursively adding the strongest associated variant of the previous association test as a covariate. Up to five iterations of forward-selection were performed, identifying up to five independently associated variants with the expression of one gene. This procedure is explained in greater detail in the next section.

I chose an unconventional way to perform the association test using the linear mixed model: in every association test, including the test for independent associations I regressed out all PEER covariates as well as the forward-selected genotypes prior to testing associations. The residuals specific to every round of forward selection were then tested with the linear mixed model. This is in contrast to the usual procedure using forward-selected genotypes as covariates in the model when testing for associations. The reason to implement this procedure was that in some cases high correlation ($r > 0.99$) between genotypes and covariates led to incorrect effect estimation by the linear mixed model and hence to strong false positive association results. The aforementioned procedure overcomes this issue, but also means a slight loss of detection power.

The method explained in the previous paragraph overcomes the problem of highly correlated variables in a linear model and hence delivers reliable detection of independent association signals. A caveat of the method is that it does not give easily interpretable effect size estimates for independent association signals. I solved this by using a two-step approach:

1. Identify the independent lead variants using the residual-based approach mentioned above.

2. Re-run the linear mixed model for all iterations of forward-selection using normalised fluorescence intensity measures as phenotypes, the PEER factors and the previously (forward) selected lead variants as covariates in one model to test for association with genetic variants.

3. Extend the results from the residual-based analysis with the effect size estimations of the full model in 2.

With this approach of re-estimation of effect sizes I take advantage of the robustness of identification of independent lead variants with the residual-based analysis, as well as estimating interpretable effect sizes, using the full linear mixed model with the defined independent associations. The unit of the effect size after re-estimation is the normalised binary logarithm of fluorescence intensity ($log_2(FI)$).

Aggregation of studies When using expression and genotyping data from more than one study for eQTL tests the expression and genotype data were stacked along the sample axis. The covariates were added, so that when using three datasets with 10 PEER factors each, in total 30 covariates were used in the joint association test. Covariate values for unobserved samples were set to zero. With this setup also the batch effect caused by joining datasets was captured. After performing eQTL association tests, the genome wide significance level was assessed in the following way: for every probe and every iteration of forward-selection, Bonferroni correction was performed individually. Next, the strongest associated variant was selected individually for every probe and every iteration of forward-selection - if for example five iterations of forward-selection were performed, then five lead (or tag) variants were selected. The false discovery rate (FDR) was calculated for every iteration of forward selection independently using the Benjamini-Hochberg procedure [122]. Therefore, with five iterations of forward selection, the Benjamini-Hochberg procedure was executed five times on the individual sets of Bonferroni-corrected variants. Finally, a threshold on the FDR of 0.01 was selected to define the genome-wide significance threshold.

The effects of LD on association tests

Linkage disequilibrium (LD) is a manifestation of the modes of inheritance. It measures whether two variants are inherited independently from each other or - if in LD - two variants are inherited together more frequently than by chance. The LD is measured as the coefficient of determination (r^2) and ranges from 0 to 1, with 1 indicating perfect LD between two variants. Typically LD reduces with the distance between variants reflecting the increasing chance of recombination events during meiosis. LD manifests as a correlation of genotyping data and causes an eQTL association test to link differences in transcript levels to a set of variants rather than a single one. Hence, variants in high pairwise LD necessarily yield similar levels of association strength thereby creating challenges when attempting to identify the functionally causal variant. As genotype correlation due to LD is consistent for common variants across similar populations, results of eQTL association tests are often reduced to reporting the lead variant at a position. This is the variant for which the strongest association was found in terms of p-values [123, 124].

Since gene regulation is a complex procedure it is likely that, for a single gene, multiple variants control gene expression independently. It is therefore biologically relevant to disentangle the associated variants by their correlation structure using conditional eQTL tests (see section **3.2.1**, page **30**). To indicate that the lead variant is not necessarily the causal variant, the lead variant and variants in high pairwise LD are called a signal. Conditional analyses can reveal independent association signals, each tagged by their own lead variant. The above forward-selection conditional analysis was applied for all eQTL analyses.

Validation of preprocessing pipeline

The expression and genotyping data of non-activated monocytes from the WTCHG study were used to check the validity of the preprocessing pipeline [112]. Association tests for eQTL identification were performed using the original preprocessed data [112] and the newly preprocessed data using the analysis pipeline described above.

The resulting associations were compared by probe identifier. Furthermore, LD between the lead variants which were associated with the same probe in the two differently preprocessed datasets was calculated. In case of high pairwise LD ($r^2 > 0.8$) the two association signals were assumed to be identical. If $r^2 < 0.8$, then it was assumed that the two preprocessing procedures yielded independent association signals. For this analysis only the primary lead variants were considered.

For associations with $p < 10^{-50}$ mostly the same identical signals were found to be associated with the probes (Fig. **3.1a** and Fig. **3.1b**). The weaker the association signals, the more likely it was that the two analysis approaches identified independent primary association signals (LD values of $r^2 < 0.8$). This observation does not imply that the association signals found in the original preprocessing were lost in the new preprocessing, rather it indicates a re-ordering of the association landscape, where the order of strongest signal, second strongest signal, etc. were swapped. This swapping indicates the sensitivity of eQTL analysis to data preprocessing.

In total, 88% of associated expression probes (eProbes) in the original preprocessing were also eProbes using the new preprocessing pipeline. Amongst the eProbes of the original and newly preprocessed data 23% had the same strongest signal with $r^2 > 0.8$. For the remaining 77% the newly identified strongest association signal was independent (Fig. **3.1a** and Fig. **3.1b**).

To demonstrate that variants that were associated in the original preprocessing were still associated in the new preprocessing, I extracted the raw p-values of the lead variants in the original preprocessing and the p-values calculated using the new preprocessing (Fig. **3.1c**). This analysis confirms that eQTL datasets are sensitive to preprocessing and that preprocessing alone can lead to the swapping the order of independent signals.

The results depicted in Fig. **3.1** show the plausibility of results produced by the new preprocessing pipeline when compared with the original results reported for the WTCHG study. The analysis also illustrates how sensitive eQTL data is to preprocessing. This in turn indicates the crucial importance of homogenising data generated by different groups by a unified preprocessing analysis before merging data.

eQTL results after preprocessing

After successful preprocessing of the expression and genotyping data a cis-eQTL analysis was performed for all datasets individually. For every tested gene, variants within 1 Mb up- and down-stream of the gene body, including the gene body itself were tested for association with the gene's transcript levels. An overview over the number of eQTLs identified for each of the cell types is presented in table **3.3**. Only primary association signals (ignoring results from conditional analyses) are displayed.

Using the same data, an eQTL analysis was performed by aggregating the expression and genotyping data by cell type across the studies. The resulting increase in sample size for e.g. monocytes led to the discovery of additional eQTLs as shown in table **3.4**. The

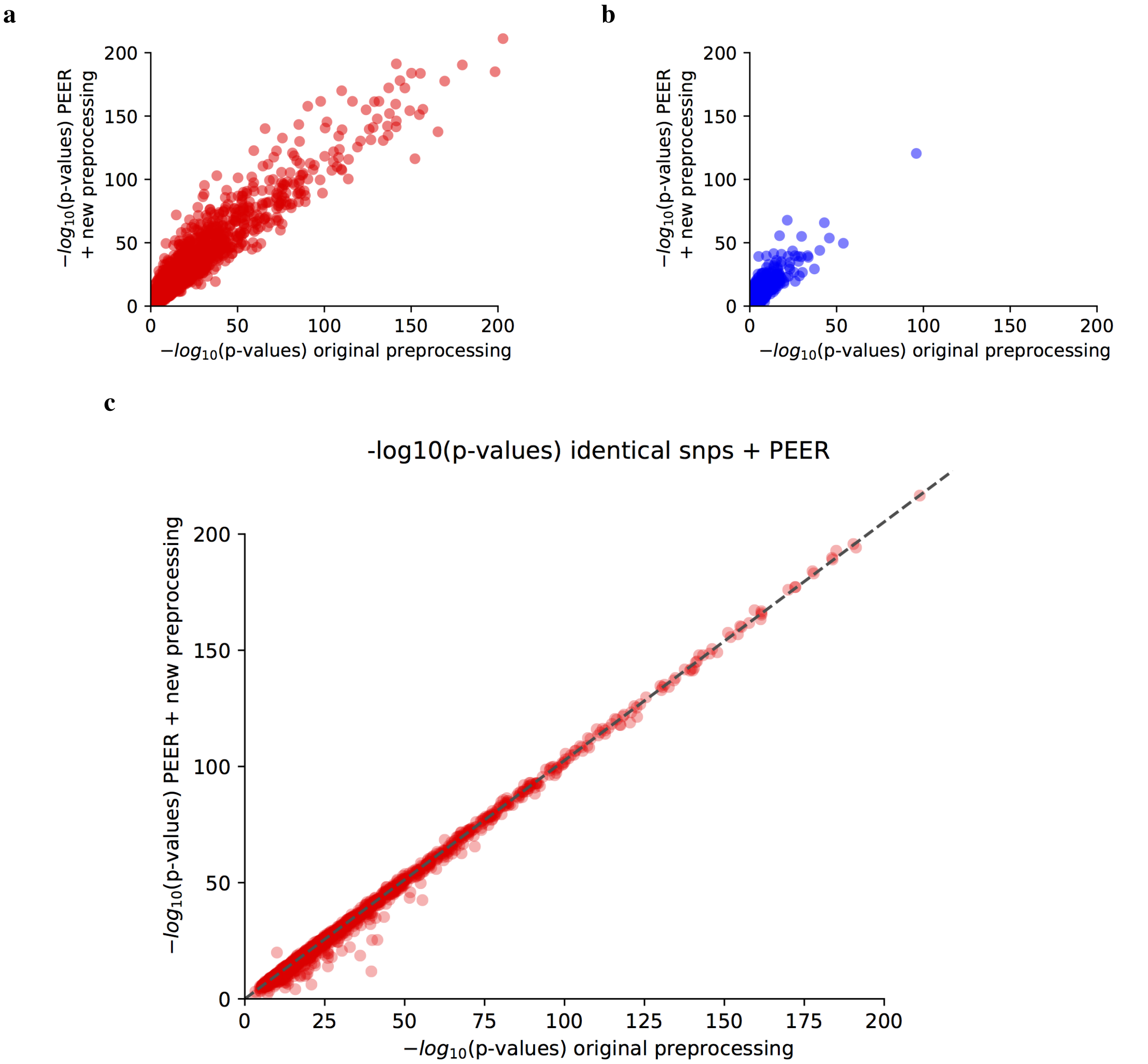


Fig. 3.1 Comparison of the association results obtained with the newly and original preprocessed data. Fig. **3.1a** eProbe associations: Red dots are associations where the lead variants identified by the two analysis are in high LD ($r^2 > 0.8$). Fig. **3.1b** eProbe associations: Blue dots are associations where lead variants were identified in both analysis but the respective lead variants are in LD with $r^2 < 0.8$. Fig. **3.1c** Replication of association p-values after new preprocessing. Association p-values of original and the newly preprocessed dataset are displayed on the horizontal and vertical axis, respectively. Each dot represents a single lead variant - eProbe combination. For the new preprocessing 10 PEER factors have been included.

Cell type	CD Marker	Study dataset	# samples	# probes tested	# eQTLs	# eGenes
Monocyte	CD14	CEDAR	301	17,270	2,192	1,866
Monocyte	CD14	Cardiogenics	758	11,271	4,577	4,014
Monocyte	CD14	WTCHG	421	16,160	4,474	3,686
Macrophage	MAC∗	Cardiogenics	599	11,136	4,148	3,674
Neutrophil	CD15	CEDAR	303	17,094	1,081	923
Neutrophil	CD15	WTCHG	109	16,249	1,177	1,001
B-cell	CD19	CEDAR	298	17,175	1,230	1,058
B-cell	CD19	WTCHG	285	16,380	2,854	2,429
T-cell	CD4	CEDAR	309	17,056	2,359	2,012
T-cell	CD8	CEDAR	304	17,098	1,891	1,619
Platelet	PLT∗∗	BLUEPRINT	151	8,653	1,019	849
Platelet	PLT∗∗	CEDAR	236	16,973	458	395

Table 3.3 Primary eQTL signals per blood cell type. ∗macrophages were obtained by culturing of CD14+ monocytes and were paired samples with the monocytes from the Cardiogenics study; ∗∗platelets were obtained from whole blood by centrifugation and no monoclonal antibody was used for their purification.

increase in power by merging datasets from the different studies will be discussed in the following section. The Cardiogenics cohort consists in healthy individuals and others that were diagnosed with coronary artery disease. Analyses with and without the affected sub-cohort showed that the gain in power outweighed the added noise which led me to the decision to use the full Cardiogenics dataset for all further analyses.

Cell type	CD marker	# samples	# probes tested	# eQTLs	# eGenes
Monocyte	CD14	1,480	10,113	5,368	4,731
Macrophage	MAC	599	11,136	4,148	3,674
Neutrophil	CD15	412	14,986	1,831	1,539
B-cell	CD19	583	14,805	2,556	2,183
T-cell	CD4	309	17,056	2,359	2,012
T-cell	CD8	304	17,098	1,891	1,619
Platelet	PLT	388	8,293	1,031	862

Table 3.4 Primary eQTL signals per cell type after merging of the data from the different studies

3.2.2 Increase of power to detect eQTLs

To show the effects of an increased sample size in eQTL analyses, the results gained from merging the monocyte datasets from the CEDAR, Cardiogenics and WTCHG studies were compared to the results of the original studies. This analysis shows the relation between increased sample size and the number of eQTLs detected and the proportion of replicated eQTL associations identified in the original individual studies.

The preprocessing of the data as well as the criteria for the selection of variants and genes were identical across analyses, e.g. the datasets of all five studies were preprocessed as described in section **3.2.1** and cis-eQTL associations with gene expression levels were tested including all variants in proximity (<1 Mb up- or down-stream of the gene body) or within the gene body. Only chip bead probe signals and genotyping variants passing the quality control criteria (section **3.2.1**) were used. In order to achieve the same minimum number of gene expression observations for at least three individuals homozygous for the minor allele in all analyses, a minor allele homozygosity count (MAC) threshold was used instead of a threshold on variant MAF. This approach improves the robustness of expression estimation for the group of individuals homozygous for the minor allele. For analyses in this section a MAC-threshold of 3 was used, requiring at least three observations of homozygous minor carriers.

As mentioned in section **3.2.1**, slight alterations in the input data, for example caused by differences in preprocessing, may result in the identification of apparently different lead variants. Therefore and because of LD between variants replication of eQTL associations cannot be reduced to comparing lead variants. To overcome this, pairwise LD was calculated between lead variants identified in the merged monocyte eQTL dataset and the ones found in monocyte eQTL data from the individual studies. An association signal was assumed to be identical if for the same probe identifier the pairwise LD amongst the associated variants was $r^2 > 0.8$.

First, the merged dataset for monocytes was used as a discovery set, revealing that the analysis of the merged dataset replicated the largest portion of eQTLs observed in the analysis of the data of the individual studies (first three bars in Fig. **3.2a**). The increased number of eQTLs observed in the analysis of the merged data compared with the analysis of the data from the individual studies are mainly explained by the stringent threshold of requiring at least three homozygous individuals for the minor allele. The sparsity of eQTLs in the analysis of the data of the original studies at lower minor allele frequencies is illustrated by a lower density of blue data points in Fig. **3.2b**.

The following conclusions can be drawn for the re-analysis of the data for monocytes from the three original studies (Cardiogenics, CEDAR, WTCHG) and the analysis of the

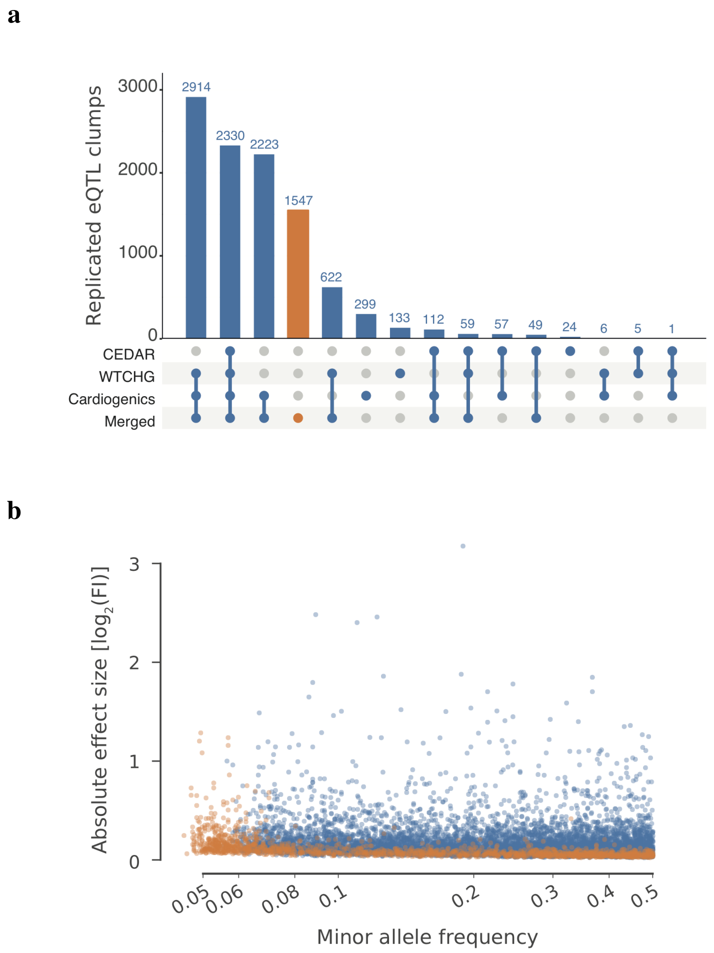


Fig. 3.2 Replication and gain of power for eQTLs detection in monocytes. The replication of eQTLs observed in the analysis of the data from the three original studies with 'merged' representing the results obtained after aggregated data. Fig. **3.2a** The top bar diagram shows the number of eQTL replications in the merged data on the vertical axis and the bottom panel shows to which studies the replication refers. The column in orange illustrates the new eQTL signals not previously observed by analysis of the data of the original studies. Fig. **3.2b** Graph showing the correlation between MAF and effect size (β, $[log_2(FI)]$) of the eQTL on the horizonal and vertical axis respectively. New, previously unobserved eQTLs are in orange and replicated other are in blue. β is displayed on the vertical and MAF on the horizontal axis.

merged data for monocytes of these three studies. First, 94.94% of the original eQTLs are replicated in the analysis of the merged data. However the remaining 5.06% were not detected in the analysis of the merged data (Fig. **3.2a**); e.g. 299 eQTLs identified in the Cardiogenics data remained unobserved in the analysis of the merged data and in the analysis of the CEDAR and WTCHG data. The most likely explanations for this lack of replication are: i) none of the eQTL studies performed to date are of a size that the power for identification of associations is saturated and whether an eQTL is identified or not remains an issue of sampling, ii) the original association may have been false-positive [125], iii) the eQTLs identified only in Cardiogenics may be specific to the cohort which included in addition to samples from 395 healthy individuals also samples from almost an equal number (363) of individuals with a history of coronary artery disease and iv) relatively weak association signals from the analysis of the Cardiogenics data may be 'washed-out' as a consequence of the merging of the data [126].

With the increase in sample size the number of eQTLs with relative small effects on transcript levels increases and this is observed across the entire spectrum of MAFs (Fig. **3.2b**). Finally 1,547 new eQTLs were observed in the analysis of the merged data which were unobserved in the analysis of the original data and a substantial portion of these new ones were observed for low MAFs.

Altogether, merging of eQTL study data can improve power but reprocessing of the measurement data is important to minimise the risk of additional noise being introduced by slight differences in preprocessing approaches.

3.2.3 Low-frequency variants

Under normal evolutionary assumptions one would reason that rare variants, which are generally recent and population-specific [127, 128] have a greater chance of exerting large effects on expression than common variants [45]. This observation is generally confounded with the lack of power to detect associations of rare variants with weak effect sizes.

The chance of detecting eQTLs with low MAFs and large effect size is correlated with the number of individuals included in the analysis. The merged dataset generated as part of the research for my PhD provided a unique opportunity to search for low frequency variants with large effect sizes on transcript levels in monocytes. In the initial analysis which focused on the replication of associations between studies presented in section **3.2.2**, the analysis for eQTLs was constrained to relatively high MAFs by setting a threshold of having at least three individuals homozygous for the minor allele being included in the analysis.

In a next analysis step this high threshold was removed and all 7,191,953 variants passing quality control and with a MAF > 0.01 were included in the association test. The results

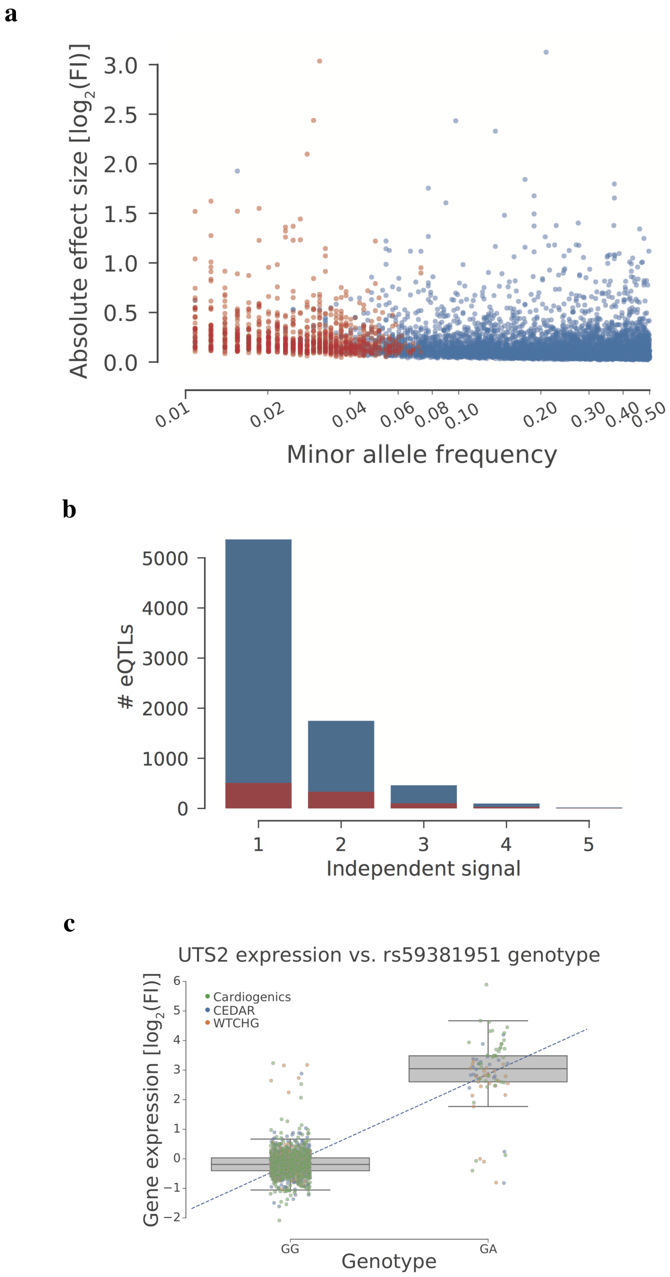


Fig. 3.3 Low-frequency variants in the merged monocyte dataset. Orange are rare variants with a MAF $<5\%$ in the UK10K cohort [100]. Fig. **3.3a** Effect size and MAF estimated from the sample genotypes of all lead variants in the merged dataset. Fig. **3.3b** Number of eQTLs found in each iteration of forward-selection. Fig. **3.3c** Example of a rare variant with a strong association and a large effect size. Illumina bead array probe 6860048 was associated with variant 1:7962498:C:G, $p = 1.94594x10^{-273}$, effect size = 2.50104 $log_2(FI)$, MAF: 0.0294118 in study population. The associated variants were annotated with the MAF in the UK10K cohort [100]. This annotation was used for visualisation purposes as an upper limit for the selection of low-frequency variants. Only variants with a MAF < 5% in the UK10K cohort were considered low-frequency (Fig. **3.3a**)).

of this association analysis, which are shown in Fig. **3.3a** identified 7,693 eQTL variants (up to five independent signals per probe, FDR < 0.01). When annotating the lead variants with MAF estimates from the UK10K cohort [100] (MAF_{UK10K}), 1,134 eQTLs of the lead variants had a $MAF_{UK10K} < 0.05$. The conditional analysis of the signals in these 1,134 loci, as described in section **3.2.1** revealed independent secondary and up to quinary association signals. Low-frequency variants were found across the full spectrum of independent signals (Fig. **3.3b**).

A subset (328) of the eQTLs associated with low frequency variants were based on the comparison of the transcript levels in monocytes from individuals homozygous for the major allele and heterozygous individuals. It could be reasoned that the absence of observations from carriers of the homozygous minor allele increases the risk of observing a false-positive eQTL, but that this risk diminishes if the variant localises in a functionally active element. Therefore, in a next analysis step the eQTL variants were integrated with ENSEMBL information about epigenetic states [129] and data on transcription factor binding sites from the ENCODE and Roadmap projects [130–132]. This data integration yielded annotation for 451 of the 1,134 variants and the eQTL variant with the highest β value and lowest p-value of association (2.43702 and $1.94594x10^{-273}$, respectively) was for UTS2 (Fig. **3.3c**). The lead variant rs59381951 at this locus is localised 49 kb upstream of the TSS of gene UTS2 and has an epigenetic label as *active promoter flanking* region in monocytes, macrophages, neutrophils, CD4+ and CD8+ T lymphocytes and B lymphocytes.

UTS2 encodes a highly conserved oligopeptide of 124 amino acids with a well defined biological role [134]. It acts through binding its G-protein coupled receptor UTS2R, previously known as GPR14, which has been extensively investigated for its role as a vasoconstrictor [135]. According to *RefSeq* UTS2 is only expressed in brain tissue [134]. Interestingly, the retrieval of the RNA-seq expression data from the BLUEPRINT project shows wider expression in haematopoiesis (Fig. **3.4a**) and expression data from the Cardiogenics dataset show detectable levels of UTS2 in monocytes (Fig. **3.4b**) and macrophages (Fig. **3.4c**).

3.2.4 Comparison to other large eQTL analyses

Blood, being the most accessible tissue, has been attractive to develop the principle of eQTL studies. Most eQTL studies on blood have however been performed with RNA obtained from samples of whole blood. A recent meta-analysis of whole blood studies by the eQTLgen consortium (Lude Franke, University of Groningen, the Netherlands) included data from 14,000 individuals across 18 original eQTL studies. To inform the design of future eQTL studies I compared the results from their whole blood eQTL data with the results I obtained from the merged data for monocytes. This comparison was made possible because the

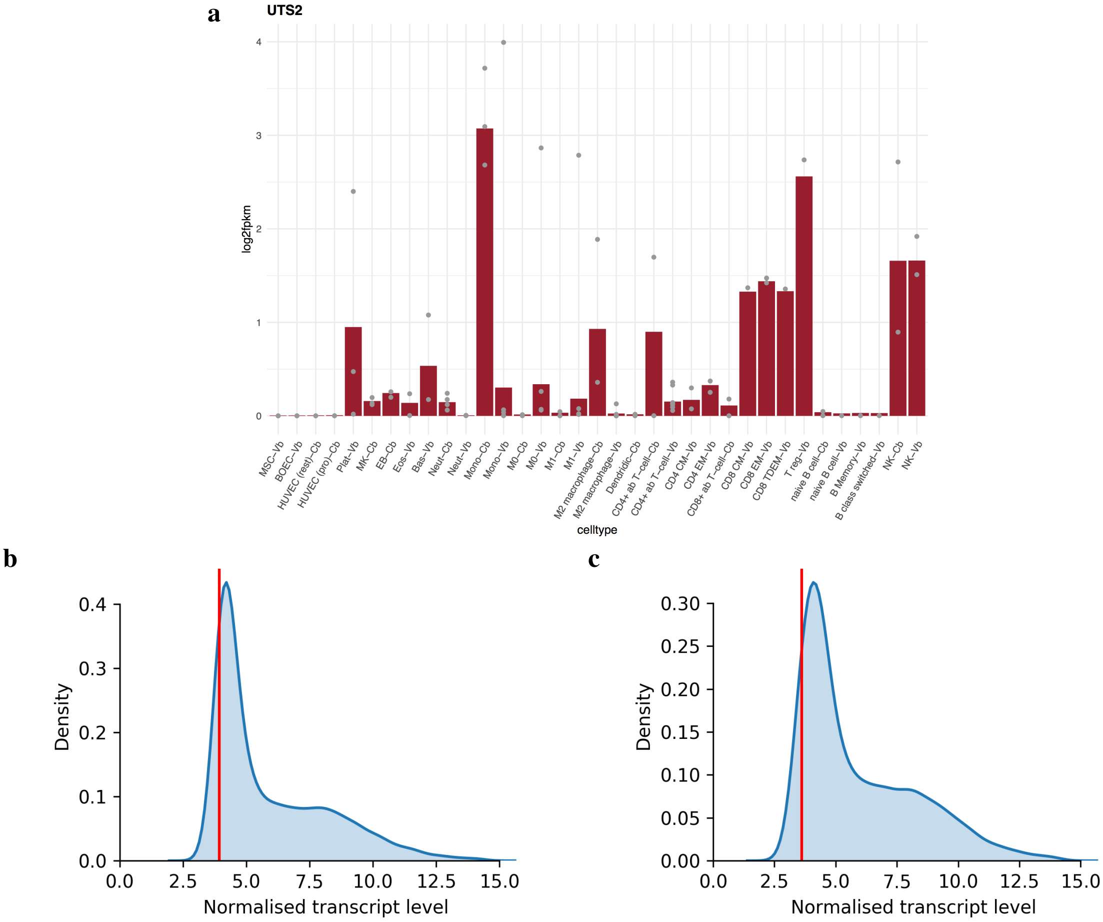


Fig. 3.4 UTS2 transcript levels Fig. **3.4a** in purified blood cell types, measured in multiple individuals (grey dots), red bars are mean $log_2(FPKM)$ measures of transcription. Values below 1 are considered background noise. Taken from BLUEPRINT RNAexpress [133]. Fig. **3.4b** Histogram of the median across samples of normalised mean of the binary logarithm of fluorescence intensity (MFI) of all genes measured in monocytes as part of the Cardiogenics expression dataset. Red vertical line indicates median normalised MFI of UTS2. Fig. **3.4c** identical to Fig. **3.4b** but displaying Cardiogenics expression data on macrophages.

eQTLgen consortium generated a reference for the mapping of microarray probes across several array platforms, including the Illumina Bead Chip Human-Ref-8 v3 and HumanHT-12 v4.0.

To make the two association analyses comparable, both the eQTLgen and the merged monocyte results were constrained to those eQTLs that were tested in both analyses. The imputation of genotypes for the two studies was not identical. Therefore variants found to be associated in either of the two eQTL analyses were linked based on their pair-wise LD value. Variants with with a pairwise $r^2 > 0.8$ were treated as identical. LD between variants was calculated using whole-genome sequence data from 7,493 genetically independent individuals of European ancestry whose DNA samples were sequenced as part of the rare diseases pilot study for the 100,000 Genomes Project, which was performed by the NIHR BioResource team at the University of Cambridge. The comparison of the associations identified by the eQTLgen and the merged monocyte eQTL studies restricted to primary association signals only and two different criteria were used for the LD-clumps of lead variants to asses whether there was replication of the eQTL association or not:

Firstly, a genome-wide significance FDR < 0.01 was used in both the detection and the replication analysis and 92.0% of the eQTLs in the merged monocyte data were found to replicate in the eQTLgen data (first bar in Fig. **3.5**). Secondly, a genome-wide significance threshold was used in the merged monocyte discovery data and a lenient p-value of 0.05 was used in the eQTLgen data, yielding replication of 97.1% of associations (Third bar in Fig. **3.5**). This more lenient p-value for replication is motivated by the assumption that once an association has been identified in a discovery experiment then less evidence is necessary for its replication [87]. These results suggest that almost all eQTLs discovered in an association study using purified monocytes from 1,480 individuals can also be observed when using whole-blood eQTL data from 10 times more individuals.

The sample size of the eQTLgen meta-analysis is far greater than the one for the monocyte dataset reported in this thesis. Therefore, the results of one of the eQTL studies used as part of eQTLgen (Fehrmann et al. [136], 1,227 individuals) was compared with eQTLs of purified monocytes from 1,480 individuals. Only results obtained with probes present on the bead chips of both studies were used and lead variants were clumped with their neighbouring high-LD variants. Applying a genome wide significance threshold (FDR < 0.01), 65.4% of the associations found in the whole blood study were also detected in the results obtained with monocytes. Conversely only 50.9% of the monocyte eQTLs were detected in the whole blood study at the stringent genome-wide significance level, which increased to 80.7% at a more lenient threshold of $p < 0.05$.

