Chapter 1

Introduction

1 Introduction

1.1 Human genetic variation

Genetic variation describes differences in DNA sequences across individuals that are inherited from maternal and paternal chromosomes. Variation also arises through factors such as errors in DNA replication, incomplete DNA repair, or through the controlled development of the highly variable immune receptor genes (MHC, T cell receptor) (Barnes and Lindahl, 2004, Shiina et al., 2009).

In studying population-level variation, we identify associations between the frequency of genetic variants and physiological differences. On a cellular level, we study how every cell in the human body contains the same DNA molecule yet different tissues carry out highly specialised functions. On a molecular level, sequence variation can affect gene expression and epigenetic functionality. Human genetics now encompasses the study of multiple layers of biological processes, which can represent intermediate steps through which variants ultimately affect organismal phenotypes.

The most common type of genetic variation, and the focus of this thesis, is known as a single nucleotide polymorphism (SNP) where the type of nucleotide at one position varies across individuals. In humans, although there are four possible nucleotide combinations (A, T, G, C), in general only two of the possible four nucleotides are ever seen in a population, and one individual carries two copies (alleles) on each diploid chromosome (Casci, 2010, McDaniell et al., 2010). Variants are classified by the occurrence of the least frequent (minor) allele within a population. Common variants occur with minor allele frequency (MAF) ≥ 5% and rare variants are often defined as occurring with a MAF of less than 1%. A second class of variation is structural variation including insertions-deletions (indels), block substitutions, inversions and copy number variants (Frazer et al., 2009).

SNPs are not inherited independently but are correlated, resulting in the systematic association and correlation of alleles at nearby loci (Slatkin, 2008). This structure is known as linkage disequilibrium (LD) and is variable across populations of different ancestries. The International HapMap Project defined LD regions in 269 individuals of four different populations including Yoruba in Ibadan, Nigeria (YRI), Utah with northern and western European ancestry (CEU), Han Chinese in Beijing (CHB) and Japanese in Tokyo (JPT) (International HapMap Consortium, 2005). Alleles of SNPs within the same LD block are inherited more frequently together in the same haplotype. A set of highly correlated loci (high LD) is known as a haplotype block, the boundaries of which are associated with

recombination hot spots. Within haplotype blocks, recombination is infrequent. In humans, haplotypes range in size from a few kb to over 100 kb (Wall and Pritchard, 2003, Daly et al., 2001). Despite the observation of a few large blocks, most European population haplotypes are smaller, between 5-20 kb (Wall and Pritchard, 2003). This discovery had wider implications for genetic association studies described in detail in Section 1.2.

1.2 Identification of trait-associated genetic variants using genome-wide association studies

Identification of LD patterns, the establishment of public databases containing millions of curated SNPs and emerging microarray technologies together transformed genetic studies (International HapMap Consortium, 2005, Sachidanandam et al., 2001). At the beginning of the GWAS era, genotyping arrays could be designed based on known LD structure to contain probes assaying approximately 500,000 “tag” SNPs, which captured the majority of common European variation without directly genotyping every variant (Barrett and Cardon, 2006). Later came the development of imputation methods, where high-quality haplotypes from reference populations were and still are used to estimate variant alleles that have not been directly genotyped (Huang et al., 2015). Using reference haplotypes such as those available from the UK10K, 1000 Genomes projects or both combined now enables association tests of tens of millions of variants (UK10K. Consortium et al., 2015, 1000 Genomes Project Consortium et al., 2015, Huang et al., 2015). With the falling costs of whole-genome sequencing, we are also moving to using next-generation sequencing technologies to sequence all sites, which vastly improves the accuracy of rare or private variant detection (Bomba et al., 2017).

Collectively these approaches are called genome-wide association studies (GWAS). For the analysis of diseases, GWAS identify discordant variant allele frequencies between cases and controls, where the association of a higher allele frequency with a disease suggests this is a risk factor. GWAS can also be applied to quantitative traits commonly using linear regression to test for association of the variant with increasing or decreasing trait values. In most studies, variants with additive effects are evaluated, where there is a linear and uniform increase in the trait value/disease risk with each copy of the effect allele (Bush and Moore, 2012).

For each variant, an independent statistical test is applied meaning that for a genome-wide approach, multiple tests are implemented. This greatly increases the probability of detecting false positive associations. When using a p value threshold of 0.05, there is a 5% probability of rejecting the null hypothesis by chance, which equates to a high number of observations if

performing millions of tests. Therefore, it is advisable to use a more stringent p value threshold. Based on the International Hapmap Consortium estimation of the number of common (MAF ≥ 5%) independent variants across the genome in a European population, a significance p-value threshold of 5×10^{-08} was suggested to control for multiple testing in GWAS (International HapMap Consortium). Alternatively, for a specific cohort, the Bonferroni correction can be used, where the threshold of 0.05 is divided by the number of independent tests. Alternative methods are discussed in Chapter 2 and implemented in Chapter 4.

GWAS have transformed the study of complex traits and diseases by enabling the unbiased screening for significant genetic variants on a genome-wide scale. Hundreds of risk/trait-associated loci have now been identified. As of the 10th October 2017, the NHGRI-EB GWAS catalog contains 52,491 unique variant-trait associations (MacArthur et al., 2017). This high number reflects the genetic architecture of complex traits in that they are multifactorial and explained by many variants influencing genes and pathways that are biologically relevant to the trait (polygenic) (Visscher et al., 2017). However, the overall phenotypic-variation explained by the identified loci is low, suggesting we have not been able to identify all genetic factors that constitute pre-calculated heritability estimates (Visscher et al., 2017). This is referred to as the "missing heritability" problem, which is an important challenge in the field but not the focus of this thesis (Manolio et al., 2009).

Recently, an "omnigenic" model has been suggested in order to interpret the observation that trait heritability is spread across the whole genome, rather than clustered in key genes (Mumbach et al., 2017). This model posits that variants in highly relevant "core genes" directly affect the trait, but all genes (and variants within them) are highly interconnected through extensive networks, although a full knowledge of such connections is currently lacking (Mumbach et al., 2017). These multiple small effects cumulatively effect disease risk. The authors, however, acknowledge that GWAS provide important biological insights, such as identifying core genes and implicate pathways in which lead variants are enriched (Mumbach et al., 2017). Arguably, investigating cellular contexts of identified genes is still of value, particularly as the authors posit that these complex networks are also cell-type specific (Mumbach et al., 2017).

1.3 Challenges in gaining functional insight from GWAS

Despite the successes of GWAS in identifying many trait-associated variants, there remain some key challenges. This main focus of this thesis is in the functional interpretation of the frequency and effect size spectrum of loci that is currently detectable by GWAS. This includes mainly common variants with modest effect sizes or in some cases low-frequency variants with intermediate effects (McCarthy et al., 2008). Mechanistic interpretation represents a major bottleneck in the GWAS to function process. Biological hypotheses are more straightforward when genetic variants are located within coding regions, particularly if the gene function is known and relates to a relevant phenotype and the variation results in a change in amino acid sequence (non-synonymous) (Vasquez et al., 2016).

However, with the advent of GWAS, somewhat surprisingly, it became apparent that more than 90% of trait-associated SNPs were located in non-coding regions of the genome rather than within genic exons (Maurano et al., 2012, Vasquez et al., 2016). This complicates biological interpretation and linking of downstream consequences to the effect on the overall phenotypic trait.

In addition, whilst LD enabled early successes of GWAS by allowing the assessment of tag SNPs, it complicates a definitive identification of the causal SNP(s). Causal SNPs are those that underlie the true trait association and of all variants in the locus demonstrate the best model fit to the phenotype (Battle and Montgomery, 2014). Distinguishing the true causal variants from highly correlated proxy SNPs (those with an $r^2 > 0.8$) is extremely complex as these will likely fit the phenotype equally as well as the true causal variant (Battle and Montgomery, 2014). Larger sample sizes, high-density genotyping, imputation with a high-quality reference panel or whole-genome sequencing all increase the number of variants identified and therefore the likelihood of identifying the causal variant (Battle and Montgomery, 2014). However, even the various statistical approaches for fine-mapping causal variants are limited in cases of high correlation between variants (Chun et al., 2017). Ultimately, functional experiments are required to fully resolve such loci.

There are multiple approaches that attempt to address each of these challenges. This thesis will focus on those that aim to assign function to genetic loci, which can also aid identification of causal variants in some cases. I discuss the type of data and approaches in detail below.

1.4 Assigning function to genetic loci

1.4.1 Understanding the non-coding regulatory genome

Describing the biology of non-coding SNPs requires an understanding of the function of the regulatory genome. While we are unable to predict this function from DNA sequence, through the efforts of large-scale consortia such as ENCODE, ROADMAP and BLUEPRINT, we know now that much of the non-coding genome performs a regulatory function (Encode Project Consortium, 2012, Roadmap Epigenomics Consortium et al., 2015, Adams et al., 2012). There are multiple different layers of (epi)genomic function. The data made available through such consortia can be used to investigate the context of non-coding genetic variation. Below, I summarise our current knowledge of key concepts of epigenomics function and gene regulation.

1.4.1.1 Transcription initiation at promoters

Transcription is a highly regulated process where RNA polymerase (RNAP) enzymes generate an RNA molecule that is complementary to the sequence of DNA. Transcription is initiated at core promoters, which are DNA segments of between 50 and 100 bp (Roy and Singer, 2015). Here, the core transcription machinery including RNAP and general transcription factors (GTFs) assembles. There are various RNAP enzymes, RNA polymerase II (Pol II) transcribes protein-coding genes as well as the non-coding RNAs, small-nucleolar (sn)RNA and micro(mi)RNA (Guiro and Murphy, 2017). Studies utilising cell-free systems identified six GTFs, TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH (Roeder, 1996, Roy and Singer, 2015). GTFs recognise specific elements of the core promoter through sequence-specific DNA binding. Classification of mammalian promoters based on canonical elements is complex as many do not contain such sequences, which include the TATA box, Initiator (Inr) element, the TFIIB recognition element (BRE) and downstream promoter element (DPE) (Roy and Singer, 2015). For example, only 5-7% of eukaryotic promoters contain a TATA box, and as such there are many cases of non-canonical core promoters (Roy and Singer, 2015). These can contain unmethylated CpG islands or ATG deserts (low occurrence of ATG trinucleotides). Particular chromatin modifications can also mark mammalian promoters, which I discuss in detail below.

Initiation is an important regulated step in transcription. Recently, the association of rs34481144 with severe risk of influenza in humans was shown to involve the disruption of promoter activity as a result of the change in one nucleotide from G (protective) to A (risk) (Allen et al., 2017). rs34481144 resides with the 5' UTR of the interferon induced transmembrane protein 3 gene, *IFITM3*. Through a series of elegant experiments, the risk allele was shown to be associated with lower *IFITM3* gene expression, lower promoter

activity and lower promoter binding of the innate immune interferon, IRF3 and disruption of a CpG methylation site in $CD8^+$ T cells, where reduced methylation increased binding of the insulator factor CTCF. Carriers of the risk allele had lower numbers of $CD8^+$ T cells in the airways during influenza infection, suggesting how reduced *IFITM3* expression (due to reduced promoter activity and demethylation) could increase susceptibility to severe infection and providing evidence for a role of *IFITM3* in the cellular response to infection (Allen et al., 2017). Therefore, sequence-specificity is important to the recruitment of factors required for promoter activity and can be affected by SNPs. This example also highlights a potential role for DNA methylation in regulating gene expression.

1.4.1.2 Regulation of transcription by enhancers and other regulatory elements

Transcription is also regulated by the activity of distal regulatory sequences located upstream or downstream of the promoter (Heinz et al., 2015). These cognate regulatory elements are known as enhancers that activate transcription (Roy and Singer, 2015). Enhancers were originally identified using plasmid-based assays as sequences of no more than 100 bp that could drive gene expression (Banerji et al., 1981, Banerji et al., 1983, Krijger and de Laat, 2016). Enhancer-gene interaction can be promiscuous but also selective and may not necessarily be between the nearest gene (Javierre et al., 2016, Mumbach et al., 2017, Krijger and de Laat, 2016). STARR-seq, a massively parallel reporter assay that enables the assessment of all genome-wide candidate enhancers through the ability of these sequences to drive transcription, was used to show that there were two different clusters of enhancer sequences that separately activated housekeeping genes and developmental genes (Zabidi et al., 2015).

Silencers have similar properties to enhancers but instead act to inhibit transcription. Insulators are boundary elements that inhibit the spreading of transcription and chromatin interactions between neighbouring genomic regions (Gaszner and Felsenfeld, 2006, Ali et al., 2016). CTCF is a key factor in mediating insulation (Ali et al., 2016). Therefore, the spatial and temporal control of gene expression by distal regulators represents another layer of regulation and functionality of the non-coding genome (Ong and Corces, 2011).

The enrichment of SNPs in enhancer regions is now well established and commonly used as a method to assign functionality to non-coding SNPs (Farh et al., 2015, Huang et al., 2017b, Musunuru et al., 2010, Chen et al., 2016a). Multiple examples of SNPs modifying enhancer activity are discussed throughout this thesis and my investigation into disease risk loci in Chapter 2 adds further examples to the many already demonstrated.

1.4.1.3 Transcription factors

Transcription factors regulate gene expression through the sequence-directed binding to DNA at either promoters or regulatory elements such as enhancers. Multiple transcription factors bound at enhancers interact with components such as the Mediator complex or the general TF, TFIID to help recruit RNA polymerase II (Kagey et al., 2010). Looping out of intervening DNA enables interaction between enhancers and promoters. Other factors, such as the cohesin complex can act as scaffold proteins to ensure the stability of these interactions (Kagey et al., 2010, Schmidt et al., 2010). A study that assayed the binding of over 100 transcription factors in colorectal cancer (CRC) LoVo cells found that TFs were bound in clusters across the genome; 75% of the TF peaks were localised in 0.8% of the genome, consistent with previous observations that TF act combinatorially (Yan et al., 2013). Almost all clusters were formed around cohesin, demonstrating the importance of the cohesin complex in enabling complex TF binding (Yan et al., 2013).

The initial selection of enhancers during the differentiation of specific cell lineages is controlled by pioneer transcription factors such as the haematopoietic-specific master regulator, PU.1 (Heinz et al., 2015). Pioneer factors can bind to their cognate motifs prior to any transcriptional activity or chromatin modification and at sites of DNase I inaccessibility (Heinz et al., 2010, Pham et al., 2013). Although PU.1 is an important factor for multiple haematopoietic cell types, PU.1 binding was shown to be cell-type specific (Pham et al., 2013, Heinz et al., 2010). Cooperative binding of PU.1 with other collaborative transcription factors together establish the cell-type specific transcriptional signatures that support lineage-specific differentiation (Heinz et al., 2015, Pham et al., 2013, Adams and Workman, 1995). For example, PU.1 is required for the generation of the general myeloid progenitor and the common lymphoid progenitor but different co-factors are associated with PU.1 at cognate binding sites between macrophages and B cells (Heinz et al., 2010). For example, C/EBP and AP-1 motifs were highly enriched within macrophage-specific distal PU.1 sites whilst E2A, EBF, Oct and NF-κB motifs were enriched in B cell specific PU.1 sites. These additional TFs both had roles in macrophage and B cell differentiation respectively (Heinz et al., 2010). In a PU.1 deficient myeloid progenitor cell line, the absence of PU.1 resulted in a reduced genome-wide C/EBPβ binding pattern. No corresponding PU.1 motifs were found in the C/EBPβ binding sites that remained. Restoration of PU.1 expression in this cell line using a fusion protein, increased PU.1 binding and the number of induced C/EBPβ-bound sites, 75% of which were now co-bound by both TFs and enriched for the PU.1 motif (Heinz et al., 2010). The importance of combinatorial TF binding was confirmed by evaluating the effects of naturally occurring motif mutations in PU.1 and C/EBPα between two different mouse strains (Heinz et al., 2013). Loss of binding of one TF as a result of motif disruption led to the corresponding loss of the second TF and vice versa (Heinz et al., 2013). It is suggested that

co-binding of these TFs enables competition with nucleosomes to maintain open chromatin and establish the required cell-type specific binding (Heinz et al., 2010).

Some enhancers require additional co-factors to become fully activated, particularly in response to external or internal signals. Cell type-specific responses to the same stimuli can be achieved through the collaboration between pioneer factors, which first select enhancer sites in the respective cell types and open chromatin (Mullen et al., 2011, Heinz et al., 2015). Following this, a second tier of signal-dependent TFs can bind to these previously established enhancers ensuring that a specific subset of regulatory elements is activated in different cell types (Mullen et al., 2011, Heinz et al., 2010, Ghisletti et al., 2010). Multiple studies have provided evidence for a relatively small number of TFs that interact and bind with pioneer factors to determine cell type specific differentiation and signalling responses by directing the genes to which signalling TFs bind. For example, Mullen *et al.* (2011) used ChIP-seq to show that TGFβ signalling is mediated by Smad2/3, but only 1% of Smad3 binding sites were occupied in more than one cell type between embryonic stem cells, pro-B cells and myotubes (Mullen et al., 2011). Further, they showed that cell-type specific signalling responses were the result of Smad2/3 co-occupying distinct sites with cell-type specific master/pioneer TFs; Oct4 in ES cells, PU.1 in pro-B cells and Myod1 in myotubes (Mullen et al., 2011). Similar cooperative interactions were shown *in vivo* where 61% of NF-κB binding sites in strain-specific mice were already bound by PU.1 and CEBPα before Toll-like receptor 4 (TLR4) stimulation (Heinz et al., 2013).

In summary, cooperative binding of a relatively small and defined group of TFs establishes cell-type specificity of gene expression, lineage differentiation and response to external and internal stimuli. Given the importance of TF in these processes, it is often investigated whether a SNP disrupts TF binding motifs, many examples of which are discussed throughout this thesis.

1.4.1.4 Transcription elongation and RNA processing

Transcriptional regulation is not restricted to initiation. For many mammalian genes, high levels of transcription initiation were observed, but this was not correlated with a high level of gene expression (Guenther et al., 2007). This is due to post-initiation regulation where negative elongation factors can cause Pol II promoter-proximal pausing. Pol II can be released by factors such as the Positive transcription elongation factor (P-TEFb) (Rahl et al., 2010, Zhou et al., 2012). This mechanism is thought to enable fine-tuning in transcription to produce the optimal level of cellular gene transcription as some genes will progress to productive elongation but not all (Zhou et al., 2012). Regulation at this stage also influences

processes that can be coupled to transcription such as 5' mRNA capping, splicing and 3' cleavage and polyadenylation (Zhou et al., 2012).

Splicing is the removal of introns within genes to produce a mature processed RNA. Alternative splicing is widespread, occurring with up to 94% multiexonic human genes (Chen et al., 2014). The process can generate multiple transcripts from a single gene as a result of exon skipping, alternative 3' acceptors, alternative 5' donor sites or intron retention (Figure 1.1) (Chen et al., 2014, Nilsen and Graveley, 2010). Splicing can be tissue and developmental-stage specific and is important in disease, with 15% of disease-causing mutations being located in splice sites (Chen et al., 2014). Mutations in splicing factor genes occur at high frequency in haematological cancers (Chen et al., 2014). Extensive transcript diversity as a result of alternative splicing was recently shown in haematopoietic progenitor and precursor cell populations, where 7,881 novel splice junctions were discovered as well as 2,301 alternative splicing events (Chen et al., 2014). In many cases transcript changes were not associated with detectable changes in gene expression, showing that increasing cell-type commitment during lineage differentiation involves the use of alternative transcript isoforms. Therefore, a full understanding of development diversity requires an assessment of all transcriptome effects not just those at the gene level (Chen et al., 2014).

Two methods to quantify splicing events are summarised in Figure 1.1. Both of these methods were used in the BLUEPRINT consortium and as such as used in the analysis of variant function throughout this thesis. Accurate splicing quantification requires RNA-seq data. This is a technique that uses next-generation sequencing to quantify genome-wide gene expression profiles, where high gene expression is represented by an increased number of reads mapping to the corresponding gene location in the reference genome (Marioni et al., 2008). Reads across splicing junctions can also be counted, as is employed in the splicing annotation method, referred to as percent splice in (Figure 1.1) (Chen et al., 2016a). Alternatively, the relative expression levels of all known and annotated transcripts, as defined by GENCODE for example, can be estimated using RNA-seq reads across the gene body (Figure 1.1) (Chen et al., 2016a).

Splicing and donor-acceptor sites are highly sequence specific and therefore could be disrupted by genetic variants (Figure 1.1). In addition, branch points, exonic and intronic splicing enhancers/silencers and mRNA secondary structures can also be influenced by SNPs and result in splicing changes (Hiller et al., 2006). For example, the multiple sclerosis risk SNP, rs17612638 (G) abrogates an exonic splicing silencer, which normally functions to repress the use of a 5' splice site of exon 4 of the *PTPRC* gene (Lynch and Weiss, 2001). This gene encodes a receptor of the protein tyrosine phosphatase family, also known as

CD45, which expressed all nucleated haematopoietic cells (Lynch and Weiss, 2001, Nakano et al., 1990). The immune-related function of this gene suggests that disruption of the tightly regulated exon 4 and resultant alternative transcripts may underlie the observed MS risk (Lynch and Weiss, 2001).