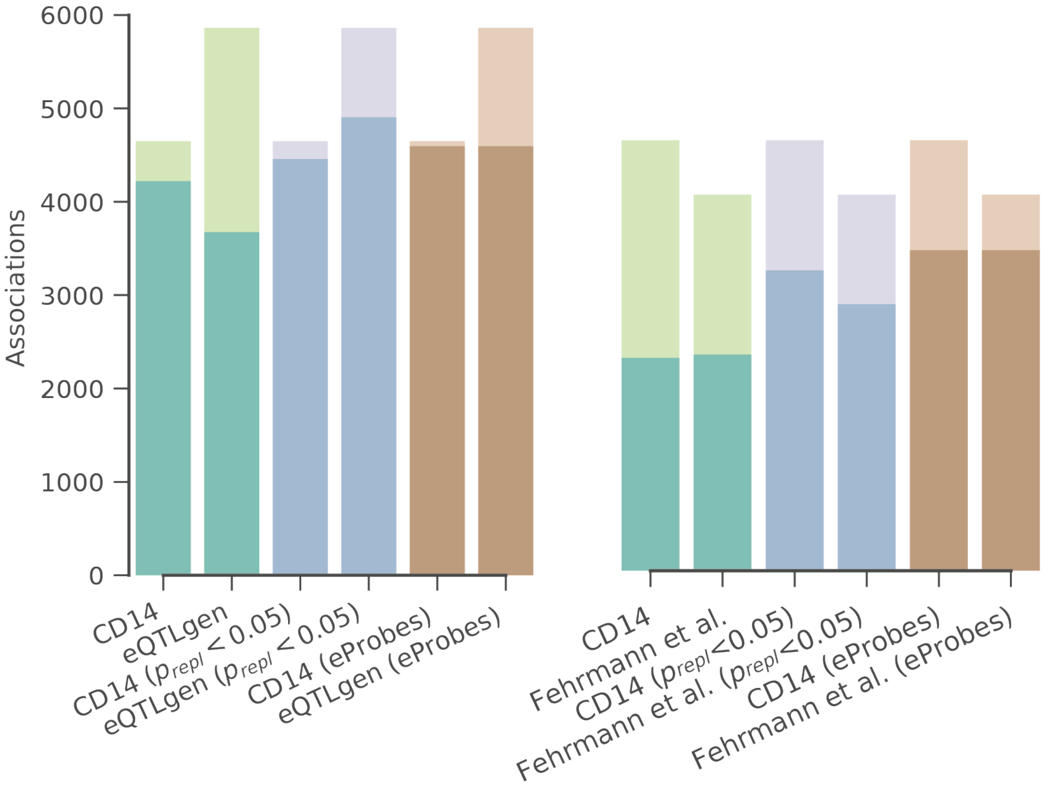


Fig. 3.5 Replication of eQTL results between merged monocyte data (CD14) and eQTLgen (left panel) and Fehrmann et al. [136] (right panel). The labels on the x-axis are the names of the replication datasets. Dark bars display the amount of replication, light bars are the numbers of eQTLs detected for the commonly tested microarray probes. Green bars: the genome wide significance threshold of FDR < 0.01 was used for replication. Purple bars: a nominal p-value threshold of 0.05 was used for replication. Brown bars: all probes that were also associated with any variant in the other dataset are displayed as replicating.

Overall, this analysis has shown that eQTL results obtained with whole blood RNA are capable of replicating most of the associations observed by eQTL using RNA from purified cells. There are several reasons why the whole blood eQTL results are less informative when used in an integrated fashion such as for GWAS colocalisation studies. Firstly it is unclear from which cell type the eQTL signal in whole blood arises. Secondly, a subset of eQTLs exert their effect on transcript levels in different cell types and sometimes with opposite directionality [87]. These aspects of shared eQTLs between cell types will be discussed in more detail in the next chapter. An alternative approach to the use of RNA obtained from whole blood or purified populations of cells is to apply single cell expression data in eQTL studies [137]. This is a promising approach, as the cell type and cell state can be assessed for every measured cell individually post hoc. These advantages come with the drawbacks of the high cost of single cell RNA sequencing and the sparsity of the number of reads obtained per cell.

3.3 Discussion

In this chapter I have presented a joint analysis of four independent eQTL studies. The four studies made use of the same bead chip array by Illumina to measure relative RNA abundance and genotyping results were obtained by Illumina genotyping tests or by whole genome sequencing at low coverage. The use of these very similar experimental approaches made the joint analysis possible. To ensure compatibility between the experimental data and to minimise noise resulting from the batch effects, the expression data was preprocessing through a single analysis pipeline specifically designed for this study. The 88% concordance between the original eQTL results from the WTCHG study for monocytes and the results obtained after the reprocessing of the same raw expression data validated the new analysis method.

After this quality control, experiment data from all four studies was reprocessed and data on the same cell type (monocytes, neutrophils, platelets) was combined across studies to increase the power to detect eQTL signals at variants with lower MAFs.

Indeed, the eQTL results obtained with the aggregated data for monocytes showed more power compared to the results from the three individual (Cardiogenics, CEDAR, WTCHG) studies. With increased sample size, additional associations were identified not only with common variants but also with less frequent ones. This also indicated that thanks to the unified preprocessing little noise was added while aggregating the data. A total of 15,851 independent eQTLs were identified across the seven cell types. Merging of data for the same type of cell improved the robustness of the detected associations. It is postulated that a considerable portion of the original eQTL variants which did not replicate were likely to be false positives or associations that were unique to one sample collection, e.g. half of the participants in the Cardiogenics study had a history of a cardiovascular event before enrolment [88, 89] and the participants in the WTCHG study had been enrolled over a very short time period to reduce seasonal variations in the expression levels of a subset of genes [126]. Altogether, the joint and unified analysis has resulted in a unique resource of robust eQTL associations for six different types of blood cells and also for ex vivo generated macrophages. These results are of considerable value for colocalisation studies to map variants identified by GWAS to their target genes.

Such colocalisation studies can also be performed with eQTL results from whole blood studies. One of the advantages of whole blood studies over studies with purified cells is that the former type is simpler to scale and therefore cheaper. Interestingly when using the results from the recent eQTLgen meta-analysis study which included whole blood expression results from 14,000 individuals, 92.0% of the associations found in the eQTL study performed were replicated with the monocytes from 1,480 individuals. This indicates that whole blood eQTL studies, when large enough, can capture eQTL signals which emanate from a relatively rare type of leukocytes (monocytes make up 2-10% of whole blood). A distinct disadvantage of bulk whole blood eQTL studies over those with purified cells remains because of our limited ability to deconvolute the whole blood effect sizes to the different cell types. A method has been developed by Westra et al. [138] to estimate the mediating cell type in whole blood cis-eQTLs. It relies on the abundance estimates of the constituent cell types in whole blood and gives a p-value of how likely a selected cell type is mediating the whole blood eQTL. The authors suggest to use results from this method as a supplement to eQTL studies performed on purified cell types. Its power to detect cell type mediated eQTLs depends on the abundance of the cell type in whole blood and may require whole blood datasets of thousands of individuals.

In this chapter I have shown the technical approach used to generate the unified eQTL analysis and how to relate its results to those obtained by whole blood eQTL studies. The next chapter will focus on the identification of eQTL signals shared between cell types, the enrichment of eQTL associated variants in nucleosome depleted elements, the presence of eQTL variants in long-range promoter interacting elements and colocalisation of eQTL variants with variants identified by a GWAS for blood cell indices.

Chapter 4

Application of expression quantitative trait loci

Expression quantitative trait loci (eQTLs) provide information about how genetic background is involved in a cellular phenotype, and can explain about 31% of expression variance [139]. eQTL results give an insight into genetic control of gene expression, but the gain of knowledge from direct interpretation of associations is limited. Gene regulation is complex and relies on several layers of control, and only one of them is genetic background. To help understand genetic control, I will compare eQTL results across the seven cell types and highlight cell type specific associations. Next, I will put the eQTL results into bigger context, overlapping them with genomic regions which are known to be important for gene regulations. Apart from the direct interpretation, eQTL results have been used as a tool to help improve the understanding of results from genome-wide association studies (GWAS) by means of colocalisation. I will perform a colocalisation analysis to explain how a GWAS variant mechanistically exerts its effect on plateletcrit - a measured blood index derived from a routine full blood count analysis as recently reported by Astle et al. [45]. Finally, I will draw a link between gene regulation, three dimensional chromatin structure and gene constraint against coding loss of function mutations.

4.1 Cell type specificity

Gene expression, as well as the gene-regulatory landscape, can be cell type specific [28]. Some genes are expressed more homogeneously across cell types than others, but some genes are highly cell type-specific. Gene expression atlases have been created which contain expression levels of genes across cell types and tissues.

Since the existence of an eQTL is necessarily linked to the expression of the associated gene in a cell type, it is expected that a subset of eQTLs are cell type-specific. In efforts such as GTEx [77] a wide range of tissues were studied. eQTL replication was used to find patterns of tissue specificity and the inference of genetic pathways. Fairfax et al. [87] have studied the differences in eQTLs using monocyte and B-cell data and they found cell type specific cis-eQTLs, as well as switching of effect sizes between cell types. In a later study, the same authors [112] found similar results when comparing monocytes in different activation states. In this section, I will identify cell type specific associations, as well as pleiotropic ones, which replicate across all cell types.

LD-clumping

Before I could compare eQTLs across cell types, it was necessary to ensure that the results in the different cell types were made comparable. eQTLs are associations between DNA variants and the expression level of genes. Every independent association signal is summarised by the strongest associated variant or so-called tag or lead variant (section **3.2.1**). Due to the noise of the input data an association of a gene with variant a detected in cell type A can be the same association as in cell type B, even if the eQTL results for cell type B give an association with variant $b \neq a$. In those cases, the equality of association signals is estimated by calculating the pairwise linkage disequilibrium (LD) between a and b. If a and b are in high LD ($r^2(a,b) > 0.8$) then the association signal in cell types A and B is assumed to be identical. LD was calculated as the correlation of genetic variants based on whole-genome sequencing data from 7,493 genetically unrelated individuals from Northern European ancestry. Their DNA samples were analysed by whole genome sequencing as part of their participation in the NIHR BioResource - Rare Diseases project (currently unpublished data, Ouwehand, NIHR BioResource, University of Cambridge, UK). Using this LD-based criterion it was possible to clump variants together and to make association signals comparable across different types of blood cells.

Technically, the comparison was performed as follows: in the first iteration all gene-variant pairs that tag a conditionally independent association signal were collected in every cell type. These pairs were aggregated across all datasets. Next, association p-values and effect sizes of all the aforementioned gene-variant pairs were extracted from eQTL results in all cell types. Not all gene-variant pairs were tested in all cell types, because not all genes are expressed in all cell types. All the variants that were associated with one gene across all cell types were clumped by LD as described above to establish replication of association signals.

I then defined replication for an LD clump if one of its variants was associated at genome wide significance level in at least one cell type (discovery set) and variants in the same clump had a minimum nominal p-value <0.05 in the other cell types (replication set). This criterion was applied across all genome wide significant associations of all cell types to generate a replication matrix of signals.

In total 15,851 independent eQTL LD-clumps were identified across the seven cell types. Splitting LD-clumps by replication, the biggest proportion of eQTLs (4,682) was found to be cell type specific (Fig. **4.1a**), but 281 eQTLs replicated in all cell types (rightmost bar in Fig. **4.1a** and orange bar in Fig. **4.1b**). Platelets, apart from being the cell type with the fewest identifiable eQTLs, showed strong dissimilarity to all the other investigated cell types. This is illustrated in Fig. **4.1b**, where the biggest set of eQTLs that platelets share with any of the cell types is the set of eQTLs that are shared across all cell types. On the other hand cell types that are closely related in terms of haematopoeitic ontology like monocytes and macrophages tend to share more eQTLs with each other than with any other cell type.

Using this eQTL replication matrix I created a hierarchical cluster of the seven cell types using the cosine distance metric (Fig. **4.2**). The clustering dendrogram gives a plausible representation of the haematopoietic tree. This replication result shows the robustness of the eQTL datasets, which were derived from studies performed at different laboratories.

4.1.1 Effect size across cell types

So far I have investigated eQTL replication in terms of p-values of associations, for biological interpretation the effect size and directionality of the effect are of interest. Effect size measures the difference in mean expression levels in the presence of the minor allele. In other words, the magnitude of the effect indicates by how much the expression is altered by 'adding a minor allele', and effect directionality describes if carriers of the minor allele on average have higher or lower expression of a gene, compared to expression levels measured in individuals lacking the minor allele.

Now, I will compare effect sizes across the seven different cell types. Previous studies [87] have already shown that effect sizes may differ between cell types and even between different activation states of the same cell type [112]. I analysed the replication of effect directionalities in eQTLs that replicate across all seven cell types. Of the 281 eQTLs that replicate across all cell types (Fig. **4.1b**) there were 30 genes whose association did not have the same effect directionality in all cell types (Fig. **4.3**). This difference in effect direction was not an artefact of technical nature because it was not caused by PEER covariates, nor by inaccuracies from misalignment of genetic allele order across cell types, nor by expression data preprocessing. As technical noise was excluded, these observations most likely have biological cause. Gene expression regulation is a complex process and expression is rarely controlled by a single locus. The variability of effect magnitude and the change in effect

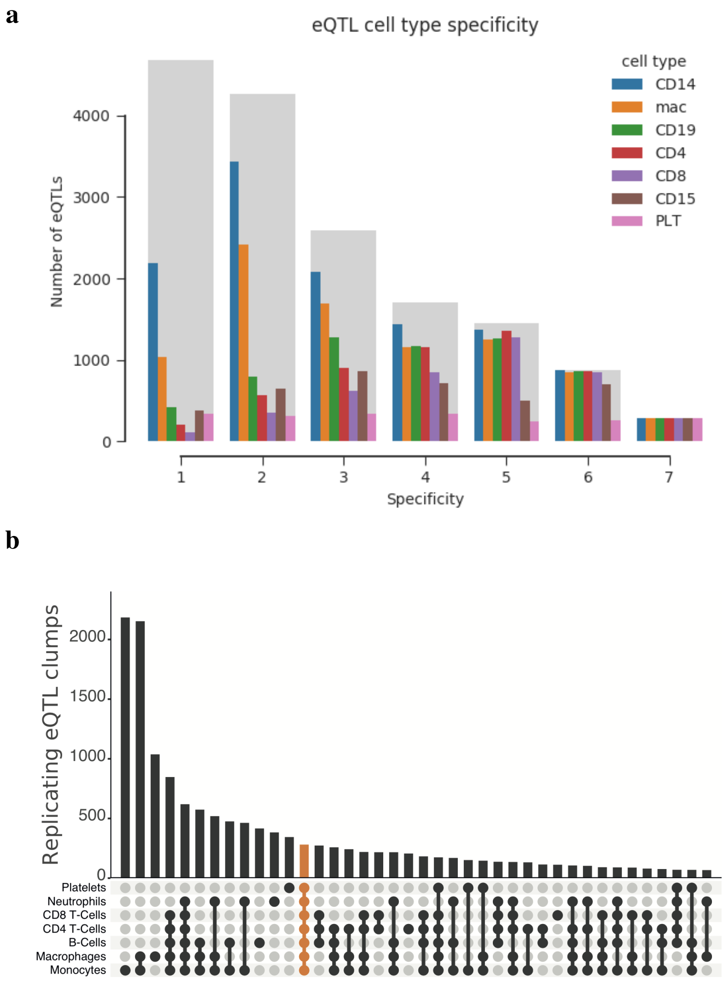


Fig. 4.1 eQTL datasets for seven types of blood cells. Fig. **4.1a** Cell type specificity of eQTLs. The number of eQTLs per cell type are on the y axis and the x axis indicates in how many cell types the respective eQTLs have been replicated. The coloured bars represent the number of eQTLs per cell type and summation of these numbers across the cell types are represented by the grey bar. Fig. **4.1b** eQTLs shared between cell types. The top panel shows the number of eQTLs per category on the y axis and the matrix in the bottom panel shows the different categories of eQTLs based on whether they are shared between different types of cells or not.

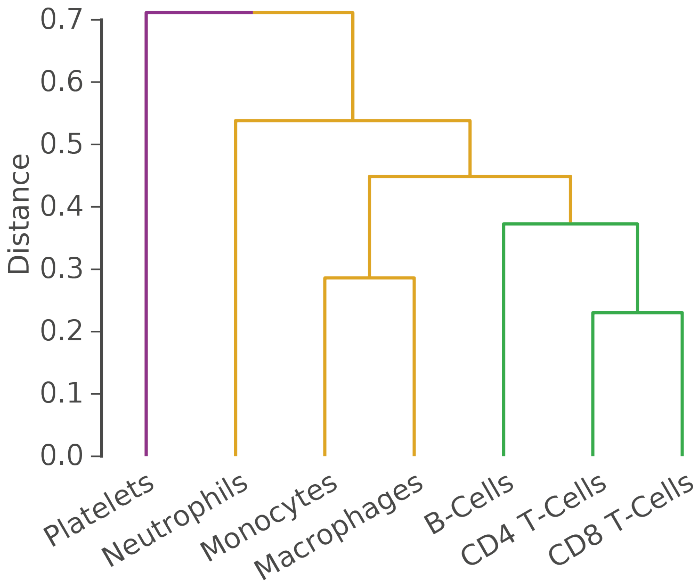


Fig. 4.2 Recapitulation of the haematopoietic tree by use of the eQTL datasets. Hierarchical clustering of cell types based on a non-supervised clustering using the eQTL datasets for the seven types of blood cells.

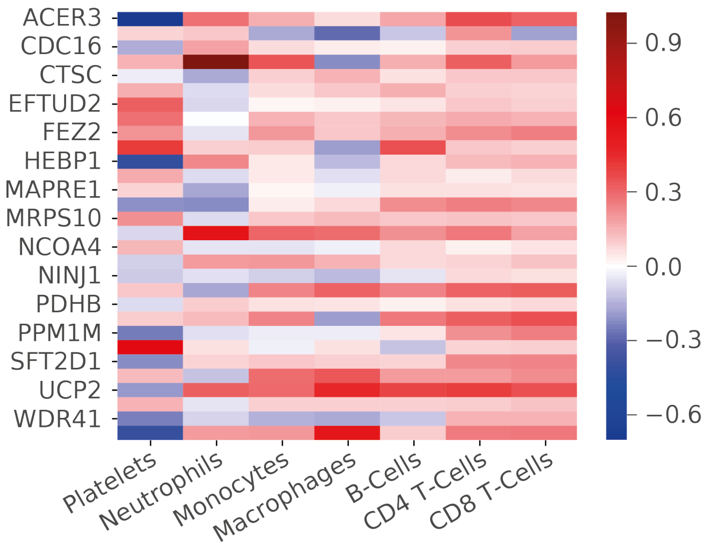


Fig. 4.3 Pleiotropic eQTLs with opposite effect direction in different cell types. Displayed are the eQTL effect sizes (β, $[log_2(FI)]$) (red to blue colour intensity) and directionality (positive and negative β values in red and blue respectively). Cell types and gene names on horizontal and vertical axis, respectively.

direction are strong indications of the complex interplay of many factors in gene regulation and the differences between cell types in how gene transcription is initiated and maintained. An important molecular mechanism underlying differences in gene transcription detected by eQTLs is the differential level of binding of the transcription factors at the position of the lead variants. The different states of the seven cell types is to a large extent determined by differences in the abundance of key transcription factors which determine cell state [81, 140]. Therefore I reason that it is highly unlikely that the lead variants of the 281 eQTLs detected in all seven cell types are occupied by the same composite of transcription factors. Further experimental studies are warranted to improve the understanding which molecular mechanisms underlie the shared eQTLs.

The observed sharing or lack of sharing of eQTLs between blood cell types does highlight the limitations of eQTL data generated with RNA obtained from samples of whole blood. Such eQTL data is more cost-effect to produce for large sample sizes [138], but signal deconvolution between a variant and the expression of a gene in a single type of cell is challenging and the correct estimation of its eQTL effect magnitude is not possible. In samples of whole blood taken from different individuals the ratio between the cell types differs and these differences are for a large extent (up to 20%) genetically explained [45]. Additionally, the differences in RNA levels between non-nucleated cells (red cells, platelets) and the nucleated ones are extensive adding further confounders to the measured effect size in an whole blood eQTL study.

4.2 eQTLs in regulatory regions

More than 90% of all the lead variants identified by GWAS for common diseases and medically relevant quantitative traits are localised in the non-coding portion of the genome [43]. Similarly, eQTL variants are by their nature localised in the intronic or intergenic space. It has been shown that a subset of eQTL lead variants are localised in nucleosome-depleted (accessible) regions of the genome as determined by DNAse hypersensitivity assays [141] or more recently by sequencing assays to identify transposase-accessible chromatin sites (ATAC-seq). Furthermore, it has been shown that eQTL variants that lie in regulatory active regions, such as enhancers, are more likely to replicate across studies and hence are potentially more robust findings [142]. In this section I will discuss the overlap of eQTLs with accessible elements of nucleosome-depleted chromatin using ATAC-seq data. Gene expression regulatory factors, such as transcription factors, bind in regions of accessible chromatin [143]. Therefore, variants in accessible chromatin are potentially of higher relevance to gene expression regulation. The ATAC-seq data for monocytes, CD4 T-Cells and

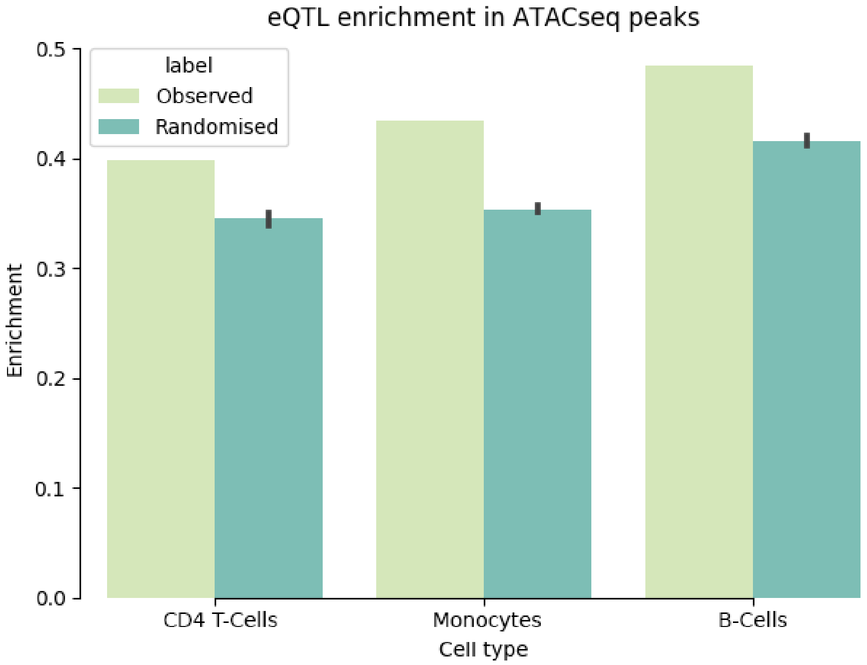


Fig. 4.4 Enrichment of eQTL variants in nucleosome depleted regions (accessible chromatin) of monocytes, CD4 T-Cells, and B-Cells. ATAC-seq data was used to define nucleosome depleted regions.

B-Cells was generated and preprocessed by the laboratory of Dr Mattia Frontini (Department of Haematology, University of Cambridge). To determine the extent of overlap between eQTL variants and accessible chromatin of the matching cell types I determined the level of enrichment using the GoShifter software (version 0.2) [144].

GoShifter extends the associated eQTL variants by adding their LD-proxies ($r^2 > 0.8$) using 1,000 Genomes data [145] (LD matrices used from BEAGLE package release 2 [146]) and the amount of overlap of the extended set of variants with the epigenetic annotation is calculated. In order to test for statistical significance, GoShifter randomises the position of the epigenetic annotation and calculates the overlap for every iteration of randomisation. For this experiment I used the ATAC-seq data for the corresponding cell types and executed GoShifter with 1,000 random permutations. I identified significant enrichment ($p < 0.001$) for all three cell types (Fig. **4.4**).

These results suggest that about 40% of LD-extended eQTLs lie in accessible chromatin (Fig. **4.4**). The results are comparable with the results by Gaffney et al. who observed 40% overlap between eQTLs and DNAse hypersensitivity sites for lymphoblastoid B cell lines [141]. There are some plausible reasons why the remaining 60% of eQTL variants do not overlap with accessible chromatin. Firstly, the genotyping data is collected in coarse manner

- using genotyping arrays and limited sample sizes affects the correct the identification of causal variants. Secondly, differences in transcript levels in a fully differentiated cell type like the neutrophil may reflect gene transcription at precursor or even progenitor state. These earlier cell types which reside in the bone marrow have a significantly different landscape of nucleosome depleted elements [28]. Third, eQTL variants may be localised in micro-RNA sites and the sensitivity of ATAC-seq to detect such sites may be limited. Fourth, layers of gene regulation may exist outside ATAC-seq elements, which remain to be discovered. Finally, this analysis relies on the current working model that variants exert their effects through differential binding of transcription factors, which may be an incomplete model.

In summary I have shown that the vast majority of eQTLs is cell type specific, that there is considerable sharing of eQTLs between haematopoeitic closely related cell types (e.g. monocytes and macrophages; CD4 and CD8 T cells). The robustness of the eQTL data presented here is confirmed by the recapitulation of the haematopoetic tree. Finally, I obtained evidence that eQTL variants are strongly enriched in nucleosome depleted regions of the genome. In section **4.4.1** I will present the results of an analyses which explores a possible link between eQTLs and chromatin conformation.

4.3 Colocalisation with GWAS signals

A decade of GWAS has found more than 50,000 variants associated with the risk of common diseases and medically relevant quantitative traits [147]. The next big challenge after the identification of the genetic architecture of many of the common diseases is to determine through which cells and tissues these variants exert their effect on phenotype. Crucially, this involves linking the associated variants to gene(s) and molecular pathways. Multiple approaches are being exploited to functionally annotate the observed associations. Firstly, extensive genome annotation, such as patterns of cytosine methylation in CpG islands, chromatin accessibility, modifications of histone proteins and binding sites for transcription factors have been used to assign function to the non-coding fraction of the genome. Consortium based international efforts (Encode, Roadmap and BLUEPRINT) [6, 148, 149] have led to the generation of high resolution reference maps of the functional annotation of the genome for hundreds of different cell types. Using this data it was shown that GWAS variants in enhancer regions of the genome explain 19-46% of heritable variation in blood cell traits [45]. Secondly, statistical methods have been developed for the fine-mapping of associated variants [106, 108, 109, 104, 103, 107] allowing the selection of causal variants from all associated ones at a locus (see: [150]). Interestingly, some of these fine-mapping methods integrate functional genome annotations in their inference [105]. A third way to link GWAS associated variants to the gene(s) through which they exert their effect is to use colocalisation of GWAS variants with eQTL associations [73, 72, 75]. Colocalisation identifies statistically significant similarity between a GWAS association signal and an eQTL association signal. The biologically causal association can be inferred by linking the different alleles of a trait-associated variant to differences in expression levels of a gene. Additionally, the colocalisation links higher or lower levels of mRNA with higher or lower levels of the relevant trait. Linking a GWAS variant with a gene is the first step in order to map an observed GWAS association to a molecular pathway.

GWAS Cell Type	GWAS trait name	eQTL cell type
Platelet	Platelet count	platelet
	Mean platelet volume	platelet
	Platelet distribution width	platelet
	Plateletcrit	platelet
Myeloid white cell	Monocyte count	monocyte
	Neutrophil count	neutrophil
	Sum neutrophil eosinophil counts	neutrophil
	Sum basophil neutrophil counts	neutrophil
	Neutrophil percentage of granulocytes	neutrophil
	Myeloid white cell count	monocyte, neutrophil, macrophage
	Granulocyte percentage of myeloid white cells	neutrophil
Lymphoid white cell	Lymphocyte count	CD4, CD8 T-cell, B-cell
Compound white cell	White blood cell count	monocyte, neutrophil, macrophage, CD4, CD8 T-cell, B-cell
	Monocyte percentage of white cells	monocyte
	Neutrophil percentage of white cells	neutrophil
	Lymphocyte percentage of white cells	CD4, CD8 T-cell, B-cell

Table 4.1 GWAS blood cell traits selected for colocalisation with eQTLs

I have used the aggregated eQTL dataset to link GWAS variants associated with blood cell indices [45] to genes. For this analysis I have made use of a recently developed analysis approach for colocalisation developed by Verena Zuber at the Stegle group at the EMBL-EBI (Hinxton, UK). This method requires that prior to execution, regions of interest have to be defined within which colocalisation is being tested. I selected regions where GWAS and eQTL lead variants were less than 20 kb apart. As a first step of colocalisation, fine-mapping of the GWAS and the eQTL results was performed individually, using FINEMAP (version 1.1) [109]. FINEMAP is a Bayesian algorithm that calculates, for every tested variant, the posterior probability of it being causal. It takes combinations of variants into account and can find multiple independent causal variants, which correspond to conditionally independent associations. FINEMAP uses summary statistics either from the GWAS or eQTL association tests and requires a matrix of pairwise LD (correlation) of the tested variants. For this analysis, I obtained access to the GWAS genotyping data and calculated the correlation of the tested variants in the eQTL dataset and in the GWAS dataset individually. In order to estimate the posterior probability correctly, FINEMAP needs estimations of LD from the same set of

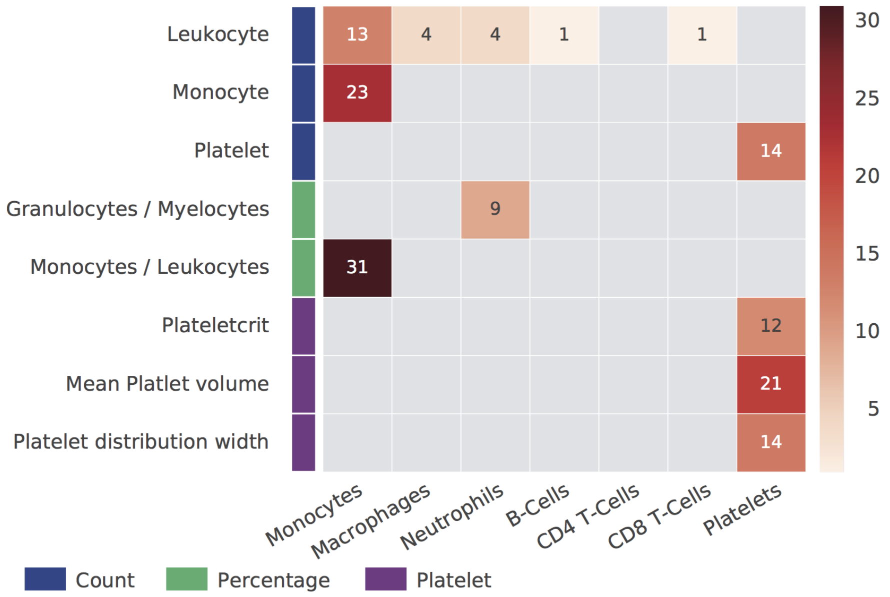


Fig. 4.5 Colocalisation of eQTLs and GWAS signals for blood cell traits. The type of blood cell trait is coded by colour in the left column. The number of observed colocalisation signals for a certain trait type are given and this number is also reflected by the colour intensity (for scale see right column). Grey boxes indicate the absence of colocalisation or that the GWAS trait was not tested for colocalisation with the eQTLs in the cell type.

samples that was used for the initial association tests. Highly correlated variants (r >0.95 or r <-0.95) variants yield numerical instabilities in FINEMAP's calculation, therefore pairwise highly correlated variant combinations had to be reduced to a single tagging variant. This was done in a uniform way for eQTL and GWAS data, always selecting the identical tag variant for one block of high-LD variants, in order to produce comparable results. After executing FINEMAP on eQTL and on GWAS data, I calculated the joint posterior probability of colocalisation $p_{joint}(SNP_i) = p_{eQTL}(SNP_i) * p_{GWAS}(SNP_i)$. This joint posterior is the probability of colocalisation for every variant. A positive result (colocalisation) was defined as $p_{joint}(SNP_i) \geq 0.5$.

For the colocalisation, 16 of 36 blood cell traits from the GWAS were selected. The combinations of eQTL cell types and GWAS traits that were tested against each other were selected on the basis of the most likely biological relevance (table **4.1**).