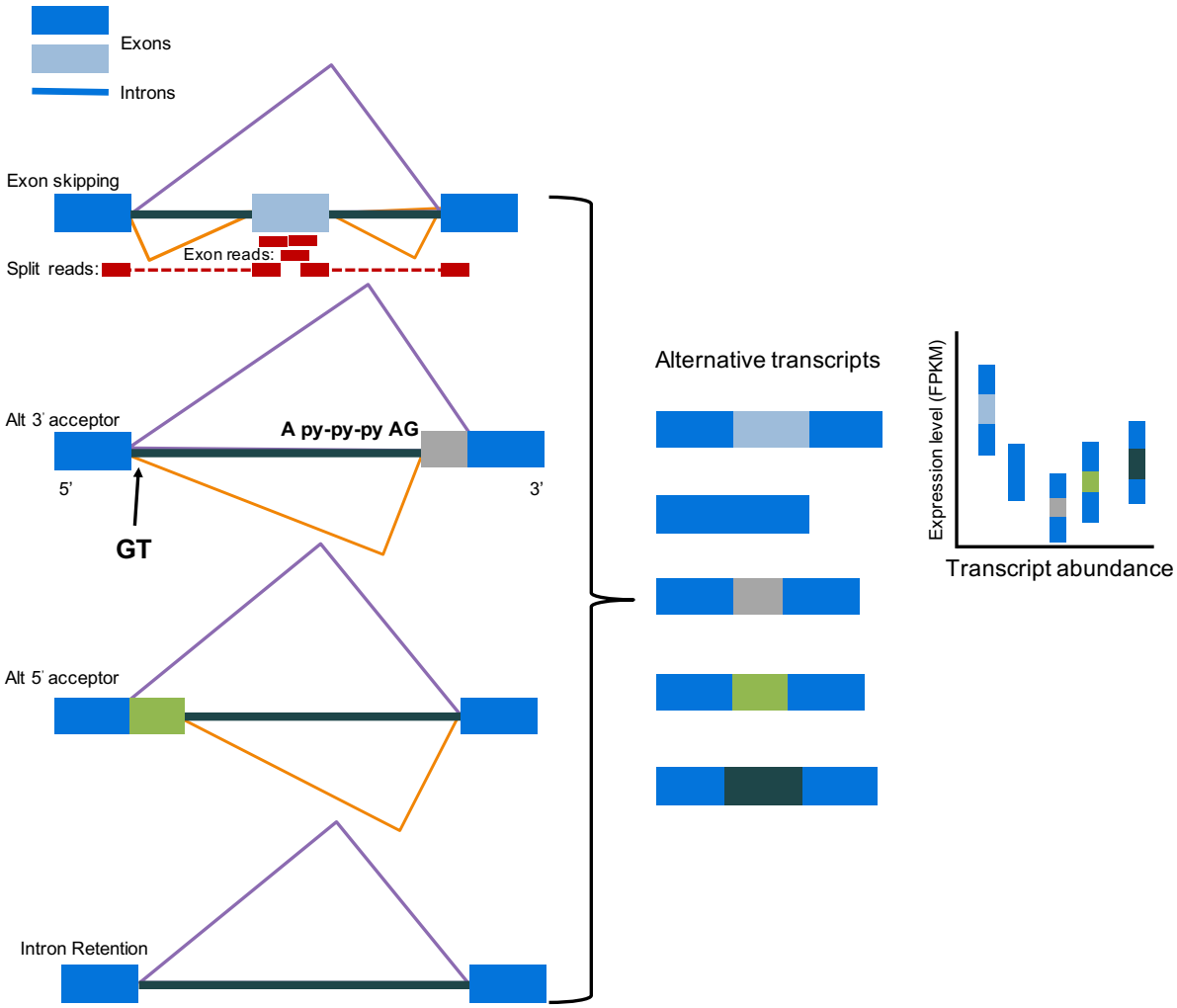

Figure 1.1: Alternative splicing mechanisms produces multiple distinct transcripts
Schematic summarises the different molecular changes involved in alternative splicing and the different possible RNA transcripts. The cognate vertebrate splicing donor site contained in the 5' intron sequence (GT) is also shown along with the 3' splicing acceptor site (AG). The polypyrimidine tract (py-py-py) is a region high in C and T/U pyrimidines. Upstream of this tract is the branch point, which includes an A nucleotide and is important in the splicing molecular mechanism. RNA-seq can be used to quantify the reads (shown in red) across the splicing junctions. The examples above show split reads across two introns and reads within an exon, which both support exon inclusion. Splicing can also be assessed by quantifying the expression of the known alternative transcripts (right) by counting reads expressed in fragments per kilobase of transcript per million fragments sequenced (FPKM). Adapted from (Chen et al., 2014, Nilsen and Graveley, 2010). The percent splice-in method portrayed above is similar to that described by Geuvadis consortium and the information in the figure above was adapted from the (Geuvadis, 2010) webpage listed in the references.

1.4.1.5 Chromatin structure

The regulatory processes described above do not navigate a simple linear DNA sequence, but a complex three-dimensional structure known as chromatin. For DNA to fit into an approximate 10 μm-diameter nucleus it is highly condensed in a nucleoprotein complex (Nieto Moreno et al., 2015). 147 bp of DNA is wrapped 1.7 turns around the histone protein octamer, which is known as a nucleosome (Figure 1.2) (Luger et al., 1997). Octamers comprise two H3-H4 and two H2A-H2B dimers and histone H1 (Figure 1.2) (Luger et al., 1997). Nucleosomes are repeating units (Figure 1.3) and this structure allows further supercoiling and condensation into functional structural domains (Lavelle, 2014). Chromatin remodellers disassemble local compacted nucleosomes to allow access for Pol II and other cofactors, which is essential for active gene expression. This is a state generally referred to as “open chromatin”, whereas “closed chromatin” generally refers to genes and regulatory elements that are inaccessible due to the compact structure (Figure 1.3) (Bannister and Kouzarides, 2011). Further compaction beyond this leads to the formation of constitutively closed heterochromatin containing repressed genes.

Chemical modification of the core histone proteins or protruding amino-terminal tails is also an important regulatory mechanism and confers function to chromatin. Histone modifications are chemical groups that are added to specific residues in the histone protein sequence by chromatin modifying proteins (Figure 1.2). Possible modifications include histone phosphorylation, acetylation, methylation and ubiquitylation. The charges associated with certain modifications, such as the negatively charged phosphorylation, can affect the interactions between histones and, it has been suggested, with the negatively charged DNA phosphate backbone changing the local compaction of DNA (Bannister and Kouzarides, 2011). In addition, these groups can act as molecular “flags” for the binding of histone chaperones, other functional cofactors or additional chromatin remodellers. These proteins contain domains which can recognise modifications, for example, CHD1 binds to H3K4me3 through the chromodomain and the heterochromatin protein, HP1, binds to methylated lysine 9 on histone H3 (Flanagan et al., 2005, Bannister et al., 2001). Proteins containing bromodomains bind to acetyl-lysine modifications and subsequently initiate transcription, therefore targeting these domains offers an attractive potential for specific therapeutics in inflammation, viral infection and in regulating oncogene expression (Filippakopoulos and Knapp, 2014).

Histone modifications are dynamic, can be altered in response to intracellular and extracellular stimuli, and regulate multiple processes beyond chromatin structure and transcription including DNA repair, replication and recombination (Bannister and Kouzarides, 2011). Chromatin structure within genic regions can also influence alternative splicing (as

discussed above) through kinetic coupling with transcription whereby nucleosomes act as obstacles, promoting Pol II pausing and influencing exon inclusion/exclusion (Kadener et al., 2001, Schor et al., 2009, Bintu et al., 2012).

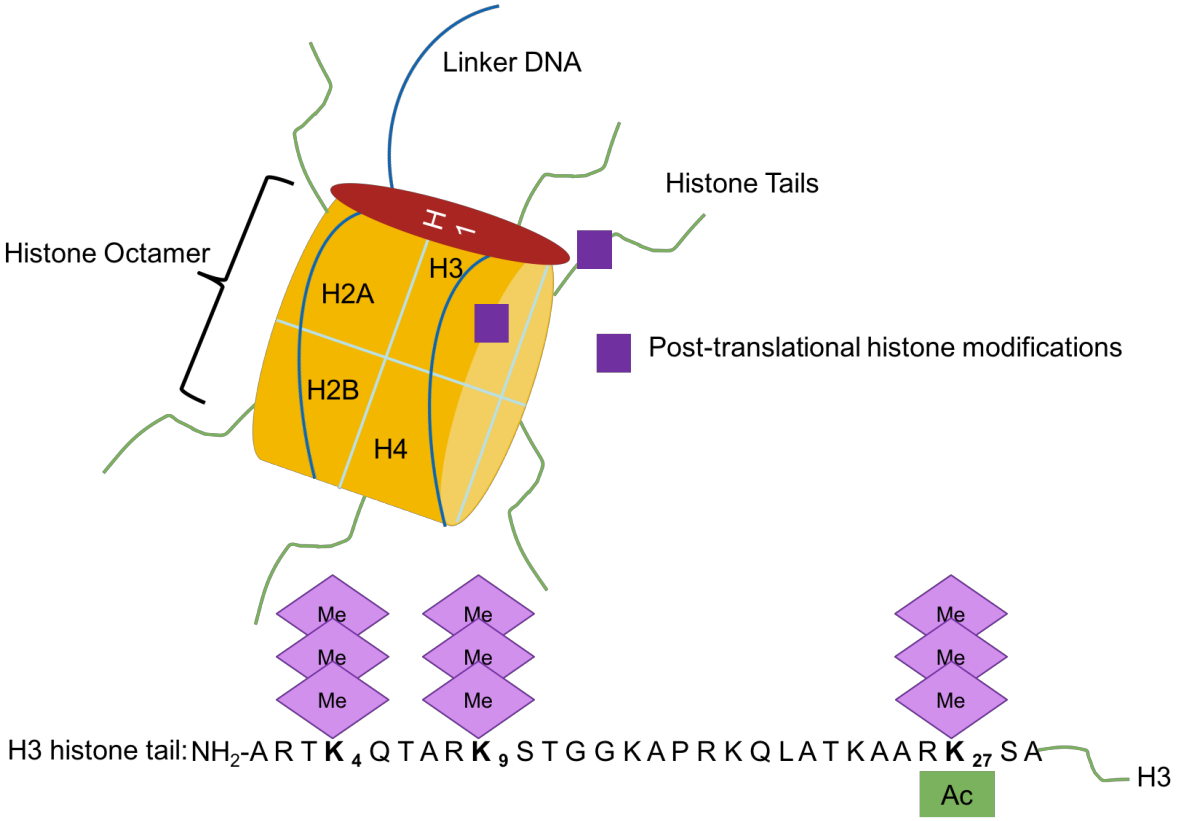

Figure 1.2: Histone structure and modifications
Nucleosomes are protein structure units consisting of approximately 147 bp of DNA (dark blue) wrapped around the octameric protein structure containing two copies of each of the core histones H2A, H2B, H3 and H4 (yellow). Histone H1 is a linker histone that stabilises higher order structure of chromatin and protects the DNA from nuclease digestion. Most histone modifications (dark purple) occur on the N-terminal histone tails (green). Modifications considered in this thesis are shown below for the histone tail of the H3 core histone. The notation of, for example, H3K4me3 refers first to the histone H3, then to the lysine residue that is fourth in the sequence counting from the N-terminus and then to the chemical modification itself, here a tri-methylation of the lysine residue. Modifications also occur within the core globular protein structure. Adapted from (Fullgrabe et al., 2011).