From those combinations of eQTL cell types and GWAS traits, a set of candidate regions was selected for colocalisation as follows: a candidate region was defined as a region ranging from one megabase up- and downstream of an eQTL gene harbouring a GWAS lead variant within 20 kb of the lead eQTL variant. The joint posterior colocalisation probability was calculated for all variants that lie in those candidate regions. Overall, 58 out of the 13,230 eQTL genes were found to be colocalising with one or several of the GWAS traits (Fig. **4.5**). As expected, monocytes (CD14) showed the highest number of colocalisations, because it is the eQTL dataset derived from the largest number of individuals (table **3.4**). Colocalisation signals were observed for all four monocyte count related traits (Fig. **4.5**).

Platelets were the cell type with the second highest number of colocalisations (Fig. **4.5**). There, the highest number of colocalisations was found for mean platelet volume. In the following section I give an example of one of these colocalisation results in greater detail.

4.3.1 ABCC4

All biological experiments on ABCC4 described in this section have been performed by Dr. Tadbir Bariana during her PhD studies.

The lead eQTL variant rs4148436 of the ABCC4 locus in platelets has a minor allele frequency (MAF) of 0.39 and the effect of size (β) of this association is 0.644 $log_2(FI)$. The ABCC4 gene is widely transcribed in blood cells (Fig. **4.6**), with extremely high levels in megakaryocytes, platelets and erythroblasts. The transcript encodes a member of the superfamily of ATP-binding cassette (ABC) transporters. ABCC4 is also a prototype for the subfamily of ABC transporters known for their role in multi-drug resistance and particularly resistance for anti-retroviral drugs used for the treatment of infection with HIV [151]. Studies with human platelets confirmed the high abundance of ABCC4 and it has been suggested that it plays a role in the transport of cyclic guanosine monophosphate (cGMP) across the platelet external membrane and possibly the membrane of dense granules [152]. Further studies suggest also a possible role in the transport of cyclic adenosine monophosphate (cAMP) and ADP [153]. These results obtained with human platelets have however not been confirmed in recent studies with platelets from Abcc4 knock-out mice [154, 155]. This discrepancy between observations made with human platelets versus murine ones is either explained by the anti-ABCC4 antibodies not being specific [152] or there are differences in the function of ABCC4 between mice and man.

Interestingly, whole genome sequencing of cases with unexplained bleeding disorders as part of the NIHR BioResource - Rare Diseases initiative identified a proband (A) to be a homozygous carrier of a mutation leading to a premature stop codon at residue 743 (p.743fs*2). Pedigree studies showed another affected member (B) to be a homozygous carrier of the same mutation (Fig. **4.7**) and a heterozygous carrier of the mutation (F) without bleeding. Sequencing of the RNA obtained from the probands' platelets confirmed reduced levels of the ABCC4 transcript in affected members (A,B) (p=0.002). The question whether the bleeding phenotype was entirely explained by the lack of ABCC4 is not answered by this

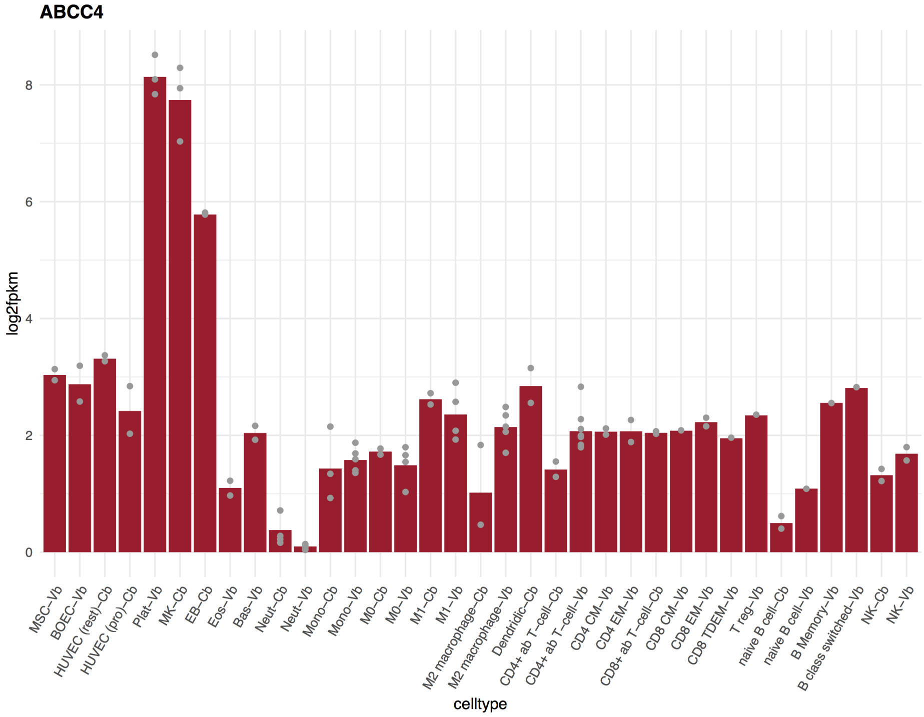


Fig. 4.6 ABCC4 transcript levels in purified human blood cell types, measured in multiple individuals (grey dots), red bars are mean $log_2(FPKM)$ measures of transcription. Values below 1 are considered background noise. Taken from BLUEPRINT RNAexpress [133]

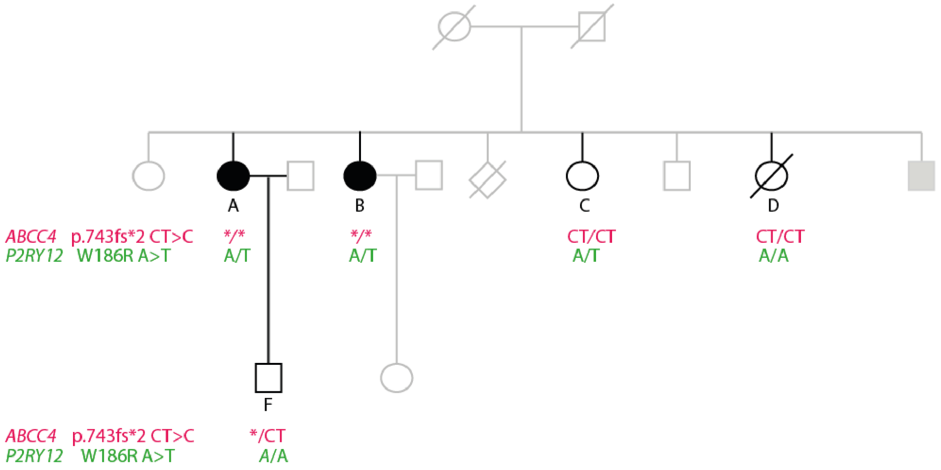


Fig. 4.7 Pedigree chart with carriers of the ABCC4 and P2RY12 mutations. Individuals affected by bleeding are indicated by filled circles. The genotype of the ABCC4 variant is indicated in pink where */* indicates homozygosity for the p.743fs*2 CT>C mutation causing a premature stop codon, */CT are heterozygous carriers of this mutation and the reference allele, and CT/CT are homozygous for the reference allele. The genotypes for the P2RY12 mutation is displayed in green, where A is the reference allele.

pedigree study because the two affected members also carry a p.Trp186Arg missense variant in P2RY12, the gene encoding one of the two G-protein coupled receptors for ADP.

In addition, pedigree member (C) also carries the P2RY12 mutation but she has no history of bleeding indicating that this mutation in isolation is not linked to bleeding. The platelets from the two affected pedigree members showed significantly increased levels of cAMP ($p < 0.0001$) and levels were normal in the other studied family members. This observation is consistent with results obtained with the platelets from Abcc4-/- mice [155]. In summary, the study in this pedigree suggests that the absence of ABCC4 is associated with bleeding, but we cannot exclude that the mutation in P2RY12 is amplifying the function defect.

GWAS for blood cell indices identified variant rs4148441 in the ABCC4 locus as being associated with platelet count [54] and this observation was replicated in the more recent GWAS [45]. The latter study had far greater power to detect associations with platelet count and also included association studies for additional indices (mean platelet volume (MPV), plateletcrit (platelet count x MPV) and the distribution width of MPV (PDW) in each individual). This analysis revealed that the effect of the SNP was limited to the count and crit of platelets. The directionality of the effect was identical for both traits and the size of the effect was very similar for both traits with β values (count: 0.035 standard deviations (SD) and crit: 0.042 SD).

Altogether, these three observations of an eQTL for ABCC4, an association signal for the count and crit of platelets and the bleeding phenotype in cases who lack functional ABCC4

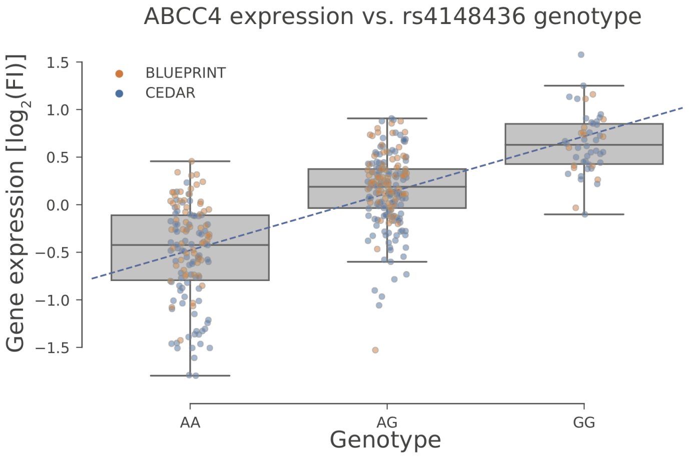
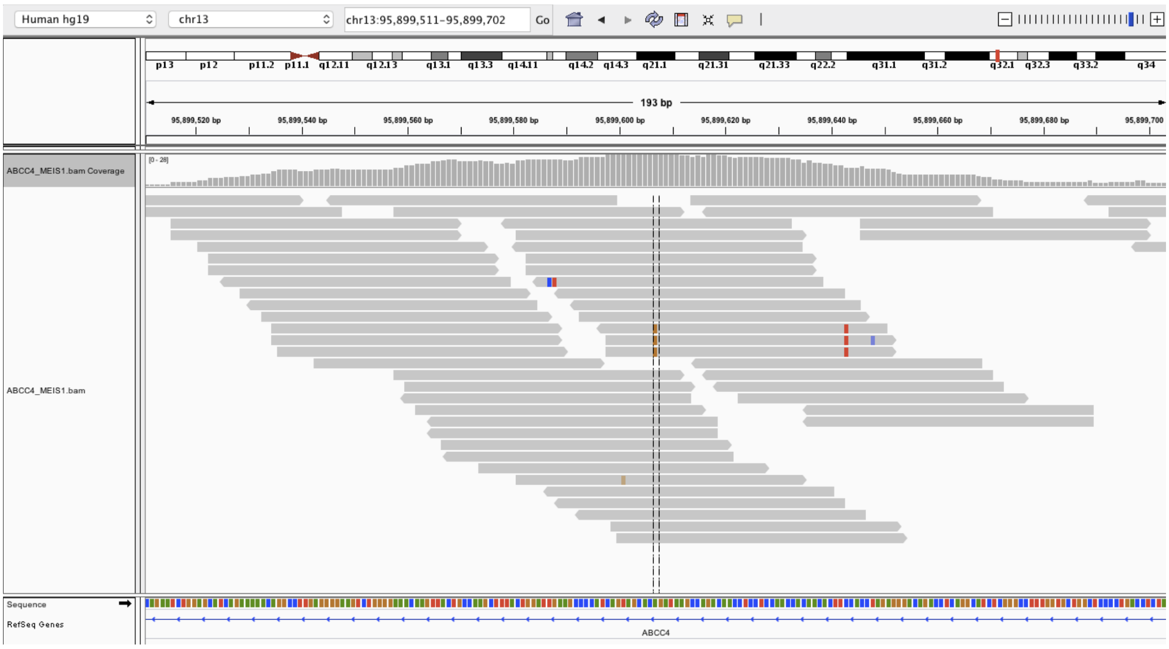


Fig. 4.8 ABCC4 transcript levels in platelets. The normalized gene expression levels are on the y axis and the results of samples for the three genotype groups for the eQTL lead variant rs4148436 are shown with the reference and alternate allele being A and G respectively. The box and whisker plots indicate the average transcript level (horizontal line within box) and its quartiles (box outline) with the β indicated by the blue line). Association $p = 3.82256x10^{-58}$, effect size = 0.6445 $log_2(FI)$.

prompted further detailed studies of the mechanism underlying the associations observed in the GWAS and eQTL study.

Firstly, in the merged platelet eQTL dataset I observed an association of ABCC4 expression levels with variant rs4148436, with the minor (G) allele being associated with a higher level of ABCC4 transcript in platelets compared to the major allele (A) (Fig. **4.8**). Secondly, to better understand the possible molecular mechanism underlying the differences in gene transcription I overlaid this region with functional annotation data for megakaryocytes, the precursor cell for platelets from the BLUEPRINT epigenome project [28]. This showed that the variant localises in a nucleosome depleted element (as indicated by an ATAC-seq signal) with positive marks for acetylation (ac) at lysine at position 27 (K27) of histone protein 3 (H3) (H3K27ac) and for monomethylation (me) at K4H3 (H3K4me), which are typical marks for an enhancer. Third, integration of annotation from binding sites for transcription factors [16, 156, 54] showed that the variant localises in a binding site for the transcription factor MEIS1 in megakaryocytes [156]. The chromatin immunoprecipitation combined with massive parallel sequencing (ChIP-seq) experiments to identify the MEIS1 binding sites was performed with the megakaryocytic cell line CHRF-288-11 [156]. Close inspection of the ChIP-seq results obtained with the CHRF cells showed the presence of both alleles at the position of variant rs4148436, which enabled the assessment of differential binding of MEIS1 to either allele (Fig. **4.9**). Therefore the original sequencing files were retrieved



Human hg19

chr13

chr13:95,899,511-95,899,702

Fig. 4.9 Differential MEIS1 binding of the A and G alleles of variant rs4148436. Screenshot of a genome browser displaying the reads of the WASP-realigned MEIS1 ChIP-seq data from Nurnberg et al. [156]. The position of the eQTL lead variant rs4148436 is in between the vertical dashed lines. Reads carrying the reference (A) allele are in grey and the non-reference ones (G) in brown at the position between the dashed lines.

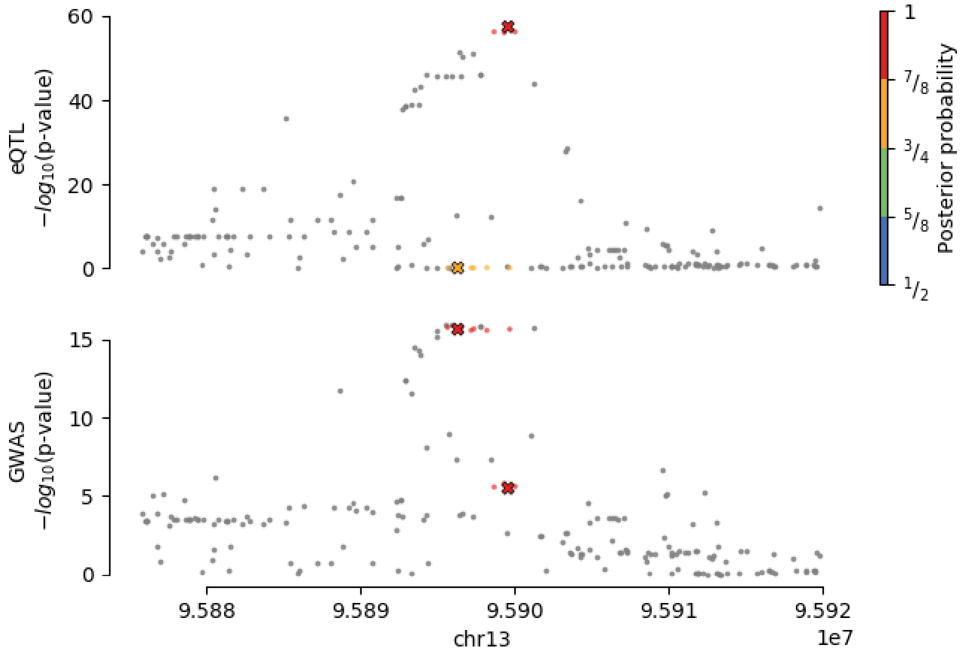


Fig. 4.10 Colocalisation of the ABCC4 eQTL and GWAS association signal for plateletcrit. Manhattan plots depict the results from the eQTL (top) and GWAS (bottom) analyses. The FINEMAP per-SNP posterior probabilities for the eQTL and GWAS results are coded from blue to red via green and orange. Variants were pruned for linkage disequilibrium (LD) before the use in FINEMAP - an X indicates the tested tag variant with a posterior probability > 0.5. p-value of association and genomic positions (in 10 Mb) on vertical and horizontal axis, respectively

and re-aligned using WASP (version 0.2.2) [157] to correct for alignment biases. After realignment read counts carrying either allele of the variant were calculated and a χ^2-test showed significant differences in binding of MEIS1 to the two alleles ($p = 2.27x10^{-7}$). The minor (G) allele, which is associated with higher ABCC4 transcript levels yielded lower binding probability of MEIS1 compared to the major (A) allele. MEIS1 is a homeobox transcription factor with a well documented and important role in haematopoiesis [158] and megakaryopoiesis [159]. Altogether, the annotation of the ABCC4 locus showed the eQTL variant rs4148436 to be the most likely driver of the differential levels of gene transcription, with reduced binding of MEIS1 being associated with higher ABCC4 transcript levels in platelets. This observation is compatible with the notion that MEIS1 has a repressive function at this particular position in the genome.

The aforementioned colocalisation analysis revealed that the lead eQTL variant of ABCC4 (rs4148436) colocalises with the genome-wide significant association of plateletcrit observed in the GWAS by Astle et al. (Fig. **4.10**). In the calculation, FINEMAP found two independent associated variants for both the GWAS and the eQTL. The weaker eQTL signal (yellow cross in the top panel) is the same variant as the stronger GWAS signal (Fig. **4.10**: bottom panel, higher red cross) and vice versa. The joint posterior probability for the lead variant in eQTLs (rs4148436, top panel, red cross) was $p_{joint} = 0.73$, which indicates colocalisation ($p_{joint} > 0.5$). Also, the p-value of the association of variant rs4148436 with plateletcrit is genome-wide significant ($5.222x10^{-10}$). From a comparison of the effect directionalities for the eQTL and in the GWAS associations it can be inferred that individuals with higher ABCC4 transcript levels of have lower plateletcrit.

Summary

In this section I have discussed an example where results from multiple association studies were integrated in order to investigate the regulation of the expression of ABCC4 in megakaryocytes and its effects on platelet phenotypes. By using extensive genome annotation data it could be shown that the lead eQTL variant has high chances of being the causal regulatory variant explaining the association observed by GWAS for plateletcrit. Additionally, the same variant causes allele-specific difference in the binding of the homeobox transcription factor MEIS1, suggesting that it acts as a repressor of ABCC4 transcription in megakaryocytes [156]. The observation that the eQTL and GWAS variants in ABCC4 colocalise strongly suggests that there is a causal relationship between plateletcrit and the levels of ABCC4 transcript present in platelets. The mechanism by which differences in ABCC4 transcript level regulate plateletcrit needs to be explored and it could be possible that the prothrombotic function of platelets is also altered by the same lead SNP.

4.4 eQTL and PCHiC

Some of the findings of this section were included in a publication which I co-authored as shared first author [160].

Promoter-capture HiC (PCHiC) is a technique to identify long-range interactions between promoters and distant DNA elements. Together with collaborators from the Fraser laboratory at the Babraham Institute (Cambridge, UK), I have characterised promoter interacting regions (PIRs) [160], where I have focused on their potential role in gene transcription regulation. The analysis approach will be reviewed and the results discussed.

4.4.1 PCHiC and data integration

Prior to any analysis, the PCHiC experimental data had to be re-annotated and prepared. Briefly, the PCHiC experiment was designed to test the long-range interactions of promoters with other regions of the genome - promoter interacting regions (PIRs). A critical step of the PCHiC is the fragmentation of DNA into defined regions (fragments) with the restriction enzyme HindIII [160]. This resulted in fragments of an average length of 4 kilobases, thereby defining the resolution of the assay. To achieve high sensitivity for the detection of PIRs, interacting fragments were enriched by a pull-down step before sequencing using all annotated promoters (Ensembl v.75) as baits [160]. Pulled-down HindIII fragments and their interacting elements were sequenced and any fragment interacting with a promoter bait was called a PIR. The PCHiC experiment was performed with 17 primary haematopoietic cell types. As part of preprocessing, interactions measured between directly adjacent fragments were removed from further analysis. This step was necessary to maintain specificity of interactions because HindIII digestion may be incomplete due to ineffective restriction. As a result interactions between a promoter bait and its adjacent 5' and 3' fragments cannot be measured, creating a so-called "blind spot" for interactions around the promoter.

In the initial analyses some annotation inconsistencies were detected in the first version of the dataset. I therefore re-annotated bait regions with gene names, by overlapping them with the transcription starting sites (TSS) of genes as defined in Ensembl v.75. Some baits covered the TSS of multiple genes and were therefore annotated with multiple gene names. This re-annotation was used in all of the following sections.

4.4.2 eQTLs indicate regulatory function of PIRs

The existence of a three-dimensional interaction (PIR) does not necessarily convey an expression regulatory function. Identifying eQTL variants in PIRs provides direct evidence of functionality. The largest portion of eQTLs is localised proximal to the TSS (Fig. **4.12c**). The chance of identifying eQTL variants in genomic regions further from the TSS - beyond the blind spot of PIRs - is therefore limited. Notwithstanding I analysed the presence of eQTL variants in PCHiC PIRs.

Several analyses were performed.

PIRs as a prior for eQTLs

For a typical eQTL analysis the association test is limited to 1 Mb windows on either side of the gene body (Fig. **4.11**). The PCHiC data allowed to restrict the test to subsections of the aforementioned window and the inclusion of more distant promoter-interacting elements

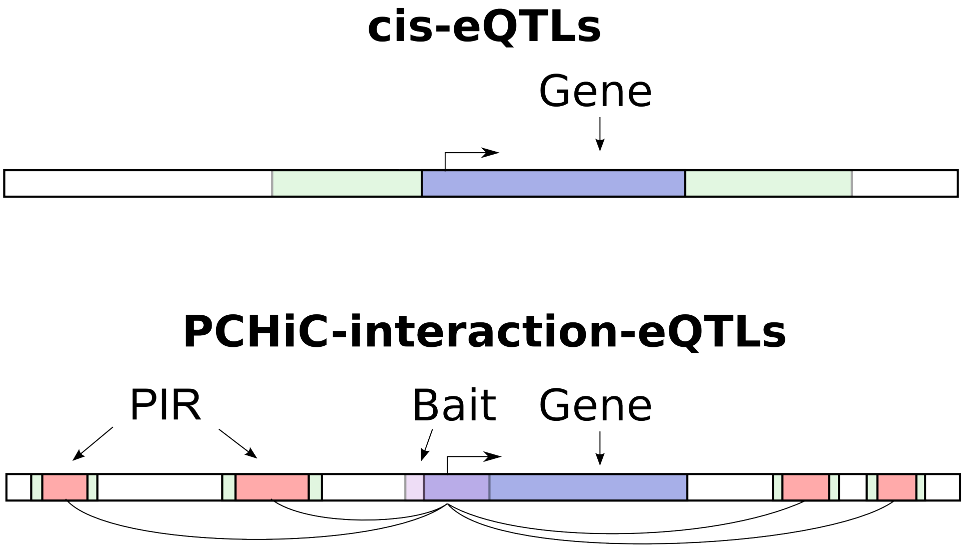


Fig. 4.11 Schema of PCHiC-based eQTL calling. Green areas: regions from which variants are selected for an association test with the gene highlighted in the respective schema; red areas: PIRs; transparent red areas: bait fragments; blue areas: gene bodies; bent arrows: gene TSS. On the top panel the default approach for cis-eQTL tests is displayed - all variants within 1Mb up- and downstream of the gene body (including the gene body) are tested. On the bottom panel the PCHiC interaction-based approach is displayed - all variants localising in PIRs (±500bp) are tested against the transcript levels of the gene. Proportions of displayed annotations are not to scale.

which are localised outside this window. I therefore tested whether PIRs are informative priors for eQTLs and calculated the enrichment of identifiable eQTLs in PIRs. To achieve this the typical cis-eQTL testing approach was modified and instead of testing the variants in a window greater or equal to 2 Mb, only variants localised in the gene-specific PIRs were used for association testing. The PIRs were extended by 500 bp on either side in order to capture variants lying just outside the HindIII fragments, accounting for arbitrariness of the restriction sites of HindIII in the gene regulatory context. Additionally, to allow for a consistent randomisation procedure (see below) PCHiC interactions further than 1 Mb from the promoter bait were excluded from the following analysis. The analysis was performed with cell type-matched interaction and eQTL data and the association test was performed using a 10% FDR threshold. This analysis procedure identified 899 and 577 eQTLs in PIRs for monocytes and B-cells, respectively. The lower number of eQTLs for the latter type of cell reflects the smaller eQTL sample set for B-cells compared with monocytes (n=432 vs n=283). To determine the significance of these findings the observed number of identifiable eQTLs was compared to the number of eQTLs obtained from a randomisation of PIRs.

For this analysis a set of random PIRs (rPIRs) were generated using the following analysis approach:

1. Sample all bait-PIR pairs as prepared for the first association test and shuffle the gene names of the baits.
2. Remap the bait and its interactions to its newly assigned gene promoter whilst maintaining distances between the bait and its PIRs.
3. Apply mirroring of the genomic positions of rPIRs around the bait if the newly assigned gene lies on the opposite strand compared to the original gene

The randomisation was performed 1,000 times and association tests were performed for every iteration. Enrichment of associations in the observed PCHiC data over rPIRs was calculated at various distances from the TSS. The results of the analysis (Fig. **4.12a** and Fig. **4.12b**) show a significant enrichment of eQTLs in promoter-interacting elements identified by PCHiC ($p < 0.001$). The following conclusions can be drawn from this experiment. First, the portion of eQTLs localised in PIRs is relatively small compared to the total number of eQTLs for a cell (e.g. the typical analysis showed eQTLs for 3,686 genes in monocytes and only 899 eQTLs in PIRs were identified). This relative low fraction is because most eQTL associations are localised in a narrow window around the TSS (Fig. **4.12c** and Fig. **4.12d**). Because of this proximity of eQTLs to the TSS and the PCHiC blind spot a

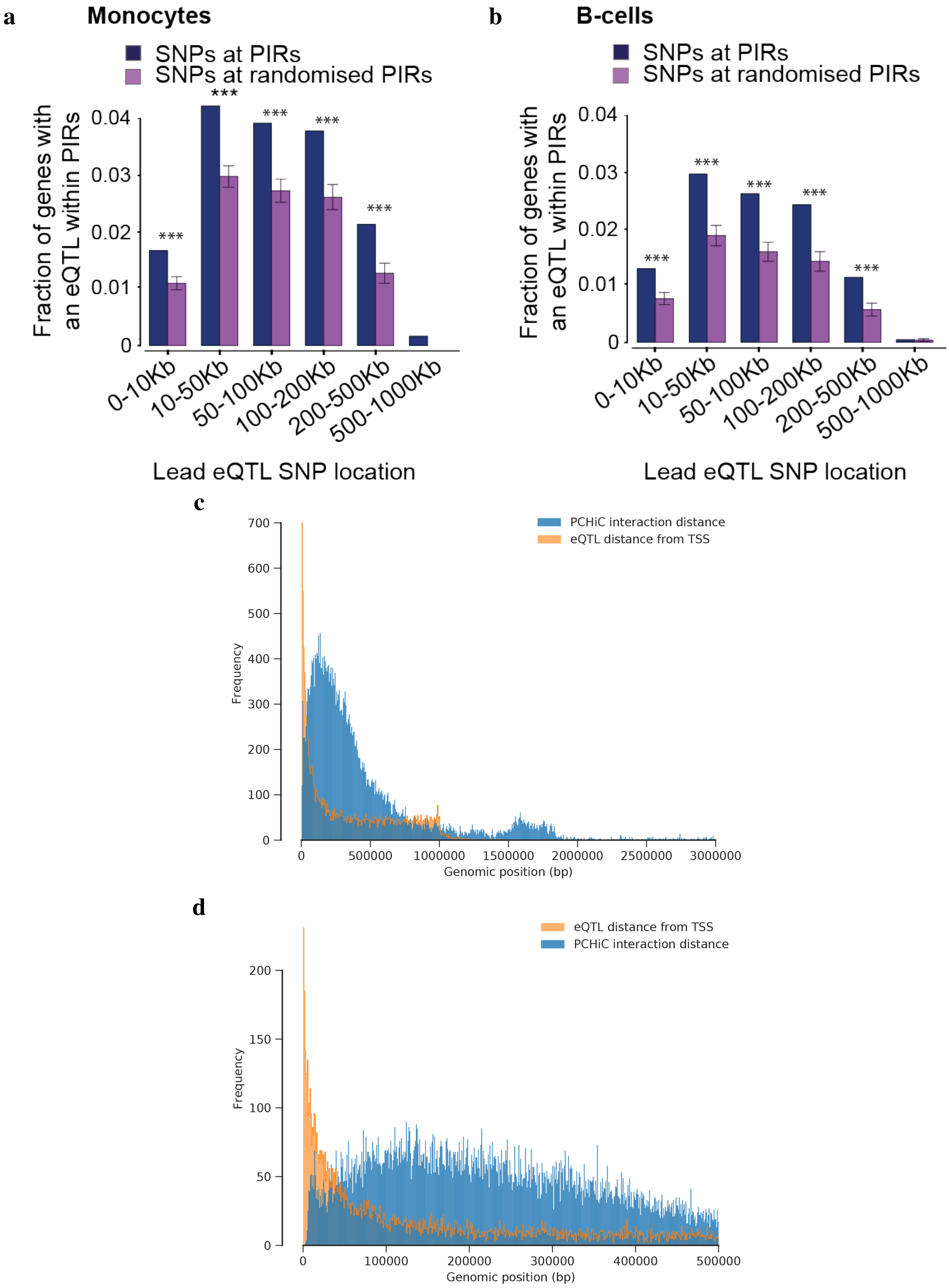


Fig. 4.12 Localisation of eQTL lead variants and their enrichment in promoter interacting elements (PIRs). Fig. **4.12a** and Fig. **4.12b** display the enrichment of eQTL in PIRs in monocytes and B-cells, respectively. Relative enrichment and distance between lead variant and TSS is displayed on the vertical and horizontal axis, respectively. $*$ indicates significant enrichment over rPIRs obtained by permutation testing with $*p < 0.05$; $**p < 0.01$; $***p < 0.001$. Fig. **4.12c** Distribution of distances from TSS to eQTL lead variant for monocytes (in orange) and from the bait centre to PIR centre in blue. Fig. **4.12d** A zoomed-in version of Fig. **4.12c**.

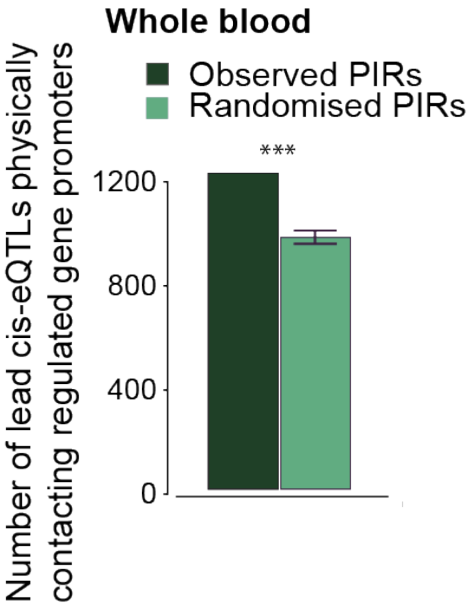


Fig. 4.13 Enrichment of cis-eQTLs in promoter interacting elements using whole blood eQTL data. $***$ indicates significance of enrichment at $p < 0.001$ for observed over randomized PIRs as obtained by permutation testing.

large portion of enhancers containing eQTLs cannot be observed. Secondly, the number of identifiable eQTLs in PIRs reduces rapidly with increasing distance from the TSS.

Overall, the analysis has shown a significant enrichment of eQTLs in PIRs indicating that a proportion of the long-range promoter-interacting elements are functional.

Enrichment of whole blood eQTLs in PIRs

The results of eQTL enrichment presented above are based on the Wellcome Trust Centre for Human Genetics (WTCHG) datasets for monocytes and B-cells, which was based on the analysis of expression and genotyping date of only 432 and 283 individuals, respectively. The enrichment analysis was repeated with an eQTL meta-analysis dataset generated with RNA samples obtained from whole blood of individuals [101]. This study comprised data from 5,311 samples and identified 6,418 eQTLs. There are several explanations for the far lower number of eQTLs detected in this meta-analysis compared with the results from the eQTL analysis presented in this thesis (section **3.2.2**).

The interaction data was merged across all 17 blood cell types [160] for the analysis of the whole blood eQTL study. The same PIR randomisation procedure as mentioned above was applied and an enrichment of eQTLs in PIRs over rPIRs was also observed for the whole blood eQTL study (Fig. **4.13**).

Examples of eQTLs acting over megabases distances

Having established that there are eQTLs exerting their effects on the expression of genes which are far from the gene body, I searched for additional illustrative examples of eQTLs in PIRs which were beyond the arbitrarily chosen 1 Mb window around the gene body. For this analysis I used the full set of association data instead of the restricted one used for the enrichment analysis. Amongst many others, I detected two eQTLs localised in PIRs, both of which exert an effect on the transcription of the same NCOA4 gene (Fig. **4.14**); one of the eQTL variants is 5 Mb upstream of the TSS of NCOA4. This type of long-distance eQTLs are not detected in the typical cis-eQTL analysis, and would most probably have been missed in a trans-eQTL analysis due to their relatively low association strength. Another interesting example was identified in B-cells where the PIR-localised eQTL variant rs17561058 regulates the transcript levels of both NDUFAF4 and ZBTB2 (Fig. **4.15**). This associated variant lies at distance greater than 10 Mb from both these genes. The PCHiC interaction data suggests that the associations with both genes may be in cis but further biological experiments are required to confirm this assumption.

In conclusion, using PIRs as priors reduces the number of performed eQTL association tests by reducing the genomic space from which variants are sampled. This in turn reduces the multiple testing correction that has to be applied to the association p-values. Having knowledge of these functionally relevant long-range interactions is important to improve the accuracy by which disease and trait associated variants identified by GWAS are linked to candidate genes and molecular pathways.

4.4.3 Variation in transcript levels and promoter connectivity

The variation in the transcript levels within the population varies widely between genes. The associations observed in the typical and PIR-based eQTL analyses explain a part of this variation. As discussed before, the contribution of eQTL variants in regulatory elements far from the TSS remain unobserved because of the limitations imposed by the correction for multiple testing. Furthermore, many trans-acting eQTL variants cannot be identified because of lack of power for most eQTL studies, which have been performed in relatively small samples of individuals. To explore other possible parameters regulating the variation in transcript levels I posed the following questions:

- Is there a correlation between the level of connectivity of a promoter, its interacting elements, and the variance in transcript levels?

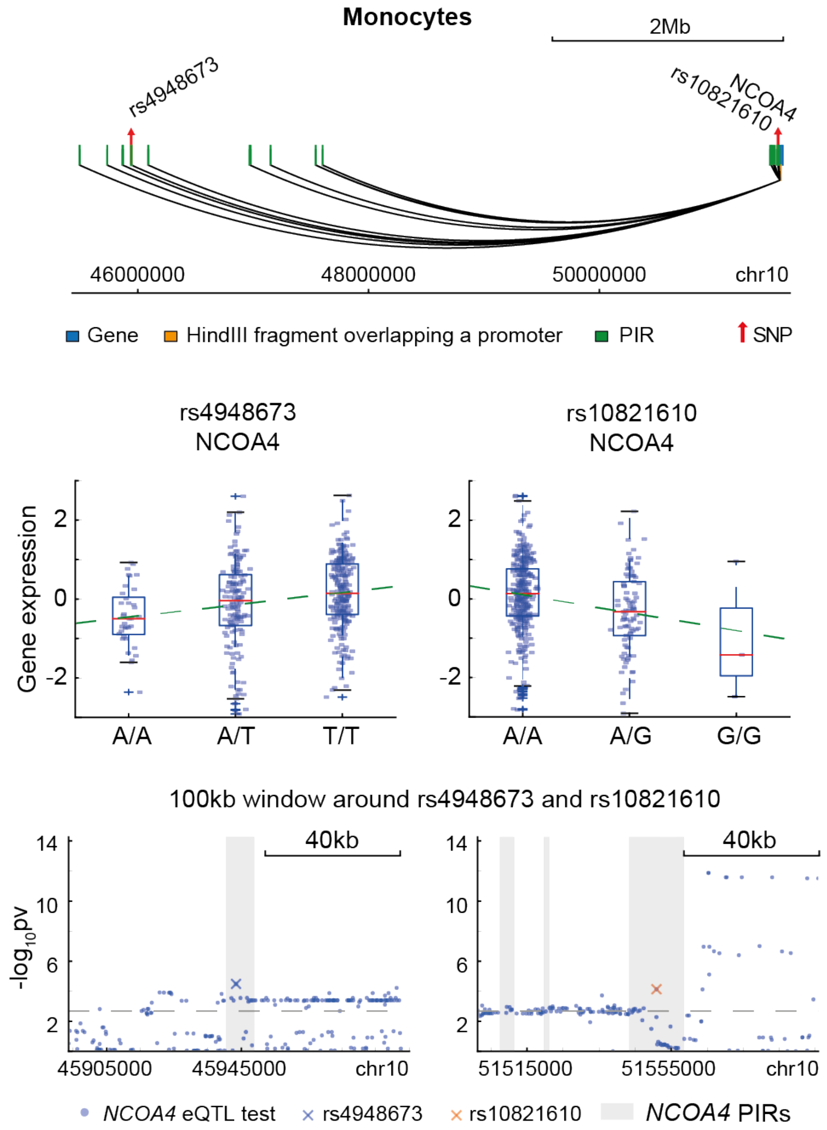


Fig. 4.14 Two independent eQTL variants exerting an effect on the transcript levels of the gene NCOA4 in monocytes. Top panel: A schema depicting a 5 Mb element on the long arm of chromosome 10 from 5’ to 3’, highlighting the eQTL variants rs4948673, rs10821610 and the gene NCOA4. Middle panel: Association between genotypes and NCOA4 transcript levels in monocytes for eQTL variants rs4948673 (left figure) and rs10821610 (right figure). The box and whisker plots indicate the transcript levels as function of genotyope; median transcript level in red, the effect size in green (both in $log_2(FI)$) and the first to third quartile as box. Bottom panel: Manhattan plots depicting the two eQTL variants with nominal p values of association on the vertical axis and chromosome position on the horizontal axis. The horizontal grey dashed line indicates the significance threshold.

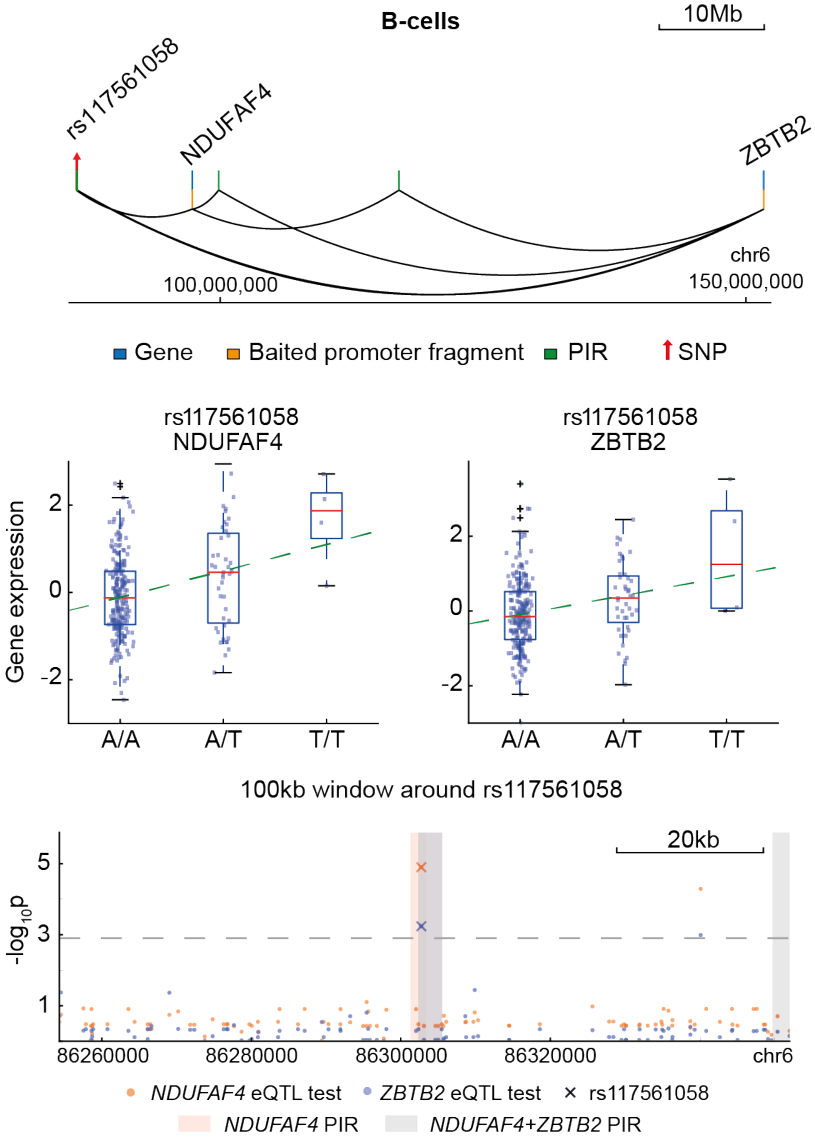


Fig. 4.15 An eQTL variant exerting an effect over the transcript levels of two genes in B-cells. Top panel: A schema depicting a 53 Mb element on the long arm of chromosome 6 with depicted from 5' to 3' the eQTL variant rs117561058 and the genes NDUFAF4 and ZBTB2. Middle panel: Association between genotypes of rs117561058 and transcript levels in B-cells for NDUFAF4 (left figure) and ZBTB2 (right figure). The box and whisker plots indicate the transcript levels as function of genotyope; median transcript level in red, the effect size in green (both in $log_2(FI)$) and the first to third quartile as box. Bottom panel: Manhattan plots depicting the eQTL signals for NDUFAF4 and ZBTB2 with p value of association on the vertical axis and chromosome position on the horizontal axis. The horizontal grey dashed line indicates the significance threshold.

- Is there a correlation between the level of tolerance for loss of function variants and the variance in transcript levels?

The results of the experiments are presented in section **4.4.3** and section **4.5**.

Before testing the relationships, gene expression variance data had to be preprocessed. Variance of gene expression across samples is confounded with the mean gene expression level (mean-bias, Fig. **4.16a**). This bias is particularly strong for genes with low expression values compared to the variance at higher expression levels. Three approaches were reviewed to reduce mean-bias and the most promising was selected by visually comparing the variance after data preprocessing with the uncorrected variance. The coefficient of variation and a normalisation by lowess regression were computed with two different parameter settings. An implementation of the lowess regression in the statsmodels python package (version 0.8.0) [161] was used with a fraction parameter of 0.1, indicating that 10% of the dataset was used to fit the model. The mean-bias in the variance data was reduced by calculating the residual after the lowess regression (Fig. **4.16b**). In spite of the normalisation a mean-bias remained but the effect of the mean expression on the bulk of the expression variance was successfully reduced.

The number of elements interacting with promoters varies (Fig. **4.17**). Less than 37% of promoters have more than 10 interacting elements. Inspection of the PCHiC interactions for the ABCC4 promoter shows 54 and 34 interactions in megakaryocytes and erythroblasts, respectively but the number of interactions is distinctively lower in other myeloid and in lymphoid cells. Using this interaction data the normalised gene expression variance was tested for correlation with the number of interactions with gene promoters under the assumption that the variance of transcript levels was inversely correlated with the number of promoter interactions. The number of PIRs (interactions) per gene was calculated from the PCHiC data. The number of interactions was not found to be correlated with the normalised expression variance data (Fig. **4.18**) (r=0.006904).

eQTL effect size and promoter connectivity

An alternative approach to explore whether promoter connectivity influences gene transcription is to consider the effect sizes that were calculated for the eQTLs. In contrast to the expression variance, the magnitude of the effect (β) is an indicator of the genetically explainable variation of gene expression. Therefore, higher β values indicate that the expression of a gene is modulated more strongly by a cis-acting eQTL variant. When testing the correlation between the β values and the number of PIRs per gene no apparent correlation was observed for the dataset generated for monocytes.

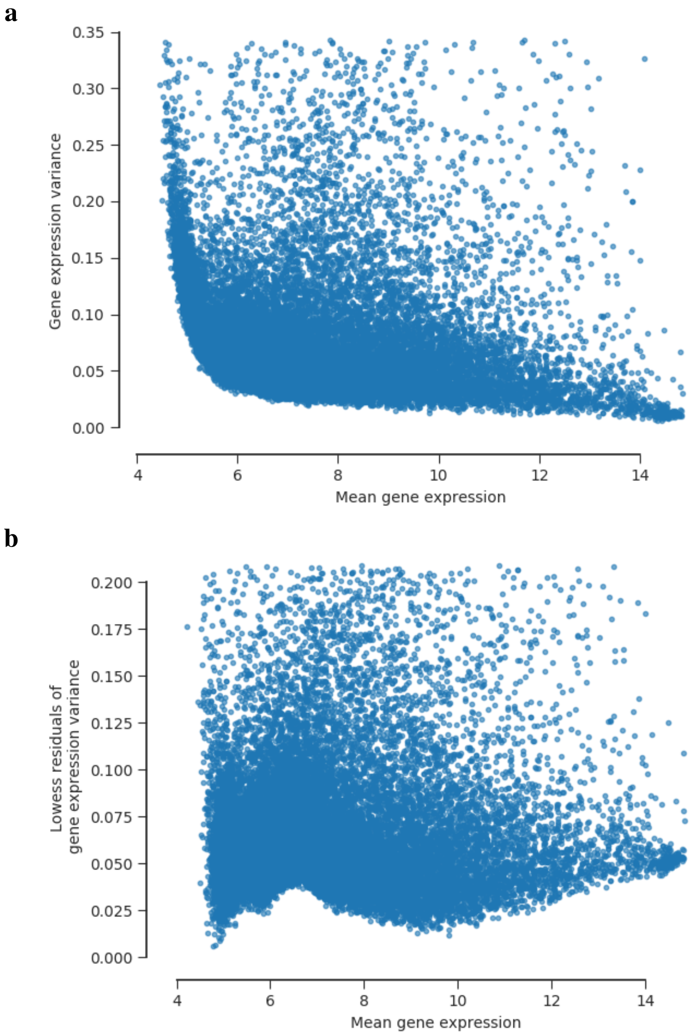


Fig. 4.16 Normalisation of estimated gene transcript levels by lowess regression. Fig. **4.16b** The mean gene expression levels $[log_2(FI)]$ for all array probes and the gene expression variance levels are plotted on the horizontal and vertical axis, respectively. Fig. **4.16b** The mean gene expression levels $[log_2(FI)]$ against the residuals of gene expression variance after regressing out the mean-bias effect with lowess regression. Gene expression data for monocytes from the WTCHG eQTL study were used.

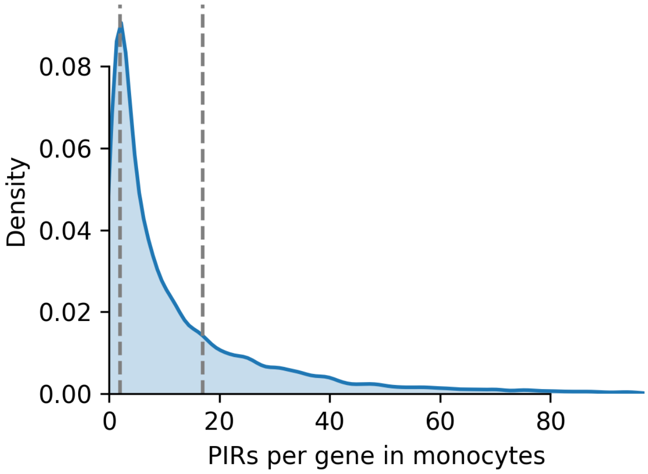


Fig. 4.17 Histogram of PIRs per gene in monocytes. The Histogram is truncated at the 99th percentile (97 PIRs per gene) for display purposes. At maximum there are 411 PIRs per gene. Grey vertical lines indicate the first and third quantile.

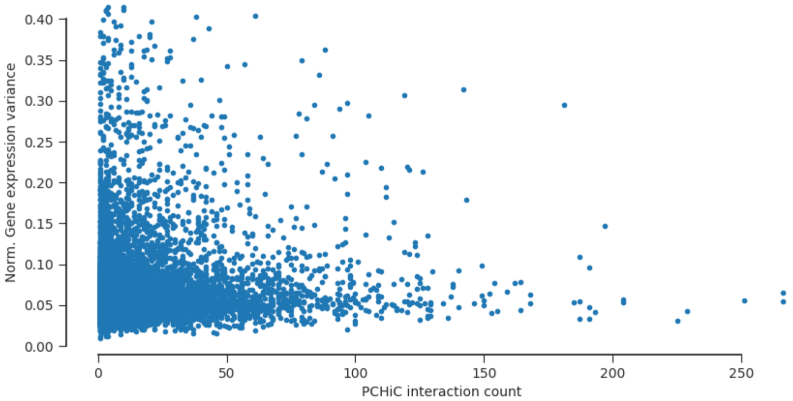


Fig. 4.18 Correlation of normalised gene expression variance with the number of PIRs per gene. Lowess-normalised gene expression variance on x axis; PIR count on y axis.

Summary

In summary the variance in expression levels is not correlated with the number of connections of the promoters with their interacting elements. There are several possible explanations for this lack of correlation. Firstly, not all interactions are observed because of the blind spot of the PCHiC experiment and there is therefore an underestimation of the number of interactions. Secondly, the level of promoter connectivity may be an important factor for the spatiotemporal regulation of expressing but has no bearing over the levels of variance. Nevertheless, I have observed eQTLs in promoter interacting elements validating the function of hundreds of promoter interacting elements, thereby confirming their role in gene regulation. Further studies are now required to better define the portion of nucleosome-depleted elements which are functional.

4.5 Gene constraint scores and transcript variant levels

A loss of function (LoF) variant (e.g. premature stop codon, alternative splice variant) on one allele is tolerated for a large portion of genes. This was exemplified by the sequencing of individuals with a high level of consanguinity in which 1,317 genes with LoFs on both alleles were identified [162]. At the other end of the spectrum are genes which do not tolerate single alleles of LoF variants because of their detrimental effect on an individual's fitness and / or ability to reproduce. The sequencing of the coding fraction or so called exome in more then 60 thousand individuals has provided an opportunity to calculate a score of LoF intolerance of thousands of genes longer than 300 amino acids [163]. The pLI score ranges between 0 and 1 with many transcription factors having a high score (FLI1, 0.82; GATA1, 0.84; MEIS1, 0.99; TAL1, 0.82) and genes reviewed in this chapter showing a range of pLI values (ABCC4, 0.00; NCOA4, 0.01; NDUFAF4, 0.48; ZBTB2, 0.98). The pLI score is of great value when labelling high impact rare coding variants with pathogenicity level information, e.g. a premature stop codon on one allele in a gene with a high pLI score (e.g. GATA1) is more likely to cause pathology compared to a gene with a low value (e.g. ABCC4).

Genes with high pLI values are protected from acquiring LoF mutations and are under negative selection (*purifying selection*) to remove variants introduced as a consequence of naturally occurring variation. It is therefore reasonable to postulate that genes with a high pLI have a lower variance of transcript levels than genes with a low pLI.

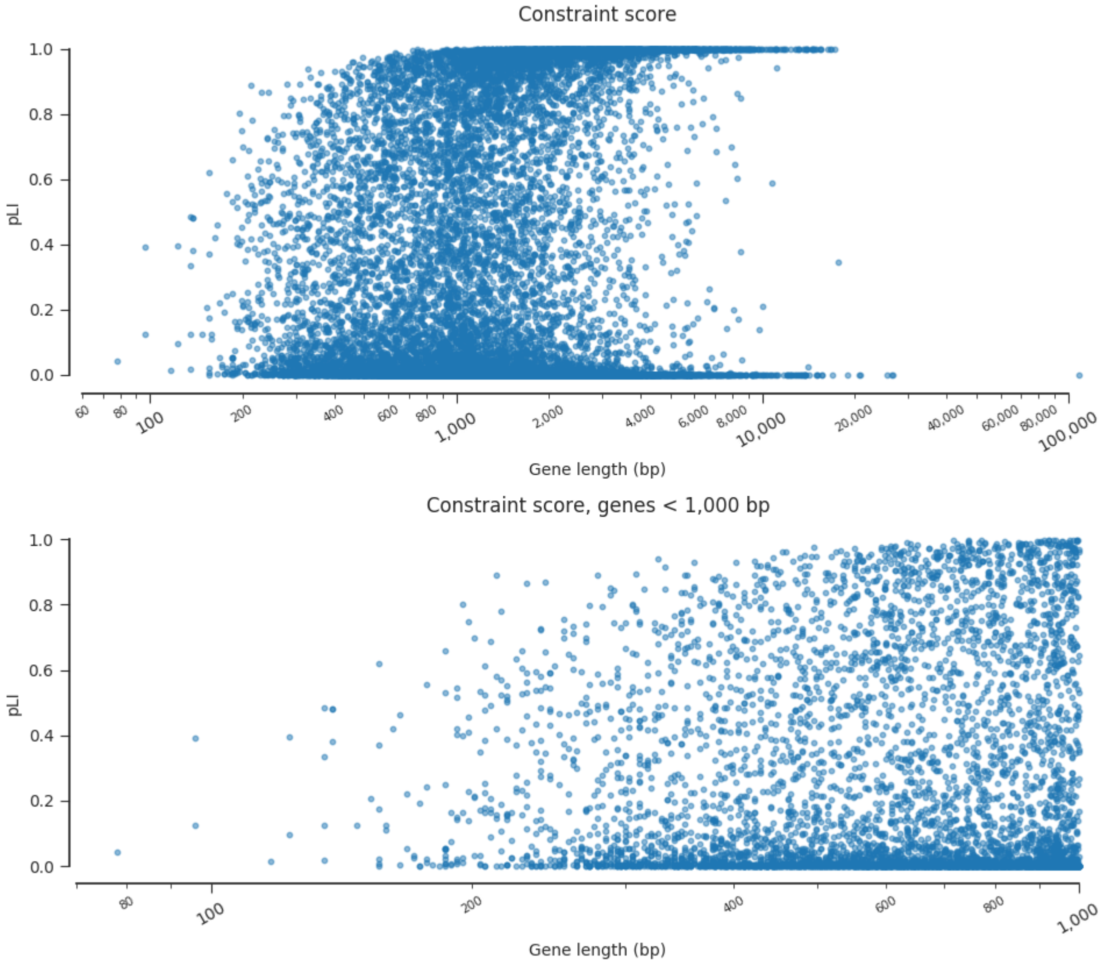


Fig. 4.19 Correlation between pLI score and length of the canonical transcript of a gene. The pLI scores and gene (transcript) length in base pairs (bp) are on the vertical and horizontal axis respectively. Top and bottom panels show results for all genes and for genes of < 1000 bp in length.

4.5.1 The pLI score

There is a correlation between the pLI score and the length of the transcript [163]. The distribution of pLI scores was plotted against gene length, showing that for short genes the pLI value does not reach probabilities of 1.0 (Fig. **4.19**, upper panel). It is not plausible that all short genes have low pLI scores (Fig. **4.19**, lower panel) hence the correlation analysis between pLI scores and variation in transcript levels was limited to genes with a canonical transcript length greater than 1 kb. It is noteworthy that due to the calculation of the pLI score genes with longer transcripts tend to have pLI values at either of the extremes (Fig.Fig. **4.19**, upper panel) compared to genes with shorter transcripts.

4.5.2 eQTL effect size and pLI

Interestingly, an analysis using GTEx [77] eQTL data by Lek et al. showed that there is an enrichment of eQTLs in genes with a low pLI score and likewise a depletion of eQTLs in genes with high pLI scores (Fig. 3d in [163]). In a further analysis I explored whether the effect sizes of eQTL associations were correlated with pLI scores. The analysis was expanded from the binary question of association or no association to whether the order of pLI scores was related to the order of eQTL effect magnitudes. The results of this analysis using data from seven types of blood cells are shown in Fig. **4.20**. For all cell types a negative correlation between eQTL effect magnitudes and the pLI scores were observed. The evidence of this inverse correlation was strongest for the monocyte eQTL dataset which was based on the largest number of individuals. The result from this analysis is in concordance with the analysis by Lek et al. using the GTEx data [163]. In a final step an analysis was performed on the correlation between normalised gene expression variance data (section **4.4.3**) and the pLI score and no association was detected.

4.5.3 Summary

In agreement with Lek et al. [163], I observed a relationship between eQTLs and the probability of LoF intolerance. I extended their analysis by showing that the order of eQTL effect magnitudes is negatively correlated with the order of pLI values. From this observation I conclude that genes which are more tolerant for LoF variants tend to have higher eQTL effect sizes across a wide range of common and low-frequency eQTL variants (any variant with MAF > 1%). The observed correlations for the different types of cells between effect size and pLI value is mainly driven by genes with pLI scores close to 0 or 1 (Fig. 3.16a). I

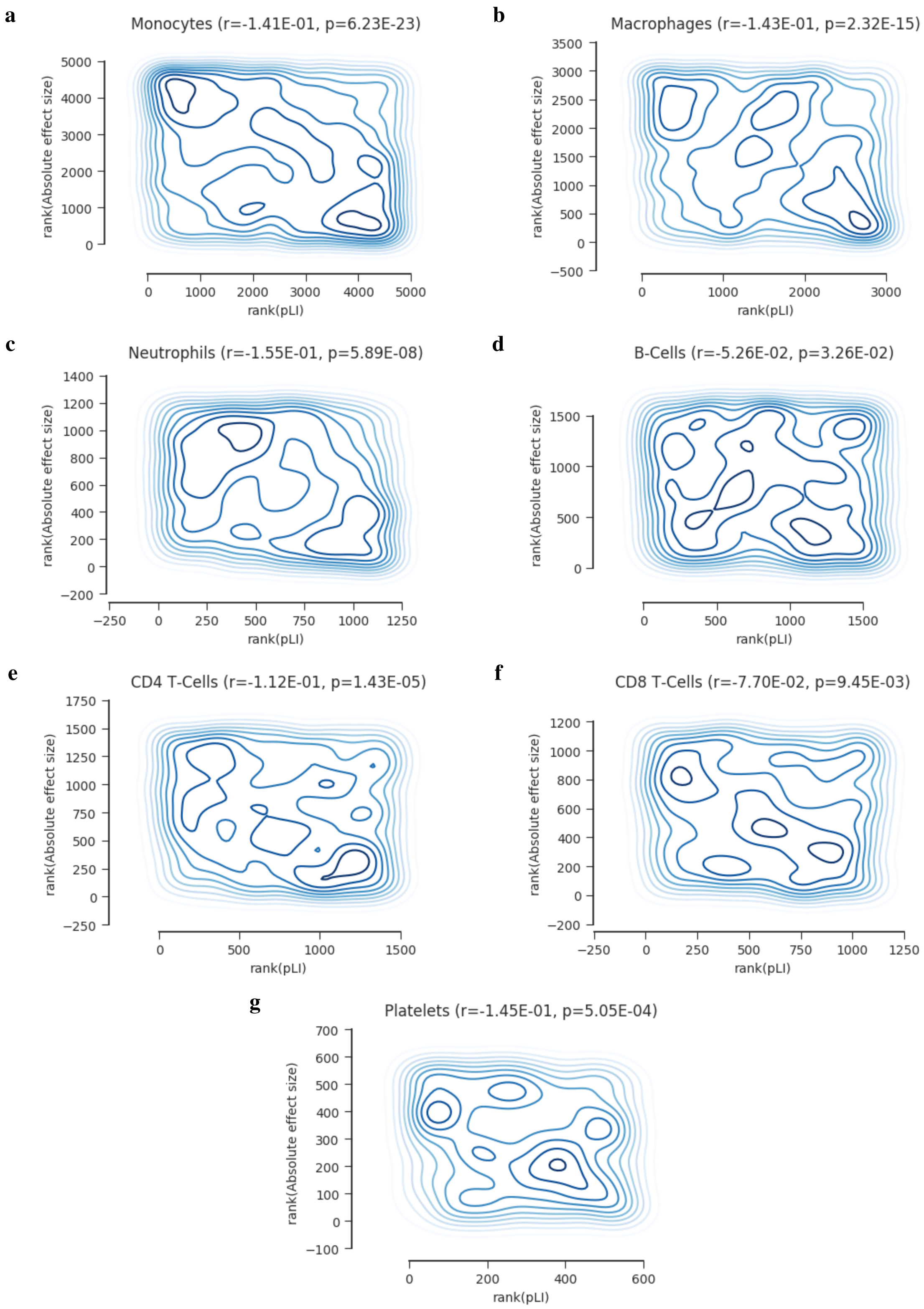


Fig. 4.20 Kernel density estimation plots of the correlation between ranks of the eQTL absolute effect sizes and the ranks of the pLI score. The effect size $[log_2(FI)]$ and pLI values are on the vertical and horizontal axis, respectively. Negative correlations are observed for all seven types of blood cells (panels **a-g**) and the correlation values (r) and p-values are given in brackets.

did not find a consistent relationship between pLI values and normalised gene expression variance for the different cell types.

4.6 Discussion

In this chapter I have annotated, interpreted and applied eQTL results from Chapter chapter **3**. I have shown that by using eQTL data of seven different types of blood cells the process of haematopoietic lineage commitment can be recapitulated. This observation is compatible with the notion that a large portion of the eQTLs are localised in cell type specific regulatory elements and only exert their effect in a single type of cell. In contrast I have also highlighted the existence of shared eQTLs exerting their effect across different types of blood cells. It has previously been shown that the effect direction can switch between cell types by comparing the eQTL data from monocytes and B-cells [87]. I replicated and expanded on these findings by showing a larger set of eQTLs where the association of the same variant and the same gene had opposite effect directions in different cell types.

eQTLs are empirically identified regulatory variants. Given the current understanding of gene regulation one would expect them to localise in regions, that are biologically marked as gene regulatory important, such as accessible chromatin, enhancers and transcription factor binding. I observed a significant enrichment of eQTL variants in regions of accessible chromatin and after an extension of the lead variants with their LD-proxies approximately 40% of eQTLs mapped to a nucleosome-depleted region identified by ATAC-seq. This observation is corroborating earlier findings [141]. The mechanisms by which the remaining eQTL variants affect gene regulation and transcription requires further research. Possible areas of interest are the role of methylation of CpG islands, micro-RNAs and the possible effects of haplotypes on 3-dimensional chromatin configuration. It may also be possible that there is a lack of sensitivity of ATAC-seq and similar methods like DNase1 hypersensitivity profiling in identifying nucleosome-depleted elements.

Independently of how a variant modulates gene expression, eQTL data can be used as a means to aid the functional annotation of DNA variants found by GWAS to by using colocalisation. Thousands of such variants have been identified since the completion of the first comprehensive GWAS study in 2007 [44]. Most of these variants ($> 90\%$) are localized in the non-coding space and establishing a reliable link between variants and genes is critically important for functional follow-up studies. eQTL variants are unequivocally linked to a gene, or sometimes more than one gene. Colocalisation of eQTL and GWAS variants is therefore an effective approach to functionally annotate GWAS variants. By applying a colocalisation analysis, I successfully linked 60 of 2,069 variants identified in a recent GWAS study for blood indices [45] to genes. One of these colocalisation signals occurred at the ABCC4 locus, where the plateletcrit associated variant rs4148441 colocalises with different levels of the ABCC4 transcript in platelets. These differences in transcript level were correlated with different levels of MEIS1 binding at the regulatory element in which the lead eQTL variant was localised. The allele of variant rs4148436 with a lower level of signal in the ChIP-seq experiment for MEIS1 binding was associated with higher ABCC4 transcript levels, possibly indicating of MEIS1 having a repressive role at this position [156]. The observation of a colocalisation of the GWAS and eQTL variant confirms the assumption from studies with human and murine platelets that the transporter ABCC4 plays an important role in platelet biology.

In recent years it has become possible to query long range interactions between different regions of the genome. Using the specialised PCHiC technique the interactions of promoters with distant elements were determined in 17 blood cell types. To assess the regulatory importance of promoter interacting regions (PIRs), I used eQTL data to show a significant enrichment of regulatory variants in PIRs [160]. I showed that it is possible to use an alternative definition of cis-eQTLs, where instead of defining proximity using linear distance between a gene's TSS and a variant, PIRs were used. Associations that putatively act in cis, spanning up to 10 Mb between the eQTL variant and TSS, could be identified. In addition to having identified a large number of the long-range eQTLs, this finding also validates the function of hundreds of PIRs. The question whether associations found within PIRs are truly cis acting - exert their effect directly on the expression of a gene - can only be answered by additional biological experiments.