Genome-wide profiling of histone-bound regions indicated that specific histone modifications are associated with specialised functional genomic regions including promoters or enhancers. As such, these approaches have transformed the way we now identify functional genomic regions (Barski et al., 2007, Hon et al., 2009). To identify these regions the technique, chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq), uses antibodies specific to a histone modification (or transcription factor) to enrich crosslinked protein-DNA fragments for bound-regions, which are then sequenced (Barski et al., 2007, Schmidt et al., 2009). Bound genomic regions are identified by pile-ups of sequence reads (referred to as “peaks”), which provide a quantitative measurement of genome-wide protein binding (Figure 1.4).

Insights from these genome-wide profiles include the observation that H3K4me3 preferentially associates with promoters and marks regions of active transcription (Hon et al., 2009). Chromatin signatures at promoters were found to be similar across cell types but in contrast, H3K4me1 associated with cell-type specific enhancers (Heintzman et al., 2009 2009). However, many H3K4me1-associated enhancer regions were later found to be inactive when tested in reporter assays, leading to the discovery that active enhancers are marked by a combination of H3K4me1 and H3K27ac (Figure 1.2-1.3) (Creyghton et al., 2010). Instead, H3K4me1 alone marks poised enhancers that may not necessarily be active but could reflect molecular ‘memory’ of previous activation (Heinz et al., 2015, Creyghton et al., 2010). For example, many inactive haematopoietic stem cell developmental genes were found to be regulated by distal enhancers enriched with H3K4me1 (Creyghton et al., 2010, Cui et al., 2009). H3K27ac, which is deposited by both p300 and CREB binding protein (CBP) can also mark active promoters, when not in conjunction with H3K4me1 (Creyghton et al., 2010).

Clearly the context of chromatin functional state has important consequences for molecular function. For example, using STARR-seq, it was observed that although many sequences possessed the capacity to act as enhancers, many were endogenously repressed (Zabidi et al., 2015, Krijger and de Laat, 2016). The multiple layers of transcriptional regulation and chromatin context are summarised in Figure 1.3. Also shown is the high levels of 5-methyl cytosine (5mC) in closed chromatin, contributing to gene repression (Figure 1.3) (Jones, 2012). Recent advances in genome-wide DNA methylation mapping techniques have highlighted the varied roles of this epigenetic mark depending on the genomic context and interpretation of the functional effect requires appreciation of multiple genomic factors (Jones, 2012).

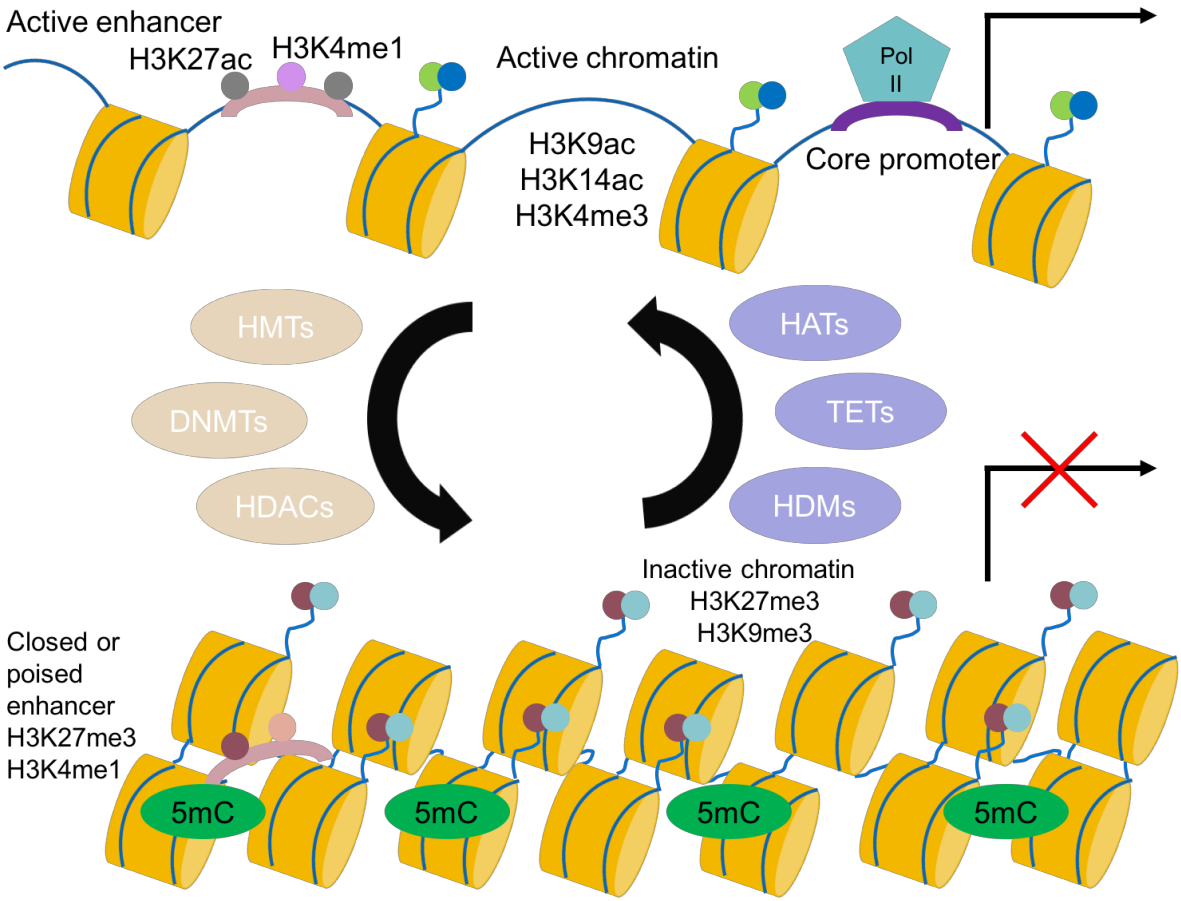

Figure 1.3: Multiple layers of gene regulation
This schematic summarises the many molecular processes that control transcription. The level of DNA compaction controls access of DNA-binding cofactors. In the bottom panel, DNA is highly compacted and hypermethylated at cytosine residues (5mC) preventing access to transcriptional cofactors and repressing gene expression. Histone remodelling proteins (purple) can open chromatin allowing access to other cofactors (top panel). This leads to activation of RNA polymerase II and transcription initiation at the core promoter. Enhancer-bound cofactors can also influence transcription of distal genes through long-range interactions as a result of DNA looping and clustering. DNMT = DNA methyltransferase. HAT = histone acetyltransferase. HDAC = histone deacetylase. HDM = histone demethylase. HMT = histone methyltransferase. TET = ten-eleven translocation. Adapted from (Greco and Condorelli, 2015).

1.4.1.6 Higher-order chromatin structure

With the advent of chromatin conformation capture techniques came the ability to study the three-dimensional spatial genomic structure on a global scale, showing that regulatory loops are widespread and provide another mechanism for transcriptional regulation (Dekker et al., 2002). Chromatin conformation capture (3C) and adaptations of this approach including 4C, 5C, Hi-C, ChIA-PET and promoter-capture HiC (PcHiC), identify long-range interactions by formaldehyde cross-linking of genomic regions located close in physical space (de Wit and de Laat, 2012). Similar to ChIP-seq, these fragments are sequenced and mapped to the reference genome, thereby identifying fragments connecting distally located elements. Chromatin conformation techniques differ by the resolution of interactions detected. For example, genome-wide approaches such as HiC revealed chromatin loops on a larger scale (100kb to 5Mb) referred to as topologically associated domains (TADs) (Lieberman-Aiden et al., 2009, Dixon et al., 2012, Krijger and de Laat, 2016). TADs are more likely to be tissue-invariant but sub-TADs (median size of ~185 kb) and regulatory loops that form within TADs are more tissue-specific and dynamic (Dixon et al., 2012, Phillips-Cremins et al., 2013, Krijger and de Laat, 2016). Stabilisation of TADs requires CTCF and cohesin whereas regulatory loops also require additional tissue-specific TFs (Krijger and de Laat, 2016, Phillips-Cremins et al., 2013, Kagey et al., 2010).

The physical partitioning of the genome into these architectural domains correlates well with genomic function including actively transcribed or repressed genes (Symmons et al., 2014). A definitive causal relationship between promoter-enhancer chromatin looping and gene expression was demonstrated by inducing looping between the beta-globin gene and corresponding super enhancer (locus control region), which resulted in significantly upregulated beta-globin gene expression (Deng et al., 2014).

Connections between distal enhancers and gene targets complicate assignment of genes to regulatory SNPs. HiC data can be used to identify target genes of distal regulatory SNPs. PcHiC is used predominantly in this thesis and achieves higher resolution in comparison to HiC by enriching fragments for genome-wide promoter-mediated interactions using an array with promoter-probes of all cellular genes (Mifsud et al., 2015). This approach was recently used to identify the interacting regions of 31,253 promoters in 17 primary human haematopoietic cells (Javierre et al., 2016). Interactions were found to be highly cell-type specific, recapitulating the haematopoietic tree and interacting regions were enriched in GWAS disease variants (Javierre et al., 2016). Using this data, the 6q23 locus, associated with RA and psoriasis, was found to interact with the promoter of the most proximal gene, *TNFAIP3*, but also with the promoter of *IL20RA,* located 680 kb upstream (McGovern et al., 2016). The risk allele of the likely causal SNP in this locus, rs6927172, correlated with

increased gene expression of *IL20RA*, increased binding of both enhancer-associated histone marks and the TF, NFκB (McGovern et al., 2016). In this case, monoclonal therapy against IL-20 has been shown to be effective for both diseases (McGovern et al., 2016). On a genome-wide scale, an independent but similar capture approach, HiChIP, was used to map disease SNP target genes (Mumbach et al., 2017). Instead of focusing on promoter interactions, HiChIP is a protein-centric technique that was recently used with H3K27ac as a bait to assay interactions in T cell populations (Mumbach et al., 2017). Using H3K27ac interaction maps, 2,597 target genes were identified for 684 autoimmune disease variants (Mumbach et al., 2017, Trynka, 2017). Only 14% of the mapped target genes represented the closest gene to the GWAS variant. This demonstrates the utility of interaction data to identify target genes, which is important in the translation of GWAS to the clinic.
Capture techniques can be used to identify SNP target genes, but long-range interactions could themselves be disrupted by these variants. Disruption of TF binding has long been suggested as the predominant mechanism underlying regulatory variation (Pai et al., 2015). However, only a minority, 10-20%, of GWAS SNPs were found to be located within TF binding motifs (of 823 variants assessed), suggesting other regulatory mechanisms may underlie genetic associations (Farh et al., 2015). Evidence of allele-specific interactions has been observed, using ChIA-PET of CTCF and Pol II in different human cell lines. For example, 50 loci showed allele-specific tandem loops (loops coordinated by two CTCF motifs positioned in a tandem manner) that contained phased SNPs within the gene body (Tang et al., 2015). 44% of these loci displayed allele-specific expression (Tang et al., 2015). The authors also showed that the asthma-associated SNP, rs12936231, disrupted a CTCF motif and CTCF binding further abrogating looping and chromatin topology, which they postulated could represent the primary molecular event underlying the locus (Tang et al., 2015). Similar observations have been made combining H3K27ac HiChIP interaction data from primary human cells with available genome phasing (Mumbach et al., 2017). The authors observed 4.2% of loops exhibited allelic bias (FDR < 0.05) where risk alleles either disrupted or increased enhancer-gene interactions (Mumbach et al., 2017). Thorough examination of the allelic bias of chromatin interactions in a larger population-scale cohort is needed to establish this as a widespread disease-relevant regulatory mechanism.