Finally, I used the effect size of eQTL associations as a proxy for the relative tightness[1] of the regulation of gene transcription and also related it to the gene's probability of LoF intolerance (pLI) [163]]. A high impact coding mutation present on a single allele of a gene with a high pLI value is more likely to lead to changes in phenotype if compared with similar events in genes with a low pLI. The presence of such mutational events present on a single allele of genes encoding transcription factors (ETV6, FLI1, GATA1, GATA2, MECOM, RUNX1) are associated with abnormal haematopoiesis. The observed inverse correlation between pLI values and the eQTL effect magnitudes for all cell types and the relative depletion of eQTLs from this category of loci with high pLI scores corroborates the notion that LoF mutations on both alleles are generally incompatible with viability of the organism. The results from the analysis presented here extend on the observation by Lek et al. [163].

[1] Phenomenologically, tightly controlled genes are regulated to have one level of expression, allowing little variation [164]

Overall, the findings presented in this chapter show the wide utility of eQTL datasets obtained with purified populations of blood cells. I illustrated the utility by several example experiments and I have made the datasets available to others to perform colocalisation studies with variants identified in GWAS for cardiovascular and other diseases.

Chapter 5

Genetic variant effect prediction using deep learning

The software presented in this chapter was developed in collaboration with Ziga Avsec, PhD student in the Gagneur group at the Technical University of Munich, Germany. It is published as a pre-print, which I co-authored as shared first author [165]. Throughout the chapter I will highlight which sections were developed by whom to clarify my contributions.

In this chapter I will present *Kipoi*. This software has been designed to facilitate sharing and re-use of trained machine learning models in genomics. I will motive its development and focus on explaining parts of the software that I implemented, such as the effect estimation of DNA variants. Additionally, I will describe tools and features I have implemented to visualise datasets using machine learning models.

5.1 Deep learning in genomics

Technological advances in DNA sequencing and associated high-throughput assays have catalysed the generation of large scale genomic and molecular profiling studies. The flood of data has created the need for powerful computational techniques that can capture complex relationships in high-dimensional data. *Predictive statistical* and *machine learning models* are increasingly developed and applied to genomic data.

In statistics and machine learning a wide range of techniques and approaches have been developed to model data. A model class that has recently gained considerable attention is the *neural network*. It gained popularity in fields like image- and speech-recognition due to its plasticity in modelling data, because of its capabilities to capture non-linear relationships. In genomics the first examples of neural networks have been published starting from 2001 classifying cancers based on gene expression profiles[166–170]. Following from these early pioneering contributions, a number of lessons have been learned and model architecture as well as training data selection have been optimised. State-of-the-art models can very successfully predict transcription factor binding, chromatin accessibility, RNA splicing efficiency [171–175], and a wide range of other molecular traits from inputs such as DNA sequence. These models are trained for example on transcription factor binding data produced from ChIP-seq experiments that large international projects such as ENCODE [6] have generated.

In the training phase the model learns a hierarchical representation of the training data. In the case of a model for transcription factor binding, DNA sequence motifs are learned together with how combinations thereof control the binding probability of a transcription factor. Such trained models can therefore be regarded as a compressed representation of the information contained in the training data. Once trained, the models can be used to make predictions. In other words the information extracted from the training dataset is transferred and applied to new, potentially unseen data. Trained models therefore hold the promise to allow for probing regulatory dependencies *in silico*, which among other applications enables interpreting functional variation in individual genomes. Due to these properties and capabilities of trained predictive models, their availability and sharing is beneficial to researchers and users.

Apart from making predictions such predictive models can be used to estimate the importance of subsets of their input as well as the effect of perturbations of the input. For DNA sequence-based models this enables the effect prediction of genetic variants on the specific model task - for example transcription factor binding. For a model to predict variant effect estimates it is not necessary that the query variant was seen in the training data - it infers the predictions from similar training sequences. Leveraging this aspect predictive models can be turned into tools for genetic variant annotation, integrating data from various experiments on different molecular profiles.

5.1.1 Existing model zoos

Despite the importance of trained predictive models, it is surprisingly difficult to share and exchange models effectively. Existing approaches to sharing models after publication are diverse, scattering models in author's websites, code repositories or other sources, rather than collecting all published models in one common repository. This makes it hard to find, explore, adapt, and to correctly install existing models. Currently, model repositories (*model zoos*) have been set up, but they are mostly specific to the technical modelling *framework* and programming language the model was implemented in. An overview of some popular model zoos is given in table **5.1**. All these repositories offer different sets of features, but in their common denominator they are all online storages of models without measures to ensure the model's functionality.

Name	Description	Bioinformatics models	URL
ModelZoo	Multiple frameworks No standardised model download No standardised data processing	No	https://modelzoo.co/
Tensorflow model zoo	Restricted to Tensorflow models No standardised model download No standardised data processing	No	https://github.com/tensorflow/models
Caffe model zoo	Restricted to Caffe models Standardised model download No standardised data processing	No	http://caffe.berkeleyvision.org/model_zoo.html
Keras model zoo	Restricted to Keras models Standardised model download Standardised data processing	No	https://github.com/fchollet/deep-learning-models

Table 5.1 The biggest existing model repositories (model zoos).

An alternative approach the above is OpenML (https://docs.openml.org/), where datasets and models are shared on an online platform. Machine learning pipelines can be designed and executed, which simplifies reproducibility of results and collaborations. Calculations are either executed on the OpenML server centrally or locally. OpenML requires data to be in tabular format so that it can directly be applied to pipelines and models. This has the disadvantage that raw data has to be processed prior to data submission, obscuring data preprocessing and complicating the application of OpenML functionality on user data.

All of the approaches above improve the availability of models and OpenML in particular improves reproducibility, but none of them can setup the required software dependencies or simplify and help with data preprocessing or model execution. This greatly hampers researchers from using published models on their own data. Most machine learning models require a numerical data matrix as input which has to be generated from diverse file formats. Given the range of file types in bioinformatics and the different ways to interpret them, preprocessing of genomics data is a more complex task than preprocessing in popular fields of machine learning, such as image recognition. Therefore, the availability of model-specific data preprocessing is critical for the functionality and ease of use of the model. There are several approaches to help with the interaction between bioinformatic data and deep learning such as dragoNN [176] and concise [177], but both are limited to one particular deep learning framework and are mainly designed to help training neural networks rather than using existing, trained ones.

Training and use of deep learning models is technically challenging due to the required computational power. For practical reasons, these calculations are commonly performed in graphical processing units (*GPU*s). Their access and use is controlled by specialised software libraries, which are not readily available in current operating systems. In order to simplify and abstract the technicalities, many frameworks have independently been developed and are incompatible with each other: Tensorflow [178], PyTorch [179], Theano [178], CNTK [180], etc. Additionally, these frameworks are heavily under development so that models, which are created using one version of the framework cannot be used in combination with a new software version that is released only a few months later. To ensure functionality and accessibility of models, it is therefore necessary to track the precise software requirements for every model individually. None of the existing model zoos implement strategies or standards for software dependency management, data preprocessing or automatic testing of models.

5.2 Kipoi - model zoo for genomics

To address the need of a platform for sharing trained models, that is capable of handling software dependencies of models as well as data preprocessing, Ziga Avsec and I have designed *Kipoi*. The name Kipoi, greek for "gardens", indicates our aim to integrate a wide range of models and to help the enhancement ("growing") of existing models in Kipoi. Kipoi is able to address the needs of genomics and at the same time also defines an application programming interface (*API*), that enables use of all models, irrespective of their software framework or programming language, with the same set of commands. Additionally, Kipoi offers predefined functionality for the most common deep learning frameworks - Keras [181], Tensorflow [178], and PyTorch [179] - which facilitates the contribution of those models to the Kipoi model zoo. Based on Kipoi's design, we defined standards for sharing of trained models. At the same time, models in Kipoi are not limited to deep learning or to any deep learning framework. At the time of writing the model zoo already contains over 2,000 trained genomics models.

Kipoi is a collaborative effort, including Ziga Avsec and me as the main contributors. Fruitful discussions on the software design and feasibility tests were performed with the help of Johnny Isreali, Nancy Xu, and Avanti Shrikumar from Prof. Anshul Kundaje's group at Stanford University, USA and Jun Cheng from Prof. Julien Gagneur's group at the TU Munich, Germany. Ziga Avsec implemented the core Kipoi API, model management, and automated testing. I implemented genetic variant effect prediction and model interpretation functionality, all of which will be discussed in detail in the following sections.

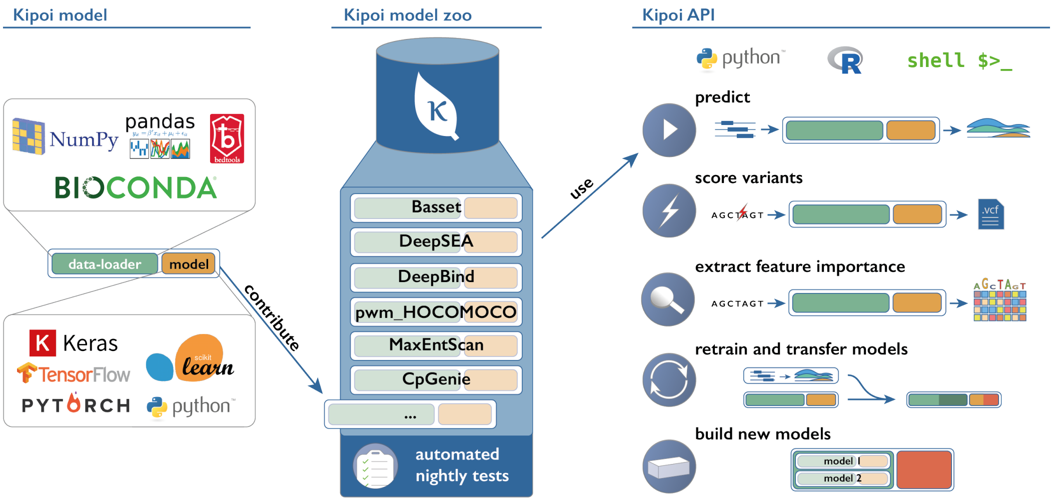


Fig. 5.1 Graphical abstract of the design of Kipoi. A Kipoi model consists in a data-loader and the core model. Kipoi models are stored in a repository (model zoo) and their execution is tested nightly. Models in the model zoo can be used for predicting on custom data, to score variants, to extract feature importance, to retrain and transfer models, and to build new models.

5.2.1 Design and implementation

The three main conceptual elements in Kipoi are: the Kipoi model, the repository (model zoo) and the application programming interface (API) (Fig. **5.1**).

Kipoi model

In a trade-off between generality and requirements of a powerful API, we have decided to split *Kipoi models* into a *data-loader* and a *core model*. The data-loader and the core model have one configuration file each that captures software dependencies, a basic description, and a specification of input/output-properties. All data regarding each Kipoi model has to physically reside in one directory.

Data-loader The data-loader is a piece of code written by the model contributor, which preprocesses bioinformatics input data. In the most generic form it is a python function or class that takes input parameters from the user, which may be file paths or alphanumerical values. The output is a python dictionary containing model input data (a numpy [182] array) and additional meta-data. All objects contained in the dictionary may only be instances of a python list, dictionary, numpy array, or base datatype (string, integer, float, etc.). This output is required by the Kipoi API to have a standardised interface between all data-loaders and all models. Furthermore, the python dictionary produced by the data-loader has to contain the `inputs` value which contains the model input data and can directly be passed to the model's `predict_on_batch` function.

To facilitate the implementation of data-loaders, all the preprocessing can be done with any software available on the coda or pip package managers. To comply with the Kipoi API they can then be wrapped in a python code by the model contributor.

Core model The Kipoi API at its bare minimum is designed to handle software requirements of models and to generate model predictions. The core model is a python class which encapsulates the original model as it is defined in the respective machine learning framework. The python class is required to have one function: `predict_on_batch`, which takes one argument and returns model predictions. The model class may have additional functions for intermediate layer activation extraction and for gradient calculation, as mentioned in section **5.2.2**.

If the model is written in the frameworks Keras [181], PyTorch [179], Tensorflow [178], or scikit-learn [183], then the respective default class in Kipoi can be used for the integration of the model. Otherwise, the model contributor has to write a python class that complies with the definitions within Kipoi. If the model is not written in python or in a python-compatible framework then the model can be integrated by writing a wrapper class in python.

Kipoi model zoo

The repository (*Kipoi model zoo*), implemented by Ziga Avsec, contains Kipoi models. All models in the zoo are tested nightly to ensure their functionality and the correctness of the software dependencies. Nightly tests are preformed at midnight according to the central european timezone to ensure a fixed testing interval of 24 hours. Kipoi models are versioned, hence updates of models (their configuration, data-loader, or the core model) are tracked and users can choose to use one specific version to ensure reproducibility.

The Kipoi model zoo relies on git and gitlfs for versioning and storing models.

The available models can either be listed using the Kipoi API or using the website: https://www.kipoi.org.

Kipoi API

The Kipoi API, which Ziga Avsec has implemented, unifies commands across frameworks to simplify the interchangeable execution of models as will be discussed in section **5.2.2**. Furthermore, it enables automatic download and installation, testing and execution of models in the model zoo. The automated software setup relies on the package managers conda and pip. Kipoi creates a new conda virtual environment for every model, which makes it possible to install software in the exact version required by one model, without interfering with previous installations of software. The new environment is then set up as specified in the data-loader and core model configuration files.

Model testing

To ensure reproducibility and functionality, every Kipoi model comes with an example (toy) dataset. This dataset is used for testing the model and data-loader functionality and compatibility. For the test itself Kipoi attempts to generate model predictions using the toy data. If this succeeds, the model passes the tests. If any warnings or errors occur in the process then the test has failed. The model zoo maintainers are informed which models have failed the tests, so that the original model contributors can be contacted. It is the contributor's responsibility to maintain a model's functionality. If the contributor refuses to do so, the model zoo maintainers may attempt to solve the problem themselves, but if that fails the zoo maintainers remove the model from the list of accessible models. Previous functional versions of the model remain accessible.

The model tests are executed for all models in an automated way, which also asserts that model and data-loader dependencies are correctly defined. Therefore, for every model a new empty conda environment is generated, the dependencies are installed and the test prediction is performed. If all passes, then the model has passed the automated test. Model testing and its automatisation was implemented by Ziga Avsec.

5.2.2 Usage

The application of machine learning models on data is a highly unified process using Kipoi. Given conda, git and gitlfs are installed and set up, Kipoi creates the new model environment using the command `kipoi env create <ModelName>`, where `<ModelName>` is replaced the name of the requested model. Once the environment has been created it has to be activated prior to use by `source activate <ModelEnvironmentName>`. After activating the environment, the model can be used from the python, R, or command-line interface.

Model prediction

Leveraging our API design, there is one single command (`kipoi predict <ModelName>`) to execute prediction of any model on custom data. A user therefore only has to familiarise themselves once with the syntax of Kipoi and can then use all models in the same way, irrespective of the model's programming language or framework. The only argument of `kipoi predict` that has to be adapted for every model, apart from the model name, is the set of arguments that are passed on to the data-loader (the `-dataloader_kwargs` argument).

Data-loaders are per definition model-specific. As they preprocess data and convert it to a format that is compatible with the model, all the input files and parameters necessary for model prediction have to be passed on to the data-loader. For example, the *DeepSEA* model [172] predicts epigenetic features from DNA sequence. This DNA sequence could potentially come from any source, but in order to cover the main use-cases the data-loader was designed to use a combination of standard bioinformatics file formats: one `.fasta` file and one `.bed` file. The `.fasta` file may contain the sequence data from a reference genome and the `.bed` file would then contain the definition of query regions within the sequences defined in the `.fasta` file. With this combination of the two files, the data-loader can extract requested DNA sequences from the `.fasta` file and generate python objects that can be used directly for model execution. The data-loader is provided by the model contributor and should aim to make it possible to execute the model with standard bioinformatics file formats in a straightforward way. The consecutive execution of data-loader and model are all handled in the background by Kipoi.

The output from the model prediction can either be saved to a text or a `hdf5` file in a standardised format, or, if the python or R API are used, the returned data object can be accessed directly.

Intermediate layer activation

For neural networks it is possible to evaluate the internal state of the model at prediction time. This can be used as a tool to open the *black box* (see section **5.4**) in order to see how the model reacts to input and how predictions are calculated. For example in convolutional neural networks that use DNA sequence as an input, the first layer acts as a motif detector, which can be visualised by extracting the activation states of that layer [171].

Technically, extracting layer activation works differently in every framework and at least basic knowledge of the respective framework is required to do so. For Kipoi, I have integrated functionality to extract the activation state of any layer in all deep learning frameworks that are supported out-of-the-box: Keras [181], PyTorch [179] and Tensorflow [178]. If the model contributor uses a different framework then they would also have to implement this function using the respective framework. Giving the user simplified accessibility to intermediate layer activation helps understanding the specific model. It is also one of the basic tools for quality control for the trained models.

Gradient calculation

In neural network frameworks, the calculation of gradients of the output with respect to the input is one of the core features. It is used to train the model by gradient descent, updating the model weights in a way that minimises the defined loss function (see section **1.4.1**). During training, the gradient is therefore informative for the model how its parameters have to be adjusted to improve its predictions (by defining the direction and strength of corrections in the parameter space). The visualisation of gradients with respect to model input has also proven to be a powerful tool to highlight influential regions of the input (see section **5.4**). It can therefore be informative for the user to access this information in order to interpret their data with a given trained model. An example is given in section **5.4**.

Despite gradient calculation being a core element of deep learning frameworks, it is often complicated to find documentation on how to perform the calculation explicitly and in-depth knowledge of the framework is required. To facilitate this further, I have implemented a consistent way to perform gradient calculation using three popular deep learning frameworks: Keras [181], PyTorch [179] and Tensorflow [178]. The unified command to calculate gradients is `kipoi grad <ModelName>`, which is the same for all three frameworks and resembles the structure of the `kipoi predict <ModelName>` command. The output can either be saved to a text or a `hdf5` file, or, if the python or R API are used, the returned data object can be accessed directly. Availability of this universal command greatly facilitates advanced model and data visualisation, which would otherwise only be available for machine learning developers.

5.2.3 Contributing models

Kipoi's functionality relies on the standardised way of model contribution. Once a new machine learning model of any kind has been developed, it can be integrated into the Kipoi model zoo. All software required to load the model and to execute the data-loader has to be (made) available via the package managers conda or pip. We have defined configuration files that list all the software dependencies including versions and other vital information for Kipoi.

The requirements in order to contribute a model are:

- Make sure that all software required to load the model and to execute the data-loader is available via the package managers conda or pip
- Write the model and data-loader configuration files
- Write the data-loader

- If the modelling framework is not supported out of the box: write a python model class
- Supply a small sample dataset which can be used as a toy example and for testing
- Make sure the model is fully functional and produces correct results at all times

Great care has to be taken that all configuration files are correct. We have therefore equipped Kipoi with model testing functionality, which can be used to verify the successful setup and execution of the data-loader and the model. Before a model can be added to the model zoo it has to successfully pass both tests. We recommend that the model contributor performs these tests prior to submission, to ensure functionality, integrity and avoid delays in accepting the model in the zoo.

Once the model is ready for submission, the default approach for adding new data to git repositories has to be followed: the model repository https://github.com/kipoi/models/ has to be *forked* using git. Then the new model has to be copied into the forked repository. Finally, the new model can be added and a *pull request* (i.e.: merging of git repositories or branches) to the original Kipoi model repository can be made. The pull request will inform all maintainers of the repository about the attempt to add the new model. After review by us maintainers either changes to the model may be requested or we accept the pull request. After accepting, the new model is part of the Kipoi model zoo.

5.3 Genetic variant effect prediction

Currently, a variety of models are trained to predict various molecular measurements of the cell [171–175, 184] purely based on DNA sequence. For models that have DNA sequence as one of their inputs it is possible to predict the effect of DNA variants on the task of a model. For example, assuming a model is trained to predict transcription factor binding from DNA sequence, it is then possible to predict the effect of DNA variants on the binding of the respective transcription factor with that model. One approach to do so, *in-silico mutagenesis* (*ISM*), produces model predictions twice, once using DNA input sequence carrying the reference allele, once using the mutated DNA sequence carrying the alternative allele. Differences between the two sets of model predictions are then contrasted and scored. The result is the estimated variant effect. This approach was used by Alipanahi et al. [171] and Zhou et al. [172] in slight variations of the scoring functions. Here I implemented a generalised way to predict ISM-based variant effect scores for any DNA-sequence-based model. To enable a model for variant effect prediction, only the model configuration files have to be updated and no extra implementation time from the model contributor is required.

5.3.1 In-silico mutagenesis

In-silico mutagenesis (ISM) is a perturbation-based approach for variant effect prediction. It perturbs the model input and for every perturbation, model predictions are calculated. Differences of the model output with respect to the input perturbations are therefore used to estimate effect of the perturbation. Prior to calculation, ISM selects a region of the genome containing the position of the variant of interest, typically the region is centred on the variant. The perturbation is then performed by replacing the reference allele with the alternative allele. It is therefore important to distinguish between single nucleotide variants (*SNV*s) and insertions and deletions (*indel*s). Perturbation for SNVs are base substitutions in the input DNA sequence. Indels, per definition, alter the sequence length. This has different implications for the implementation, which I will discuss in the next section.

Contrasting predictions from reference and alternative sequences can be done in various ways: the difference between model outputs, the difference of the log-odds of the model outputs, or combinations thereof can be calculated. ISM predicts the effect of a variant on the model task. If a model, for example, predicts the binding probability of one transcription factor to a given DNA sequence, then the most appropriate way to contrast predictions is by calculating the difference of the log-odds of the model predictions. The result will then be informative of the effect of the variant on the binding of the transcription factor for which the model was trained. Various different ways of contrasting ISM predictions were used by Alipanahi et al. [171] and Zhou et al. [172] and are available in Kipoi.

5.3.2 Implementation

Despite its simplicity, not all DNA-sequence based models are equipped with ISM effect estimation, as it is the case for MaxEntScan, HAL, and labranchor [184, 174, 175]. The implementation time for the algorithm as well as the user interface is not negligible. Moreover, among the published models that bring variant effect prediction with them, scores are calculated in different ways [171, 172]. This leads to ambiguity and makes results hard to compare.

Leveraging the standardisation of Kipoi, I implemented ISM variant effect prediction as Kipoi *plug-in*. This plug-in can be applied to any trained model in the Kipoi zoo that uses DNA sequence as input. Input and output files of the variant effect prediction are in *VCF* format, the standard file format for storing variants. The output VCF files are annotated with the calculated variant effect. My implementation wraps around the data-loader and the model to generate valid datasets for the perturbation of model inputs (Fig. **5.2**). The only requirements for data-loaders are that meta-data is returned alongside the model input

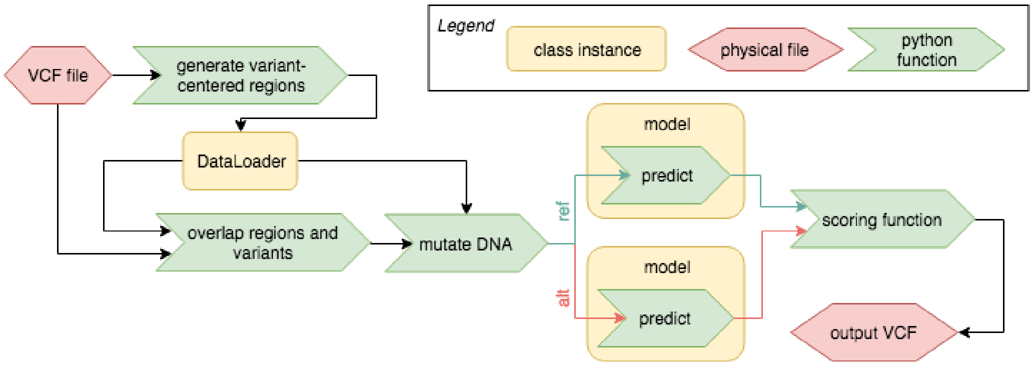


Fig. 5.2 ISM-based variant effect prediction for data-loaders that accept `bed` files.

data. For variant effect prediction the metadata must annotate every model input sample with the genomic region it has been generated from. Additionally, if the DNA-sequence input of the model is in one-hot encoded format, then the encoding must be: $A = [1,0,0,0]$, $C = [0,1,0,0]$, $G = [0,0,1,0]$, $T = [0,0,0,1]$.

At the moment variant effect prediction is implemented for SNVs, as variant scoring for indels cannot be implemented in the same general and "non-invasive" way as described below. This is because indels alter the DNA sequence length and most models have a fixed input sequence length, other solutions therefore have to be found for indel effect prediction in future.

To cater for the different requirements of models, I have designed two alternative ways to select genomic testing regions on the basis of the query VCF file:

Variant-centred approach

In agreement with the strategies in Alipanahi et al. [171] and Zhou et al. [172] I attempt to centre the model input on the query variant. This has the advantage that it gives the model the most sequence context on either side of the variant to perform inference (Fig. **5.2**).

1. The VCF file is read and a genomic region of the input length of the model, centred on the variant, is generated
2. The data-loader is executed on these regions and produces model input data as well as meta-data (genomic position of the region, etc.)
3. Leveraging the meta-data, the VCF is overlapped again with the data generated by the data-loader
4. Based on the overlap a *reference* (ref) and *alternative* (alt) set of the model input data is generated, by base substitution

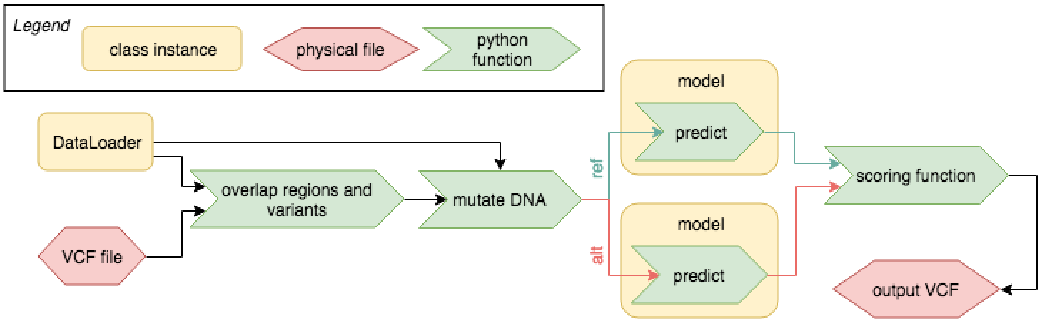


Fig. 5.3 ISM-based variant effect prediction for any model and data-loader. Not generating variant-centred regions.

5. Model predictions are calculated on both input datasets
6. The effect size is calculated in a scoring function that contrasts the predictions and returns one score
7. The input VCF is annotated with the calculated scores

This approach works for all trained models for which the data-loader can generate model input data in any region of the genome. Additionally, the data-loader is required to accept files in the `.bed` file format.

Overlap-based approach

Some models, such as the HAL and labranchor [174, 175] models for splicing, can only predict values for specific regions of the genome and therefore the previous approach fails. To overcome this limitation, I implemented an alternative procedure, where the initial step of the overlap of data-loader data with the VCF file is different (Fig. **5.3**):

1. The data-loader is executed and produces model input data as well as meta-data (genomic position of the region, etc.)
2. Leveraging the meta-data the VCF is overlapped again with the data generated by the data-loader
3. Based on the overlap a *reference* (ref) and *alternative* (alt) set of the model input data is generated, by base substitution
4. Model predictions are calculated on both input datasets
5. The effect size is calculated in a scoring function

6. The input VCF is annotated with the calculated scores

Since this procedure does not have any control over the regions that are being investigated, it is not possible to centre the tested region on the variant. Also, it can not be guaranteed that all variants in the VCF file are tested.

Scoring functions

In the previous sections scoring functions were mentioned as a step that converts the model predictions for reference and alternative input sequences into a single effect score. The user can select one or more of the following functions to annotate the VCF file:

- reference: Model prediction for the reference allele
- alternative: Model prediction for the alternative allele
- difference: Difference between model predictions for the alternative allele and reference allele
- logit-reference: Logit of the model prediction for the reference allele
- logit-alternative: Logit of the model prediction for the alternative allele
- logit-difference: Difference between the logits of the model predictions for the alternative allele and reference allele
- DeepSEA-effect: An additional score used in Zhuo et al. [172]:
  `abs(logit-difference) * abs(difference)`

This set of scores covers most use cases, but if the user wants to calculate a different score then that scoring function (written in python) can be supplied at execution time.

5.3.3 Usage

For the user most of the above is handled transparently. In order for Kipoi to execute the model prediction, variant effect prediction requires a VCF file and all the other input files and settings that are needed for model prediction (section **5.2.2**). The VCF file location is defined using the `-vcf_path` argument and the additional files are defined using the `-dataloader_kwargs` argument.

Another application of variant effect prediction are mutation maps (see section **5.4.1**). *Mutation maps* help to understand the effects of variants in the context of their surrounding genetic region.

To exemplifying both the modularity and the power of Kipoi's design we have built an *ensemble model* using variant effect prediction on existing Kipoi models. `KipoiSplice4` (https://kipoi.org/models/KipoiSplice/4/), an ensemble model designed and implemented by Ziga Avsec and me, predicts splice variant effects based on outputs of four other Kipoi models. Therefore, first, variant effect prediction is applied on the MaxEntScan/3prime, MaxEntScan/5prime, labranchor, and HAL models [184, 174, 175]. Effect scores are then collected and fed into a logistic regression model that then returns one score for every tested splice variant. The model can only be applied to variants close to splice sites as MaxEntScan/3prime, MaxEntScan/5prime, labranchor, and HAL can only perform predictions in regions close to splice sites. KipoiSplice4 showcases how Kipoi users may create new models using existing trained models as building blocks.

5.3.4 Summary

Unifying variant effect prediction in Kipoi improves comparability among the results generated by different models and enhances models that don't have variant effect prediction capabilities, at no extra cost. It does not impose any functional limitations on data-loaders or models and has very moderate requirements on the model and data-loader design. The premise to have so few requirements for the data-loader and model also made it impossible to add support for indel effect prediction. For indels the model input sequence length is altered and since in general, models have a fixed input sequence length, a more invasive implementation will have to be used.

One of the big advantages of this implementation is that it works for any kind of DNA-sequence based model, be it a neural network, a support vector machine or any other kind of model.

Performing variant effect prediction using machine learning models can help resolve uncertainty of the causal variant due to linkage disequilibrium (LD). In association tests LD hampers the identification of the causal variants, and often multiple variants in high LD show similarly high association strength. Machine learning models trained on thousands of regions of the genome learn regulatory important motifs and sequences. This makes it possible to assess the effect of high-LD variants independently in the context of regulatory important regions.

5.4 Tools for model and data interpretation and visualisation

The more complex machine learning algorithms get, the harder it is to follow the reasoning of how the algorithm came to a prediction. The machine learning model in those cases turns into a *black box*, where all the reasoning is hidden from the user and the output is produced from an input by an unknown function. Neural networks and especially deep neural networks can reach much higher levels of complexity than traditional methods like decision trees or support vector machines and hence their reasoning is much harder to comprehend. Opening the black box of a machine learning model is not purely of academic interest. For a complex trained model it is hard to ensure the robustness and generalisation of its predictions, as generally the transformation from model input to model output is complex and not well understood. Additionally, most neural networks return point estimates of their predictions without confidence intervals. All of the above hamper the application of complex trained models in settings, such as in medicine, where correctness of predictions is of utmost importance and uncertainty estimates of predictions are essential.

To address this problem in neural networks, different approaches have been proposed and implemented: in image recognition one of the first approaches, *saliency maps*, was to use the gradient of the model output with respect to the model input [185] as an indication of which areas of the input were influential for the output. The idea was to understand how minimal changes of the output would reflect on the input. Therefore, the most influential parts of the input have the strongest gradients. Using the gradient as a measure to understand models, as well as input data, is particularly convenient as all software frameworks for neural networks implement the calculation of gradients, since they are a core element for training the neural network. Several variations and enhancements of saliency maps have been published [186–189].

Approaches such as saliency maps are transferable to genomics. For models that use DNA sequence as input, influential regions of the sequence that drive the model output, such as important motifs, have stronger gradients. In Kipoi I have implemented a consistent way to execute gradient calculation, leveraging the internal functions of software frameworks for neural networks (see section **5.2.2**). Fig. **5.4** shows an example of a visualisation of gradients calculated for a model input.

Gradients of the output with respect to model input describe how the output changes as the model input increases, calculated at every position of the input. In other words, the gradients are calculated using the input data as a baseline. Alternative algorithms have been developed that calculate the importance of input features with regard to a user-defined

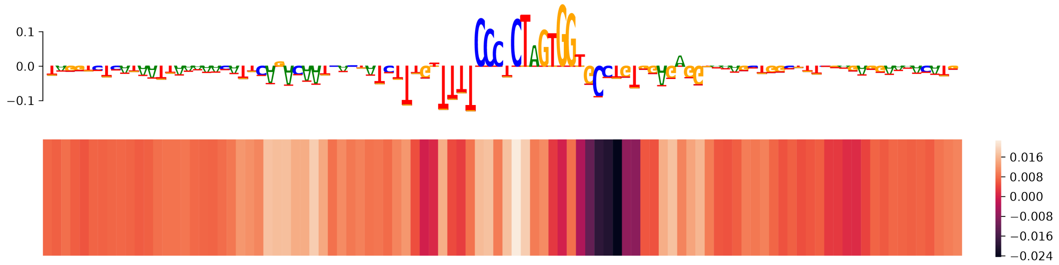


Fig. 5.4 Displays data of the genomic region chr22:28,711,970-28,712,070 (reverse complement, hg19) scored by the gradients of FactorNet trained on CTCF binding data (Kipoi model: `FactorNet/CTCF/meta_RNAseq_Unique35_DGF`). The letter height of the top panel is the gradient of the output with respect to the input and on the bottom is a heatmap displaying the gradient of the output with respect to chromatin accessibility.

baseline and therefore are potentially more informative for capturing all influential features [188, 189]. In Kipoi, I have integrated the existing python implementation of DeepLIFT [189], which is capable of estimating feature importance with respect to a baseline. For the user it is therefore possible to interpret their input data using a trained model and DeepLIFT without writing a line of python code.

5.4.1 Mutation maps

Variant effect prediction, as it is implemented in Kipoi, calculates the effect of variants as defined in a VCF. At times it is relevant to visualise the context of the variant and the effects of variants in the proximity of the query variant. Convolutional neural networks (CNNs) predict by finding patterns in the input sequence. CNNs used in genomics therefore base their predictions for example on binding motifs of transcription factors, which are learned from the data during training. Important patterns of the input sequence can either be found by the interpretation tools mentioned above, or by using *mutation maps*. Mutation maps calculate variant effects for every base in the input sequence and mutate to all possible alternative alleles. It is therefore possible to relate the effect of one variant to the effect of all other possible variants in its proximity. Fig. **5.5** highlights the effect of a query variant (chr11:5246970:A:C) and the effects of all variants in its proximity on models trained on GATA2 binding. The letter height in the seqlogo-plot on top of the heatmaps shows the cumulative magnitude of the predicted variant effects at the letter's position. As a result the GATA motif is highlighted as a region where mutations are predicted to strongly affect GATA2 binding. I implemented the visualisation of mutation maps in Kipoi in order to make it easily accessible for users.

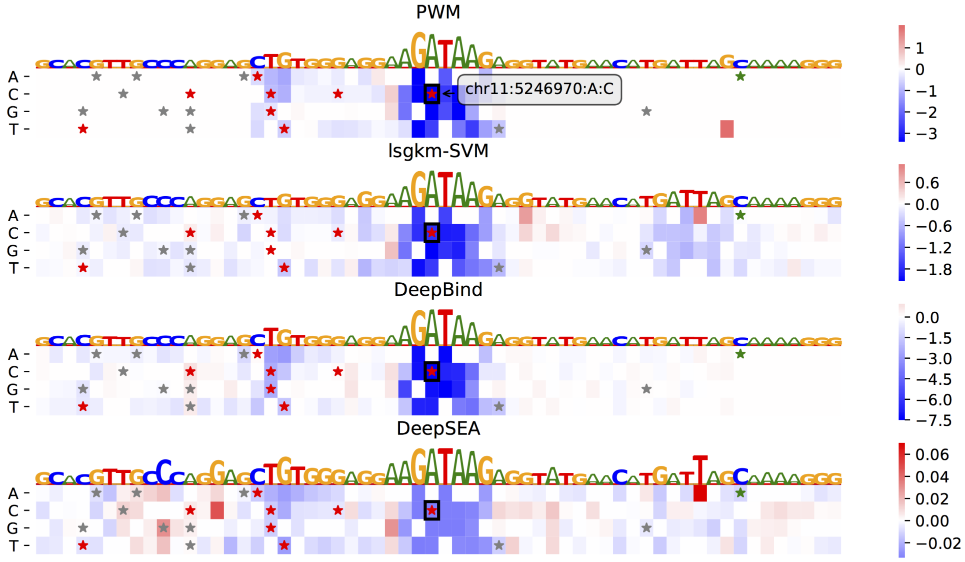


Fig. 5.5 Mutation map (heatmap of variant effects, blue negative effect, red positive effect) for query variant rs35703285 (chr11:5246970:A:C) and the predicted GATA2 binding difference between alleles for 4 different models. The letter height is the sum over all effects (absolute value) of all possible variants in one position. The black boxes in the mutation maps highlight the position and the alternative allele of the query variant. Additionally, stars highlight variants annotated in the human variant database ClinVar with red: (likely) pathogenic, green: likely benign, grey: uncertain or conflicting significance, other.

Notably, mutation maps are not limited to neural network models, but can only be applied to models that used DNA sequence as input. In Fig. **5.5** mutation maps are calculated for non-neural network models such as position weight matrices (PWMs) and a support vector machine model (lsgkm-SVM), as well as DeepBind and DeepSEA, which are both neural networks.

Mutation maps are calculated by iteratively applying variant effect prediction on the whole input sequence. This may cause considerable computational costs when applied to many regions of the genome. It is therefore relevant to have alternative approaches like gradient-based methods that predict importance of all input features in a single calculation.

5.4.2 Summary

All of the above methods, including visualising the activation states of intermediate layers (see section **5.2.2**), attempt to help understand model predictions. For genomics it is difficult to find easily understandable measures of model interpretability, as genomics data is not as naturally interpretable as images. Furthermore, different genomic and epigenetic measurements are in a complex relationship with each other. As an example I showed the gradients with respect to input data for the FactorNet model [173] (Fig. **5.4**). The FactorNet model aims to predict transcription factor binding probabilities based on DNA sequence, also taking chromatin accessibility and mappability into account. Fig. **5.4** shows the gradients of a CTCF transcription factor binding model with respect to a genomic region that overlaps an empirically found CTCF binding site. Interpreting the gradients on the input DNA sequence is rather straightforward - the binding motif is highlighted. But already the visualisation of the gradients on the chromatin accessibility data is hard to interpret. Regions of high interest are highlighted, but it is hard to intuitively understand whether that pattern of gradients makes sense and how it relates to biological properties of the cell. This highlights that model and data interpretability is still in its infancy and more effort has to be put into finding better ways to validate and interpret models.

5.5 Discussion

The lack of a universal platform for sharing trained machine learning models, in particular for the field of genomics, hampers exchange, re-use and reproducibility. Such a platform would have to manage software dependencies and allow for easy setup and execution of said models. Together with a fellow PhD student, Ziga Avsec (TU Munich), and other international collaborators from Stanford University, we have developed Kipoi. Kipoi is a model repository ("model zoo") which allows the contribution of trained models irrespective of their underlying software framework and programming language. Apart from simplest possible sharing and installation of models, Kipoi has an application programming interface (API) which allows the access and execution of all models in a unified way, hiding model-specific commands from the user.

After the setup of a model, Kipoi's core features facilitate the execution of model prediction using new data. To do so, Kipoi models - trained machine learning models submitted to the Kipoi model zoo - are a bundle of a data preprocessing pipeline ("data-loader") and the core model. Data-loaders encapsulate all data preprocessing steps necessary to convert standard bioinformatics files into a model-compatible format. Data-loaders are written by model contributors and are a cornerstone for Kipoi's simplicity of use.

Kipoi's API defines a consistent approach that generalises access to core functions of machine learning frameworks. Most machine learning models are based on a framework that facilitates model development. This framework defines how the model can be executed, which function has to be called in order to perform predictions and how model parameters can be accessed. In Kipoi we have currently integrated four popular machine learning frameworks: Keras [181], PyTorch [179], Tensorflow [178], and scikit-learn [183]. For those frameworks we have implemented full integration for all Kipoi API functions required for model prediction and model interpretation. If a model contributor wants to add a model to the zoo that is not based on any of those frameworks, they have to, at the least, define how Kipoi can execute model prediction. An example of such a model is the `lsgkm-SVM` model, which I have successfully contributed to the model zoo.

Leveraging the standards and components that we have defined in Kipoi, I was able to implement variant effect prediction for single nucleotide variants (SNVs) in a way that does not require any adaptions of model or data-loader code. With my approach way, any model that uses DNA sequence as input can be used to score variant effects. The procedure uses VCF files, a standard bioinformatics file format for genetic variants, as input and produces annotated VCF files as output. The non-invasiveness of the implementation causes no extra implementation cost for the model contributor. In order to enable a trained model for variant effect prediction, only a few settings in the model and data-loader configuration files have to be added.

In addition to variant effect prediction I have implemented unified access to model interpretation functionality irrespective of the underlying software framework. Especially, for the deep learning frameworks Keras [181], PyTorch [179], and Tensorflow [178] I have written abstracting functions that hide and unify framework-specific function calls in order to extract intermediate layer activation and calculate gradients with respect to model input. These features are commonly used to open the "black box" of machine learning models and to interpret input data in the context of the model task. Without this abstraction in-depth knowledge of the machine learning framework is necessary in order to perform aforementioned calculations. In the Kipoi API, one function call performs the corresponding calculations in each framework.

Kipoi can be used to enhance and build new models. Since Kipoi ensures model functionality and facilitates model setup, it also greatly reduces setup time for model developers, who plan to enhance or transfer an existing model to a new one. The data-loader makes it unnecessary to spend time on correct preprocessing of input data. Additionally, new Kipoi ensemble models can be built by integrating existing Kipoi models into a new one. An example of those is `KipoiSplice4` (https://kipoi.org/models/KipoiSplice/4/), an ensemble model designed and implemented by Ziga Avsec and me, that predicts splice variant effects based on outputs of four other Kipoi models.

We envision Kipoi to become the main platform for sharing published machine learning models in genomics and beyond, in analogy to how experimental data is shared on platforms like the Gene Expression Omnibus (GEO) [190]. The flexibility of Kipoi's design allows the integration of any model using any underlying software framework and programming language. Specialised functionalities which are not currently covered by Kipoi can easily be added as Kipoi plug-ins, which may work on all or on a subset of models in the model zoo.

Chapter 6

Conclusion

In my PhD I have studied the effects of non-coding variants using different techniques: I have performed eQTL analyses to identify gene expression regulatory variants, characterised them with genome annotation, chromatin conformation and colocalised them with GWAS results. Additionally, with a fellow PhD student Ziga Avsec from Prof. Julien Gagneur's group at the Technical University of Munich, Germany, I have developed a software, (*Kipoi*), which facilitates sharing of trained machine learning models in genomics and beyond. As part of that I have focussed on the implementation of techniques for the effect prediction of genomic variants using trained models and tools to understand the reasoning for predictions within the models.

In an effort to aggregate all available eQTL datasets of purified human blood cells, I have compiled eQTL datasets from four studies on seven purified blood cell types. The aim was to generate a resource that enables the detection of currently undetected gene expression-associated variants and increase robustness of the known findings. I re-processed the individual datasets uniformly, merged them and performed cis-eQTL analyses, taking batch effects into account. For monocytes the collection was the largest with 1,480 samples while for other cell types sample sets were smaller (table **3.4**). Bigger sample sizes enabled higher confidence in the identification of associations, as well as the detection of variants with more subtle effects. Additionally, the analysis has been extended from common to low-frequency variants, identifying additional regulatory variants. Comparing eQTLs of purified monocytes with eQTLs from whole blood, I discovered that the majority of gene-variant associations can be detected in both. This was the case for a whole blood eQTL dataset of similar size, as well as for one of ten times the sample size of the aggregated monocyte dataset. The two main advantages of purified cell types are: firstly, that the estimated effect size of eQTLs detected in whole blood is derived from a mix of cell types, averaging over potentially different effect magnitudes and directionalities of the same gene-variant association in different cell types, and secondly, data from purified cell types enables the identification of eQTLs in a cell type specific manner.

Expression QTL datasets from purified cell types can be used to identify cell type specific regulatory variants, interpret and annotate genomics datasets, and help identify potentially causal genes for trait and disease. By clustering the replication of eQTLs across cell types I found that cell type specificity of eQTLs can recapitulate the ontology of these seven heamatopoietic cells. This alignes with the cell type specificity of gene expression regulatory elements [6, 191], thereby indicating that similarity of gene expression regulation relates to blood cell differentiation. Comparing eQTL effect directionality across cell types, I could reproduce previous findings [87, 112] that effect directionality of the same gene-variant association may differ between cell types. This confirms that gene regulation is a complex process that requires multiple DNA-binding proteins to work together in a cell type specific manner. The change of effect directionality also has practical implications on the viability of effect size estimates in whole blood eQTL results, as mentioned in the last paragraph. In an effort to link genes to blood cell traits I have colocalised eQTL results with results from a recent GWAS study on blood cell indices [45]. I found colocalisation between 60 (out of 2,069) GWAS variants and 58 (out of 13,230) eQTL genes. One of those was variant rs4148436, associated with plateletcrit, which could be linked to ABCC4 expression. Loss of ABCC4 was independently found to be a potential cause for impaired haemostasis resulting in bleeding by my former PhD colleague, Dr. Tadbir Bariana at the Department of Haematology. Together, these observations highlight the relevance of ABCC4 in the platelet biology and function. Gene expression is regulated by many variables, one of those being the binding of transcription factors to DNA [12]. Molecular interactions between these bound transcription factors and other DNA-binding proteins contribute to cell type specific three dimensional interaction patterns between promoters and expression regulatory elements of the genome, thereby driving cell type specific programmes of gene transcription. In my analyses I found that promoter interacting regions identified by promoter capture HiC are enriched for expression regulatory variants, supporting the aforementioned hypothesis. I published these results as a co-first author in Javierre et al. [160].

Biological processes in a cell are complex and generally non-linear. It is therefore necessary to apply complex algorithms to understand and model cellular regulation. With increasing amounts of available biological measurements it has been possible to use machine learning models for the study of genomics. This led to increasing numbers of published machine learning models in genomics. However, the simple use and rapid enhancement and improvement of existing models is hindered by the lack of a consistent standard for sharing of trained models. Together with Ziga Avsec, I have developed Kipoi, which defines standards for sharing, and for facilitating installation and execution of trained models. To date, Kipoi contains more than 2,000 models, which are freely accessible. To enhance Kipoi's core functionality I have implemented a method that enables any DNA-based predictive models to estimate effects of genomic variants. To do this I mutate DNA sequence inputs to models and calculate the difference of the model predictions. This turns existing machine learning models in genomics into variant annotation tools. Thanks to Kipoi's standardisation all these functions can be executed in a user-friendly environment. We could show the power and the simplicity of use of Kipoi in a preprint in which I am second co-first author [165].

In order to make the output of my research available to the community I am currently preparing a manuscript in which I will publish the eQTL results as well as a website for interactive exploration of these results. I hope that this will help researchers to relate and annotate their data with the one presented here. I have already made my data available for colocalisation studies that will be performed on the MEGASTROKE [192] GWAS, and on two still unpublished GWAS, one on coronary artery disease and one on a new category of blood cell traits.

Kipoi is already publicly available and in use. The project's immediate future is ensured as grants have been secured for a full-time PhD student or post doctoral fellow. We hope that Kipoi will serve as a new standard for the publication of machine learning models in genomics, and that it will therefore simplify access to published models, and facilitate their comparison, re-use, and enhancement.

References

[1] Geoffrey M. Cooper and Robert E. Hausman. *The Cell A Molecular Approach*. ASM Press, Washington, DC 20036 USA, 4th edition, 2007. ISBN 0-87893-219-4.

[2] J.D. Watson, T.A. Baker, S.P. Bell, A. Gann, M. Levine, and R. Losick. *Molecular Biology of the Gene*. COLD SPRING HARBOR LABORATORY PRESS, Cold Spring Harbor, New York, 7th edition, 2014. ISBN 9780321762436.

[3] Jamal Tazi, Nadia Bakkour, and Stefan Stamm. Alternative splicing and disease. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, 1792(1):14–26, 2009. ISSN 09254439. doi: 10.1016/j.bbadis.2008.09.017. URL http://dx.doi.org/10.1016/j.bbadis.2008.09.017.

[4] Xiang-Dong Fu and Manuel Ares. Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics*, 15(10):689–701, 2014. ISSN 1471-0056. doi: 10.1038/nrg3778. URL http://www.ncbi.nlm.nih.gov/pubmed/25112293.

[5] Greg Elgar and Tanya Vavouri. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends in Genetics*, 24(7):344–352, 2008. ISSN 01689525. doi: 10.1016/j.tig.2008.04.005. URL http://linkinghub.elsevier.com/retrieve/pii/S0168952508001510.

[6] The ENCODE Project Consortium, Ian Dunham, Anshul Kundaje, Shelley F Aldred, Patrick J Collins, Carrie a Davis, Francis Doyle, Charles B Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R Lajoie, Stephen G Landt, Burn-Kyu Bum-Kyu Lee, Florencia Pauli, Kate R Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M Simon, Lingyun Song, Nathan D Trinklein, Robert C Altshuler, Ewan Birney, James B Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Jason Ernst, Terrence S Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C Hardison, Robert S Harris, Javier Herrero, Michael M Hoffman, Sowmya Iyer, Manolis Kelllis, Pouya Kheradpour, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K Marinov, Angelika Merkel, Ali Mortazavi, Stephen C J Stephanie L Parker, Timothy E Reddy, Joel Rozowsky, Felix Schlesinger, Robert E Thurman, Jie Wang, Lucas D Ward, Troy W Whitfield, Steven P Wilder, Weisheng Wu, Hualin S Xi, Kevin Y Yip, Jiali Zhuang, Bradley E Bernstein, Eric D Green, Chris Gunter, Michael Snyder, Michael J Pazin, Rebecca F Lowdon, Laura a L Dillon, Leslie B Adams, Caroline J Kelly, Julia Zhang, Judith R Wexler, Peter J Good, Elise a Feingold, Gregory E Crawford, Job Dekker, Laura Elinitski, Peggy J Farnham, Morgan C Giddings, Thomas R Gingeras, Roderic Guigó, Timothy J Tomothy J Hubbard, Manolis Kellis, W James Kent, Jason D Lieb, Elliott H Margulies, Richard M Myers, John a Starnatoyannopoulos, Scott a Tennebaum, Zhiping Weng, Kevin P

White, Barbara Wold, Yanbao Yu, John Wrobel, Brian a Risk, Harsha P Gunawardena, Heather C Kuiper, Christopher W Maier, Ling Xie, Xian Chen, Tarjei S Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J Coyne, Timothy Durham, Manching Ku, Thanh Truong, Matthew L Eaton, Alex Dobin, Andrea Tanzer, Julien Lagarde, Wei Lin, Chenghai Xue, Brian a Williams, Chris Zaleski, Maik Röder, Felix Kokocinski, Rehab F Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto Chakrabortty, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Cédric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J Luo, Eddie Park, Jonathan B Preall, Kimberly Presaud, Paolo Ribeca, Daniel Robyr, Xiaoan Ruan, Michael Sammeth, Kuljeet Singh Sandu, Lorain Schaeffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Hao Wang, Yoshihide Hayashizaki, Alexandre Reymond, Stylianos E Antonarakis, Gregory J Hannon, Yijun Ruan, Piero Carninci, Cricket a Sloan, Katrina Learned, Venkat S Malladi, Matthew C Wong, Galt P Barber, Melissa S Cline, Timothy R Dreszer, Steven G Heitner, Donna Karolchik, Vaness M Kirkup, Laurence R Meyer, Jeffrey C Long, Morgan Maddren, Brian J Raney, Linda L Grasfeder, Paul G Giresi, Anna Battenhouse, Nathan C Sheffield, Kimberly a Showers, Darin London, Akshay a Bhinge, Christopher Shestak, Matthew R Schaner, Seul Ki Kim, Zhuzhu Zhancheng Zhengdong Zhang, Piotr a Mieczkowski, Joanna O Mieczkowska, Zheng Liu, Ryan M McDaniell, Yunyun Ni, Naim U Rashid, Min Jae Kim, Sheera Adar, Tianyuan Wang, Deborah Winter, Damian Keefe, Vishwanath R Iyer, Kljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Jason Gertz, Jost Vielmetter, E Christopher Partridge, Katherine E Varley, Clarke Gasper, Anita Bansal, Shirley Pepke, Preti Jain, Henry Amrhein, Kevin M Bowling, Michael Anaya, Marie K Cross, Michael a Muratet, Kimberly M Newberry, Kenneth McCue, Amy S Nesmith, Katherine I Fisher-Aylor, Barbara Pusey, Gilberto DeSalvo, Suganthi Sreeram Balasubramanian, Nicholas S Davis, Sarah K Meadows, Tracy Eggleston, J Scott Newberry, Shawn E Levy, Devin M Absher, Wing H Wong, Matthew J Blow, Axel Visel, Len a Pennachio, Laura Elnitski, Hanna M Petrykowska, Alexej Abyzov, Bronwen Aken, Daniel Barrell, Gemma Barson, Andrew Berry, Alexandra Bignell, Veronika Boychenko, Govanni Bussotti, Claire Davidson, Gloria Despacio-Reyes, Mark Diekhans, Iakes Ezkurdia, Adam Frankish, James Gilbert, Jose Manuel Gonzalez, Ed Griffiths, Rachel Harte, David a Hendrix, Toby Hunt, Irwin Jungreis, Mike Kay, Ekta Khurana, Jing Leng, Michael F Lin, Jane Loveland, Zhi Lu, Deepa Manthravadi, Marco Mariotti, Jonathan Mudge, Gaurab Mukherjee, Cedric Notredame, Baikang Pei, Jose Manuel Rodriguez, Gary Saunders, Andrea Sboner, Stephen Searle, Cristina Sisu, Catherine Snow, Charlie Steward, Electra Tapanari, Michael L Tress, Marijke J van Baren, Stefan Washieti, Laurens Wilming, Amonida Zadissa, Zhang Zhengdong, Michael Brent, David Haussler, Alfonso Valencia, Alexandre Raymond, Nick Addleman, Roger P Alexander, Raymond K Auerbach, Keith Bettinger, Nitin Bhardwaj, Alan P Boyle, Alina R Cao, Philip Cayting, Alexandra Charos, Yong Cheng, Catharine Eastman, Ghia Euskirchen, Joseph D Fleming, Fabian Grubert, Lukas Habegger, Manoj Hariharan, Arif Harmanci, Susma Iyenger, Victor X Jin, Konrad J Karczewski, Maya Kasowski, Phil Lacroute, Hugo Lam, Nathan Larnarre-Vincent, Jin Lian, Marianne Lindahl-Allen, Renqiang

Min, Benoit Miotto, Hannah Monahan, Zarmik Moqtaderi, Xinmeng J Mu, Henriette O'Geen, Zhengqing Ouyang, Dorrelyn Patacsil, Debasish Raha, Lucia Ramirez, Brian Reed, Minyi Shi, Teri Slifer, Heather Witt, Linfeng Wu, Xiaoqin Xu, Koon-Kiu Yan, Xinqiong Yang, Kevin Struhl, Sherman M Weissman, Scott a Tenebaum, Luiz O Penalva, Subhradip Karmakar, Raj R Bhanvadia, Alina Choudhury, Marc Domanus, Lijia Ma, Jennifer Moran, Alec Victorsen, Thomas Auer, Lazaro Centarin, Michael Eichenlaub, Franziska Gruhl, Stephan Heerman, Burkard Hoeckendorf, Daigo Inoue, Tanja Kellner, Stephan Kirchmaier, Claudia Mueller, Robert Reinhardt, Lea Schertel, Stephanie Schneider, Rebecca Sinn, Beate Wittbrodt, Jochen Wittbrodt, Gaurav Jain, Gayathri Balasundaram, Daniel L Bates, Rachel Byron, Theresa K Canfield, Morgan J Diegel, Douglas Dunn, Abigail K Ebersol, Tristan Frum, Kavita Garg, Erica Gist, R Scott Hansen, Lisa Boatman, Eric Haugen, Richard Humbert, Audra K Johnson, Ericka M Johnson, Tattayana M Kutyavin, Kristin Lee, Dimitra Lotakis, Matthew T Maurano, Shane J Neph, Fiedencio V Neri, Eric D Nguyen, Hongzhu Qu, Alex P Reynolds, Vaughn Roach, Eric Rynes, Minerva E Sanchez, Richard S Sandstrom, Anthony O Shafer, Andrew B Stergachis, Sean Thomas, Benjamin Vernot, Jeff Vierstra, Shinny Vong, Molly a Weaver, Yongqi Yan, Miaohua Zhang, Joshua a Akey, Michael Bender, Michael O Dorschner, Mark Groudine, Michael J MacCoss, Patrick Navas, George Stamatoyannopoulos, John a Stamatoyannopoulos, Kathryn Beal, Alvis Brazma, Paul Flicek, Nathan Johnson, Margus Lukk, Nicholas M Luscombe, Daniel Sobral, Juan M Vaquerizas, Serafim Batzoglou, Arend Sidow, Nadine Hussami, Sofia Kyriazopoulou-Panagiotopoulou, Max W Libbrecht, Marc a Schaub, Webb Miller, Peter J Bickel, Balazs Banfai, Nathan P Boley, Haiyan Huang, Jingyi Jessica Li, William Stafford Noble, Jeffrey a Bilmes, Orion J Buske, Avinash O Sahu, Peter V Kharchenko, Peter J Park, Dannon Baker, James Taylor, and Lucas Lochovsky. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012. ISSN 1476-4687. doi: 10.1038/nature11247. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3439153{&}tool=pmcentrez{&}rendertype=abstract.

[7] Don W. Cleveland, Yinghui Mao, and Kevin F. Sullivan. Centromeres and kinetochores: From epigenetics to mitotic checkpoint signaling. *Cell*, 112(4):407–421, 2003. ISSN 00928674. doi: 10.1016/S0092-8674(03)00115-6.

[8] Centromere DNA. Sandwalk - centromere dna. http://sandwalk.blogspot.co.uk/2008/05/centromere-dna.html. Accessed: 2016-03-02.

[9] UT Southwestern. Facts about telomeres and telomerase: Shay/wright lab - ut southwestern. http://www.utsouthwestern.edu/labs/shay-wright/research/facts-about-telomeres-telomerase.html. Accessed: 2016-03-03.

[10] Noncoding dna. https://www.boundless.com/biology/textbooks/boundless-biology-textbook/evolution-and-the-origin-of-species-18/evolution-of-genomes-127/noncoding-dna-512-13092/. Accessed: 2016-03-02.

[11] Nickolay Neznanov, Akihiro Umezawa, and Robert G Oshima. A Regulatory Element within a Coding Exon Modulates Keratin 18 Gene Expression in Transgenic Mice. *The Journal of Biological Chemistry*, 272(44):27549–27557, 1997.

[12] Attila Reményi, Hans R Schöler, and Matthias Wilmanns. Combinatorial control of gene expression. *nature structural & molecular biology*, 11(9):812–815, 2004. doi: 10.1038/nsmb820.

[13] Juan M. Vaquerizas, Sarah K. Kummerfeld, Sarah A. Teichmann, and Nicholas M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10:252 EP –, 04 2009. URL https://doi.org/10.1038/nrg2538.

[14] Mark J Solomon, Pamela L Larsen, and Alexander Varshavsky. Mapping Protein-DNA Interactions In Vivo with Formaldehyde : Evidence That Histone H4 Is Retained on a Highly Transcribed Gene. *Cell*, 53:937–947, 1988.

[15] Elaine R Mardis. Next-Generation DNA Sequencing Methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402, 2008. doi: 10.1146/annurev.genom.9.081307.164359.

[16] Marloes R Tijssen, Ana Cvejic, Anagha Joshi, Rebecca L Hannah, Rita Ferreira, Ariel Forrai, Dana C Bellissimo, S Helen Oram, Peter A Smethurst, Nicola K Wilson, Xiaonan Wang, Katrin Ottersbach, Derek L Stemple, Anthony R. Green, Willem H. Ouwehand, and Berthold Gottgens. Article RUNX1 , FLI1 , and SCL Binding in Megakaryocytes Identifies Hematopoietic Regulators. *Developmental Cell*, 20:597–609, 2011. doi: 10.1016/j.devcel.2011.04.008.

[17] Robin K Richmond, David F Sargent, Timothy J Richmond, Karolin Luger, and Armin W Ma. Crystal structure of the nucleosome core particle at particle at 2.8 A resolution. *Nature*, 7:251–260, 1997.

[18] Timothy J Richmond and Curt A Davey. The structure of DNA in the nucleosome core. *Nature*, 423:145–150, 2003.

[19] Roger D. Kornberg. Chromatin structure: A repeating unit of histones and dna. *Science*, 184:868–871, 1974.

[20] Tony Kouzarides. Review Chromatin Modifications and Their Function. *Cell*, pages 693–705, 2007. doi: 10.1016/j.cell.2007.02.005.

[21] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin , DNA-binding proteins and nucleosome position. *Nature methods*, 10(12), 2013. doi: 10.1038/nmeth.2688.

[22] Vicky W Zhou, Alon Goren, and Bradley E Bernstein. Charting histone modifications and the functional organization of mammalian genomes. *Nature Publishing Group*, 12(1):7–18, 2011. ISSN 1471-0056. doi: 10.1038/nrg2905. URL http://dx.doi.org/10.1038/nrg2905.

[23] Menno P Creyghton, Albert W Cheng, G Grant Welstead, Tristan Kooistra, Bryce W Carey, Eveline J. Steinea, Jacob Hanna, Michael A. Lodato, Garrett M. Frampton, Phillip A. Sharp, Laurie A. Boyer, Richard A. Young, and Rudolf Jaenisch. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *PNAS*, 107(50):21931–21936, 2010. doi: 10.1073/pnas.1016071107.

[24] Axel Visel, Edward M Rubin, Len A Pennacchio, Genomics Division, and Lawrence Berkeley. Genomic Views of Distant-Acting Enhancers Axel. *Nature*, 461(7261): 199–205, 2009. doi: 10.1038/nature08451.Genomic.

[25] Ivan Krivega and Ann Dean. Enhancer and promoter interactions — long distance calls. *Current Opinion in Genetics & Development*, 22(2):79–85, 2012. doi: 10.1016/j.gde.2011.11.001.Enhancer.

[26] Britta A M Bouwman and Wouter De Laat. Getting the genome in shape : the formation of loops , domains and compartments. *Genome Biology*, pages 1–9, 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0730-1. URL http://dx.doi.org/10.1186/s13059-015-0730-1.

[27] Denes Hnisz, Brian J Abraham, Tong Ihn Lee, Ashley Lau, Violaine Saint-andre, Alla A Sigova, Heather A Hoke, and Richard A Young. Resource Super-Enhancers in the Control of Cell Identity and Disease. *Cell*, 155:934–947, 2013. doi: 10.1016/j.cell.2013.09.053.

[28] Romina Petersen, John J Lambourne, Biola M Javierre, Luigi Grassi, Roman Kreuzhuber, Heather Elding, Johanna P Van Geffen, Tao Jiang, Dace Ruklisa, Isabel M Rosa, Ana R Tome, Samantha Farrow, Jonathan Cairns, Abeer M Al-subaie, Sofie Ashford, Antony Attwood, Joana Batista, Heleen Bouman, Frances Burden, Fizzah A Choudry, Laura Clarke, Paul Flicek, Stephen F Garner, Matthias Haimel, Carly Kempster, Vasileios Ladopoulos, An-sofie Lenaerts, Paulina M Materek, Harriet Mckinney, Stuart Meacham, Daniel Mead, Magdolna Nagy, Christopher J Penkett, Augusto Rendon, Denis Seyres, Benjamin Sun, Salih Tuna, Marie-elise Van Der Weide, Steven W Wingett, Joost H Martens, Oliver Stegle, and Sylvia Richardson. Platelet function is modified by common sequence variation in megakaryocyte super enhancers. *Nature Communications*, (May), 2017. doi: 10.1038/ncomms16058.

[29] Bryan R. Lajoie, Job Dekker, and Noam Kaplan. The Hitchhiker's guide to Hi-C analysis: Practical guidelines. *Methods*, 72:65–75, 2015. ISSN 10462023. doi: 10.1016/j.ymeth.2014.10.031. URL http://linkinghub.elsevier.com/retrieve/pii/S1046202314003582.

[30] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science (New York, N.Y.)*, 295(5558):1306–11, 2002. ISSN 1095-9203. doi: 10.1126/science.1067799. URL http://www.ncbi.nlm.nih.gov/pubmed/11847345.