1.4.1.7 Non-coding RNA regulation

90% of the genome is transcribed into non-coding RNAs including ribosomal, transfer-RNAs, long non-coding RNAs and microRNAs, compared to 2-3% transcribed to protein (Roy and Singer, 2015, Lee, 2012). miRNAs are short (19-24 nucleotides) and function to cleave or repress complementary mRNA post-transcriptionally where binding is mediated by the RNA-induced silencing complex (RISC) (Hrdlickova et al., 2014). The translation of more than half of protein-coding genes is regulated by miRNAs (Hrdlickova et al., 2014). Many lncRNAs,

which consist of a heterogeneous group of RNAs more than 200 nucleotides, are thought to regulate expression of protein-coding genes (Harrow et al., 2012, Hrdlickova et al., 2014). lncRNAs exhibit cell-type specific expression and widespread regulatory functions through interaction with DNA, RNA or protein enabling the control of processes such as gene silencing, RNA maturation and transport, protein production and chromatin remodelling (Derrien et al., 2012, Hrdlickova et al., 2014).

Non-coding RNAs have been implicated in a range of neurodegenerative, cardiovascular and autoimmune diseases as well as cancer (Hrdlickova et al., 2014). Disease SNPs have been shown to confer risk by disrupting the function of non-coding RNAs, for example, by altering RNA expression or by changing binding sites in target genes. rs57095329 is associated with systemic lupus erythematosus (SLE) and located in the promoter of microRNA, miR-146a (Luo et al., 2011, Hrdlickova et al., 2014). Increased SLE risk is associated with lower miR-146a expression levels, observed in peripheral blood leukocytes (Luo et al., 2011). Upregulated type I interferon pathway activity is known to occur in SLE pathogenesis and miR-146a functions as a negative regulator of this activity, explaining how a decreased miRNA expression could increase disease risk (Luo et al., 2011, Tang et al., 2009). Non-coding RNA function and target gene interaction is another important regulatory function to consider in genetic function studies. Figure 1.4 summarises how all of the described epigenomic data can be used to annotate function of trait-associated variants and in part aid the prediction of putative causal SNPs.

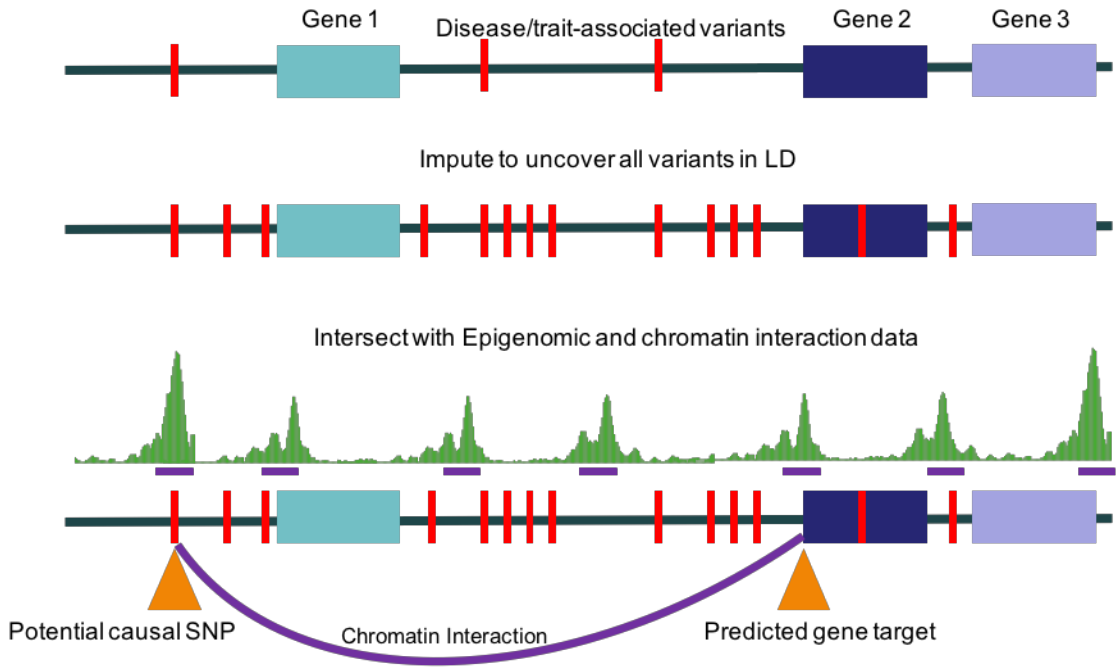

Figure 1.4: Annotating genetic variants with epigenomic function
Schematic summarises initial steps in predicting molecular mechanisms of trait-associated SNPs. Imputation, targeting genotyping or use of whole-genome sequencing data identifies all variants in LD. Disease-associated SNPs are intersected with epigenomic regions such as chromatin modification or transcription factor binding (ChIP-seq binding peaks in green). Combined with high-resolution chromatin interaction data, putative target genes can be identified. Further techniques to identify function such as quantitative trait studies are discussed below. Figure based on (Krijger and de Laat, 2016).

1.4.2 Quantitative trait loci studies with molecular phenotypes

Annotating the genome with epigenomic data (Figure 1.4), while helping to highlight molecular function, is prone to chance overlaps. Alternatively, using epigenomic data as a quantitative trait in association mapping can identify, with statistical confidence, specific variants (and those in high LD) associated with disrupting a molecular function. If a genomic locus is associated with both a disease or complex trait and with a molecular phenotype such as gene expression, this is a strong indicator of possible causal mechanism (Nica and Dermitzakis, 2013).

Variation in gene expression can arise from environmental factors, epigenetic effects, random biological noise and genetic effects. QTL mapping uncovers the genetic basis of variation in quantitative phenotypes in a similar approach to GWAS. Smaller cohorts can reduce power and therefore, rather than genome-wide, the number of variants tested in a QTL study is constrained within a genomic window surrounding each molecular feature. These QTLs are referred to as cis-QTLs, which are SNPs that act locally to the feature being investigated (Nica and Dermitzakis, 2013). The definition of "local" can vary between studies; a window of 1 Mb either side of the start and end of the feature was used in the Chen *et al.* (2016) study. This approach limits the burden of multiple testing if all genome-wide variants were assessed. Depending on the assay, the expression of all genes (~22,000) can be tested for cis-QTLs.

Early studies showed heritability of gene expression, chromatin modifications and transcription factor binding and identified that eQTLs (SNPs associated with gene expression variation) were fairly widespread, with some observations of up to 30% of genes having an eQTL in lymphoblastoid cell lines (LCL) (Stranger et al., 2007, Price et al., 2011, Grundberg et al., Pickrell et al., 2010, Montgomery and Dermitzakis, 2011, McDaniell et al., 2010, Pai et al., 2015). With increasing sample sizes and denser genotypes or sequenced data, the number of discovered eQTLs has increased. For example, the latest G. TEx analysis of RNA-seq gene expression across 44 tissues with 449 donors identified 152,869 cis-eQTLs for 19,725 genes corresponding to 50.3% and 86.1% of all known lincRNA and protein-coding genes respectively (G. TEx Consortium, 2017).

Cis-eQTLs are enriched at gene start sites and variants upstream of the TSS are observed to have greater effect sizes than those in gene bodies, suggesting that SNPs regulating transcription have a larger impact than those that may regulate post-transcriptional processes (G. TEx Consortium, 2017). However, splice site QTLs or those that introduce a stop codon do have a high impact on downstream consequences (G. TEx Consortium, 2017). Early eQTL studies demonstrated high cell-type specificity, Dimas *et al.* (2009)

identified that 69-80% of eQTLs across three cell types, LCLs, primary fibroblasts and umbilical T cells were cell type specific (N = 75) (Dimas et al., 2009). Similar tissue specificity has been later confirmed in primary cell types (Chen et al., 2016a). Cell-type specificity can also manifest as opposing direction of effects of the same QTLs in different contexts. For example, Raj *et al.* (2014) identified 7000 shared eQTLs between monocytes and T cells (Raj et al., 2014). The effect size for most eQTLs, defined as the most significant SNP per gene, was concordant across the two cell types but for 42 genes, the most significant SNP had opposing directions where the allele with increased expression in one cell and decreased in the other (Raj et al., 2014). QTL studies in stimulated cell types have shown that context specificity not only applies to different cell types but also to different active states. Specific QTLs were only detected in activated immune cells when stimulated by, for example, bacterial components (LPS) or inflammatory cytokines (IFN-γ) (Fairfax et al., 2014, Naranbhai et al., 2015, Kim-Hellmuth et al., 2017, Alasoo et al., 2017).

eQTLs are often used to integrate with GWAS SNPs to identify gene targets. Zhu *et al* (2016) used a Mendelian randomization method adapted for summary statistics to analyse complex trait and disease GWAS and blood eQTL data (N = 5311) and subsequently identified 126 loci for which there was evidence of pleiotropy between gene expression and complex trait variance (Zhu et al., 2016). Here, pleiotropy describes genetic loci associated with two traits that may not be linked via a causal mechanism where the variant affects a phenotype through an endophenotype such as gene expression. Importantly, for approximately 60% of the colocalised cases, the regulated gene target, as identified by an eQTL, was not the nearest gene to the sentinel GWAS SNP (Zhu et al., 2016). Therefore, identifying gene targets based on proximity may lead to incorrect assignment.