[31] Borbala Mifsud, Filipe Tavares-Cadete, Alice N Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W Wingett, Simon Andrews, William Grey, Philip a Ewels, Bram Herman, Scott Happe, Andy Higgs, Emily LeProust, George a Follows, Peter Fraser, Nicholas M Luscombe, and Cameron S Osborne. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, (April):1–12, 2015. ISSN 1061-4036. doi: 10.1038/ng.3286. URL http://www.nature.com/doifinder/10.1038/ng.3286.

[32] Erez Lieberman-aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O

Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, and Leonid A Mirny. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 33292(October):289–294, 2009.

[33] Aylwyn Scally. ScienceDirect The mutation rate in human evolution and demographic inference. *Current Opinion in Genetics & Development*, 41:36–43, 2016. ISSN 0959-437X. doi: 10.1016/j.gde.2016.07.008. URL http://dx.doi.org/10.1016/j.gde.2016.07.008.

[34] Ian M Campbell, Tomasz Gambin, Shalini N Jhangiani, Megan L Grove, Narayanan Veeraraghavan, Donna M Muzny, Chad A Shaw, Richard A Gibbs, Eric Boerwinkle, Fuli Yu, and James R Lupski. Multiallelic Positions in the Human Genome: Challenges for Genetic Analyses. *Human Mutation*, pages 1–4, 2015. doi: 10.1002/humu.22944.

[35] Tanita Casci. Population genetics: SNPs that come in threes. *Nature reviews. Genetics*, 11(1):8, 2010. ISSN 1471-0056. doi: 10.1038/nrg2725. URL http://dx.doi.org/10.1038/nrg2725.

[36] Jeff M. Hall, Ming K. Lee, Beth Newman, Jan E. Morrow, Lee A. Anderson, Bing Huey, and Mary-Claire King. Linkage of Early-Onset Familial Breast Cancer to Chromosome 17q21. *Science*, 250, 1990.

[37] Steven A. Narod, Jean Feunteun, Henry T. Lynch, Patrice Watson, Theresa Conway, Jane Lynch, and Gilbert M. Lenoir. Familial breast-ovarian cancer locus on chromosome 17q12-q23. *The Lancet*, 338:82–83, 1991.

[38] Jean Feunteun and Gilbert M. Lenoir. BRCA1, a gene involved in inherited predisposition to breast and ovarian cancer. *Biochimica et biophysica acta*, 1242, 1996.

[39] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38 (16):1–7, 2010. ISSN 03051048. doi: 10.1093/nar/gkq603.

[40] William McLaren, Bethan Pritchard, Daniel Rios, Yuan Chen, Paul Flicek, and Fiona Cunningham. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics (Oxford, England)*, 26 (16):2069–70, 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq330. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2916720{&}tool=pmcentrez{&}rendertype=abstract.

[41] Wenqiang Shi, Oriol Fornes, Anthony Mathelier, and Wyeth W. Wasserman. Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Research*, 44(21):10106–10116, 2016. ISSN 13624962. doi: 10.1093/nar/gkw691.

[42] Ashley K. Tehranchi, Marsha Myrthil, Trevor Martin, Brian L. Hie, David Golan, and Hunter B. Fraser. Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. *Cell*, 165(3):730–741, 2016. ISSN 10974172. doi: 10.1016/j.cell.2016.03.041. URL http://dx.doi.org/10.1016/j.cell.2016.03.041.

[43] Dirk S. Paul, Nicole Soranzo, and Stephan Beck. Functional interpretation of non-coding sequence variation: Concepts and challenges. *BioEssays*, 36(2):191–199, 2014. ISSN 02659247. doi: 10.1002/bies.201300126.

[44] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(June), 2007. doi: 10.1038/nature05911.

[45] William J Astle, Heather Elding, Tao Jiang, Willem H Ouwehand, Adam S Butterworth, Nicole Soranzo, William J Astle, Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L Mann, and Daniel Mead. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease Resource The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*, 167:1415–1429, 2016. doi: 10.1016/j.cell.2016.10.042.

[46] Derek Klarin, Qiuyu Martin Zhu, Connor A Emdin, Mark Chaffin, Steven Horner, Brian J Mcmillan, Alison Leed, E Michael, Chris C A Spencer, François Aguet, Ayellet V Segrè, Kristin Ardlie, Amit V. Khera, Virendar K. Kaushik, Pradeep Natarajan, CARDIoGRAMplusC4D Consortium, and Sekar Kathiresan. Genetic Analysis in UK Biobank Links Insulin Resistance and Transendothelial Migration Pathways to Coronary Artery Disease. *Nature Genetics*, 49(9):1392–1397, 2017. doi: 10.1038/ng.3914.Genetic.

[47] Helen R Warren and UK Biobank. Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nature Genetics*, 49(3), 2017. doi: 10.1038/ng.3768.

[48] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562, 2018. doi: 10.1038/s41586-018-0579-z.

[49] Lloyd T Elliott, Kevin Sharp, Fidel Alfaro-almagro, Sinan Shi, Karla L Miller, Gwenaëlle Douaud, Jonathan Marchini, and Stephen M Smith. Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature*, 562, 2018. doi: 10.1038/s41586-018-0571-7.

[50] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):1001–1006, 2014. ISSN 03051048. doi: 10.1093/nar/gkt1229.

[51] Christa Meisinger, Holger Prokisch, Christian Gieger, Nicole Soranzo, Divya Mehta, Dieter Rosskopf, Peter Lichtner, Norman Klopp, Jonathan Stephens, Nicholas A. Watkins, Panos Deloukas, Andreas Greinacher, Wolfgang Koenig, Matthias Nauck, Christian Rimmbach, Henry Völzke, Annette Peters, Thomas Illig, Willem H. Ouwehand, Thomas Meitinger, H.-Erich Wichmann, and Angela Döring. A Genome-wide

Association Study Identifies Three Loci Associated with Mean Platelet Volume. *AJHG*, 84:66–71, 2009. doi: 10.1016/j.ajhg.2008.11.015.

[52] Nicole Soranzo, Augusto Rendon, Christian Gieger, Chris I Jones, Nicholas A Watkins, Stephan Menzel, Jonathan Stephens, Holger Prokisch, Wendy Erber, Simon C Potter, Sarah L Bray, Philippa Burns, Jennifer Jolley, Mario Falchi, Brigitte Ku, Jeanette Erdmann, Heribert Schunkert, Nilesh J Samani, Thomas Illig, Stephen F Garner, Angela Rankin, Christa Meisinger, John R Bradley, Swee Lay Thein, Alison H Goodall, Tim D Spector, Panos Deloukas, and Willem H Ouwehand. A novel variant on chromosome 7q22.3 associated with mean platelet volume, counts, and function. *Blood*, 113(16):3831–3838, 2019. doi: 10.1182/blood-2008-10-184234.

[53] Nicole Soranzo, Tim D Spector, Massimo Mangino, Brigitte Kühnel, Augusto Rendon, Alexander Teumer, Christina Willenborg, Benjamin Wright, Li Chen, Mingyao Li, Perttu Salo, Benjamin F Voight, Philippa Burns, Roman A Laskowski, Yali Xue, Stephan Menzel, David Altshuler, John R Bradley, Suzannah Bumpstead, Mary-susan Burnett, Joseph Devaney, Angela Döring, Roberto Elosua, Stephen E Epstein, Wendy Erber, Mario Falchi, Stephen F Garner, Mohammed J R Ghori, Alison H Goodall, Rhian Gwilliam, Hakon H Hakonarson, Alistair S Hall, Naomi Hammond, Christian Hengstenberg, Thomas Illig, Inke R König, Christopher W Knouff, Ruth Mcpherson, Olle Melander, Vincent Mooser, Matthias Nauck, Markku S Nieminen, Christopher J O Donnell, Leena Peltonen, Simon C Potter, Holger Prokisch, Daniel J Rader, Heribert Schunkert, Stephen M Schwartz, Jovana Serbanovic-canic, Juha Sinisalo, David S Siscovick, Klaus Stark, Ida Surakka, Jonathan Stephens, John R Thompson, Uwe Völker, Henry Völzke, Nicholas A Watkins, George A Wells, H-erich Wichmann, David A Van Heel, Chris Tyler-smith, Swee Lay Thein, Sekar Kathiresan, Markus Perola, Muredach P Reilly, Alexandre F R Stewart, Jeanette Erdmann, Nilesh J Samani, Christa Meisinger, Andreas Greinacher, Panos Deloukas, Willem H Ouwehand, and Christian Gieger. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nature Genetics*, 41(11):1182–1190, 2009. ISSN 1061-4036. doi: 10.1038/ng.467. URL http://dx.doi.org/10.1038/ng.467.

[54] C Gieger, A Radhakrishnan, A Cvejic, W Tang, E Porcu, G Pistis, J Serbanovic-Canic, U Elling, AH Goodall, Y Labrune, LM Lopez, R Mägi, S Meacham, Y Okada, N Pirastu, R Sorice, A Teumer, K Voss, W Zhang, R Ramirez-Solis, JC Bis, D Ellinghaus, M Gögele, JJ Hottenga, C Langenberg, P Kovacs, PF O'Reilly, SY Shin, T Esko, J Hartiala, S Kanoni, F Murgia, A Parsa, J Stephens, P van der Harst, C Ellen van der Schoot, H Allayee, A Attwood, B Balkau, F Bastardot, S Basu, SE Baumeister, G Biino, L Bomba, A Bonnefond, F Cambien, JC Chambers, F Cucca, P D'Adamo, G Davies, RA de Boer, EJ de Geus, A Döring, P Elliott, J Erdmann, DM Evans, M Falchi, W Feng, AR Folsom, IH Frazer, QD Gibson, NL Glazer, C Hammond, AL Hartikainen, SR Heckbert, C Hengstenberg, M Hersch, T Illig, RJ Loos, J Jolley, KT Khaw, B Kühnel, MC Kyrtsonis, V Lagou, H Lloyd-Jones, T Lumley, M Mangino, A Maschio, IM Leach, B McKnight, Y Memari, BD Mitchell, GW Montgomery, Y Nakamura, M Nauck, G Navis, U Nöthlings, IM Nolte, DJ Porteous, A Pouta, PP Pramstaller, J Pullat, SM Ring, JI Rotter, D Ruggiero, A Ruokonen, C Sala, NJ Samani, J Sambrook, D Schlessinger, S Schreiber, H Schunkert, J Scott, NL Smith, H Snieder, JM Starr, M Stumvoll, A Takahashi, WH Tang, K Taylor,

A Tenesa, S Lay, Thein, A Tönjes, M Uda, S Ulivi, DJ van, Veldhuisen, PM Visscher, U Völker, HE Wichmann, KL Wiggins, G Willemsen, TP Yang, J Hua, Zhao, P Zitting, JR Bradley, GV Dedoussis, P Gasparini, SL Hazen, A Metspalu, M Pirastu, AR Shuldiner, L Joost van Pelt, JJ Zwaginga, DI Boomsma, IJ Deary, A Franke, P Froguel, SK Ganesh, MR Jarvelin, NG Martin, C Meisinger, BM Psaty, TD Spector, NJ Wareham, JW Akkerman, M Ciullo, P Deloukas, A Greinacher, S Jupe, N Kamatani, J Khadake, JS Kooner, J Penninger, I Prokopenko, D Stemple, D Toniolo, L Wernisch, S Sanna, AA Hicks, A Rendon, MA Ferreira, WH Ouwehand, and N Soranzo. New gene functions in megakaryopoiesis and platelet formation. *Nature*, 480(7376):201–208, 2011. ISSN 0028-0836. doi: 10.1038/nature10659. URL http://dx.doi.org/10.1038/nature10659.

[55] P van der Harst, W Zhang, I Mateo, Leach, A Rendon, N Verweij, J Sehmi, DS Paul, U Elling, H Allayee, X Li, A Radhakrishnan, ST Tan, K Voss, CX Weichenberger, CA Albers, A Al-Hussani, FW Asselbergs, M Ciullo, F Danjou, C Dina, T Esko, DM Evans, L Franke, M Gögele, J Hartiala, M Hersch, H Holm, JJ Hottenga, S Kanoni, ME Kleber, V Lagou, C Langenberg, LM Lopez, LP Lyytikäinen, O Melander, F Murgia, IM Nolte, PF O'Reilly, S Padmanabhan, A Parsa, N Pirastu, E Porcu, L Portas, I Prokopenko, JS Ried, SY Shin, CS Tang, A Teumer, M Traglia, S Ulivi, HJ Westra, J Yang, JH Zhao, F Anni, A Abdellaoui, A Attwood, B Balkau, S Bandinelli, F Bastardot, B Benyamin, BO Boehm, WO Cookson, D Das, PI de Bakker, RA de Boer, EJ de Geus, MH de Moor, M Dimitriou, FS Domingues, A Döring, G Engström, GI Eyjolfsson, L Ferrucci, K Fischer, R Galanello, SF Garner, B Genser, QD Gibson, G Girotto, DF Gudbjartsson, SE Harris, AL Hartikainen, CE Hastie, B Hedblad, T Illig, J Jolley, M Kähönen, IP Kema, JP Kemp, L Liang, H Lloyd-Jones, RJ Loos, S Meacham, SE Medland, C Meisinger, Y Memari, E Mihailov, K Miller, MF Moffatt, M Nauck, M Novatchkova, T Nutile, I Olafsson, PT Onundarson, D Parracciani, BW Penninx, L Perseu, A Piga, G Pistis, A Pouta, U Puc, O Raitakari, SM Ring, A Robino, D Ruggiero, A Ruokonen, A Saint-Pierre, C Sala, A Salumets, J Sambrook, H Schepers, CO Schmidt, HH Silljé, R Sladek, JH Smit, JM Starr, J Stephens, P Sulem, T Tanaka, U Thorsteinsdottir, V Tragante, WH van Gilst, LJ van Pelt, DJ van Veldhuisen, U Völker, JB Whitfield, G Willemsen, BR Winkelmann, G Wirnsberger, A Algra, F Cucca, AP D'Adamo, J Danesh, IJ Deary, AF Dominiczak, P Elliott, P Fortina, P Froguel, P Gasparini, A Greinacher, SL Hazen, MR Jarvelin, KT Khaw, T Lehtimäki, W Maerz, NG Martin, A Metspalu, BD Mitchell, GW Montgomery, C Moore, G Navis, M Pirastu, PP Pramstaller, R Ramirez-Solis, E Schadt, J Scott, AR Shuldiner, GD Smith, JG Smith, H Snieder, R Sorice, TD Spector, K Stefansson, M Stumvoll, WH Tang, D Toniolo, A Tönjes, PM Visscher, P Vollenweider, NJ Wareham, BH Wolffenbuttel, DI Boomsma, JS Beckmann, GV Dedoussis, P Deloukas, MA Ferreira, S Sanna, M Uda, AA Hicks, JM Penninger, C Gieger, JS Kooner, WH Ouwehand, N Soranzo, and JC. Chambers. Seventy-five genetic loci influencing the human red blood cell. *Nature*, 492(7429):369–375, 2012. ISSN 0028-0836. doi: 10.1038/nature11677. URL http://dx.doi.org/10.1038/nature11677.

[56] Matthew V. Rockman and Leonid Kruglyak. Genetics of global gene expression. *Nature Reviews Genetics*, 7(11):862–872, 2006. ISSN 1471-0056. doi: 10.1038/nrg1964. URL http://www.nature.com/doifinder/10.1038/nrg1964.

[57] A.A. Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012. URL http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/.

[58] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. GCTA : A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011. ISSN 0002-9297. doi: 10.1016/j.ajhg.2010.11.011. URL http://dx.doi.org/10.1016/j.ajhg.2010.11.011.

[59] C. Lippert, F. P. Casale, B. Rakitsch, and O. Stegle. LIMIX: genetic analysis of multiple traits. *bioRxiv*, page 003905, 2014. doi: 10.1101/003905. URL http://biorxiv.org/content/early/2014/05/21/003905.abstract.

[60] Po-ru Loh, George Tucker, Brendan K Bulik-sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, Nick Patterson, and Alkes L Price. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Publishing Group*, 47 (3):284–290, 2015. ISSN 1061-4036. doi: 10.1038/ng.3190. URL http://dx.doi.org/10.1038/ng.3190.

[61] H. Guo, M. D. Fortune, O. S. Burren, E. Schofield, J. a. Todd, and C. Wallace. Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Human Molecular Genetics*, 24(12):3305–3313, 2015. ISSN 0964-6906. doi: 10.1093/hmg/ddv077. URL http://www.hmg.oxfordjournals.org/cgi/doi/10.1093/hmg/ddv077.

[62] Avinash Das, Michael Morley, Christine S. Moravec, W. H. W. Tang, Hakon Hakonarson, Euan a. Ashley, Jeffrey Brandimarto, Ray Hu, Mingyao Li, Hongzhe Li, Yichuan Liu, Liming Qu, Pablo Sanchez, Kenneth B. Margulies, Thomas P. Cappola, Shane Jensen, and Sridhar Hannenhalli. Bayesian integration of genetics and epigenetics detects causal regulatory SNPs underlying expression variability. *Nature Communications*, 6:8555, 2015. ISSN 2041-1723. doi: 10.1038/ncomms9555. URL http://www.nature.com/doifinder/10.1038/ncomms9555.

[63] Tomi Pastinen, Robert Sladek, Scott Gurd, Alya Sammak, Bing Ge, Pierre Lepage, Karine Lavergne, Amelie Villeneuve, Tiffany Gaudin, Helena Brandstrom, Andrei Verner, Jade Kingsley, Eef Harmsen, Damian Labuda, Kenneth Morgan, Marie-claude Vohl, Anna K Naumova, Daniel Sinnett, and Thomas J Hudson. A survey of genetic and epigenetic variation affecting human gene expression. *Physiol Genomics*, 16: 184–193, 2003.

[64] Matthew Freeman, Angela Risch, Christoph Plass, Graham Casey, Mariella De Biasi, and Chris Carlson. Principles for the post-GWAS functional charactarization of cancer risk loci. *Nat Genet*, 43(6):513–518, 2011. doi: 10.1038/ng.840.Principles.

[65] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, Susan Sullivan, Jason Bridge, Harold Magazine, John Syron, Johnelle Fleming, Laura Siminoff, Heather Traino, Maghboeba Mosavel, Laura Barker, Scott Jewell, Dan Rohrer, Dan Maxim,

Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Eric Hudson, Kristin Feenstra, Leslie Sobin, James Robb, Phillip Branton, Greg Korzeniewski, Charles Shive, David Tabor, Liqun Qi, Kevin Groch, Sreenath Nampally, Steve Buia, Angela Zimmerman, Anna Smith, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, Mark Cosentino, Norma Diaz-Mayoral, Mary Kennedy, Theresa Engel, Penelope Williams, Kenyon Erickson, Kristin Ardlie, Wendy Winckler, Gad Getz, David DeLuca, Daniel MacArthur, Manolis Kellis, Alexander Thomson, Taylor Young, Ellen Gelfand, Molly Donovan, Yan Meng, George Grant, Deborah Mash, Yvonne Marcus, Margaret Basile, Jun Liu, Jun Zhu, Zhidong Tu, Nancy J Cox, Dan L Nicolae, Eric R Gamazon, Hae Kyung Im, Anuar Konkashbaev, Jonathan Pritchard, Matthew Stevens, Timothèe Flutre, Xiaoquan Wen, Emmanouil T Dermitzakis, Tuuli Lappalainen, Roderic Guigo, Jean Monlong, Michael Sammeth, Daphne Koller, Alexis Battle, Sara Mostafavi, Mark McCarthy, Manual Rivas, Julian Maller, Ivan Rusyn, Andrew Nobel, Fred Wright, Andrey Shabalin, Mike Feolo, Nataliya Sharopova, Anne Sturcke, Justin Paschal, James M Anderson, Elizabeth L Wilder, Leslie K Derr, Eric D Green, Jeffery P Struewing, Gary Temple, Simona Volpi, Joy T Boyer, Elizabeth J Thomson, Mark S Guyer, Cathy Ng, Assya Abdallah, Deborah Colantuoni, Thomas R Insel, Susan E Koester, a Roger Little, Patrick K Bender, Thomas Lehner, Yin Yao, Carolyn C Compton, Jimmie B Vaught, Sherilyn Sawyer, Nicole C Lockhart, Joanne Demchok, and Helen F Moore. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 2013. ISSN 1061-4036. doi: 10.1038/ng.2653. URL http://www.nature.com/doifinder/10.1038/ng.2653.

[66] Robert Klein. Power analysis for genome-wide association studies. *BMC Genetics*, 8 (1):58+, 2007. ISSN 1471-2156. doi: 10.1186/1471-2156-8-58. URL http://dx.doi.org/10.1186/1471-2156-8-58.

[67] Paul Schliekelman. Statistical power of expression quantitative trait loci for mapping of complex trait loci in natural populations. *Genetics*, 178(4):2201–2216, 2008. ISSN 00166731. doi: 10.1534/genetics.107.076687.

[68] Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, Beverley Balkau, Barbara Heude, Guillaume Charpentier, Thomas J. Hudson, Alexandre Montpetit, Alexey V. Pshezhetsky, Marc Prentki, Barry I. Posner, David J. Balding, David Meyre, Constantin Polychronakos, and Philippe Froguel. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130): 881–885, 2007. ISSN 00280836. doi: 10.1038/nature05616.

[69] Scott Smemo, Juan J. Tena, Kyoung-Han Kim, Eric R. Gamazon, Noboru J. Sakabe, Carlos Gomez-Marin, Ivy Aneas, Flavia L. Credidio, Debora R. Sobreira, Nora F.Wasserman, Ju Hee Lee, Vijitha Puviindran, Davis Tam, Michael Shen, Joe Eun Son, Niki Alizadeh Vakili, Hoon-Ki Sung, Silvia Naranjo, Rafael D. Acemel, Miguel Manzanares, Andras Nagy, Nancy J. Cox, Chi-Chung Hui, Jose Luis Gomez-Skarmeta, and Marcelo A. Nobrega. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, 507, 2014. doi: 10.1038/nature13138.

[70] Melina Claussnitzer, Simon N. Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, Isabel S. Sousa, Jacqueline L. Beaudry, Vijitha Puviindran, Nezar A. Abdennur, Jannel Liu, Per-Arne Svensson, Yi-Hsiang

Hsu, Daniel J. Drucker, Gunnar Mellgren, Chi-Chung Hui, Hans Hauner, and Manolis Kellis. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *The new england journal of medicine*, 373(10):895–907, 2015. doi: 10.1056/NEJMoa1502214.

[71] Aharon Brodie, Johnathan Roy Azaria, and Yanay Ofran. How far from the SNP may the causative genes be? *Nucleic Acids Research*, 44(13):6046–6054, 2016. ISSN 13624962. doi: 10.1093/nar/gkw500.

[72] Claudia Giambartolomei, Damjan Vukcevic, Eric E. Schadt, Lude Franke, Aroon D. Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genetics*, 10(5), 2014. ISSN 15537404. doi: 10.1371/journal.pgen.1004383.

[73] Farhad Hormozdiari, Martijn Van De Bunt, Ayellet V Segre, Xiao Li, Jong Wha J Joo, Michael Bilow, Jae Hoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin. Colocalization of GWAS and eQTL Signals Detects Target Genes. *The American Journal of Human Genetics*, pages 1–16, 2016. doi: 10.1016/j.ajhg.2016.10.003.

[74] Lu Chen, Bing Ge, Francesco Paolo Casale, Louella Vasquez, Tony Kwan, Diego Garrido-Martin, Stephen Watt, Ying Yan, Kousik Kundu, Simone Ecker, Avik Datta, David Richardson, Frances Burden, Daniel Mead, Alice L. Mann, Jose Maria Fernandez, Sophia Rowlston, Steven P. Wilder, Samantha Farrow, Xiaojian Shao, John J. Lambourne, Adriana Redensek, Cornelis A. Albers, Vyacheslav Amstislavskiy, Sofie Ashford, Kim Berentsen, Lorenzo Bomba, Guillaume Bourque, David Bujold, Stephan Busche, Maxime Caron, Shu-Huang Chen, Warren Cheung, Oliver Delaneau, Emmanouil T. Dermitzakis, Heather Elding, Irina Colgiu, Frederik O. Bagger, Paul Flicek, Ehsan Habibi, Valentina Iotchkova, Eva Janssen-Megens, Bowon Kim, Hans Lehrach, Ernesto Lowy, Amit Mandoli, Filomena Matarese, Matthew T. Maurano, John A. Morris, Vera Pancaldi, Farzin Pourfarzad, Karola Rehnstrom, Augusto Rendon, Thomas Risch, Nilofar Sharifi, Marie-Michelle Simon, Marc Sultan, Alfonso Valencia, Klaudia Walter, Shuang-Yin Wang, Mattia Frontini, Stylianos E. Antonarakis, Laura Clarke, Marie-Laure Yaspo, Stephan Beck, Roderic Guigo, Daniel Rico, Joost H.A. Martens, Willem H. Ouwehand, Taco W. Kuijpers, Dirk S. Paul, Hendrik G. Stunnenberg, Oliver Stegle, Kate Downes, Tomi Pastinen, and Nicole Soranzo. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells Resource Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell*, (167):1398–1414, 2016. doi: 10.1016/j.cell.2016.10.026.

[75] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher, and Jian Yang. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, (March), 2016. ISSN 1061-4036. doi: 10.1038/ng.3538. URL http://www.nature.com/doifinder/10.1038/ng.3538.

[76] Silva Kasela, Kai Kisand, Liina Tserel, Epp Kaleviste, Anu Remm, Krista Fischer, T Esko, Harm-jan Westra, Benjamin P Fairfax, Seiko Makino, Julian C Knight, and Lili Milani. Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4 + versus CD8 + T cells. *PLoS genetics*, 2017. doi: 10.1371/journal.pgen.1006643.

[77] Gtex Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550, 2017. doi: 10.1038/nature24277.

[78] Brandon L Pierce, Lin Tong, Maria Argos, Kathryn Demanelis, Farzana Jasmine, Muhammad Rakibuz-zaman, Golam Sarwar, Tariqul Islam, Hasan Shahriar, Tariqul Islam, Mahfuzar Rahman, Muhammad G Kibriya, Lin S Chen, and Habibul Ahsan. Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. *Nature Communications*, pages 1–12, 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03209-9. URL http://dx.doi.org/10.1038/s41467-018-03209-9.

[79] Yukihide Momozawa, Julia Dmitrieva, Emilie Theatre, Valrie Deffontaine, Souad Rahmouni, Benoit Charloteaux, Francois Crins, Elisa Docampo, Mahmoud Elansary, Ann-Stephan Gori, Christelle Lecut, Rob Mariman, Myriam Mni, Cecile Oury, Ilya Altukhov, Dmitry Alexeev, Yuri Aulchenko, Leila Amininejad, Gerd Bouma, Frank Hoentjen, Mark Lowenberg, Bas Oldenburg, Marieke J. Pierik, Andrea E. vander Meulen-de Jong, C. Janneke van der Woude, Marijn C. Visschedijk, Clara Abraham, Jean-Paul Achkar, Tariq Ahmad, Ashwin N. Ananthakrishnan, Vibeke Andersen, Carl A. Anderson, Jane M. Andrews, Vito Annese, Guy Aumais, Leonard Baidoo, Robert N. Baldassano, Peter A. Bampton, Murray Barclay, Jeffrey C. Barrett, Theodore M. Bayless, Johannes Bethge, Alain Bitton, Gabrielle Boucher, Stephan Brand, Berenice Brandt, Steven R. Brant, Carsten Buning, Angela Chew, Judy H. Cho, Isabelle Cleynen, Ariella Cohain, Anthony Croft, Mark J. Daly, Mauro D'Amato, Silvio Danese, Dirk De Jong, Goda Denapiene, Lee A. Denson, Kathy L. Devaney, Olivier Dewit, Renata D'Inca, Marla Dubinsky, Richard H. Duerr, Cathryn Edwards, David Ellinghaus, Jonah Essers, Lynnette R. Ferguson, Eleonora A. Festen, Philip Fleshner, Tim Florin, Andre Franke, Karin Fransen, Richard Gearry, Christian Gieger, Jurgen Glas, Philippe Goyette, Todd Green, Anne M. Griffiths, Stephen L. Guthery, Hakon Hakonarson, Jonas Halfvarson, Katherine Hanigan, Talin Haritunians, Ailsa Hart, Chris Hawkey, Nicholas K. Hayward, Matija Hedl, Paul Henderson, Xinli Hu, Hailiang Huang, Ken Y. Hui, Marcin Imielinski, Andrew Ippoliti, Laimas Jonaitis, Luke Jostins, Tom H. Karlsen, Nicholas A. Kennedy, Mohammed Azam Khan, Gediminas Kiudelis, Krupa Krishnaprasad, Subra Kugathasan, Limas Kupcinskas, Anna Latiano, Debby Laukens, Ian C. Lawrance, James C. Lee, Charlie W. Lees, Marcis Leja, Johan Van Limbergen, Paolo Lionetti, Jimmy Z. Liu, Gillian Mahy, John Mansfield, Dunecan Massey, Christopher G. Mathew, Dermot P. B. McGovern, Raquel Milgrom, Mitja Mitrovic, Grant W. Montgomery, Craig Mowat, William Newman, Aylwin Ng, Siew C. Ng, Sok Meng Evelyn Ng, Susanna Nikolaus, Kaida Ning, Markus Nothen, Ioannis Oikonomou, Orazio Palmieri, Miles Parkes, Anne Phillips, Cyriel Y. Ponsioen, Uros Potocnik, Natalie J. Prescott, Deborah D. Proctor, Graham Radford-Smith, Jean-Francois Rahier, Soumya Raychaudhuri, Miguel Regueiro, Florian Rieder, John D. Rioux, Stephan Ripke, Rebecca Roberts, Richard K. Russell, Jeremy D. Sanderson, Miquel Sans, Jack Satsangi, Eric E. Schadt, Stefan Schreiber, Dominik Schulte, L. Philip Schumm, Regan Scott, Mark Seielstad, Yashoda Sharma, Mark S. Silverberg, Lisa A. Simms, Jurgita Skieceviciene, Sarah L. Spain, A. Hillary Steinhart, Joanne M. Stempak, Laura Stronati, Jurgita Sventoraityte, Stephan R. Targan, Kirstin M. Taylor, Anje ter Velde, Leif Torkvist, Mark Tremelling, Suzanne van Sommeren, Eric Vasiliauskas, Hein W. Verspaget, Thomas Walters, Kai Wang, Ming-Hsi Wang, Zhi Wei, David Whiteman, Cisca Wijmenga, David C. Wilson, Juliane

Winkelmann, Ramnik J. Xavier, Bin Zhang, Clarence K. Zhang, Hu Zhang, Wei Zhang, Hongyu Zhao, Zhen Z. Zhao, Mark Lathrop, Jean-Pierre Hugot, Rinse K. Weersma, Martine De Vos, Denis Franchimont, Severine Vermeire, Michiaki Kubo, Edouard Louis, Michel Georges, and The International IBD Genetics Consortium. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nature Communications*, 9(1):1–18, 2018. doi: 10.1038/s41467-018-04365-8.

[80] Rick Jansen, Jouke-jan Hottenga, Michel G Nivard, Abdel Abdellaoui, Bram Laport, Eco J De Geus, Fred A Wright, Brenda W J H Penninx, and Dorret I Boomsma. Conditional eQTL analysis reveals allelic heterogeneity of gene expression. *Human Molecular Genetics*, 26(8):1444–1451, 2018. doi: 10.1093/hmg/ddx043.

[81] Stuart H Orkin and Leonard I Zon. Hematopoiesis : An Evolving Paradigm for Stem Cell Biology. *Cell*, 132:631–644, 2008. doi: 10.1016/j.cell.2008.01.025.

[82] L. Pauling, H. A. Itano, S. J. Singer, and I. C. Wells. Sickle cell anemia, a molecular disease. *Science*, 110(2865):543–548, nov 1949. doi: 10.1126/science.110.2865.543. URL https://doi.org/10.1126/science.110.2865.543.

[83] David Bryder, Derrick J. Rossi, and Irving L. Weissman. Hematopoietic Stem Cells. *The American Journal of Pathology*, 169(2):338–346, 2006. doi: 10.2353/ajpath.2006.060312.

[84] Martin Korbling and Emil J Freireich. Twenty-five years of peripheral blood stem cell transplantation. *Blood*, 117(24):6411–6417, 2011. doi: 10.1182/blood-2010-12-322214.

[85] By Jean C Y Wang, Monica Doedens, and John E Dick. Primitive Human Hematopoietic Cells Are Enriched in Cord Blood Compared With Adult Bone Marrow or Mobilized Peripheral Blood as Measured by the Quantitative In Vivo SCID-Repopulating Cell Assay. *Blood*, 89(11), 1997.

[86] Jun Seita and Irving L Weissman. Hematopoietic stem cell : self-renewal versus differentiation. *Wiley Interdiscip Rev Syst Biol Med*, 2, 2010. doi: 10.1002/wsbm.86.

[87] B P Fairfax, S Makino, J Radhakrishnan, K Plant, S Leslie, a Dilthey, P Ellis, C Langford, F O Vannberg, and J C Knight. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat Genet*, 44(5):502–510, 2012. ISSN 1061-4036. doi: 10.1038/ng.2205. URL http://www.ncbi.nlm.nih.gov/pubmed/22446964.

[88] Matthias Heinig, Enrico Petretto, Chris Wallace, Leonardo Bottolo, Maxime Rotival, Han Lu, Yoyo Li, Rizwan Sarwar, Sarah R Langley, Anja Bauerfeind, Oliver Hummel, Young-ae Lee, Svetlana Paskas, Carola Rintisch, Kathrin Saar, Jason Cooper, Rachel Buchan, Elizabeth E Gray, Jason G Cyster, Cardiogenics Consortium, Jeanette Erdmann, Thomas Mu, Timothy J Aitman, Francois Cambien, David Clayton, John A Todd, Norbert Hubner, and Stuart A Cook. A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk Matthias. *Nature*, 467(7314):460–464, 2010. ISSN 0028-0836. doi: 10.1038/nature09386. URL http://dx.doi.org/10.1038/nature09386.

[89] Sophie Garnier, Vinh Truong, Jessy Brocheton, Tanja Zeller, Maxime Rovital, Philipp S Wild, Andreas Ziegler, The Cardiogenics Consortium, Thomas Munzel, Laurence Tiret, Stefan Blankenberg, Panos Deloukas, Jeannette Erdmann, Christian Hengstenberg, Nilesh J Samani, Heribert Schunkert, Willem H Ouwehand, Alison H Goodall, Francois Cambien, and David-Alexandre Trégouët. Genome-Wide Haplotype Analysis of Cis Expression Quantitative Trait Loci in Monocytes. *PLoS Genetics*, 9(1):1–11, 2013. doi: 10.1371/journal.pgen.1003240.

[90] Lu Zhou, Rajesh Somasundaram, Rosa F Nederhof, Gerard Dijkstra, Klaas Nico Faber, Maikel P Peppelenbosch, and Gwenny M Fuhler. Impact of Human Granulocyte and Monocyte Isolation Procedures on Functional Studies. *Clinical and Vaccine Immunology*, 19(7):1065–1074, 2012. doi: 10.1128/CVI.05715-11.

[91] Stefan Amisten. *A Rapid and Efficient Platelet Purification Protocol for Platelet Gene Expression Studies*, pages 155–172. Springer New York, New York, NY, 2012. ISBN 978-1-61779-307-3. doi: 10.1007/978-1-61779-307-3_12. URL https://doi.org/10.1007/978-1-61779-307-3_12.

[92] Diagram showing the development of different blood cells from haematopoietic stem cell to mature cells. https://upload.wikimedia.org/wikipedia/commons/f/f0/Hematopoiesis_simple.svg.

[93] D. Stites, A. Terr, and T. Parslow. *Medical immunology*. Lange Medical Publishers, Stamford, 1997. ISBN 9780838562789.

[94] A.V. Hoffbrand and P.A.H. Moss. *Essential Haematology*. Wiley-Blackwell, Oxford, 6th edition, 2011. ISBN 9781405198905.

[95] Christopher Bishop. *Pattern recognition and machine learning*. Springer, New York, 2006. ISBN 0-387-31073-8.

[96] Addison Greenwood. *Science at the frontier*. National Academy Press, Washington, D.C, 1992. ISBN 978-0-309-04592-6.

[97] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

[98] Frank Rosenblatt. The perceptron - a perceiving and recognizing automaton, 1957.

[99] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and Understanding Recurrent Networks. *arXiv preprint*, pages 1–12, 2015. ISSN 978-3-319-10589-5. doi: 10.1007/978-3-319-10590-1_53. URL http://arxiv.org/abs/1506.02078.

[100] UK10K. Uk10k. https://www.uk10k.org/.

[101] Harm-Jan Westra, Marjolein J. Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W. Christiansen, Benjamin P. Fairfax, Katharina Schramm, Joseph E. Powell, Alexandra Zhernakova, Daria V. Zhernakova, Jan H. Veldink, Leonard H. Van den Berg, Juha Karjalainen, Sebo Withoff, André G. Uitterlinden, Albert Hofman, Fernando Rivadeneira, Peter a. C. 't Hoen, Eva Reinmaa, Krista Fischer, Mari Nelis, Lili Milani, David Melzer, Luigi Ferrucci, Andrew B. Singleton, Dena G. Hernandez, Michael a. Nalls, Georg Homuth, Matthias Nauck, Dörte

Radke, Uwe Völker, Markus Perola, Veikko Salomaa, Jennifer Brody, Astrid Suchy-Dicey, Sina a. Gharib, Daniel a. Enquobahrie, Thomas Lumley, Grant W. Montgomery, Seiko Makino, Holger Prokisch, Christian Herder, Michael Roden, Harald Grallert, Thomas Meitinger, Konstantin Strauch, Yang Li, Ritsert C. Jansen, Peter M. Visscher, Julian C. Knight, Bruce M. Psaty, Samuli Ripatti, Alexander Teumer, Timothy M. Frayling, Andres Metspalu, Joyce B. J. van Meurs, and Lude Franke. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243, 2013. ISSN 1061-4036. doi: 10.1038/ng.2756. URL http://www.nature.com/ng/journal/v45/n10/full/ng.2756.html.

[102] Fabian Grubert, Judith B Zaugg, Lars M Steinmetz, Michael Snyder, Fabian Grubert, Judith B Zaugg, Maya Kasowski, Oana Ursu, Damek V Spacek, and Alicia R Martin. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions Article Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, 162(5): 1051–1065, 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2015.07.048. URL http://dx.doi.org/10.1016/j.cell.2015.07.048.

[103] Yongtao Guan and Matthew Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Annals of Applied Statistics*, 5(3):1780–1815, 2011. ISSN 19326157. doi: 10.1214/11-AOAS455.

[104] Joseph K Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *bioRxiv*, 2014. doi: dx.doi.org/10.1101/000752.

[105] G Kichaev, Wy Yang, S Lindstrom, F Hormozdiari, E Eskin, Al Price, P Kraft, and B Pasaniuc. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS genetics*, 10(10), 2014. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004722.

[106] Wenan Chen, Beth R Larrabee, Inna G Ovsyannikova, Richard B Kennedy, Iana H Haralambieva, Gregory A Poland, and Daniel J Schaid. Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics Society of America*, 200(July):719–736, 2015. doi: 10.1534/genetics.115.176107.

[107] Kyle Kai-How Farh, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J Housley, Samantha Beik, Noam Shoresh, Holly Whitton, Russell J H Ryan, Alexander A Shishkin, Meital Hatan, Marlene J Carrasco-Alfonso, Dita Mayer, C John Luckey, Nikolaos A Patsopoulos, Philip L De Jager, Vijay K Kuchroo, Charles B Epstein, Mark J Daly, David A Hafler, and Bradley E Bernstein. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343, feb 2015. ISSN 0028-0836. URL http://dx.doi.org/10.1038/nature1383510.1038/nature13835http://www.nature.com/nature/journal/v518/n7539/abs/nature13835.html{#}supplementary-information.

[108] BIMBAM. Bimbam. http://www.haplotype.org/bimbam.html.

[109] Christian Benner, Chris C A Spencer, Aki S Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. Genetics and population analysis FINEMAP : efficient variable selection using summary data from genome-wide association stud-

ies. *Bioinformatics (Oxford, England)*, 32(January 2016):1493–1501, 2016. doi: 10.1093/bioinformatics/btw018.

[110] ArrayExpress. Arrayexpress – functional genomics data. https://www.ebi.ac.uk/arrayexpress/.

[111] Vivek Naranbhai, Benjamin P. Fairfax, Seiko Makino, Peter Humburg, Daniel Wong, Esther Ng, Adrian V. S. Hill, and Julian C. Knight. Genomic modulators of gene expression in human neutrophils. *Nature Communications*, 6(May):7545, 2015. ISSN 2041-1723. doi: 10.1038/ncomms8545. URL http://www.nature.com/doifinder/10.1038/ncomms8545.

[112] Benjamin P Fairfax, Peter Humburg, Seiko Makino, Vivek Naranbhai, Daniel Wong, Evelyn Lau, Luke Jostins, Katharine Plant, Robert Andrews, Chris McGee, and Julian C Knight. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science (New York, N.Y.)*, 343(6175):1246949, 2014. ISSN 1095-9203. doi: 10.1126/science.1246949. URL http://classic.sciencemag.org/content/343/6175/1246949.full.

[113] European Genome-phenome Archive. European genome-phenome archive. https://www.ebi.ac.uk/ega/home.

[114] The cambridge nihr bioresource. https://www.cambridgebioresource.group.cam.ac.uk/.

[115] Hendrik G Stunnenberg, The International, Human Epigenome, and Martin Hirst. Essay The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell*, 167:1145–1149, 2016. doi: 10.1016/j.cell.2016.11.007.

[116] Pan Du, Warren A Kibbe, and Simon M Lin. lumi : a pipeline for processing Illumina microarray. *Bioinformatics*, 24(13):1547–1548, 2008. doi: 10.1093/bioinformatics/btn224.

[117] Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012. doi: 10.1093/bioinformatics/bts034.

[118] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3):1–8, 2012. doi: 10.1038/nprot.2011.457.

[119] Janine Arloth, Daniel M Bader, Simone Röh, and Andre Altmann. Re-Annotator : Annotation Pipeline for Microarray Probe Sequences. pages 1–13, 2015. doi: 10.1371/journal.pone.0139516.

[120] Sanger Imputation Service. Sanger imputation service. https://imputation.sanger.ac.uk.

[121] Srijan Sen and Margit Burmeister. Hardy – Weinberg analysis of a large set of published association studies reveals genotyping error and a deficit of heterozygotes across multiple loci. *Human Genomics*, 3(1):36–52, 2008.

[122] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 57(1):289–300, 1995.

[123] Sarah L. Spain and Jeffrey C Barrett. Strategies for fine-mapping complex traits. *Human Molecular Genetics*, 24(July):111–119, 2015. doi: 10.1093/hmg/ddv260.

[124] Sierra S Nishizaki and Alan P Boyle. Mining the Unknown : Assigning Function to Noncoding Single Nucleotide Polymorphisms. *Cell*, 33(1):34–45, 2017. ISSN 0168-9525. doi: 10.1016/j.tig.2016.10.008. URL http://dx.doi.org/10.1016/j.tig.2016.10.008.

[125] Hyun Min Kang, Chun Ye, and Eleazar Eskin. Accurate Discovery of Expression Quantitative Trait Loci Under Confounding From Spurious and Genuine Regulatory Hotspots. *Genetics Society of America*, 1925(December):1909–1925, 2008. doi: 10.1534/genetics.108.094201.

[126] Xaquin Castro Dopico, Marina Evangelou, Ricardo C Ferreira, Hui Guo, Marcin L Pekalski, Deborah J Smyth, Nicholas Cooper, Oliver S Burren, Anthony J Fulford, Branwen J Hennig, Andrew M Prentice, Anette-g Ziegler, Ezio Bonifacio, Chris Wallace, and John A Todd. Widespread seasonal gene expression reveals annual differences in human immunity and physiology. *Nature Communications*, 6(May): 1–13, 2015. doi: 10.1038/ncomms8000. URL http://dx.doi.org/10.1038/ncomms8000.

[127] Wenqing Fu, Timothy D O Connor, Goo Jun, Hyun Min Kang, Goncalo Abecasis, Suzanne M Leal, Stacey Gabriel, Mark J Rieder, David Altshuler, Jay Shendure, Deborah A Nickerson, Michael J Bamshad, Nhlbi Exome, Sequencing Project, and Joshua M Akey. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493:4–9, 2013. doi: 10.1038/nature11690.

[128] Samuel P Dickson, Kai Wang, Ian Krantz, Hakon Hakonarson, and David B Goldstein. Rare Variants Create Synthetic Genome-Wide Associations. *PLoS Biology*, 8(1), 2010. doi: 10.1371/journal.pbio.1000294.

[129] Ensembl regulatory build 90. Ensembl regulatory build 90. ftp://ftp.ensembl.org/pub/release-90/regulation/homo_sapiens/RegulatoryFeatureActivity/.

[130] ENCODE. Encode/roadmap transcription factor binding sites. http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/wgEncodeRegTfbsClusteredWithCellsV3.bed.gz.

[131] The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(October):636–641, 2004.

[132] Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature*, 518, 2015. doi: 10.1038/nature14248.

[133] Luigi Grassi. Blood rnaexpress. https://blueprint.haem.cam.ac.uk/bloodatlas/.

[134] GeneCards. Ut2 gene - genecards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=UTS2, 2018.

[135] Robert S Ames, Henry M Sarau, Johathan K Chambers, Robert N Willette, Nambi V Aiyar, Anne M Romanic, Calvert S Louden, James J Foley, Charles F Sauermelch, Robert W Coatney, Zhaohui Ao, Jyoti Disa, Stephen D Holmes, Jeffrey M Stadel, John D Martin I, Wu-schyong Liu I, George I Glover I, Shelagh Wilson, Dean E Mcnulty I, Catherine E Ellis, Nabil A Elshourbagy, Usman Shabon, John J Trill, Douglas W P Hay, Eliot H Ohlstein, Derk J Bergsma, and Stephen A Douglas. Human urotensin-II is a potent vasoconstrictor and agonist for the orphan receptor GPR14. *Nature*, 401 (September), 1999.

[136] Rudolf S N Fehrmann, Ritsert C Jansen, Jan H Veldink, Harm-jan Westra, Danny Arends, Jan Bonder, Jingyuan Fu, Patrick Deelen, Harry J M Groen, Asia Smolonska, Rinse K Weersma, Robert M W Hofstra, Wim A Buurman, Sander Rensen, Marcel G M Wolfs, Mathieu Platteel, Alexandra Zhernakova, Clara C Elbers, Eleanora M Festen, Gosia Trynka, Marten H Hofker, Christiaan G J Saris, Roel A Ophoff, Leonard H Van Den Berg, and David A Van Heel. Trans-eQTLs Reveal That Independent Genetic Variants Associated with a Complex Phenotype Converge on Intermediate Genes , with a Major Role for the HLA. *PLoS Genetics*, 7(8), 2011. doi: 10.1371/journal.pgen.1002197.

[137] Monique G P Van Der Wijst, Harm Brugge, Dylan H De Vries, Patrick Deelen, and Morris A Swertz. Single-cell RNA sequencing identifies celltype- specific cis-eQTLs and co-expression QTLs. *Nature Genetics*, 50(April), 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0089-9. URL http://dx.doi.org/10.1038/s41588-018-0089-9.

[138] Harm-Jan Westra, Danny Arends, Tõnu Esko, Marjolein J. Peters, Claudia Schurmann, Katharina Schramm, Johannes Kettunen, Hanieh Yaghootkar, Benjamin P. Fairfax, Anand Kumar Andiappan, Yang Li, Jingyuan Fu, Juha Karjalainen, Mathieu Platteel, Marijn Visschedijk, Rinse K. Weersma, Silva Kasela, Lili Milani, Liina Tserel, Pärt Peterson, Eva Reinmaa, Albert Hofman, André G. Uitterlinden, Fernando Rivadeneira, Georg Homuth, Astrid Petersmann, Roberto Lorbeer, Holger Prokisch, Thomas Meitinger, Christian Herder, Michael Roden, Harald Grallert, Samuli Ripatti, Markus Perola, Andrew R. Wood, David Melzer, Luigi Ferrucci, Andrew B. Singleton, Dena G. Hernandez, Julian C. Knight, Rossella Melchiotti, Bernett Lee, Michael Poidinger, Francesca Zolezzi, Anis Larbi, De Yun Wang, Leonard H. van den Berg, Jan H. Veldink, Olaf Rotzschke, Seiko Makino, Veikko Salomaa, Konstantin Strauch, Uwe Völker, Joyce B. J. van Meurs, Andres Metspalu, Cisca Wijmenga, Ritsert C. Jansen, and Lude Franke. Cell Specific eQTL Analysis without Sorting Cells. *PLOS Genetics*, 11(5):e1005223, 2015. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005223. URL http://dx.plos.org/10.1371/journal.pgen.1005223.

[139] Luke R Lloyd-Jones, Alexander Holloway, Allan Mcrae, Jian Yang, Kerrin Small, Jing Zhao, Biao Zeng, Andrew Bakshi, Andres Metspalu, Manolis Dermitzakis, Greg Gibson, Tim Spector, Grant Montgomery, Tonu Esko, Peter M Visscher, and Joseph E Powell. The Genetic Architecture of Gene Expression in Peripheral Blood. *The*

American Journal of Human Genetics, 100(2):228–237, 2017. ISSN 0002-9297. doi: 10.1016/j.ajhg.2016.12.008. URL http://dx.doi.org/10.1016/j.ajhg.2016.12.008.

[140] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. Simple Combinations of Lineage-Determining Transcription Factors Prime cis - Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4):576–589, 2010. ISSN 1097-2765. doi: 10.1016/j.molcel.2010.05.004. URL http://dx.doi.org/10.1016/j.molcel.2010.05.004.

[141] Daniel J Gaffney, Jean-Baptiste Veyrieras, Jacob F Degner, Roger Pique-Regi, Athma a Pai, Gregory E Crawford, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Dissecting the regulatory architecture of gene expression QTLs. *Genome Biology*, 13(1):R7, 2012. ISSN 1465-6906. doi: 10.1186/gb-2012-13-1-r7. URL http://genomebiology.com/2012/13/1/R7.

[142] Christopher D. Brown, Lara M. Mangravite, and Barbara E. Engelhardt. Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs. *PLoS Genetics*, 9(8), 2013. ISSN 15537390. doi: 10.1371/journal.pgen.1003649.

[143] Michael J Guertin and John T Lis. Mechanisms by which transcription factors gain access to target sequence elements in chromatin. *Current Opinion in Genetics & Development*, 23(2):116–123, 2013. doi: 10.1016/j.gde.2012.11.008.Mechanisms.

[144] Gosia Trynka, Harm-jan Westra, Kamil Slowikowski, Xinli Hu, Han Xu, Barbara E Stranger, Robert J Klein, Buhm Han, and Soumya Raychaudhuri. Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *The American Journal of Human Genetics*, 97(1):139–152, 2015. ISSN 0002-9297. doi: 10.1016/j.ajhg.2015.05.016. URL http://dx.doi.org/10.1016/j.ajhg.2015.05.016.

[145] 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, (526), 2015. doi: 10.1038/nature15393.

[146] BEAGLE 1000 genomes pairwise LD map. Beagle 1000 genomes pairwise ld map. http:/data.broadinstitute.org/srlab/BEAGLE/1kG-beagle-release2/pairwise_ld, 2016.

[147] GWAScatalog. Gwas catalog. https://www.ebi.ac.uk/gwas/.

[148] Chloe M Rivera and Bing Ren. Mapping Human Epigenomes. *Cell*, 155(1):39–55, 2013. ISSN 0092-8674. doi: 10.1016/j.cell.2013.09.011. URL http://dx.doi.org/10.1016/j.cell.2013.09.011.

[149] Alison Abbott. Europe to map the human epigenome. *Nature*, 477(7366):518–518, sep 2011. doi: 10.1038/477518a. URL https://doi.org/10.1038/477518a.

[150] Anubha Mahajan, Daniel Taliun, Matthias Thurner, Neil R. Robertson, Jason M. Torres, N. William Rayner, Anthony J. Payne, Valgerdur Steinthorsdottir, Robert A. Scott, Niels Grarup, James P. Cook, Ellen M. Schmidt, Matthias Wuttke, Chloé Sarnowski, Reedik Mägi, Jana Nano, Christian Gieger, Stella Trompet, Cécile Lecoeur,

Michael H. Preuss, Bram Peter Prins, Xiuqing Guo, Lawrence F. Bielak, Jennifer E. Below, Donald W. Bowden, John Campbell Chambers, Young Jin Kim, Maggie C. Y. Ng, Lauren E. Petty, Xueling Sim, Weihua Zhang, Amanda J. Bennett, Jette Bork-Jensen, Chad M. Brummett, Mickaël Canouil, Kai-Uwe Ec kardt, Krista Fischer, Sharon L. R. Kardia, Florian Kronenberg, Kristi Läll, Ching-Ti Liu, Adam E. Locke, Jian'an Luan, Ioanna Ntalla, Vibe Nylander, Sebastian Schönherr, Claudia Schurmann, Loïc Yengo, Erwin P. Bottinger, Ivan Brandslund, Cramer Christensen, George Dedoussis, Jose C. Florez, Ian Ford, Oscar H. Franco, Timothy M. Frayling, Vilmantas Giedraitis, Sophie Hackinger, Andrew T. Hattersley, Christian Herder, M. Arfan Ikram, Martin Ingelsson, Marit E. Jørgensen, Torben Jørgensen, Jennifer Kriebel, Johanna Kuusisto, Symen Ligthart, Cecilia M. Lindgren, Allan Linneberg, Valeriya Lyssenko, Vasiliki Mamakou, Thomas Meitinger, Karen L. Mohlke, Andrew D. Morris, Girish Nadkarni, James S. Pankow, Annette Peters, Naveed Sattar, Alena Stančáková, Konstantin Strauch, Kent D. Taylor, Barbara Thorand, Gudmar Thorleifsson, Unnur Thorsteinsdottir, Jaakko Tuomilehto, Daniel R. Witte, Josée Dupuis, Patricia A. Peyser, Eleftheria Zeggini, Ruth J. F. Loos, Philippe Froguel, Erik Ingelsson, Lars Lind, Leif Groop, Markku Laakso, Francis S. Collins, J. Wouter Jukema, Colin N. A. Palmer, Harald Grallert, Andres Metspalu, Abbas Dehghan, Anna Köttgen, Goncalo R. Abecasis, James B. Meigs, Jerome I. Rotter, Jonathan Marchini, Oluf Pedersen, Torben Hansen, Claudia Langenberg, Nicholas J. Wareham, Kari Stefansson, Anna L. Gloyn, Andrew P. Morris, Michael Boehnke, and Mark I. McCarthy. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics*, oct 2018. doi: 10.1038/s41588-018-0241-6. URL https://doi.org/10.1038/s41588-018-0241-6.

[151] John D. Schuetz, Michele C. Connelly, Daxi Sun, Sheela G. Paibir, Patricia M. Flynn, R.V. Srinivas, Alok Kumar, and Arnold Fridland. MRP4 : A previously unidentified factor in resistance to nucleoside-based antiviral drugs. *Nature medicine*, 5(9):4–7, 1999. URL https://www.ncbi.nlm.nih.gov/pubmed/10470083.

[152] Gabriele Jedlitschky, Konstanze Tirschmann, Lena E Lubenow, Hendrik K Nieuwenhuis, Jan W N Akkerman, Andreas Greinacher, and Heyo K Kroemer. The nucleotide transporter MRP4 (ABCC4) is highly expressed in human platelets and present in dense granules, indicating a role in mediator storage. *Blood*, 104(12):3603–3611, 2004. doi: 10.1182/blood-2003-12-4330.Supported.

[153] Frans G M Russel, Jan B Koenderink, and Rosalinde Masereeuw. Multidrug resistance protein 4 (MRP4 / ABCC4): a versatile efflux transporter for drugs and signalling molecules. *Cell*, 4(March), 2008. doi: 10.1016/j.tips.2008.01.006.

[154] Satish B Cheepala, Aaron Pitre, Yu Fukuda, Kazumasa Takenaka, Yuanyuan Zhang, Yao Wang, Sharon Frase, Tamara Pestina, T Kent Gartner, Carl Jackson, and John D Schuetz. The ABCC4 membrane transporter modulates platelet aggregation. *Blood*, 126(20):2307–2320, 2015. doi: 10.1182/blood-2014-08-595942.

[155] Benoit Decouture, Elise Dreano, Tiphaine Belleville-Rolland, Orjeta Kuci, Blandine Dizier, Amine Bazaa, Berard Coqueran, Anne-Marie Lompre, Cecile V. Denis, Jean-Sebastien Hulot, Christilla Bachelot-Loza, and Pascale Gaussem. Impaired platelet activation and cAMP homeostasis in MRP4-deficient mice. *Blood*, 126(15):1823–1831, 2015. doi: 10.1182/blood-2015-02-631044.

[156] Sylvia T Nuernberg, Augusto Rendon, Peter A Smethurst, Dirk S Paul, Katrin Voss, Jonathan N Thon, Heather Lloyd-jones, Jennifer G Sambrook, Marloes R Tijssen, Haemgen Consortium, Joseph E Italiano Jr, Panos Deloukas, Berthold Gottgens, Nicole Soranzo, and Willem H Ouwehand. PLATELETS AND THROMBOPOIESIS A GWAS sequence variant for platelet volume marks an alternative DNM3 promoter in megakaryocytes near a MEIS1 binding site. *Blood*, m(24):4859–4869, 2012. doi: 10.1182/blood-2012-01-401893.

[157] Bryce Van De Geijn, Graham Mcvicker, Yoav Gilad, and Jonathan K Pritchard. WASP : allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods*, 12(11), 2015. doi: 10.1038/nmeth.3582.

[158] Bob Argiropoulos, Eric Yung, and R Keith Humphries. Unraveling the crucial roles of Meis1 in leukemogenesis and normal hematopoiesis. *GENES & DEVELOPMENT*, 21 (604):2845–2849, 2007. doi: 10.1101/gad.1619407.3.

[159] Ana Cvejic, Jovana Serbanovic-canic, Derek L Stemple, and Willem H Ouwehand. The role of meis1 in primitive and definitive hematopoiesis during zebrafish development. *haematologica*, 96(2):190–198, 2011. doi: 10.3324/haematol.2010.027698.

[160] Biola M Javierre, Oliver S Burren, Steven P Wilder, Roman Kreuzhuber, Chris Wallace, Mikhail Spivakov, Peter Fraser, Biola M Javierre, Oliver S Burren, Steven P Wilder, Steven M Hill, Sven Sewitz, Kate Downes, Luigi Grassi, Myrto Kostadima, Paula Freire-pritchett, Fan Wang, The Blueprint Consortium, Hendrik G Stunnenberg, John A Todd, Daniel R Zerbino, and Oliver Stegle. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Resource Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, pages 1369–1384, 2016. doi: 10.1016/j.cell.2016.09.037.

[161] Skipper Seabold and Josef Perktold. Statsmodels : Econometric and Statistical Modeling with Python. In *9th PYTHON IN SCIENCE CONF.*, pages 57–61, 2010.

[162] Danish Saleheen, Pradeep Natarajan, Irina M. Armean, Wei Zhao, Asif Rasheed, Sumeet Khetarpal, Hong-Hee Won, Konrad J. Karczewski, Anne H. O'Donnell-Luria, Kaitlin E. Samocha, Benjamin Weisburd, Namrata Gupta, Mozzam Zaidi, Maria Samuel, Atif Imran, Shahid Abbas, Faisal Majeed, Madiha Ishaq, Saba Akhtar, Kevin Trindade, Megan Mucksavage, Nadeem Qamar, Khan Shah Zaman, Zia Yaqoob, Tahir Saghir, Syed Nadeem Hasan Rizvi, Anis Memon, Nadeem Hayyat Mallick, Mohammad Ishaq, Syed Zahed Rasheed, Fazal-ur-Rehman Memon, Khalid Mahmood, Naveeduddin Ahmed, Ron Do, Ronald M. Krauss, Daniel G. MacArthur, Stacey Gabriel, Eric S. Lander, Mark J. Daly, Philippe Frossard, John Danesh, Daniel J. Rader, and Sekar Kathiresan. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature*, 544(7649):235–239, 2017. doi: 10.1038/nature22034.Human.

[163] Monkol Lek, Konrad J Karczewski, V Eric, Andrew J Hill, Beryl B Cummings, Taru Tukiainen, Anne H O Donnell-luria, James S Ware, Grace Tiao, Maria T Tusie-luna, Ben Weisburd, and Hong-hee Won. Analysis of protein-coding genetic variation in 60,706 humans. *Nature Publishing Group*, 536(7616):285–291, 2016. ISSN 0028-0836. doi: 10.1038/nature19057. URL http://dx.doi.org/10.1038/nature19057.

[164] Rolf Lutz and Hermann Bujard. Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Research*, 25(6):1203–1210, 1997.

[165] Žiga Avsec, Roman Kreuzhuber, Johnny Israeli, Nancy Xu, Jun Cheng, Avanti Shrikumar, Abhimanyu Banerjee, Daniel S Kim, Lara Urban, Anshul Kundaje, Oliver Stegle, and Julien Gagneur. Kipoi: accelerating the community exchange and reuse of predictive models for genomics. *bioRxiv*, pages 1–31, 2018. doi: 10.1101/375345. URL http://dx.doi.org/10.1101/375345.

[166] Javed Khan, Jun S. Wei, Markus Ringer, Lao H. Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R. Antonescu, Carsten Peterson, and Paul S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 486:673–679, 2001.

[167] Tatsuya Ando, Miyuki Suguro, Takeshi Kobayashi, Masao Seto, and Hiroyuki Honda. Multiple fuzzy neural network system for outcome prediction and classification of 220 lymphoma patients on the basis of molecular profiling. *Cancer Science*, 94(10): 906–913, 2003. ISSN 13479032. doi: 10.1111/j.1349-7006.2003.tb01374.x.

[168] Michael C O Neill and Li Song. Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. *BMC Bioinformatics*, 12:1–12, 2003.

[169] Nikhil R Pal, Kripamoy Aguan, Animesh Sharma, and Shun-ichi Amari. Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering. *BMC Bioinformatics*, 18:1–18, 2007. doi: 10.1186/1471-2105-8-5.

[170] Lawrence P Petalidis, Anastasis Oulas, Magnus Backlund, Matthew T Wayland, Lu Liu, Karen Plant, Lisa Happerfield, Tom C Freeman, Panayiota Poirazi, and V Peter Collins. Improved grading and survival prediction of human astrocytic brain tumors by artificial neural network analysis of gene expression microarray data. *American Association for Cancer Research*, 7(May):1013–1025, 2008. doi: 10.1158/1535-7163.MCT-07-0177.

[171] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, 33(8):831–838, July 2015.

[172] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, 12(10):931–934, 2015.

[173] Daniel Quang and Xiaohui Xie. FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data, 2017.

[174] Alexander B Rosenberg, Rupali P Patwardhan, Jay Shendure, and Georg Seelig. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell*, 163(3):698–711, October 2015.

[175] J M Paggi and G Bejerano. A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. *bioRxiv*, 2017.

[176] Johnny Israeli et al. Dragonn. https://kundajelab.github.io/dragonn/, 2018.

[177] Žiga Avsec, Mohammadamin Barekatain, and Jun Cheng. Modeling positional effects of regulatory sequences with spline transformations increases prediction accuracy of deep neural networks. *bioRxiv*, 2017. doi: dx.doi.org/10.1101/165183.

[178] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale machine learning on heterogeneous distributed systems. March 2016.

[179] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[180] Frank Seide and Amit Agarwal. CNTK : Microsoft 's Open-Source Deep-Learning Toolkit. In *KDD '16 - 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 2945397, 2016. ISBN 9781450342322. doi: dx.doi.org/10.1145/2939672.2945397.

[181] François et al. Chollet. Keras. https://github.com/keras-team/keras, 2015.

[182] Travis E. Oliphant. *Guide to NumPy*. CreateSpace Independent Publishing Platform, USA, 2nd edition, 2015. ISBN 151730007X, 9781517300074.

[183] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[184] Gene Yeo and Christopher B Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, 11(2-3):377–394, 2004.

[185] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2013.

[186] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for Simplicity: The All Convolutional Net. *ICLR*, pages 1–14, 2015.

[187] Sebastian Bach, Alexander Binder, Grégoire Montavon, and Frederick Klauschen. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE*, pages 1–46, 2015. doi: 10.1371/journal.pone.0130140.

[188] Mukund Sundararajan, Ankur Taly, Qiqi Yan, and Mountain View. Gradients of Counterfactuals. *ICLR*, pages 1–19, 2017.

[189] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences. *arXiv preprint*, 2017.

[190] R Edgar. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1):207–210, 2002.

[191] Chen Chen, Shihua Zhang, and Xiang-sun Zhang. Discovery of cell-type specific regulatory elements in the human genome using differential chromatin modification analysis. *Nucleic Acids Research*, 41(20):9230–9242, 2013. doi: 10.1093/nar/gkt712.

[192] MEGASTROKE Consortium. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nature Genetics*, 50(April), 2018. doi: 10.1038/s41588-018-0058-3.