QTL studies also allow the integrated study of genetic effects on gene expression, chromatin and TF binding, which has provided many insights into the mechanism of gene regulation. 55% of eQTLs in LCLs overlapped with DNase I hypersensitivity QTLs marking open chromatin, suggesting that a subset of eQTLs may influence gene expression through disruption of chromatin modification or transcription factor binding (Degner et al., 2012). Three studies measuring chromatin state, modification, TF binding and Pol II occupancy, provided initial evidence of high variability in enhancer function as well as suggesting that TF binding was the primary mechanism underlying modification of regulatory chromatin (Table 1.1) (Kasowski et al., 2013, McVicker et al., 2013, Kilpinen et al., 2013). These observations were confirmed and expanded by two recent studies in LCLs that assayed genome-wide binding of histone modifications, PU.1 and Pol II binding (Grubert et al., 2015, Waszak et al., 2015). Both studies showed extensive local correlation of molecular features in defined genomic windows (< 1Mb), which Waszak *et al.* (2015) referred to as variable chromatin

modules (VCMs). Interestingly, SNP-mediated changes in the local chromatin state were also correlated with those observed more distally in regions located up to 200 kb away (Denker and de Laat, 2015). This coordination was shown to result from physical interaction; Grubert *et al.* (2015) showed that 15% of proximal hQTLs were associated with changes at distal histone modifications that were connected by long-range chromatin interactions by using HiC and ChIA-PET data. Distal hQTLs were enriched within TADs and the majority of local-distal QTL pairs occurred between different enhancers (Grubert et al., 2015, Koch, 2015). Both studies provided evidence that TF activity underpinned chromatin variation, which in turn correlated with gene expression, in 99% of cases positively (Waszak et al., 2015). A single genetic variant could, therefore, propagate to multiple correlated features, perhaps explaining why a degree of chromatin variation cannot be correlated with proximal effects (Waszak et al., 2015). Single disease SNPs could therefore disrupt an entire coordinated molecular system, supporting the use of epigenomic data including chromatin interactions in identifying disease mechanisms and target genes and thus demonstrating the power of QTL studies to provide medically relevant insights as well as improve our understanding of genomic regulation (Koch, 2015, Denker and de Laat, 2015).

Chen *et al.* (2016) extended these efforts by assaying gene expression, splicing, DNA methylation, H3K4me1 and H3K27ac QTLs across multiple primary human cell types; monocytes, $CD4^+$ T cells and neutrophils (Chen et al., 2016a). Of the 20,403 genes assessed across the three cell types between 33.9-39.3% of genes had an eQTL. Here, an average of 9.89% of methylation probes, 25.7% of H3K4me1 peaks and 11.5% of H2K27ac peaks had at least one QTL. Confirming previous observations, there was a high degree of cell-type specificity to all marks (Dimas et al., 2009, Chen et al., 2016a). Particularly, hQTLs were highly cell specific, as expected for enhancer function. By considering lead SNPs and those in high LD ($r^2 \geq 0.8$), ~43.4% of eQTLs were also hQTLs, confirming previous observations of high correlation between chromatin and gene expression. For 18.4% of the genes, a splicing QTL effect was identified, but these were largely independent of eQTLs, shown by a low concordance of lead QTLs for the respective traits ($r^2 < 0.1$). There was also a high degree of colocalisation with autoimmune disease, which is discussed in detail in Chapter 2. The details of key QTL studies are summarised in Table 1.1.

In summary, observations from QTL studies and genome-wide approaches discussed above both support the role of key TFs underpinning chromatin state effects and gene expression, at least for a subset of sites. For regulatory QTLs that cannot be explained by TF binding or correlated with gene expression effects, it remains to be shown whether these effects could be explained by disruption of long-range interactions or whether there is extensive redundancy between enhancers removing downstream consequences of genetic disruption.

Author	Cell type	Stimulated/Resting	Molecular Trait	Trait Assay	Cohort
(Dimas et al., 2009)	LCL	Resting	Gene expression	Microarray	75
(Kasowski et al., 2010)	LCL	Resting	NF-kB, Pol II	ChIP-seq	10
(Maranville et al., 2011)	LCLs	Glucocorticoids	Gene expression	Microarray	114
(Degner et al., 2012)	YRI LCL	Resting	Open chromatin	DNase-seq	70
(Barreiro et al., 2012)	Dendritic cells	*M.tuberculosis*	Gene expression	Microarray	65
(Westra et al., 2013)	Whole blood	Resting	Gene expression	Microarray	5311
(Battle et al., 2014)	Whole blood	Resting	Gene expression	RNA-seq	922
(Lappalainen et al., 2013)	LCL	Resting	Gene expression, miRNA	RNA-seq	452-462
(Kasowski et al., 2013)	LCL	Resting	H3K27ac, H3K4me1, H3K4me3, H3K36me3 and H3K27me3, CTCF, SA1	ChIP-seq	19
(Kilpinen et al., 2013)	LCL	Resting	H3K4me1, H3K4me3, H3K27ac, H3K27me3, TFIIB, Pu.1, MYC, Pol II	ChIP-seq	8 + 2 trios
(McVicker et al., 2013)	YRI LCL	Resting	H3K4me1, H3K4me3, H3K27ac, H3K27me3, Pol II	ChIP-seq	10
(Ding et al., 2014)	CEU LCL	Resting	CTCF	ChIP-seq	51
(Raj et al., 2014)	$CD4^+$ T cell, Monocytes	Resting	Gene expression	Microarray	461
(Fairfax et al., 2014)	Monocytes	LPS (2h), LPS (24h), IFNγ (24h)	Gene expression	Microarray	262-414
(Lee et al., 2014)	Dendritic cells	LPS (5hr), influenza (10hr), IFNβ (6.5hr)	Gene expression	Microarray	534
(Naranbhai et al., 2015)	Neutrophils	Resting	Gene expression	Microarray	101
(Kumasaka et al., 2016)	CEU LCL	Resting	Open chromatin	ATAC-seq	24
(Caliskan et al., 2015)	PBMCs	Rhinovirus	Gene expression	Microarray	98
(Waszak et al., 2015)	CEU LCL	Resting	PU.1, Pol II, H3K4me1, H3K4me3, H3K27ac	ChIP-seq	47
(Chen et al., 2016a) BLUEPRINT	Monocytes, neutrophils, $CD4^+$ T cells	Resting	H3K27ac, H3K4me1, gene, splicing, methylation	ChIP-seq, RNA-seq, 450K	Up to 197
(Joehanes et al., 2017)	Whole blood	Resting	Gene and exon expression	Microarray	5257
(Kim-Hellmuth et al., 2017)	Monocytes	LPS, MDP, 5'-ppp-dsRNA (90min, 1 h)	Gene expression	Microarray	134
(Alasoo et al., 2017) (preprint)	iPSC differentiated macrophages	IFNγ (18h), *Salmonella* (5h), IFNγ + *Salmonella*	Gene expression, chromatin accessibility/open chromatin	RNA-seq, ATAC-seq	86, 42
(G. TEx Consortium, 2017)	44 Multiple tissues	*post mortem*	Gene expression	RNA-seq	449
Watt et al., 2018 (in preparation)	Neutrophils	Resting	H3K4me3, H3K27me3, PU.1, CEBPB, CTCF	ChIP-seq	22-110

Table 1.1: Summary of key blood quantitative trait loci studies

1.5 Functional, cellular and immune phenotypes

Beyond molecular phenotypes, heritable genetic variation has been observed in cellular and functional phenotypes. Examples include the levels of a broad range of blood cell types and surface receptor expression levels quantified using FACs-based immunophenotyping (Orru et al., 2013, Roederer et al., 2015) as well as cytokine production and circulating cytokine levels (Brodin et al., 2015, Ahola-Olli et al., 2017). These additional phenotypes allow comprehensive insights into immune functions and disease risk.

The Human Functional Genomics Project (HFGP) has collated an array of deeply phenotyped individuals with information such as microbiome composition, immune responses against human pathogens and disease status (autoimmune, diabetes, Lyme's disease, gout) (Netea et al., 2016, Li et al., 2016b). Li *et al.* (2016) demonstrated how host genetics plays a major role in the variation of immune cell cytokine responses from either whole blood, peripheral blood mononuclear cells (PBMCs) or macrophages stimulated *ex vivo* in a healthy population (Li et al., 2016b). Interestingly, the authors observed that the cytokine with the strongest inter-individual variation was IL6. Variants in this pathway have been previously associated with a multitude of diseases (Chapter 2). This further supports the functional importance of this cytokine in immune responses. In total, 17 novel genome-wide significant QTLs were associated with the production of mostly monocyte- or T cell-specific cytokines. cQTLs were enriched in regions under selective pressure, in ENCODE monocyte-specific enhancers, in infectious disease SNPs (for monocyte-derived cytokine QTLs) and in autoimmune disease SNPs (for T cell-derived cQTLs) (Li et al., 2016b). Similar autoimmune disease- and complex trait- loci enrichments were identified using 27 SNPs associated with circulating levels of 41 different cytokines from an independent GWAS in a large healthy cohort of up to 8,293 Finnish individuals (Ahola-Olli et al., 2017). Continuing on the efforts to measure protein-level traits, 38 variants were associated with immunoglobulin levels (IgA, IgG, IgM), which are effector molecules of the adaptive immune system (Jonsson et al., 2017). Similarly, these variants also had known roles in autoimmune diseases and haematopoietic malignancies.

An exemplary study demonstrated how the combination of multiple pieces of genetic, molecular and functional evidence can resolve complex autoimmune disease risk loci, in this case, the *TNFSF13B* gene locus encoding the cytokine B cell activating factor (BAFF) (Steri et al., 2017). An indel variant was associated with multiple sclerosis and systemic lupus erythematosus in a Sardinian cohort, as well as with 18 different endophenotypes including B cell and monocyte counts (Steri et al., 2017). The variant produces an alternative polyadenylation site and a 3' UTR truncated transcript, which resulted in both a gene expression and protein translation effect, the latter due to the presence of fewer miRNA

binding sites. Ultimately this culminated in an increased level of soluble BAFF. Elevated BAFF levels were observed prior to disease diagnosis in separate preclinical samples, which was clear evidence of the causal relationship between higher BAFF protein levels and autoimmune disease (Steri et al., 2017). This clearly shows the power of combining functional and molecular phenotypes with longitudinal and clinical datasets when evaluating causal relationships between functional and disease phenotypes.

There are more and more studies recognising the importance of multiple phenotypes in facilitating functional interpretation of GWAS loci and in providing basic biological insights. Very recently, the Hi-HOST Phenome Project have generated a catalog cellular GWAS associations using 79 phenotypes in response to live pathogens in 528 LCLs and identified 17 genome-wide significant loci (Wang et al., 2017a). The cellular phenotypes measured included readouts of endocytosis, endosomal trafficking, cell signalling, cell death, cytokine production as well as the molecular readouts of transcriptional regulation (Wang et al., 2017a). In addition, the Enhancing GTEx (eGTEx) project was recently announced, wherein a bid to describe the effect of variation from “molecule to individual”, other intermediate measurements such as protein expression and telomere length will be assayed in the wide range of tissue types from this project in addition to gene expression and molecular phenotypes (eGTEx Project, 2017).

In future, similar efforts will likely be extended to multiple primary cell types and greater sample sizes providing rich resources for functionally annotating genetic loci.

1.6 Recall-by-genotype studies

Recall-by-genotype (RbG) studies are genotyped-directed experimental phenotyping investigations representing downstream hypothesis-driven approaches to investigate functional mechanisms (Corbin et al., 2017). They have emerged as the primary choice for designing experiments to further investigate the function of observations first identified in large-scale genetic studies. They allow greater functional resolution with smaller sample sizes compared to hypothesis-free GWAS (Figure 1.7).

RbG test a small number of predicted causal variants (between 1 and 10) selected from the integration of GWAS-associated variants, functional studies and statistical methods such as fine-mapping. Similar to GWAS, RbG studies have the advantage of utilising genetic variants that have arisen from the random allocation of alleles at conception, which cannot, in turn, be influenced by the traits of interest (Section 1.7.1). A further advantage of RbG studies are that they are designed to query causal relationships in selected stratified groups based on previously observed biological associations. This increases the precision of functional insight

in a cost-effective, efficient manner (Corbin et al., 2017). I demonstrate the implementation and utility of a RbG study in Chapter 4.

1.7 Haematopoiesis as a paradigm for genetics

Haematopoiesis is the production of all mature blood cell types including thrombocytes (platelets), erythrocytes (red blood cells), myeloid cells (monocytes, macrophages, neutrophils) and lymphocytes (B cells and T cells) (Figure 1.5). Self-renewing haematopoietic stem cells (HSC) in the bone marrow differentiate to lineage-committed progenitor cells, which further differentiate into mature cells (Orkin and Zon, 2008, Vasquez et al., 2016). Chromatin regulation is important in this differentiation process and mutations in factors mediating histone modification and chromatin architecture result in myeloid malignancies (Woods and Levine, 2015). Chromatin was recently shown to be highly dynamic during lineage specification with 17,035 enhancers established *de novo* mainly after commitment of the first lineage progenitor (Lara-Astiaso et al., 2014). TFs are key to the activity of these enhancers, full activation of which preceded lineage-specific gene expression programmes (Lara-Astiaso et al., 2014). Therefore, haematopoiesis represents a model system for the study of all stages of stem cell development as well as chromatin formation, transcription factory activity and the cell-type specificity of these processes.

Mature haematopoietic cells perform vital biological roles including oxygen transport (red blood cells), blood clotting (platelets) and immune responses (myeloid and lymphoid cells). Sustained haematopoiesis occurs under homeostatic conditions as well as during infection, (Orkin and Zon, Amulic et al., 2012). Dysregulated blood cell function is a known factor in the aetiology of a wide variety of diseases. Understanding the biological context of disease-dysregulated processes can highlight important haematopoietic pathways and novel genes in haematopoiesis and mature cell function. The role of these cells in disease and function is discussed in detail in Chapter 2 and 3 of this thesis.

Haematopoiesis and mature blood cells are both relatively experimentally tractable. Whole blood is easily accessible from a high number of individuals and from this specific cell populations can be isolated with high purity and relative technical ease. The evolutionary conservation of haematopoiesis also facilitates study in model organisms. As a result, haematopoiesis is one of the best-characterised mammalian cellular differentiation systems.

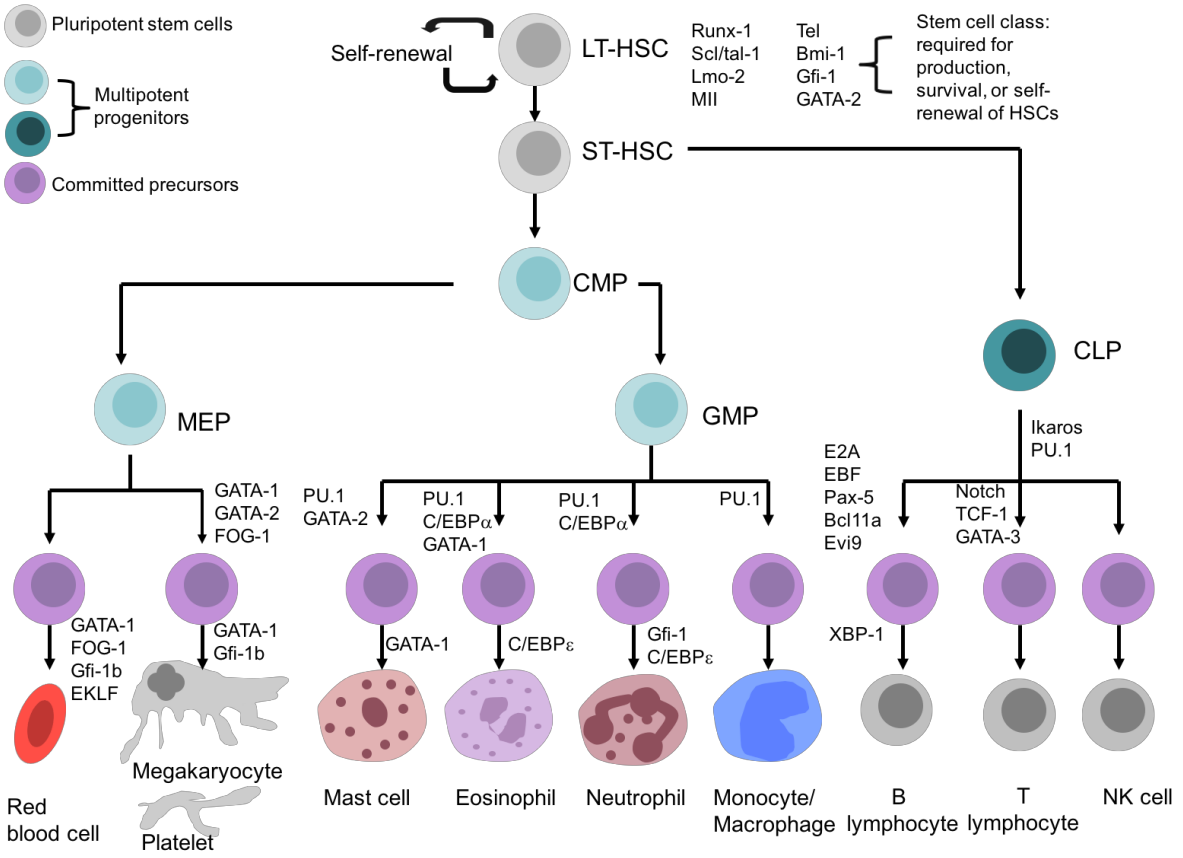

Figure 1.5: Haematopoiesis and the involvement of essential transcription factors
Differentiation of self-renewing haematopoietic stem cells to form all mature cells (red blood cell, platelet, mast cell, eosinophil, neutrophil, monocyte and macrophage, B and T lymphocytes, NK cells). The transcription factors required for each stage were discovered using conventional gene knockouts that resulted in a blockage of haematopoietic differentiation. LT-HSC: long-term haematopoietic stem cell, ST-HSC: short-term haematopoietic stem cell, CMP: common myeloid progenitor; CLP: common lymphoid progenitor, MEP: megakaryocyte/erythroid progenitor, GMP: granulocyte/macrophage progenitor. Additional TFs, not shown here, were predicted using a highly sensitive ChIP-seq protocol to be involved in 16 differentiation stages (Lara-Astiaso et al., 2014). Figure adapted from (Orkin and Zon, 2008).

Blood cell phenotypes such as full blood counts (FBC), are also readily measured by automated haematology analysers (Chami and Lettre, 2014, Astle et al., 2016). The deviation from normal size, physical characteristics or number of blood cells is diagnostic for human disease such as infection, anaemia, thrombotic diseases or haematological disorders (Table 1.2) (Vasquez et al., 2016, Soranzo et al., 2009). FBC is therefore routinely measured as part of clinical diagnosis and assessment of general health (Chami and Lettre, 2014). Table 1.2 summarises the full range of phenotypes that can be measured with recent analysers such as the Sysmex system (Astle et al., 2016, Vasquez et al., 2016, Sysmex Corporation).

Blood cell traits vary across healthy individuals and part of this variation is due to genetic factors (Pilia et al., 2006, Evans et al., 1999, Garner et al., 2000, Chami and Lettre, 2014). Therefore, studying naturally occurring genetic variation of circulating mature blood cell counts is a common and successful strategy used to gain insight into the regulation of haematopoiesis (Table 1.2). This approach has yielded many insights, not only in identifying novel haematopoietic regulators but also for the wider field of human genetics. For example, blood GWAS has been successful in identifying novel regulators of haematopoiesis (Gieger et al., 2011, van der Harst et al., 2012, Bielczyk-Maczynska et al., 2014). Previously unknown genes identified from GWAS of RBCs and platelets displayed haematopoietic phenotypes in model organisms (Vasquez et al., 2016).

Up until 2016, blood GWAS only explained a fraction of variation in the population (4-10%) and high-powered cohorts for studying myeloid and lymphoid parameters were lacking (Vasquez et al., 2016, Gieger et al., 2011, van der Harst et al., 2012). The recent large GWAS using data from the UK biobank cohort (N = 173,480) investigated a high number of traits, 36 in total (Table 1.2) (Astle et al., 2016). 2,706 independent variants were identified, representing a ten-fold increase in the number of known loci that included hundreds of rare variants with high effects sizes (Vasquez et al., 2016, Kim-Hellmuth and Lappalainen, 2016). Most of the sentinel variants were highly specific across red blood cell, white cell and platelet traits and enriched in corresponding cell-type specific enhancers. Coding variants were enriched with Mendelian disease mutations, a demonstration of how important clinical insight can be gleaned from large-scale GWAS. Plausible molecular mechanisms were identified through integration with the BLUEPRINT QTL data for 276 blood trait variants that colocalised with at least one molecular QTL (Astle et al., 2016). It was estimated that a higher proportion of variance in the blood indices was explained by the common autosomal genotypes from this study, for example between 5-21% of variance in white cell traits (Astle et al., 2016). The full UK Biobank cohort of 500,000 individuals could identify further significant variants explaining trait variance (Collins, 2012).

	Trait [Units]	Description	Determination	Example Diseases/disorders
RBC	Red blood cell count [per pL]	Count of RBCs per unit volume of blood	Impedance (measured)	Anaemia, polycythemia vera
HGB	Haemoglobin concentration [g/dl]	Concentration of Hb per unit volume of blood	Light absorbance (measured)	
HCT	Hematocrit [%]	Volume fraction of blood occupied by red cells	Impedance (measured)	
MCV	Mean corpuscular haemoglobin concentration [fL]	Mean volume of RBCs	(HCT/RBC)×10 (derived)	
RDW	Red cell distribution width [fL]	Coefficient of variation of red cell volume distribution	CV of impedance measured red cell volume distribution (measured)	
MCH	Mean corpuscular haemoglobin [pg]	Average mass of Hb per red cell	(HGB/RBC)×10 (derived)	
MCHC	Mean corpuscular haemoglobin concentration [g/dL]	Concentration of Hb per unit of volume occupied by red cells	(HGB/HCT)×100 (derived)	
PLT	Platelet count [per nL]	Count of platelets per unit volume of blood	Impedance (measured)	Essential thrombocythemia, thrombotic Thrombocytopenic purpura
MPV	Mean platelet volume [fL]	Mean volume of platelets	(PCT/PLT)×10000 (derived)	
PDW	Platelet distribution width [fL]	Spread of the platelet volume distribution (PDV)	Impedance: Coefficient of variation of PDV (measured)	
PCT	Plateletcrit [%]	Volume fraction of blood occupied by platelets	Impedance (measured)	
WBC	White blood cell count [per nL]	Aggregate count of white cells per unit volume of blood	Impedance (measured)	Autoimmune/immunological, infection, inflammation, leukaemia
NEU	Neutrophil count [per nL]	Count of neutrophils per unit volume of blood	(NEUT%×WBC)/100% (derived)	Myelodysplasia, bacterial infections
LYM	Lymphocyte count [per nL]	Aggregate count of lymphoid cells per unit volume of blood	(LYMPH%×WBC)/100% (derived)	Lymphoma, viral infections
MON	Monocyte count [per nL]	Count of monocytes per unit volume of blood	(MONO%×WBC)/100% (derived)	Myelomonocytic leukaemia, chronic infections (tuberculosis)
EOS	Eosinophil count [per nL]	Count of eosinophils per unit volume of blood	(EO%×WBC)/100% (derived)	Allergies, asthma, parasitic infections
BAS	Basophil count [per nL]	Count of basophils per unit volume of blood	(BASO%×WBC)/100% (derived)	Hyperthyroidism, myeloproliferation disorders

Table 1.2: Summary of the main haematological indices, measurement unit and related disorders
Adapted from (Vasquez et al., 2016, Astle et al., 2016). Additional traits were also tested in the Astle *et al.* (2016) GWAS, that included for example the percentage of granulocytes that is made up by neutrophils. I list the main traits measuring mature blood cell counts here that are routinely measured and have been explored in previous studies.

1.7.1 Genetics, correlation and causation

Correlation between blood indices and increased risk of certain diseases such as obesity, stroke and cardiovascular diseases has been observed (del Zoppo, 1998, Poitou et al., 2011, Ensrud and Grimm, 1992, Hoffman et al., 2004, Boos and Lip, 2007). However, correlation does not necessarily show causation as epidemiological and observational relationships can be subject to confounding factors, measurement error, bias or reverse causation (where the disease state influences the endophenotype such as blood indices).

Genetics, with the exception of somatic mutations, is pre-determined at birth where variants are segregated randomly and independently of other traits (Evans and Davey Smith, 2015). In this way, confounding and reverse causation are both reduced as genetics precedes any biological effect or outcome (Evans and Davey Smith, 2015). We can also measure genetic variants with high precision, reducing measurement error that can occur in observational studies. Approaches have therefore been developed that use genetic variants (instrumental variables) that are known to influence a biological intermediate (exposure), which itself affects disease risk. In this case, the studied variants should also be related to the risk of the disease. This approach can assess both the causality of biological intermediates and quantify the size of the causal effect and is referred to as Mendelian Randomization (Evans and Davey Smith, 2015). There are certain assumptions that must not be violated in these analyses, which in some cases can be challenging to definitively confirm. These are summarised in Figure 1.6.

This approach was implemented by Astle *et al.* (2016) to test for causal relationships between blood indices and each of a group of six autoimmune, three cardiometabolic and five neuropsychiatric diseases. Positive correlations were found between eosinophil count and rheumatoid arthritis and asthma, with a weaker effect between neutrophil indices and asthma. Interestingly, there was a reduced likelihood for causality between red blood cell, white blood cell, granulocyte and neutrophil counts and risk of coronary heart disease (CHD), despite previously reported correlations (Wheeler et al., 2004, Astle et al., 2016).

Overall, studying the process of haematopoiesis and mature cell function increases our understanding of basic biology. Concomitantly, it also offers the potential to use blood cell traits as disease biomarkers and tractable intermediate phenotypes in genetic studies and functional follow-ups.

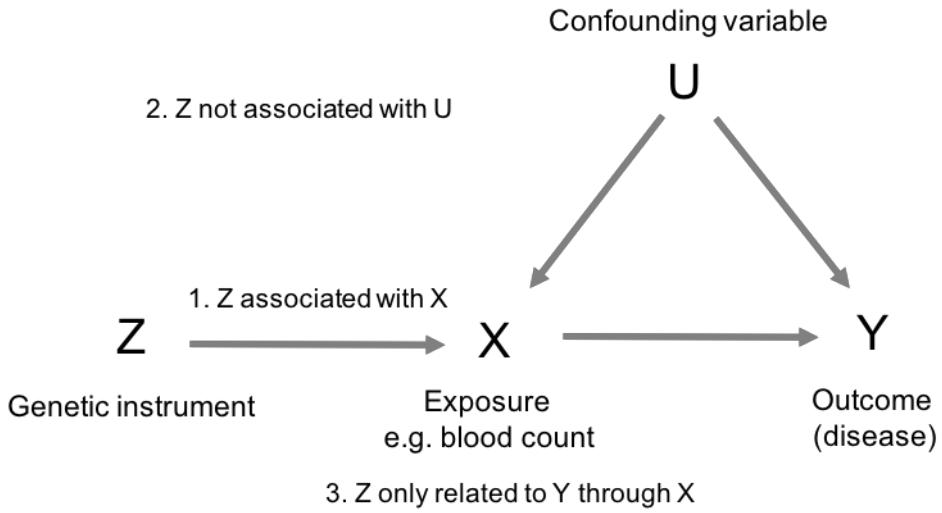

Figure 1.6 Mendelian randomization methodology and assumptions
Schematic summarising a causal relationship between an exposure and an outcome/disease assessed by using genetic variants (Z) that are associated with the exposure and under causality are also associated with disease. Causal relationships are depicted with arrows. The three assumptions are also given. Adapted from (Evans and Davey Smith, 2015).

1.8 Aims of this thesis

In this thesis, I use a combination of genetic and genomics approaches I have discussed to resolve functional consequences of genetic variation whilst also understanding the biology of haematopoietic cells. These are summarised in Figure 1.7.

In Chapter 2, I discuss how these approaches have increased our understanding of autoimmune diseases. I apply the lessons learnt from these studies to diseases that are not traditionally classified as immune-mediated. I use epigenomic phenotypes to resolve mechanisms of risk loci and also explore potential insight into pathways or genes that could provide future therapeutic avenues for these diseases. I demonstrate that the combination of genomic and genetic approaches provides hypothesis-free identification of genes and pathways dysregulated in disease, representing an early step in identifying new therapeutic avenues.

Following from this, I apply GWAS to novel neutrophil phenotypes with an overall aim of expanding the phenotype repertoire by providing additional functional datasets with which to annotate trait- or disease- associated loci. Finally, I implemented a recall-by-genotype study to perform an in-depth investigation into two genetic loci where there was previous evidence of an association with neutrophil count. Throughout my thesis, I demonstrate the application of varied but complementary approaches in gaining biological insight from genetic associations.

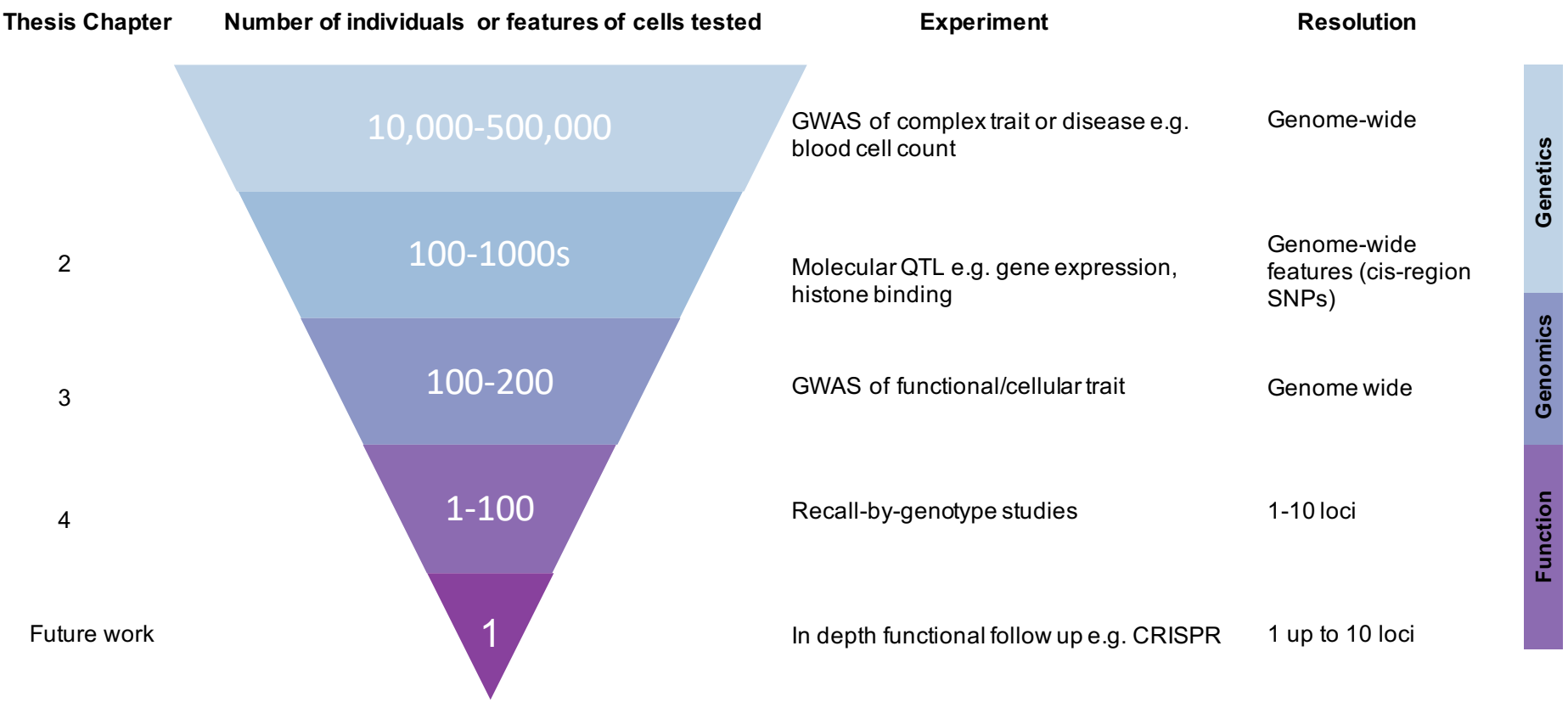

Figure 1.7 Approaches to investigate functional mechanisms of genetic variants
Schematic summarises the type of experiments, number of individuals required, resolution of variants investigated and the chapters of this thesis where the techniques are used